

Business Documentation Classifier and Analyzer

Dr. Gayatri Hegde¹, Dr. Nusrat Parveen², Soham Bait³, Kamlesh Kale⁴, Aditya Kushwaha⁵, Anuj Karambalkar⁶

^{1,2,3,4,5,6}Computer Science & Business System, Department of Engineering & Technology, Bharati Vidyapeeth Deemed University Navi Mumbai, India

Abstract—The Business Documentation Classifier and Analyzer powered by AI is an intelligent automation solution aimed at the classification, extraction and analysis of business documentations and financial information contained within various documents like invoice, receipt, report, etc. This document classification uses a combination of OCR i.e., Tesseract, AWS Textract, ML and NLP i.e., LayoutLMv3, spaCy, Regex, to convert unstructured document data into usable insights. After extracting this information, it can be stored in a hybrid database model of MongoDB. The dashboard is built using React.js and Tailwind CSS, and visualization of financial analytics is done using Plotly and Chart.js. The backend is built using Node.js, Express.js and Flask to manage all APIs, ML pipelines and logic, and uses Gemini API for intelligent summarization. Finally, this solution is deployed on Vercel and AWS, using CI/CD pipelines to support scalability, reliability and rapid data-driven financial decision-making for the modern business.

Index Terms—OCR, Tesseract, AWS Textract, Machine Learning (ML), Natural Language Processing (NLP), LayoutLMv3, spaCy, Regex, data extraction, Supabase, SQL, a MongoDB hybrid database, React.js, Tailwind CSS, Plotly, Chart.js, Node.js, Express.js, Flask, Gemini API, summary/document aggregation, and implementation via Vercel using AWS infrastructure and CI/CD pipelines.

I. INTRODUCTION

Organizations today are faced with a massive number of financial documents; for example, they deal with many invoices, receipts, purchase orders as well as reports. The process to manage and analyze that information using manual methods is labor intensive, error-prone and inefficient which delays the ability to provide timely financial reports, and also causes delays in the decision-making process of an organization. The traditional method of entering data and verifying manually does not scale well or provide

the accurate information that is needed to ensure organizations have real-time visibility into all financial activities.

AI-Powered Business Document Classifier & Analyzer utilizes AI, ML and NLP technologies to assist businesses with classifying documents and extracting financial information.

Using Optical Character Recognition (OCR), including Tesseract OCR and AWS Textract, the system processes images of both scanned and digitally stored documents to read the text and output either structured or unstructured data. The layout of the documents is analyzed by LayoutLMv3 and spaCy in order to appropriately identify & extract the following keys: invoice number, invoice date, vendor name and total amount, making it possible to convert unstructured text into structured actionable financial data.

The resulting structured and unstructured data is presented in a real-time interactive dashboard created with the following technology stack: React.js, Tailwind CSS and Chart.js. The dashboard provides users with dynamic views of financial trend analysis; financial comparisons; and, key performance metrics. The back-end system architecture forms an entire ecosystem for processing and analyzing data as well as APIs for moving data from one system to another, built on a technology stack that contains Node.js, Express, and Flask (Python). This hybrid database architecture allows for the maximum availability and scalability of our businesses and end-users, utilizing the technology of MongoDB to achieve this. The system also has built-in capabilities that allow it to detect anomalies and predict financial trends through the integration of AI-drive analytics using the Gemini APIs.

In summary, this system will provide an automated, intelligent, and reliable way to manage financial

documents, reducing the amount of human labor required, the number of errors made, and enabling more intelligent, data-driven decisions in business through its automated workflow capabilities.

II. LITERATURE REVIEW

Recent research highlights the growing role of Artificial Intelligence in automating document processing and analysis. Studies such as AI Document Analysis: How AI Transforms Document Processing [1] and Automated Document Analysis in Qualitative Research [2] demonstrate how Machine Learning (ML) and Natural Language Processing (NLP) enhance scalability, cost reduction, and efficiency in document handling, though they still face issues like bias and limited interpretative flexibility. Advanced approaches like Exploring AI-driven Approaches for Unstructured Document Analysis [3] and The Smart Document Processing with Artificial Intelligence [4] emphasize hybrid AI models for improved classification and automation, despite challenges in data quality and complexity. Targeted solutions such as OCRMiner [5] improve OCR accuracy for invoice analysis, while Healthcare Document Analysis [6] applies structured data methods for efficient information extraction, albeit with ethical concerns. Foundational research like Qualitative Document Analysis [7] supports accessible and practical methods, though limited by human selectivity. Innovations like Donut — OCR-free Document Understanding Transformer [8] introduce Vision Transformers for faster OCR-free processing, and Document AI: Benchmarks, Models and Applications [9] standardize AI-driven document reading using deep learning, despite OCR-related limitations. Dataset contributions such as MIDD [10] enhance named entity recognition with layout diversity, while broader reviews like Deep Learning for Anomaly Detection [11] provide insights into cost-effective AI detection methods, reinforcing AI's transformative impact on document analysis and automation.

III. PROBLEM STATEMENT

In today's business world companies deal with a number of documentations & financial papers like invoices, receipts, purchase orders and reports every day. Doing this manually takes a lot of time is work

and people make mistakes often leading to wrong data entry, slow approvals and incorrect financial information. Also, financial data comes in formats. Scanned PDFs, images and text that isn't organized. Making it hard to get, sort and analyse properly.

Now there isn't a single automated system that can smartly sort, get and analyse financial information from different types of papers accurately. This limitation stops businesses from using real-time information to make decisions. Improve how they work.

Here are the main goals:

1. Automatically sort papers like invoices and receipts using AI.
2. Accurately get details like invoice numbers and dates from papers using advanced technology.
3. Show trends and insights in easy-to-understand dashboards.
4. Reduce mistakes and work by automating document handling and data entry.
5. Help organizations make accurate and informed financial decisions.
6. Improve workflow speed, accuracy and scalability with automation and analytics.

The proposed AI-Powered Business Documentation Classifier and Analyzer aim to solve these problems by offering an automated and smart solution, for managing financial documents today. The AI-Powered Business Documentation Classifier and Analyzer will help businesses classify, extract and analyse data from documents thereby enabling data-driven decision making and operational efficiency. This solution will provide financial insights, automate document classification and reduce errors and effort in document handling and data entry processes.

IV. EXISTING SYSTEM

There are some existing platforms like Oculous and Google Document AI that can automatically analyse documents. However, they have some problems that make them not very useful for looking at financial information in a detailed way and making pictures with the data.

Oculous is a platform that companies pay to use. It is mainly for financial companies. It can look at things like invoices and bank statements. Get information

from them using artificial intelligence and special computer programs that can read text from pictures. It is good at checking if the information is correct. It is like a box that you cannot see inside so you do not know how it is making its decisions. It is mostly just looking at the information. Saying if it is right or wrong it does not really help you understand what the information means or show you any trends. Also, it is not very flexible. It is hard to use it for things that it was not specifically made for. And it is pretty expensive small companies or schools might not be able to afford it.

Google Document AI is different it is a platform that runs on the internet and uses computer programs to understand what is in documents like invoices and receipts. It is really good at getting the text out of these documents. It does not really help you understand what the information means in terms of money. You can teach it to recognize types of information but it is not very good, at making pictures or comparing information. Usually, you have to take the information it gives you and use another program to understand it. Because it only runs on the internet some people might be worried that their information is not safe and you have to use Google's computers to run it.

V. DEVELOPED SYSTEM

The AI-Powered Business Documentation Classifier and Analyzer use intelligence, OCR and data visualization to analyze business and financial documents. It automates the process from classifying documents to extracting data storing it and doing, in-depth analysis. This system relies on automation and data to make business decisions. The architecture of the system is based on four pillars. These pillars are accuracy, adaptability and giving businesses financial insights. The AI-Powered Business Documentation Classifier and Analyzer have automation. It provides data-driven decision-making. The system ensures accuracy and adaptability. It provides financial intelligence.

A. Document Classification

The classification feature helps the system figure out what kind of documents it is looking at, like invoices, receipts, balance sheets, payment slips and financial reports. It uses tools to recognize patterns and

understand what the words mean so it can automatically tell what type of document it is and what is in it.

- Purpose:

The system does this so people do not have to sort and categorize documents by hand.

- Functionality:

The system looks at how the document's laid out and what the words say so it can put each file into the right group, like "Invoice" "Receipt" or "Bank Statement".

- Impact:

This makes things go faster makes sure the right information gets to the people and helps businesses deal with a lot of different financial documents.

- Distinctive

Aspect: This system is different from others because it does not just look at tags or templates it actually understands what the words mean so it can even handle documents it has never seen before or ones that are not, like the others.

B. Differentiated Storage

Financial documents have both structured and unstructured data. The system uses a way to store data. This way helps organize the information into the database formats for easy access and analysis.

- Purpose:

The goal is to handle types of data well. It also needs to be able to grow and be reliable.

- Functionality:

The system stores structured data like invoice IDs transaction dates and total amounts in tables. This makes it easy to query and report. Unstructured data like notes, remarks or extra information is stored in a document database.

- Impact:

This way of storing data makes it faster to find information. It also helps with analysis. No important information whether structured or not is lost during processing.

- **Distinctive Aspect:**

The systems storage layer helps combine table and document data. This ensures that data is synced in time for analysis and dashboard visualization.

The system handles documents well by managing data diversity. It stores structured data like invoice IDs and transaction dates. Unstructured data, like notes are also stored. This helps with analysis and data retrieval. The system is scalable and reliable.

C. Information Extraction

The system is about extraction, which is really the main part of it. This is where the system takes information from financial documents and turns it into data that a machine can read and understand. The main goal of this feature is to be precise, consistent and aware of the context.

- The purpose of this is to get the financial details from different kinds of documents no matter what they look like.
- The system does this by pulling out information like the invoice number, the date the vendors details, what was bought, taxes and the total amount owed. It does more than just recognize text. It also understands what certain words mean, like "subtotal" and "due date" so it can figure out what the document is saying.
- This makes a difference because it means we get clean and accurate financial data that is ready to be looked at and we do not have to fix it by hand as much.
- What is really special about this extraction process is that it does not just look at the text. It also tries to understand what each piece of information means and how it fits into the picture. It can handle documents, in languages and it can even deal with documents that are not laid out in a standard way or are not very clear.

D. Financial Analysis and Visualization

When we get the data and put it in order the systems analysis feature turns it into information that we can actually use. This helps people see what is happening with their money track how they are spending it and understand how things are going and if there are any problems.

- **Purpose:**

To help people make financial decisions based on the facts.

- **Functionality:**

The analysis part of the system takes all the data we have and compares it makes pictures like bars and lines to help us see what is going on and finds things that do not seem right like expenses or things that do not add up. It also helps us make sense of reports by summarizing the parts.

- **Impact:**

Companies can use this system to keep an eye on how they're doing financially find mistakes and predict what might happen next. All from one place where they can interact with the information.

- **Distinctive Aspect:**

This system is different from tools that just give us a static summary. It gives us pictures that change comparisons, across time periods and explanations of what it all means in simple language so financial intelligence and financial decision-making and financial health are all easy to understand and use.

VI. SYSTEM ARCHITECTURE

When we get the data and put it in order the systems analysis part turns it into information that we can actually use. This helps people see what is happening with their money track how they are spending it and understand how things are going and if there are any problems.

- The main goal is to help people make financial decisions based on the facts.
- What it does is it takes all the data we have and compares it makes pictures like bars and lines to help us see what is going on and finds things that do not seem like weird expenses or things that do not add up. It also helps us make sense of the reports we get.
- What it means for businesses is that they can check how they are doing with money away find mistakes and guess what might happen next. All on a screen that is easy to use.
- What makes this system special is that it does not just give us pictures of what is going on with our

money. Instead, it gives us pictures that change comparisons that show us how things are different from one time, to another and explanations that are easy to understand. All of this helps us see what is going on with our financial data.

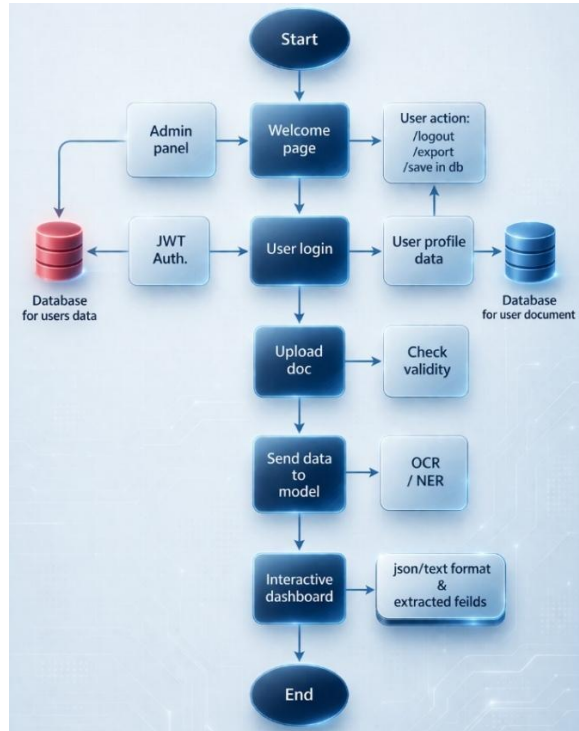


Fig. 1. System Diagram

VII. TECHNIQUES AND TECH STACK

A. Techniques

The proposed system follows a structured pipeline for processing and analyzing business documents. Initially, text is extracted from uploaded documents using OCR techniques, such as Tesseract OCR or AWS Textract, enabling the conversion of scanned or image-based content into machine-readable form. The extracted text is then refined through preprocessing steps, where noise, unnecessary spaces, and irrelevant characters are removed to improve data quality.

Subsequently, natural language processing techniques are applied to interpret and organize the content. Models such as LayoutLMv3, along with the spaCy library, are utilized to classify document types—including invoices, receipts, and reports—and to identify essential fields such as invoice numbers, dates, and total amounts. Following this, rule-based methods using regular expressions and Python scripts

are employed to validate the extracted information, ensuring accuracy and consistency in formatting.

The processed data is then stored in suitable database system of MongoDB, enabling real-time synchronization. For user interaction and interpretation, the system presents the extracted information through a web-based dashboard developed using React.js, Tailwind CSS, Plotly, and Chart.js. Finally, analytical features are incorporated to generate meaningful insights and comparisons, supporting informed and data-driven decision-making for business operations.

B. Tech Stack

The implementation of the proposed system is supported by a combination of frontend, backend, and data processing technologies. The user interface is developed using React.js, which enables the creation of a dynamic and modular design, while Tailwind CSS is used to achieve responsive and consistent styling across different devices. For effective data representation, visualization libraries such as Plotly and Chart.js are incorporated to generate interactive charts and dashboards.

On the server side, Node.js and Express.js are employed to manage routing, handle API requests, and execute core application logic. In parallel, Python with the Flask framework is utilized to integrate and execute artificial intelligence models, particularly for natural language processing tasks. Libraries such as Pandas, NumPy, and Matplotlib support data manipulation, analysis, and financial computations within the system.

To manage data efficiently, database solutions of MongoDB is used. Structured and unstructured data are stored using this platform. It has enabled real-time data updates and synchronization. For document processing and language understanding, tools including Tesseract OCR and AWS Textract are applied to extract textual content from documents, while LayoutLMv3, spaCy, and regular expression techniques are used for classification and entity extraction.

Furthermore, external APIs are integrated to enhance system capabilities. In particular, the Gemini API is utilized to support advanced document & text summarization, contributing to improved interpretation and analysis of financial documents.

VIII. OUTCOMES

The DOCUSENSE system helps businesses with four things:

- It sorts documents, like invoices and receipts into categories.
- It stores data in a way.

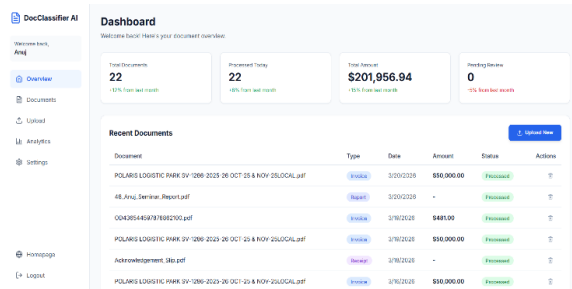


Fig. 2. Database Snippet

- It pulls out information in extracted form.

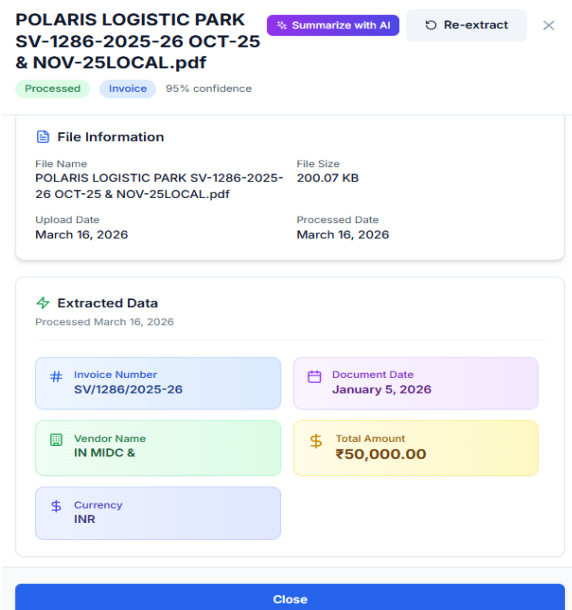


Fig. 3. Extract

- It analyzes financial data.

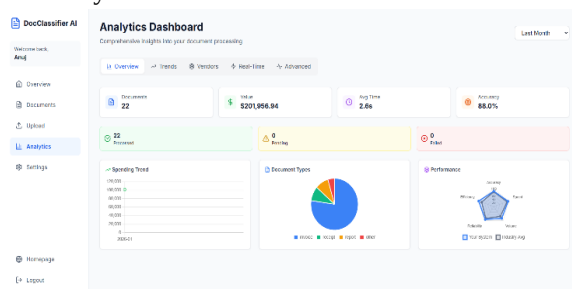


Fig. 4. Analyzer

The system uses machine learning and natural language processing to sort documents. It uses Optical Character Recognition and Regex to extract details from documents.

The extracted data is then shown on a dashboard. This dashboard provides insights. Helps with decision making. Using DOCUSENSE businesses can reduce work. It also helps improve accuracy and enables financial management. The DOCUSENSE system provides real-time insights. It helps businesses make decisions with data. The system stores data efficiently for access. It. Visualizes data. The DOCUSENSE system helps with analysis of financial documents as well.

Below are the output screenshots of the system:

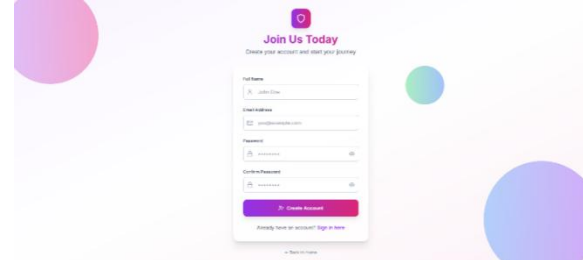


Fig. 5. User Registration

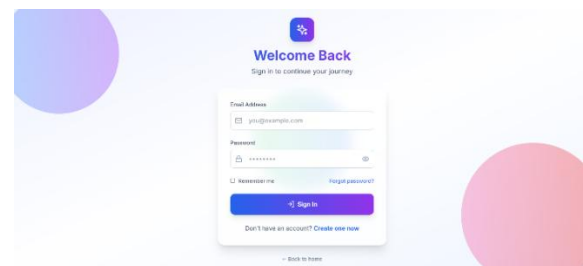


Fig. 6. User Sign In

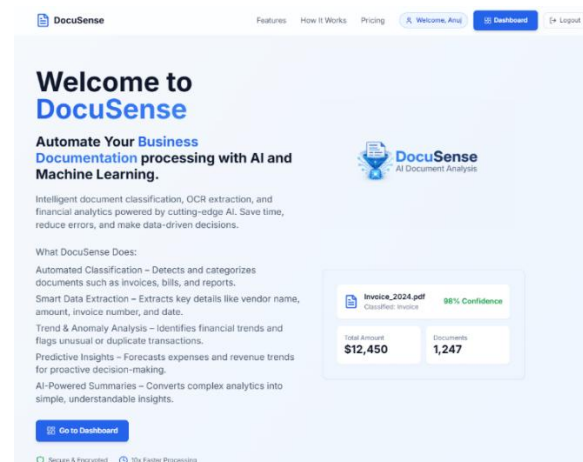


Fig. 7. User Dashboard

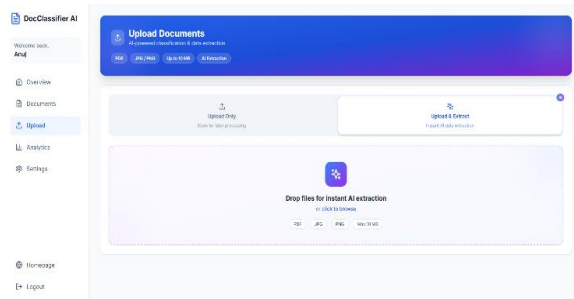


Fig. 8. Document Upload

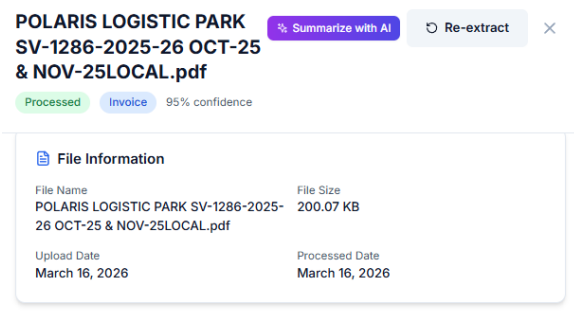


Fig. 9. Classify

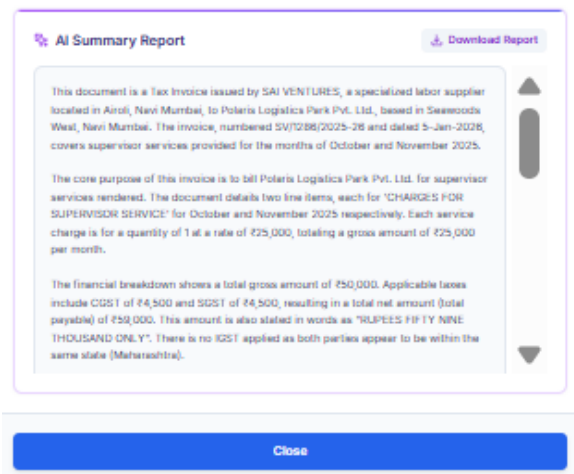


Fig. 10. Summarize

IX. OUR SYSTEM ADVANTAGE:

The system AI-Powered Business Documentation Classifier and Analyzer is really good because it can be changed to fit our needs. It is also easy to see what is going on. It can handle a lot of work. This system uses things like OCR from Tesseract and AWS Textract and NLP from spaCy and LayoutLMv3 and ML-based analytics. It does more than just get the information from documents. The system can also

classify documents make summaries look at trends make predictions and find things that are not normal.

The system has a dashboard that's easy to use and it shows us financial information in real time. This dashboard was built with React.js, Chart.js and Plotly

It can handle varied types of data and it supports a system that uses MongoDB.

Our system is different from others because it does not have to be connected to the internet to work. It can work offline. It gives us control over our data.

The AI-Powered Business Documentation Classifier and Analyzer use Gemini API to help it understand what it is reading. This makes it really smart. It can give us deep analysis, like Oculous and Google Document AI cannot.

X. CONCLUSION:

The Business Documentation Classifier and Analyzer demonstrate how artificial intelligence can improve the handling of varied business documents. By combining OCR, natural language processing, and machine learning, the system automates tasks such as document classification, data extraction, and financial analysis. This reduces manual effort, minimizes errors, and enables quicker, more reliable decision-making. Developed using React.js, Node.js, Flask, and AWS, the platform ensures efficient performance, security, and real-time processing capabilities.

Overall, the system offers a practical and user-friendly approach to store, classify, and summarize the variety of business documentations and perform financial document analysis. It allows organizations to move away from traditional, manual workflows and adopt data-driven processes. As a result, businesses can manage their documents more effectively while improving accuracy, speed, and operational efficiency.

REFERENCES

- [1] Haponik, "How AI is Revolutionizing Document Analysis: A Comprehensive Guide," *Addepto Blog*, 2025.
- [2] J. Doe and R. Lee, "Automated Document Analysis in Qualitative Research: Exploring the Potential and Challenges," *AI & Society Journal*, vol. 40, no. 2, pp. 155–170, 2025.

- [3] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven Approaches for Unstructured Document Analysis and Future Horizons," *Journal of Big Data*, vol. 11, no. 1, p. 59, 2024.
- [4] R. Gupta and N. Sharma, "The Smart Document Processing with Artificial Intelligence," *AI Research Digest*, vol. 12, no. 3, pp. 45–59, 2024.
- [5] D. Kim and S. Park, "Information Extraction from Scanned Invoice Images (OCRMiner)," *International Journal of Computer Applications*, vol. 975, no. 8887, pp. 45–53, 2022.
- [6] T. Moilanen, M. Sivonen, K. Hipp, H. Kallio, O. Papinaho, M. Stolt, R. Turjamaa, A. Häggman-Laitila, and M. Kangasniemi, "Developing a Feasible and Credible Method for Analyzing Healthcare Documents as Written Data," *SAGE Open Nursing*, vol. 8, 2022.
- [7] H. Morgan, "Conducting a Qualitative Document Analysis," *The Qualitative Report*, vol. 27, no. 1, pp. 64–77, 2022.
- [8] D. Kim, M. Cho, and J. Song, "Donut—OCR-free Document Understanding Transformer," *arXiv preprint arXiv:2111.15664*, 2021.
- [9] L. Cui, Y. Xu, T. Lv, and F. Wei, "Document AI: Benchmarks, Models and Applications," *arXiv preprint arXiv:2111.08609*, 2021.
- [10] D. P. Baviskar, S. Ahirrao, and K. Kotecha, "Multi-Layout Invoice Document Dataset (MIDD): A Dataset for Named Entity Recognition," *Data*, vol. 6, no. 7, p. 78, 2020.
- [11] L. Ruff, R. Vandermeulen, N. Görnitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep Learning for Anomaly Detection: A Survey," *arXiv preprint arXiv:1901.03407*, 2019.