Decognize: Prescription Digitization Using Knowledge Graphs

Project Team

Muhammad Sharjeel Akhtar	20P-0101
Mahad Ashraf	20P-0563

Session 2020-2024

Supervised by

Mr. Muhammad Sohaib Khan

Department of Computer Science

National University of Computer and Emerging Sciences
Peshawar, Pakistan

June, 2022

Student's Declaration

We declare that this project titled "*Decognize: Prescription Digitization Using Knowledge Graphs*", submitted as requirement for the award of degree of Bachelors in Computer Science, does not contain any material previously submitted for a degree in any university; and that to the best of our knowledge, it does not contain any materials previously published or written by another person except where due reference is made in the text.

We understand that the management of Department of Computer Science, National University of Computer and Emerging Sciences, has a zero tolerance policy towards plagiarism. Therefore, We, as authors of the above-mentioned thesis, solemnly declare that no portion of our thesis has been plagiarized and any material used in the thesis from other sources is properly referenced.

We further understand that if we are found guilty of any form of plagiarism in the thesis work even after graduation, the University reserves the right to revoke our BS degree.

Muhammad Sharjeel Akhtar Signature: ________________

Mahad Ashraf Signature: ________________

Verified by Plagiarism Cell Officer
Dated:

Certificate of Approval

The Department of Computer Science, National University of Computer and Emerging Sciences, accepts this thesis titled *Decognize: Prescription Digitization Using Knowledge Graphs*, submitted by Muhammad Sharjeel Akhtar (20P-0101), and Mahad Ashraf (20P-0563), in its current form, and it is satisfying the dissertation requirements for the award of Bachelors Degree in Computer Science.

Supervisor

Mr. Muhammad Sohaib Khan Signature: ________________

Name of FYP Coordinator

FYP Coordinator
National University of Computer and Emerging Sciences, Peshawar

Dr. Head of Computer Science Department

HoD of Department of Computer Science
National University of Computer and Emerging Sciences

Acknowledgements

Your acknowledgments here

Muhammad Sharjeel Akhtar

Mahad Ashraf

Abstract

0.1 Research Problem

- Problem: Inefficient healthcare data management for prescriptions.
- Challenge: Illegible handwriting, medical jargon and Knowledge Graph
- Consequence: Errors in healthcare due to traditional OCR systems.
- Goal: Develop NLP-based system for accurate prescription transcription.

0.2 Methodology

- OCR: Enhancement: Improve scanned image quality and transcribe medical handwriting accurately using advanced models.
- NLP-Based Enhancement: Use of NLP with enhance our Project in overall readability.
- Bio-NER: Identify drugs, diseases, and procedures in prescriptions.
- Bio-NEL (BERN): Identify drugs, diseases, and procedures in prescriptions.
- BERT Text Enhancement: Improve readability by disambiguating text and providing definitions/synonyms.
- Zero-Shot Relation Extraction: Identify relationships like drug-dosage links.
- Approach: This approach combines OCR and NLP, with BERN adding specialized biomedical knowledge and context from a biomedical knowledge graph.

- Knowledge Graph: Collect medical data, construct a structured graph using standardized vocabularies, and enable efficient data retrieval.
- Node Embedding for Context: Use techniques prescriptions within the knowledge graph.

0.3 Key Results

- Accurate Data Extraction
- Efficient Object Detection
- Comprehensive Bio-Medical Knowledge Graph
- User-friendly Interface
- Documentation and Training Resources
- Enhanced Healthcare Data Management
- Enhanced Decision-Making
- Improved Patient Safety
- Contribution To Healthcare
- Scalability

0.4 Conclusion

The project outlined a comprehensive solution that combines Optical Character Recognition (OCR) enhancement and Natural Language Processing (NLP) to significantly improve the handling of medical data, particularly in the context of handwritten prescriptions and medical documents. The OCR enhancement focuses on improving the quality of scanned images and transcribing medical handwriting with precision, leveraging advanced models to ensure data accuracy.

Within the NLP-based enhancement, the implementation of Bio-NER and Bio-NEL (BERN) techniques empowers the system to identify drugs, diseases, and medical procedures within prescriptions. Moreover, the use of BERT Text Enhancement further enhances the readability of the extracted information by disambiguating text and providing definitions and synonyms, ensuring that the data is not only accurate but also comprehensible.

One of the project's standout features is the Zero-Shot Relation Extraction, which can identify complex relationships such as drug-dosage. Additionally, the creation of a Knowledge Graph, constructed from standardized vocabularies, adds a layer of structured data organization and facilitates efficient data retrieval. Node Embedding for Context further aids in understanding entity significance and relationships within the knowledge graph, which enhances the overall depth and context of the extracted medical data.

By combining OCR and NLP, incorporating biomedical knowledge, and using innovative techniques for data processing and retrieval, the project significantly contributes to the advancement of medical information management and readability, benefiting healthcare professionals and patients alike.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

State (1) the purpose of the investigation, (2) the problem being investigated, (3) the background (context and importance) of the problem (citing previous work by others), (4) your thesis and general approach, and (5) the criteria for your study's success **?**.

- Develop NLP-based system for accurate prescription transcription.
- Inefficient healthcare data management for prescriptions.
- Errors in healthcare due to traditional OCR systems.
- This approach combines OCR and NLP, with BERN adding specialized biomedical knowledge and context from biomedical knowledge graph
- Improve scanned image quality and transcribe medical handwriting accurately using advanced models.

1.1 Purpose

In response to the evolving landscape of healthcare and the growing demand for efficient, data-driven solutions, our project represents a significant step forward in addressing a wide array of critical needs. From enhanced medical record digitization and improved healthcare decision-making to creating new research opportunities and enabling significant time and efficiency savings, our innovative approach is poised to provide a multi-faceted solution. Moreover, our project is not merely a technological endeavor; it is a commitment to enhancing accessibility, fostering competitive recognition, and making a tangible contribution to local healthcare. This introduction sets the stage for a project that not only leverages cutting-edge technology but also places the wellbeing of patients and the efficiency of healthcare practitioners at the forefront, ultimately leading to an improved and more responsive healthcare ecosystem.

1.1.1 Motivation

The motivation behind this project is rooted in a profound recognition of the critical healthcare challenges that persist in our current healthcare system. By focusing on aspects such as improving data management, preventing medication errors, and ensuring swift access to precise patient information, our project aims to address fundamental issues that impact the quality of care. Furthermore, by facilitating informed decision-making and reducing the burdens of manual data entry, we strive to enhance the efficiency and effectiveness of healthcare delivery. The project also seeks to bridge the gap between healthcare and technology research, opening up new possibilities for innovation and advancement in the field. Handwritten prescription handling is a particularly challenging aspect of healthcare documentation, and our approach leverages Knowledge Graph technology to tackle this issue head-on. Ultimately, the overarching goal is to significantly enhance patient safety and healthcare quality, driving a positive transformation in the way healthcare data is managed, leading to safer, more efficient, and higher-quality healthcare outcomes.

graphicx

Chapter 2

Review of Literature

2.1 OCR Enhancement Using Tesseract and OpenCV (2023)

This paper explores the enhancement of OCR using Tesseract and OpenCV with a focus on medical handwriting recognition. The goal is to improve data management in healthcare records.

2.1.1 Body:

The methodology integrates image pre-processing and text detection. Results indicate satisfactory accuracy in well-preprocessed images but struggles with complex backgrounds and artifacts.

2.1.2 Conclusion:

While effective in recognizing characters, Tesseract's accuracy is hindered by poor image quality and the need for meticulous preprocessing, which may limit its applicability in complex cases.

2.2 OCR with TensorFlow and Custom ResNet Model (2021)

This research focuses on enhancing OCR using TensorFlow and a custom ResNet model, aiming to improve healthcare data management and patient safety.

2.2.1 Body:

This research focuses on enhancing OCR using TensorFlow and a custom ResNet model, aiming to improve healthcare data management and patient safety. The methodology enhances model robustness with data augmentation and custom architecture. The results demonstrate robustness and suitability for the specified task.

2.2.2 Conclusion:

2.3 Biomedical Knowledge Graph Construction with NLP (2021)

This paper introduces the construction of a biomedical knowledge graph with NLP, emphasizing text extraction from biomedical documents.

2.3.1 Body:

The methodology involves OCR, BERN, and zero relation extraction. Successful establishment of a Neo4j knowledge graph is highlighted.

2.3.2 Conclusion:

While versatile in applications, limitations include potential inaccuracies in the zero-shot relation extractor, requiring expert validation.

2.4 Handwritten Text Recognition System with RNN (2018)

This paper discusses the creation of a handwritten text recognition system using RNN and CTC layers, focusing on biomedical applications.

2.4.1 Body:

The methodology employs RNN and data preprocessing. It successfully recognizes handwritten words but faces challenges with non-dictionary words.

2.4.2 Conclusion:

The system demonstrates potential for biomedical applications but has limitations in diversity and recognition accuracy, especially for non-standard words.

2.5 Multilingual Handwritten Prescription Recognition (2022)

This research focuses on using CNNs, RNNs, and LSTMs to recognize and translate handwritten prescriptions in multiple languages.

2.5.1 Body:

Machine learning techniques like CNNs and LSTMs are employed to achieve efficient recognition. Sensitivity to variations in handwriting styles is highlighted.

2.5.2 Conclusion:

While efficient, variations in handwriting styles and the quality and diversity of training data can impact performance.

2.6 Machine Learning Algorithms for Prescription Text (2022)

This study compares various machine learning algorithms for recognizing text on medical prescriptions.

2.6.1 Body:

The methodology involves image scanning and CNN-based feature extraction, leading to successful recognition of medical prescriptions.

2.6.2 Conclusion:

While successful, the study acknowledges the need for further investigation into alternative machine learning algorithms for comprehensive comparison.

2.7 Online Cursive Handwritten Medical Word Recognition (2020)

This research presents an online cursive handwritten medical word recognition system.

2.7.1 Body:

The methodology utilizes a bidirectional LSTM network and data augmentation techniques, showcasing successful utilization.

2.7.2 Conclusion:

The system exhibits recognition efficiency but is limited to reading only the trained data, lacking adaptability.

2.8 Medical Prescription Recognition with Machine Learning (2021)

This paper introduces a medical prescription recognition system using image processing techniques and machine learning.

2.8.1 Body:

The methodology employs machine learning for prescription recognition but faces limitations due to reliance on a small dataset.

2.8.2 Conclusion:

The system exhibits low accuracy and is restricted to reading only one line at a time.

2.9 Medical Prescription Identification with Neural Networks (2021)

This study develops a medical prescription identification solution using neural networks.

2.9.1 Body:

The methodology involves neural networks and knowledge-based matching. Limitations include limited dataset usage and low accuracy.

2.9.2 Conclusion:

The system exhibits low accuracy levels and is limited in dataset usage, which hinders its applicability.

2.10 Hello

Chapter 3

Project Vision

3.1 Problem Statement

3.1.1 Problem

Inefficient healthcare data management for prescriptions.

3.1.2 Challenge

Illegible handwriting , medical jargon and Knowledge Graph

3.1.3 Consequences

Errors in healthcare due to traditional OCR systems.

3.1.4 Goal

Develop NLP-based system for accurate prescription transcription

3.2 Business Opportunity

3.2.1 Efficient Data Digitization:

Streamline the conversion of handwritten medical notes into digital records, saving time and reducing errors.

3.2.2 Enhanced Decision-Making:

Improve data accuracy for quicker, informed healthcare decisions, boosting patient care.

3.2.3 Research and Efficiency:

Enable advanced research and analytics with biomedical knowledge graphs, driving medical innovation.

3.2.4 Cost Savings:

Automate manual data entry, significantly cutting operational costs for healthcare institutions.

3.2.5 Competitive Edge:

Adopting this technology positions healthcare providers as industry leaders.

3.2.6 Scalable Deployment:

Tailor the solution for diverse healthcare institutions, from clinics to hospitals.

3.2.7 Global Reach:

Multilingual support ensures accessibility in international markets.

3.3 Objectives

- To reduce error percentage in prescriptions readability.
- To create an improved OCR system which could later on deployed on other real-life-domains as well.
- To allow user to save and access their prescription data conveniently.

3.4 Project Scope

- Global OCR in Healthcare: Booming market, 15.4
- Recent Projects: Automated doctor prescription by Nano Net Technologies Inc and Neurodata Group.
- OCR in Healthcare in Pakistan: Active research by Seerat Rani, Abd Ur Rehman, Beenish Yousaf, Hafiz Tayyab Rauf, Emad Abouel Nasr, and Seifedine Kadry.
- Summary: OCR enhancing healthcare in Pakistan through innovation and integration.

3.5 Constraints

3.5.1 Data Privacy and Compliance:

The project must adhere to stringent healthcare data privacy regulations, such as HIPAA and GDPR, which can limit data sharing and access.

3.5.2 Data Quality Variability:

Variability in the quality of handwritten medical notes and scanned images may impact the accuracy of OCR and NLP, especially in cases of poor handwriting or low-resolution scans.

3.5.3 Multilingual and Handwriting Variability:

Healthcare documents may be written in multiple languages and various handwriting styles, posing a challenge for accurate recognition and translation.

3.5.4 Hardware and Connectivity:

Accessibility to compatible hardware and a reliable internet connection may be a constraint, especially for smaller healthcare providers with limited resources.

3.5.5 Initial Investment:

Implementing and customizing the system may require a substantial initial financial investment, which can be a constraint for budget-constrained healthcare institutions.

3.5.6 User Training and Adoption:

Healthcare professionals and staff may require training to effectively use the technology, and resistance to adopting new systems can be a constraint.

3.5.7 Integration with Existing Systems:

Seamless integration with existing Electronic Health Record (EHR) systems, with their unique standards and formats, can pose integration challenges.

3.5.8 Maintenance and Updates:

Ongoing maintenance, updates, and IT support are crucial for system reliability, and the associated costs and resource requirements can be constraints.

3.5.9 Scalability:

Ensuring that the system can scale to accommodate growing data volumes and user loads without performance degradation can be a challenge.

3.5.10 Vendor Lock-In:

Dependence on a specific vendor for the technology may limit flexibility and pose long-term constraints.

3.5.11 Data Sovereignty:

Geopolitical factors and data sovereignty laws may restrict the storage location and cross-border transfer of healthcare data.

3.5.12 Ethical and Legal Considerations:

The application of advanced technologies in healthcare data management may raise ethical and legal concerns, potentially constraining certain aspects of the project's implementation.

3.5.13 Accessibility in Remote Areas:

Healthcare institutions in remote or underserved areas may face challenges in accessing the technology due to infrastructure limitations.

3.6 Stakeholders Description

3.6.1 Healthcare Providers:

Healthcare institutions, including hospitals, clinics, and private practices, are primary stakeholders. They are interested in efficient data management, enhanced patient care, and cost savings.

3.6.2 Patients:

Patients are indirect stakeholders as they benefit from improved data accuracy, which contributes to better healthcare decision-making, reduced medication errors, and enhanced quality of care.

3.6.3 Healthcare Professionals:

Physicians, nurses, and other healthcare staff are critical stakeholders. They use the system to access and manage patient data, impacting their daily workflows and decision-making.

3.6.4 Health IT Vendors:

Companies providing healthcare information technology solutions, EHR systems, and data management software are stakeholders, as the project may complement or compete with their offerings.

3.6.5 Pharmaceutical and Biotech Companies:

Stakeholders in this sector can benefit from the biomedical knowledge graph for drug research and development.

3.6.6 Telemedicine Platforms:

Telemedicine providers are stakeholders interested in enhancing data management capabilities and accessibility in remote healthcare services.

3.6.7 Regulatory Authorities:

Regulatory bodies like the FDA, HIPAA, and GDPR play a crucial role in shaping the project's compliance and data privacy aspects.

3.6.8 Data Security Experts:

Cybersecurity and data privacy experts are stakeholders in ensuring the security and privacy of patient data within the project.

3.6.9 Research Institutions:

Academic and research institutions may benefit from the project's data for medical research and analytics, making them indirect stakeholders.

3.6.10 Investors and Funders:

Individuals or organizations investing in the project or providing funding are stakeholders with an interest in the project's financial success.

3.6.11 Technology Providers:

Suppliers of hardware, software, and infrastructure required for the project are stakeholders.

3.6.12 Data Scientists and Analysts:

Professionals with expertise in data analysis and NLP are essential stakeholders in developing, maintaining, and improving the project's capabilities.

3.6.13 Ethical and Legal Advisors:

Experts in healthcare ethics and legal matters are stakeholders, ensuring that the project complies with ethical and legal standards.

3.6.14 Patients' Advocacy Groups:

Groups advocating for patient rights and privacy are indirect stakeholders, with an interest in how the project impacts patient data.

3.6.15 Consulting and Integration Firms:

Firms providing consulting and integration services for the project's implementation are stakeholders.

graphicx

Chapter 4

Software Requirements Specifications

This chapter will have the functional and non functional requirements of the project.

4.1 List of Features

- **Optical Character Recognition (OCR):**
  - Accurate recognition of handwritten medical notes.
  - Image preprocessing for optimal OCR results.
  - Multilingual support for diverse patient populations.
- **Natural Language Processing (NLP):**
  - Extraction and categorization of medical entities (e.g., drugs, diseases, procedures).
  - Text disambiguation for improved readability.
  - Multilingual NLP capabilities.
- **Biomedical Knowledge Graph:**
  - Construction of a structured graph using standardized vocabularies.
  - Linking medical entities to enrich data.

- Facilitation of data retrieval for research and analytics.

- **Integration with Electronic Health Records (EHR):**
  - Seamless integration with existing EHR systems.
  - Compatibility with various EHR standards and formats.

- **Data Privacy and Security:**
  - Compliance with healthcare data privacy regulations (e.g., HIPAA, GDPR).
  - Encryption and secure storage of patient data.
  - Access control and audit trails.

- **Scalability:**
  - Ability to scale with growing data volumes and user loads.
  - Support for healthcare institutions of all sizes, from clinics to large hospitals.

- **User Training and Support:**
  - Training resources and documentation for healthcare professionals.
  - Ongoing technical support and assistance.

- **Multilingual Support:**
  - Recognition and translation of multiple languages.
  - Accommodation of diverse patient populations.

- **Data Quality Improvement:**
  - Image enhancement to improve OCR accuracy.
  - Data validation and correction mechanisms.

- **Efficient Data Retrieval:**
  - Fast and accurate data retrieval for healthcare professionals.
  - Advanced search capabilities for research and decision-making.

- **Cost and Time Savings:**
  - Automation of manual data entry processes.
  - Reduction of operational costs for healthcare institutions.
- **Comprehensive Reporting:**
  - Generation of detailed reports for data analysis.
  - Customizable reporting options.
- **Regulatory Compliance Tools:**
  - Tools for ensuring compliance with healthcare regulations.
  - Regular updates to maintain compliance with evolving laws.
- **Adaptive Learning:**
  - Machine learning algorithms that adapt and improve over time.
  - Continuous optimization for better recognition accuracy.
- **Customization and Integration:**
  - Tailoring the system to suit the unique needs of each healthcare institution.
  - APIs for seamless integration with other healthcare systems.
- **Cross-Platform Accessibility:**
  - Accessible via web browsers and mobile devices.
  - Cloud-based solutions for remote access.
- **Research and Analytics Tools:**
  - Access to a rich biomedical knowledge graph for research and innovation.
  - Analytics features for healthcare data-driven decision-making.

4.2 Functional Requirements

- The system shall accurately recognize handwritten medical notes using OCR.
- It shall perform image preprocessing to optimize OCR results.
- OCR shall support multiple languages for diverse patient populations.
- The system shall extract and categorize medical entities, including drugs, diseases, and procedures using NLP.
- It shall disambiguate text for improved readability.
- NLP capabilities shall support multiple languages.
- The system shall construct a structured biomedical knowledge graph using standardized vocabularies.
- It shall link medical entities to enrich data.
- The system shall facilitate data retrieval for research and analytics.
- It shall seamlessly integrate with existing Electronic Health Records (EHR) systems.
- The system shall be compatible with various EHR standards and formats.
- It shall comply with healthcare data privacy regulations, such as HIPAA and GDPR.
- The system shall encrypt and securely store patient data.
- Access control and audit trails shall be implemented.
- The system shall be scalable to accommodate growing data volumes and user loads.
- It shall support healthcare institutions of all sizes, from clinics to large hospitals.
- The system shall provide training resources and documentation for healthcare professionals.

- Ongoing technical support and assistance shall be available.
- The system shall recognize and translate multiple languages.
- It shall accommodate diverse patient populations.
- The system shall enhance images to improve OCR accuracy.
- It shall include data validation and correction mechanisms.
- The system shall provide fast and accurate data retrieval for healthcare professionals.
- It shall offer advanced search capabilities for research and decision-making.
- The system shall automate manual data entry processes.
- It shall reduce operational costs for healthcare institutions.
- The system shall generate detailed reports for data analysis.
- It shall offer customizable reporting options.
- The system shall provide tools for ensuring compliance with healthcare regulations.
- It shall include regular updates to maintain compliance with evolving laws.
- The system shall use machine learning algorithms that adapt and improve over time.
- Continuous optimization shall be implemented for better recognition accuracy.
- The system shall allow tailoring to suit the unique needs of each healthcare institution.
- It shall offer APIs for seamless integration with other healthcare systems.
- The system shall be accessible via web browsers and mobile devices.
- It shall offer cloud-based solutions for remote access.
- It shall grant access to a rich biomedical knowledge graph for research and innovation.
- It shall include analytics features for healthcare data-driven decision-making.

4.3 Quality Attributes

- **Accuracy:** The system must ensure high accuracy in recognizing handwritten medical notes and categorizing medical entities.
- **Reliability:** Healthcare professionals must rely on the system for accurate data recognition and retrieval.
- **Security:** Robust data security measures must be in place to protect patient data and comply with privacy regulations.
- **Usability:** Healthcare professionals should find the system user-friendly and easy to navigate.
- **Interoperability:** The system should seamlessly integrate with various Electronic Health Record (EHR) systems and healthcare standards.
- **Maintainability:** The system must allow for easy updates, maintenance, and continuous improvement.
- **Compliance:** Strict adherence to healthcare data privacy regulations, such as HIPAA and GDPR, is essential.
- **Scalability:** The system should scale to handle increasing data volumes and user loads as healthcare institutions grow.
- **Performance:** The system should deliver fast OCR and NLP processing, ensuring efficient data management.
- **Customization:** Healthcare institutions should be able to customize the system to meet their unique needs.

4.4 Non-Functional Requirements

- **Performance:** The system shall process OCR and NLP tasks within a maximum response time of 2 seconds.

- **Availability:** The system should have a minimum uptime of 99.9

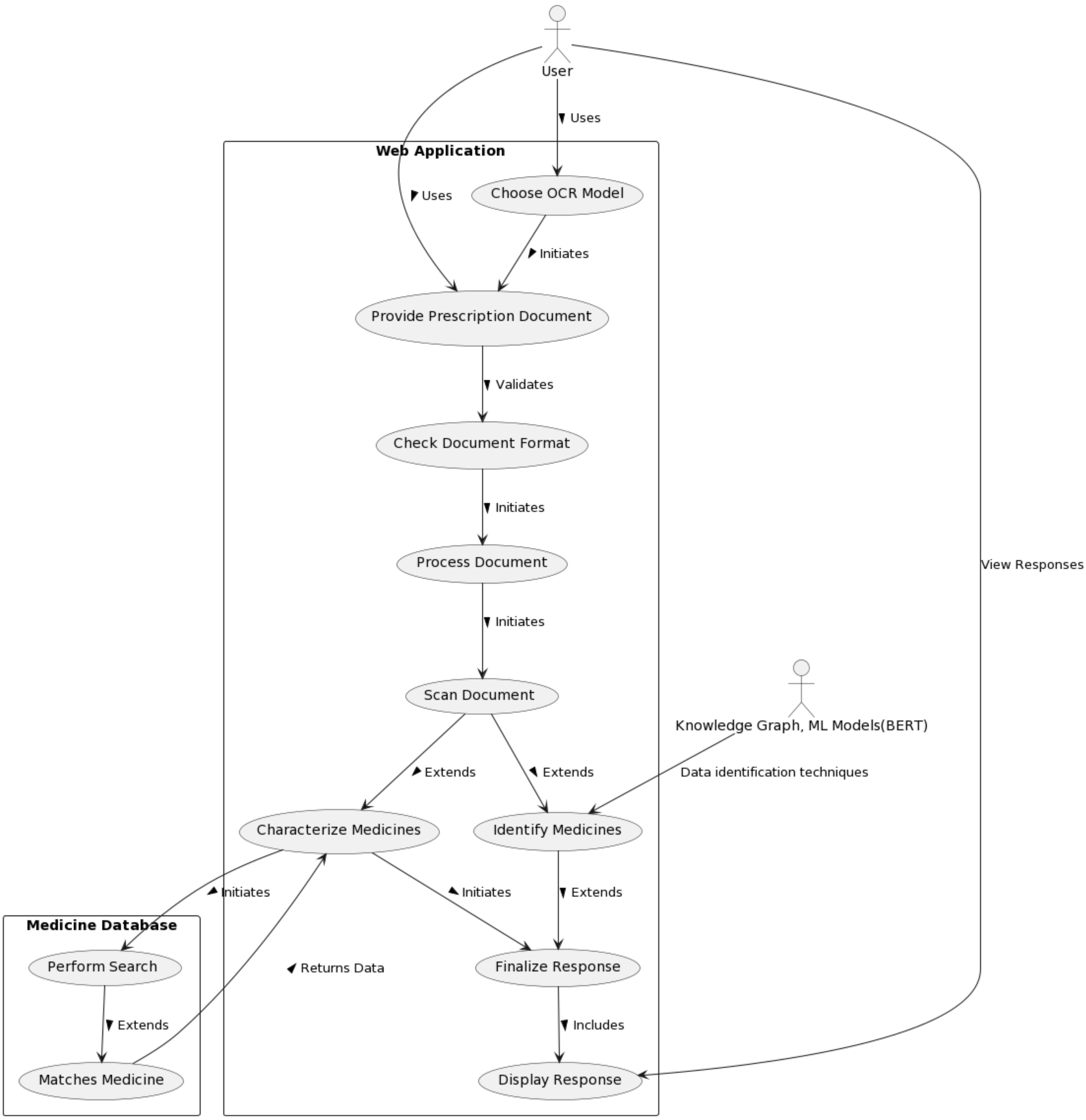
- **Security:** Data transmitted between the system and users shall be encrypted using industry-standard encryption protocols.

- **Scalability:** The system should support a minimum of 100 concurrent users without performance degradation.

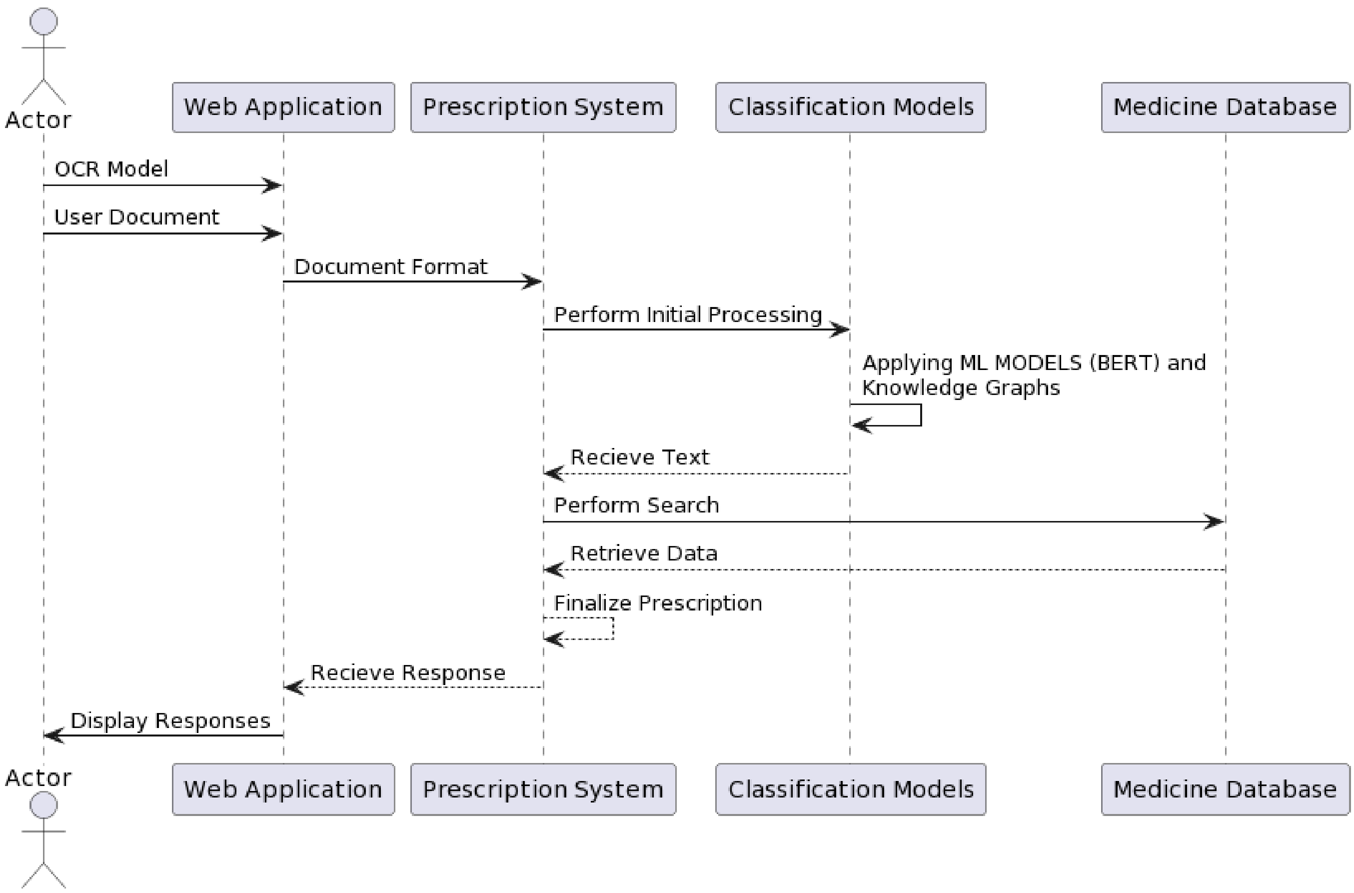
- **Usability:** The system's user interface shall comply with accessibility standards to accommodate users with disabilities.

- **Regulatory Compliance:** The system shall continuously update to remain compliant with evolving healthcare data privacy and security regulations.

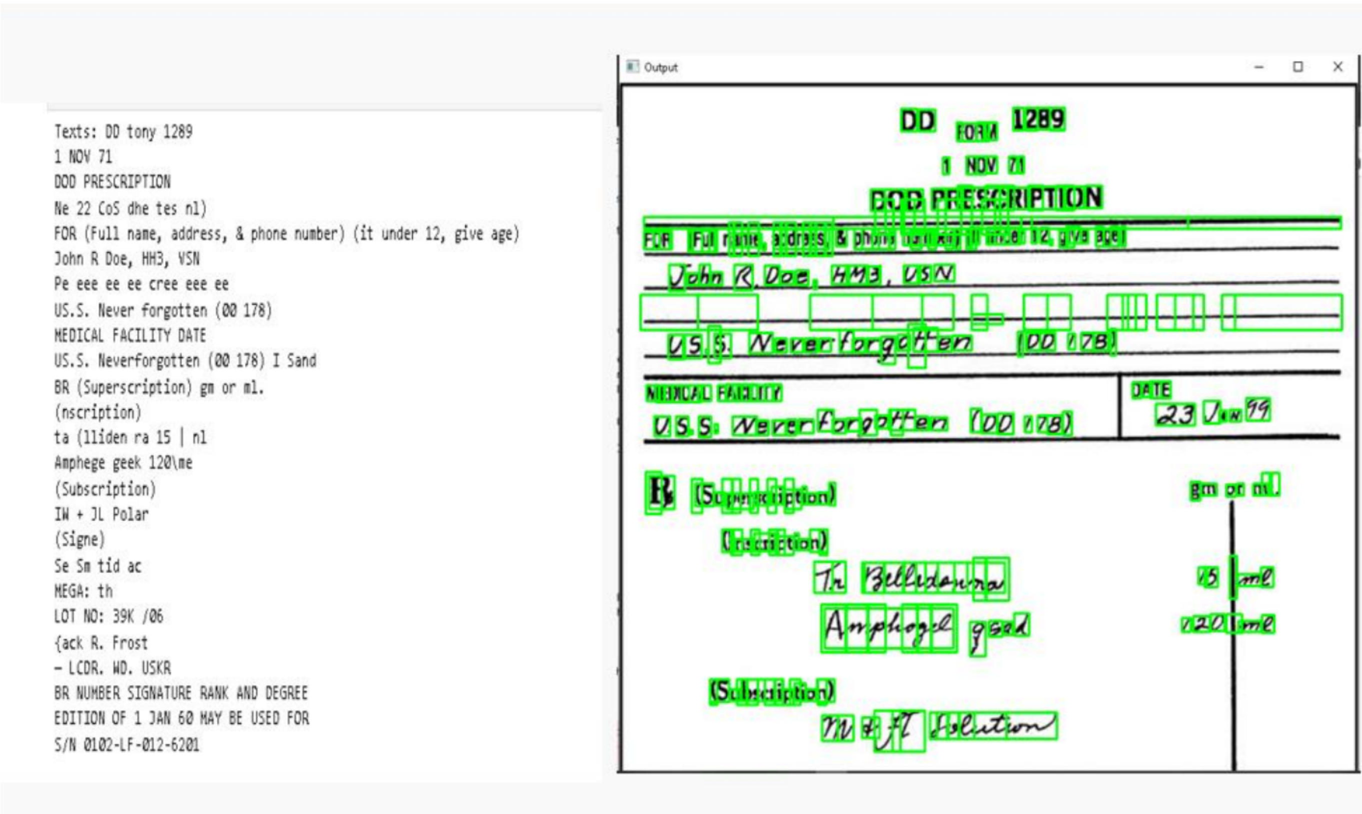
4.5 Use Cases/ Use Case Diagram



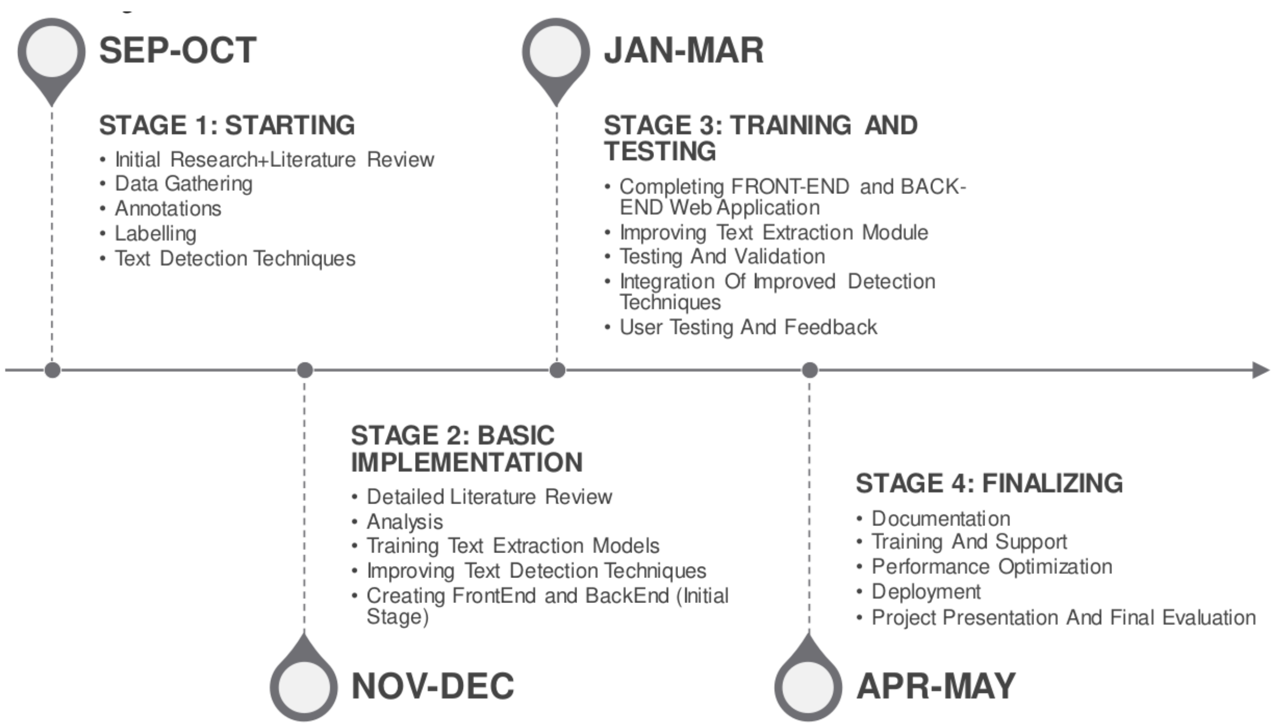
4.6 Sequence Diagrams/System Sequence Diagram



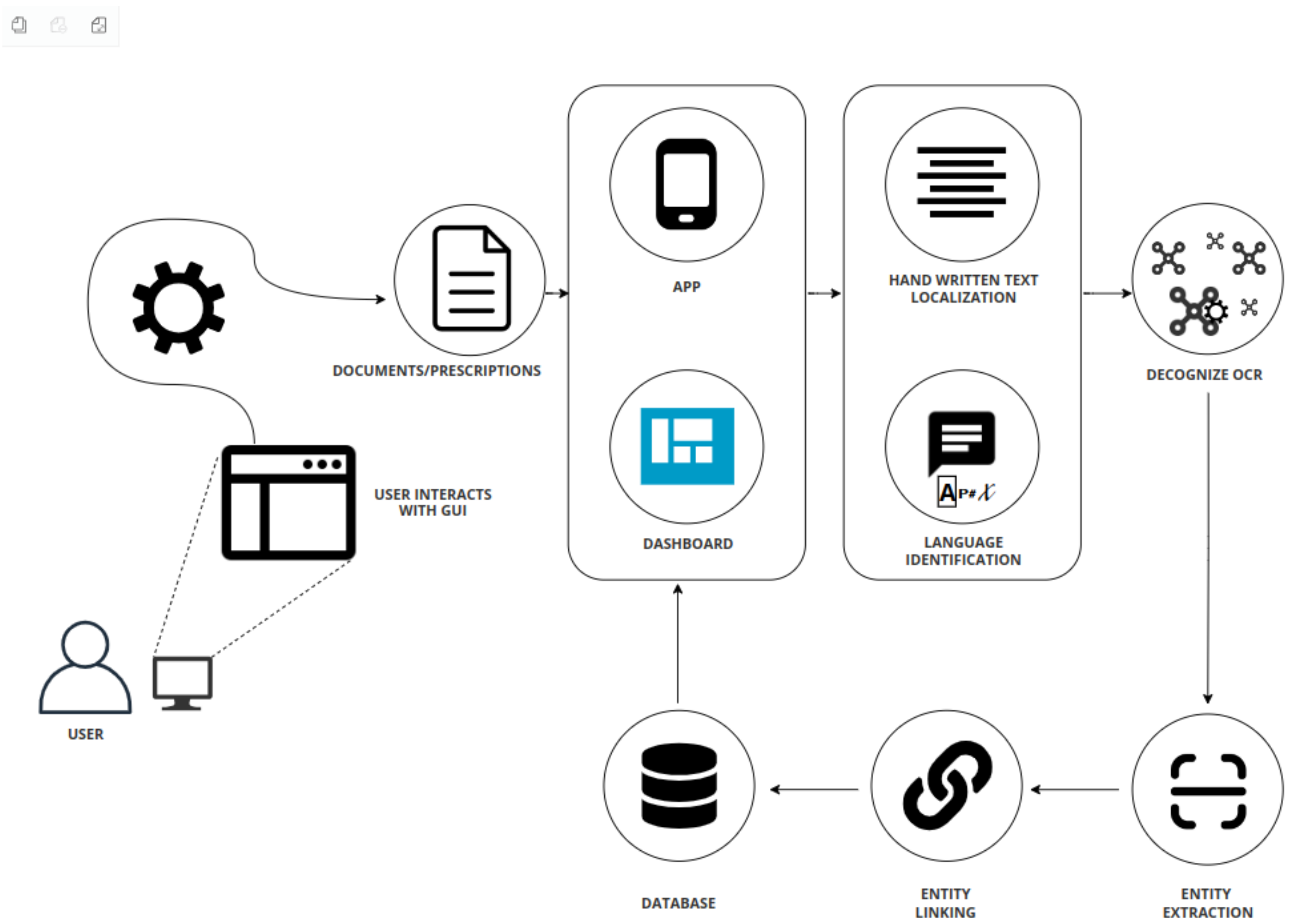
4.7 Test Plan (Test Level, Testing Techniques)



4.8 Software Development Plan

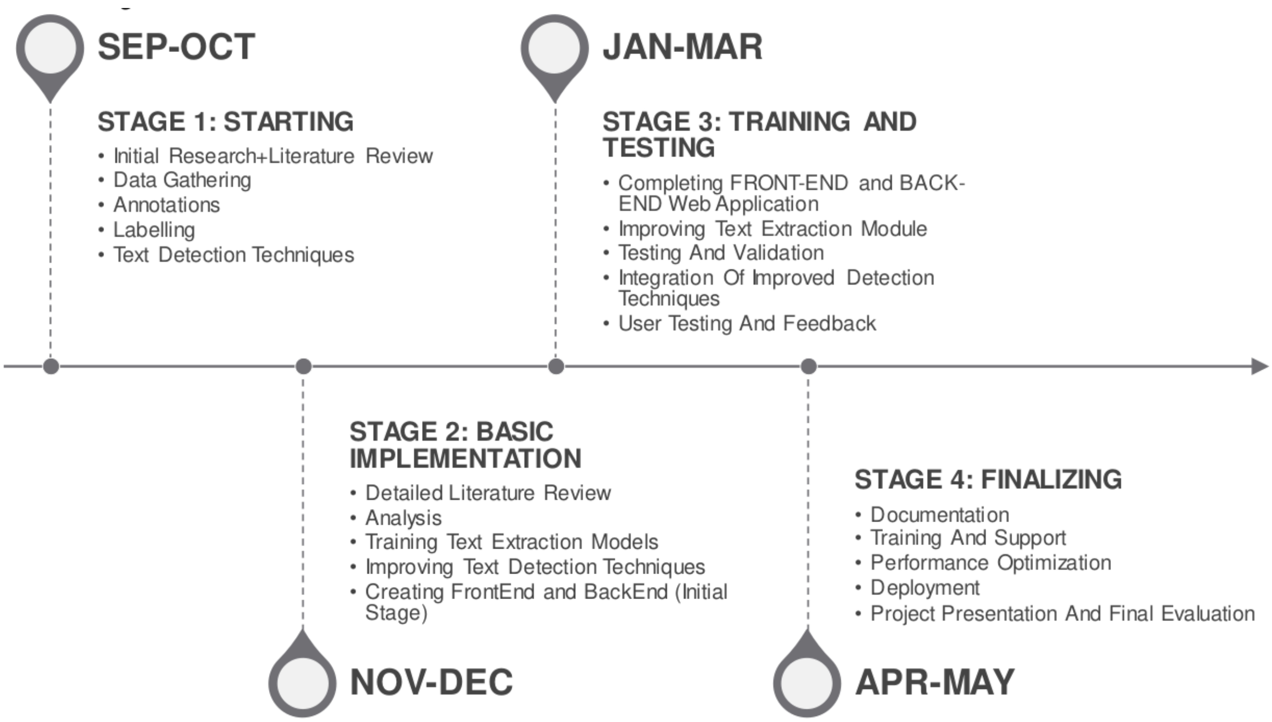


4.9 UI Screens



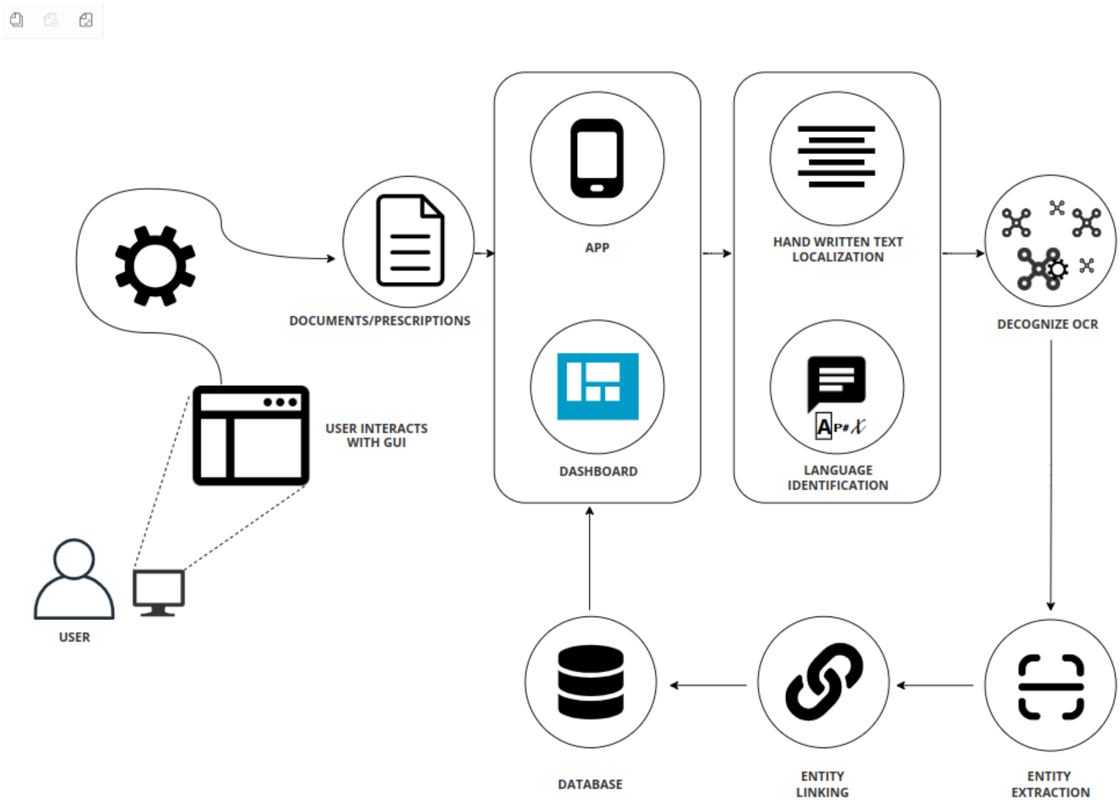
Chapter 5

Iteration Plan



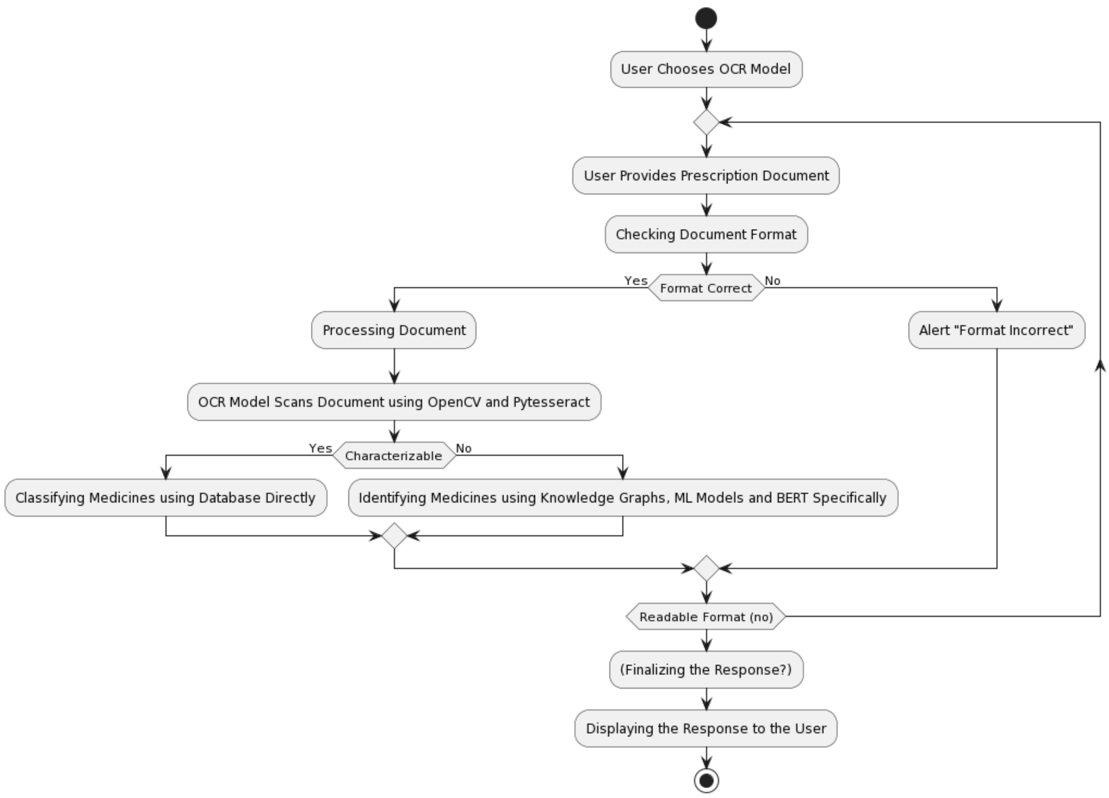
Chapter 6

Iteration 1

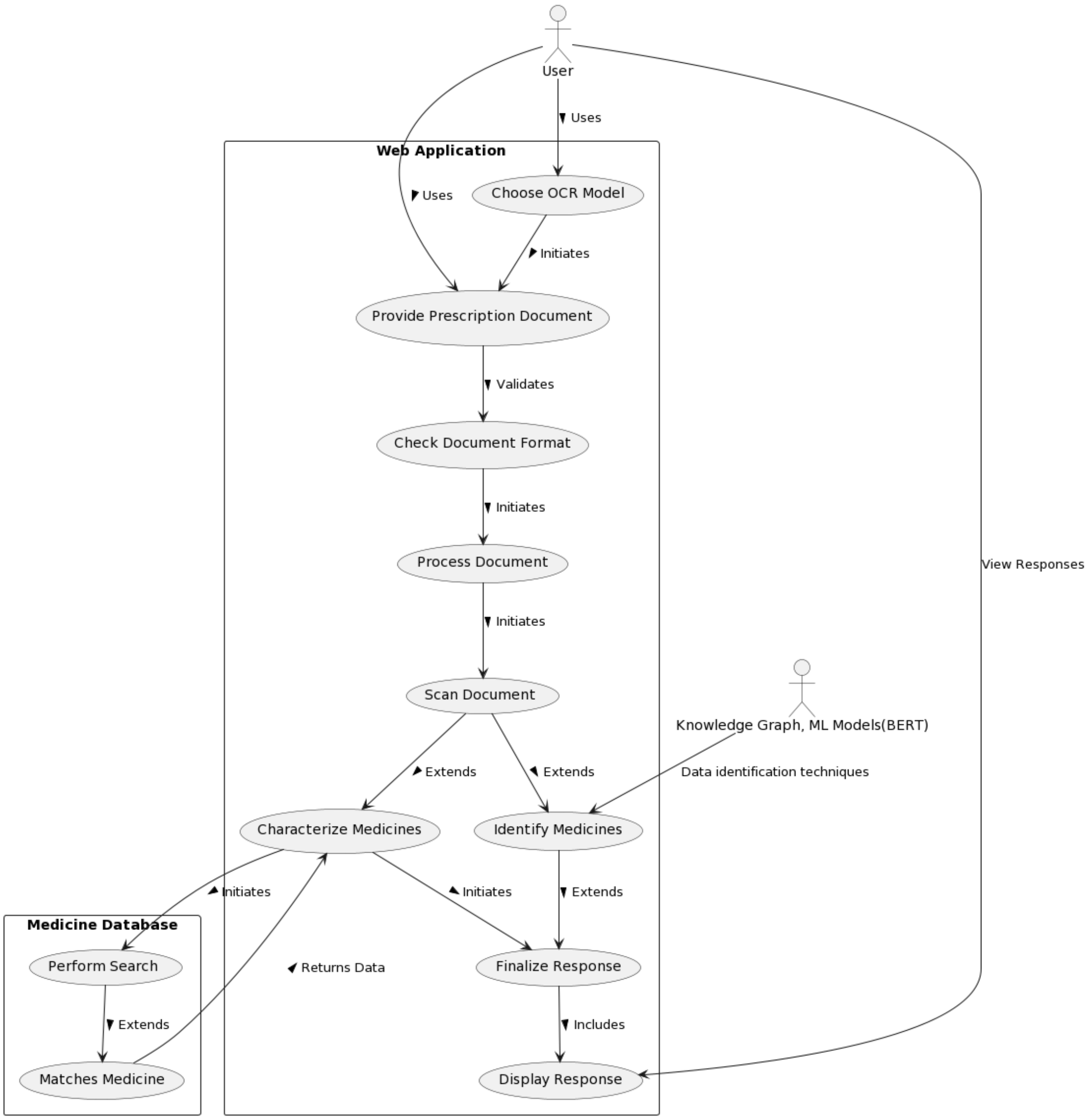


structural design

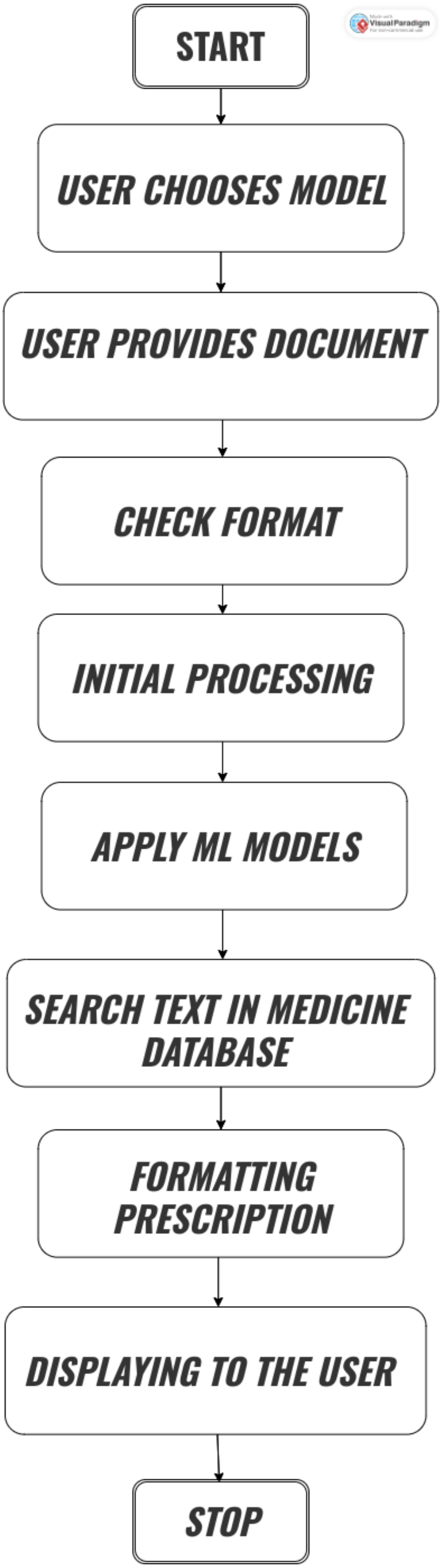
6.1 Domain Model/ Class Diagram



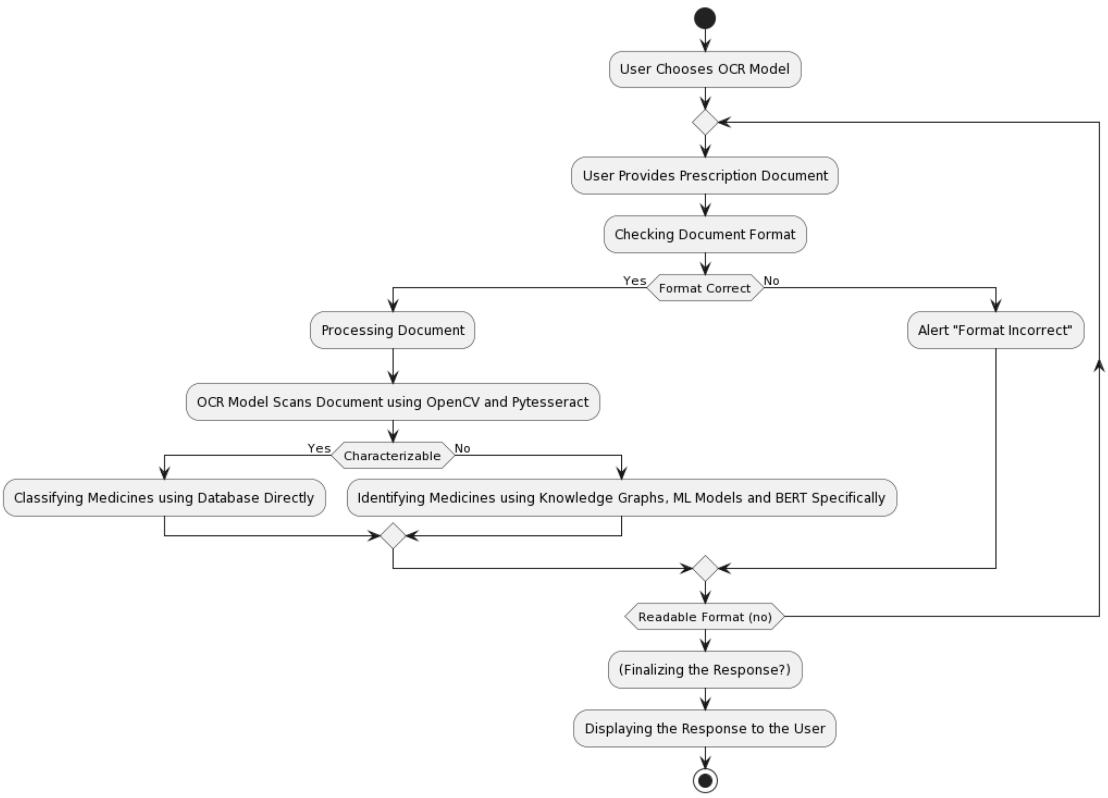
6.2 Component Diagram



6.3 Layer Diagram

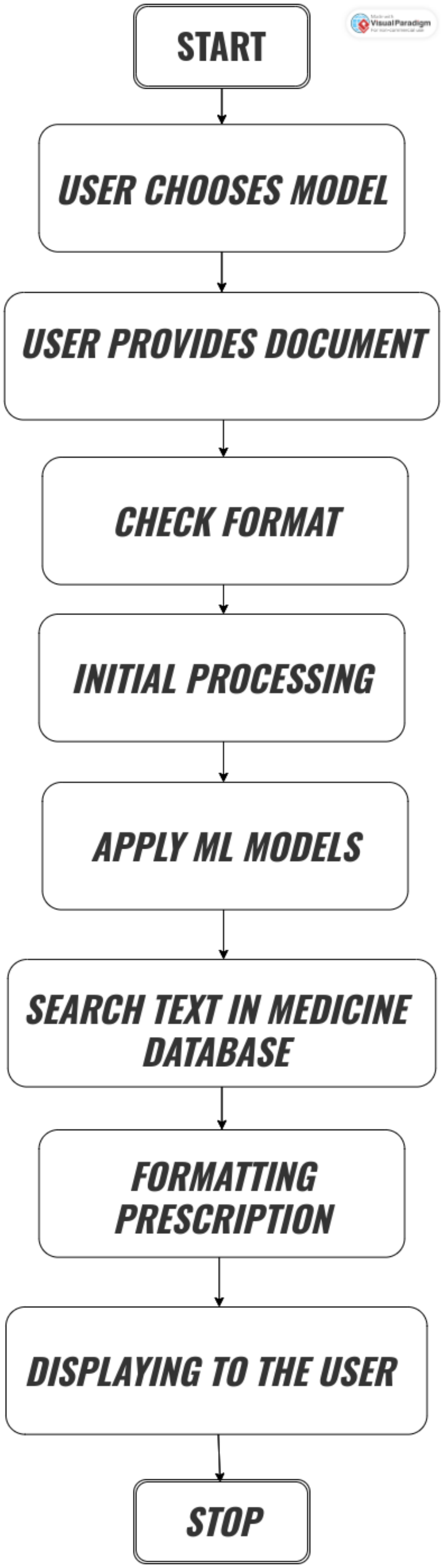


6.4 Structure Chart

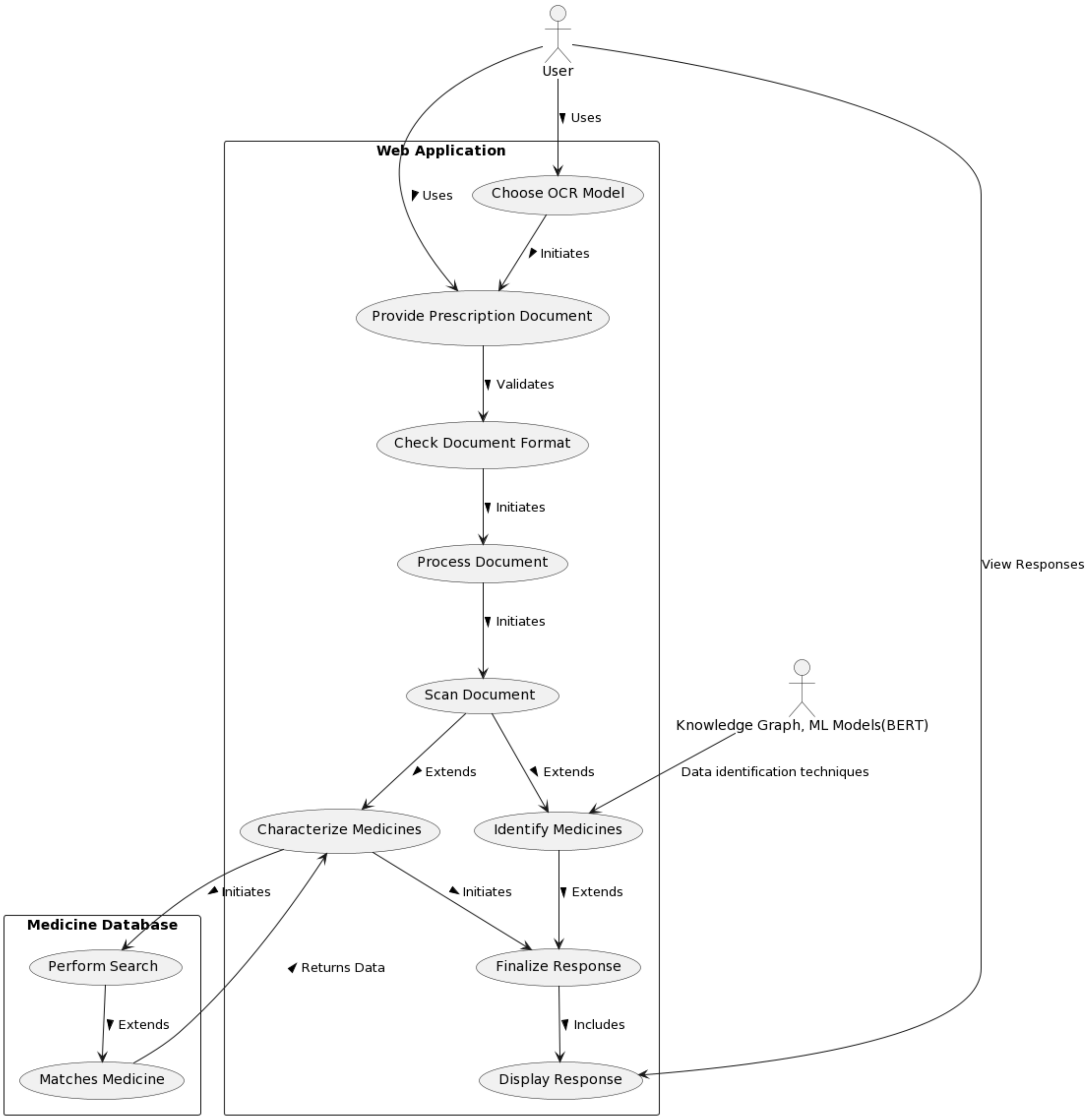


Behavior Design

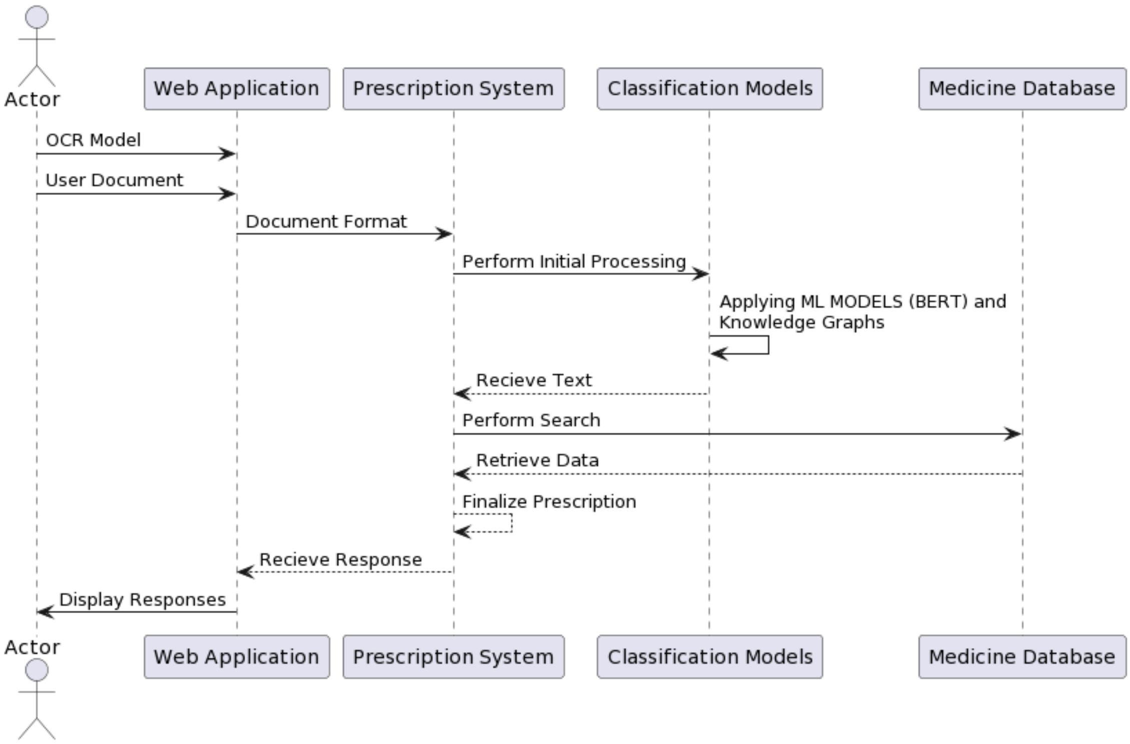
6.5 Flow Diagram



6.6 Data Flow Diagram (DFD)



6.7 Data Dictionary



6.8 Activity Diagram