

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

Project ID: 2022-110

Project Proposal Report

De Silva K.H.K.L. – IT19006994

B.Sc. (Hons) Degree in Information Technology

Department of Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

January 2022

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

Project ID: 2022-110

Project Proposal Report

De Silva K.H.K.L – IT19006994

Supervised by – Mr. Jeewaka Perera

Co – Supervisor - Ms. Laneesha Ruggahakotuwa

B.Sc. (Hons) Degree in Information Technology

Department of Software Engineering


Sri Lanka Institute of Information Technology

Sri Lanka

January 2022

## Declaration

We 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.

| 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.

Name of supervisor: Mr. Jeewaka Perera

Name of co-supervisor: Ms. Laneesha Ruggahakotuwa

.....

.....

Signature of the supervisor:

Date

(Mr. Jeewaka Perera)

.....

.....

Signature of the supervisor:

Date

(Ms. Laneesha Ruggahakotuwa)

## **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.

## **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.

## **Abstract**

According to recent studies, covid19 has 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 of the 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 proposing 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*

## Table of Contents

|  |      |
|--|------|
| Declaration.....   | i    |
| List of Figures .....  | vii  |
| List of Tables .....   | viii |
| List of Abbreviations .....  | ix   |
| List of Appendices .....   | x    |
| 1. Introduction .....  | 1    |
| 1.1 Background Study .....   | 1    |
| 1.2 Literature Survey.....   | 3    |
| 1.3 Research Gap .....   | 7    |
| Worldview Mobile Complete .....  | 9    |
| 1.4 Research Problem .....   | 10   |
| 2. Objectives.....   | 12   |
| 2.1 Main Objective .....   | 12   |
| 2.2 Specific Objectives .....  | 12   |
| 3. Methodology.....  | 14   |
| 3.1 Project Overview.....  | 14   |
| 3.2 System Overview Diagram .....  | 15   |
| 3.3 System Overview.....   | 16   |
| 3.4 Design Phase - Individual Component .....                                | 17   |
| 3.4.1 Text Extraction and Annotate relevant data from Medical Documents..... | 17   |
| 3.4.2 Entity extraction from the captured data .....                         | 18   |
| 3.4.3 Ideas to improve the model accuracy .....                              | 18   |
| 3.5 Software Development Process.....  | 19   |
| 3.6 Feasibility Study .....  | 20   |
| 3.7 Technology Selection .....   | 22   |
| 3.8 Work Breakdown Structure and Gantt Chart .....                           | 23   |
| 3.8.1 Work Breakdown Structure .....   | 23   |
| 3.8.2 Gantt Chart.....   | 24   |
| 4. Project Requirements .....  | 25   |
| 4.1 Functional Requirements.....   | 25   |
| 4.2 Non-Functional Requirements.....   | 26   |
| 5. Business Potential .....  | 27   |

|     |  |    |
|-----|--|----|
| 1.1 | Targeted Audience.....                         | 27 |
| 1.2 | Benefits from the system .....                 | 27 |
| 6.  | Description of personnel and facilities .....  | 28 |
| 7.  | Budget and budget justification (if any) ..... | 29 |
| 8.  | Conclusions and Recommendations.....           | 30 |
|     | Reference List.....                            | 31 |
|     | Appendices.....                                | 34 |



## List of Figures

|  |           |
|--|-----------|
| <i>Figure 1.4.1: Summary of the responses whether healthcare issues occur during pandemic.....</i>   | <i>10</i> |
| <i>Figure 1.4.2: Summary of the responses whether healthcare automation is critical.....</i>         | <i>10</i> |
| <i>Figure 1.4.3: Summary of the responses on drawbacks in manually entering data into EHR .....</i>  | <i>11</i> |
| <i>Figure 1.4.4: Summary of the responses on the importance of a medical document scanner.....</i>   | <i>11</i> |
| <i>Figure 3.1.1: Project Overview Diagram .....</i>  | <i>14</i> |
| <i>Figure 3.2.1: System Overview Diagram.....</i>  | <i>15</i> |
| <i>Figure 3.4.1.1: Steps followed in Textual Data extraction from Medical Document Scanner .....</i> | <i>17</i> |
| <i>Figure 3.5.1: The Software Development Life Cycle.....</i>  | <i>19</i> |
| <i>Figure 3.8.1.1: Work Breakdown Structure.....</i>   | <i>23</i> |
| <i>Figure 3.8.2.1: Gantt Chart.....</i>  | <i>24</i> |

## List of Tables

| <b>Table</b>   | <b>Page</b> |
|--|-------------|
| <i>Table 1.3.1: Summary of the responses on drawbacks in manually entering data into EHR..</i> | 8           |
| <i>Table 2.3.2: Limitations of the existing systems in the market .....</i>                    | 9           |
| <i>Table 6.1: Resource personnel for Development.....</i>                                      | 28          |
| <i>Table 6.2: Resource personnel for External Supervising.....</i>                             | 28          |
| <i>Table 7.1: Budget and Budget Justification.....</i>   | 29          |

## List of Abbreviations

| Abbreviations | Description                   |
|---------------|-------------------------------|
| EHR           | Electronic Health Record      |
| OCR           | Optical Character Recognition |
| OMR           | Outside Medical Records       |
| CNN           | Convolution Neural Network    |
| NLP           | Natural Language Processing   |
| NER           | Named Entity Recognition      |

## List of Appendices

| Appendix | Description  | Page |
|----------|--|------|
|          | <i>Appendix - A Additional Survey Responses gathered during the Research Survey...</i> | 34   |
|          | <i>Appendix - B Supervisor and Co-Supervisor's Endorsement.....</i>                    | 36   |

# **1. Introduction**

## **1.1 Background Study**

The world is facing numerous challenges in the field of healthcare due to COVID19. To control the spread of COVID-19, numerous countries had to close their borders, implement lockdowns, and employ social distance. 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 COVID19 has forced the countries, government bodies and researchers to rethink in applying modern digital solutions to the healthcare domain [1]. In recent years researchers are focusing more on the use of Blockchain and Machine Learning approaches for digital transformation in the healthcare field [1].

As a result of the pandemic, the amount of digitally stored patient data has grown significantly [2]. Data surveillance, telemedicine, remote pharmaceutical diagnosis, and strategy innovation are all part of the digital healthcare ecosystem that need 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 use Blockchain and Machine Learning-Based Healthcare framework which provides healthcare services to medical practitioners and patients while staying at home 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, leakage of sensitive data can be overcome with the use of blockchain-based electronic health records [3]. Conversion of the existing patient data into electronic health records is a challenging task since most of the medical documents such as lab test reports, prescriptions from hospitals are in printed format. Converting these data into Electronic Health Records (EHR) and entering these details into blockchain often needs 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 staying at home. Virtual conversational agents like chatbots will serve as virtual healthcare assistants to make things easier.

This proposal paper will give an overall idea of the proposed solution and focus more on the Medical Document Scanner component which is based on Deep Learning 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 COVID19.

The document is organized as follows. Section 1 includes Introduction, Background Study and Literature review, Research gap and overview of the Research Problem. Section 2 provides an overview of the objectives. Section 3 will explore the Methodology. The final sections will explain the business requirements and budget justification and conclude the document with the Reference List and Appendices.

## 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 the research paper published in the year 2019 by the members of IEEE namely WENYUAN XUE, QINGYONG LI, AND QIYUAN XUE titled "Text Detection and Recognition for Images of Medical Laboratory Report with a Deep Learning Approach" [4] suggested a text detection approach with the use of patch-based training strategy and a concatenation structure which can combine the features of the deep and shallow layers in Neural network. This study was carried out 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 the research paper "A Study on Optical Character Recognition Techniques" [5] looked at a theoretical and mathematical model to address one of the most difficult challenges in optical character identification: scale, translate and rotate quality in optical alphabet detection. In addition, the potential deployments of OCR algorithms were investigated.

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 Srinidhi Karthikeyan, Alba G. Seco de Herrera, Faiyaz Doctor and Asim Mirza developed a deep neural network-based self-supervised pre-training

model on their research work on "An OCR Post-correction Approach using Deep Learning for Processing Medical Reports" [6]. This bi-directional encoder has been designed to predict concealed text and fill in gaps in non-transcribable areas of the page. Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu published a study [7] that emphasized the usage of a single deep neural network for fast text detection. They have created a new model named "TextBoxes," which is a fully convolutional network that detects text from beginning to end. In cluttered backgrounds, this proposed text can create very steady and efficient word suggestions. Here the three major tasks such as text detection, word spotting, and end-to-end recognition were validated through evaluations and comparisons on benchmark datasets.

Moreover, Eva D'hondt, Cyril Grouin and Brigitte Grau presented a novel approach in their research work on "Generating a Training Corpus for OCR Post-Correction Using Encoder-Decoder Model" [8] for automatic correction of orthographic errors caused by OCR driven text. In this study, they presented a zero annotated data strategy for OCR post-correction. Instead of learning from annotated data, they employed the biLSTMs encoder-decoder model and introduced their material. To create a robust character-based language model, the model was trained using clean or substantially cleaned data. Because of the organizational and technological constraints that exist, information in Outside Medical Records (OMR) is underutilized.

According to the study "Salience of Medical Concepts of Inside Clinical Texts and Outside Medical Records for Referred Cardiovascular Patients" [9] by Sungrim Moon, Sijia Liu, David Chen, Yanshan Wang, Douglas L. Wood, Rajeev Chaudhry, Hongfang Liu, and Paul Kingsbury, it has found the clinical concepts contained in OMR are beneficial for Cardiovascular medicine. The study can be regarded as the first step toward automated data extraction from OMRs generated by a variety of healthcare providers.

The methods of detecting text can be divided into two groups: texture-based approaches and region-based methods [9]. According to the research paper on "Text



Detection in Natural and Computer-Generated Images" done by Azmi Can Özgen, Mandana Fasounaki, Hazim Kemal Ekenel of Istanbul Technical University [10] it can be seen, that encouraging results can be produced when a variety of image processing algorithms are applied in text detection. In the research paper, they 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. "Convolution Neural Network for Text Mining and Natural Language Processing " journal by N I Widiastuti Department of Informatics Engineering, Universitas Komputer Indonesia [11] provides an overview on the application of Convolution Neural Network 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.

In the year 2019, Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani [12] proposed an NLP based free-text extraction from notes related to chronic diseases. The study investigates the difficulties of NLP approaches as well as clinical narratives. The analysis found that machine learning is becoming more popular, 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 handled directly using machine learning algorithms.

In 2019 Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, Anirban Chakraborty did a research study on "OCR-VQA: Visual Question Answering by Reading Text in Images" [13]. They 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.

In the year 2019, Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza researched "Assessing the Impact of OCR Quality on Downstream NLP Tasks" [14] and 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.

As per the above-mentioned readings and the thorough the 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 [5], [7], [10], [11], [13], [14], 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 [6] 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 [4] 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 review work on performance examination utilizing mathematical models was done in the research study [5]. However, the accuracy and identification are insufficient for practical deployment. As a result, for improved outcomes, the current study should include more OCR techniques and algorithms.

The initial step toward autonomous data extraction from OMRs was completed by the research study [9]. 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.

| Reference ID | Modelled for Healthcare Domain-Specific words | Transcribe the scanned documents/images where text is skewed | Automatic word suggestions | Automatic data correction |
|--------------|---|--|----------------------------|---------------------------|
| Research [4] | ✓   | ✗  | ✓                          | ✓                         |
| Research [5] | ✗   | ✗  | ✗                          | ✗                         |
| Research [6] | ✗   | ✓  | ✓                          | ✓                         |
| Research [9] | ✓   | ✗  | ✗                          | ✗                         |
| Our Solution | ✓   | ✓  | ✓                          | ✓                         |

*Table 1.3.1: Summary of the responses on drawbacks in manually entering data into EHR*

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.

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.

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

*Table 2.3.2: Limitations of the existing systems in the market*

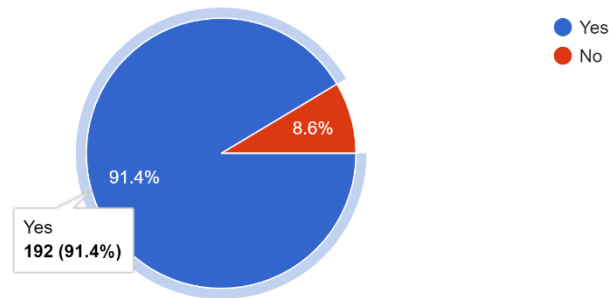
Most current medical document scanners scan the document's image and convert it to a soft copy. However, many of these methodologies have not been addressed in textual data extraction from such publications. Most other document scanners, on the other hand, are built for general use and not specifically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

#### 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% says that they face healthcare issues during the pandemic.

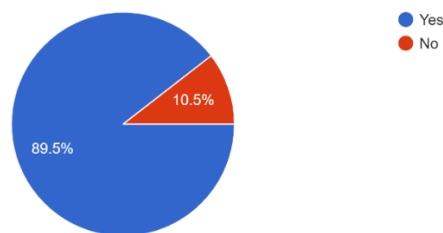
Do you have any healthcare issues as a result of the COVID-19 pandemic?  
210 responses



*Figure 1.4.1: Summary of the responses whether healthcare issues occur during pandemic*

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

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



*Figure 1.4.2: Summary of the responses whether healthcare automation is critical*

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

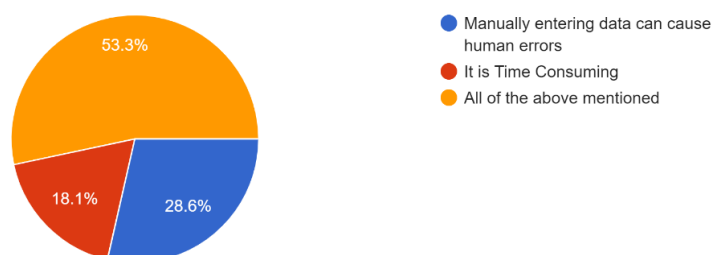


Figure 1.4.3: Summary of the responses on drawbacks in manually entering data into HER

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

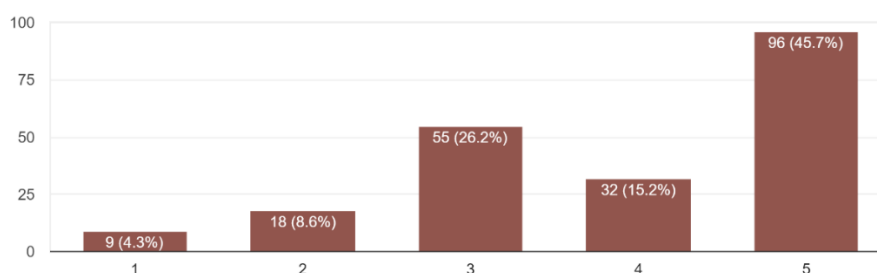


Figure 1.4.4: Summary of the responses on the importance of a medical document scanner

In addition, most of the 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.

## **2. Objectives**

### **2.1 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's 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 distance.

### **2.2 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 of the medical documents such as lab test reports are in printed format, and it is a 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.

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

Most of the 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 the proposed solution to capture data from such illegible papers.

### 3. Methodology

#### 3.1 Project Overview

The suggested system is designed to meet the challenges that the healthcare domain confronts during the COVID19 pandemic, as well as to provide healthcare solutions that ensure service continuity while people remain at home and maintain social distance. The proposed distributed healthcare framework would include secure patient health record management and pharmaceutical diagnostic capabilities.

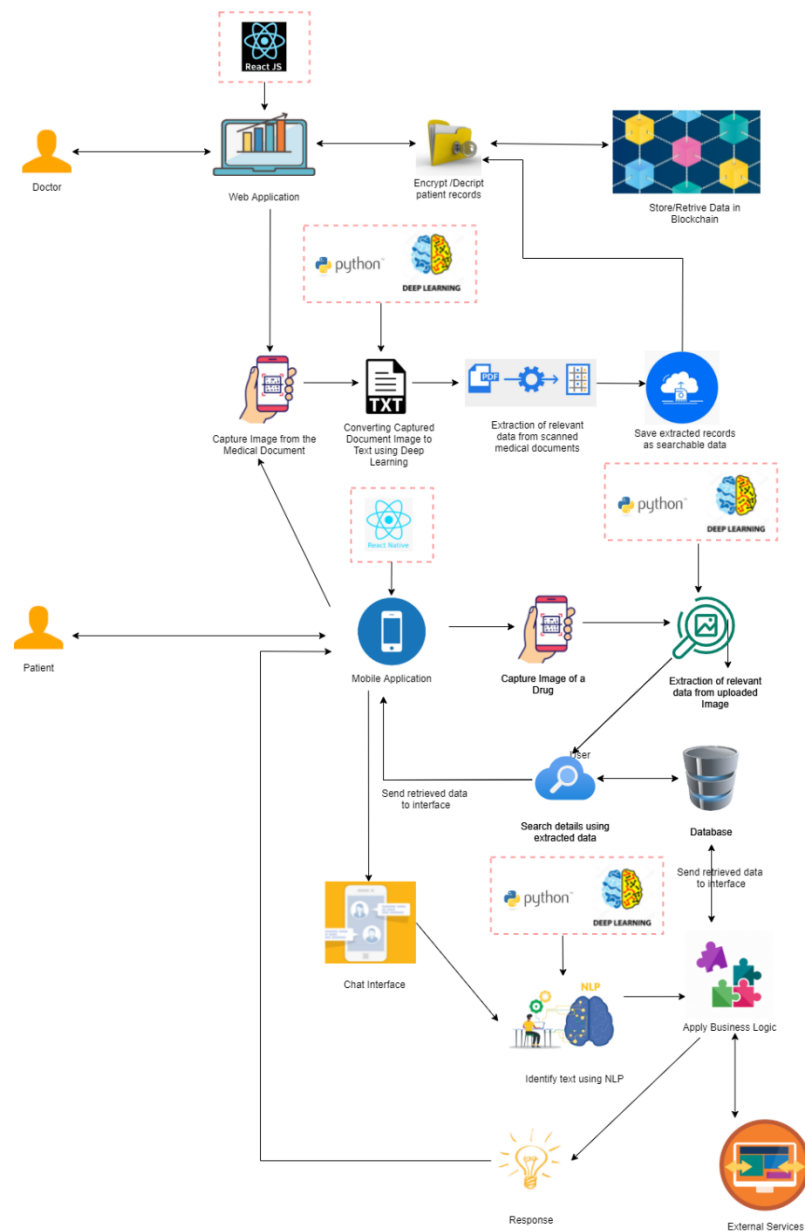


Figure 3.1.1: Project Overview Diagram

## 3.2 System Overview Diagram

Figure 3.2.1 depicts the System Overview Diagram of the Medical Document Scanner component

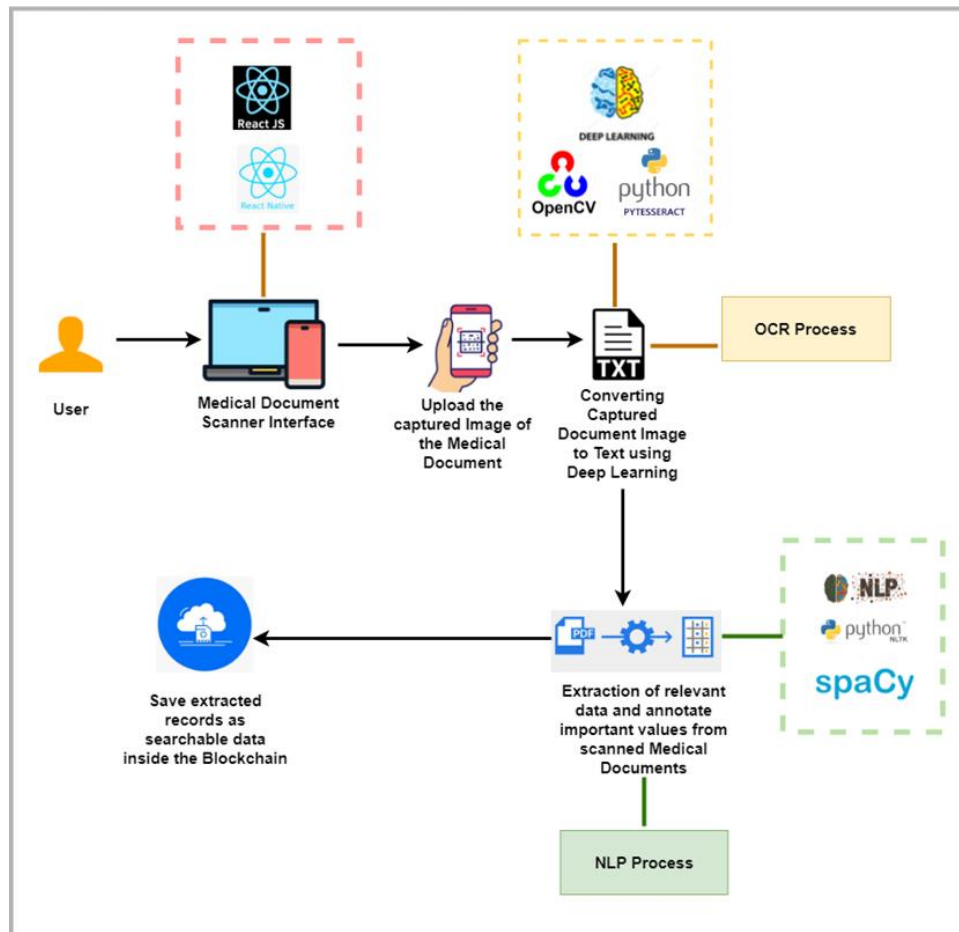


Figure 3.2.1: System Overview Diagram

### 3.3 System Overview

This component works as a Medical Document Scanner, which can extract textual data from printed Medical Documents such as Lab Test reports, restructure them and annotate the important values. 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 documents 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 data will then be transformed into a searchable or editable format and stored within the Blockchain.

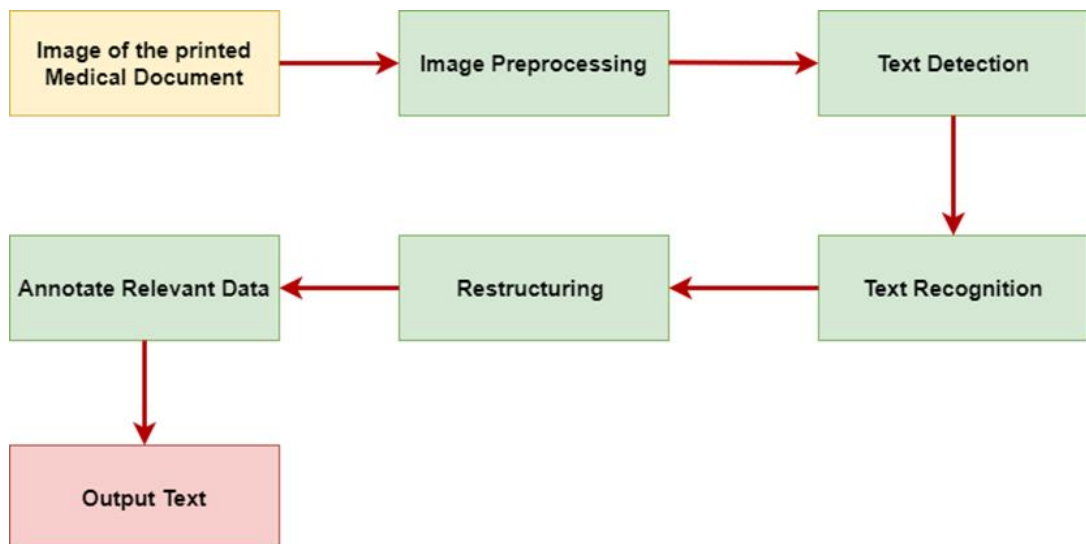
### 3.4 Design Phase - Individual Component

#### 3.4.1 Text Extraction and Annotate relevant data from Medical Documents

Optical Character Recognition (OCR) is a technology that captures text present in images into a machine-readable format. Image preprocessing is the initial stage in the optical character recognition pipeline, and it removes any image deformities such as noise. Text detection is the second stage, which captures the regions where text is present. Under the Text recognition step, the module will recognize the text present in the area which is detected previously under the text detection step.

Finally, the image will be reorganized to appear identical to the original image. Optical character recognition in deep learning will be used for textual extraction.

An output text will be retrieved at the end of the optical character recognition processing.



*Figure 3.4.1.1: Steps followed in Textual Data extraction from Medical Document Scanner*

### **3.4.2 Entity extraction from the captured data**

Following text extraction from the taken image, the text will be cleaned using several methods before being given to a Natural Language Processing model trained for Named Entity Recognition (NER). Important values and entities will then be retrieved from the extracted data. The data from the image will be extracted and converted into a Data frame by this model. The data frame will then be converted to content, and the named entities from the model will be obtained. Then each word will be tagged, token data frames will be joined with Pytesseract data, and the bounding box image annotating the important values will be obtained.

### **3.4.3 Ideas to improve the model accuracy**

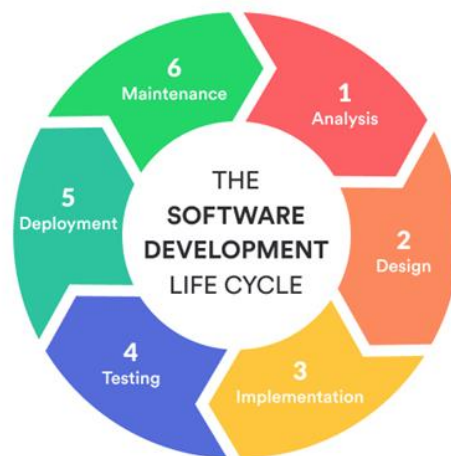
1. More training data should be added.

The model's accuracy may be improved by using the same framework and adding more training data.

2. Change data preparation framework for cleaning data

### 3.5 Software Development Process

The software development lifecycle 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. 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 3.5.1: The Software Development Life Cycle [15]*

### **3.6 Feasibility Study**

- **Economic Feasibility**

The proposed solution is aimed at hospital chains across the country, and physicians, medical practitioners, and patients would all benefit upon 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.

### **3.7 Technology Selection**

- The textual extraction of data from medical documents will be done using optical character recognition techniques using Deep Learning. This will be accomplished with the help of OpenCV and Pytesseract.
- The medical documents will be loaded using OpenCV technology, and the text will be extracted using Pytesseract.
- The deep learning model will be trained to annotate the important values using Natural Language Processing techniques. To extract the name entities, the Named Entity Recognition Pipeline will be used with the Spacy technology.
- Since we continue to research the components and technologies to be employed, certain technologies can be subject to change.

### 3.8 Work Breakdown Structure and Gantt Chart

#### 3.8.1 Work Breakdown Structure

The following Figure 3.8.1.1 depicts the work breakdown structure for the development of the Medical Document Scanner.

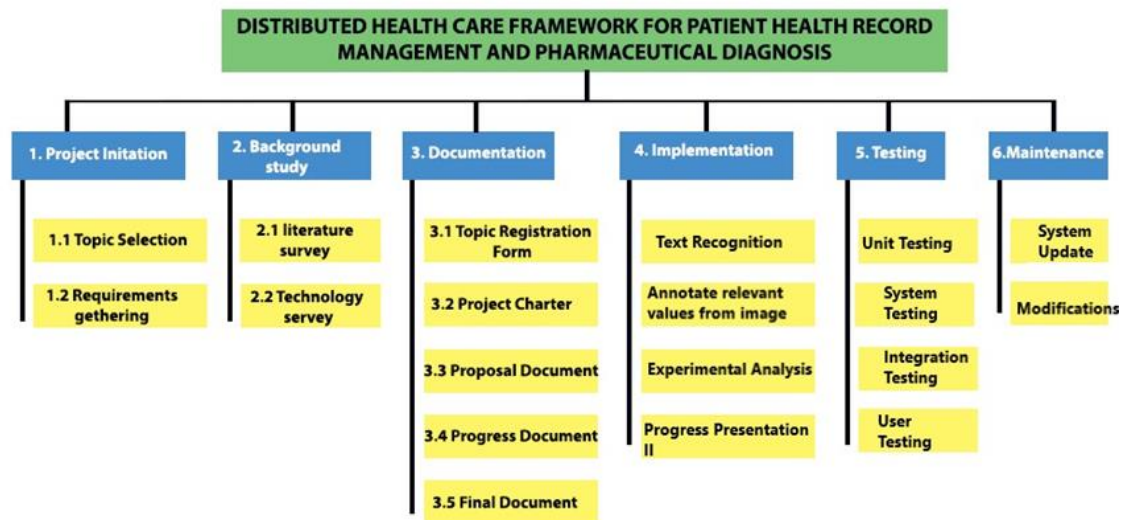


Figure 3.8.1.1: Work Breakdown Structure

### 3.8.2 Gantt Chart

The following Figure 3.8.2.1 shows the Gantt Chart for the development of the Medical Document Scanner component.

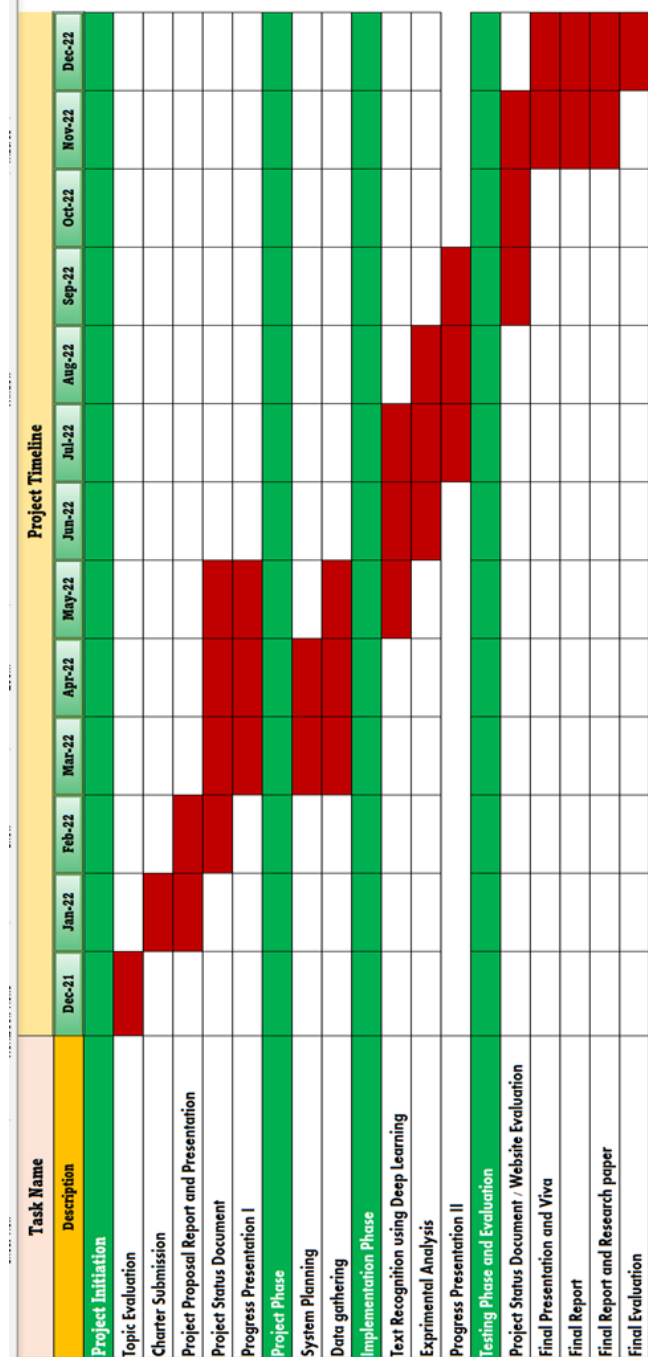


Figure 3.8.2.1: Gantt Chart

## **4. Project Requirements**

### **4.1 Functional Requirements**

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

Textual data will be captured extracted from captured images of printed medical documents such as lab test 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 will be rearranged in this module to appear the same as the original picture.

3. Extract text from distorted documents

With time, most of the printed medical documents get distorted. This suggested module will also be trained in how to extract data from skewed documents.

4. Annotate important values from the recognized text

Not only that, in this component important values will be identified and annotated based on the gathered data.

## **4.2 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 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.

## **5. Business Potential**

### **1.1 Targeted Audience**

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

### **1.2 Benefits from the system**

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

## 6. Description of personnel and facilities

Resource personnel for the development team and the tasks assigned to them are as follows:

| Registration Number | Name                    | Assigned Task  |
|---------------------|-------------------------|--|
| IT19004778          | Wickramarathna W.G.M.S. | <ul style="list-style-type: none"><li>• Development and Testing for subcomponent based on Blockchain</li><li>• Integration of the relevant component to the final system</li></ul>     |
| IT19006994          | De Silva K.H.K.L        | <ul style="list-style-type: none"><li>• Development and Testing of Medical Document Scanner Subcomponent</li><li>• Integration of the relevant component to the final system</li></ul> |
| IT19111766          | Lekamalage U.L.V.M.     | <ul style="list-style-type: none"><li>• Development and Testing of Drug Identifier subcomponent</li><li>• Integration of the relevant component to the final system</li></ul>          |
| IT19043388          | Chathuranga S.J         | <ul style="list-style-type: none"><li>• Development and Testing of Virtual Chatbot subcomponent</li><li>• Integration of the relevant component to the final system</li></ul>          |

*Table 6.1: Resource personnel for Development*

Resource personnel assisting externally are as follows:

| Name                        | Designation   | Workplace  |
|-----------------------------|---|--|
| Dr. Muditha Vidanapathirana | Doctor of Medicine, MA, MBBS, DLM, FFFLM (UK), Professor of Forensic Medicine University of Sri Jayewardenepura Sri Lanka | Department of Forensic Medicine, University of Sri Jayewardenepura Sri Lanka |

*Table 6.2: Resource personnel for External Supervising*



## 7. Budget and budget justification (if any)

Table 7.1 shows the Budget allocation for the research and associated tasks and that includes Project Planning, Document Preparation, Internet charges, Hosting charges and other expenses.

| Resource type                        | Amount (LKR) | Amount (USD) |
|--------------------------------------|--------------|--------------|
| Document Preparations<br>(Hard Copy) | Rs. 500      | \$2.48       |
| Internet usage for research          | Rs. 2000     | \$9.91       |
| Hosting Charges (Server)             | Rs. 3800     | \$18.83      |
| Other Expenses<br>(Travelling)       | Rs. 1500     | \$7.43       |
| Total                                | Rs. 7800     | \$38.65      |

*Table 7.1: Budget and Budget Justification*

## **8. Conclusions and Recommendations**

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 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 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 handle the challenges that arise during a worldwide pandemic.

## Reference List

- [1] Jabarulla, M.Y. and Lee, H.N., 2021, August. A blockchain and artificial intelligence-based, patient-centric healthcare system for combating the COVID-19 pandemic: opportunities and applications. In *Healthcare* (Vol. 9, No. 8, p. 1019). Multidisciplinary Digital Publishing Institute.
- [2] Karthikeyan, S., de Herrera, A.G.S., Doctor, F. and Mirza, A., 2021. An OCR Post-correction Approach using Deep Learning for Processing Medical Reports. *IEEE Transactions on Circuits and Systems for Video Technology*.
- [3] Zhao, Y., Cui, M., Zheng, L., Zhang, R., Meng, L., Gao, D. and Zhang, Y., 2019. Research on electronic medical record access control based on blockchain. *International Journal of Distributed Sensor Networks*, 15(11), p.1550147719889330.
- [4] Xue, W., Li, Q. and Xue, Q., 2019. Text detection and recognition for images of medical laboratory reports with a deep learning approach. *IEEE Access*, 8, pp.407-416.
- [5] Sahu, N. and Sonkusare, M., 2017. A study on optical character recognition techniques. *The International Journal of Computational Science, Information Technology and Control Engineering*, 4, pp.1-14.
- [6] Karthikeyan, S., de Herrera, A.G.S., Doctor, F. and Mirza, A., 2021. An OCR Post-correction Approach using Deep Learning for Processing Medical Reports. *IEEE Transactions on Circuits and Systems for Video Technology*.
- [7] Liao, M., Shi, B., Bai, X., Wang, X. and Liu, W., 2017, February. Textboxes: A fast text detector with a single deep neural network. In *Thirty-first AAAI conference on artificial intelligence*.

- [8] D'hondt, E., Grouin, C. and Grau, B., 2017, November. Generating a training corpus for OCR post-correction using encoder-decoder model. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1006-1014).
- [9] Moon, S., Liu, S., Chen, D., Wang, Y., Wood, D.L., Chaudhry, R., Liu, H. and Kingsbury, P., 2019. Saliency of medical concepts of inside clinical texts and outside medical records for referred cardiovascular patients. *Journal of Healthcare Informatics Research*, 3(2), pp.200-219.
- [10] Özgen, A.C., Fasounaki, M. and Ekenel, H.K., 2018, May. Text detection in natural and computer-generated images. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
- [11] Widiastuti, N.I., 2019, November. Convolution neural network for text mining and natural language processing. In *IOP Conference Series: Materials Science and Engineering* (Vol. 662, No. 5, p. 052010). IOP Publishing.
- [12] Sheikhalishahi, S., Miotto, R., Dudley, J.T., Lavelli, A., Rinaldi, F. and Osmani, V., 2019. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2), p.e12239.
- [13] Mishra, A., Shekhar, S., Singh, A.K. and Chakraborty, A., 2019, September. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 947-952). IEEE.
- [14] van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B. and Colavizza, G., 2020. Assessing the impact of OCR quality on downstream NLP tasks.
- [15] Admin, "Software development life cycle (SDLC) - winklix - software development blog," *Winklix*, 18-Sep-2019. [Online]. Available:

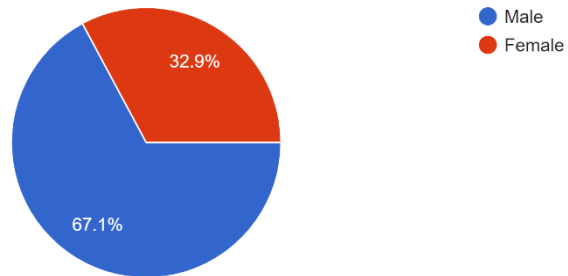
<https://www.winklix.com/blog/software-development-life-cycle-sdlc/>. [Accessed: 05-Feb-2022].

## Appendices

### Appendix A – Additional Survey Responses gathered during the Research Survey

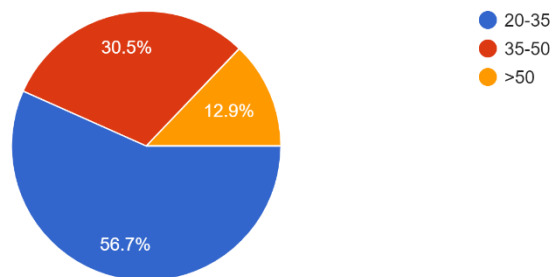
Gender

210 responses



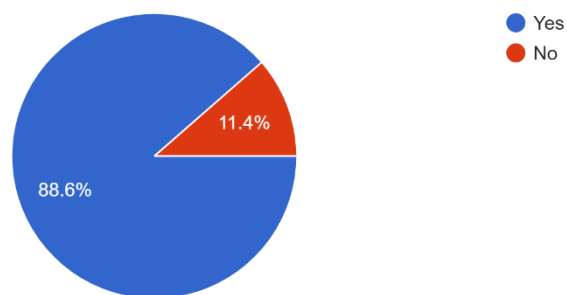
Age Group

210 responses



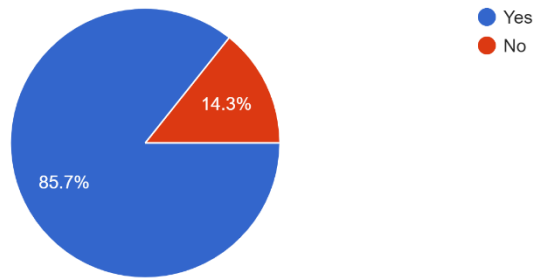
Do you maintain a personal medical log book?

210 responses



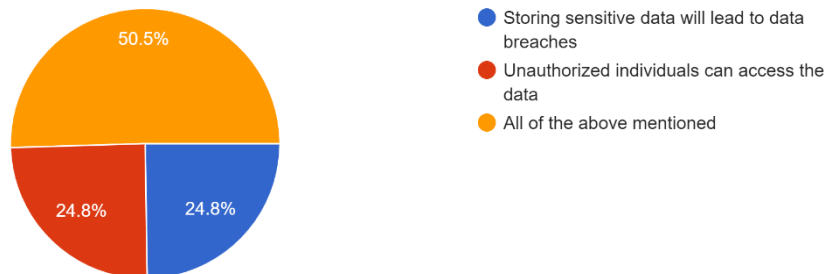
If Yes, Have you ever misplaced your medication history logbook?

210 responses



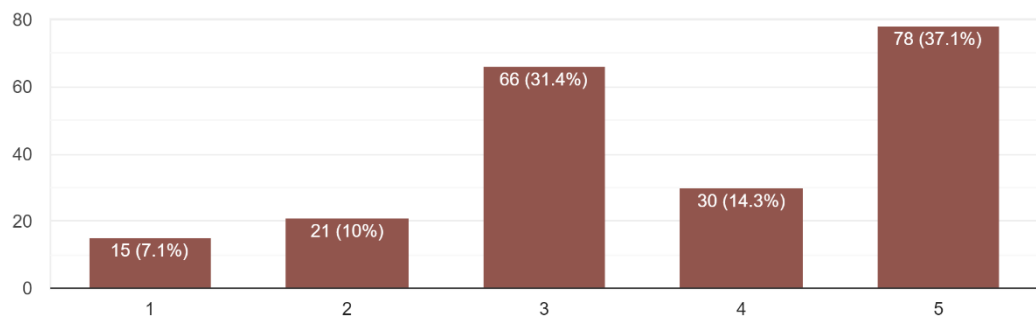
Which problem do you think will occur if we propose an automated solution to store patient data in digital format?

210 responses



"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?

210 responses



## Appendix B – Supervisor and Co-supervisor’s Endorsement

