Proposal

for

Final Year Project

Computer & Information Systems Engineering Department

Document Digitization and Medical Record Retrieval in Healthcare

Kinza Hameed CS-036

Maher Fatima CS-061

Samrah Mazhar CS-054

NED University of Engineering & Technology

1. Project Identification

A. Reference Number (for office use only)

C	S	-	1	7		

B. Project Title

Document Digitization and Medical Record Retrieval in Healthcare

C. Project Internal Advisor

Name	**Dr. Majida Kazmi**
Designation	**Assistant Professor**

D. Project Internal Co-Advisor

Name	**Ms. Lubaba Rehman**
Designation	**Research Assistant**
Mobile #	**03462360605**

E. Project External Advisor

Name	-	
Designation	-	
Organization	-	
Mobile #	-	Email -

F. Student Team

S.No.	Roll No.	Name	CGPA	Email
1.	CS-036	Kinza Hameed	3.8	hameed4000750@cloud.neduet.edu.pk
3.	CS-061	Maher Fatima	3.50	fatima4005187@cloud.neduet.edu.pk
2.	CS-054	Samrah Mazhar	3.44	mazhar4000585@cloud.neduet.edu.pk

G. Sponsoring Organization (if any)

Neurocomputation lab, NCAI, Karachi, Pakistan

H. Keywords

- Optical Character Recognition
- Artificial Intelligence
- Medical data
- Document image processing
- Data retrieval and Automated Up-dation
- Web scrapping

I. Project Idea

☐ New ☐ Modification to a previous project

☐ Extension of a previous project

2. ABSTRACT

Patients are challenged in their daily lives by maintaining their multiple medical records in huge number of piles. The resources available for them are limited and localized. Each record is filled with information about the patient such as their identity, medical history, and laboratory results. Given the importance of health records, it is essential to keep them up to date so that optimal care can be provided when necessary. However the vast amount of records makes updating them a difficult job. Doctors face difficulty to check the medical history for their patients. Hospitals often face challenges in sifting through voluminous healthcare records and extracting relevant information from physical documents.

Fortunately there is a system needed for health records digitization, which can streamline information and keep it current. A system that merges the medical history of the respective patients into one single document and update the medical history of the patient automatically to cloud-based storage.

Document Digitization and Medical Record Retrieval in Healthcare is a compatible with all OPERATING SYTEM application with a goal to automate the collection of medical data through medical reports. The data will be processed and retrieved in a format by OCR after the uploading of the medical report, analysis of extracted medical data through OCR medical history of the patient will be updated on the basis of matching algorithm. The output will be easily transformed into human readable documents by means of standard techniques allowing the users to graphically compare the results against the input image of the document.

Project Background and Literature Review

This system is made to analyze images of paper documents and recognize written characters and letters within them. It then records the recognized characters in the same sequence as they appear on the scanned image which can then be saved in a digital format. This process effectively digitizes the information from the scanned document and allows the user to save it to a database for later use. In the healthcare industry, OCR technology could help digitize document types including but not limited to:

- ✓ Clinical trial documents
- ✓ Patient reports containing clinical data or medical records
- ✓ Prescription slips or receipts that may be used to verify a patient received medication
- ✓ Lab notebooks from clinical trials or other experiments

Patient reports, likely from outside clinics or physicians, may also contain helpful information about likely side effects of a drug and how individual patients may respond. It is particularly important to incorporate all relevant patient information when new side effects are discovered or if a patient has had allergic reactions to it in the past.

Healthcare institutions can benefit from this information by having access to every past reaction to a given drug and using that awareness to improve patient care.

All procedures and results from pharmacology experiments and other healthcare fields of study are recorded in lab notebooks, which usually contain notes from multiple weeks of work. These notes would then be digitized and saved to a database as new results are found.

Resemble English OCR work:

Mohammed Javed et al [1] presents a paper where he gives an idea of learning and identifying the font size using simple line height features directly from the compressed text documents at the line level. In the model, mixed case text documents are taken and segmented into compressed text lines and the ascender height and the lint height features are extracted which are intern used to obtain the regression line of the pattern. The modal gains an overall accuracy of 99.67%.

Chaudhuri et al [2] proposes a system where the image is captured which is ten subjected to text correction, line segmentation, word and character segmentation, zone detection, text graphics separation using some modified and conventional. The characters are separated using zonal information and shape characteristics. Then the compound characters are recognized using template matching and tree classifier. For identifying root word and suffixes a dictionary-based error-correction process is used. And an accuracy of 95.50% is obtained.

Resemblance system work:

In 2015 Electrical and Electronic Measurements Laboratory made “An automatic document processing system for medical data extraction” in which they implemented an automatic document processing system for the extraction of data contained in medical laboratory results printed on paper. [3]

The first advantage of their proposed method consists in the reduction of time expenditure when creating digital records from printed documents. In general, extra time burdens for using clinical information systems and lack of flexibility have slowed the adoption of electronic medical records. It should be noted that, in the intended use of their system, the documents can be processed automatically in batches whereas human intervention, for checking and eventual correction of the results, can be postponed to a second time. The second important advantage is that the reported error rate of 5%, which aggregates all kinds of errors encountered in their experimentation, is acceptable if compared with procedures that include manual intervention.

Their conclusion was that the described system would benefit from improvements in the reduction of character recognition error rate. In fact, more successful character recognition will increase the number of recognized test names and reduce the errors in test results. Improving the OCR task implies also the use of better quality printed sources and a refinement of the pre-processing stages in accordance with the particular OCR engine employed.

Motivation and Need

Medical History and medical forms are extremely important documents. They are commonly found in the offices of hospitals, clinics, and healthcare providers. Each record is filled with information about the patient such as their identity, medical history, and Results of different Medical Test. Given the importance of health records, it is essential to keep them up to date so that optimal care can be provided when necessary. Keeping updated and

accurate health records is a matter of utmost necessity. If hospitals and providers do not have current information on their patients then they cannot provide the proper care. Having incorrect medical information can waste time and even risk lives. Instead of having employees handle large amounts of paper records through manual data entry you can cut costs and save time and resources with OCR for health records.

So we decided to make a system that will be help patients to maintain a file of medical records and it will also automate the job of doctor of going through multiple pages of patient medical files into a single file.

Our main challenge is to make an OCR module that will work by converting scanned documents, which are essentially image files, into machine encoded text files. Scans and PDFs are image files and they cannot be edited or searched through since they are not formatted in text. OCR formats the files into text, rendering them text editable and searchable. Applied to health records, this would mean that you can scan all the medical documents and OCR them for greater accessibility. All the medical information extracted and analyzed will be inserted automatically in the medical history of patient on the basis of some matching algorithm. By means of standard techniques users can graphically compare the results against the input image of the document and can simply locate their medical history using a search function.

3. Objectives

We are going to implement a mobile application in which user will upload a scanned document of his/her history in the mobile application and whenever needed it is exported to the respective Doctor, irrespective of the hospital.

Now, the reports can be in hardcopy or the patients will be accessing his or her report through a URL provided by the hospital. The reports' PDF will be uploaded if it's present in hardcopy or if URL is provided the image of URL will be uploaded and medical data will then be available through web scraping.

Now here our first objective is to design an OCR engine that has the capability to gather all the data written in the report. This would be a challenging task as the reports are printed through different printers like inkjet, laser and dot matrix printers, and to recognize the characters in that format will require some algorithm.

Our second challenge would be to design a module for placement of medical data in the right domain and this will require some matching algorithm. Here the challenging task is that reports can be different; there is no universal format for medical report. All the hospitals produce medical report in their own format. So, our matching algorithm must be that much efficient that it places the extracted and analyzed medical data in right places, like it should have the ability to recognize test name, result and unit.

4. Methodology and Equipment/Tools

- **METHODOLOGY**

The main stages that allow the processing of a printed page containing the laboratory results are: Image preprocessing, in which the document readability is enhanced; layout analysis, in which the document layout is analyzed in order to identify columns and rows containing the information to be extracted, data extraction and classification, in which text returned by the OCR is analyzed syntactically and semantically exportation in a format of the extracted data, and finally updating of the medical history.

I) Preprocessing

In this phase the image is prepared for subsequent processing steps. In particular, equalization, binarization and suppression of long lines are required to ease layout analysis and OCR.

II) Layout analysis

In the layout analysis phase, the image is subdivided into blocks; text rows are processed by the OCR, and the text is compared with a list of column headers. The column headers identify those table columns that contain pertinent medical data.

III) Data extraction

In this phase data are extracted from table cells. The cells are processed by the OCR, which is configured to recognize a different set of characters according to the cell type.

IV) Exportation in a format and Automated Updation in medical record

In this phase, the data extracted from the document are saved in an output file. This output will conceive in order to share medical laboratory results in a simple manner. Other than integration into databases, the output can be easily transformed into human readable documents by means of standard techniques. It should be noted, however, that the information on the coordinates of the extracted information is preserved in the results, so allowing the users to graphically compare them against the input image of the document.

- **Web Scrapping Methodology:**
  When the patient will upload a receipt of his test, including URL, ID and password to access the test reports, the OCR will retrieve this information and use web scrapping to obtain the test results the methodology of web scrapping will be as follows:

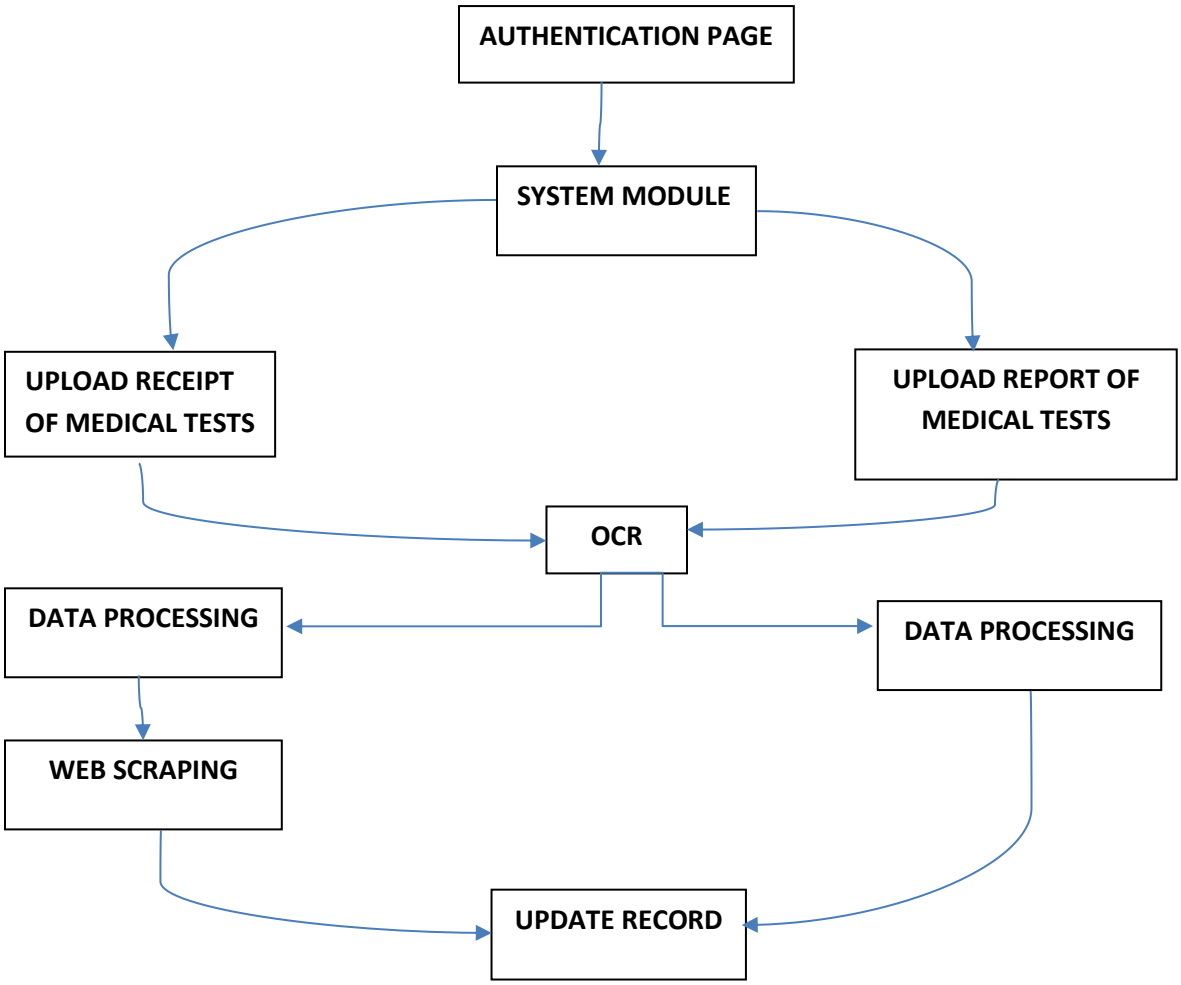
  (i) Find the URL that you want to scrape
  First, you should understand the requirement of data according to your project. A webpage or website contains a large amount of information. That's why scrap only relevant information. In simple words, the developer should be familiar with the data requirement.

  (ii) Inspecting the Page
  The data is extracted in raw HTML format, which must be carefully parsed and reduce the noise from the raw data. In some cases, data can be simple as name and address or as complex as high dimensional weather and stock market data.

  (iii) Write the code
  Write a code to extract the information, provide relevant information, and run the code.

  (iv) Store the data in the file
  Store that information in database. This stored date will update the record of patient's medical history.

DATAFLOW DIAGRAM



AUTHENTICATION:

It will ask for the user's authenticity to get services from the system module by giving its own authentication keys while sign up.

SYSTEM MODULE:

It will be the main home page of the application from which user can upload or view his/her medical record according to his needs. Also he/she can export its data to make sure that his new doctor knows about his history.

UPLOAD RECEIPT OF MEDICAL TESTS:

Whenever we take tests, hospital gave us the receipt on which the id, password and URL is printed on which the report has to be uploaded so it will take picture or pdf of receipt.

UPLOAD REPORTS OF MEDICAL TESTS:

Patients also have their record as PDF file or as a hard copy so here patient can upload his/her files.

OCR:

Both the above uploading module will go to OCR and OCR will extract the relevant information and use it according to the system needs.

DATA PROCESSING:

Data will now be extracted and being analyzed and categorized according to the specific requirements.

WEB SCRAPING:

When the system receives the receipt information then by using web scraping, it will extract and store the content in the system.

UPDATE RECORD:

The records will be automatically updated after every new record is added to the Application.

4B. TOOLS

- Python
- React native
- Artificial algorithm
- PHP API
- MYSQL
- Server

5. Key Milestones and Deliverables

No.	Elapsed time (in months) from start of the project	Milestone	Deliverables
1.	1 month	Requirement Gathering	----
2.	2 month	Research of Algorithm to be used	Document of SRS use for the concept will be delivered
3.	2.5 month	Data Collection	Sample Database will be maintain
4.	3 month	Data processing and Analysis	----

5.	4.5 month	Design and Implementation	Mid-year Report and APK
6.	5 Month	Integration of Modules	----
7.	6 month	System Testing, Verification & Validation	First Prototype
8.	7 month	Modification if required	Second Prototype
9.	8.5 month	Integration & Acceptance Testing	Final Prototype
10.	9 month	Merging of all remaining work and documents	Final Report, Research paper and Project

Expected Outcome

Observed outcome of project:

- To create a single, continuous record for a patient provides a holistic view of overall health for better diagnosis and lifetime treatment.
- To automatically retrieve report results from the hospital website and update the patient's health record.
- To provide accurate, up-to-date, and complete medical information about patients at the point of care.
- To improve in the quality of care, a reduction in medical errors, and other improvements in patient-level measures that describe the appropriateness of care.

6. Direct Customers / Beneficiaries of the Project

- Paramedical staff
- Patients
- Doctors

7. Consent of Advisors

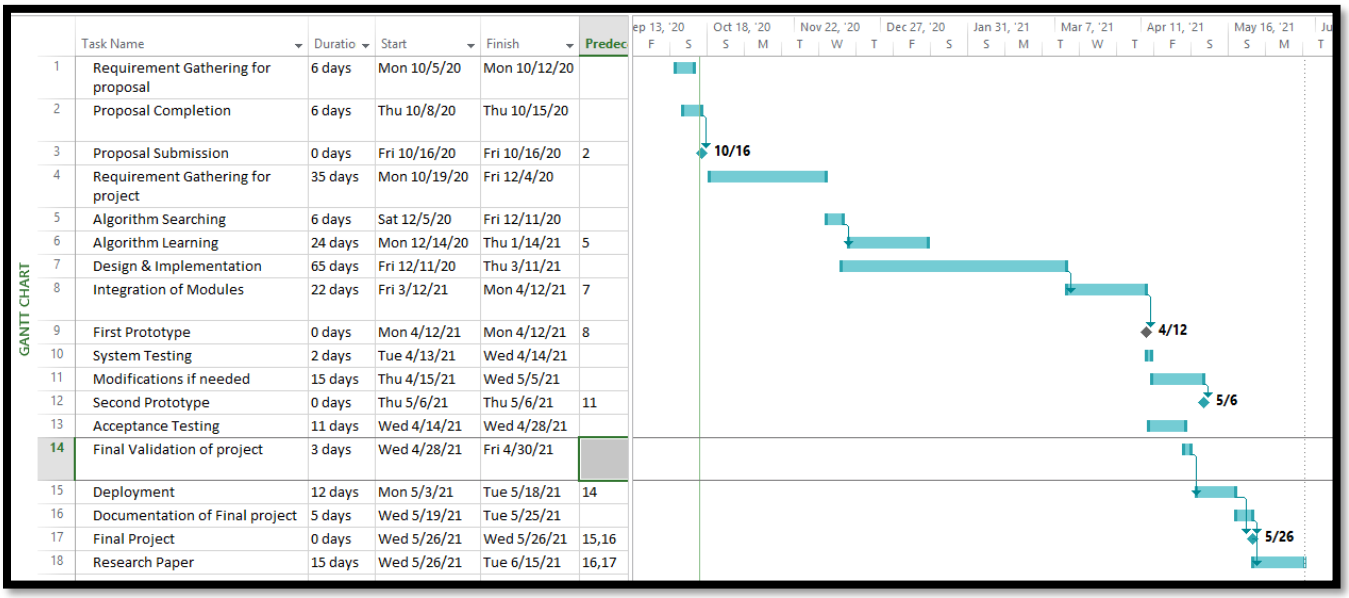
Consent of the Internal Advisor Signature: _____

Consent of the Co-Internal Advisor Signature: _____

Consent of the External Advisor (if any) Signature: _____

8. Reviewers Committee's Comments

9. Project Schedule / Milestone Chart



GANTT CHART

	Task Name	Duratio	Start	Finish	Predec
1	Requirement Gathering for proposal	6 days	Mon 10/5/20	Mon 10/12/20	
2	Proposal Completion	6 days	Thu 10/8/20	Thu 10/15/20	
3	Proposal Submission	0 days	Fri 10/16/20	Fri 10/16/20	2
4	Requirement Gathering for project	35 days	Mon 10/19/20	Fri 12/4/20	
5	Algorithm Searching	6 days	Sat 12/5/20	Fri 12/11/20	
6	Algorithm Learning	24 days	Mon 12/14/20	Thu 1/14/21	5
7	Design & Implementation	65 days	Fri 12/11/20	Thu 3/11/21	
8	Integration of Modules	22 days	Fri 3/12/21	Mon 4/12/21	7
9	First Prototype	0 days	Mon 4/12/21	Mon 4/12/21	8
10	System Testing	2 days	Tue 4/13/21	Wed 4/14/21	
11	Modifications if needed	15 days	Thu 4/15/21	Wed 5/5/21	
12	Second Prototype	0 days	Thu 5/6/21	Thu 5/6/21	11
13	Acceptance Testing	11 days	Wed 4/14/21	Wed 4/28/21	
14	Final Validation of project	3 days	Wed 4/28/21	Fri 4/30/21	
15	Deployment	12 days	Mon 5/3/21	Tue 5/18/21	14
16	Documentation of Final project	5 days	Wed 5/19/21	Tue 5/25/21	
17	Final Project	0 days	Wed 5/26/21	Wed 5/26/21	15,16
18	Research Paper	15 days	Wed 5/26/21	Tue 6/15/21	16,17

10. Project Approval Certificate

Recommendation of FYP Coordinator Signature: _____

Approval by the Chairman Signature: _____

REFERENCES

- [1]M. Javed, P. Nagabhushan and B.B. Chaudhuri, *Automatic detection of font size straight from run length compressed text documents*, 2014.
- [2]B.B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system", *Pattern recognition*,vol. 31, no. 5, pp. 531-549, 1998.
- [3]Electrical and Electronic Measurements Laboratory, Department of Electrical and Information Engineering (DEI), Politecnico di Bari, Via E. Orabona 4, 70125 Bari, Italy