MedSync: Data Integration & Predictive Health Analysis

[1]Gaurav Choughule, [2]Yash Sodaye, [3]Tejas Shelke, [4]Sahil Wani, [5]Nilesh Bhelkar
[1]UG Scholar, [2]UG Scholar, [3]UG Scholar, [4]UG Scholar, [5]Assistant Professor
[1]Department of Artificial Intelligence & Data Science,
[1]Rajiv Gandhi Institute of Technology, Mumbai, India
[1]gauravchoughule7@gmail.com, [2]yashsodaye13@gmail.com, [3]tejas.shelkee@gmail.com,
[4]sahilwani31103@gmail.com, [5]nilesh.bhelkar@mctrgit.ac.in

***Abstract*—There are considerable difficulties in the management and digitization of medical data within the healthcare industry, especially concerning physical prescriptions and patient records. MedSync is a cutting-edge platform aimed at making this process more efficient by utilizing Optical Character Recognition (OCR) technologies, such as Amazon Textract, to extract text from medical documents with precision. By incorporating machine learning methods like Cosine Similarity and TF-IDF, the system improves data organization and accessibility while offering users pertinent information about medicines, such as their composition, usage instructions, side effects, and ratings. Moreover, MedSync includes a chatbot driven by AI for patient questions, which enhances the accessibility of healthcare information. With its predictive analysis component, the system can identify potential health risks early on by examining historical medical records. MedSync seeks to improve the efficiency, accuracy, and accessibility of healthcare data management by implementing these technologies. The methodology, results, and effects of MedSync in the digitization and standardization of medical records are discussed in this paper, with a view to enhancing patient care and facilitating data-driven decision-making.**

***Keywords*— Medical Data Digitization, Optical Character Recognition (OCR), Amazon Textract, Machine Learning, Cosine Similarity, TF-IDF, Predictive Analysis, Healthcare Chatbot, Data Standardization, Electronic Health Records (EHR).**

I. Introduction

Data science is a rising and trending field of technology in recent years. Under data science comes data cleaning, storing, management and similar operations. Data is very valuable irrespective of the domain; whether it's in IT, business, education, medicine, security etc. Out of these, medicine or healthcare is a very volatile field where data management should be done precisely since the information obtained and stored here is very delicate. Moreover, this data voluminous and it becomes a hassle to administer it without the risk of losing it.

With the rise of Artificial Intelligence & Machine Learning, their application is widely observed in various sectors. Machine Learning and Data Science go hand-in-hand. It contributes by predicting future outcomes on the basis of past data. This is called predictive analysis. Machine learning methods are highly used for predictive analysis. Domains like construction, botany, education, social service & medicine are the ones where predictive analysis is frequently used [2]. Healthcare is the most prominent or vital field of application. Thus, prediction and automation in analysis of data is a contribution of ML that strongly supports healthcare sector. Patients usually obtain their medical data like prescriptions or lab reports in physical format. To keep a record of this data, falls under the same issue of data management as mentioned previously. To tackle this issue, automation and proper data management is essential. This solution is efficiently provided by our project MedSync.

MedSync is a healthcare platform with the primary objective of handling heaps of physical medical data like prescriptions. This is achieved by digitizing the prescriptions from paper to textual data. Digitization also helps in easier storage and management of medicinal records. Further, the system provides the general information of the prescribed medicines like its composition, uses, side-effects and ratings. This helps the patient to easily remember and understand the medicine's effects far after a doctor or hospital visit. For general queries, the system also has a chatbot which clears any medical concern or doubts of the patient.

This paper is based on our research and implementation in the process of achieving our final system – Medsync. The rest of the paper deals with the related work, literature survey, methodology, results and conclusion of the project.

II. Methods

OCR & Amazon Textract: By examining patterns in pictures, optical character recognition (OCR) technology transforms handwritten or printed text into a machine-readable format. Usually, it includes character segmentation, feature extraction, preprocessing (noise reduction, binarization), and machine learning model classification.

Amazon Textract is a cutting-edge OCR service that extracts text, handwriting, and structured data from documents using deep learning. It is much more efficient in processing complicated texts, like as prescriptions, than typical OCR because it can recognize key-value pairs, comprehend layouts, and extract tables. Textract uses context-aware text extraction and AI-driven handwriting recognition to improve accuracy.

Cosine Similarity: Cosine similarity is a metric that measures the similarity between two non-zero vectors by computing the cosine of the angle between them. It ranges from -1 (opposite) to 1 (identical) and is given by:

$$\cos(\theta) = \frac{A \cdot B}{|A||B|}$$

In MedSync, cosine similarity is used to match user-entered medicine names with the closest spelling in the dataset. By converting names into vector representations, the system identifies the most similar medicine and retrieves its usage, side effects, and ratings. This technique ensures accurate results even when users make spelling errors.

TF-IDF (Term Frequency-Inverse Document Frequency): This numerical statistic indicates the significance of a word within a document in a collection or corpus.

Term Frequency (TF): Assesses the frequency of a term's occurrence in a document.

Inverse Document Frequency (IDF): Assesses the rarity of a term throughout the whole corpus.

Each word in a document is assigned a weight by TF-IDF, with greater weights given to words that occur frequently in that document but are infrequent across the entire collection. This aids in recognizing the words that best exemplify what a document is about.

Predictive Analysis: Predictive analysis is to use historical data and based on it use statistical machine learning methods to predict future outcomes. In the context of healthcare sector, this technology can be used to detect the possible diseases or issues a patient might suffer in the future with respect to his previous health records and reports. However, good quality of data and precise models are essential for effective outputs of predictive analysis.

III. Related Work

Rathod and Sari (2020) have outlined their plan for digitizing health data in [1]. To digitize physical data, such as general medical records, they have adopted Handwritten Character Recognition (HCR). The goal was to automate the process of digitizing handwritten material so that databases could store and retrieve it with ease. The initiative effectively illustrated its goal, and by utilizing HCR, they have increased the accuracy and efficiency of health data management. However, HCR finds it difficult to deal with identification of characters in cursive script. This is the only flaw of the technique.

Brief studies on machine learning-based predictive analysis have been compiled by the authors in [2]. It was determined that the medical profession accounts for 56.7% of the use of machine learning techniques in predictive analysis. Additionally, supervised learning encompasses 70% of the techniques employed, with Random Forest (RF) being the most widely utilized technique.

In order to increase the effectiveness of administering an Electronic Health Record (EHR) system, the authors of [3] have suggested a system that integrates a 3D human body model for the visualization of health data and automates the process of digitizing paper-based medical information. The initiative makes use of OCR's advantages to digitize medical records. After testing more than 200 medical reports, a 100% success rate in digitizing medical records was attained. However, the model was only tested on medical records written in Chinese script.

Kadadi and Agrawal (2014) examined the difficulties of data integration and interoperability in big data in their research in [4]. The goal is to investigate data integration methods, deal with huge data complexity, and create an integration architecture to overcome these obstacles. In order to overcome these problems, the authors suggest a number of tools and technologies, including Hadoop, KARMA, and JNBridgePro. However, the effectiveness of the proposed solutions is uncertain due to lack of real-world case studies.

The authors of [6], Sinha and Jenckel (2019), examined the OCR models in an unsupervised way using Generative Adversarial Networks (GANs). By employing GAN-generated text line images that closely resemble the original input images from OCR output text, the experiment shows enhanced validation of OCR models. This allows the model to be assessed even in situations where the softmax layer is unavailable. Although, it is challenging to compare the generated images with the original input when such styling is crucial because the created images lacked information like font style (bold or italic, for example).

Saluja and Punjabi (2019), the authors of [7], have created a technique to enhance OCR mistake correction for Indic languages by utilizing sub-word embeddings to take into consideration intricate linguistic principles. The project trains a model that can fix OCR errors in Indic languages using sub-word embeddings, such as n-grams. In terms of F-scores and Word Error Rates (WER), the fastText-based model fared better than the baseline models, indicating that it was efficient at handling words that were not in the lexicon. However, the availability of high-quality dates has a significant impact on the results. Due to variations between the training and test sets, the approach also demonstrates an increase in WER for certain Malayalam terms that are not part of the vocabulary.

All the above papers demonstrate the use variety of technologies like OCR, HCR, GANs, Hadoop, FHIR for data digitization & standardization and use of machine learning algorithms for predictive analysis.

During the development timeline of the project, a few models were considered before finalizing the text digitization module of the project.

1. Tesseract OCR - Regularized Expression:

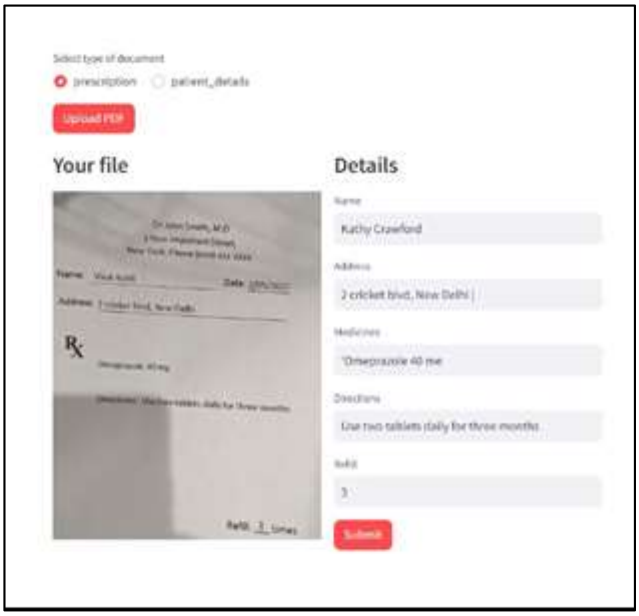
Working:

Input prescription is converted from pdf to Image using pdf2image library. Further image is preprocessed using OpenCV2 library. Then, text is extracted from the image by passing it through Tesseract OCR Engine. Using RegEx library, the text is categorized into classes like 'medicine_name', 'patient_name', etc. which is returned into JSON format. A FastAPI backend server managed the data extraction requests and returning JSON Object. In this way, the extracted text is stored in JSON format for convenient storage.

Drawbacks:

The text extraction process only worked accurately on prescriptions with a specific format. Thus, it results in limited scope and lack of scalability.

Output:



Select type of document

● prescription ○ patient_details

Upload PDF

Your file

Details

Name

Kathy Crawford

Address

2 cricket blvd, New Delhi |

Medicines

'Omeprazole 40 me

Directions

Use two tablets daily for three months

Refill

3

Submit

Figure 1. Tesseract OCR – RegEx Model Output

2. CNN-biLSTM Model:

Working:

Image is preprocessed using Pillow & OpenCV library. A pre-trained convolutional model is used to predict the prescription details from the text which assists the core Tesseract OCR.

Drawbacks:

The accuracy of the output was considerably low as the extracted text was observed to be ‘gibberish’.

The Output is observed in Figure 3.

Output:

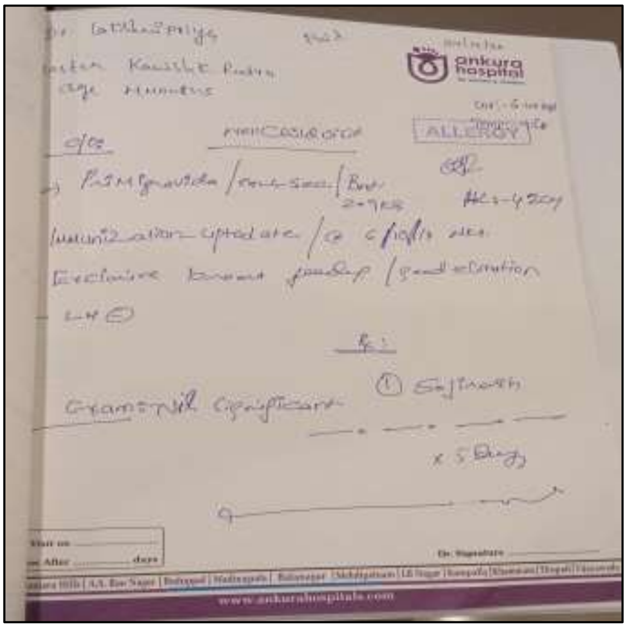
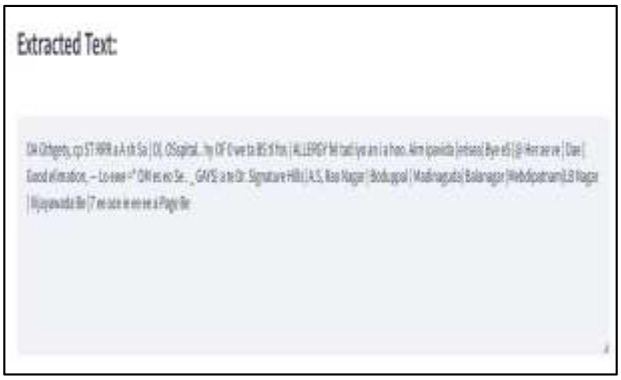

ankura hospital
ALLERGY
Kaushik
Immunization uptodate
Examination significant
Visit on
After days
Dr. Signature
www.ankurahospitals.com

Figure 2. Prescription A



Extracted Text:

Dk Orhgety, cp ST RRR a A ck Sa | Ol. OSgrital... hy OF Cwe ta 85 tt fm | ALLERGY M tad iye an i a heo. Alm ipavida |enkeas| Bye e5 | @ Hen ae ve | Dae | Eaod afimation, – Lo-eee =" OM es wo Se.. , GAYS: a te Dr. Signature Hills | A.S. Rao Nagar | Boduppal | Madinaguda| Balanagar |Mehdipatnam|LB Nagar | Vijayawada Be | 7 ws oce ie ee ee a Page Be

Figure 3. Inaccurate Output

3. Amazon Textract:

Working:

This is the model we resorted to in the end. It is simply an OCR-powered tool for text extraction. User inputs the prescription image and all the text from the image is precisely extracted.

Advantages:

Highest accuracy of results among all the models implemented during the timeline.

There is no requirement to build a model from scratch and maintenance of it.

IV. Methodology

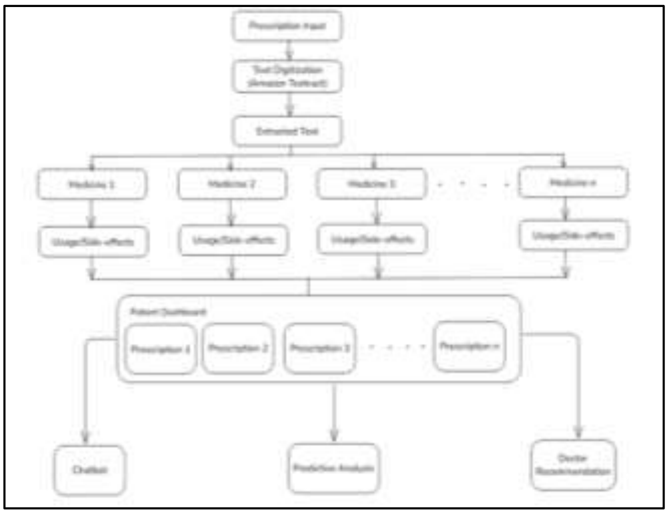


Figure 4. Flow Diagram

Figure 4 represents the flow diagram of MedSync. It can be divided into the following parts:

Text Digitization:

User or the Patient inputs his prescription by clicking its photograph. Thus, the uploaded prescription is in jpg, jpeg or heif format. Once uploaded, OCR powered Amazon Textract accurately extracts the textual information from the input image. This text is then displayed on an output window and is stored in the system.

Medicine Information Finder:

User inputs the name of any medicine either from his prescription or of his choice. The system compares the entered name with the medicine database using Cosine Similarity & TF-IDF and provides information on that medicine. This information includes medicine composition, uses, side-effects, manufacturer, ratings, etc.

Chatbot:

Users can ask any medical or health related query, which will be answered in an elaborate manner or as per the requirements of the user. The chatbot is powered by Groq API.

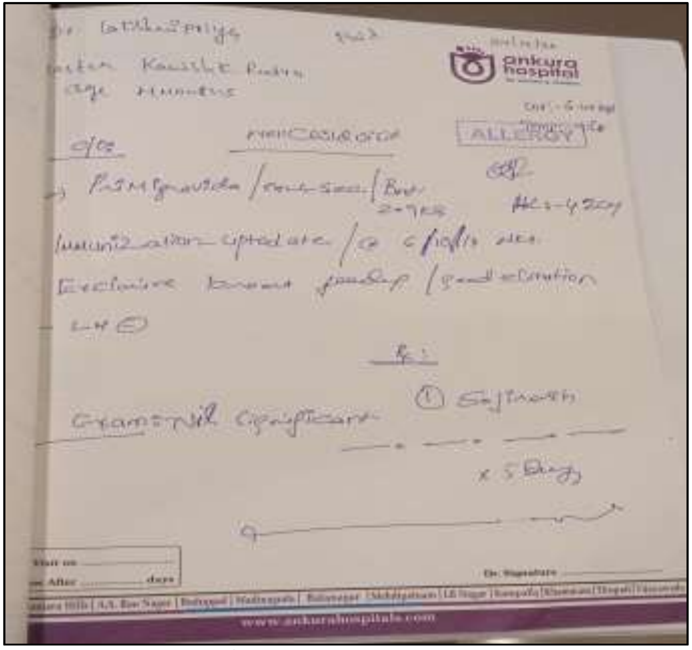
Doctor Recommendation:

User inputs his medical issue or disease along with his preferred location. The system provides a list of doctors with specialization in the given type of health issue and in the entered location of the patient.

V. RESULTS AND DISCUSSION

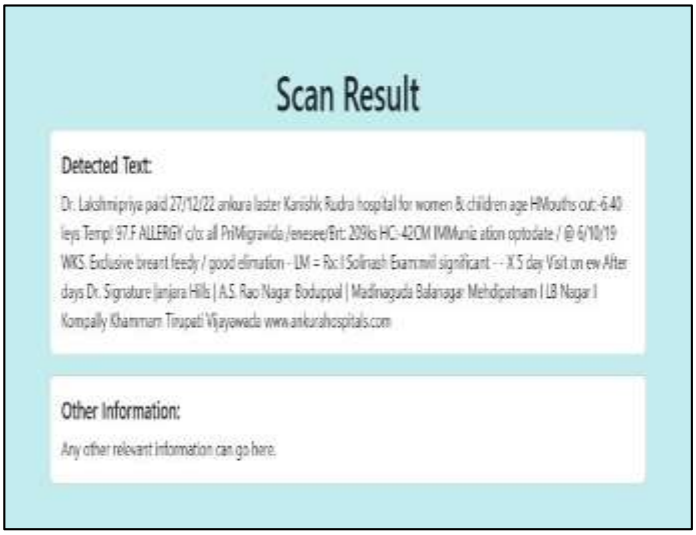
After the successful implementation of finalized techniques and tools, we obtained the following results:

Text Digitization:



Dr. Lakshmipriya
Patient Kanishk Rudra
Age 4 Months
ankura hospital
ALLERGY
c/o
→ Primigravida / [illegible] / Birth 2.9 kgs HC - 42cm
Immunization uptodate / @ 6/10/19 wks
Exclusive breast feeding / good elimination
LM
Rx:
① [illegible]
Examination significant
x 5 Days
Visit on
After ___ days
Dr. Signature
www.ankurahospitals.com

Figure 5. Prescription A



Scan Result

Detected Text:

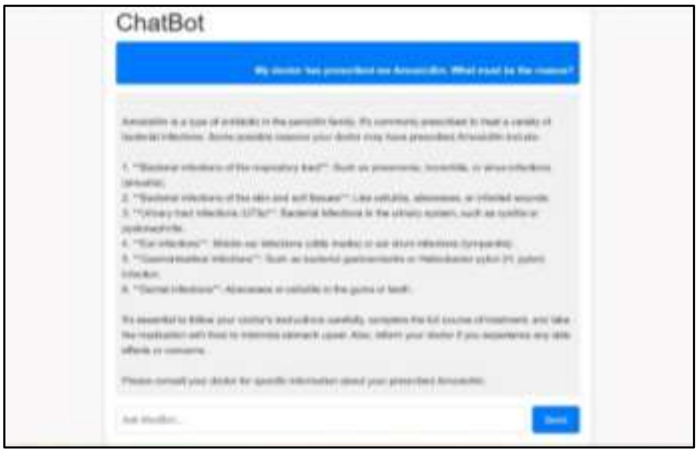
Dr. Lakshmipriya paid 27/12/22 ankura laster Kanishk Rudra hospital for women & children age HMouths cut:-6.40 leys Temp! 97.F ALLERGY c/o: all PriMigravida /enesee/Brt: 209ks HC:-42CM IMMuniz ation optodate / @ 6/10/19 WKS. Exclusive breant feedy / good elimation - LM = Rx: I Solinash Examnil significant - - X 5 day Visit on ew After days Dr. Signature Janjara Hills | A.S. Rao Nagar Boduppal | Madinaguda Balanagar Mehdipatnam I LB Nagar I Kompally Khammam Tirupati Vijayawada www.ankurahospitals.com

Other Information:

Any other relevant information can go here.

Figure 6. Digitized Text Using Amazon Textract

Chatbot:



ChatBot

My doctor has prescribed me Amoxicillin. What must be the reason?

Amoxicillin is a type of antibiotic in the penicillin family. It's commonly prescribed to treat a variety of bacterial infections. Some possible reasons your doctor may have prescribed Amoxicillin include:

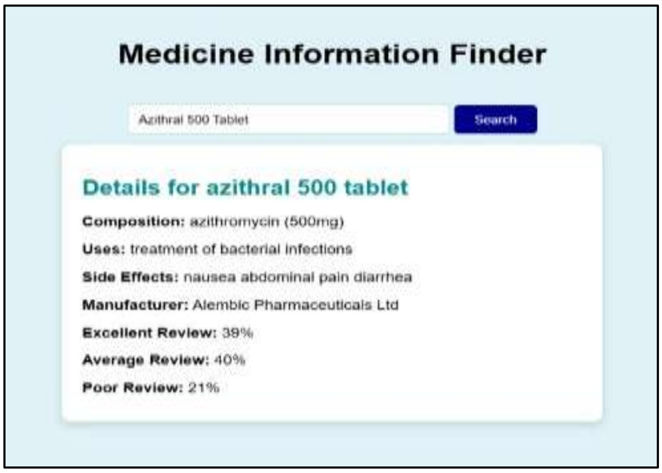
1. **Bacterial infections of the respiratory tract**: Such as pneumonia, bronchitis, or sinus infections (sinusitis).
2. **Bacterial infections of the skin and soft tissues**: Like cellulitis, abscesses, or infected wounds.
3. **Urinary tract infections (UTIs)**: Bacterial infections in the urinary system, such as cystitis or pyelonephritis.
4. **Ear infections**: Middle ear infections (otitis media) or ear drum infections (tympanitis).
5. **Gastrointestinal infections**: Such as bacterial gastroenteritis or Helicobacter pylori (H. pylori) infection.
6. **Dental infections**: Abscesses or cellulitis in the gums or teeth.

It's essential to follow your doctor's instructions carefully, complete the full course of treatment, and take the medication with food to minimize stomach upset. Also, inform your doctor if you experience any side effects or concerns.

Please consult your doctor for specific information about your prescribed Amoxicillin.

Figure 7. Chatbot

Medicine Information Finder:



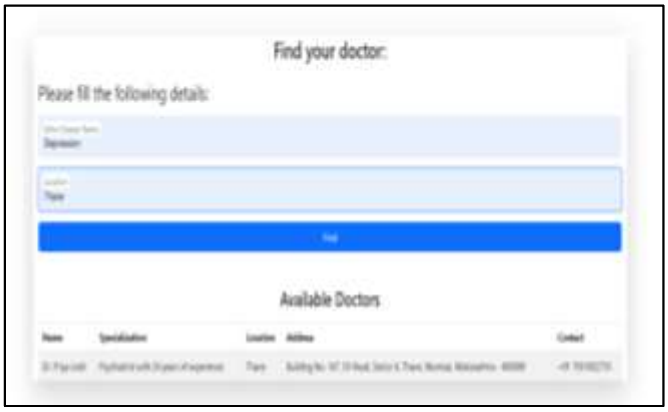
Medicine Information Finder

Details for azithral 500 tablet

Composition: azithromycin (500mg)
Uses: treatment of bacterial infections
Side Effects: nausea abdominal pain diarrhea
Manufacturer: Alembic Pharmaceuticals Ltd
Excellent Review: 39%
Average Review: 40%
Poor Review: 21%

Figure 8. Medicine Information

Doctor Recommendation:



Find your doctor:

Please fill the following details:

Enter Disease Name: Depression
Location: Thane

Available Doctors

Name	Specialization	Location	Address	Contact
Dr. [illegible]	Psychiatrist with [illegible] years of experience	Thane	Building No. [illegible], Thane, Mumbai, Maharashtra - [illegible]	+91 [illegible]

Figure 9. Doctor Recommendation

VI. CONCLUSION

The project MedSync acts as an easy and user-friendly platform for the management of medical prescriptions of patients along with its complementary features of medicine information finders, healthcare chatbot and doctor recommendation.
This is achieved after overcoming the shortcomings of the systems & techniques from the literature survey and initially implemented text digitization models.

REFERENCES

[1]. Rathod & Sarita. (2020). Converting non-digitized health data to digital format. Asian Journal of Convergence in Technology, 6(1), 10-13.

[2]. Loola B., Khadija, O. T., & Souissi N. (2020). Predictive analysis using machine learning: Review of trends and methods. 2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 1-6.

[3]. Liu, N., et al. (2020). A new data visualization and digitization method for building electronic health record. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2980-2982.

[4]. Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of data integration and interoperability in big data. 2014 IEEE International Conference on Big Data (Big Data), 38-40.

[5]. Borisov, V., Minin, A., Basko, V., & Syskov, A. (2018). FHIR data model for intelligent multimodal interface. 2018 26th Telecommunications Forum (TELFOR), 420-425.

[6]. Sinha, A., Jenckel, M., Bukhari, S. S., & Dengel, A. (2019). Unsupervised OCR model evaluation using GAN. 2019 International Conference on Document Analysis and Recognition (ICDAR), 1256-1261.

[7]. Saluja, R., Punjabi, M., Carman, M., Ramakrishnan, G., & Chaudhuri, P. (2019). Sub-word embeddings for OCR corrections in highly fusional Indic languages. 2019 International Conference on Document Analysis and Recognition (ICDAR), 160-165.