

DISTRIBUTED HEALTH CARE FRAMEWORK FOR PATIENT HEALTH RECORD MANAGEMENT AND PHARMACEUTICAL DIAGNOSIS

De Silva K.H.K.L.

(IT19006994)

B.Sc. (Hons) Degree in Information Technology

Specializing in Software Engineering

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

September 2022

DISTRIBUTED HEALTH CARE FRAMEWORK FOR PATIENT HEALTH RECORD MANAGEMENT AND PHARMACEUTICAL DIAGNOSIS

De Silva K.H.K.L.

(IT19006994)

Dissertation submitted in partial fulfillment of the requirements for the BSc (Hons) in
Information Technology Specializing in Software Engineering

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

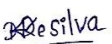
Sri Lanka

September 2022

Declaration

I declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name	Student ID	Signature
De Silva K.H.K.L.	IT19006994	

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

.....

Date :.....

Signature of the supervisor:

(Mr. Jeewaka Perera)

.....

Date :.....

Signature of the supervisor:

(Ms. Laneesha Ruggahakotuwa)

Abstract

According to recent studies, covid19 has had a significant impact on the global healthcare field and the pandemic has shown limitations in the existing digital healthcare technologies. Countries must rethink how to overcome limits to ensure service continuity while people remain at home, maintaining social distance. Transformation to Electronic Health Records (EHR) is an important solution but it can cause sensitive data leakages. Data privacy can be ensured using blockchain-based frameworks, however, manually entering patient health records onto the blockchain can lead to human errors. Most medical documents such as medical laboratory test reports and prescriptions are in printed format and textual information extraction from the said printed documents and converting them to EHR is challenging. The proposed Distributed Healthcare framework may scan and extract relevant data from patient medical documents. Textual data will be collected from medical documents using Optical Character Recognition Technology in Deep Learning while Natural Language Processing will be used to extract the relevant values from the extracted data. The optical character recognition pipeline will be utilized for text recognition, and image pre-processing will be employed to address difficulties with low image quality. Natural Language Processing techniques are used to extract relevant medical entities and values from the data extracted through OCR. Implementation of the proposed web and mobile application will help to automatically enter data into the blockchain and will minimize the errors caused by manual data entering.

Keywords: Deep Learning, Optical Character Recognition, Natural Language Processing, Machine Learning, Text Detection, Text Recognition, Healthcare

Acknowledgements

The author would like to thank Mr. Jeewaka Perera (Faculty of Computing, Sri Lanka Institute of Information Technology, Sri Lanka) and Ms. Laneesha Ruggahakotuwa (Faculty of Computing, Sri Lanka Institute of Information Technology, Sri Lanka) for the continued supervision, encouragement, and support.

Dedication

The author would like to dedicate this material to the research community, which is working tirelessly to discover solutions to sustain better outcomes in the field of healthcare.

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iii
Dedication	iv
Table of Contents	v
List of Tables.....	vii
List of Figures	viii
List of Abbreviations.....	ix
1. Introduction.....	1
1.1 Background Study	1
1.2 Literature Survey	3
1.3 Research Gap	8
Worldview Mobile Complete.....	9
1.4 Research Problem	11
1.5 Research Objectives	14
1.5.1 Main Objective	14
1.5.2 Specific Objectives	14
2. Methodology	16
2.1 Project Overview	16
2.2 System Overview Diagram	17
2.3 System Overview	18
2.4 Software Development Process	19
2.5 Feasibility Study	21
2.6 Requirements Gathering	23
2.6.1 Functional Requirements	23
2.6.2 Non-Functional Requirements	24
2.7 Technology Selection	25
2.8 Commercialization aspects of the product	26
2.8.1 Targeted Audience	26
2.8.2 Benefits from the system	26
2.9 Implementation	27

2.9.1	Project Set Up	27
2.9.2	Data Preparation and Text Extraction	29
2.9.3	Data Pre-processing and Cleaning	29
2.9.4	Train Named Entity Recognition Model.....	29
2.9.5	Predictions	30
2.9.6	Develop a Medical Document Scanner Web Application.....	30
2.9.7	Improve the Model performance	33
2.10	Testing	33
2.10.1	Unit Testing	33
2.10.2	Integration Testing.....	33
2.10.3	System Testing.....	34
2.11	Work Breakdown Structure and Gantt Chart	40
2.11.1	Work Breakdown Structure	40
2.11.2	Gantt Chart	41
3.	Results and Discussion	42
3.1	Results.....	42
3.1.1	Outputs of the Medical Document Scanner	42
3.2	Research Findings	44
3.3	Discussion	44
3.4	Future Work	45
6.	Conclusion	46
	References	47
	Appendices	50

List of Tables

Table 1.3.1: Comparison of the Developed Solution with existing Research studies..	8
Table 1.3.2: Limitations of the existing systems in the market.....	9
Table 2.10.3.1: Test Case 01.....	34
Table 2.10.3.2: Test Case 02.....	35
Table 2.10.3.3: Test Case 03.....	36
Table 2.10.3.4: Test Case 04.....	38
Table 2.10.3.5: Test Case 05.....	39

List of Figures

Figure 1.4.1: Responses on healthcare issues occurring during the pandemic.....	11
Figure 1.4.2: Response summary on the importance of healthcare automation	11
Figure 1.4.3: Responses on drawbacks in manually entering data into EHR	12
Figure 1.4.4: Responses on the importance of a medical document scanner.....	12
Figure 2.1.1: Project Overview Diagram.....	16
Figure 2.2.1: System Overview Diagram.....	17
Figure 2.4.1: The Software Development Life Cycle.....	19
Figure 2.9.1.1: Virtual environment setup.....	27
Figure 2.9.1.2: Pytesseract installation.....	28
Figure 2.9.6.1: Medical Document Scanner Interface.....	30
Figure 2.9.6.2: Image with the located coordinates.....	31
Figure 2.9.6.3: Image with the Bounding Boxes with the Tag name.....	32
Figure 2.9.6.4: Image of the table with extracted Named Entities.....	32
Figure 2.11.1.1: Work Breakdown Structure.....	40
Figure 2.11.2.1: Gantt Chart.....	41
Figure 3.1.1.1: How four coordinates are detected using Python.....	42
Figure 3.1.1.1.2: Bounding Boxes image with the Tag name	43
Figure 3.1.1.3: Extracted Named Entities in tabular format.....	43
Figure 3.3.1: Accuracy of the trained Spacy Pipeline.....	44

List of Abbreviations

Abbreviation	Description
EHR	Electronic Health Records
ML	Machine Learning
OCR	Optical Character Recognition
NLP	Natural Language Processing
OMRs	Outside Medical Records
CNN	Convolution Neural Networks
NER	Entity Recognition
SDLC	Software Development Lifecycle
OpenCV	Open-Source Computer Vision Library

1. Introduction

1.1 Background Study

The world is facing numerous challenges in the field of healthcare due to COVID-19. To control the spread of COVID-19, numerous countries had to close their borders, implement lockdowns, and employ social distancing. The epidemic has had an unforeseen worldwide impact, not just in terms of economics, but also in terms of healthcare systems generating difficulty for healthcare workers in identifying and monitoring mass populations [1]. The pandemic has exposed the importance of the digitalization of the healthcare industry and the limitations of the existing outdated systems. Hence COVID-19 has forced countries, government bodies and researchers to rethink applying modern digital solutions to the healthcare domain. In recent years researchers are focusing more on the use of Blockchain and Machine Learning approaches for digital transformation in the healthcare field. As a result of the pandemic, the amount of digitally stored patient data has grown significantly. Data surveillance, telemedicine, remote pharmaceutical diagnosis, and strategy innovation are all part of the digital healthcare ecosystem that needs to be facilitated. In addition, leveraging digital platforms to combat COVID-19 and future pandemics while developing a more patient-centric and mainstreamed digital healthcare ecosystem is essential to design.

Here we address these issues by proposing an approach to using Blockchain and Machine Learning-Based Healthcare framework which provides healthcare services to medical practitioners and patients while staying at home and maintaining social distancing.

The proposed solution facilitates the healthcare services given as follows:

1. Blockchain-based distributed healthcare framework for securely storing and accessing patient data
2. A Deep Learning and Natural Language Processing based medical document scanner to prevent the errors that are due to manually entering data

3. An Image Processing based drug identification module for remote pharmaceutical diagnosis
4. A Natural Language Processing and Machine Learning based virtual chatbot in healthcare assistance

In most countries accessing patient data is exceedingly difficult due to the unavailability of Electronic Health Records. Problems in the Medical Industry such as poor data sharing and leakage of sensitive data can be overcome with the use of blockchain-based electronic health records [2].

Conversion of existing patient data into electronic health records is a challenging task since most of the medical documents such as lab test reports, and prescriptions from hospitals are in printed format. Converting these data into Electronic Health Records (EHR) and entering these details into the blockchain often need to follow the manual data-entering procedure. But manual entering data is often time-consuming and error prone. Here we address these issues by proposing an approach to extract structured data from the photographed images of the medical documents. Frequent Visit to doctors or pharmacists for pharmaceutical diagnosis during a pandemic is not safe. Remote pharmaceutical diagnosis in the proposed solution will help to identify the drug and learn about its side effects and dosage even while staying at home. Virtual conversational agents like chatbots will serve as virtual healthcare assistants to make things easier.

This report will give an overall idea of the proposed solution and focus more on the Medical Document Scanner component which is based on Optical Character Recognition and Natural Language Processing. The next chapters will elaborate and provide a more in-depth insight into the said component. This research aims to deliver patient-centric healthcare services to combat COVID-19.

1.2 Literature Survey

For the past few years, several studies have been conducted on the digitalization of the field of healthcare and how to provide healthcare facilities to the public while mitigating the challenges associated with it. This section highlights a variety of studies that have been conducted on the use of deep learning and natural language processing-based techniques for document scanning to overcome the challenges in the manual entering of data. This will also go over a range of studies that have been done on text detection using various technologies and approaches. This evaluation will include both generic document scanning technologies and medical domain-specific document scanning approaches, both of which will be quite useful in the review.

According to [3], an essential milestone in the advancement of contemporary medicine is the introduction of electronic health records. But, due to the limitations of EHR systems, comprehensive health records are not frequently accessible during treatment. Therefore, here the authors suggested a text detection approach with the use of a patch-based training strategy and a concatenation structure which can combine the features of the deep and shallow layers in the Neural network. This study was conducted to improve the accuracy of multilingual text recognition. According to the research, a patch-based training technique has been applied to the medical laboratory report and outputs the bounding boxes that contain texts. The text is then printed after the concatenation structure is inserted into the recognizer.

The authors of [4] show that a medical laboratory report becomes a type of crucial record for healthcare providers to use in patient evaluation and treatment. Electronic medical records are easier to maintain than paper ones, which are currently ubiquitous in the contemporary healthcare system. However, there continues to be a huge demand for past medical laboratory report identification, particularly in underdeveloped nations. Here the authors use a method for collecting data from medical laboratory reports using textual image processing. The table sections and words of a report are initially segmented using top-down pipelines after being provided with an image of the document. Although the system achieves satisfactory results still it was capable of extracting text from documents only. However, the presented research study did not include entity extraction.

A key challenge in Optical Character Recognition is the inability of the current OCR algorithms to correctly transcribe the scanned documents where text is skewed or distorted. The authors of [5] developed a deep neural network-based self-supervised pre-training model for their research work. This bi-directional encoder has been designed to predict concealed text and fill in gaps in non-transcribable areas of the page. The suggested model, however, has not been trained for the healthcare domain-specific words and has not improved the quality of the images.

A recent study [6] revealed that the COVID-19 epidemic is still exerting extreme pressure on the service sector. For any organisation to operate efficiently, getting the appropriate information at the right time is essential. Businesses now generate more complex data per minute. The unstructured data is digitally mastered, and the OCR procedure is conducted. The accuracy of scanning scanned and handwritten materials, where the text may be distorted, blurred, or unintelligible, is a significant problem for the OCR technique. Here, the authors demonstrate how OCR may facilitate the electronic extraction of printed materials, including applications and medical records, to access critical data that was available in the past.

According to the study [7], the prevalence of machine-illegible information and the restricted system accessibility in healthcare, obtaining usable and relevant information from these Outside Medical Records (OMRs) in a timely way is a difficult undertaking. Here the authors have found the clinical concepts contained in OMR are beneficial for Cardiovascular medicine with the use of techniques in Optical Character Recognition (OCR) and Natural Language Processing (NLP) to extract data from the computer-readable OMRs. The study can be regarded as the first step toward automated data extraction from OMRs generated by a variety of healthcare providers.

According to the research paper [8], it can be seen, that encouraging results can be produced when a variety of image-processing algorithms are applied to text detection. In the research paper, the authors proposed a modular strategy for text detection. It has been mentioned in the research findings that satisfactory performance can be achieved even without the use of the deep learning approach. Although OCR aids in promoting the precisions, it decreases recall performance. This results from the removal of non-

text sections together with certain crucial text parts. Therefore, the writers must avoid it in this instance to enhance overall outcomes.

The research study [9], provides an overview of the application of Convolution Neural Networks in text mining. CNN can be used for text mining with sentiment analysis, classification of documents or with its semantic representation, according to this study. The research findings indicate that CNN has not fully addressed several issues in the text mining and NLP domains. It occurs because CNN is utilised in handling different issues, including document categorization, or NLP situations involving entities and relationships.

In the year 2019, a group of researchers proposed an NLP-based free-text extraction from notes related to chronic diseases [10]. The study investigates the difficulties of NLP approaches as well as clinical narratives. The analysis found that machine learning is becoming more popular, and deep learning technologies are still in their preliminary stages. According to the study's findings, developing natural language processing (NLP) approaches is critical for automatically converting clinical text into structured data that can be managed directly using machine learning algorithms.

The authors of [11], provided a unique task of visual question answering by reading text in photographs as part of their study, which included a large-scale dataset as a baseline. They proposed a large-scale dataset, OCRVQA–200K, to assist a systematic approach to research this novel issue. The gathered dataset opens numerous fascinating research options both for document image analysis and contributes to the research community on text recognition.

According to [12], the authors used popular, out-of-the-box tools to perform a series of extrinsic assessment tasks such as named entity recognition, retrieval of information, dependency parsing, segmentation, and topic modelling. According to the study, when compared to those trained on human-corrected text, poor OCR quality has an increasing influence on prediction models and has conducted a large-scale analysis of the impact of OCR errors on numerous NLP tasks.

The research study [13] employed NER to extract text from historical documents using OCR techniques. According to the authors, the fundamental difficulty with this work

is that the OCR process produces output text that has grammatical and spelling problems. Aged papers may also have historical variances, which might affect how well the NER process works. Two prominent datasets in German and French were compared to earlier state-of-the-art models by the authors in this study, and they suggest a framework to overcome the NER challenge for past data.

According to the study [14], NER models are often built using the Bi-directional LSTM architecture. The limitations of the sequential nature and the modelling of single input restrict the full usage of global information from the greater scope in both the complete phrase and the entire text. The authors overcome these two shortcomings in this study and provide a model enhanced with hierarchical contextualised depiction at the document levels.

The study [15] shows that clinical NER technology is a cornerstone for learning the contents of digitized medical data. Traditional clinical NER techniques are severely hampered by feature extraction. Additionally, these approaches disregard the lengthy contextual relationships and consider NER as a sentence-level problem. To address the issue, the authors suggest an attention-based neural network design in this research. The document's pre-trained mode with neural recognition is used to extract the document's overall content.

As per the above-mentioned readings and the thorough complete literature review, we can see several scholars have already worked and are working on identifying alternative ways for Text Detection and Recognition and extraction of essential data from it. And several researchers are still studying them and continue to do so. Furthermore, some researchers are still investigating them and will continue to do so in the future. Some research projects yielded promising outcomes, while others had research gaps that need to be filled. (In the next section, we will talk about the Research Gap.) The study found that the same goal can be accomplished using several technologies and methodologies. Every strategy has advantages and disadvantages. The review addressed a wide range of textual data extraction techniques, including Deep Learning and Natural Language Processing.

As a result of the reviewed research studies, it is apparent that textual extraction from printed documents is a significant requirement in the healthcare domain, and researchers should spend more time in this area to develop better solutions.

1.3 Research Gap

Most available research papers and research studies focus primarily on textual data extraction from generic documents [8], [9], [11], [12], but most research studies ignore approaches for data extraction from medical and healthcare-specific documents such as lab test reports and printed prescriptions. One of the most difficult aspects of textual data extraction is extracting text from skewed or occluded materials. However, most research papers have not used adequate ways to retrieve data from such distorted texts. Even though the research [5] focused on OCR post-correction techniques, the model has not been trained for domain-specific data sets. Medical terminologies are different from the general vocabulary, so it is necessary to train the models with healthcare domain-specific terms to achieve better outcomes. The precision of research [3] is around 98.6%, however, the proposed technique is tailored to the Chinese language, and there are limitations to adopting the same approach for other languages. The initial step toward autonomous data extraction from OMRs was completed by the research study [7]. However, the study was limited to a single institute, and future research should broaden the scope to include a multi-site institution to extract essential clinical narrative information.

Table 1.3.1: Comparison of the Developed Solution with existing Research studies (Research Gap)

Reference ID	Modelled for Healthcare Domain-Specific words	Extract Text from low-quality images	Extract important Named Entities from unstructured text	Make the extracted text into an editable format
Research [3]	✓	✗	✓	✓
Research [5]	✗	✓	✓	✓
Research [7]	✓	✗	✗	✗
Our Solution	✓	✓	✓	✓

Most of the research that has been done is designed for generic usage. In healthcare, there are so many domain-specific words and document scanners developed for general usage that will not produce better results. Our proposed solution will be tailored to the healthcare industry and will provide better results than generic solutions.

Table 1.3.2: Limitations of the existing systems in the market

Name of the Application	Available Format / Platform	Limitations
Worldview Mobile Complete	Mobile Applications designed for iPhones and iPads	<ul style="list-style-type: none"> • Textual data will not be extracted • Do not capture the important values or entities • Data will be stored locally
EncryptScan by HIPAA	Mobile Application designed for iOS & Android	<ul style="list-style-type: none"> • Textual data will not be extracted • Do not capture the important values or entities
Abby FineReader PDF	Windows 10	<ul style="list-style-type: none"> • Generic Document scanner • Not specifically designed for healthcare-related documents
VueScan	Windows, macOS, and Linux	<ul style="list-style-type: none"> • Generic Document scanner • Not specifically designed for healthcare-related documents

Most of the printed documents will be of mediocre quality so our proposed method emphasizes a post-correction approach for improved results. Grammatical mistakes and spelling mistakes are widespread in text documents therefore our solution will follow the approaches for word correction and data correction. Not only have we considered the existing literature, but we have also considered the existing systems

that are already on the market, and we have highlighted the limitations of those systems. It will assist in the identification of research gaps in the related area.

Most current medical document scanners scan the document's image and convert it to a soft copy. However, many of these methodologies have not been addressed in textual data extraction from such publications. Most other document scanners, on the other hand, are built for typical use and not specifically for the healthcare industry.

These are the research gaps in the field of healthcare that have been found, and our proposed solution will help to bridge those gaps.

1.4 Research Problem

The pandemic has exposed healthcare's limitations, emphasizing the importance of digitalization. Most medical papers are in printed format and extracting information from them and transferring them to electronic health records takes a lot of time. Manually entering these data into Blockchain is a risky task that frequently results in human errors. As a result, an automated method for extracting textual data from printed medical records and converting them to editable and searchable formats should be introduced. A public survey was conducted to gather information on the healthcare problems that emerged during the covid19 epidemic. According to the survey, about 91.4% say that they face healthcare issues during the pandemic.

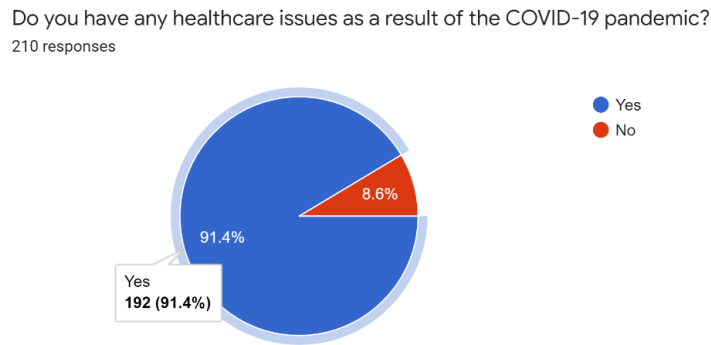


Figure 1.4.1: Responses on healthcare issues occurring during the pandemic

And about 89.5% of the participants do believe that healthcare automation is critical during the pandemic.

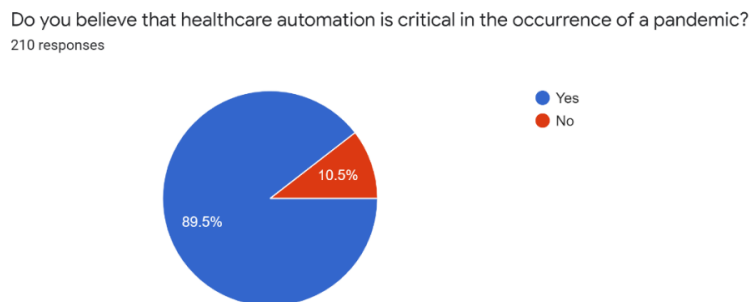


Figure 1.4.2: Response summary on the importance of healthcare automation

About 53.3% of most participants think that manually entering data and transferring them into Electronic Health records can cause errors and is typically a time-consuming procedure.

What are the drawbacks of manually entering data and transferring it to an electronic health record (EHR)?
210 responses

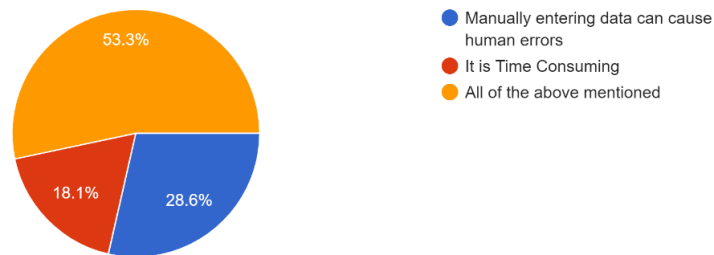


Figure 1.4.3: Responses on drawbacks in manually entering data into EHR

About 45.75% of the majority responded that the introduction of medical document scanners is of high importance. Most of the available document scanner applications are for general usage and such scanners will not work efficiently when it comes to the field of healthcare since there are so many domain-specific words in the medical field. Hence there is a market need to implement document scanners specifically trained for textual data extraction from Medical Documents.

What is the importance of introducing a medical document scanner? Please rate your preference on a scale of 1 to 5.
210 responses

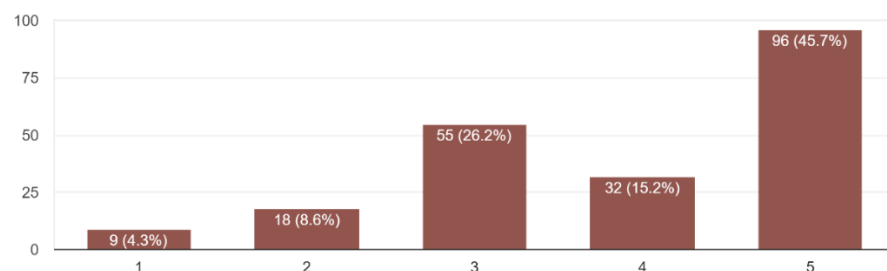


Figure 1.4.4: Responses on the importance of a medical document scanner

In addition, most printed medical documents deteriorate or get skewed over time. Data extraction from such documents is a challenging task. In such cases, post-correction procedures and automatic text prediction methods should be used.

1.5 Research Objectives

1.5.1 Main Objective

The primary objective of implementing a healthcare framework is to address the healthcare difficulties that may occur because of the COVID-19 pandemic. The pandemic exposed healthcare shortcomings, and this framework will automate the existing healthcare services. The proposed solution's key objective is to securely store patients' healthcare information while protecting users' privacy and to provide healthcare services for Medical Documents Scanning, Conversational Chatbot for Virtual Assisting and remote pharmaceutical diagnosis. The proposed solution's principal goal is to provide secure healthcare facilities for Medical Practitioners and Patients while maintaining social distancing.

1.5.2 Specific Objectives

The following are the specific objectives that must be completed to achieve the main goal. This section will go through the specific goals of the Medical Document Scanner component in greater detail.

1. Prevent the errors that they cause when manually entering data into Blockchain

Most medical documents such as lab test reports are in printed format, and data extraction and converting them to EHR is a complicated process. Manually inputting data from such documents and entering those data into Blockchain is a time-consuming and error-prone process. These issues will be prevented with the suggested medical document scanner.

2. Automatically extract structured data from the captured images and annotate relevant data from the medical documents using Text Recognition

With the aid of Deep Learning Techniques, the proposed solution will extract text from captured images of medical documents and convert it to text. The captured data will be in an editable or searchable format, making it easy to enter data into the blockchain.

3. Extract important entities from the recognized text

After the text has been captured, the appropriate entities and values will be annotated using Natural Language Processing algorithms.

Given below are the entities which will correctly be identified from the developed solution.

1. Age of the Patient
2. Report Date
3. Patient Name
4. Test/Profile Name (Name of the Laboratory Report)
5. Test Results
6. Additional Comments

4. Correctly transcribe documents where text may be skewed or illegible

Most printed documents are obscure over time. Text recognition technologies are significantly more challenging to use to extract data from such documents. Special strategies will be utilized in this solution to capture data from such illegible papers.

2. Methodology

2.1 Project Overview

The suggested system is designed to meet the challenges that the healthcare domain confronts during the COVID-19 pandemic, as well as to provide healthcare solutions that ensure service continuity while people remain at home and maintain social distancing. The proposed distributed healthcare framework would include secure patient health record management and pharmaceutical diagnostic capabilities.

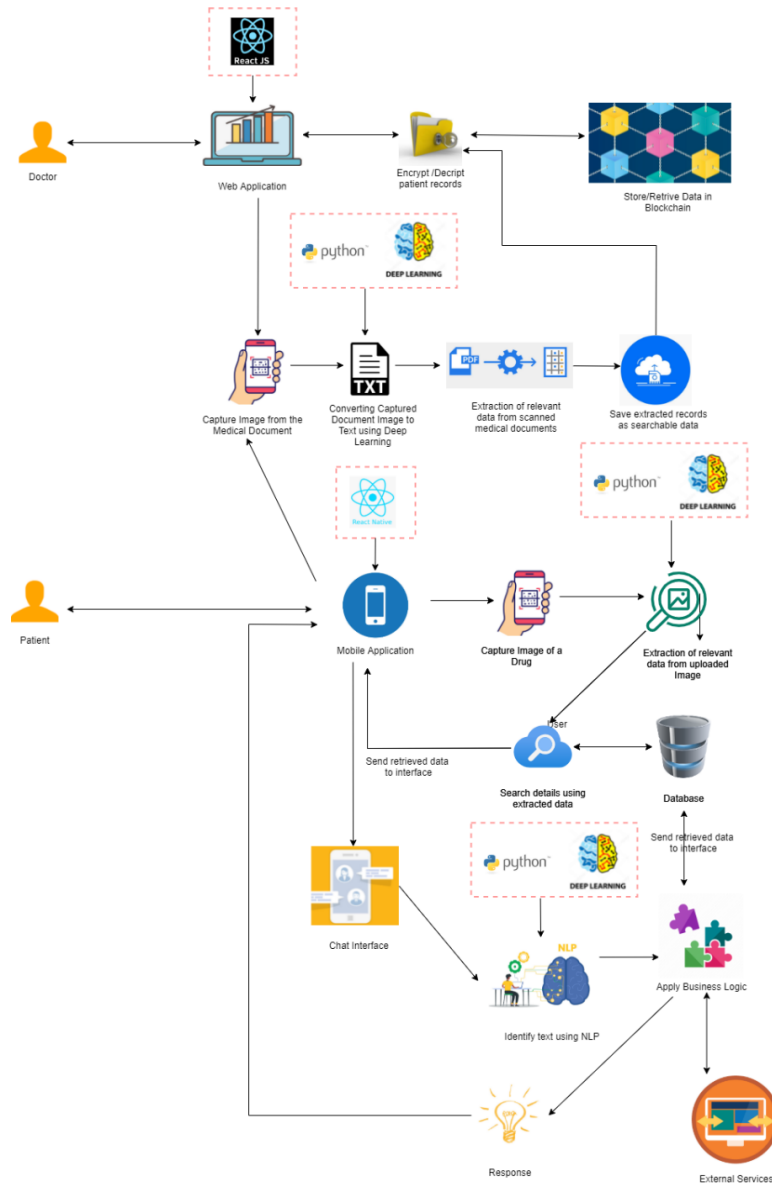


Figure 2.1.1: Project Overview Diagram

According to Figure 2.1.1, there are four major components in the Healthcare Application developed during this research. The system includes a Blockchain component, a Medical Document Scanner, Virtual Medical Chat Bot, and a Drug Identifier as the four major components. The blockchain component will provide capabilities for safe access and data sharing while securely storing patient data. The Medical Document Scanner will scan the clinical laboratory test reports and scan and extract important named entities from the scanned documents. Patient inquiries will be answered by the medical chatbot, which will also provide reminders to take medication based on the specifics of the prescription. Drug Identifier will recognise the medication using images of the tablets and deliver the necessary information.

2.2 System Overview Diagram

Figure 2.2.1 depicts the System Overview Diagram of the Medical Document Scanner component.

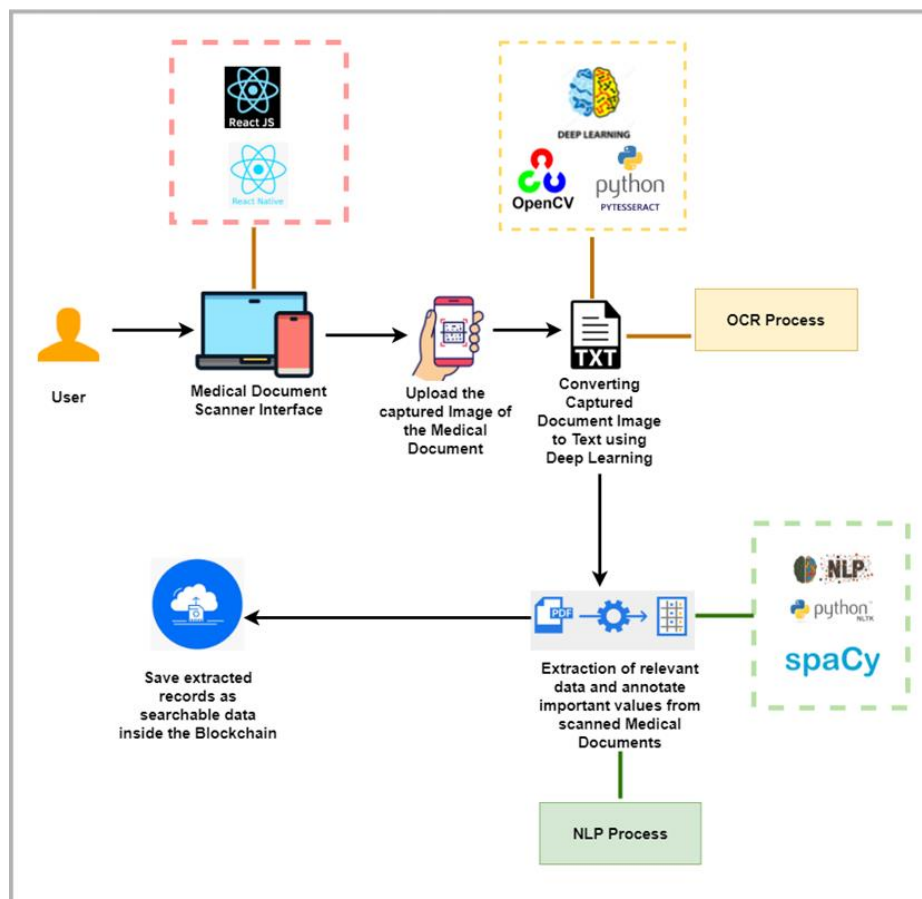


Figure 2.2.1: System Overview Diagram

2.3 System Overview

Figure 2.2.1 shows the Medical Document Scanner component, which can extract textual data from printed Medical Documents such as Clinical Laboratory Test Reports and restructure them to annotate the important values and extract defined Named Entities from the extracted Text. This component will be added to the blockchain component to reduce the amount of data that must be manually entered into the blockchain and can minimize the errors caused due to human errors.

A medical document scanner can be used by a doctor, medical practitioner, or any other authorized entity that can add or modify data inside the Blockchain.

Due to security concerns, Blockchain has access control procedures in place for extremely sensitive patient data. Eligible users can use the web application and access the Medical Document Scanner interface to upload a captured image of a medical document.

With the use of the Optical Character Recognition technique in Deep Learning Textual data will be extracted from the captured image and converted into a text document. Important values will be annotated with the use of the same technique.

One of the drawbacks of existing document scanners is that it extracts text word by word and does not provide a meaningful idea. As a result, the proposed model will be trained to extract data and restructure it in a way that is similar to the original image as well as to provide a meaningful idea. Techniques in Natural Language processing will be used for this.

The Bounding Boxes will be drawn, and the predicted entities will be tagged. Finally, the data will then be transformed into a searchable or editable format and stored within the Blockchain.

2.4 Software Development Process

The Software Development Lifecycle (SDLC) divides the operations of the software development process into small steps. Out of the several diverse types of software development models, agile methodology is ideal for continual expansion over several iterations. The requirements of the proposed solution will change gradually over time, especially as the development process begins.

The agile methodology's incremental and iterative nature aids in the continuous changes that occur over time. According to Figure 2.4.1 Requirements gathering, Analysis, Design, Coding, Testing, and Maintenance are the six main steps in the agile software development cycle.

Each iteration will result in a finished product. There are several distinct types of Agile Methodologies and SCRUM is the most common and popular one. SCRUM is a framework for agile project development that will be utilized throughout the research. The team will have daily stand-up calls to receive a daily update on the project's development. SCRUM is the ideal approach since it can adapt to frequent changes, and the project is susceptible to frequent modifications.

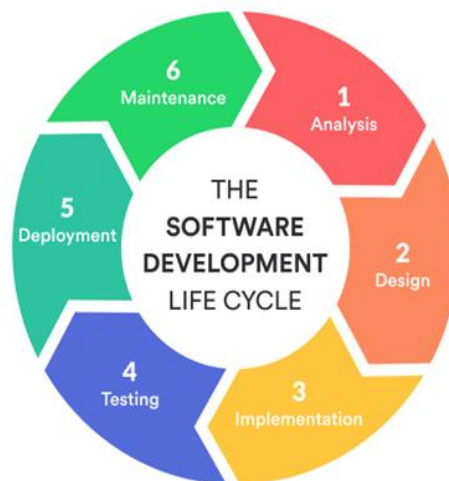


Figure 2.4.1: The Software Development Life Cycle

During the Analysis phase, research gaps were identified, and research analysis was conducted. In a similar vein, during the Designing phase of the SDLC, a viable solution is proposed, and the system is designed based on that. Following the design, the system implementation began with the use of best-suited tools and technologies while following the best practices.

After the implementation and testing, the system testing phase began. This phase included unit testing, integration testing, functional testing, and performance testing. In the deployment phase, the customer receives the finished product in a real-time production setting. The product is prepared for usage after it has been deployed.

2.5 Feasibility Study

- **Economic Feasibility**

The proposed solution is aimed at hospital chains across the country, and physicians, medical practitioners, and patients would all benefit from the system's completion. The usage of a medical document scanner will reduce the amount of data that needs to be manually entered into the blockchain, as well as the mistakes that can arise as a result. Most importantly, the system is a full software solution that does not require any hardware components. As a result, the proposed solution will be executed successfully and at a low cost. The costs incurred at each stage, namely

- a. Planning and Design Cost
- b. Document preparation costs
- c. Hosting charges
- d. Internet usage costs.

- **Technical Feasibility**

Deep Learning approaches for Optical Character Recognition will be utilized for Text Recognition, while Natural Language Processing techniques will be used to annotate data and reorganize the document in a way that is comparable to the original picture. The sub-components will then be combined into a solitary product that will be hosted on a server. To ensure a successful implementation, everyone should extensively research these modern technologies before beginning implementation, ensuring that the proposed solution is technically possible.

- **Operational Feasibility**

The proposed solution would operate effectively in the field of healthcare, and the system will benefit both healthcare professionals and patients. The present limitations in the healthcare domain will be reduced by this technology. The Medical Document Scanner component will help to eliminate the drawbacks

of manually inputting data into the system. The suggested component will be beneficial to doctors and medical practitioners.

- **Schedule Feasibility**

The proposed solution is expected to be completed within a year. The scope of the study and its sub-components have been narrowed and fine-tuned accordingly. The intended system will be implemented on time, and the system will be feasible according to the schedule.

2.6 Requirements Gathering

2.6.1 Functional Requirements

1. Extract textual data from captured images of the medical documents:

Textual data should be extracted from the captured images of printed medical documents such as Clinical Laboratory Text Reports.

2. Restructure the image to appear the same as the original image:

Normal text recognition models capture raw data by scanning word by word. The captured text must be rearranged in this module to appear the same as the original picture.

3. Draw Bounding Box around the recognized Named Entities and Tagging:

The names of the identified Named Entities should be tagged next to the Bounding Box that has been drawn around them.

4. Extract text from distorted documents:

With time, most printed medical documents get distorted. This suggested module is trained to extract data from skewed documents. Normal text recognition models capture raw data by scanning word by word. The captured should be rearranged in this module to appear the same as the original picture.

5. Extract Text and display the captured name entities in a tabular view:

Important Named Entities such as Age, Date, Patient Name, Test, Result and Additional comments should be extracted and displayed in a tabular format.

6. Display the extracted Named Entities and make them into editable format:

Before transmitting the data to Blockchain, any problems that affect the results' correctness or that include spelling or grammar should be fixed. Therefore, the captured textual data should be in an editable format.

2.6.2 Non-Functional Requirements

1. Security

Instead of keeping the collected textual data on a centralized server, the captured data will be kept inside the blockchain. Data breaches will be avoided, and the data will be stored safely.

2. Availability

This proposed system will be deployed in blockchain and will be accessible 24/7 and can authorized parties can access it from anywhere without any restriction.

3. Usability

Doctors, medical practitioners, and patients will benefit from the proposed solution. Therefore, the system will consider the usability aspects such as satisfaction and efficiency.

4. Accuracy

The proposed method would reduce data entry by hand and, as a result, will ensure that the system is accurate.

5. Performance

This proposed solution will be implemented to provide a quick response within a specified period and to function at an elevated level of efficiency.

2.7 Technology Selection

In the Computer Vision module, the document is first scanned, the text's placement is determined, and then the text is extracted from the image. Then in Natural language processing, the entities from the text are extracted, do the necessary text cleaning, and parsed the entities from the text.

Optical character recognition, a Deep Learning approach, is used to extract textual data from Clinical Laboratory Test reports. Python libraries such as

1. OpenCV,
2. NumPy,
3. Pytesseract

are used to achieve this. Using OpenCV technology, the medical records were loaded, and Pytesseract is utilized to extract the content. Named Entity Recognition is conducted using Natural Language Processing.

Python libraries in Natural Language processing such as

1. Spacy,
2. Pandas
3. Regular expressions

Was used for the extraction of Named Entities. The front-end web application is developed using technologies like HTML, JavaScript, and ReactJs. The preparation and enhancement of images were done using various OpenCV techniques.

2.8 Commercialization aspects of the product

2.8.1 Targeted Audience

The proposed solution is aimed at the field of healthcare, and the proposed system's target audience includes physicians, healthcare workers, and patients.

2.8.2 Benefits from the system

1. Securely storing, and accessing scattered patient data across several EHRs (Electronic Health Records)
 2. Medical Document Scanner to extract text from medical documents and annotate and extract important entities from the captured text
 3. Identify drugs using the image and provide adequate information such as dosage, side effects and many more
 4. Virtual conversational medical chatbot to communicate with patients while giving daily reminders to take medication on time
 5. 24/7 service with no or minimum downtime
 6. Provide distributed healthcare services to end-users across the island
- High data security with required access control protocols

2.9 Implementation

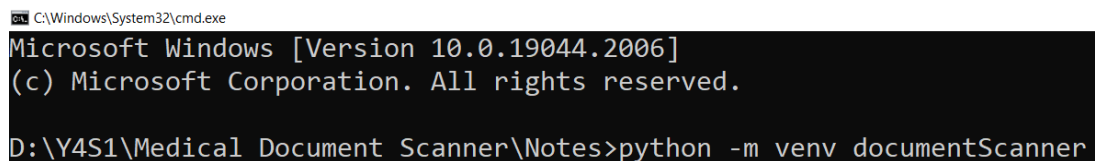
The developed system is a Medical Document Scanner which extracts text and customized Named Entities from Clinical Laboratory Test Reports. This project makes use of Computer Vision and Natural Language processing which are two key technologies in Data Science and Machine Learning. As the project combines two important technologies, it has been broken down into multiple stages of development for ease of understanding. Each of the steps that make up the stages of development is given below and will be covered in more detail in separate subsections.

1. Project Set Up
2. Data Preparation and Text Extraction
3. Data Pre-processing and Cleaning
4. Train Named Entity Recognition Model
5. Predictions
6. Develop a Medical Document Scanner Web Application
7. Improve the Model performance

2.9.1 Project Set Up

Setting up a project entails completing the required installations and prerequisites. Python and its dependencies must be installed as the initial step. Anyone can download the required installers and source files for Python from the official website.

The Virtual Environment should be established as the next step. The dependencies of each project are independent of the requirements of other projects, and a virtual environment is helpful when working on a shared system and resolving the issue of access privileges to install the software. A virtual environment can be created using the following command as given in Figure 2.9.1.1.

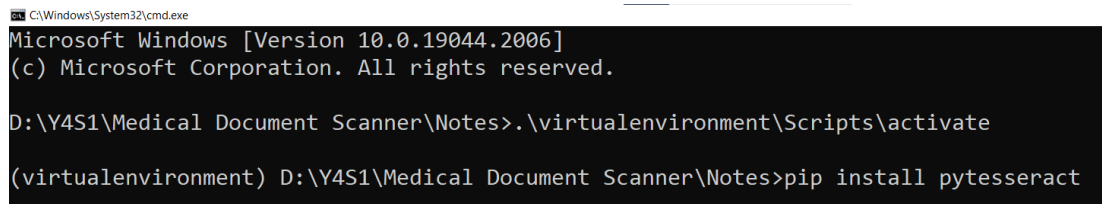


```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19044.2006]
(c) Microsoft Corporation. All rights reserved.

D:\Y4S1\Medical Document Scanner\Notes>python -m venv documentScanner
```

Figure 2.9.1.1: Virtual environment setup

After the virtual environment set-up is done Tesseract OCR should be installed in the computer before the installation of the Pytesseract library. Tesseract OCR is an open-source text recognition OCR engine developed by Google. Pytesseract can be installed using the below command given in Figure 2.9.1.2 after Tesseract OCR has been successfully installed.

A screenshot of a Windows command prompt window. The title bar shows 'C:\Windows\System32\cmd.exe'. The window content displays the following text: 'Microsoft Windows [Version 10.0.19044.2006] (c) Microsoft Corporation. All rights reserved. D:\Y4S1\Medical Document Scanner\Notes>. \virtualenvironment\Scripts\activate (virtualenvironment) D:\Y4S1\Medical Document Scanner\Notes>pip install pytesseract'.

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19044.2006]
(c) Microsoft Corporation. All rights reserved.

D:\Y4S1\Medical Document Scanner\Notes>. \virtualenvironment\Scripts\activate
(virtualenvironment) D:\Y4S1\Medical Document Scanner\Notes>pip install pytesseract
```

Figure 2.9.1.2: Pytesseract installation

All necessary libraries should be installed into the virtual environment after Pytesseract is installed. The following libraries need to be installed:

1. NumPy - One of the most popular Python packages for scientific computation. It offers a multidimensional array object along with variants like filters and matrices that may be used for different mathematical operations.
2. pandas - A tool for data cleaning and analysis in data science and machine learning
3. SciPy - Offers additional useful functions for processing applications, statistics, and optimizations.
4. Matplotlib - A cross-platform data visualisation and graphical plotting library
5. Pillow - Python Imaging Toolkit, which provides support for viewing, processing, and storing a wide variety of image file types
6. OpenCV-python - Used in computer vision

The Spacy Library can be installed as the last step. Spacy is an open-source library for NLP and it will be used for training the NER model. From the official website, the set-up can be downloaded and under the Trained Pipeline the English language should be selected before starting the installation since this project focuses on extracting Entities from Clinical Laboratory Reports given in the English language. Testing should be done once the installations are finished to ensure that all the packages were correctly installed in the virtual environment.

2.9.2 Data Preparation and Text Extraction

Gathering a data set is the main thing to be done in the data preparation stage. The dataset which is been used in this research includes 260 clinical laboratory test reports from 24 Egyptian laboratories [16]. It is applied in the study of text extraction from clinical laboratory test reports. Images of the lab test reports from scanners, smartphones, and pdf-to-image scanning are included in the dataset.

After getting the dataset the text should be extracted from all the images with the use of Pytesseract. Pytesseract operates in a specific order. There are five levels, including levels 1, 2, 3, 4, and 5. The page is defined at Level 1. The block is defined at Level 2. The paragraphs are defined at Level 3. Level four determines the lines while the last level specifies the words. Here in Pytesseract, the block is recognised after the page has been identified. Similar to this, Pytesseract locates paragraphs, lines, and ultimately words before sending them to the deep learning model. After that, data frames are created from the image's retrieved text. The data was then stored in CSV format after the text in the data frame was cleaned. Labelling should begin after all the data has been stored in the CSV format and the BIO/IOB format is used for labelling. The prefix "I" indicates that the tag is within the chunk, while the letter "B" indicates that the word is at the beginning of the chunk. The term does not belong to the chunk, as indicated by the prefix "O."

2.9.3 Data Pre-processing and Cleaning

The data should be pre-processed using certain methodologies before training the Named Entity Model. First, must load the data and convert it into pandas. After cleaning the data from all the images then the data can be organised into groups. When the data is converted to the Spacy format the data can be split into the Training set and the Testing set.

2.9.4 Train Named Entity Recognition Model

There are several preconfigured models in the Spacy library, and these models can be used to train the Named Entity Recognition model according to the requirements of the research. The spacy train command on the command line is the suggested method

for training the Spacy pipeline. All settings and hyperparameters must be included in a single configuration file. The data should be in pickle format for model training. Training the NER model can begin after data preparation. The model can be stored for further usage after the training is complete.

2.9.5 Predictions

The trained NER model should first be loaded for the predictions. The imported NER model can be utilized to draw the bounding boxes and tag the predicted entities after the text has been extracted from Pytesseract. In this manner, a complete prediction pipeline can be created.

2.9.6 Develop a Medical Document Scanner Web Application

Following is how the Document Scanner web application ought to operate. First, there should be an interface to upload the image of the laboratory report as shown in Figure 2.9.6.1.

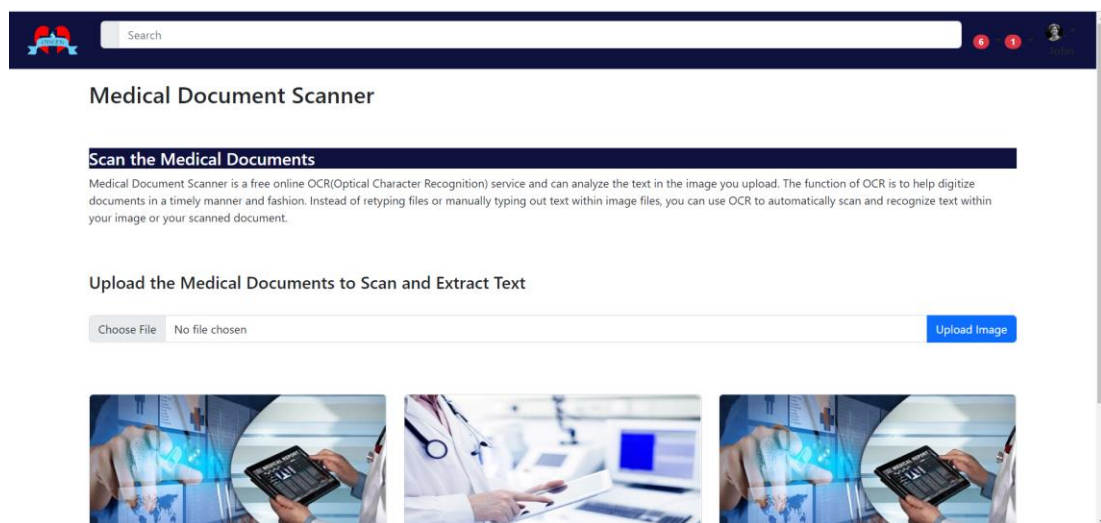


Figure 2.9.6.1: Medical Document Scanner Interface


The image is then internally passed into the document scanner and that scanner will return the four points. If an incorrect prediction is received, the corners can be adjusted with the use of JavaScript as shown in Figure 2.9.6.2.

Upload the Medical Documents to Scan and Extract Text

No file chosen

Wrap Document and Extract Text

Located the Coordinates of Document using OpenCV



PATIENT NAME	Registered	05/05/2020	Collected	05/05/2020
Visit Number	Age	Gender	Referred By	Client ID
	28 Year	Male	Prof. Dr.	

Test Name	Result	Unit	Reference Range
Complete Blood Picture			
Hemoglobin	13.9	g/dl	12.5 - 17.5
Hematocrit (PCV)	41.4	%	41 - 52
RBC Count	5.19	millions / mm ³	4.5 - 6.0
MCV	79.8	fL	88 - 100
MCH	28.8	pg	27 - 33
MCHC	35.5	g/dl	31 - 37
RDW-CV	13.8	%	11.5 - 14
Platelet Count (EDTA Blood)	248	thousands / mm ³	150 - 450
Total Leucocyte Count (EDTA Blood)	7.8	thousands / mm ³	4 - 11
	Percent Values	Absolute Values	
Neutrophils	58.3 %	4.53 x10 ⁹ /L	2 - 7
Lymphocytes	28.6 %	2.22 x10 ⁹ /L	1 - 4.8
Monocytes	10.8 %	0.84 x10 ⁹ /L	0.2 - 1
Eosinophils	1.9 %	0.15 x10 ⁹ /L	0.1 - 0.45
Basophils	0.4 %	0.03 x10 ⁹ /L	0 - 0.1
Other Cells:			
Comment:			
Relative monocytosis.			
Follow up is recommended.			

Professor of clinical pathology,
Faculty of medicine, Cairo university

19014
al-mokhtabar.com

CAP
Cairo Pathology

حب نفسك

Figure 2.9.6.2: Image with the located coordinates

Then the wrap document extract text button should be clicked. The Bounding Box image with the tag names will then show up on the following page as given in Figure 2.9.6.3.

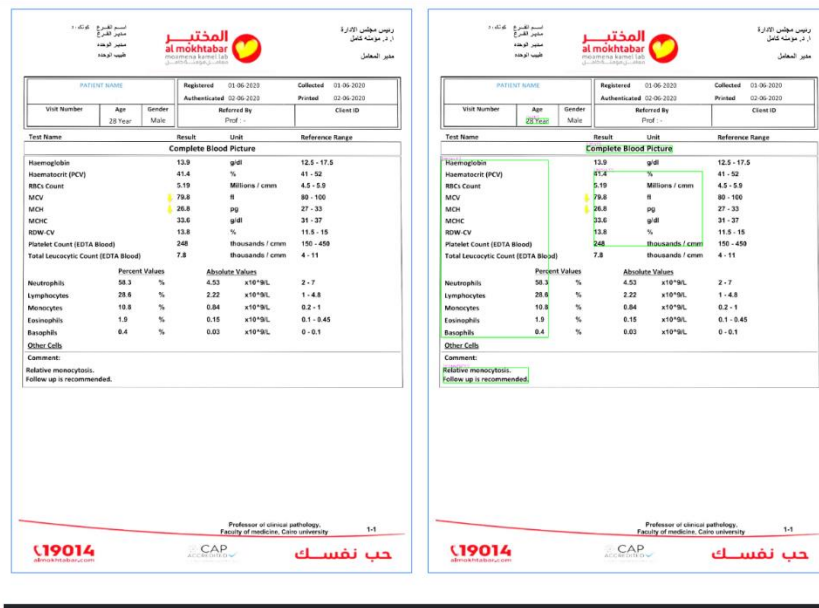


Figure 2.9.6.3: Image with the Bounding Boxes with the Tag name

Similar to this, a tabular representation of all retrieved Named Entities will be shown as given in Figure 2.9.6.4.

Since this table is editable, any errors that might have occurred due to erroneous predictions can be fixed before the data is sent to the blockchain. Before developing the web application, Flask should be installed in the virtual environment. Python in Flask can be used to predict the document coordinates. Additionally, JavaScript can be used to change the coordinates.

Named Entities		
Entities	Extracted Text	
AGE	59 y	⊕
DATE	07/06/2021	⊕
PATIENTNAME		
TEST	Chemistry Unit	⊕
RESULT	Fasting Plasma Glucose 111', 'Plasma Glucose 2Hrs Pp 103', 'Glycated Haemoglobin', 'Serum Urea', '17', 'Serum Creatinine', 'Serum Uric Acid 7.0	⊕
COMMENTS		

Figure 2.9.6.4: Image of the table with extracted Named Entities

2.9.7 Improve the Model performance

Some image processing techniques can be used on the image to enhance the performance of the model. The image may be processed using image processing techniques including morphological transformation, Gaussian Blur, and detail enhancement. The magic colour can be used on an image after procedures to modify brightness and contrast have been applied. When extracting text from low-quality or skewed documents, improving the picture quality will enhance the model's performance.

2.10 Testing

The test results from the developed application are highlighted in this section. Various testing techniques are required for the system at various stages of the development life cycle. These tests aid in detecting any system vulnerabilities. Testing is a challenging and crucial step in the development of the application. Usability, performance, security, and functional and non-functional elements are all included in application testing. The testing will raise the product's quality and it is critical to spot the system's weaknesses early on. Bugs and issues can be solved by preparing the test cases for each function.

2.10.1 Unit Testing

Each module is evaluated independently to ensure that it meets all the standards and has all the necessary functions. The components can be readily merged with other modules if they are error-free. Individual tasks including uploading an image, locating contours, extracting named entities, generating bounding boxes in the image, and model training processes are all assessed separately under the unit testing.

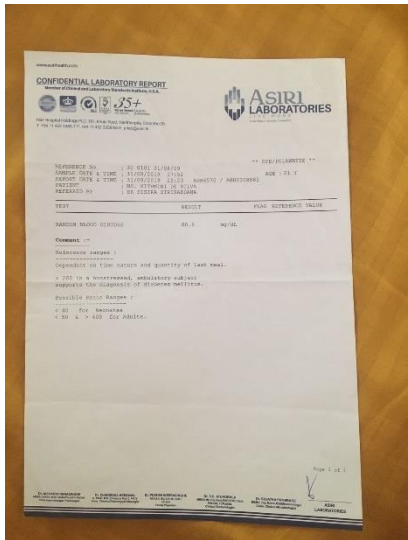
2.10.2 Integration Testing

All the individual components are linked during integration testing and evaluated as a single unit. Integration testing is required to confirm that all functionalities perform properly once all the components have been integrated.

2.10.3 System Testing

System testing is done to see if the system's actual outputs match what was anticipated. Here, system testing is conducted for images of various qualities. The test cases used for system testing are listed below.

Table 2.10.3.1: Test Case 01

Test Case No	Test Case 01
Pre-requirements	PC or a Laptop with an internet connection
Description	Testing whether the system predicts the coordinates of the image correctly for the images with a background
Test Procedure	<ol style="list-style-type: none"> 1. Visit the Document Scanner Web page 2. Click on choose file 3. Select an image of a clinical laboratory report with a background 4. Select upload image
Input	<p>The image of the laboratory report with a background.</p> 

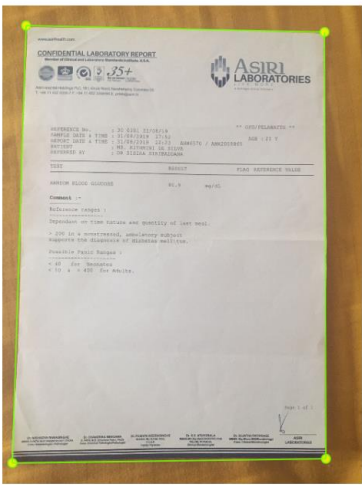
Expected Output	Automatically mark the four coordinates of the image
Actual Result	<p>Located the Coordinates of Document using OpenCV</p> 
Result of Test Case	Pass

Table 3.10.3.2: Test Case 02

Test Case No	Test Case 02
Pre-requirements	PC or a Laptop with an internet connection
Description	Testing whether the “wrap document and extract text” button appears when an image is uploaded to the scanner.
Test Procedure	<ol style="list-style-type: none"> 1. Visit the Document Scanner Web page 2. Click on choose file 3. Select an image of a clinical laboratory report with a background
Input	The image of the laboratory report with a background.

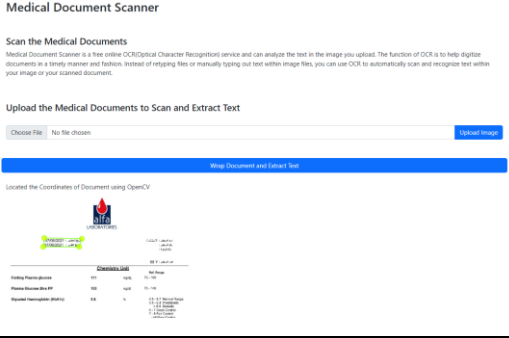
Expected Output	The “Wrap document and extract text” button should appear when an image is uploaded to the scanner.
Actual Result	<p>The “Wrap document and extract text” button appeared when an image is uploaded to the scanner.</p> 
Result of Test Case	Pass

Table 4.10.3.3: Test Case 03

Test Case No	Test Case 03
Pre-requirements	PC or a Laptop with an internet connection
Description	Testing whether the Bounding Boxes are drawn around the extracted Named Entities of the image which is uploaded
Test Procedure	<ol style="list-style-type: none"> 1. Visit the Document Scanner Web page 2. Click on choose file 3. Select an image of a clinical laboratory report with a background 4. Click the “Wrap Document and Extract Text Button” 5. Then navigate to the prediction’s web page

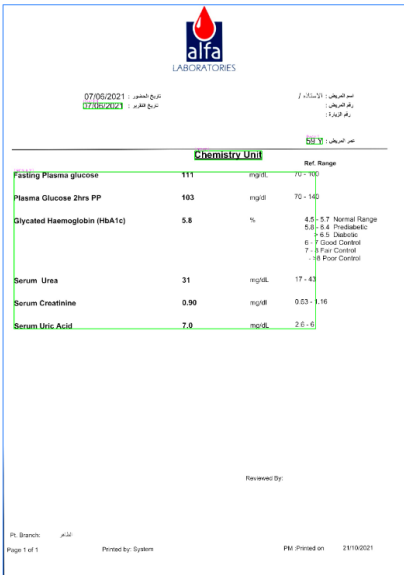
Input	The image of the laboratory report with a background.
Expected Output	Output image of the laboratory report with the Bounding Boxes drawn around the pre-defined Named Entities of the image.
Actual Result	<div>Image of the laboratory report with the Bounding Boxes drawn around the pre-defined Named Entities displayed on the prediction page.</div> <div></div>
Result of Test Case	Pass

Table 5.10.3.4: Test Case 04

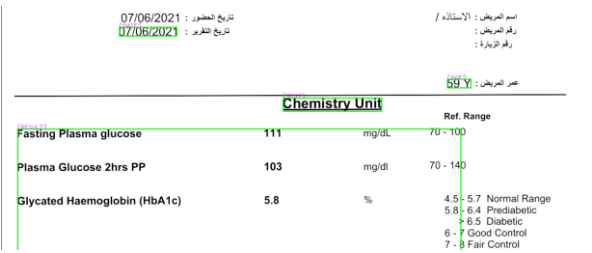
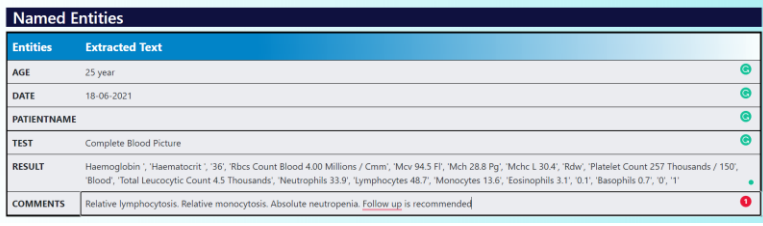
Test Case No	Test Case 04
Pre-requirements	PC or a Laptop with an internet connection
Description	Testing whether the Tag names appear with the Bounding Boxes drawn around the extracted Named Entities of the image which is uploaded
Test Procedure	<ol style="list-style-type: none"> 1. Visit the Document Scanner Web page 2. Click on choose file 3. Select an image of a clinical laboratory report with a background 4. Click the “Wrap Document and Extract Text Button” 5. Then navigate to the prediction’s web page
Input	The image of the laboratory report with a background.
Expected Output	Output image of the laboratory report with the Bounding Boxes with the Tag name should be displayed.
Actual Result	<p>Image of the laboratory report with the Tag Names displayed on the prediction page.</p>  <p>The screenshot shows a clinical laboratory report. At the top, there is patient information: 'اسم المريض : الأستاذة /' (Patient Name: Mrs. /), 'رقم المريض : ' (Patient ID:), and 'رقم الزيار : ' (Visit ID:). Below this, there is a table of chemistry results. The table has columns for 'Chemistry Unit', 'mg/dL', and 'Ref. Range'. The results are: 'Fasting Plasma glucose' (111 mg/dL, Ref. Range 70 - 100), 'Plasma Glucose 2hrs PP' (103 mg/dl, Ref. Range 70 - 140), and 'Glycated Haemoglobin (HbA1c)' (5.8 %, Ref. Range 4.5 - 5.7 Normal Range, 5.8 - 6.4 Prediabetic, 6.5 - 7 Diabetic, 7 - 8 Fair Control). Bounding boxes are drawn around the text, and tag names like 'Chemistry Unit' are displayed.</p>
Result of Test Case	Pass

Table 6.10.3.5: Test Case 05

Test Case No	Test Case 05														
Pre-requirements	PC or a Laptop with an internet connection														
Description	Testing whether the Extracted Named Entities correctly display in a tabular format														
Test Procedure	<ol style="list-style-type: none"> 1. Visit the Document Scanner Web page 2. Click on choose file 3. Select an image of a clinical laboratory report with a background 4. Click the “Wrap Document and Extract Text Button” 5. Then navigate to the prediction’s web page 														
Input	The image of the laboratory report with a background.														
Expected Output	The extracted Named Entities should appear in a table.														
Actual Result	<p>Extracted Named Entities appear in a table</p>  <p>The screenshot shows a table titled 'Named Entities' with two columns: 'Entities' and 'Extracted Text'. The table contains the following data:</p> <table border="1"> <thead> <tr> <th>Entities</th><th>Extracted Text</th></tr> </thead> <tbody> <tr> <td>AGE</td><td>25 year</td></tr> <tr> <td>DATE</td><td>18-06-2021</td></tr> <tr> <td>PATIENTNAME</td><td></td></tr> <tr> <td>TEST</td><td>Complete Blood Picture</td></tr> <tr> <td>RESULT</td><td>Haemoglobin ' , 'Haematocrit ' ; 36 ' ,Rbcs Count Blood 4.00 Millions / Cmm' , 'Mcv 94.5 Ff' , 'Mch 28.8 Pg' , 'Mchc L 30.4' , 'Rdw' , 'Platelet Count 257 Thousands / 150' , 'Blood' , 'Total Leucocytic Count 4.5 Thousands' , 'Neutrophils 33.9' , 'lymphocytes 48.7' , 'Monocytes 13.6' , 'Eosinophils 3.1' , '0.1' , 'Basophils 0.7' , '0' , '1'</td></tr> <tr> <td>COMMENTS</td><td>Relative lymphocytosis. Relative monocytosis. Absolute neutropenia. Follow up is recommended</td></tr> </tbody> </table>	Entities	Extracted Text	AGE	25 year	DATE	18-06-2021	PATIENTNAME		TEST	Complete Blood Picture	RESULT	Haemoglobin ' , 'Haematocrit ' ; 36 ' ,Rbcs Count Blood 4.00 Millions / Cmm' , 'Mcv 94.5 Ff' , 'Mch 28.8 Pg' , 'Mchc L 30.4' , 'Rdw' , 'Platelet Count 257 Thousands / 150' , 'Blood' , 'Total Leucocytic Count 4.5 Thousands' , 'Neutrophils 33.9' , 'lymphocytes 48.7' , 'Monocytes 13.6' , 'Eosinophils 3.1' , '0.1' , 'Basophils 0.7' , '0' , '1'	COMMENTS	Relative lymphocytosis. Relative monocytosis. Absolute neutropenia. Follow up is recommended
Entities	Extracted Text														
AGE	25 year														
DATE	18-06-2021														
PATIENTNAME															
TEST	Complete Blood Picture														
RESULT	Haemoglobin ' , 'Haematocrit ' ; 36 ' ,Rbcs Count Blood 4.00 Millions / Cmm' , 'Mcv 94.5 Ff' , 'Mch 28.8 Pg' , 'Mchc L 30.4' , 'Rdw' , 'Platelet Count 257 Thousands / 150' , 'Blood' , 'Total Leucocytic Count 4.5 Thousands' , 'Neutrophils 33.9' , 'lymphocytes 48.7' , 'Monocytes 13.6' , 'Eosinophils 3.1' , '0.1' , 'Basophils 0.7' , '0' , '1'														
COMMENTS	Relative lymphocytosis. Relative monocytosis. Absolute neutropenia. Follow up is recommended														
Result of Test Case	Pass														

2.11 Work Breakdown Structure and Gantt Chart

2.11.1 Work Breakdown Structure

The following Figure 2.11.1.1 depicts the work breakdown structure for the development of the Medical Document Scanner.

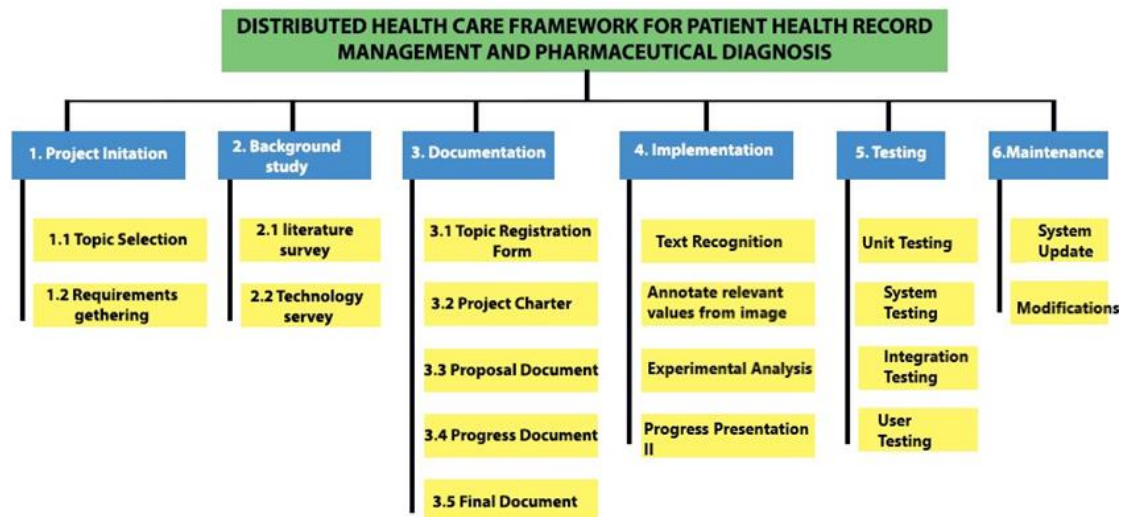


Figure 2.11.1.1: Work Breakdown Structure

2.11.2 Gantt Chart

The following Figure 2.11.2.1 shows the Gantt Chart for the development of the Medical Document Scanner component.

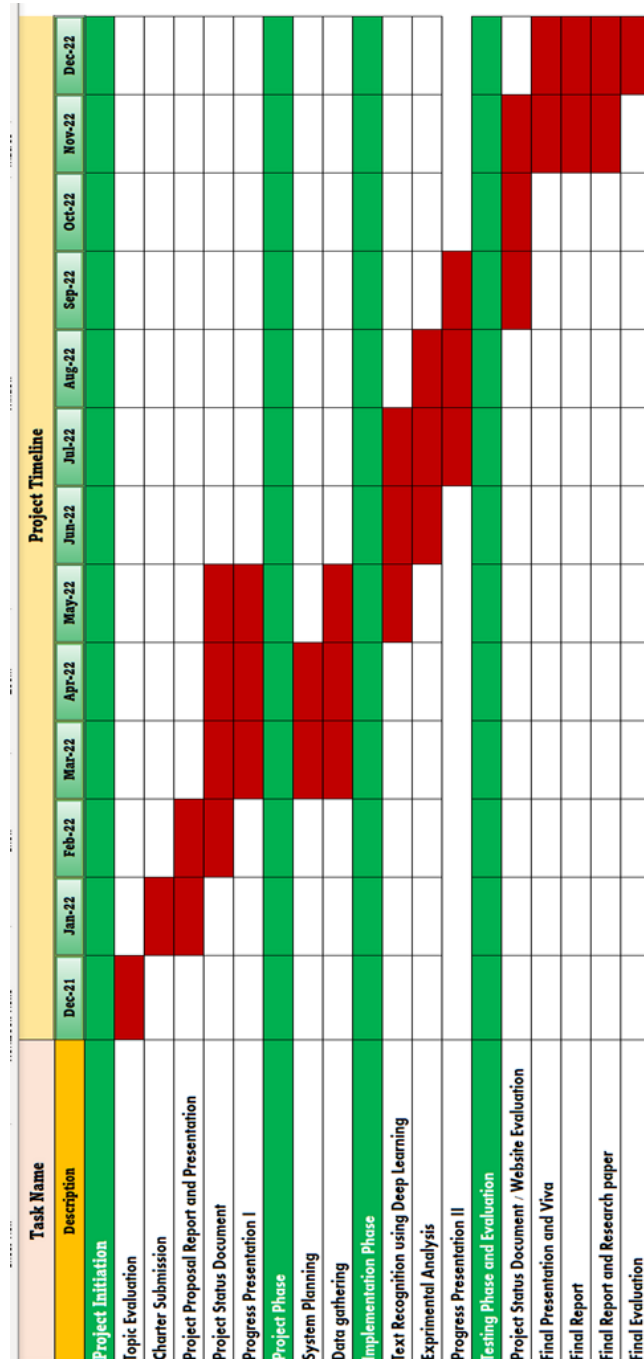


Figure 2.11.2.1: Gantt Chart

3. Results and Discussion

3.1 Results

This section summarises the results obtained from the research conducted to implement a Medical Document Scanner using Optical Character Recognition and Natural Language Processing. Since the application is developed for the healthcare community the accuracy of the results should be a considerable high value. Here, the clinical laboratory data that was retrieved will be stored on the blockchain. Medical professionals will trace the history of the patients and use the recorded data to make diagnoses. Therefore, the data obtained should be managed with care. As a result, the accuracy of the results needs to be a high value.

3.1.1 Outputs of the Medical Document Scanner

The outcomes of the medical document scanner component are displayed below. When an image of a clinical laboratory report is uploaded the four coordinates of the image will be detected using python as given in Fig. 3.1.1.1.

Located the Coordinates of Document using OpenCV

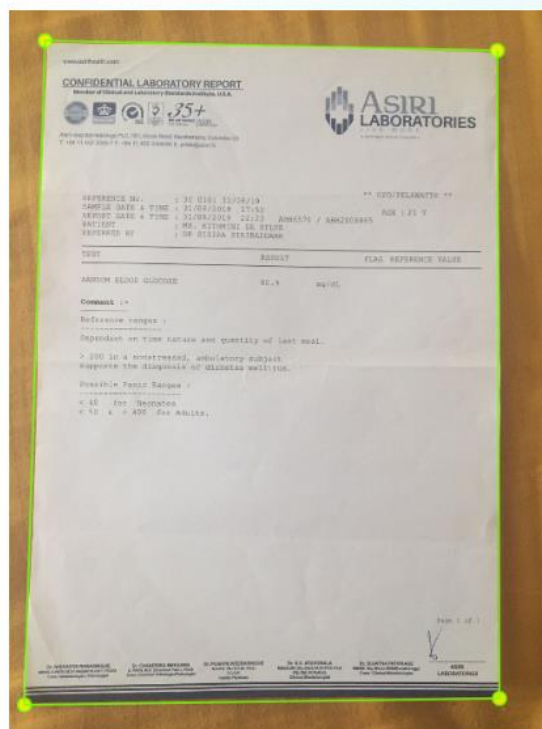


Figure 3.1.1.1: How four coordinates are detected using Python

When an image of the Clinical Laboratory Reports is uploaded, an image with bounding boxes drawn around the important named entities will be received as the output together with the relevant tag name as given in Fig. 3.1.1.2.



Figure 3.1.1.2: Image of the Document with the Bounding Boxes and the Tag name

Finally, the extracted Named Entities are displayed in a tabular format as given in Fig. 3.1.1.3. This tabular data has editing capability as well.

Named Entities		
Entities	Extracted Text	
AGE	59 y	
DATE	07/06/2021	
PATIENTNAME		
TEST	Chemistry Unit	
RESULT	Fasting Plasma Glucose 111, 'Plasma Glucose 2Hrs Pp 103', 'Glycated Haemoglobin', 'Serum Urea', '17', 'Serum Creatinine', 'Serum Uric Acid 7.0	
COMMENTS		

Figure 3.1.1.3: Extracted Named Entities in tabular format

3.2 Research Findings

The findings from the performed research study are highlighted in this section. The results of the study demonstrate that the following two factors have a significant impact on the model's accuracy.

1. The size of the training data set used
2. The quality of the image uploaded

The total accuracy of the predicted outcomes will rise as more training data are used to train the model. More training data increase the system's accuracy. Similar to this, the image's quality has a considerable influence on how accurate the predictions are. The accuracy of the textual data which is extracted from the high-quality images is higher compared to that of the low-quality images. Here, image pre-processing techniques are used to improve image quality to increase the model's capability to do predictions.

Another research finding is that Pytesseract is having limitations when extracting text from rotated or skewed documents. Pytesseract assumes text need to be aligned to get better performance. Therefore, before delivering an image to a prediction model, its quality should be improved, and the alignments should be corrected.

3.3 Discussion

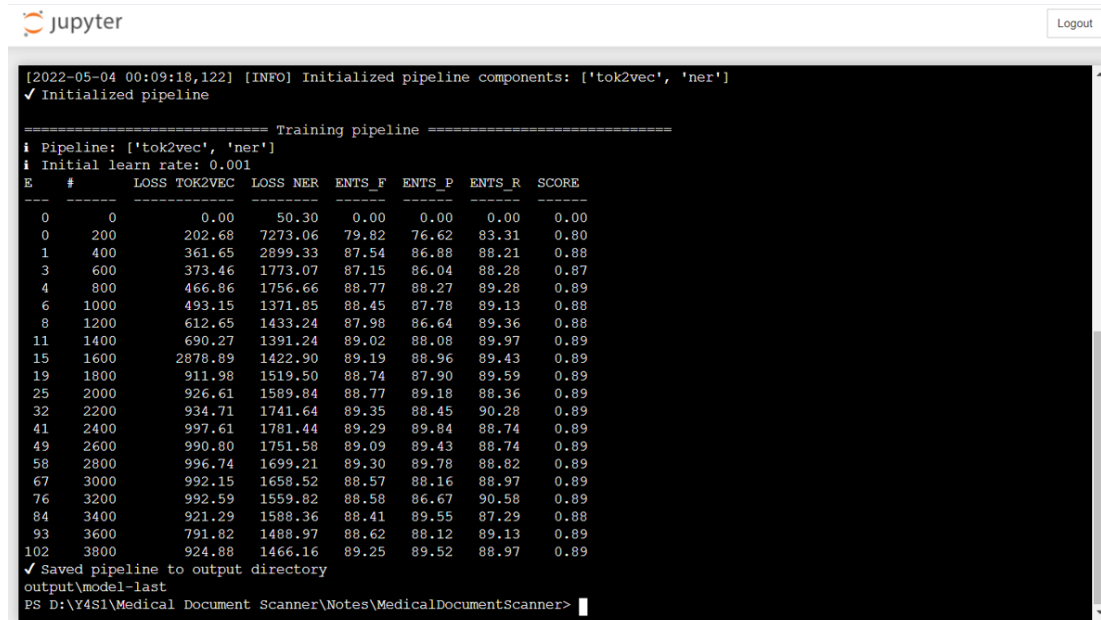


Figure 3.3.1: Accuracy of the trained Spacy Pipeline

The Named Entity Recognition Spacy Model trained for the Named Entity extraction has the following accuracies. The model has a precision (**ENTS_P**) of 89.52%. The recall (**ENTS_R**) of the model is 88.97% while the F-score (**ENTS_F**) is 89.25%.

Precision measures how accurate the trained model is. It is the proportion of all detected positives to those that were accurately classified as positives (true positives). How many of the predicted entities have the proper labels is shown by the accuracy metric [17].

$$\textbf{Precision} = \textbf{\#True_Positive} / (\textbf{\#True_Positive} + \textbf{\#False_Positive})$$

Recall determines how well the model can detect true positives. It is the proportion of expected true positives to what was tagged. How many of the predicted entities are accurate is shown by the recall measure.

$$\textbf{Recall} = \textbf{\#True_Positive} / (\textbf{\#True_Positive} + \textbf{\#False_Negatives})$$

Precision and Recall are factors that affect the F1 score. It is necessary when a balance between Precision and Recall is necessary.

$$\textbf{F1 Score} = 2 * \textbf{Precision} * \textbf{Recall} / (\textbf{Precision} + \textbf{Recall})$$

3.4 Future Work

We can add more training data to the dataset as part of our development strategy in the future. By continuing to use the same framework and including new training data, the model's accuracy can be increased. The F-score for the currently trained Named Entity Recognition model is 89.52%, although it would be preferable if it were over 90%. The quality of the image that is uploaded has a significant impact on how accurate the models are. To improve the image's quality, further image pre-processing methods can be incorporated into the model. Several additional data preparation frameworks,

particularly for the data cleaning phase, can also be used. It is anticipated that by adding such enhancements, prediction accuracy would increase in the future.

6. Conclusion

The pandemic has exposed healthcare's limitations and emphasized the significance of automating the healthcare domain. The relevance of a healthcare framework has been demonstrated through research on current literature and public surveys.

The proposed solution is a distributed healthcare framework that would securely store, access, and share patient health records across several hospitals and deliver healthcare services to patients when they are at home. A Virtual Conversational chatbot, for example, may send regular reminders to take medicine based on the most recent prescription kept in the Blockchain, as well as assist in the identification of medications based on taken photographs.

Since most medical records are printed, extracting data, and manually inputting it into the Blockchain might result in inaccuracies due to human error.

As a result, a medical document scanner model will be introduced, which will use Deep Learning and Natural Language Processing techniques to extract textual data automatically. The data will be annotated, and the significant entities will be extracted. This feature will help the user to save time and avoid mistakes.

All the above components will be combined into a single system, and this solution will, of course, bridge the gap that exists in the field of healthcare and manage the challenges that arise during a worldwide pandemic.

References

- [1] M. Y. Jabarulla and H.-N. Lee, "A Blockchain and Artificial Intelligence-Based, Patient-Centric Healthcare System for Combating the COVID-19 Pandemic: Opportunities and Applications," *MDPI*, vol. 9, no. 8, p. 1019, 2021.
- [2] Y. Zhang, M. C. L. Z. R. Zhang, L. Meng, D. Gao and Y. Zhang, "Research on electronic medical record access control based on blockchain," *International Journal of Distributed Sensor Networks*, vol. 15, no. 11, p. 1550147719889330, 2019.
- [3] W. Xue, Q. Li and Q. Xue, "Text Detection and Recognition for Images of Medical Laboratory Reports With a Deep Learning Approach," *IEEE Access*, vol. 8, pp. 407-416, 2019.
- [4] W. Xue, Q. Li, Z. Zhang, Y. Zhao and H. Wang, "Table Analysis and Information Extraction for Medical Laboratory Reports," *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pp. 193-199, 2018.
- [5] S. Karthikeyan, A. G. S. d. Herrera, F. Doctor and A. Mirza, "An OCR Post-correction Approach using Deep Learning for Processing Medical Reports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2574 - 2581, 2021.
- [6] B. Dash, "A hybrid solution for extracting information from unstructured data using optical character recognition (OCR) with natural language processing (NLP)," 2021.
- [7] S. Moon, S. Liu, D. Chen, Y. Wang, D. L. Wood, R. Chaudhry, H. Liu and P. Kingsbury, "Salience of Medical Concepts of Inside Clinical Texts and Outside

- Medical Records for Referred Cardiovascular Patients," *Journal of Healthcare Informatics Research*, vol. 3, no. 2, pp. 200-219, 2019.
- [8] A. C. Özgen, M. Fasounaki and H. K. Ekenel, "Text Detection in Natural and Computer-Generated Images," *2018 26th signal processing and communications applications conference (SIU)*, pp. 1-4, 2018.
- [9] N. I. Widiastuti, "Convolution Neural Network for Text Mining and Natural Language Processing," *IOP Conference Series: Materials Science and Engineering*, vol. 662, no. 5, p. 052010, 2019.
- [10] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *JMIR medical informatics*, vol. 7, no. 2, p. 12239, 2019.
- [11] A. Mishra, S. Shekhar, A. K. Singh and A. Chakraborty, "OCR-VQA: Visual Question Answering by Reading Text in Images," *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947-952, 2019.
- [12] D. v. Strien, K. Beelen, M. C. Ardanuy, K. Hosseini, B. McGillivray and G. Colavizza, "Assessing the Impact of OCR Quality on Downstream NLP Tasks," *12th International Conference on Agents and Artificial Intelligence*, vol. 1, pp. 484-496, 2020.
- [13] E. Boros, A. Hamdi, E. L. Pontes, L. A. Cabrera-Diego, J. G. Moreno, N. Sidere and A. Doucet, "Alleviating Digitization Errors in Named Entity Recognition for Historical Documents," *Proceedings of the 24th conference on computational natural language learning*, pp. 431-441, 2020.
- [14] Y. Luo, F. Xiao and H. Zhao, "Hierarchical Contextualized Representation for Named Entity Recognition," *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 5, pp. 8441-8448, 2020.
- [15] G. Xu, C. Wang and X. He, "Improving Clinical Named Entity Recognition with Global Neural Attention," *Asia-Pacific Web (APWeb) and Web-Age Information*

Management (WAIM) Joint International Conference on Web and Big Data, pp. 264-279, 2018.

- [16] E. Abdelmaksoud, A. Gadallah and A. Asad, "Mendeley Data," 7 January 2022. [Online]. Available: <https://data.mendeley.com/datasets/bygfmk4rx9/2>. [Accessed 10 October 2022].
- [17] "Characteristics and limitations for using custom named entity recognition," Microsoft, 19 July 2022. [Online]. Available: <https://learn.microsoft.com/en-us/legal/cognitive-services/language-service/cner-characteristics-and-limitations>. [Accessed 11 October 2022].

Appendices

Survey on Health Care issues confront during COVID-19 pandemic

Hello Everyone, We are final year Software Engineering Undergraduates at Sri Lanka Institute of Information Technology. We are conducting this research to gather information on health care problems confronted by the general public during COVID-19.

 desilvakithmini@gmail.com (not shared) [Switch account](#)



Do you have any healthcare issues as a result of the COVID-19 pandemic?

- ☐ Yes
- ☐ No

The survey was conducted among the public to gather information about healthcare issues faced during COVID-19

Do you believe that healthcare automation is critical in the occurrence of a pandemic?

☐ Yes

☐ No

Which problem do you think will occur if we propose an automated solution to store patient data in digital format?

☐ Storing sensitive data will lead to data breaches

☐ It is difficult to electronically store patient medical data printed in papers like lab test reports

☐ Both Problems are crucial

☐ Other: _____

"Since health solution has not yet proposed for pharmaceutical diagnosis, it is a must to visit the doctor even during COVID-19". Do you agree with this statement?

	1	2	3	4	5	
Strongly Disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly Agree

Submit

Clear form

The survey was conducted among the public to gather information about healthcare issues faced during COVID-19



Distributed Health Care Framework for Patient Health Record Management and Pharmaceutical Diagnosis

We are final year Software Engineering Undergraduates at the Sri Lanka Institute of Information Technology, New Kandy Road, Malabe, Sri Lanka. We are conducting this research to gather information on health care problems confronted by the medical practitioners during COVID-19. Please spare 5 minutes of your valuable time to participate in the survey. The information is being gathered solely for research purposes, and your responses are greatly welcomed.

 **desilvakithmini@gmail.com** (not shared) [Switch account](#)



* Required

Name *

Your answer

Designation *

A survey was conducted among Medical Practitioners to gather information about healthcare issues faced during COVID-19

Designation *

Your answer _____

Specialty *

Your answer _____

Hospital / Current working place *

Your answer _____

As healthcare practitioners do you face any healthcare issues during the COVID-19 pandemic? *

☐ Yes

☐ No

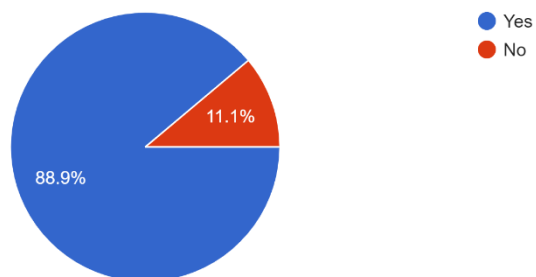
Is there any electronic health record systems in your current working hospital or have you seen one in any Sri Lankan hospital? *

☐ Yes

A survey was conducted among Medical Practitioners to gather information about healthcare issues faced during COVID-19

As healthcare practitioners do you face any healthcare issues during the COVID-19 pandemic?

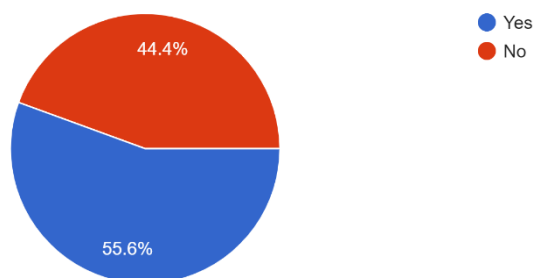
9 responses



Responses gathered by Medical Practitioners about the healthcare issues faced during COVID-19

Is there any electronic health record systems in your current working hospital or have you seen one in any Sri Lankan hospital?

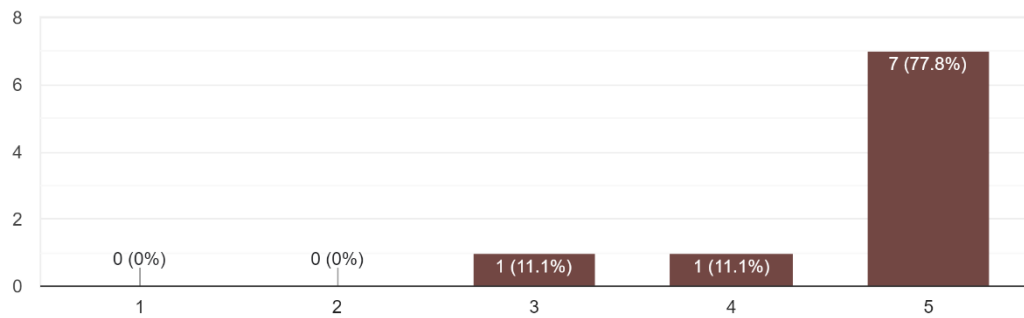
9 responses



Responses gathered by Medical Practitioners about the importance of electronic health record systems

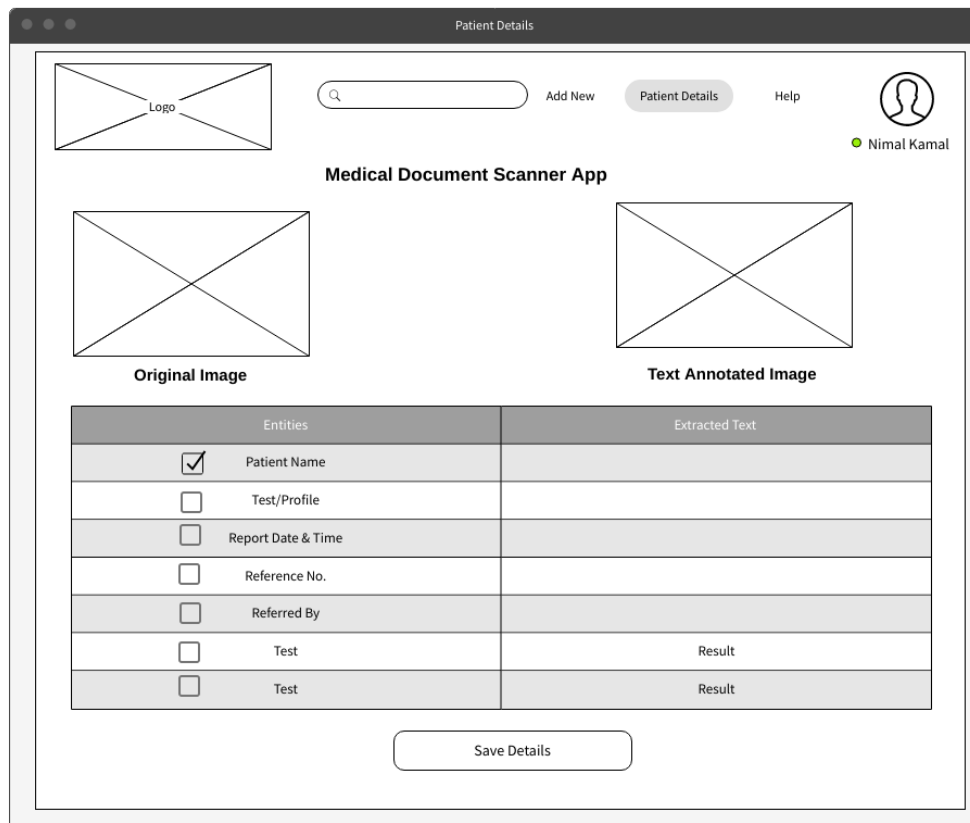
Do you agree that storing, accessing and sharing patient details in electronic format is critical in the occurrence of a pandemic?

9 responses



Responses gathered by Medical Practitioners about the importance of electronic health record systems

Low Fidelity wireframe designed for interface to upload Medical Documents



Low Fidelity wireframes designed to display the extracted Named Entities