1. Team Details

Team NAME: ASO AI TEAM

Team Members: - Thavishi Spandana

spandanathavishi@gmail.com

2. Problem Understanding and Scope

Problem Statement:

Sensitive information such as names, addresses, ID numbers, and photos is embedded across semi-structured documents like healthcare records, government IDs, and financial forms. Traditional anonymization methods often fail to handle the combination of text + images + complex layouts, leaving privacy gaps. This exposes organizations to data breaches, compliance violations, and loss of trust.

Target Documents & Formats: - Medical records, ID cards, financial statements, government forms

Types of PII to Detect & Redact: - Names, addresses, phone numbers, ID numbers, dates, barcodes, signatures, photos

User Personas: - Hospitals, government agencies, fintech companies, enterprises handling sensitive data

3. Proposed Solution & Approach

High-Level Architecture: Input Document → OCR + NLP → Vision Detection → Layout Understanding → Redaction & Export → Redacted Output.

AI/ML Models Considered:- OCR: Tesseract, EasyOCR- NLP: SpaCy, HuggingFace Transformers (NER)- Vision: YOLOv8, Detectron2, OpenCV- Layout: LayoutLMv3

Data Strategy:- Use provided datasets from hackathon resource center.- Generate synthetic data for testing (simulated IDs, forms).- Annotate sample documents for evaluation.

Innovation / USP:- Hybrid pipeline combining OCR, NLP, and Vision for robust deidentification.- Modular and open-source, designed for scalability and easy integration

4. Solution & Approach (Detailed)

Intended UI/UX Design:- Simple CLI for developers

.- REST API for integration

.- Future: Web-based dashboard.

Input & Output Format Expectations:

- Input: PDFs, images, multi-page scans.
- Output: Redacted PDFs/images + logs of redacted fields.

Accessibility / Ease of Use Considerations:

- Designed for both technical and non-technical users.
- Multilingual support planned.
- Lightweight deployment using Docker for low-resource settings