

VIT CAMPUS CONNECT

CYBER SIMULATION EXERCISE – Automating File Cleansing 
Analysis leveraging AI

September 5, 2025



Automating File Cleansing Analysis



Introduction To Case Study

Automating File Cleansing Analysis

Background

Customers often receive client data in various formats, including spreadsheets, text files, PDFs, images (such as scanned documents or screenshots), and presentations. These files usually contain client-specific details or sensitive information that must be cleansed before they can be used for analysis.

Problem Statement

The challenge is to design an automated solution that can:

- **Cleanse files** by removing or masking client logos, names, [REDACTED] any Personally Identifiable Information (PII)* so that files cannot be traced back to the client
- **Pre-process diverse file formats** into a consistent, usable structure
- **Analyse the processed data** to extract [REDACTED] generate meaningful insights for the [REDACTED] clients

This ensures consultants can work with clean, standardized, anonymized data without exposing sensitive client information.

Note: A zipped file with diverse file formats would be provided for reference testing the prototype

*PII refers to any information that can be used to directly or indirectly identify an individual such as name, address, contact details, [REDACTED] Security Number (SSN) or any other government-issued identification numbers.

Expected Outcomes

Output Deliverables

Solution Design Walkthrough*

Walkthrough of the application functional design, highlighting:

- Files cleansing workflow – Removing or masking sensitive client information (e.g., client name, client logo, PII).
- File analysis workflow – Automating the reading █████ analysis of files to extract useful information for further evaluation.

This includes but is not limited to:

- **Text extraction:** Applying OCR █████ other relevant parsing methods to read text from images, diagrams, Tables, PowerPoint slides █████ scanned pdf files (.jpeg, .png, .pptx, .xlsx, .pdf)
- **Content interpretation:** Identifying key data elements within the extracted text, such as IAM policy statements, firewall rule entries, or IDS/IPS log snippets
- **Preparation for evaluation:** Generating cleansed █████ readable text in a st █████ ardized format (*please populate the outcome in the table within shared PowerPoint template*)

A working prototype:

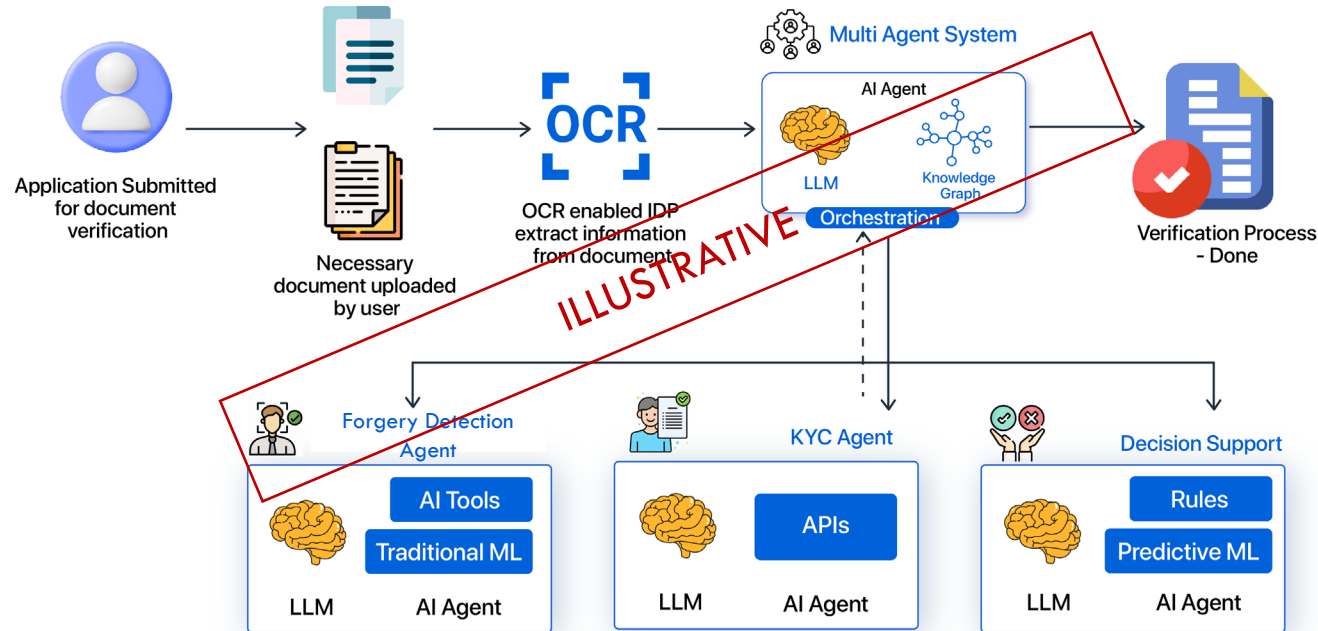
- A simple UI that allows users to upload files for cleansing █████ automated analysis
- If UI development is not feasible due to time constraints, demonstrate the backend functionality using an IDE (Integrated Development Environment) of your choice (e.g., Jupyter Notebook, PyCharm, etc.)

**Students are required to leverage the provided PowerPoint Template file to present their solution*

Sample Functional Design Diagram

Illustrative Purpose Only: The diagram below is for reference [REDACTED] not specific to the case study.

Autonomous Agents for Document Verification



Thank you

Secure greatness®

