MediTrust AI

1. Team Details

Team Name: Out Of Funds

Team Members: 1) Sarthak Rastogi(RA2311026010862)

2) Utpal Raj(RA2311056010003)

3) Naysa Madrele(RA2311026010801)

4) Keshav Tripathi(RA2311026010818)

2. Problem Understanding and Scope

Problem Statement (in your own words):

Healthcare workflows rely on semi-structured, scanned documents such as prescriptions, discharge summaries, consent forms, insurance claims, and lab reports. These carry linked visual data: a mixture of printed text, handwritten content, stamps, QR/barcodes, and critically handwritten signatures. Sharing such documents across pharmacies, insurers, referrals, or research introduces privacy and fraud risks. The challenge is to automatically detect and protect personally identifiable information (PII) particularly signatures while preserving clinical content and enabling authenticity checks. Here, "linked visual data" means any document where textual and visual cues together convey identity or authorization (e.g., signatures beside names/stamps).

Target Documents & Formats:

• Prescriptions, consent forms, discharge summaries, referral letters, insurance claim forms, lab reports.

• Scanned PDFs (multi-page), TIFF, JPEG/PNG; camera captures of paper documents.

Types of Identifiable Data (PII) to Detect & Redact:

• Handwritten signatures (patient/attendant/doctor), initials, and countersignatures.

• Patient identifiers: name, date of birth/age, medical record number (MRN), phone/email, address.

• Government IDs (e.g., Aadhaar/SSN), policy numbers, UHID, insurance IDs.

• Doctor identifiers: name/registration number when required to be hidden in sharable copies.

• Faces in embedded photos, hospital stamps/seals, barcodes/QR codes carrying identifiers.

• Dates and geo-markers when policy requires temporal deidentification.

User Personas / End Users:

• Hospital HIM/Compliance teams and medical records departments.

• Clinicians and referral coordinators sharing inter-facility records.

• Insurance TPAs and fraud investigation units processing claims.

• Clinical research/ethics committees, trial sites, and data managers.

• E-health platforms, healthtech startups, diagnostic labs, and pharmacies.

3. Proposed Solution & Approach

High-Level Architecture (with Diagram):
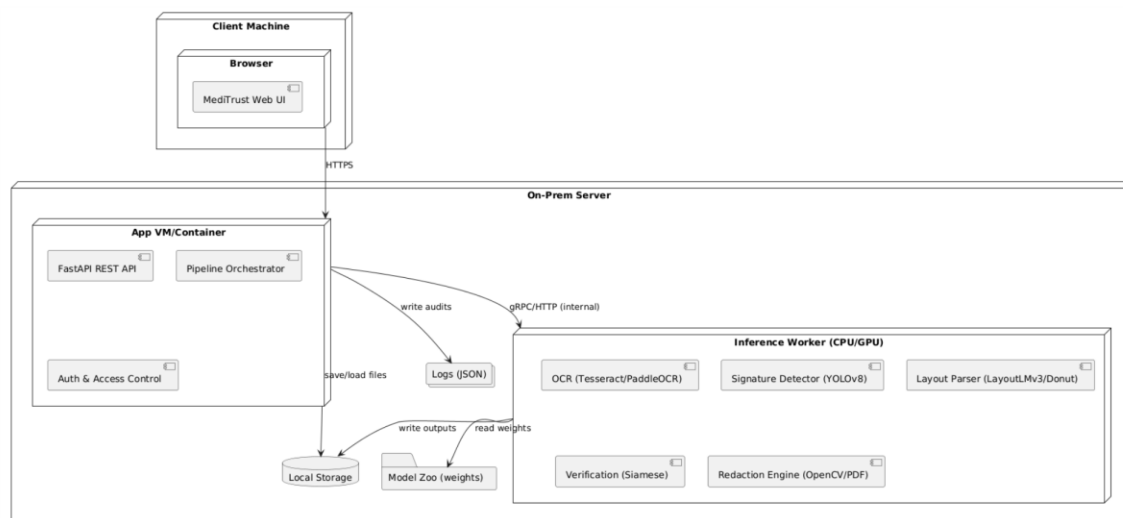
Processing flow description: 1) Ingest PDF/image

 2) Preprocess (deskew, denoise, binarize)

 3) OCR & layout parsing

4) Signature/stamp/QR detection

 5) Context-aware role classification (patient vs doctor)

 6) Policy engine applies redaction/verification

7) Generate privacy-safe PDF + logs.

AI/ML Models Considered:

• Signature detection: YOLOv8/Faster R-CNN fine-tuned on SignVerOD (victordibia/signverod).

• Layout understanding: LayoutLMv3 or Donut for region/section semantics; layoutparser for heuristics.

• OCR: Tesseract or PaddleOCR with multilingual packs (for Indian regional scripts if needed).

• Signature verification (optional): Siamese CNN embeddings for intra-doctor matching when enabled.

• Stamp/face/QR: classical CV (OpenCV), ZXing for bar/QR decoding; lightweight face detector (e.g., RetinaFace).

• PII text NER (optional): spaCy / HuggingFace transformers (ClinicalBERT/BioBERT) for textual identifiers.

Data Strategy:

• Train signature localization on SignVerOD; hold-out set for robust evaluation.

• Augment with document transforms: scaling, rotation/deskew, compression artifacts, ink bleed, low-light/phone captures.

• Synthetic overlays for stamps/seals/QRs and randomized patient/doctor text to expand coverage.

• Human-in-the-loop review via Label Studio/Doccano; active learning to prioritize uncertain detections.

• Evaluation: mAP for detection; precision/recall on redaction accuracy; latency on CPU-only inference.

Innovation / Unique Selling Point (USP):

• Role-aware signature handling: distinguish patient vs doctor signature by page region, nearby keywords, and layout.

• Privacy-first defaults: automatic redaction of patient signatures; configurable policies for doctor signatures.

• MediTrust Score: per-document confidence score combining detection certainty, policy compliance, and OCR sanity checks.

• Edge-ready: runs offline/on-prem with no third-party data egress; modular pipeline with CLI/API/UI.

4. Proposed Solution & Approach

Intended UI/UX Design (if applicable):

• Drag-and-drop web UI (or desktop) to upload PDFs/images; batch processing queue with progress.

• Preview pane shows detections (bounding boxes) and redaction result; one-click accept/correct.

• Policy selector: choose presets (Research share, Insurance share, Inter-hospital referral).

• Download: privacy-safe PDF + machine-readable JSON log of redactions and confidence scores.

Input & Output Format Expectations:

• Input: PDF (single/multi-page), TIFF, JPEG/PNG; 150–300 DPI recommended; accepts phone photos.

• Output: Redacted PDF/images; optional sidecar JSON with bounding boxes, classes, confidences, and policy applied.

• Performance targets (MVP): ≥0.8 mAP on signature detection; <2s/page on CPU; reproducible CLI + REST API.

Accessibility / Ease of Use Considerations:

• Multilingual OCR packs for Indian/English documents; UI labels in English + localizable strings.

• Color-blind safe annotations and large-text mode; keyboard shortcuts for reviewers.

• Works fully offline for compliance (HIPAA/GDPR/DPDP) with on-prem deployment option.

• Export templates aligned to insurer/research ethics committee submission formats.