



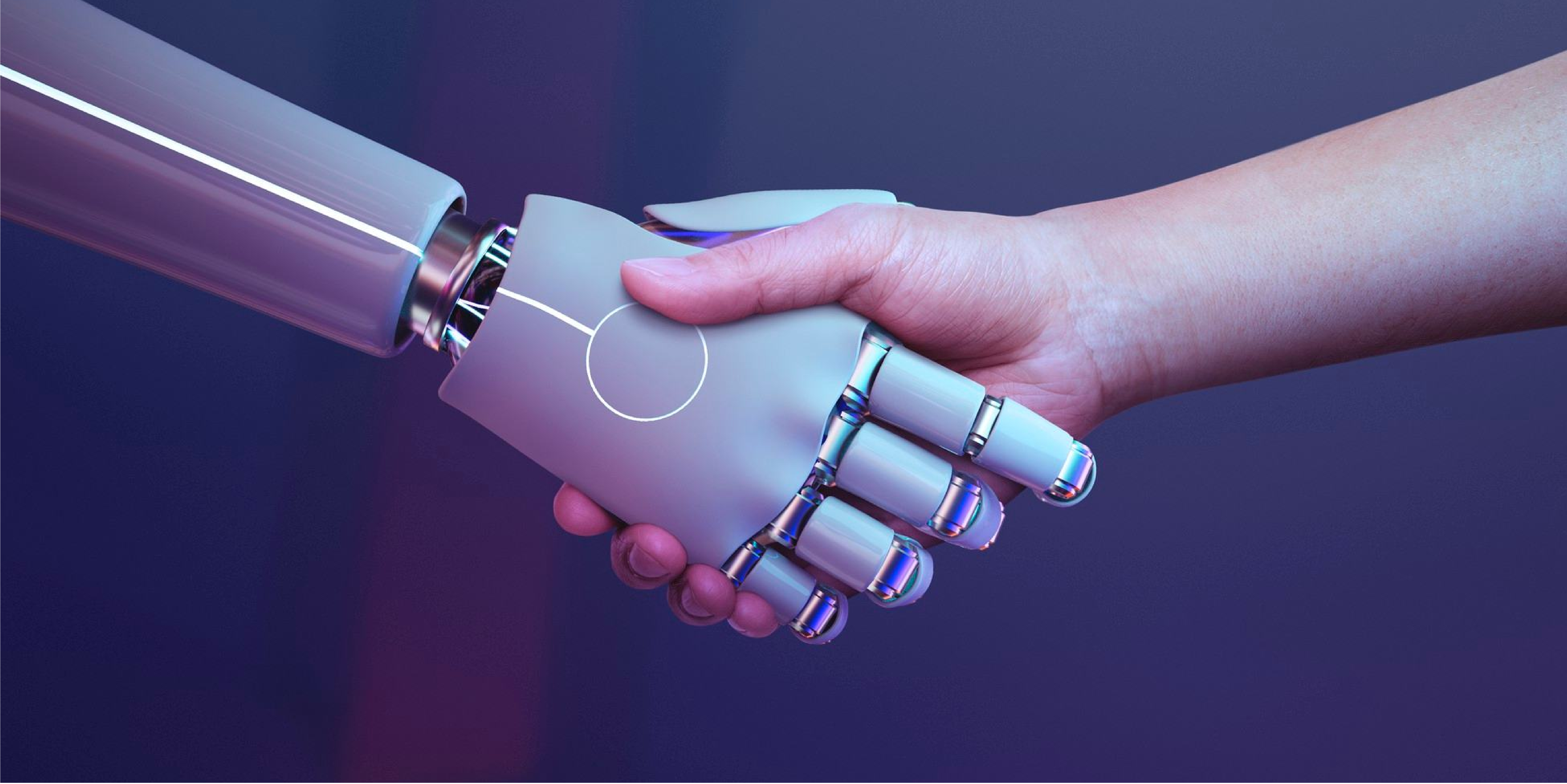
CASE STUDY

AI and Machine Learning for Enhanced Document and Image Classification with PII Detection



Index

3	___	Client Overview
3	___	Project Overview
4	___	Team Makeup
5	___	Objectives
5	___	Key Challenges Presented
6	___	The Solution
7	___	Results
7	___	Future Work



Client Background

Industry leader specializes in Data Breach Response, PII & PHI Detection, Data Security, Sensitive Information Detection, Privacy, Data Subject Access Requests, Incident Response, Data Protection, GDPR, CCPA, Incident Response, and Cybersecurity.

Project Overview

This case study demonstrates the strategic use of various AI/ML models to create a comprehensive solution that not only meets classification and PII extraction needs but also ensures data privacy and regulatory compliance, streamlining the entire document and image handling process.

Project sought to automate the classification, organization, and information extraction of large volumes of documents and images. Our client's primary goals included efficient categorization of various types of documents, precise object detection, and secure extraction of Personally Identifiable Information (PII). Given the sensitivity of the data and volume of unstructured information, the project demanded high accuracy, custom fine-tuning, and advanced AI tools.



Team Makeup

1 Project Manager

1 UX/UI

6 Backend Dev

7 Frontend Dev

3 Data Sciences

4 QA

2 Cloud Infra & DevOps

2 Engineers

Project Duration

+ 5 years

Type of Service

Managed Solution

Tech Stack

- YOLO (You Only Look Once) for real-time object detection.
- NER (Named Entity Recognition) customized for PII extraction.
- PaddleOCR for OCR tasks to recognize and digitize text.
- Donut (Document Understanding Transformer) for document layout understanding and classification.
- AWS Textract for form and table data extraction.
- BERT fine-tuned for text classification and PII entity recognition.
- PHI-3 Vision specialized for detecting Protected Health Information and Personal information
- Azure Document AI for advanced document processing and entity extraction.



Objectives

- 1 Classify Images and Documents based on their content and purpose (e.g., medical forms, invoices, insurance claims, bank statements, government ids, tax forms).
- 2 Detect Objects within images, particularly government ids like ssn, passports, drivers licenses, to identify relevant sections.
- 3 Extract PII Information securely, ensuring compliance with regulatory standards like HIPAA and GDPR.
- 4 Streamline Workflow Automation by enabling seamless integration across the AI/ML models.

Key Challenges Presented

- 1. Data Sensitivity:** Securing data was challenging. Models were deployed in a secure cloud environment, with access controls and encryption protocols.
- 2. Accuracy Demands:** Fine-tuning each model was labor-intensive but was addressed by leveraging synthetic datasets and manual labeling.
- 3. Handling Diverse Document Formats:** Complex layouts posed challenges for document parsing. Using a combination of Donut, Textract, and PaddleOCR, the system achieved robust adaptability across formats.

The Solution

1. Data Ingestion & Preprocessing

- Data Collection: Diverse datasets of medical records, forms, images, and handwritten notes were ingested.
- Preprocessing: Images were resized and enhanced for OCR clarity. Documents were categorized by type for more accurate model fine-tuning.

2. Image and Document Classification

- Donut (Document Understanding Transformer): Using Donut, documents were categorized by type (e.g., insurance forms, prescriptions, invoices, bank statements, government ids, tax forms). The model was fine-tuned on a custom dataset for improved accuracy.
- BERT: BERT was used to analyze document content to further classify documents by specific topics based on detected text patterns.

3. Object Detection with YOLO

- Custom Training: YOLO was trained on images to recognize objects like government ids, forms
- Application in Document Workflows: YOLO helped identify relevant document sections, such as patient information headers, and detected medical icons or stamps that signified importance.

4. OCR and Text Extraction

- PaddleOCR and AWS Textract: For clear text recognition, PaddleOCR handled document images with text, while AWS Textract extracted data from forms and tables with high accuracy. Textract was also optimized for different languages and layouts, critical for documents from international sources.
- PHI-3 Vision: This was employed specifically for PII detection, and to find relation between extracted PII

5. PII Detection and Redaction

- NER and BERT for PII Identification: A customized Named Entity Recognition (NER) model, integrated with BERT, helped identify PII entities (names, addresses, phone numbers) within text blocks. Fine-tuning was done on a dataset labeled with specific PII entities to improve detection accuracy.
- Azure Document AI: Document AI identified and extracted entities from complex document layouts, helping to process and redact PII from difficult forms, hand-written notes, and tables.
- Custom Logic for PII Redaction: PII entities were flagged, and an automatic redaction module was applied to all flagged areas in compliance with data protection requirements.



Results

- **High Accuracy Classification:** Custom-tuned Donut and BERT models achieved classification accuracy of over 95%, enabling faster sorting and processing.
- **Efficient Object Detection:** YOLO detected objects with 97% accuracy, significantly improving the workflow where section detection in medical imagery was needed.
- **Robust PII Extraction and Redaction:** The combination of NER, PHI-3 Vision, and Azure Document AI enabled 98% accurate PII detection, ensuring regulatory compliance.
- **Reduced Processing Time:** Automation with OCR and AI reduced document processing time by 70%, freeing up resources.
- **Improved Data Security:** Custom models for PII recognition ensured data privacy standards were upheld, reducing data breach risks.

Future Work

- **Enhanced Multilingual Support:** Fine-tuning for additional languages to support global document processing.
- **Model Optimization for Edge Devices:** To enable on-premise processing in healthcare facilities, optimizing models for lightweight deployment on edge devices is a potential next step.

Contact Us

athenaworks.com

