Invoice Automation and Processing System

Gayathri Devi T[1]**, Sheela O**[2]

[1,2]Assistant Professor, Department of Information Technology, MNM Jain Engineering College,
Chennai, India

Abstract:

This project presents an Automated Invoice Processing System developed to simplify, automate, and streamline invoice handling in a digital business environment. The system is capable of accepting invoice files in various formats, including image, PDF, and text documents. By leveraging Optical Character Recognition (OCR) and custom Python scripts, the system extracts vital information from these documents, such as invoice number, invoice date, purchase order number, vendor name, and total amount. Once extracted, the data is stored in a structured SQL database, allowing for easy access, tracking, and further analysis. To support scalability and ensure secure storage, original invoice files are also archived in cloud storage. The system integrates with Microsoft Outlook to automatically fetch emails containing invoice attachments, thereby reducing manual intervention and ensuring a continuous flow of invoice data into the system. A user-friendly React.js dashboard provides real-time access to stored invoice records and enables users to filter, search, and verify data based on specific criteria such as customer name, date, and status. This system reduces the need for manual data entry, minimizes the chances of errors, and enhances operational efficiency by automating repetitive tasks. It serves as a robust solution for organizations seeking a reliable and centralized platform to manage their invoice processing activities with improved accuracy, speed, and transparency.

Keywords: Optical Character Recognition (OCR), Invoice Automation, ERP Integration (Enterprise Resource Planning), Payment Processing.

I INTRODUCTION

1.1 OVERVIEW OF THE AUTOMATED INVOICE SYSTEM

The Automatic Invoice Processing System is designed to streamline and automate the handling of incoming invoices received via email. The system extracts invoice attachments (PDF, image, or text) from Outlook mails, processes them using Optical Character Recognition (OCR), and identifies key invoice details such as purchase order number, customer name, date, and email. Each mail is then classified and moved into designated folders (Processed, Processing, Unknown Customer, Unprocessed) based on its status. Extracted data is stored in a structured format, logged in a tracker table, and saved in a SQL database. Finally, a user-friendly dashboard allows stakeholders to view and track the status of invoices using filters such as customer name or date. This system significantly reduces manual effort, increases accuracy, and enhances the efficiency of financial document handling in an enterprise setting.

1.2 PROBLEM STATEMENT

Manual invoice processing is time-consuming, error-prone, and inefficient, especially when dealing with a large volume of emails and diverse invoice formats. Organizations often struggle with organizing, extracting, and validating invoice data, leading to delayed payments, missed records, and resource overhead. The lack of automation in email handling, attachment classification, and data extraction creates bottlenecks in financial workflows. There is a need for a robust, automated system that can process incoming emails with multiple attachments, extract relevant invoice data regardless of file type, and update a centralized database for easy tracking and monitoring. This project addresses these challenges by developing an automation invoice processing system that improves accuracy, saves time, and enhances overall productivity.

1.3 OBJECTIVE

The primary objective of this project is to develop an Automated Invoice Processing System that streamlines the end-to-end handling of invoice emails and their attachments. The system is designed to automatically fetch and read emails from Outlook that contain invoice attachments, classify them based on processing status, and organize them into appropriate folders. It uses OCR and file parsing techniques to extract text from various attachment formats, including PDFs, images, and plain text files. Key invoice details such as purchase order number, customer name, date, and email address are identified and extracted, then stored in a structured JSON format and updated into a SQL database for tracking purposes. Additionally, a tracker table is maintained to monitor the status of each email throughout the process. A user-friendly dashboard is also provided to allow filtering and reviewing of invoices based on specific criteria such as customer name, date, and status. This system is aimed at minimizing manual effort, reducing errors, and enhancing overall efficiency in invoice data processing.

II LITERATURE SURVEY

Akanksh Aparna Manjunath [1] developed an automated system for extracting data from invoices using a combination of image processing and OCR techniques. The system begins with image preprocessing using OpenCV, where steps like binarization, noise removal, and resizing enhance the clarity of invoice images. A template-based approach is then applied to detect tables and important layout components, ensuring accurate identification of structured fields. To extract text, a custom-trained Tesseract OCR engine is used, optimized specifically for invoice documents, improving recognition accuracy across various fonts and formats. Additionally, the system includes a web-based annotation tool that allows users to manually label fields such as invoice number, date, and total amount. These annotations help refine the system’s performance and adaptability to different layouts. Tested on more than 25 invoice formats, the system demonstrated an accuracy range between 85% and 95%, proving its effectiveness in handling diverse invoice structures and offering a reliable, end-to-end solution for automating invoice data extraction.

A. Antony Jenifer [2] proposed an intelligent approach for automating invoice data extraction using Natural Language Processing (NLP) techniques. The system starts by converting invoice images into machine-readable text through Optical Character Recognition (OCR). Once the raw text is extracted, NLP-based methods are applied to identify and extract critical fields such as the invoice number, date, total amount, and vendor details. This is achieved through multiple stages, beginning with text preprocessing—where the extracted content is cleaned, normalized, and prepared for further analysis. The system then uses a combination of pattern matching and rule-based logic to accurately recognize relevant fields within the unstructured text. After successful identification, the extracted data is structured and mapped into a predefined format suitable for storage or downstream processing. This method significantly reduces manual data entry, speeds up processing time, and increases the accuracy of handling invoice information. The proposed solution highlights the practical benefits of intelligent document automation and closely aligns with the goals of the current project, particularly in enhancing efficiency and minimizing human intervention in repetitive tasks.

Organizing Committee [3] introduced a competition aimed at promoting advancements in Optical Character Recognition and information extraction from scanned receipts and invoices. The study focuses on real-world challenges such as poor image quality, varying document layouts, and unstructured formats that complicate accurate data extraction. To support these research efforts, the competition provided large-scale annotated datasets and defined two main tasks: receipt OCR and key information extraction. This work is relevant to the present project as it emphasizes the development of reliable OCR systems capable of handling diverse and complex invoice formats for structured data retrieval.

Xinyu Zhou et al. [4] introduced a deep learning-based model for robust text detection in natural scene images, utilizing a single-stage, fully convolutional neural network. Unlike traditional multi-step pipelines that rely on separate components for region proposal and classification, this model streamlines the process by directly predicting text bounding boxes with the aid of integrated non-maximum suppression. The architecture is designed to handle complex scenarios, enabling the detection of text in various orientations, curved shapes, and irregular layouts. This makes the approach particularly effective for unstructured and noisy document types. In the context of the present project, this model can serve as a valuable preprocessing step for identifying text regions within invoice images before OCR is applied. By accurately localizing text areas, it enhances both the precision and reliability of the overall data extraction pipeline, especially when dealing with invoices that vary widely in design and structure.

Krieger, F., Drews, P., and Funk, B. [5] conducted research focused on extracting structured data from unstructured invoice documents using machine learning techniques. Their work involved evaluating multiple models, including Chargrid, Random Forest, and LayoutLM, to determine their effectiveness in parsing varied invoice formats. Among these, LayoutLM showed the highest accuracy, primarily because of its ability to capture and preserve the spatial layout and semantic relationships within documents. This model demonstrated strong generalization capabilities, making it particularly effective in handling unfamiliar or non-standard invoice templates. The study offers valuable insights into developing adaptive and intelligent invoice parsers that can manage diverse document structures. These contributions are highly relevant to the current project, as they support the goal of minimizing manual intervention and improving the efficiency and scalability of invoice data extraction systems.

III EXISTING SYSTEM

In the current landscape of invoice processing, most organizations follow a manual or partially automated approach. Invoices are commonly received through emails or uploaded into shared folders by vendors. Employees are then responsible for opening each email, downloading attachments, reviewing the content, and manually extracting critical details such as the invoice number, purchase order (PO) number, invoice date, customer or vendor name, and total amount.

These extracted details are typically entered into spreadsheets, Excel trackers. In certain modern environments, semi-automated tools are used to scan PDF or image files and extract text using basic Optical Character Recognition (OCR) technology. However, these tools usually operate independently and are not capable of managing the entire workflow automatically.

IV PROPOSED SYSTEM

The proposed Automated Invoice Processing System aims to overcome the limitations of existing manual and semi-automated workflows by introducing a fully integrated and intelligent solution. This system is designed to automatically handle invoice emails and attachments, extract essential data, and store it in a structured and centralized format with minimal human intervention.

The system connects directly to an email platform to fetch invoice emails and identify attachments such as PDFs, images, or text files. Advanced OCR and parsing techniques are applied to extract key details including invoice number, PO number, date, vendor name, customer name, and total amount. Extracted information is formatted in JSON and stored in a structured SQL database to ensure easy access, tracking, and reporting.

A tracker mechanism is maintained to monitor the processing status of each invoice email. Additionally, the system automatically classifies and moves emails into appropriate folders based on their status (e.g., processed, error, pending). A user- friendly dashboard built with React.js allows users to filter and review invoice data based on fields like date, customer, or status. By automating the entire end-to-end workflow — from email handling and data extraction to storage and interface visualization — the proposed system significantly reduces manual effort, improves accuracy, enhances real-time monitoring, and increases overall operational efficiency.

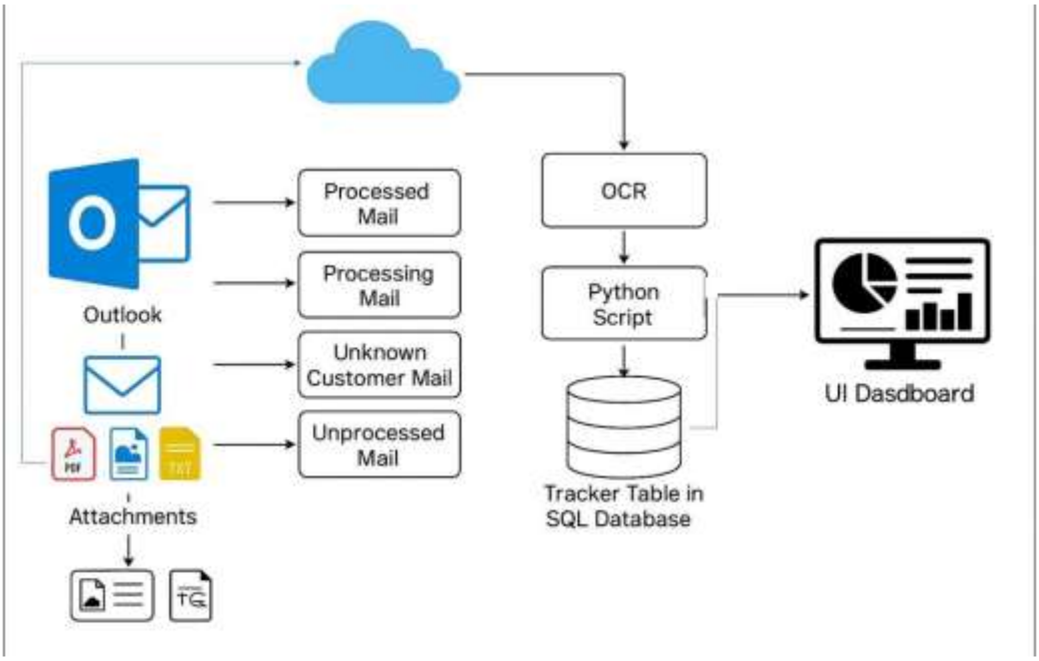

Fig. 1. Overall System Architecture

Email Input:
The system starts by fetching invoice emails from Microsoft Outlook, which may include attachments in PDF, image, or text formats.

Attachment Classification:
Emails are automatically sorted into folders such as Processed Mail, Processing Mail, Unknown Customer Mail, and Unprocessed Mail based on their processing status.

OCR and Processing:
Attachments are sent through an OCR engine to extract text. Python scripts then identify and extract key invoice fields like invoice number, date, PO number, and amount.

Database Storage:
Extracted data is saved into a tracker table in an SQL database for structured access and tracking.

Dashboard Interface:
A UI dashboard displays real-time invoice data, enabling users to filter, monitor progress, and ensure transparency.

V CONCLUSION & FUTURE ENHANCEMENT

Conclusion

The automatic email attachment processing system effectively streamlines the handling of incoming emails by automating the extraction and classification of important data. By integrating technologies such as Outlook for email retrieval, OCR for image-to-text conversion, and Python for data processing, the system reduces manual effort and improves accuracy in data handling.

Throughout the development process, a strong focus was placed on modularity, making each component independently testable and maintainable. Attachments are securely stored in the cloud, and essential information such as PO numbers, customer names, and dates are extracted and saved in a structured SQL database. This ensures that businesses can track and retrieve critical information quickly and efficiently.

One of the key highlights of the project is the dashboard, which provides real-time visibility into email processing statuses. With search and filter options by customer name or date, users can easily identify the state of each email—processed, unprocessed, unknown, or in progress. This not only improves transparency but also enhances user interaction and decision-making.

In conclusion, the system serves as a reliable solution for automating the tedious task of email and attachment management. It boosts operational efficiency, reduces human error, and lays a solid foundation for future enhancements such as advanced analytics, customer insights, or AI-driven pattern recognition.

Future Enhancement

In the future, the system can be enhanced by integrating machine learning algorithms to improve pattern recognition and data extraction accuracy. Currently, patterns are stored manually in JSON format, but a trained model could learn and adapt to new customer formats dynamically, reducing the need for manual updates. Additionally, incorporating natural language processing (NLP) could help the system understand context from email bodies, allowing it to extract more nuanced information such as intent, urgency, or action items.

Another significant enhancement would be to implement role-based access control and advanced analytics in the dashboard. This would allow different users (e.g., admins, data analysts, team leads) to access specific data views and gain insights through charts, reports, and trends over time. Integration with other business tools like ERP or CRM systems could further automate workflows, making the system a key component in enterprise-level document and communication management.

REFERENCES

1. A. A. Manjunath, “Automated Invoice Data Extraction Using Image Processing,” International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 514–521, 2023.
2. A. Jenifer, “Automatic Invoice Processing using NLP,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 10, no. 5, pp. 18–24, 2021.
3. ICDAR Organizing Committee, “ICDAR Robust Reading Challenge on Scanned Receipts OCR and Information Extraction,” Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pp. 254–263, 2019.
4. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “EAST: An Efficient and Accurate

Scene Text Detector,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2642– 2651, 2017.

5. F. Krieger, P. Drews, and B. Funk, “Automated Invoice Processing: Machine Learning-Based Information Extraction for Long Tail Suppliers,” Journal of Information Technology Management, vol. 33, no. 4, pp. 62–75, 2022.
6. A. Kumar, S. N. Omkar, and N. Kumar, “Automatic Extraction of Key Fields from Indian Financial Documents using Template Matching Technique,” Indian Journal of Computer Science and Engineering, vol. 7, no. 6, pp. 187– 191, 2017.
7. R. Annaswamy, A. Balaji, S. Singh, H. Mishra, and H. Pant, “Efficient Automated Processing of the Unstructured,” International Journal of Computer Applications, vol. 176, no. 6, pp. 10–15, 2020..
8. K. Karthick, K. B. Ravindrakumar, R. Francis, and S. Ilankannan, “Automatic Invoice Data Extraction and Entry using Deep Learning Models,”International Research Journal of Engineering and Technology (IRJET), vol. 9, no. 4, pp. 623–627, 2022.
9. D. Baviskar, S. Ahirrao, and K. Kotecha, “Multi-layout Unstructured Invoice Documents Dataset,” Data, vol. 6, no. 2, p. 25, 2021.
10. S. Long, X. He, and C. Yao, “Scene Text Detection and Recognition: The Deep Learning Era,” Foundations and Trends in Computer Graphics and Vision, vol. 8, no. 4, pp. 231–345, 2018.