**Milestone1**

**Receipt & Invoice Digitizer**

**1.Introduction:**

In today’s digital era, businesses and individuals handle a large volume of paper-based receipts and invoices. Manual handling of these documents often leads to challenges such as data loss, incorrect entries, lack of traceability, and increased operational costs. Automating the digitization of invoices is therefore a critical requirement for efficient financial management.

The Receipt & Invoice Digitizer project focuses on building an automated system that converts scanned receipts and invoices into structured digital data using Optical Character Recognition (OCR). This milestone lays the foundation by enabling file upload, image preprocessing, and raw text extraction.

**2.Problem Statement:**

Businesses and individuals handle numerous paper receipts and invoices, which are prone to loss, errors, and manual entry delays. This project builds a system that automatically scans, extracts, and digitizes information from receipts and invoices using OCR (Optical Character Recognition) and NLP-based field extraction. The digitized data is stored in a structured format, making it easy to search, analyze, and integrate with accounting or ERP systems.

**Key points:**

Manual data entry is time-consuming and error-prone

OCR enables automatic extraction of text from receipts/invoices

Persistent storage allows long-term data tracking and analytics

## **Objectives of the Milestone**

• Implement file upload.

• Preprocess images for OCR.

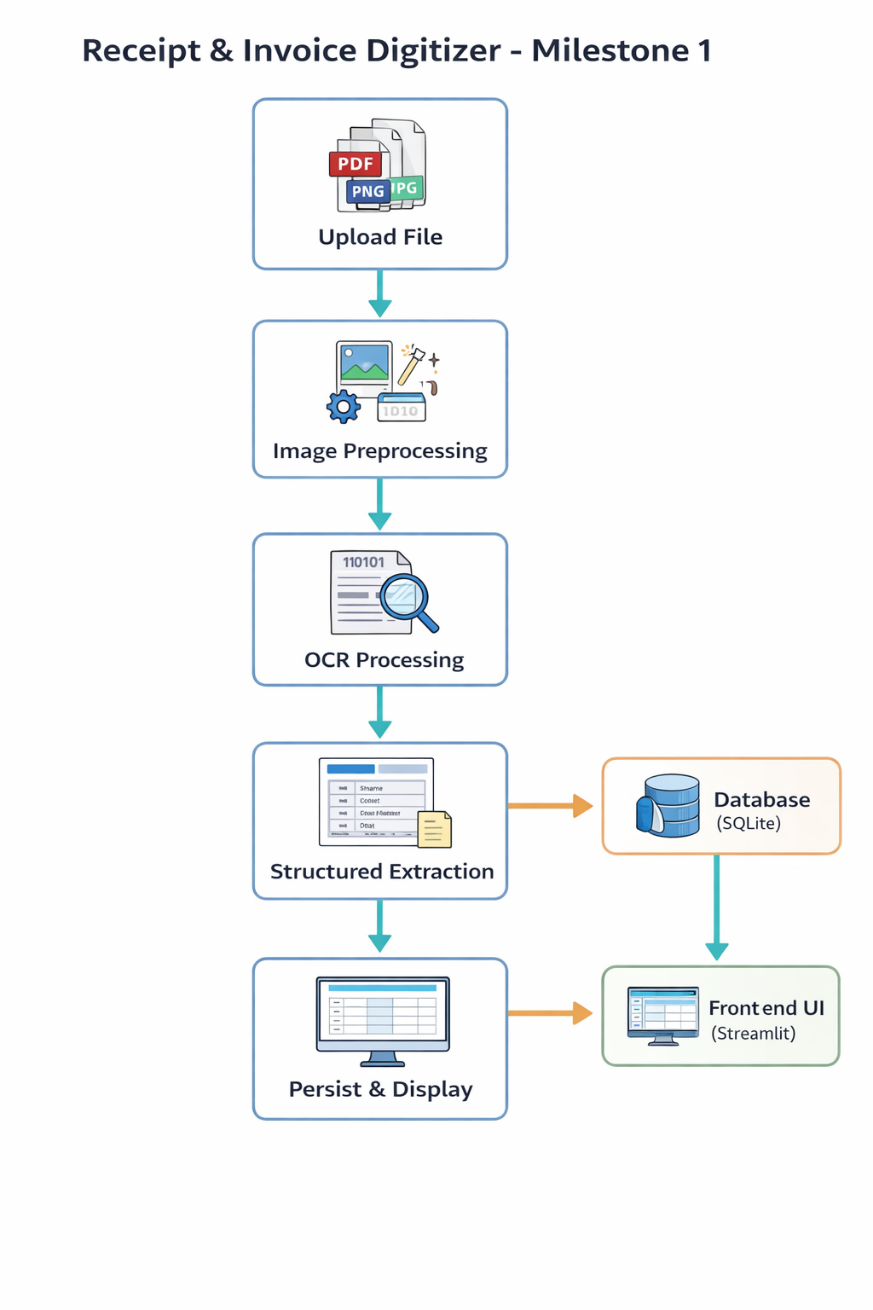
• Extract raw text from sample receipts.

**System Architecture**

The high-level architecture of the system is as follows:

**User Upload → Image Preprocessing → OCR Engine → Raw Text Output → UI Display**

Each module is loosely coupled, allowing easy extension in future milestones.



**Functional Modules Description**

**5.1 File Upload Module**

**Purpose:**

Allows users to upload invoice or receipt files through the web interface.

**Features:**

* Built using Streamlit’s file\_uploader
* Supports JPG, JPEG, PNG, and PDF formats
* Uploaded files are saved locally using unique UUID-based filenames

**Benefits:**

Prevents file name collisions

Ensures traceability of uploaded documents

**5.2 Image Preprocessing Module**

**Purpose:**

Enhances image quality to improve OCR accuracy.

**Techniques Used:**

Conversion to grayscale

Image resizing for better character recognition

Noise reduction using bilateral filtering

Adaptive thresholding for clearer text regions

**Outcome:**

Preprocessed images provide significantly improved OCR results compared to raw images.

**5.3 OCR Module**

**Purpose:** Extracts machine-readable text from preprocessed images.

**Tool Used:**

Tesseract OCR (Open Source)

**Configuration:**

OCR Engine Mode: --oem 3

Page Segmentation Mode: --psm 4

**Output:**

Raw unstructured text extracted from the invoice or receipt

**5.4 PDF Handling Module**

**Purpose:**

Enables OCR on multi-page PDF invoices.

**Approach:**

PDFs are converted to images using Poppler

Each page is processed individually

OCR results are concatenated

**Advantages:**

Supports scanned invoices

Works for invoices spanning multiple pages

**5.5 Text Structuring**

Although full structuring is implemented in later milestones, this stage includes:

Initial regex-based exploration

Identification of potential invoice fields such as:

Store name

Invoice number

Date

Total amount

This helps validate OCR quality early in development.

**5.6 Persistent Storage Module**

* SQLite used for lightweight persistent storage
* Table schema:
  + store\_name
  + invoice\_no
  + date
  + total
  + category
  + created\_at

**Advantages:**

No external DB dependency

Easy integration with Streamlit

**6. User Interface Design**

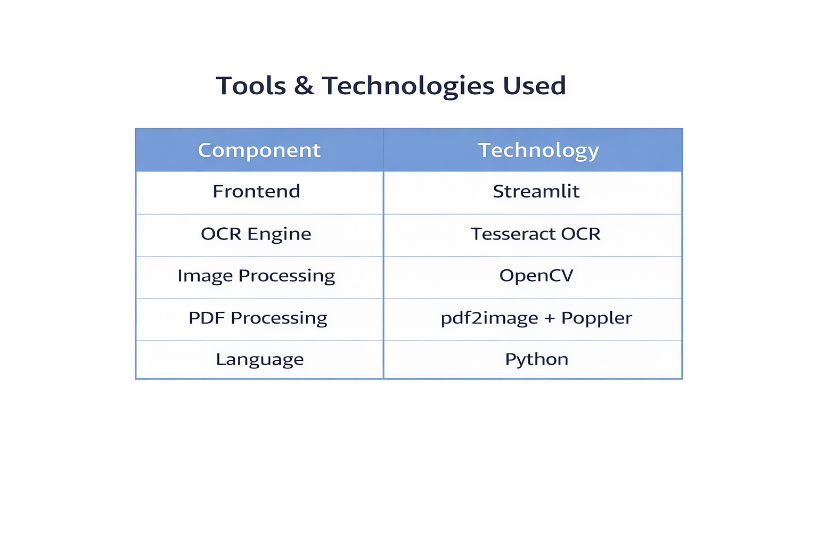
The Streamlit-based UI includes:

1. File upload section
2. Side-by-side display of original and preprocessed images
3. Structured and raw OCR output sections

**Benefits:**

* Improves transparency of OCR processing
* Helps users visually verify extraction quality

**7. Tools & Technologies Used**



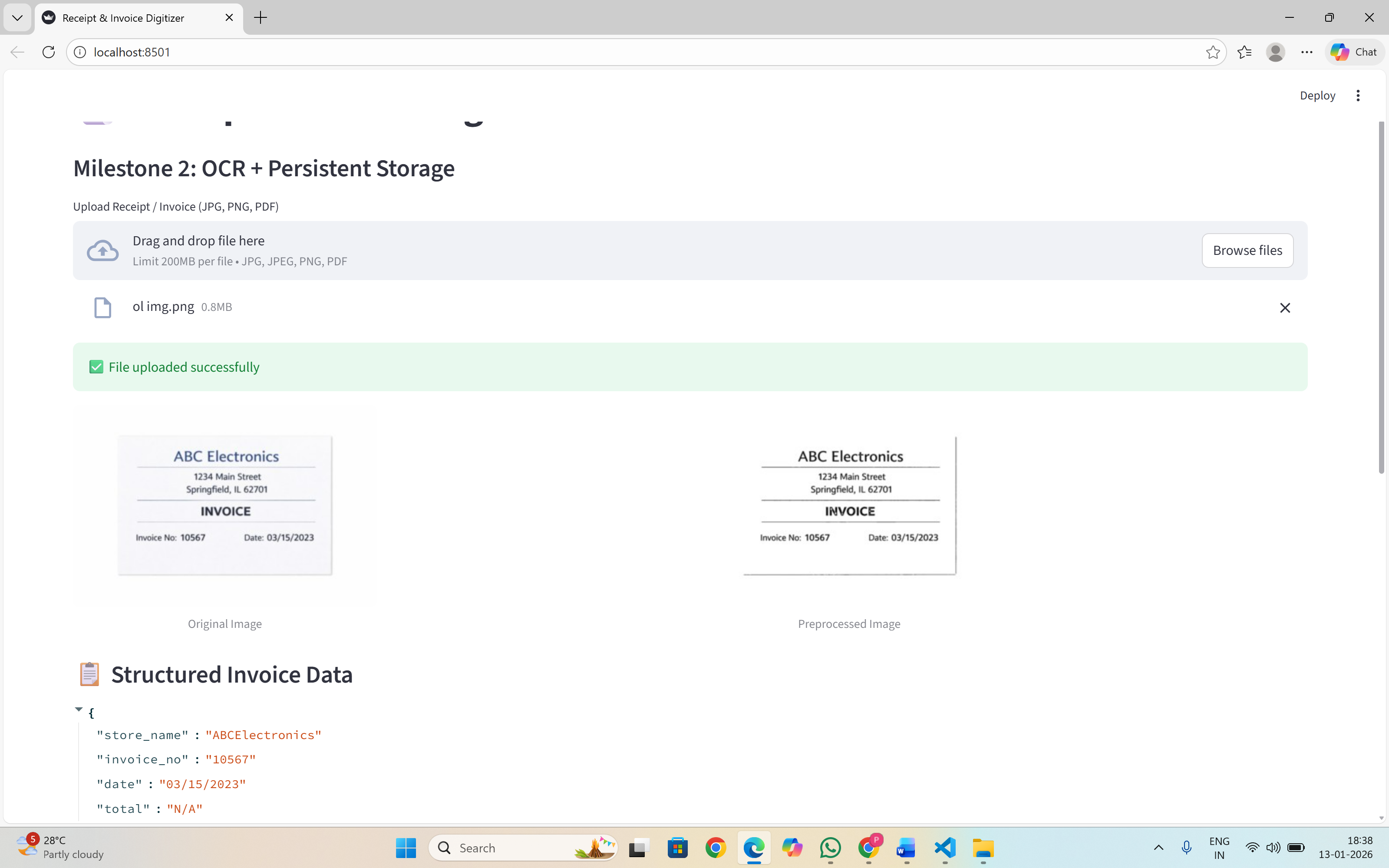
**8.Output**

OCR successfully extracts text from clear invoice images

Preprocessing significantly improves recognition accuracy

Complex layouts may require fine-tuning of OCR parameters

**Screenshots of extracted output validate the working of the pipeline.**



**9. Conclusion**

Milestone 1 successfully establishes the OCR foundation for the Receipt & Invoice Digitizer project. The system can upload files, preprocess images, extract text, and display results effectively. This milestone serves as a strong base for implementing structured extraction, duplicate detection, and persistent storage in future phases.

**10. Future Enhancements**

Advanced preprocessing techniques

Intelligent field extraction using NLP

Database persistence and analytics

Invoice classification and categorization

**Milestone 2**

**Receipt & Invoice Digitizer**

**1.Introduction:**

In today’s digital era, businesses and individuals handle a large volume of paper-based receipts and invoices. Manual handling of these documents often leads to challenges such as data loss, incorrect entries, lack of traceability, and increased operational costs. Automating the digitization of invoices is therefore a critical requirement for efficient financial management.

The Receipt & Invoice Digitizer project focuses on building an automated system that converts scanned receipts and invoices into structured digital data using Optical Character Recognition (OCR). This milestone lays the foundation by enabling file upload, image preprocessing, and raw text extraction.

**2.Problem Statement:**

Businesses and individuals handle numerous paper receipts and invoices, which are prone to loss, errors, and manual entry delays. This project builds a system that automatically scans, extracts, and digitizes information from receipts and invoices using OCR (Optical Character Recognition) and NLP-based field extraction. The digitized data is stored in a structured format, making it easy to search, analyze, and integrate with accounting or ERP systems.

**Key points:**

Manual data entry is time-consuming and error-prone

OCR enables automatic extraction of text from receipts/invoices

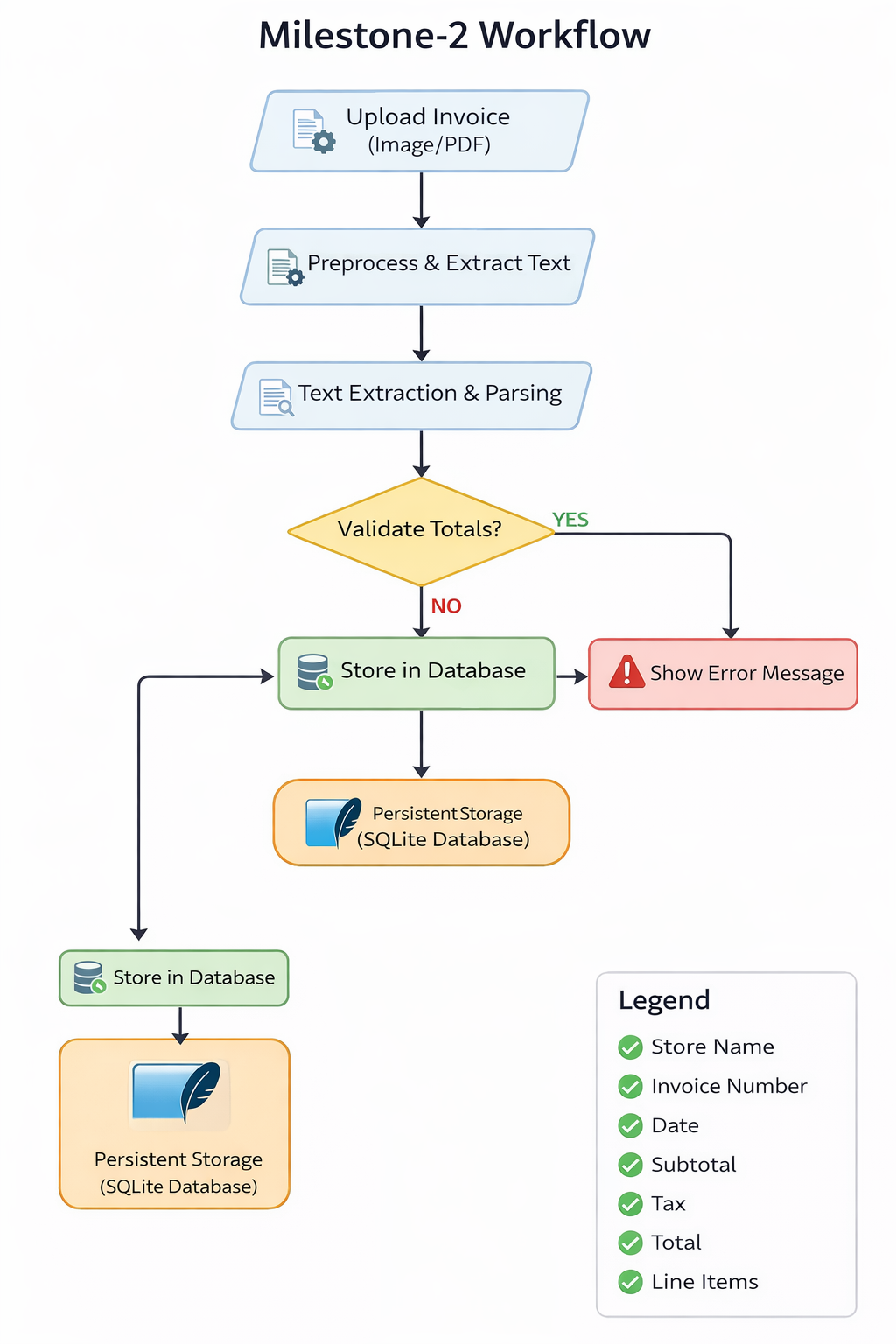
Persistent storage allows long-term data tracking and analytics

**Objectives of the Milestone**

Apply regex + NLP to parse key fields.

Validate totals and detect duplicates.

Store structured results in DB.



**Functional Modules Description**

**1. Invoice Upload Module**

**Description:**

This module allows users to upload invoice files into the system for processing.

**Functionality:**

Accepts invoice files in:

Image formats (JPG, JPEG, PNG)

PDF format

Performs file type validation before processing

Handles both single-page and multi-page PDFs

**Importance:**

Acts as the entry point of the application

Ensures only valid invoice formats are processed

Improves user experience with simple upload interface

**2. Image & Document Preprocessing Module**

**Description:**

This module enhances the quality of invoice images to improve OCR accuracy.

**Functionality:**

Converts colored images to grayscale

Resizes image for better text clarity

Applies noise reduction using bilateral filtering

Uses adaptive thresholding to highlight text

Converts PDF pages into images for OCR

**Importance:**

Reduces OCR errors caused by poor lighting or low resolution

Standardizes input format for consistent text extraction

Significantly improves recognition accuracy

**3. OCR Text Extraction Module**

**Description:**

This module extracts raw textual content from invoices using Optical Character Recognition (OCR).

**Functionality:**

Uses **Tesseract OCR engine**

Extracts text from preprocessed images

Supports different invoice layouts and fonts

Displays raw OCR output for debugging and verification

**Importance:**

Converts unstructured invoice images into machine-readable text

Forms the foundation for all further data extraction

**4. Invoice Field Extraction Module**

**Description:**

This module extracts meaningful invoice-level fields from the OCR text.

**Extracted Fields:**

Store / Vendor Name

Invoice Number

Invoice Date

Subtotal

Tax Amount

Total Amount

**Functionality:**

* Uses pattern matching and keyword-based logic
* Applies regular expressions to identify dates and monetary values
* Handles missing fields gracefully

**Importance:**

* Converts raw OCR text into structured, usable information
* Enables automated processing and storage

**5. Line Item Extraction Module**

**Description:**

This module identifies and extracts individual line items from invoices.

**Extracted Attributes:**

Item Description

Quantity

Unit Price

Item Total Amount

**Functionality:**

Analyzes OCR text line by line

Identifies numeric patterns for quantities and prices

Extracts descriptions from text preceding price values

Supports invoices with varying line item formats

**Importance:**

Enables detailed invoice analysis

Supports downstream analytics and auditing

Enhances accuracy of invoice breakdown

**6. Validation & Consistency Check Module**

**Description:**

This module verifies the correctness of extracted financial values.

**Functionality:**

Validates:

Total = Subtotal + Tax

Applies tolerance limits for rounding differences

Handles invoices without subtotal values

Displays validation success or error messages

**Importance:**

Prevents incorrect data from being stored

Improves reliability of extracted information

Builds trust in automated processing

**7. Duplicate Invoice Detection Module**

**Description:**

This module prevents duplicate invoice records from being stored.

**Functionality:**

Checks database for existing invoice numbers

Blocks insertion of duplicate invoices

Displays warning message to user

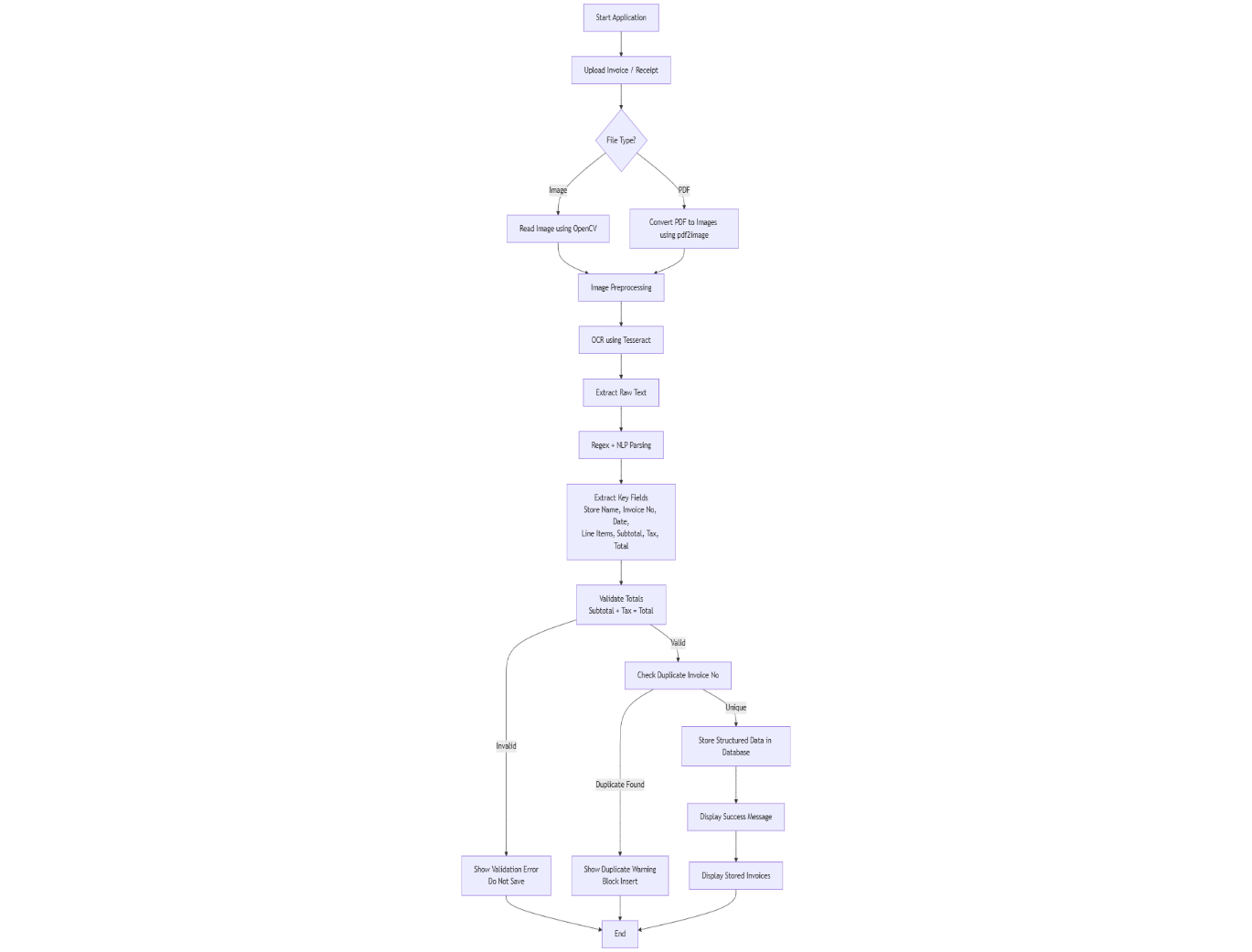
**Importance:**

Avoids data redundancy

Maintains database integrity

Prevents financial misreporting

**Flowchart of Milestone 2**



**8. Database Storage Module**

**Description:**

This module persists extracted invoice data into a database.

**Functionality:**

Uses **SQLite database**

Automatically creates tables if not present

Stores:

Invoice metadata

Line items

Raw OCR text

Ensures transactional integrity

**Importance:**

* Enables long-term storage and retrieval
* Supports reporting and auditing
* Makes the system scalable

**9. Invoice Records Viewing Module**

**Description:**

This module allows users to view all stored invoice records.

**Functionality:**

* Fetches invoice data from database
* Displays invoices in tabular format
* Supports easy review and verification

**Importance:**

Provides transparency to users

Enables manual verification if required

Enhances usability of the system

**10. User Interface & Control Module**

**Description:**

This module manages the interaction between the user and the system.

**Functionality:**

Built using **Streamlit**

Displays extracted fields clearly

Provides action buttons:

Validate Invoice

Save Invoice

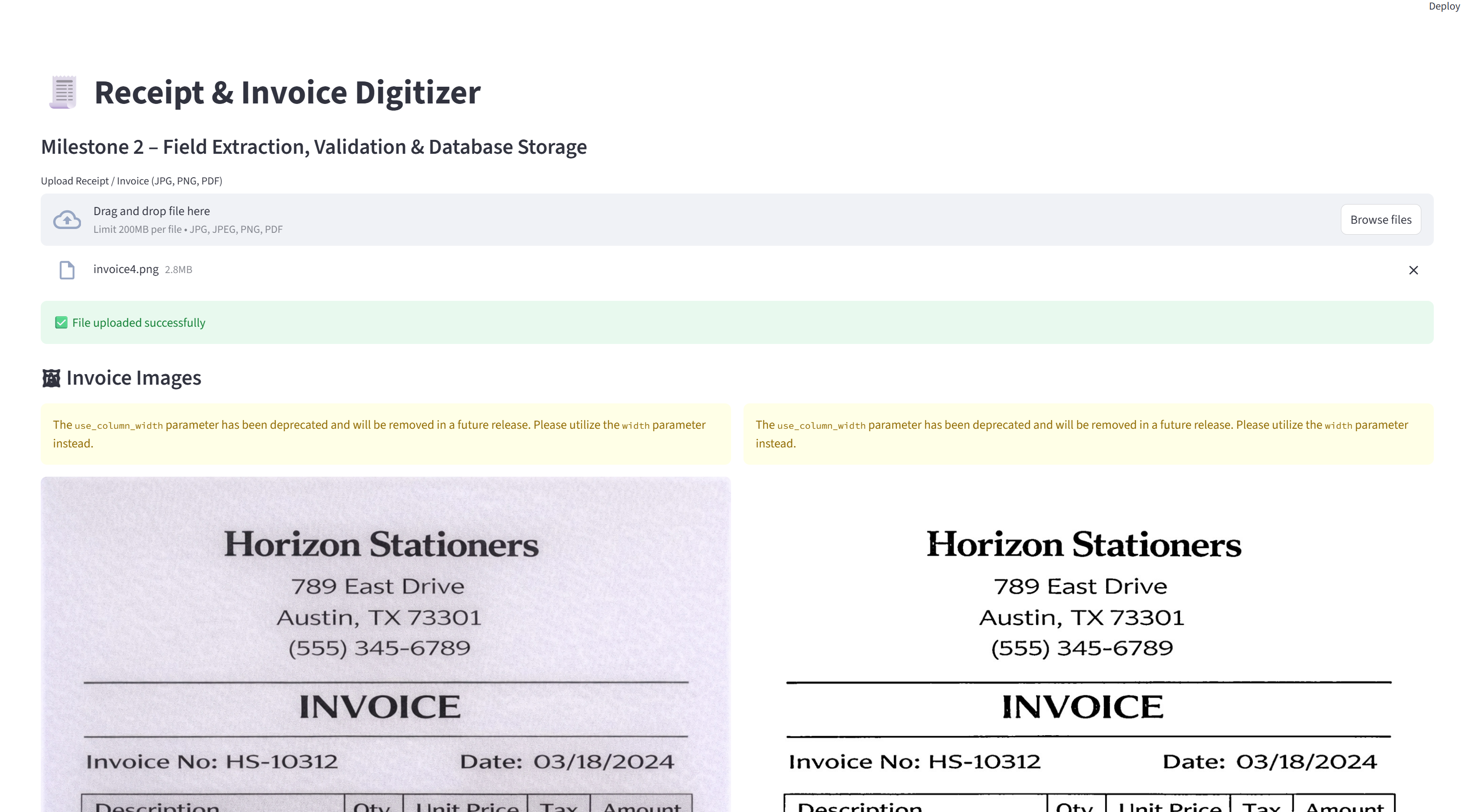
Shows real-time status messages

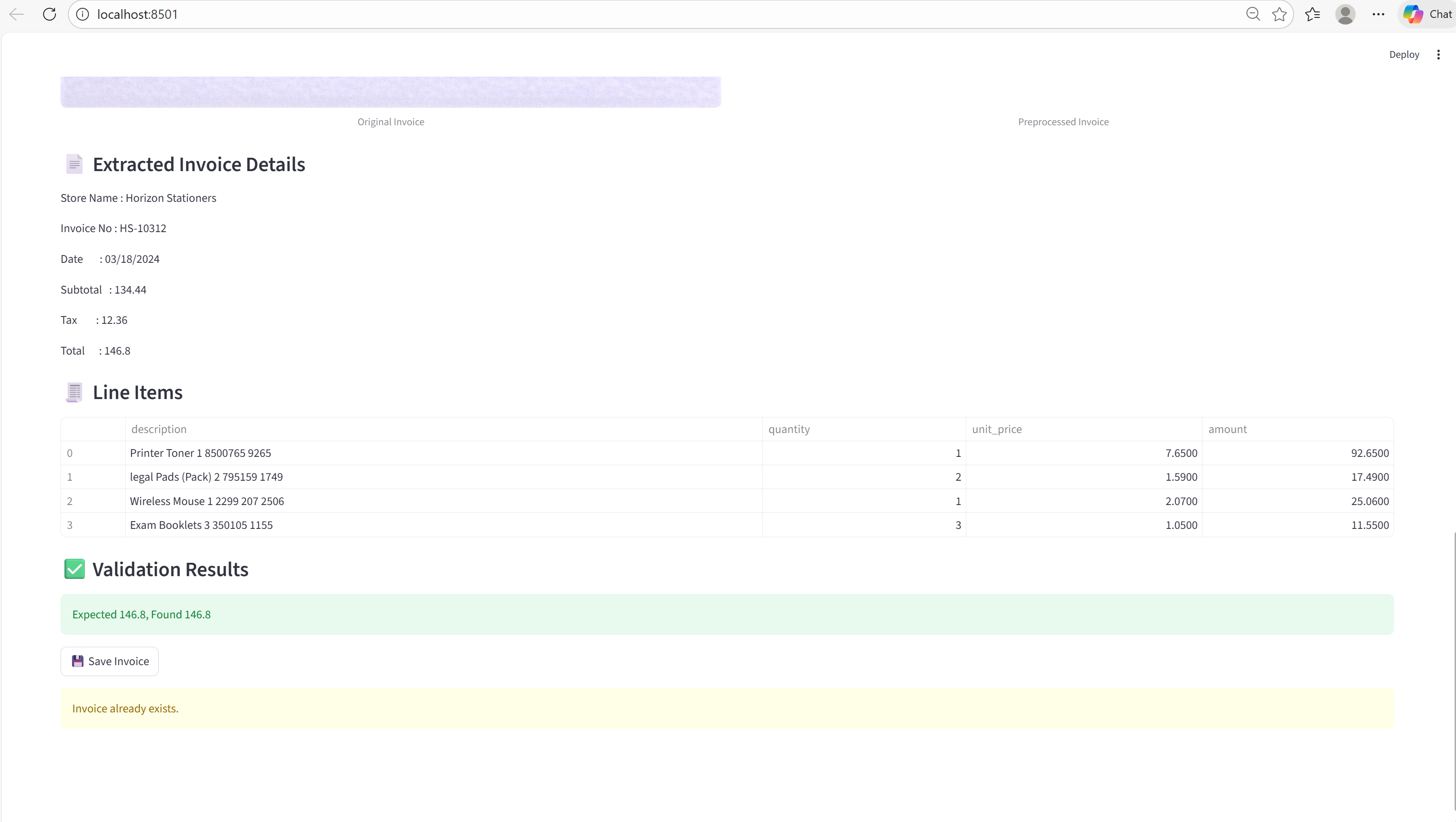
**Importance:**

Ensures smooth workflow

Reduces user errors

Makes the system easy to use even for non-technical users

**Output:  
**

****

**Conclusion:**

This milestone focused on **accuracy, reliability, and persistence**. By applying **image preprocessing techniques**, OCR accuracy was significantly improved. The extracted text was further processed using **regular expressions and rule-based logic** to identify and structure invoice fields. A **validation mechanism** was implemented to verify the correctness of financial data by cross-checking subtotal, tax, and total values.

Additionally, the system now supports **persistent storage using SQLite**, allowing extracted invoice data to be saved, checked for duplicates, and viewed later through a structured database table. The interactive **Streamlit user interface** provides a smooth workflow — from file upload to validation and database storage — making the application user-friendly and practical for real-world usage.

Overall, Milestone 2 successfully transforms raw invoice images into **verified, structured, and stored data**, laying a strong foundation for future enhancements such as NLP-based field extraction, analytics, and automated reporting in upcoming milestones.

**FUTURE ENHANCEMENTS**

Machine Learning–Based Field Extraction  
Rule-based extraction will be replaced with machine learning and NLP models to improve accuracy across different invoice layouts, formats, and vendors.

Intelligent Line Item Detection  
Advanced table detection and layout analysis techniques will be implemented to correctly extract quantities, unit prices, and amounts from complex invoices.

Multi-Language OCR Support  
The system will be enhanced to support invoices in multiple languages, enabling processing of regional and international documents.

Advanced Data Validation and Anomaly Detection  
Automated checks will be added to detect duplicate invoices, missing fields, mismatched totals, and abnormal values using intelligent validation rules.

Search, Filter, and Analytics Dashboard  
Users will be able to search and filter invoices by date, vendor, invoice number, and amount, along with viewing expense summaries and trends.

Export and System Integration  
Invoices can be exported to CSV, Excel, or PDF formats and integrated with accounting or ERP systems for seamless data flow.

User Authentication and Role Management  
Secure login functionality with role-based access (Admin and User) will be implemented for controlled access and data security.

Cloud Deployment and Scalability  
The application will be deployed on cloud platforms to support multiple users, large data volumes, and high availability.