A Major Project Synopsis on

AI-Powered Intelligent Document Processing Platform Using Generative AI

Submitted to Manipal University, Jaipur

Towards the partial fulfilment for the Award of the Degree of

MASTER OF COMPUTER APPLICATIONS

2023-2025

by

Ashutosh Tripathi

23FS20MCA00022

MANIPAL UNIVERSITY JAIPUR

Under the guidance of

Dr. Linesh Raja

Department of Computer Applications

School of AIML, IoT&IS, CCE, DS and Computer Applications

Faculty of Science, Technology and Architecture

Manipal University Jaipur

Jaipur, Rajasthan

2025

I. Introduction

In today's digital-first business environment, organizations face significant challenges in efficiently processing, classifying, and extracting valuable information from large volumes of unstructured documents. This project, undertaken as part of my internship at Celebal technologies, focuses on developing an Intelligent Document Processing Platform (IDPP) that leverages cutting-edge artificial intelligence, machine learning, and prompt engineering techniques to automate document classification and data extraction processes.

The Intelligent Document Processing Platform (IDPP) is a comprehensive solution designed to transform traditional document management systems by implementing AI-powered document classification, data extraction, and search capabilities. By utilizing advanced technologies like GPT models, Azure Document Intelligence, and vector-based search, the platform significantly reduces manual processing time while improving accuracy and consistency.

During my internship at Celebal, I worked on developing a scalable backend for intelligent document classification and data extraction, handling large volumes of PDFs, PNGs, and TIFs. I deployed the solution on Azure App Service, integrating GPT models and Azure Blob Storage for automated processing and structured storage. The project focused on prompt engineering for document classification, data extraction, and implementing confidence scoring for key-value pairs, resulting in a robust system capable of processing over 50 document types with high accuracy.

This project aims to consolidate my learning and contribute to building an efficient and reliable document processing system that can transform how organizations handle their document-intensive workflows.

Why Choose Our IDPP Solution?

At Celebal Client's IDPP project, we are revolutionizing document processing with AI-driven technologies. Our expertise in prompt engineering, data extraction, and efficient AI deployment ensures high-quality, accurate, and automated document processing.

- **Precision & Contextual Accuracy** – Vector-based search enhances document classification and retrieval.
- **Intelligent Data Extraction** – GPT and Document Intelligence deliver accurate key-value extraction.
- **Optimized Performance** – Multiprocessing implementation for efficient document processing.
- **Innovation & Customization** – Tailored document classification for 50+ document types.

Choose our solution to experience next-gen AI-powered document processing with the perfect blend of accuracy, efficiency, and scalability.

II. Problem Statement

Traditional document processing systems face significant challenges that limit their efficiency, accuracy, and scalability in real-world applications, particularly in the banking sector.

1. **Manual Document Classification Challenges**
    - High volume of documents requiring manual classification leads to operational inefficiencies.
    - Inconsistent classification results due to human error and subjective interpretation.
    - Significant time delays in processing critical customer documents.
2. **Data Extraction Limitations**
    - Manual data entry from documents is time-consuming and error-prone.
    - Unstructured documents make it difficult to extract key information in a standardized format.
    - Varying document formats and qualities complicate automated extraction efforts.
3. **Search and Retrieval Inefficiencies**
    - Traditional keyword-based search fails to understand context and semantic meaning.
    - Difficulty in finding specific information across large document repositories.
    - Limited ability to navigate to specific sections within documents containing relevant information.
4. **Document Management and Compliance Issues**
    - Challenges in maintaining proper document naming conventions and organization.
    - Difficulty in tracking document processing history and changes for compliance purposes.
    - Limited visibility into missing or incomplete customer documentation.

III. Project Objective

To address these issues, this project implements an Intelligent Document Processing Platform (IDPP) that utilizes AI-powered classification, data extraction, vector-based search, and a comprehensive document management system. The goal is to develop a scalable, efficient, and accurate platform that transforms document-intensive workflows in the banking sector.

IV. Methodology/ Planning of work

The project follows a structured approach, integrating GPT models, Azure Document Intelligence, and prompt engineering to develop a highly efficient Intelligent Document Processing Platform. The workflow is divided into multiple phases as outlined below:

Phase 1: Research & Requirement Analysis

- Understanding the limitations of existing document processing systems.
- Analyzing 54 document types for classification and data extraction requirements.
- Identifying optimal prompt engineering techniques to improve AI-generated outputs.
- Establishing accuracy benchmarks and performance metrics for the system.

Phase 2: Data Collection & Preprocessing

- Collecting sample CIF documents for testing and development.
- Preprocessing documents for classification system implementation.
- Establishing document quality standards and handling procedures.
- Creating test cases for different document types and formats.

Phase 3: Model Integration & Development

- Implementing prompt engineering for document classification and data extraction.
- Developing multiprocessing for form recognizer and OpenAI calls for improved efficiency.
- Creating PDF chunking methods to optimize thread utility.
- Building a robust backend for processing large volumes of documents.
- Implementing automated storage in Azure Blob Storage with consistent naming conventions.

Phase 4: Testing & Performance Evaluation

- Evaluating classification accuracy across all 54 document types.
- Assessing data extraction quality through confidence scoring.
- Measuring system latency, response time, and computational efficiency.
- Conducting functional testing for frontend and backend applications.

Phase 5: Optimization & Deployment

- Fine-tuning prompts for improved classification and extraction accuracy.
- Developing CI/CD pipelines using GitHub Actions.
- Deploying the final system on Azure App Service.
- Implementing comprehensive logging and exception handling.
- Documenting findings and creating a comprehensive project report.

V. Requirements for proposed work

To successfully implement the Intelligent Document Processing Platform, the project requires a combination of hardware, software, infrastructure, and tools.

1. Hardware Requirements

- High-performance servers to handle document processing and AI operations.
- Sufficient storage capacity for document repositories.
- Network infrastructure to support large file transfers.

2. Software Requirements

- Python (3.8+) – Primary programming language for backend development.
- Azure App Service – For deploying and hosting the IDPP solution.
- Azure Blob Storage – For document storage and retrieval.
- Azure Document Intelligence – For document analysis and data extraction.
- OpenAI GPT Models – For advanced prompt-based classification and extraction.
- Azure AI Search – For vector-based document search capabilities.

3. Infrastructure Requirements

- Azure Key Vault – For secure storage and retrieval of API keys.
- CI/CD Pipeline – GitHub Actions for automated testing and deployment.
- Active Directory Integration – For user authentication and management.
- IBPS System Integration – For workflow management and document routing.

4. Additional Tools & Libraries

- PyPDF2 – For PDF processing and manipulation.
- Multiprocessing – For optimizing thread utility and performance.
- Testing Frameworks – For functional and accuracy testing.
- Logging Systems – For comprehensive error tracking and performance monitoring.

VI. Bibliography/References

- **Intelligent Document Processing** Microsoft. Azure Document Intelligence Documentation. Retrieved from https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/
- **Large Language Models & Prompt Engineering** OpenAI (2023). GPT-4 Technical Report. Retrieved from https://arxiv.org/abs/2303.08774 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. [NeurIPS 2020]. Retrieved from https://arxiv.org/abs/2005.14165
- **Azure AI Search** Microsoft. Azure AI Search Documentation. Retrieved from https://learn.microsoft.com/en-us/azure/search/
- **PDF Processing & Data Extraction** PyPDF2 Documentation. Retrieved from https://pypdf2.readthedocs.io/en/latest/
- **Azure Cloud Services** Microsoft. Azure App Service Documentation. Retrieved from https://learn.microsoft.com/en-us/azure/app-service/ Microsoft. Azure Blob Storage Documentation. Retrieved from https://learn.microsoft.com/en-us/azure/storage/blobs/