SCRAPING PDF INVOICES BOT

# A PROJECT REPORT

***Submitted by***

# PRANNOV SHABARI N (220701197)

***in partial fulfillment for the course***

## OAI1903 - INTRODUCTION TO ROBOTIC PROCESS AUTOMATION

***for the degree of***

# BACHELOR OF ENGINEERING

***in***

**COMPUTER SCIENCE AND DESIGN**

# RAJALAKSHMI ENGINEERING COLLEGE RAJALAKSHMI NAGAR

**THANDALAM CHENNAI – 602 105**

**DECEMBER 2024**

# RAJALAKSHMI ENGINEERING COLLEGE CHENNAI - 602105

BONAFIDE CERTIFICATE

Certified that this project report **“The Smart PDF Scraping Bot”** is the bonafide work of **“PRANNOV SHABARI N (2116220701197)”** who carried out the project work for the subject OAI1903 - Introduction to Robotic Process Automation under my supervision.

## SIGNATURE

**Mrs. G. M. Sasikala, M.E., SUPERVISOR,**

**Assistant Professor**,

Department of

Computer Science and Engineering, Rajalakshmi Engineering College,

Rajalakshmi Nagar,

Thandalam,

Chennai- 602105.

Submitted to Project and Viva Voce Examination for the subject OAI1903 - Introduction to Robotic Process Automation held on .

**Internal Examiner External Examiner**

# ABSTRACT

This project presents an automation solution utilizing the UiPath Automation Platform to streamline the extraction and organization of invoice data. The robot is designed to efficiently scrape key information, including Company Name, Invoice Number, Date, Customer ID, Payment Terms, and Total Amount, from a dataset of 1000 invoices issued to various customers. The extracted data is systematically stored in an Excel file (*invoices\_data.xlsx*), significantly reducing manual effort and the risk of human error.

Additionally, the solution classifies processed invoices into "successfully scraped" and "unsuccessfully scraped" categories, organizing them into separate folders for easy reference and error handling. Remarkably, the RPA robot completes this high-volume data extraction task in approximately 130 minutes with near-perfect accuracy—drastically outperforming the manual process, which would require several days of effort. This project demonstrates the transformative potential of RPA in automating repetitive and labour-intensive processes, enhancing operational efficiency, and ensuring data reliability.

This RPA-driven approach highlights the potential of automation to revolutionize business processes, enabling organizations to achieve higher productivity and reliability. The project underscores the value of leveraging intelligent automation tools to simplify complex workflows, optimize resource utilization, and ensure scalability in handling large datasets.

# ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. Meganathan, B.E, F.I.E.,** our Vice Chairman **Mr. Abhay Shankar Meganathan, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) Thangam Meganathan, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. Murugesan, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. Revathy, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Design for her guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guides, **Mrs. Roxanna Samuel, M.E.,** Assistant Professor (SG), **Ms. Farjana, M.E.,** Assistant Professor (SG), **Ms. Vinothini, M.E.,** Assistant Professor (SG), Department of Computer Science and Engineering, Rajalakshmi Engineering College for their valuable guidance throughout the course of the project. We are very glad to thank our Project Coordinators, **Dr. N. Durai Murugan, M.E., Ph.D.,** Associate Professor, and **Mrs. G.M. Sasikala, M.E.,** Assistant Professor (SG), Department of Computer Science and Engineering for their useful tips during our review to build our project.

**Prannov Shabari N (220701197)**

# TABLE OF CONTENTS

**CHAPTER NO. TITLE PAGE NO.**

**ABSTRACT**

**LIST OF FIGURES**

**LIST OF ABBREVIATIONS**

1. **INTRODUCTION**

**iii vi vii 1**

* 1. [INTRODUCTION 1](#_TOC_250011)
  2. [OBJECTIVE 3](#_TOC_250010)
  3. [EXISTING SYSTEM 3](#_TOC_250009)
  4. [PROPOSED SYSTEM 4](#_TOC_250008)

1. LITERATURE REVIEW 5
2. SYSTEM DESIGN 9
   1. [SYSTEM FLOW DIAGRAM 9](#_TOC_250007)
   2. [ARCHITECTURE DIAGRAM 10](#_TOC_250006)
   3. [SEQUENCE DIAGRAM 11](#_TOC_250005)
3. PROJECT DESCRIPTION 12
   1. [MODULES 12](#_TOC_250004)
      1. INPUT HANDLING AND 12

INITIALIZATION

* + 1. [CONTENT ANALYSIS 12](#_TOC_250003)
    2. [RESULT MANAGEMENT 13](#_TOC_250002)
    3. [COMPLETION AND REPORTING 13](#_TOC_250001)

1. OUTPUT SCREENSHOTS 14
2. CONCLUSION 18

APPENDIX 19

[REFERENCES 25](#_TOC_250000)

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Name** | **Page No.** |
| 3.1 | System Flow Diagram | 14 |
| 3.2 | Architecture Diagram | 15 |
| 3.3 | Sequence Diagram | 16 |
| 5.1 | Input Dialog | 20 |

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| **ABBREVIATION** | **ACCRONYM** |
| RPA | Robotic Process Automation |
| AI | Artificial Intelligence |
| API | Application Programming  Interface |
| CV | Computer Vision |
| OCR | Optical Character Recognition |

**CHAPTER 1 INTRODUCTION**

## INTRODUCTION

In today's fast-paced business environment, managing large volumes of data accurately and efficiently is a critical challenge. Invoice processing, in particular, is a time-consuming and error-prone task when done manually, often requiring significant effort and resources. This project addresses these challenges by leveraging the UiPath Automation Platform to develop an intelligent robotic process automation (RPA) solution.

The solution is designed to extract essential data fields—Company Name, Invoice Number, Date, Customer ID, Payment Terms, and Total Amount—from a batch of 1000 customer invoices. By automating this process, the RPA robot significantly reduces manual effort, ensures near-perfect accuracy, and enhances the speed of data processing. The extracted information is systematically organized into an Excel file (*invoices\_data.xlsx*), providing a centralized repository for seamless access and analysis.

Moreover, the solution incorporates error-handling mechanisms by sorting invoices into "successfully scraped" and "unsuccessfully scraped" folders. This feature not only simplifies post-processing review but also ensures that data inconsistencies are promptly flagged for corrective action.

This project highlights the transformative potential of automation in addressing repetitive and labor-intensive tasks. By deploying this RPA solution, businesses can achieve substantial time savings, reduce operational costs, and improve data reliability, paving the way for increased efficiency and scalability in high-volume data management workflows.

## OBJECTIVE

The objective of this project is to develop an RPA solution using UiPath to automate the extraction and organization of invoice data, reducing processing time from days to 130 minutes while ensuring near-perfect accuracy. The solution aims to extract key details from 1000 invoices, store the data in a structured Excel file (*invoices\_data.xlsx*), and categorize invoices into “successfully scraped” and “unsuccessfully scraped” folders for efficient error handling. This project seeks to optimize resource utilization, reduce costs, and provide a scalable, reliable solution for managing large datasets with enhanced efficiency.

## EXISTING SYSTEM

In the current system relies on manual extraction and entry of invoice data, requiring human operators to review each invoice individually and record key details into spreadsheets or databases. This process is slow, labor-intensive, and error-prone, often taking several days to process large batches like 1000 invoices. It lacks scalability, consistency, and effective error-handling, leading to delays, inaccuracies, and higher operational costs. The inefficiencies of the existing approach highlight the need for an automated solution to enhance speed, accuracy, and productivity.

## PROPOSED SYSTEM

"The proposed system leverages the UiPath Automation Platform to automate the extraction and organization of invoice data, eliminating the need for manual processing. It efficiently extracts key details from 1000 invoices, compiles the data into a structured Excel file (*invoices\_data.xlsx*), and categorizes invoices into “successfully scraped” and “unsuccessfully.

# CHAPTER 2 LITERATURE REVIEW

### Survey on Robotic Process Automation (RPA) in Education:

While this project focuses on automating invoice data extraction using UiPath, the principles and benefits of Robotic Process Automation (RPA) extend across various sectors, including education. In educational institutions, RPA has been applied to automate repetitive tasks such as student enrolment, attendance tracking, grade reporting, and fee collection. Similar to the automation of invoice processing in this project, RPA in education ensures accuracy, reduces manual workload, and enhances operational efficiency.

Studies and surveys highlight how RPA adoption has streamlined data management in education, replacing error-prone manual processes with reliable automated systems. For example, institutions have utilized RPA to gather student data, validate application forms, and generate reports, much like how this project processes and organizes invoice data in a structured format.

This survey provides insights into how RPA can be a versatile tool across different domains. The success of automating invoice processing in this project mirrors the potential of RPA in education to handle large datasets, reduce errors, and save time, allowing staff to focus on more strategic and impactful tasks. Such comparisons underline the universality and adaptability of RPA as a solution to modern data processing challenges.

### Survey on AI-Generated Content Detection:

While this project focuses on automating the extraction and organization of invoice data using UiPath, the broader field of AI-generated content detection offers insights into ensuring data integrity and reliability in automation workflows. AI-generated content detection tools are designed to identify and differentiate between human-created and AI-produced outputs, leveraging linguistic analysis, pattern recognition, and metadata evaluation.

In the context of this project, such detection methods could be integrated into automated workflows to verify the authenticity of invoice data, ensuring that input data has not been manipulated or falsified by AI systems. This is particularly relevant for businesses dealing with large datasets where fraudulent or AI-generated invoices might compromise accuracy and operational efficiency.

Surveys and studies in AI-generated content detection emphasize its growing importance in fields like finance and data processing. Detection systems such as OpenAI's GPT detectors and other machine learning models are frequently used to assess the credibility of data sources. Integrating such technologies with RPA solutions like UiPath could enhance the robustness of automated systems, ensuring that the data processed is both authentic and reliable. This integration would not only increase trust in automated processes but also add a layer of compliance with data governance standards, further reducing risks associated with fraudulent or duplicate invoices.

### Survey on Plagiarism Detection:

Although this project focuses on automating invoice data extraction using UiPath, insights from plagiarism detection research can enhance the system's ability to verify the originality and authenticity of invoice data. Plagiarism detection tools are widely used in education, publishing, and corporate environments to identify duplicate or unoriginal content by comparing inputs against extensive databases. These systems employ algorithms to detect patterns, linguistic similarities, and metadata correlations that suggest copied or tampered data.

In the context of this project, such detection methods could be adapted to identify duplicate or fraudulent invoices within the dataset, ensuring the extracted data is both unique and accurate. For example, integrating plagiarism detection algorithms into the RPA solution could flag invoices with suspiciously similar details, aiding in fraud prevention.

Surveys and studies on plagiarism detection highlight the importance of verifying the integrity of data in automated workflows. Tools like Turnitin and Copyscape have proven effective in detecting content duplication. Applying similar principles in an RPA-driven invoice management system could further enhance data validation and bolster the system's reliability, ensuring businesses process only authentic and error-free information. Integrating such technologies with RPA solutions like UiPath could enhance the robustness of automated systems, ensuring that the data processed is both authentic and reliable. This integration could also support compliance with regulatory standards by ensuring data transparency and protecting against potential AI-driven fraud attempts.

### Summary of the intersection of RPA, AI Detection, and Plagiarism Checks:

“The Scraping PDF Invoice Bot”This project, which focuses on automating invoice data extraction using UiPath, benefits from the intersection of RPA, AI detection, and plagiarism checks to ensure data accuracy, integrity, and security. RPA automates the extraction of key information such as Company Name, Invoice Number, Date, Customer ID, Payment Terms, and Total Amount from large datasets of invoices, reducing manual effort and the risk of human error.

Integrating AI-generated content detection into the automation workflow can verify that the extracted data is authentic and has not been manipulated by AI, providing an additional layer of security. AI detection systems can identify inconsistencies or patterns suggesting that data may have been tampered with or falsified, ensuring the processed data is trustworthy.

Similarly, plagiarism detection methods can be applied to identify duplicate or fraudulent invoices within the dataset, preventing errors such as repeated entries or data manipulation. This combination of RPA with AI detection and plagiarism checks not only streamlines invoice processing but also enhances the accuracy and reliability of the entire data extraction workflow, providing businesses with a secure, efficient solution for managing large volumes of invoice data.

# CHAPTER 3 SYSTEM DESIGN

## SYSTEM FLOW DIAGRAM

The RPA workflow begins by receiving a batch of invoices and extracting key data fields. The extracted data is validated, with invalid data moved to the "unsuccessfully scraped" folder for review. Valid data is stored in an Excel file, and invoices are categorized into "successfully scraped" or "unsuccessfully scraped" folders. The process concludes with all invoices processed and categorized for easy review.

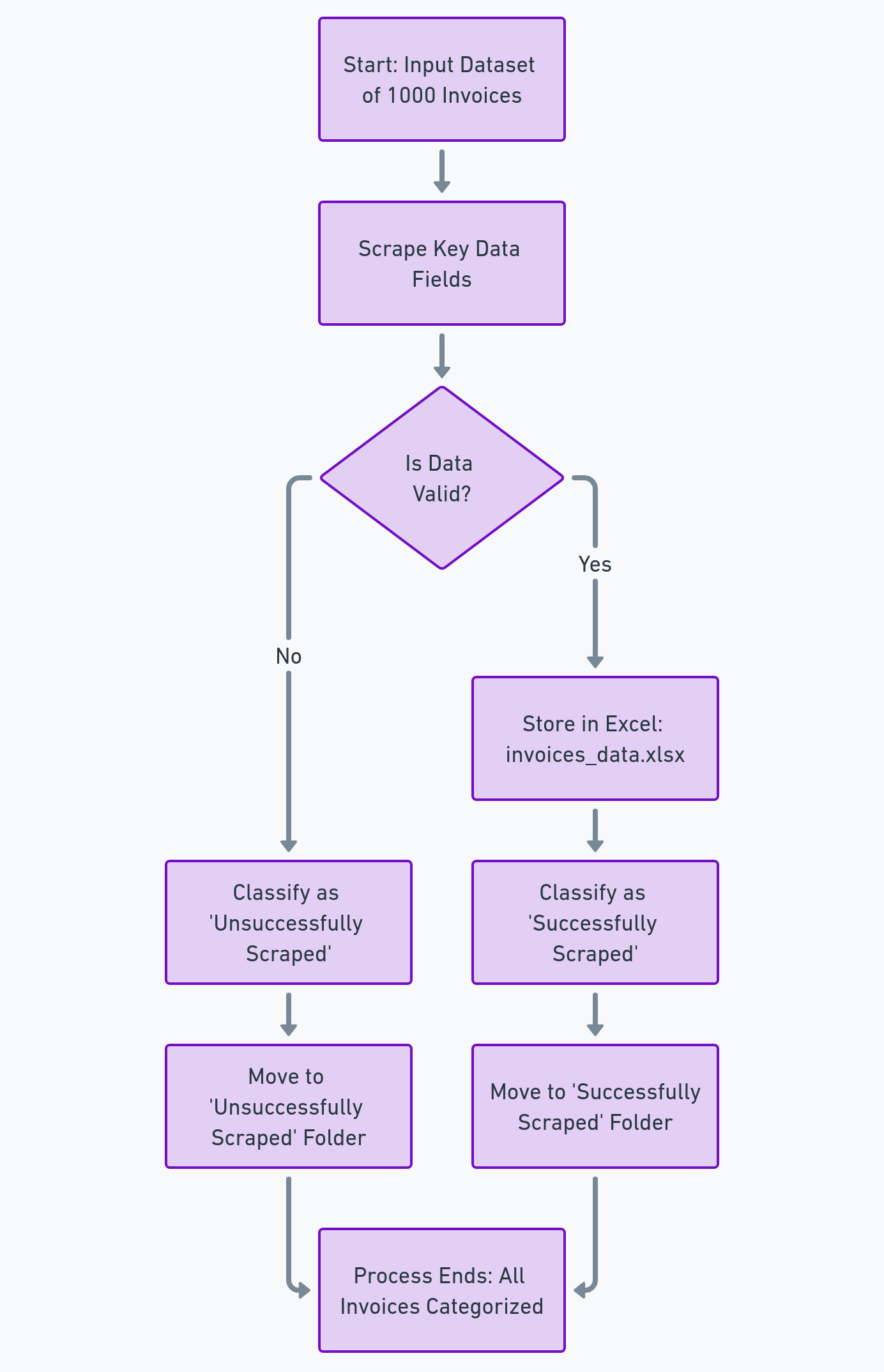


Fig 3.1 System Flow Diagram

## ARCHITECTURE DIAGRAM

The system architecture includes four layers: the Input Layer where invoices are received from a folder or external source, the ProcessingLayer where data is extracted and validated using UiPath RPA. The Storage Layer stores valid data in an Excel file and categorizes invoices, while the Error Handling Layer flags and moves invalid invoices to the "unsuccessfully scraped" folder for review.

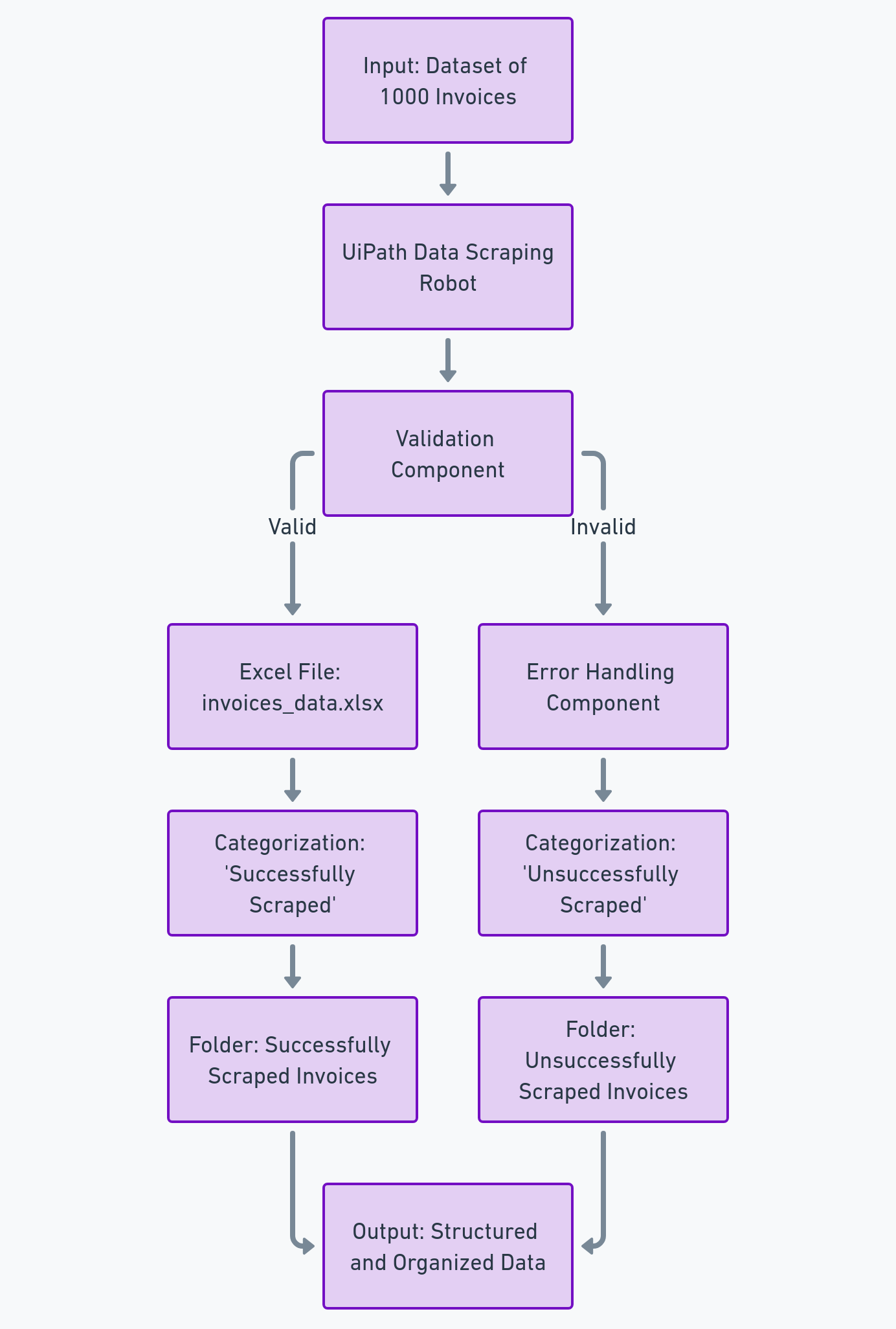


Fig 3.2 Architecture Diagram

## SEQUENCE DIAGRAM

The sequence diagram begins with the user initiating the RPA workflow, which triggers the system to fetch invoices from the input source. The UiPath RPA robot then extracts and validates the data from each invoice. Valid data is stored in an Excel file, while invalid data is flagged and moved to the "unsuccessfully scraped" folder.

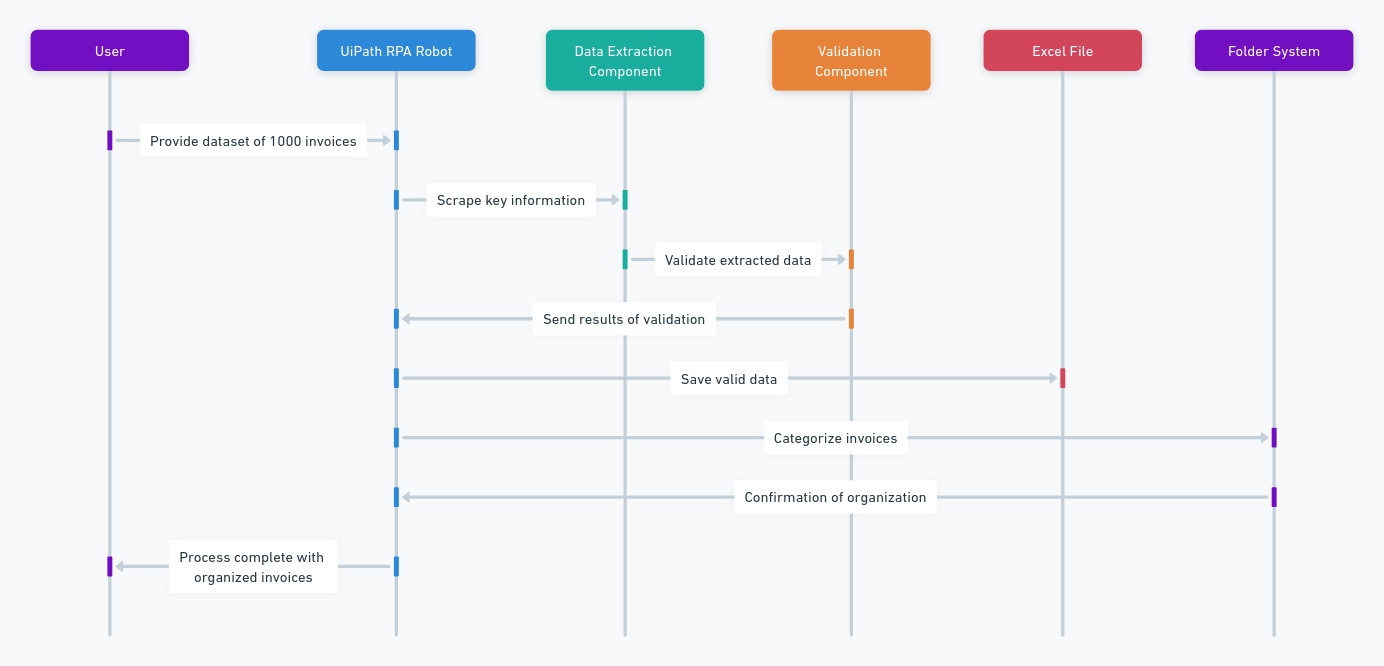


Fig 3.3 Sequence Diagram

# CHAPTER 4 PROJECT DESCRIPTION

This project uses UiPath RPA to automate the extraction of invoice data from 1000 invoices, storing the information in an Excel file and categorizing invoices based on successful or failed extraction. It reduces processing time from days to 130 minutes, improving accuracy and efficiency while eliminating manual data entry and lowering costs.

## MODULES:

* + 1. **INPUT HANDLING AND INITIALIZATION:**

### Folder Selection:

* + - * + Receive user input for the parent folder path containing the invoices.
        + Allow the user to select the folder where the invoices are stored.

### Subfolder Selection:

* + - * + List subfolders within the selected parent folder.
        + Allow the user to select the target subfolder containing the invoices for extraction..

### 4.1.1.3 Excel Report Generation:

* Dynamically create an Excel report (e.g., "invoices\_data.xlsx") within the chosen subfolder to store extracted invoice data.
* The Excel report will include fields such as Company Name, Invoice Number, Date, Customer ID, Payment Terms, and Total Amount.

## CONTENT ANALYSIS:

### AI Detection:

* + - * + Iterate through each Word document in the selected subfolder.
        + Employ advanced algorithms to detect AI-generated content.

### Plagiarism Check:

* + - * + Interface with external plagiarism detection services.
        + Conduct plagiarism checks on the assignment content.

## RESULT MANAGEMENT:

### Result Storage:

* + - * + Systematically update the Excel report with the extracted invoice data.

### Real-time Update:

* + - * + Display real-time updates of the integrity verification process.

## COMPLETION AND REPORTING:

### Plagiarism Check:

* + - * + Interface with external plagiarism detection services.
        + Conduct plagiarism checks on the assignment content.

## RESULT MANAGEMENT:

### Result Storage:

* + - * + Systematically update the Excel report with the extracted invoice data.
        + Store the processed data in a structured format to make it easy for users to view and manipulate the data later.

### Real-time Update:

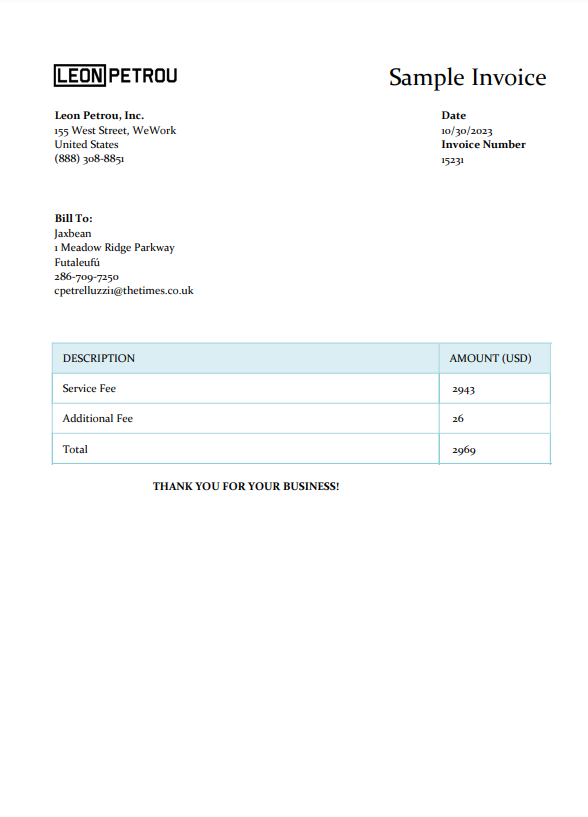
* + - * + Display real-time updates of the integrity verification process.

## COMPLETION AND REPORTING:

### Completion Message:

* + - * + Conclude the operation with a message indicating the successful completion of the invoice data extraction process.
        + Display details on the number of invoices processed, and highlight any errors encountered during extraction.

# CHAPTER 5 OUTPUT SCREENSHOTS





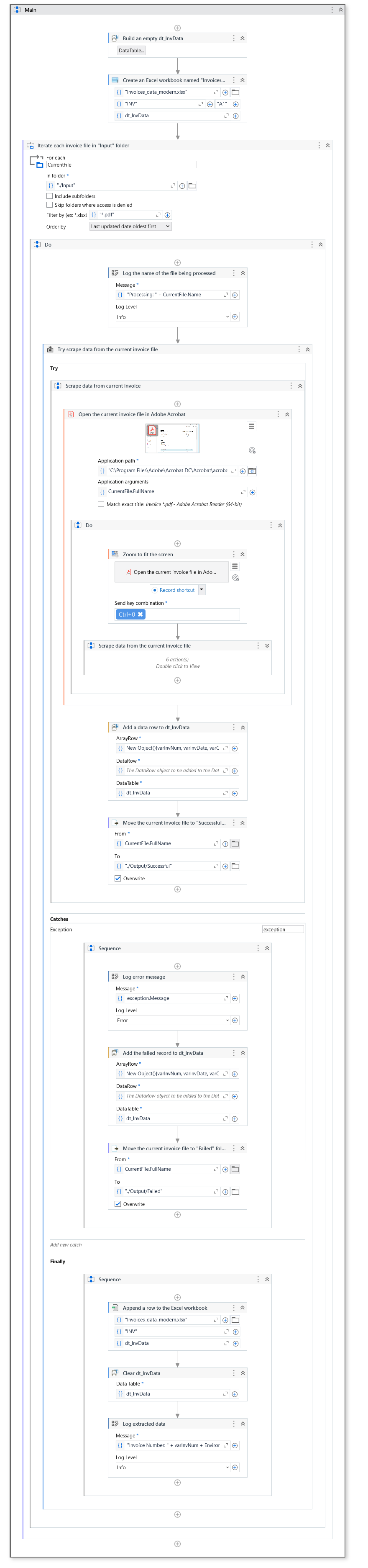


# CHAPTER 6 CONCLUSION

In conclusion, this project successfully demonstrates the significant benefits of leveraging Robotic Process Automation (RPA) through the UiPath platform to enhance the efficiency and accuracy of invoice data extraction. The automation process not only reduces the time required to process a large batch of invoices from several days to just 130 minutes, but it also ensures near-perfect accuracy by minimizing the risk of human error. By storing the extracted data in an Excel file and categorizing invoices into “successfully scraped” and “unsuccessfully scraped” folders, the solution offers seamless data organization and easy identification of errors for prompt resolution.

This project highlights the potential of RPA to transform traditional, labor-intensive processes, allowing businesses to significantly reduce operational costs while improving productivity. It ensures scalability, enabling businesses to handle increasing volumes of invoices with ease. Furthermore, the automated workflow frees up valuable human resources, allowing employees to focus on more strategic tasks. Overall, the solution not only optimizes invoice management but also demonstrates how RPA can be applied to other data-intensive processes, driving efficiency, reducing costs, and enabling businesses to scale their operations effectively. This project sets the foundation for more advanced automation applications in diverse business sectors, making it an essential step toward digital transformation.

# APPENDIX PROCESS WORK FLOW



# EFERENCES

1. Kuppusamy, Palanivel & Joseph K, Suresh. (2020). [Robotic Process](https://ijcrt.org/papers/IJCRT2006516.pdf) [Automation to Smart Education](https://ijcrt.org/papers/IJCRT2006516.pdf). 3775.
2. Patil, Dr & Mane, Vinod & Patil, Dr. (2019). [Social Innovation in](https://www.ijitee.org/wp-content/uploads/papers/v8i11/K21480981119.pdf) [Education System by using Robotic Proces](https://www.ijitee.org/wp-content/uploads/papers/v8i11/K21480981119.pdf)s Automation (Rpa). International Journal of Innovative Technology and Exploring Engineering. 8. 3757-3760. 10.35940/ijitee.K2148.0981119.
3. Elkhatat, A.M., Elsaid, K. & Almeer, S. [Evaluating the efficacy of](https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00140-5) [AI content detection tools in differentiating between human and AI-](https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00140-5) [generated text](https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00140-5). *Int J Educ Integr* **19**, 17 (2023). <https://doi.org/10.1007/s40979-023-00140-5>

[4] H. Alamleh, A. A. S. AlQahtani and A. ElSaid, "[Distinguishing](https://ieeexplore.ieee.org/document/10137767/) [Human-Written and ChatGPT-Generated Text Using Machine](https://ieeexplore.ieee.org/document/10137767/) Learning," 2023 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2023, pp. 154-158, doi: 10.1109/SIEDS58326.2023.10137767.

[5] Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. [Academic Plagiarism Detection: A Systematic Literature Review](https://dl.acm.org/doi/fullHtml/10.1145/3345317). ACM Comput. Surv. 52, 6, Article 112 (November 2020), 42 pages. <https://doi.org/10.1145/3345317>

[6] H. A. Chowdhury, D. K. Bhattacharyya, “[Plagiarism: Taxonomy,](https://arxiv.org/pdf/1801.06323.pdf) [Tools and Detection Techniques](https://arxiv.org/pdf/1801.06323.pdf)”, 19th National Convention on Knowledge, Library and Information Networking, 2018.