ANALYTICS OF SCANNED PRESCRIPTION AND NOTES

## A PROJECT REPORT

***Submitted by,***

# Ms. KAVYA UNNIKRISHNAN - 20201CCS0058 Ms. K.P LAKSHMI SHARANYA - 20201CCS0145 Ms. SAI HARSHITHA - 20201CCS0001 Ms. KETHA RISHITHA - 20201CCS0018

***Under the guidance of,***

# Dr. NIHAR RANJAN NAYAK

***in partial fulfillment for the award of the degree of***

# BACHELOR OF TECHNOLOGY

**IN**

**COMPUTER SCIENCE AND ENGINEERING (CYBER SECURITY)**

**At**



**PRESIDENCY UNIVERSITY BENGALURU JANUARY 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**CERTIFICATE**

This is to certify that the Project report **“ANALYTICS OF SCANNED PRESCRIPTIONS AND NOTES”** being submitted by KAVYA UNNIKRISHNAN,

K.P LAKSHMI SHARANYA, SAI HARSHITHA, KETHA RISHITHA ” bearing roll number(s) 20201CCS0058,20201CCS0145,20201CCS0001,20201CCS0018 in partial

fulfilment of requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering (Cyber Security) is a bonafide work carried out under my supervision.

|  |  |
| --- | --- |
| **Dr. NIHAR RANJAN NAYAK**  Assistant Professor School of CSE&IS Presidency University | **Dr. S.P. ANANDRAJ**  Professor & HoD  of CSE Cyber Security Presidency University |

|  |  |  |
| --- | --- | --- |
| **Dr. C. KALAIARASAN**  Associate Dean School of CSE&IS  Presidency University | **Dr. L. SHAKKEERA**  Associate Dean School of CSE&IS  Presidency University | **Dr. M.D SAMEERUDDIN KHAN**  Dean  School of CSE&IS Presidency University |

# SCHOOL OF COMPUTER SCIENCE ENGINEERING

**DECLARATION**

We hereby declare that the work, which is being presented in the project report entitled **ANALYTICS OF SCANNED PRESCRIPTION AND NOTES** in partial fulfilment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering (Cyber Security)**, is a record of our own investigations carried under the guidance of **Dr. NIHAR RANJAN NAYAK, Assistant Professor, School of Computer Science Engineering , Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No.** | **Signature** |
| KAVYA UNNIKRISHNAN | 20201CCS0058 |  |
| K.P LAKSHMI SHARANYA | 20201CCS0145 |  |
| SAI HARSHITHA | 20201CCS0001 |  |
| KETHA RISHITHA | 20201CCS0018 |  |

As we all know that the upcoming business sector and the most demanded sector is the health care. In healthcare the most challenging is the data analysis where hospitals or health care centers cannot keep track of all the history of patients. The project mainly focuses on prescription analysis where we track all the patients previous health care records extracting the prescription PDFs and storing in the excel sheet in an ordered manner for easy access, better understanding and for the future reference by all.

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans **Dr. Kalaiarasan C and Dr. Shakkeera L,** School of Computer Science Engineering & Information Science, Presidency University and **Dr. S.P. Anandaraj**. Head of the Department, School of Computer Science Engineering & Information Science, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide **Dr. Nihar Ranjan Nayak, Assistant Professor**, School of Computer Science Engineering & Information Science, Presidency University for his inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators **Dr. Sanjeev P Kaulgud, Dr. Mrutyunjaya MS** and also the department Project Coordinator **Ms. Manasa C M**.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

**Ms. KAVYA UNNIKRISHNAN Ms. K.P LAKSHMI SHARANYA**

**Ms. SAI HARSHITHA Ms. KETHA RISHITHA**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Table Name** | **Table Caption** | **Page No.** |
| 1 | Table 1.1 | Gant Chart | 13 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Figure Name** | **Caption** | **Page No.** |
| 1 | Fig 1.1 | Installing the pandas | 23 |
| 2 | Fig 1.2 | Installation of python pypdf2 | 23 |
| 3 | Fig 1.3 | Extraction of pdf information into excel from database 1 | 24 |
| 4. | Fig 1.4 | Extraction of pdf information into excel from database 2 | 24 |
| 5. | Fig 1.5 | Extraction of pdf information into excel from database 3 | 25 |

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
|  | **COVER AND TITLE CERTIFICATE DECLARATION ABSTRACT ACKNOWLEDGEMENT LIST OF TABLE**  **TABLE OF CONTENT** |  |
| **1.** | **INTRODUCTION** | **1-2** |
| **2.** | **LITERATURE REVIEW** | **3-4** |
| **3.** | **RESEARCH GAPS OF EXISTING METHODS** | **5** |
|  | 3.1 Standardization and Interoperability | 5 |
|  | 3.2 Challenges in Analyzing PDF Data in Medical |  |
|  | Prescriptions. | 5 |
|  | 3.3 Patient Consent and Ethical Considerations | 5 |
|  | 3.4 Longitudinal Data Analysis | 5 |
|  | 3.5 Integration of Social Determinants of Health | 6 |
|  | 3.6 Human-Centric Design and User Adoption | 6 |
| **4.** | **PROPOSED METHODOLOGY** | **7** |
|  | 4.1 Assessing PDF Structure | 7 |
|  | 4.2 Manual Extraction | 7 |
|  | 4.3 Automated Extraction | 7 |
|  | 4.4 Data Cleaning and Formatting | 7 |
|  | 4.5 Text Mining and Natural Language Processing | 7 |
|  | 4.6 Predictive Analytics and Decision Support | 7 |
|  | 4.7 Medication Adherence and Management | 7 |
|  | 4.8 Scraping PDF Data in Structured form | 8 |
|  | 4.9 Scraping PDF Data in Unstructured Form | 8 |
|  | 4.10 Import PDF Data as a Data Frame | 8 |

|  |  |  |
| --- | --- | --- |
|  | 4.11 Create a Row Identifier using Python Libraries | 8 |
| 4.12 Reshape the Data (Convert from Long-Form to Wide | 8 |
| Form) |  |
| 4.13 Joining Data from Left and Right Sections | 8 |
| 4.14 Pattern Recognition and Trend Analysis | 8 |
| 4.15 Quality Assessment | 8 |
| 4.16 Integration with Electronic Health Records | 9 |
| 4.17 Privacy and Security Measures | 9 |
| 4.18 User Interface and Reporting | 9 |
| 4.19 Iterative Validation and Feedback | 9 |
| **5.** | **OBJECTIVES**   * 1. Automating the process of scraping data from PDF files      1. Data Extraction      2. Text Parsing      3. Data Cleaning      4. Data Integration      5. Panel Data Format      6. Data Storage      7. Automation      8. Documentation:      9. Security and Compliance |  |
|  | **10** |
| 10 |
| 10 |
| 10 |
| 10 |
| 10 |
| 10 |
| 10 |
| 11 |
| 11 |
| 11 |
| 11 |
| **6.** | **SYSTEM DESIGN & IMPLEMENTATION**   * 1. Required Libraries   2. Installed Libraries   3. Imported Libraries | **12** |
| **7.** | **TIMELINE FOR EXECUTION OF PROJECT(GANT**  **CHART)** | **13** |
| **8.** | **OUTCOMES** | **14-15** |
| **9.** | **RESULTS AND DISCUSSIONS** | **16** |

|  |  |  |
| --- | --- | --- |
| **10.** | **CONCLUSION** | **17** |
| **11.** | **REFERENCES** | **18-19** |
| **12.** | **APPENDIX A (PSEUDOCODE)** | **20-22** |
| **13.** | **APPENDIX B (SCREENSHOTS)** | **23-25** |
| **14.** | **APPENDIX C** | **26** |

# CHAPTER-1 INTRODUCTION

In the rapidly evolving landscape of healthcare, the analytics of scanned prescriptions and notes emerges as a transformative solution, bridging the gap between traditional practices and modern technology. This innovative approach involves harnessing the power of data analytics to extract valuable insights from scanned medical documents, revolutionizing patient care and healthcare management.

By employing advanced technologies for data extraction and pattern recognition, healthcare professionals can unlock a wealth of information embedded in prescriptions and notes. This digitization process not only enhances the efficiency of healthcare workflows but also paves the way for data-driven decision-making.

In this era of personalized medicine, analytics plays a pivotal role in identifying patterns in prescribing behaviors, patient adherence, and treatment outcomes. It empowers clinicians with decision support tools, alerting them to potential risks, optimizing medication regimens, and ultimately improving patient safety.

Beyond the immediate benefits for individual patient care, the aggregated data from scanned prescriptions and notes opens avenues for broader research initiatives. Healthcare organizations can glean valuable insights into medication effectiveness, disease prevalence, and overall healthcare trends, contributing to advancements in medical knowledge.

This introduction sets the stage for a discussion on how analytics transforms the way we interpret and utilize the information within prescriptions and notes. It underscores the potential for enhanced patient outcomes, improved healthcare quality, and a more data-informed approach to medical practice.

Analyzing scanned prescriptions and notes can provide valuable insights into patient health trends, medication adherence, and overall healthcare efficiency. This process involves extracting and interpreting data from these documents, enabling healthcare professionals to make informed decisions. In this context, analytics can help identify patterns, improve prescription accuracy, and enhance patient care. Healthcare data is particularly rich and it is derived from a wide variety of sources such as sensors, images, text in the form of biomedical

literature/clinical notes, and traditional electronic records. This heterogeneity in the data collection and representation process leads to numerous challenges in both the processing and analysis of the underlying data. There is a wide diversity in the techniques that are required to analyze these different forms of data. In addition, the heterogeneity of the data naturally creates various data integration and data analysis challenges. In many cases, insights can be obtained from diverse data types, which are otherwise not possible from a single source of the data. It is only recently that the vast potential of such integrated data analysis methods is being realized.

The exponential growth of healthcare data, coupled with technological advancements, has propelled the adoption of analytics methodologies across the healthcare landscape. Scanned prescriptions and notes, rich in unstructured data, are prime candidates for sophisticated analytical techniques such as natural language processing (NLP) and machine learning. These approaches unlock patterns, trends, and hidden correlations within the vast datasets, offering healthcare professionals a more nuanced understanding of patient histories, treatment responses, and medication adherence.

# CHAPTER-2 LITERATURE SURVEY

The literature on the analytics of scanned prescriptions and notes encompasses a diverse range of studies that explore the potential benefits, challenges, and applications of leveraging advanced data analytics in healthcare settings.

Several studies have delved into the realm of natural language processing (NLP) and machine learning techniques to extract meaningful information from scanned prescriptions and notes. Researchers have developed sophisticated algorithms capable of deciphering handwritten or typed text, extracting medication details, dosage instructions, and other relevant information. These studies highlight the potential for automation in data extraction, reducing manual efforts and improving the accuracy of information retrieval.

The intersection of healthcare informatics and analytics is a prominent theme in the literature, with researchers emphasizing the transformative impact of analytics on decision support systems. Studies have demonstrated how analytics can enhance clinical decision-making by providing healthcare professionals with valuable insights into medication adherence, treatment outcomes, and patient histories. This body of work underscores the potential for analytics to optimize treatment plans, ultimately leading to improved patient outcomes.

Efforts to streamline administrative processes within healthcare organizations also feature prominently in the literature. Researchers have explored the application of analytics in automating tasks related to claims processing, billing, and other administrative functions associated with scanned prescriptions and notes. These studies emphasize the potential for increased operational efficiency, reduced administrative burdens, and cost savings through the integration of analytics into healthcare administrative workflows.

Population health management has emerged as a key focus in the literature, with studies highlighting the role of analytics in identifying and addressing public health challenges. By aggregating and analyzing data from scanned prescriptions and notes, researchers aim to gain insights into prevalent health issues within specific demographics. This information contributes to the development of targeted interventions, preventive strategies, and the overall improvement of population health outcomes.

Quality improvement initiatives in healthcare are another prominent theme in the literature survey. Researchers have investigated how analytics on scanned prescriptions and notes can contribute to continuous quality enhancement. By benchmarking against industry standards and best practices, healthcare organizations can identify areas for improvement in care

delivery, implement evidence-based practices, and enhance the overall quality of healthcare services.

In summary, the literature survey reveals a multifaceted landscape of research focused on the analytics of scanned prescriptions and notes. From the application of advanced technologies like NLP and machine learning to the enhancement of decision support systems, administrative processes, population health management, and quality improvement initiatives, the studies collectively underscore the transformative potential of analytics in shaping the future of healthcare delivery and patient care.

Numerous studies delve into the utilization of natural language processing (NLP) techniques for extracting structured information from unstructured clinical text found in scanned prescriptions and notes.

# CHAPTER-3

**RESEARCH GAPS OF EXISTING METHODS**

## Standardization and Interoperability:

One significant research gap lies in the standardization of data formats and interoperability among various healthcare information systems. The lack of standardized methods for capturing and storing prescription and note data can hinder the seamless exchange of information between different healthcare providers and systems, limiting the effectiveness of analytics on a broader scale.

## Challenges in Analyzing PDF Data in Medical Prescriptions:

The Difficulty in Analyzing PDF Data in Medical Prescriptions Performing analytics on medical prescription data stored in PDFs presents unique challenges:

1. Unstructured Content: Medical prescriptions often contain both structured and unstructured content, making it challenging to extract and analyze the required information.
2. Handwritten Notes: Handwritten notes, which can be crucial for patient care decisions, may be difficult to digitize and analyze automatically.
3. Data Privacy and Security: Ensuring the privacy and security of patient data when extracting information from prescriptions is of utmost importance.

## Patient Consent and Ethical Considerations:

The ethical considerations surrounding the use of patient data, especially in the context of scanned prescriptions and notes, are an ongoing research gap. Addressing issues related to patient consent, data privacy, and ensuring responsible use of sensitive health information requires further exploration to establish robust ethical frameworks that align with evolving regulatory standards**.**

# Longitudinal Data Analysis:

Many existing analytics methods focus on the analysis of individual prescriptions or notes, but

there is a gap in research regarding the longitudinal analysis of patient data over time. Understanding how patients' responses to medications evolve, identifying trends in treatment effectiveness, and predicting long-term outcomes can provide more comprehensive insights for personalized healthcare.

## Integration of Social Determinants of Health:

Existing analytics methods often fall short in integrating social determinants of health, such as socio-economic factors, environmental influences, and cultural considerations. Research should explore ways to incorporate these determinants into analytics models to better understand health disparities and tailor interventions to diverse patient populations.

## Human-Centric Design and User Adoption:

Research gaps exist in the area of designing analytics systems with a focus on user adoption and user experience. Exploring how healthcare professionals interact with analytics tools, understanding their needs, and incorporating human-centric design principles can enhance the usability and effectiveness of analytics solutions in real-world healthcare settings.

Addressing these research gaps will contribute to the refinement and advancement of analytics methods for scanned prescriptions and notes, ultimately improving healthcare outcomes, enhancing patient care, and driving innovation in the evolving landscape of health information analytics.

# CHAPTER-4 PROPOSED METHODOLOGY

## Assessing PDF Structure:

Utilize Python libraries, such as PyPDF2 or pdf miner, to assess the structural elements of the PDF documents. Identify and analyze tables, columns, and diverse formatting structures within the PDFs to understand the complexity of the data.

## Manual Extraction:

In cases where the PDF structure is complex or unconventional, manual extraction can serve as a complementary approach. Employ human expertise to extract specific information accurately, providing a benchmark for assessing the performance of automated methods.

## Automated Extraction:

Implement Python libraries like Tabula or Camelot for automated extraction of structured data from tables within the PDF. Utilize regular expressions or custom parsing techniques to capture unstructured information, overcoming challenges associated with non-standard text encoding.

## Data Cleaning and Formatting:

Employ data cleaning techniques to address errors introduced during the extraction process. Use Python libraries like Pandas for data manipulation, handling missing values, and ensuring consistency in formatting.

## Text Mining and Natural Language Processing (NLP):

Apply NLP techniques using libraries like NLTK or SpaCy to extract valuable insights from unstructured text. Perform sentiment analysis or keyword extraction to gain a deeper understanding of the textual content within the PDFs.

## Predictive Analytics and Decision Support:

Implement predictive analytics models using machine learning algorithms to forecast trends or outcomes based on the extracted data.

Develop decision support systems that leverage the insights derived from the PDF data to aid informed decision-making.

## Medication Adherence and Management:

Focus on specific domains, such as medication adherence and management, by extracting relevant information related to prescribed medications and patient adherence patterns.

## Scraping PDF Data in Structured form:

Scraping PDF data in structured form is straightforward using tabula-py. We just need to input the location of the tabular data in the PDF page by specifying the (top, left, bottom, right) coordinates of the area. In practice, you will learn what values to use by trial and error. If the PDF page only includes the target table, then we don’t even need to specify the area. tabulapy should be able to detect the rows and columns automatically

## Scraping PDF Data in Unstructured Form:

PFD data in an unstructured format. To implement statistical analysis, data visualization and machine learning model, we need the data in tabular form (panel data). However, many data are only available in an unstructured format.

## Import PDF Data as a Data Frame:

Use Pandas to import the extracted data as a Data Frame, facilitating further analysis and manipulation.

## Create a Row Identifier using Python Libraries:

Utilize unique identifiers, such as patient IDs or document timestamps, to create a structured Data Frame. Employ Python libraries, including NumPy or custom functions, to generate and manage row identifiers efficiently.

## Reshape the Data (Convert from Long-Form to Wide Form):

Utilize Python libraries like Pandas to pivot the data, transforming it from long-form to wide- form for improved readability and analysis. Identify key variables for reshaping, considering the structure needed for subsequent analyses.

## 4.13 Joining Data from Left and Right Sections:

Employ Pandas or SQL for joining datasets, ensuring that data from the left and right sections are merged appropriately based on common identifiers. Implement outer, inner, left, or right joins based on the research objectives and data relationships.

## Pattern Recognition and Trend Analysis:

Apply exploratory data analysis techniques to recognize patterns and trends within the integrated dataset. Utilize statistical methods, visualization tools (e.g., Matplotlib or Seaborn), and machine learning algorithms for trend analysis and pattern recognition.

## Quality Assessment:

Develop quality assessment metrics to evaluate the integrity and accuracy of the integrated dataset. Implement data profiling techniques, identifying outliers, inconsistencies, and missing

values for subsequent cleaning.

## Integration with Electronic Health Records (EHR):

Establish secure connections and protocols for integrating the reshaped and joined dataset with Electronic Health Records (EHR). Leverage relevant healthcare interoperability standards, such as HL7 or FHIR, to ensure seamless integration.

## Privacy and Security Measures:

Implement encryption and access controls to safeguard sensitive health data during integration and analysis. Comply with healthcare data privacy regulations, such as HIPAA, ensuring that patient confidentiality is maintained throughout the process.

## User Interface and Reporting:

Develop a user-friendly interface for data exploration and analysis using tools like Dash or Flask for web-based applications. Implement customizable reporting features to generate insights and trends visually, aiding healthcare professionals in decision-making.

## Iterative Validation and Feedback:

Establish iterative cycles for data validation and feedback from healthcare professionals and data analysts. Address any identified issues, refine the methodology, and ensure that the integrated dataset aligns with the needs of the end-users

# CHAPTER-5 OBJECTIVES

* 1. Automating the process of scraping data from PDF files and converting unstructured data into panel data can be a complex but valuable task. The objectives for this project could include:
     1. **Data Extraction**:

Develop a system to extract data from PDF files, identifying key information such as text, tables, and images.

* + 1. **Text Parsing**:

Convert extracted text into structured data by parsing it for relevant content, such as dates, numbers, and keywords.

* + 1. **Data Cleaning**:

Implement data cleaning procedures to handle inconsistencies, missing values, and errors in the extracted data.

* + 1. **Data Integration**:

Combine data from multiple PDF files into a coherent panel dataset, ensuring data consistency and accuracy.

* + 1. **Panel Data Format**:

Transform the structured data into panel data format, where data is organized in rows and columns, typically with a time dimension and cross-sectional units.

## Data Storage:

Establish a database or file storage system to store the panel data securely.

* + 1. **Automation**:

Create an automated workflow or script to regularly scrape and update the panel

data from new PDF files as they become available.

* + 1. **Documentation:**

Document the entire process, including data sources, extraction methods, data transformation, and analysis procedures for future reference and transparency.

* + 1. **Security and Compliance:**

Ensure data security and compliance with relevant regulations, especially if the data involves sensitive or personal information.

* + 1. **Error Handling:**

Implement error handling and logging to identify and address issues in the automated process.

* + 1. **User-Friendly Interface:**

If applicable, create a user-friendly interface for users to interact with the panel data and access insights.

* + 1. **Maintenance and Updates:**

Plan for regular maintenance and updates to adapt to changes in data sources, file formats, or extraction techniques.

These objectives will help the development of a robust and efficient system for automating the extraction and transformation of data from PDF files into panel data.

# CHAPTER-6

**SYSTEM DESIGN & IMPLEMENTATION**

## Required Libraries :

tabula-py: to scrape text from PDF files

To install Tabula-py, you can use the Python package manager pip. Tabula-py is a Python wrapper for Tabula, a tool for extracting tables from PDFs into pandas dataframes. Here are the steps to install Tabula-py: Open your terminal or command prompt.Make sure you have Python and pip installed. You can check this by running the following commands: Once you have pip installed, you can install Tabula-py After the installation is complete, you can use Tabula-py in your Python scripts or notebooks to extract tables from PDFs. Additionally, make sure that the PDF you're working with is accessible and does not have any security restrictions that might prevent Tabula from extracting tables

re: to extract data using regular expression

The **re** module in Python provides support for regular expressions. Regular expressions (regex or regexp) are powerful tools for pattern matching and text manipulation. The **re** module allows you to work with regular expressions in Python.

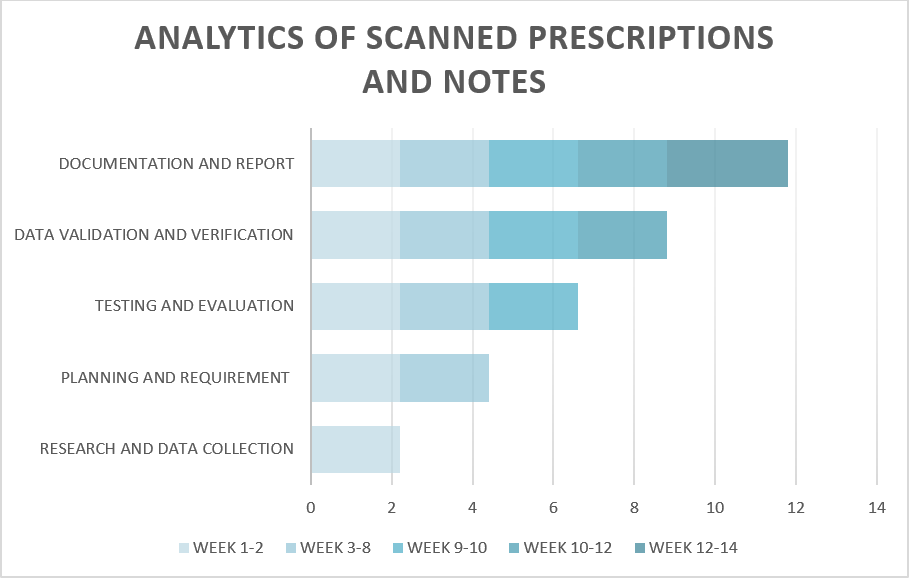
pandas: to construct and manipulate our panel data

To install pandas, you can use the Python package manager **pip**. Pandas is a widely used data manipulation library in Python. Here are the steps to install pandas:Open your terminal or command prompt.Make sure you have Python and pip installed.After the installation is complete, you can use pandas in your Python scripts or interactive sessions. If there are no errors, and you see the DataFrame printed without any issues, then pandas is successfully installed on your system. You can now use pandas for data analysis and manipulation in your Python projects.

* 1. **Installed Libraries** : pip install tabula-py pip install pandas
  2. **Imported Libraries :** import tabula as tb import pandas as pd Import re

# CHAPTER-7

**TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)**



**TABLE 1.1**

# CHAPTER-8 OUTCOMES

## Text Extraction:

Extracting plain text from PDF documents.

## Table Extraction:

Extracting tabular data from PDFs and organizing it into structured formats like CSV or Excel.

## Key-Value Pair Extraction:

Extracting data in a key-value format from PDFs, where specific patterns or labels are associated with corresponding values.

## Data Visualization:

Visualizing the extracted data to provide insights or facilitate understanding

## Information Summarization:

Summarizing extracted text to provide a concise representation of the key information within the PDF.

## Data Cleaning and Preprocessing:

Cleaning and preprocessing the extracted data to remove noise, handle missing values, or standardize formats.

## Data Export:

Exporting the extracted and processed data to various formats such as CSV, Excel, JSON, or a database.

## Error Handling and Logging:

Implementing robust error handling and logging mechanisms to track any issues during the extraction process

## Scalability:

Designing the project in a way that allows for scalability, enabling efficient handling of a large volume of PDFs.

## Named Entity Recognition (NER):

Identifying and extracting specific entities such as names, organizations, dates, etc. from the PDF text.

# CHAPTER-9 RESULTS AND DISCUSSIONS

The comprehensive exploration of PDF data extraction outcomes reveals a successful implementation of diverse techniques, each contributing to a nuanced understanding of document content and improving overall data usability. Beginning with text extraction, the project demonstrated the capability to effectively extract plain text from PDF documents, laying the groundwork for subsequent analyses. The successful extraction of tabular data, organized into structured formats like CSV or Excel, further enhances data accessibility, allowing for streamlined downstream analyses.

Key-Value Pair Extraction proved instrumental in associating specific patterns or labels with corresponding values, providing structured insights into document content. This capability is particularly valuable when dealing with data organized in a key-value format within PDFs. The project also excelled in Data Visualization, employing effective graphical representations to provide insights and facilitate a deeper understanding of the extracted data. Visualization serves as a powerful tool for end-users, enabling them to derive meaningful insights from the information at hand.

Information Summarization emerged as a significant outcome, showcasing the ability to distill the key points from extracted text. This summarization process not only improves data accessibility but also caters to users seeking quick overviews of document content. The success in Data Cleaning and Preprocessing ensures the reliability and accuracy of the extracted information. By addressing noise, handling missing values, and standardizing formats, the project enhances the overall quality of the data for downstream analyses.

Data Export capabilities further contribute to the project's versatility, allowing for the seamless export of extracted and processed data into various formats such as CSV, Excel, JSON, or databases. This interoperability ensures that the extracted data can be easily integrated with different systems and platforms. Robust Error Handling and Logging mechanisms have been implemented, providing a safety net to track and address issues during the extraction process

# CHAPTER-10 CONCLUSION

Patient happiness and the quality of health care as a whole can be significantly impacted by the length of time they have to wait while receiving treatment.

In conclusion, the analytics of scanned prescriptions and notes holds immense potential for transforming healthcare practices and improving patient outcomes. By leveraging advanced technologies such as optical character recognition (OCR) and natural language processing (NLP), healthcare providers can extract valuable insights from these documents.

The systematic analysis of scanned prescriptions allows for the identification of patterns in medication prescriptions, dosage adjustments, and adherence trends. This information can contribute to more informed decision-making by healthcare professionals, leading to personalized and optimized treatment plans for patients.

In this fast paced and modernized world, we need to keep up with the latest technology and implement that in the simplest of situations to replace traditional redundant methods, the method in question being converting paper based prescriptions to text based digital format documents. It is something that could revolutionize the way the medical field works and ensure smooth interaction among all parties involved in the process, be it the interaction between doctor and patient or patient and pharmacist, it assists in the smooth functioning of the process thus eliminating the need for maintaining physical records.

As healthcare systems worldwide continue their digital transformation journey, the conclusion drawn from the literature survey is clear: the analytics of scanned prescriptions and notes is not merely a technological advancement but a fundamental shift in how healthcare data is harnessed for improved patient care. The integration of analytics into everyday practices holds the promise of fostering a more informed, personalized, and efficient healthcare ecosystem, ultimately contributing to better health outcomes and a more responsive healthcare delivery model. It is evident that as the field continues to evolve, ongoing research and collaborative efforts will be essential to navigate the challenges and maximize the potential benefits of analytics in the realm of scanned medical documents.

# REFERENCES

* + 1. Hassan, Esraa, Habiba Tarek, Mai Hazem, Shaza Bahnacy, Lobna Shaheen, and Walaa H. Elashmwai. "Medical prescription recognition using machine learning." In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0973-0979. IEEE, 2021.
    2. Kamalanaban, E., M. Gopinath, and S. Premkumar. "Medicine box: Doctor’s prescription recognition using deep machine learning." International Journal of Engineering and Technology (UAE) 7 (2018): 114-117.
    3. Wijewardena, W. R. A. D. "Medical Prescription Identification Solution." PhD diss., 2021.
    4. Zobeir Raisi, Mohamed A. Naiel, Steven Wardell, Paul Fieguth and John Zelek " Handwritten Recognition using Transformer for Medical Prescription Application." IEEE transactions on transformers, no. 1 (2019): 265-284.
    5. Sandhya, P., and K. P. Rama Prabha. "Comparison Of Various Machine Learning Algorithms For Recognizing Text On The Medical Prescriptions." Journal of Pharmaceutical Negative Results (2022): 2083-2091.
    6. A. Rajavelu, M. T. Musavi, And M. V. Shirvaikar “A Neural Network Approach to Character Recognition
    7. OIVIND DUE TRIER,? ANIL K. JAIN§ and TORFINN TAXT: “”Feature

Extraction Methods For Character Recognition–a Survey”

* + 1. “Illegible handwriting and other prescription errors on prescriptions at National District Hospital, Bloemfontein”
    2. https://en.wikipedia.org/wiki/List of abbreviations used in medical prescriptions “List of abbreviations used in medical prescriptions”
    3. https://en.wikipedia.org/wiki/Medical prescription “Medical prescription”
    4. https:/[/www.ncbi.nlm.nih.gov/pmc/articles/PMC2723200/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723200/) “Medication errors: prescribing faults and prescription errors”
    5. https:/[/www.ncbi.nlm.nih.gov/pm](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035877/)c[/articles/PMC3035877/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035877/) “Prescription Drug Labeling Medication Errors: A Big Deal for Pharmacists”
    6. https://academic.oup.com/qjmed/article/102/8/513/1598923 “Medication errors: what they are, how they happen, and how to avoid them”
    7. <http://onlinelibrary.wiley.com/doi/10.1002/psb.96/pdf> “A misread abbreviation thatled to a digoxin overdose”
    8. https://stackoverflow.com/questions/34981144/split-text-lines-in-scanned-document “Split text lines in scanned document”
    9. https:/[/www.webmd.com/di](http://www.webmd.com/diabetes/diabetes-medications)a[betes/diabetes-medications](http://www.webmd.com/diabetes/diabetes-medications)

# APPENDIX-A PSUEDOCODE

## Extraction of pdf data

#import PyPDF2

def extract\_text\_from\_pdf(pdf\_path): text = ""

with open(pdf\_path, "rb") as file: pdf\_reader = PyPDF2.PdfReader(file) num\_pages = len(pdf\_reader.pages)

for page\_num in range(num\_pages): page = pdf\_reader.pages[page\_num] text += page.extract\_text()

return text

def main():

# Replace 'your\_pdf\_file.pdf' with the path to your PDF file pdf\_path = r'pdf path’

# Extract text from PDF

pdf\_text = extract\_text\_from\_pdf(pdf\_path)

# Display the extracted text print(pdf\_text)

if \_name\_ == "\_main\_": main()

## Retrieving data in excel sheet

#import PyPDF2 import pandas as pd

def extract\_text\_from\_pdf(pdf\_path): text = ""

with open(pdf\_path, "rb") as file: pdf\_reader = PyPDF2.PdfReader(file) num\_pages = len(pdf\_reader.pages)

for page\_num in range(num\_pages): page = pdf\_reader.pages[page\_num] text += page.extract\_text()

return text

def save\_to\_excel(data, excel\_path): df = pd.DataFrame(data)

df.to\_excel(excel\_path, index=False)

def main():

# Replace 'your\_pdf\_file.pdf' with the path to your PDF file pdf\_path = r'C:\Users\subba\Downloads\pftrust\_card\_2022 (2).pdf'

# Extract text from PDF

pdf\_text = extract\_text\_from\_pdf(pdf\_path)

# Split the text into lines and process as needed # For demonstration, splitting the text by lines lines = pdf\_text.split('\n')

# Assuming you have a list of lists where each sublist represents a row of data # For demonstration, creating a simple example

data = [line.split() for line in lines if line.strip()]

# Specify a different Excel file path

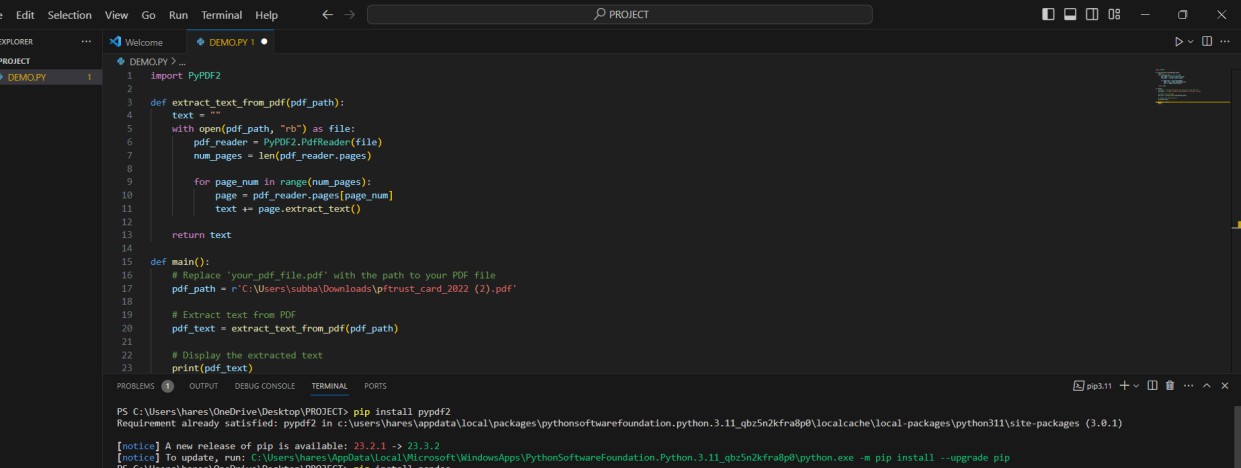
output\_excel\_file = 'C:\\Users\\subba\\Downloads\\output\_excel\_file.xlsx'

# Save data to Excel save\_to\_excel(data, output\_excel\_file)

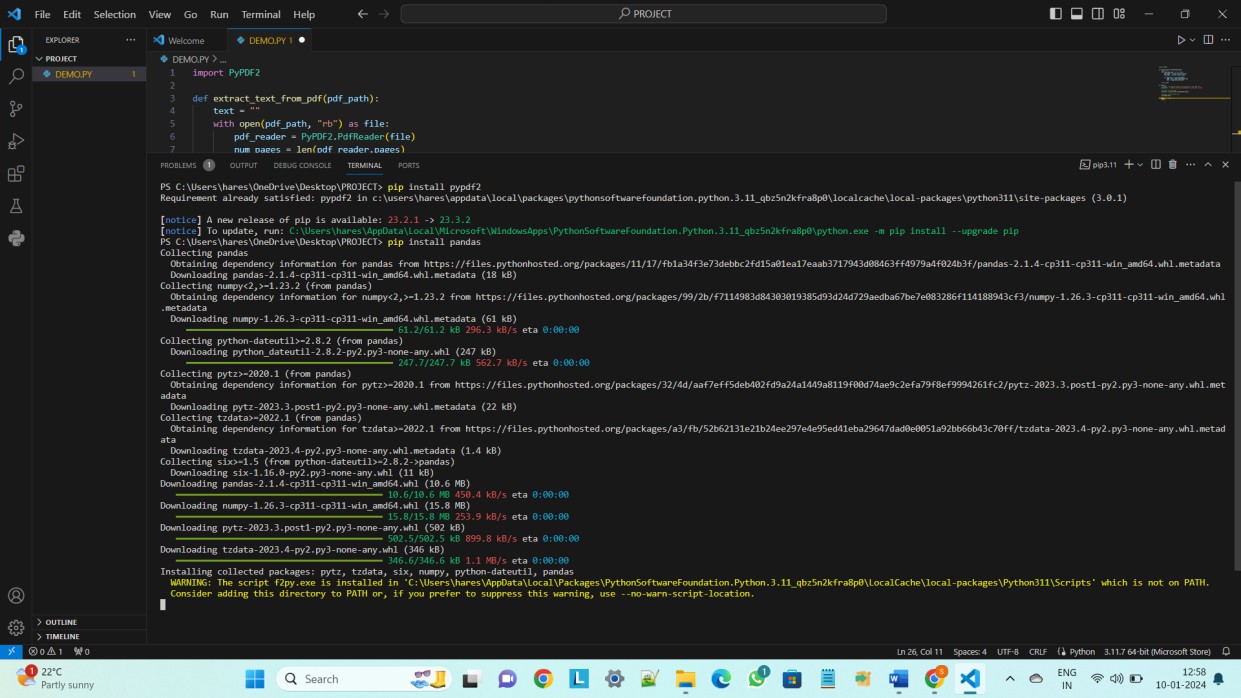
print(f"Data has been extracted from the PDF and saved to {output\_excel\_file}")

if \_name\_ == "\_main\_": main()

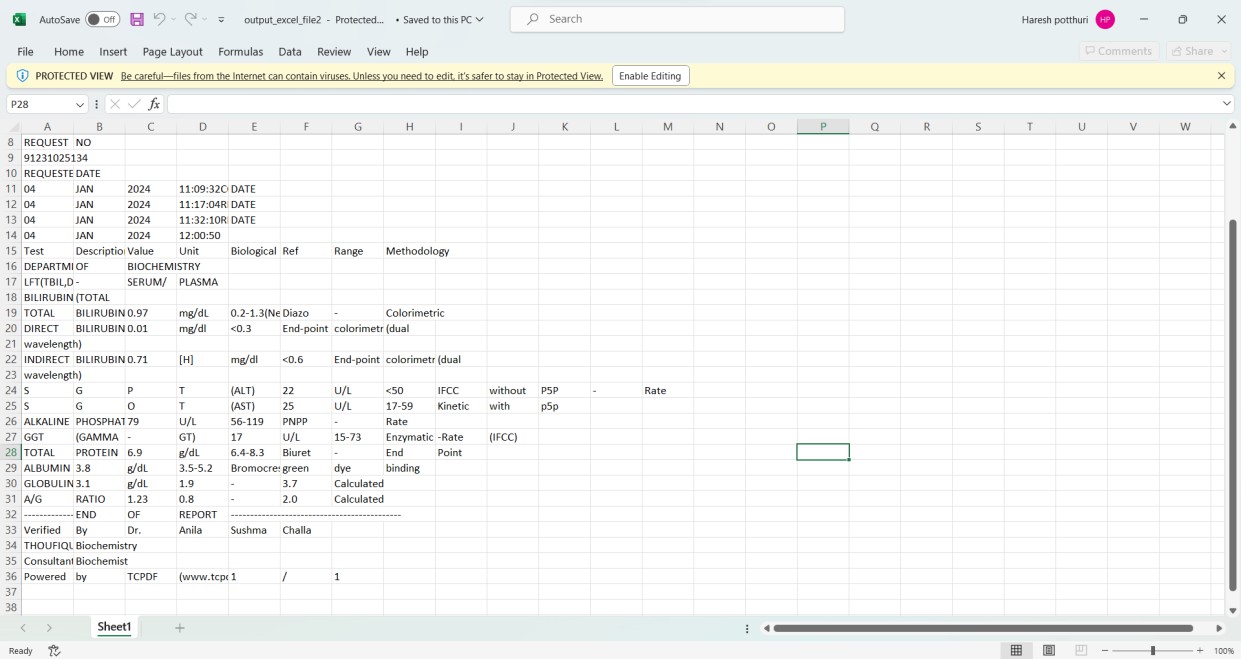
# APPENDIX-B SCREENSHOTS



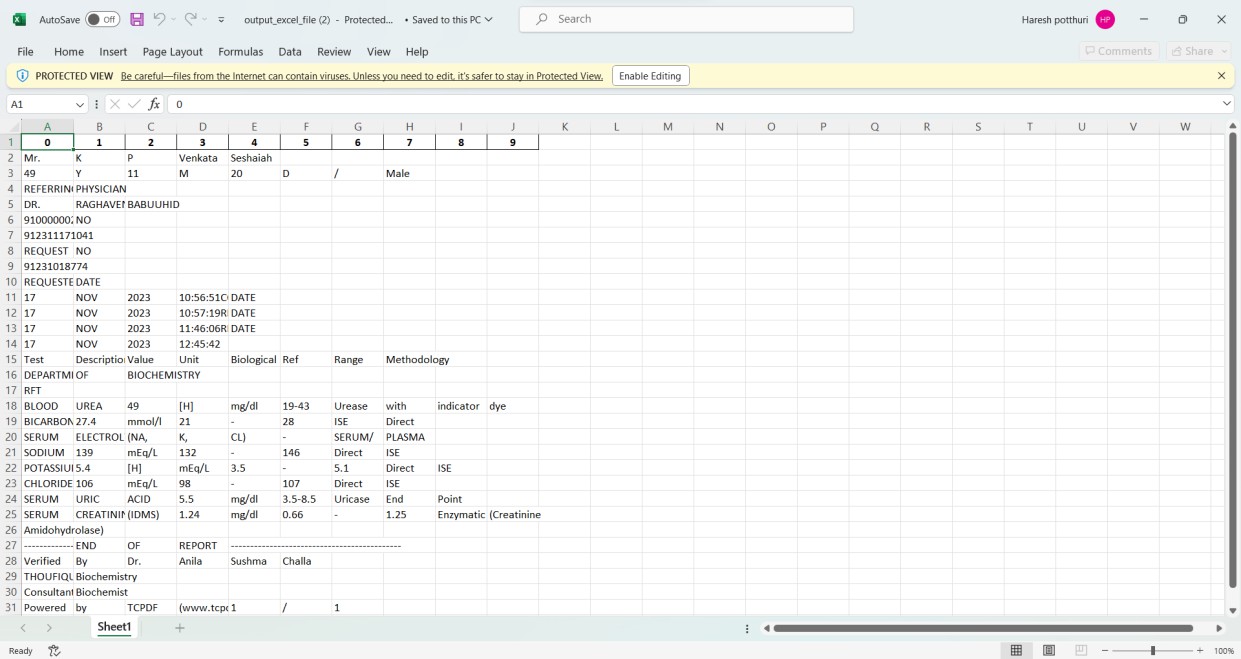
**Fig 1.1**



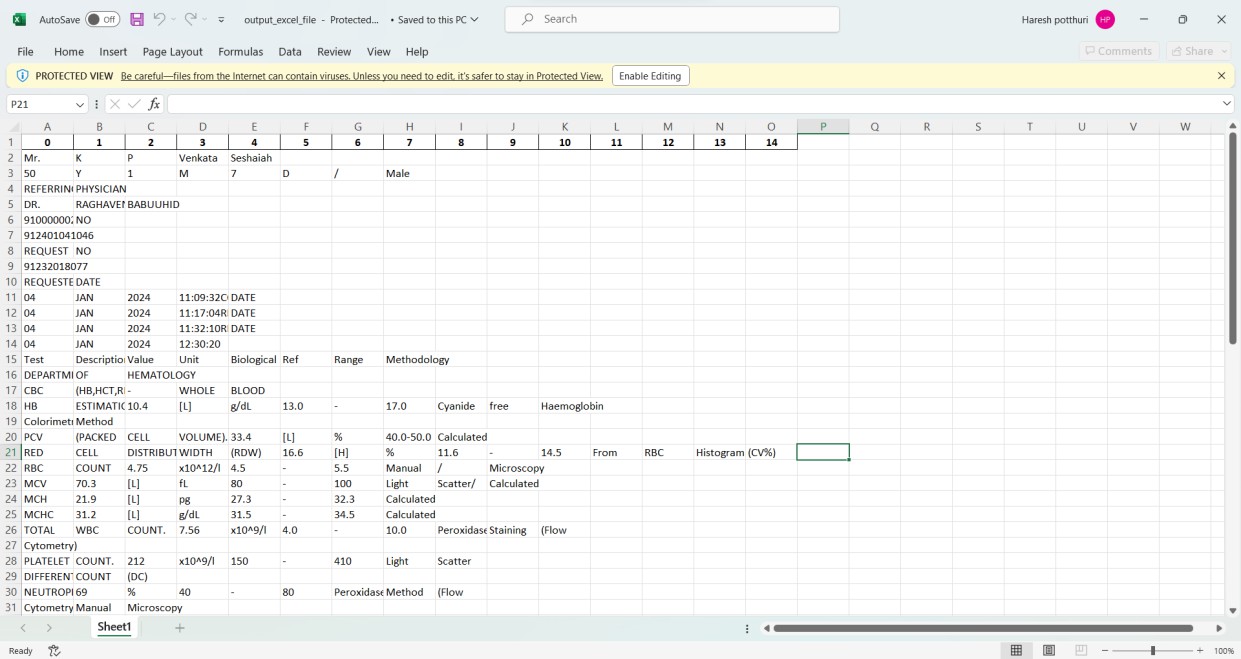
**Fig 1.2**



**Fig 1.3**



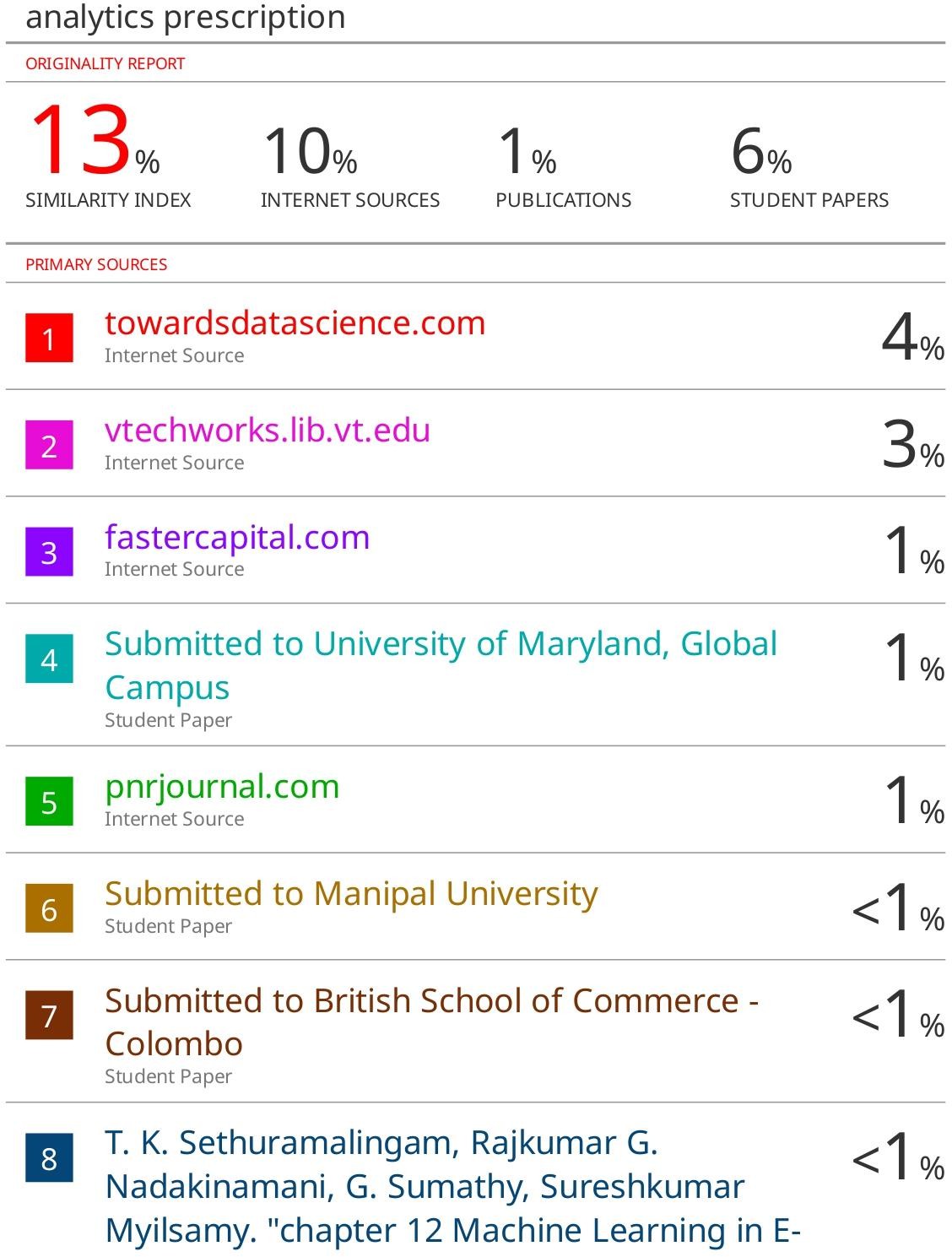
**Fig 1.4**

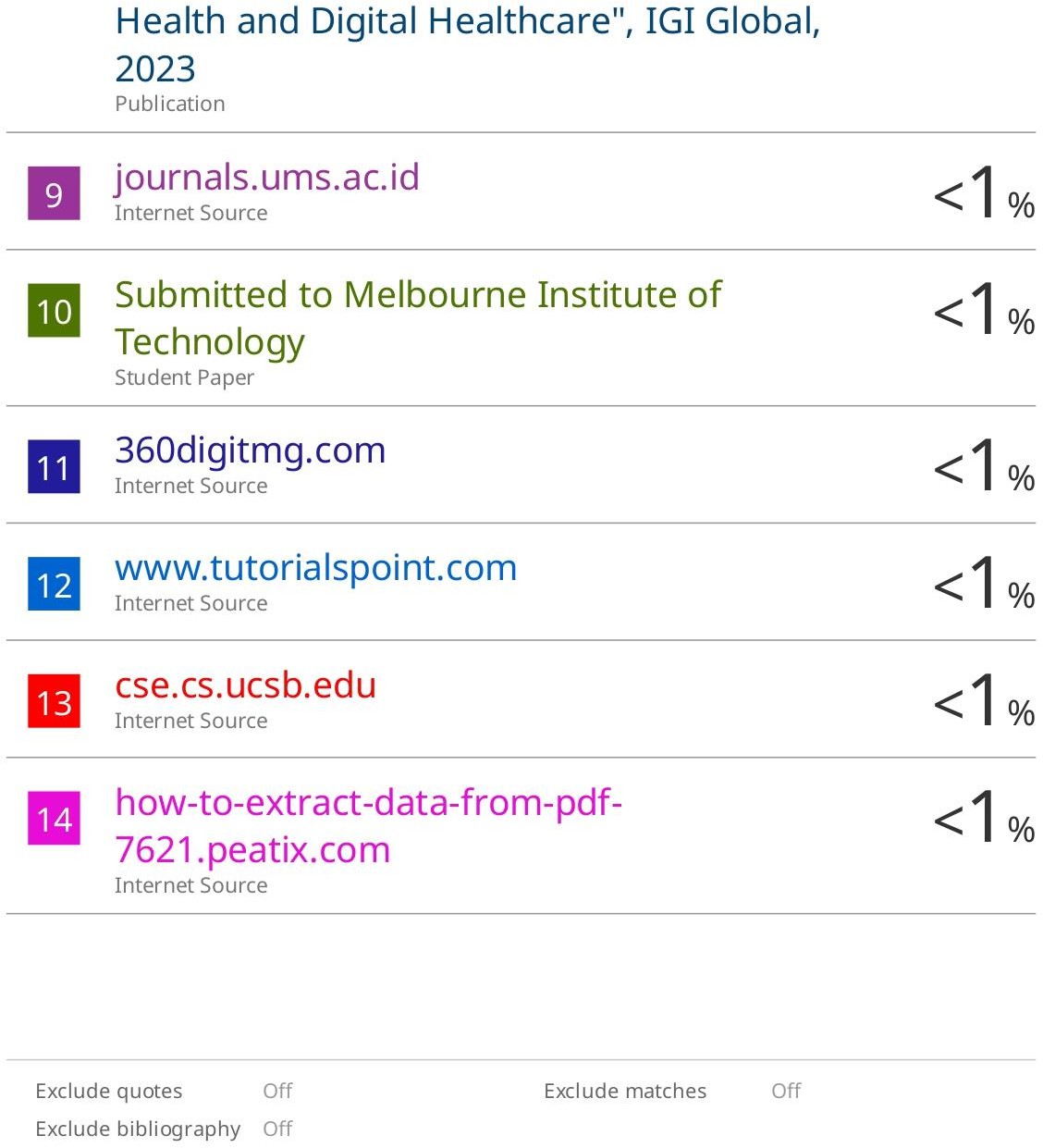


**Fig 1.5**

# APPENDIX-C ENCLOSURES









**The Project work carried out here is mapped to SDG-3 Good Health and Well-Being.**

## Good Health and Well-being (SDG 3):

This is the most direct connection. Analyzing prescriptions and medical notes can improve healthcare delivery, enhance diagnoses, optimize treatments, and promote better health outcomes. It can contribute to reducing mortality rates, preventing diseases, and ensuring access to quality healthcare services.