Document Digitization

Dr R Dinesh Kumar, G Akhila, V Harika

[1] Associate Professor, Department Of Cse, Bhoj Reddy Engineering College For Women, India.

[2,3]B. Tech Students, Department Of Cse, Bhoj Reddy Engineering College For Women, India.

ABSTRACT

This project presents a smart and efficient system for document digitization and data entry automation using a combination of Optical Character Recognition (OCR) and deep learning techniques. The proposed system primarily targets the digitization of loan application forms, which are often received in printed or handwritten formats. Using Tesseract OCR, the system extracts text from scanned images, followed by preprocessing techniques such as grayscale conversion and Otsu thresholding to enhance image clarity.

To improve accuracy, transformer-based models like BERT or T5 are incorporated for contextual text understanding and refinement. Key fields such as names, dates, and account numbers are identified and structured using Named Entity Recognition (NER) and regex-based validation. The extracted and cleaned data is then exported into Excel files in a fixed format, making it easy to review and integrate with enterprise applications.

This solution significantly reduces human effort, eliminates common manual errors, supports real-time processing, and ensures scalability for large-scale document handling.

1. INTRODUCTION

Document digitization and processing is used to improve accuracy and efficiency while minimizing human intervention in data entry tasks. The project will begin with a literature review to understand existing OCR and deep learning-based document processing techniques. The implementation will involve Teserract for image-based text recognition and Transformer-based models (such as BERT or T5) for refining extracted text and ensuring contextual accuracy. Named Entity Recognition (NER) techniques will also be explored to extract key information such as names, dates, and numerical values.

In today's data-driven world, banks deal with vast amounts of physical and digital loan forms. Traditional data entry methods are labor-intensive, time- consuming, and error-prone. To overcome these challenges, a Deep Learning-Based Smart Data Entry System is proposed for document digitization and processing. By leveraging Optical Character Recognition (OCR), this system automates text extraction, validation, and structuring from various document formats. The implementation of deep learning techniques enhances accuracy and efficiency.

2-REQUIREMENT ANALYSIS

Functional Requirements

These are the requirements that refers to the specific actions, behaviors, or tasks a system or application is designed to perform. Functional requirements describe what the system must do to achieve its objectives and typically outline features, inputs, outputs, and interactions.

Admin Module:

Admin Module:

1.User

- Login
- Upload image or document

- View the excel whether the data is entered or not
- Logout

2. Admin

- Login
- View Users
- View the excel.
- Logout

Non-Functional requirements

These are the requirements that refers to the quality attributes or characteristics of a system that do not directly relate to its specific tasks but focus on how the system performs under certain conditions. These requirements address performance, usability, reliability, scalability, and other operational aspects.

- Scalability : Ability to handle a growing number of users.
- Usability : User-friendly interface that is easy to navigate.
- Reliability : Maintain high system availability to prevent critical
periods.

Hardware Requirements

- Processor : Intel i5
- RAM : 8 GB
- Hard Disk : 512GB

Software Requirements

- Operating System : Windows 11
- Programming Language : Python
- IDE : Visual Studio Code
- Front end Technologies : HTML, CSS
- Framework : Django
- Web Server : MySQL

3-DESIGN

Architecture

Project architecture represents number of components we are using as a part of our project and the flow of request processing i.e. what components in processing the request and in which order. An architecture description is a formal description and representation of a system organized in a way that supports reasoning about the structure

Software Architecture

Software architecture design tools help to build software that does not have security issues. This is key because there are software risks in all areas of the software development process. When teams avoid software flaws or bugs, they can move forward with confidence. However, since this is not always possible, software architecture design tools also need to have the ability to find flaws during the creation of software and correct them efficiently. When using software architecture design tools that can identify flaws, you will have the ability to analyse the fundamental software design, assess the chance of an attack, figure out potential threat elements, and identify any weaknesses or gaps in existing security.

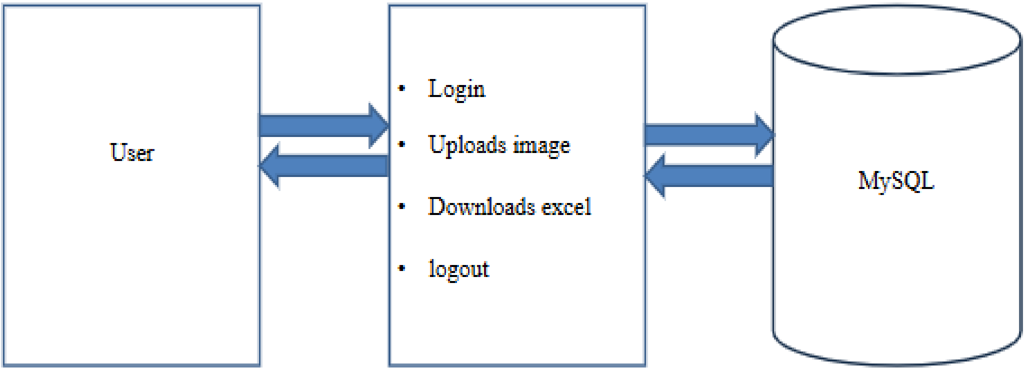


Fig. 3.1 Software Architecture

Technical Architecture

Technical Architecture is a form of IT architecture that is used to design computer systems. It involves the development of a technical blueprint regarding the arrangement, interaction, and interdependence of all elements so that system-relevant requirements are met.

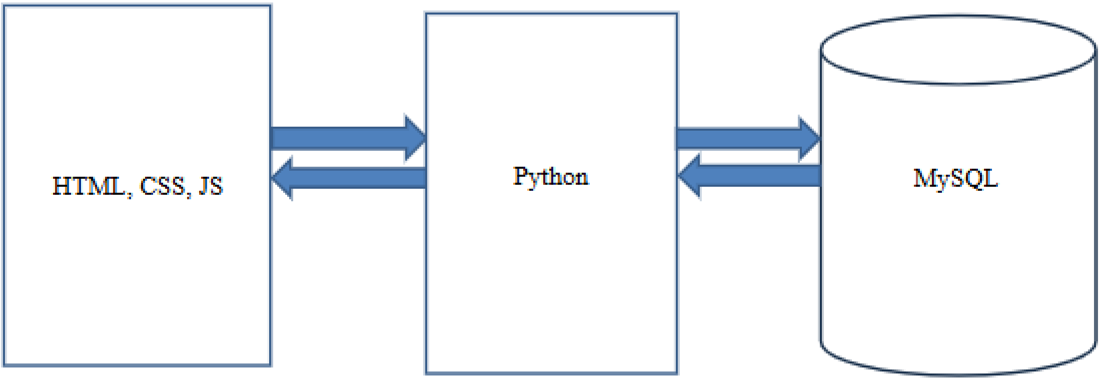


Fig.3.1 Technical Architecture

4-IMPLEMENTATION METHODOLOGY

The Document Digitization project follows a structured development approach starting with requirement analysis and system design. It uses Python, Django, Tesseract OCR, and OpenCV to extract and process text from scanned images. Preprocessing techniques like grayscale conversion and Otsu thresholding enhance OCR accuracy. Extracted data is structured using regex and exported to Excel with Pandas. The system is tested through unit, integration, and user acceptance testing, and deployed as a web app for users to upload documents and download structured Excel data efficiently.

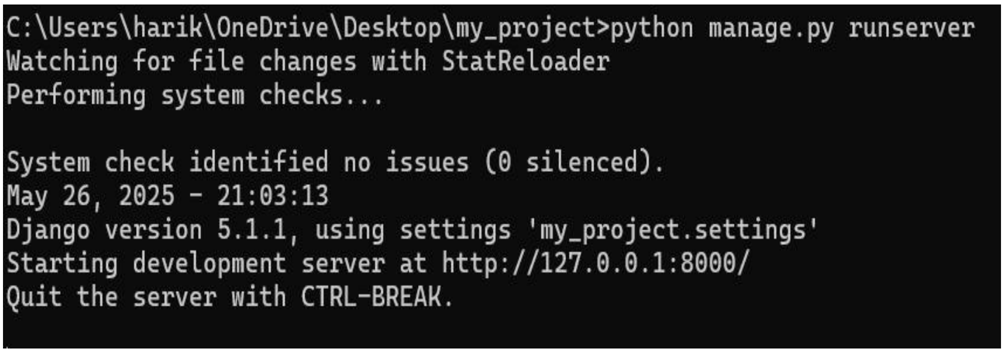
Data Extraction and Preprocessing Objective:

It is to accurately extract text from scanned documents using Tesseract OCR and enhance recognition accuracy through image preprocessing techniques like grayscale conversion and Otsu thresholding, followed by structuring the extracted text into key-value pairs for organized storage in Excel format..

Python

Python is one of the most popular programming languages now existing. The main reason for the creation of a programming language like python was to enhance the features to a large extent that were available in the present existing languages. The other reason was to invent a language which can be used easily for the developers who work a lot on media other than texts like speech, images and videos. The other important reason was to increase the built-in functions so as to reduce the number of lines in the codes and implement simplicity. Python is basically created in such a way that the garbage is involuntarily and automatically collected. The Python language can be called as a mixture of all the languages with more features added to it. It is a structured language yet it does not support the use of the semicolons at the end of each operation. Python consists of a very large standard library which consists of a huge number of built-in functions which reduce the developer's load of writing hundreds of lines to perform a single and simple task.

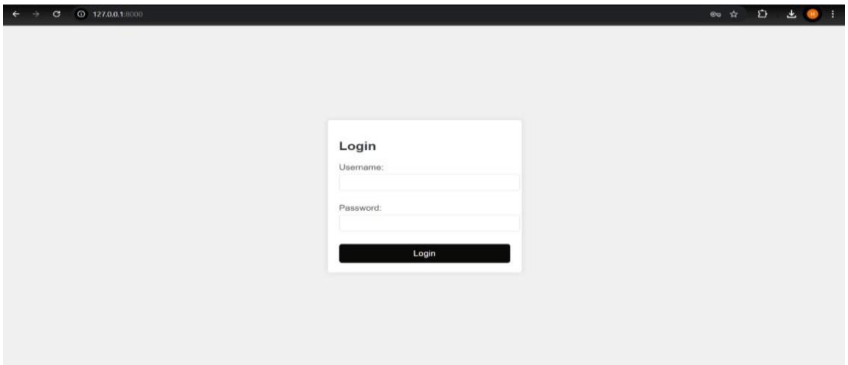
5-SCREENSHOTS



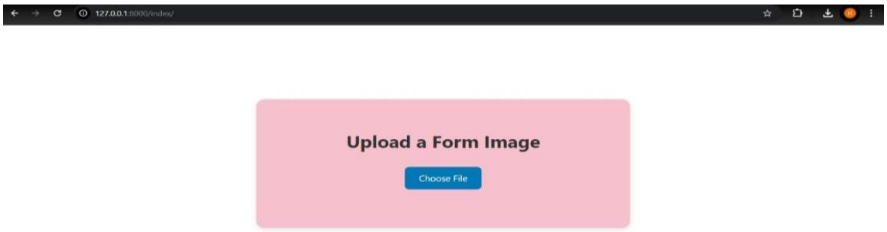
```
C:\Users\harik\OneDrive\Desktop\my_project>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).
May 26, 2025 - 21:03:13
Django version 5.1.1, using settings 'my_project.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

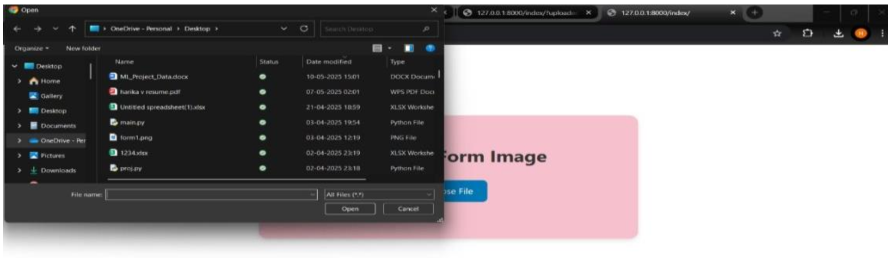
Screenshot 1 Run the web app from the command prompt



Screenshot 6.2 Login page



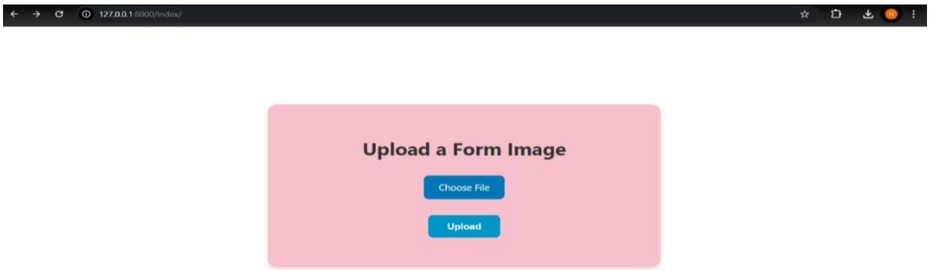
Screenshot 3 Click on Choose File



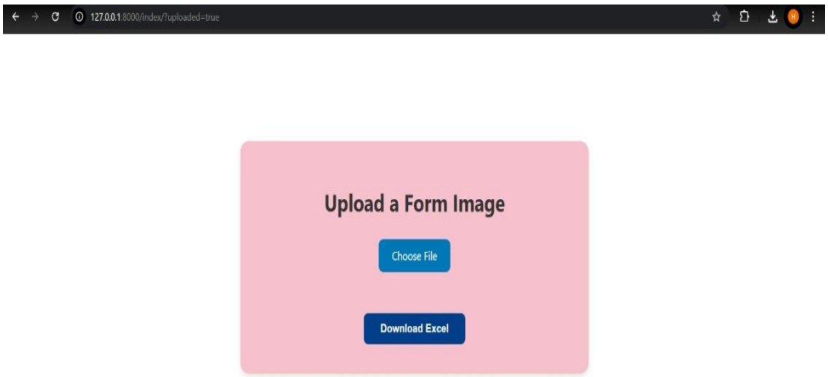
Name	Status	Date modified	Type
ML_Project_Data.docx	●	10-05-2025 15:01	DOCX Docum
harika v resume.pdf	●	07-05-2025 02:01	WPS PDF Doci
Untitled spreadsheet(1).xlsx	●	21-04-2025 18:59	XLSX Workshe
main.py	●	03-04-2025 19:54	Python File
form1.png	●	03-04-2025 12:19	PNG File
1234.xlsx	●	02-04-2025 23:19	XLSX Workshe
proj.py	●	02-04-2025 23:18	Python File

File name:
All Files (*.*)
Open
Cancel

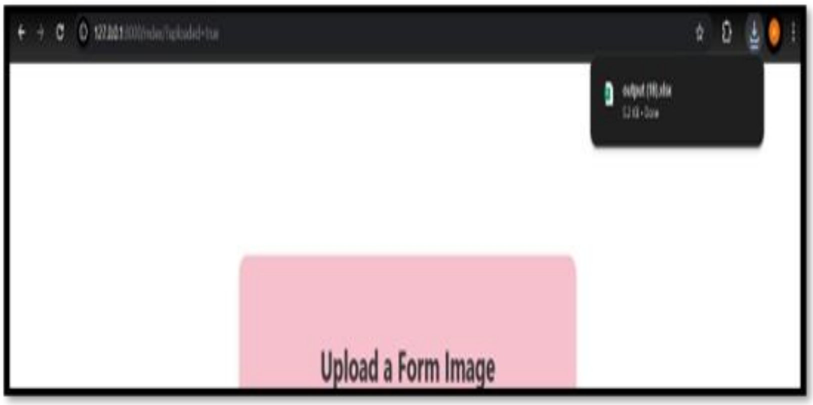
Screenshot 4 Select file to upload



Screenshot 5 Click on upload



Screenshot 6 Click on Download Excel



Screenshot 7 Downloaded Excel sheet

6-CONCLUSION

This project demonstrates an effective and intelligent approach to document digitization using OCR and deep learning techniques. By automating the extraction, validation, and structuring of data from printed and handwritten forms, the system significantly reduces manual workload and human errors. Tools like Tesseract and transformer-based models ensure high accuracy and contextual understanding, while exporting structured data to Excel enhances usability. The solution is scalable, cost-effective, and well-suited for organizations handling large volumes of documents, such as banks. Overall, this system bridges the gap between physical documentation and digital efficiency, improving operational speed and accuracy.

REFERENCES

1. **Smith, R. (2022).** *An Overview of Tesseract OCR Engine.* International Journal of Computer Vision and Image Processing, 14(1), 10-18.
   Discusses updates and capabilities of Tesseract for image-to-text processing.
2. · **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2023).** *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Journal of Artificial Intelligence Research, 72, 1-18.
   Explains BERT's use for contextual understanding and Named Entity Recognition (NER).
3. · **Vincent, W. S. (2023).** *Django for Professionals (4th ed.).* WelcomeToCode.
   A practical guide on building secure, scalable web applications using Django.
4. · **Brownlee, J. (2022).** *Deep Learning for Computer Vision.* Machine Learning Mastery.
   Covers techniques for image preprocessing and integrating deep learning in OCR tasks.
5. · **Chowdhury, S., & Rahman, M. (2022).** *Enhancing OCR Accuracy with Image Preprocessing Techniques.* Proceedings of the IEEE International Conference on Image Processing (ICIP), 1567–1571.·
   A technical paper evaluating preprocessing methods like Otsu thresholding and grayscale conversion.