OCR Driven PDF Text Extraction

BITS ZC4999T: Capstone Project

MID REPORT

By

Student Name: Nishant Chaudhary

BITS ID: 202117b3876

Capstone Project work carried out at

HCLTECH Ltd., Noida

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE

PILANI (RAJASTHAN)

December 2025.

Annexure 5G: Format for Mid semester progress evaluation Sheet

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

WORK-INTEGRATED LEARNING PROGRAMMES DIVISION

BITS ZC499T : Capstone Project Mid-Semester Progress Evaluation Sheet

ID No.	**:**	**202117b3876**
NAME OF THE STUDENT	**:**	**Nishant Chaudhary**
EMAIL ADDRESS	**:**	**202117b3876@wilp.bits-pilani.ac.in**
STUDENT'S EMPLOYING ORGANIZATION & LOCATION	**:**	**HCLTECH Ltd., Noida**
MENTOR'S NAME	**:**	**Vaishnani Dhruvkumar**
Mentor's Desgnation	**:**	**Senior Technical Lead**
MENTOR'S EMPLOYING (ORGANIZATION & LOCATION)	**:**	**HCLTECH Ltd., Pune**
MENTOR'S EMAIL ADDRESS	**:**	**vaishnanidhru.jiten@hcltech.com**
CAPSTONE PROJECT TITLE	**:**	**OCR Driven PDF Text Extraction**

1. Executive Summary:

This capstone project focuses on solving a major challenge faced by organizations: extracting text from scanned PDFs where important information is trapped within image-based documents. Manual text extraction is slow, costly, and error-prone, creating delays in compliance, auditing, reporting, and general operational workflows.

To address this, an OCR-based automated system has been developed using **Python** and **Tesseract OCR**, enabling efficient conversion of scanned PDFs into machine-readable formats such as **plain text** and **JSON**. This transformation allows easy indexing, searching, and seamless integration with enterprise applications.

The implemented solution reduces manual effort by **up to 70%**, improves reporting speed by **50%**, and enhances accessibility for legacy documentation. Planned enhancements—such as multilingual OCR, AI-driven summarization, and chatbot-based query interaction—aim to make the system scalable and intelligent for diverse industry needs.

2. Project Methodology:

The project follows an Agile Iterative Model:

- Development is divided into phases for frontend, backend, and AI integration.
- Each phase includes design, implementation, and testing.
- Regular feedback from mentors is incorporated to improve functionality.

Details of Work Done Till Date

The development of the OCR-based Document Processing System has progressed through multiple structured phases:

Phase 1 – Project Setup and Planning

The project began with defining the scope, objectives, and overall architecture for both backend automation and the user interface.
A foundational repository structure was created, including setup for Python, Tesseract OCR, and the environment needed for PDF processing.

Phase 2 – Backend OCR Pipeline Development

Core backend components were implemented, including:

- PDF-to-image conversion using pdf2image
- Text extraction logic using Tesseract OCR and pytesseract
- Initial preprocessing techniques to improve OCR accuracy

This phase ensured that scanned PDFs could be successfully read and processed.

Phase 3 – PDF & File Handling

The system was enhanced to allow users to upload scanned PDF files.
The backend now:

- Accepts PDF uploads
- Converts each page into images
- Extracts and returns the raw text

The extracted text is displayed in a simple frontend viewer for user review.

Phase 4 – Structured Output Generation (JSON/Text)

A structured output module was built to:

- Format extracted text into clean plain text
- Convert extracted results into JSON for future system integration

Error handling and logging mechanisms were added to manage corrupted or password-protected files.

Current Progress: Phase 5 – Performance Optimization & Multilingual Support Planning - (In Progress)

Work was done on optimizing OCR performance for large PDFs and multi-page files. Additionally, multilingual OCR support was planned using Tesseract language packs for regional and international document processing.

Phase 6 – AI Integration

The ongoing phase focuses on integrating AI features, starting with:

- Automated summarization of extracted text
- Intelligent text structuring
- Future chatbot-based interactive queries
  This will make the system more intelligent and capable of supporting real-time document understanding.

3. Tools & Technology Used:

- ReactJS:
  Frontend JavaScript library used for building a responsive and interactive user interface for uploading files and displaying extracted OCR results.
- Material-UI (MUI):
  Component library used with React for creating clean, modern, and responsive UI elements.
- Python:
  Primary backend programming language for handling OCR processing, PDF conversion, text extraction, and structured output generation.
- Tesseract OCR:
  Open-source OCR engine used to extract text from scanned PDFs and image-based documents.

- pytesseract:
  Python wrapper for Tesseract OCR that connects the OCR engine with the backend Python workflow.
- pdf2image:
  Used to convert PDF pages into images before passing them into the OCR engine for accurate text extraction.
- PyMuPDF (fitz):
  Library for reading PDFs, extracting metadata, handling page operations, and processing password-protected or corrupted documents.
- Python-docx (future scope):
  Planned for handling and extracting text from .docx files if needed.
- JSON:
  Used for generating machine-readable structured output from extracted text.
- Python Logging Module:
  Used for error tracking, debugging, and maintaining traceability during text extraction.
- AI/NLP Libraries (future enhancement):
  Will be used for multilingual OCR, summarization, and chatbot-based query support.

4. Scope of Work Remaining:

- Integration of AI-based question generation (newly added scope)
- UI enhancements for better user experience
- Input validation for uploads, difficulty filters
- End-to-end testing, debugging, and performance optimization
- Final documentation, screenshots, and presentation preparation
- Completion of multilingual and AI-enhancement modules (planned)

5. Plan of Work Yet to be done:

Phase	Duration	Activity	Status
OCR Accuracy Improvement	**7-Dec-2025 to 15-Dec-2025**	Improve text extraction accuracy by applying image preprocessing (noise removal, thresholding, DPI adjustments) before OCR. Test with different types of scanned PDFs.	**Complete**
Multilingual OCR Support	**16-Dec-2026 to 07-Jan-2026**	Integrate Tesseract language models to support extraction in multiple languages and test accuracy on sample multilingual documents. readability, add loaders, and refine layout consistency.	**In Progress**
Input Validations	**08-Jan-2026 to 11-Jan-2026**	Add validations for PDF uploads, preventing unsupported file types, handling empty inputs, and validating multilingual OCR selections.	**Pending**
Performance Optimization & Error Handling	**11-Jan-2026 to 20-Jan-2026**	Optimize OCR speed for large multi-page PDFs. Enhance error handling for corrupted or password-protected PDFs. Improve logging and fallback workflows.	**Pending**
UI Enhancements	**20-Jan-2026 to 25-Jan-2026**	Enhance the web interface for file upload, OCR progress indication, and clear display of extracted text and JSON output. Add responsive layout improvements.	**Pending**

BITS ZC499T : Capstone Project Mid-Semester Progress Evaluation Sheet

ID No.	**:**	**202117b3876**
NAME OF THE STUDENT	**:**	**Nishant Chaudhary**
EMAIL ADDRESS	**:**	**202117b3876@wilp.bits-pilani.ac.in**
MENTOR'S NAME	**:**	**Vaishnani Dhruvkumar**
CAPSTONE PROJECT TITLE	**:**	**OCR Driven PDF Text Extraction**

EVALUATION:

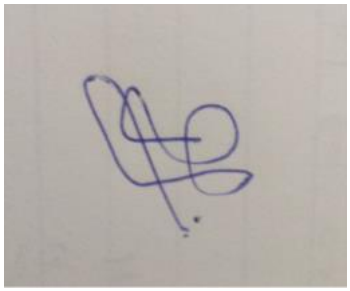
Remarks of the mentor:

The project is progressing well. The student has a clear understanding of the problem and is effectively utilizing AI. With consistent progress, the final solution is expected to be highly valuable.

CAPSTONE PROJECT PROGRESS EVALUATION *(Please put a tick (✔) mark in the appropriate box)*

EC No.	Component	Excellent	Good	Fair	Poor
1.	Capstone Project Outline	☐	✔		
2.	Work Progress & Achievements	✔	☐		
3.	Initiative and Originality	✔			
4.	Documentation & Expression	☐	✔		
5.	Research & Innovation	✔	☐		
6.	Relevance to the work environment	☐		✔	

	Mentor	Additional Examiner
Name	Vaishnani Dhruvkumar	Abhishek Dwivedi
Qualification	B.tech	B.tech
Designation	Senior Technical Lead	Senior Consultant
Employing Org and Loc.	HCLTECH Ltd, Pune	HCLTECH Ltd, Noida
Phone No. (with STD Code)	+91 8238051964	+91 7905412512
Email Address	**vaishnanidhru.jiten@hcltech.com**	Abhishek_dwivedi@hcltech.com
Date	17.12.2025	17.12.2025

		
Signature of Student	Signature of Mentor	Signature of Additional Examiner
Nishant Chaudhary	Vaishnani Dhruvkumar	Abhishek Dwivedi