FYP - Early-1

amazing — here's a **from-scratch → concrete product** guide tailored to your constraints (Jan 2026 finish, 4-person team, Python/PyTorch, open data, software-first) and aligned to your **primary 3 objectives**:

1. **High-accuracy Lung-CT nodule pipeline** (detection/segmentation + malignancy risk)
2. **Clinician-grade explainability & safety** (masks + saliency + uncertainty/human-in-loop)
3. **Grounded, multilingual reporting** (doctor report + patient-friendly summary, FHIR/ABDM-ready)

I've structured it so you can execute, track, and publish.

---

0) North-star & success gates

North-star statement

Deliver a **rural-ready, software-only** Lung-CT system that analyzes a CT study end-to-end, produces **trustworthy** clinician reports and **bilingual lay summaries**, and integrates with **ABDM/FHIR**—validated in a small pilot—by **January 2026**.

Quantitative success gates (acceptance criteria)

- **Detection/Segmentation:** Sensitivity ⩾ **95%** for nodules **⩾6 mm** at **⩽1–2 FP/scan**; per-nodule Dice ⩾ **0.80**.
- **Malignancy risk:** AUC ⩾ **0.90** (LIDC/LUNA16-style eval; external subset).
- **Explainability:** ⩾ **80%** radiologist agreement that heatmaps/masks are "helpful/consistent".
- **Reporting (doctor):** factuality error (count/size/site) **<5%** on expert review; latency **<60 s**/study on RTX 4060.
- **Reporting (patient):** bilingual readability (user study) ⩾ **80%** correct comprehension on 5 key questions.
- **Interop:** Valid **FHIR DiagnosticReport + ImagingStudy** bundles (ABDM sandbox validation).
- **Pilot UX:** operator CSAT ⩾ **4/5**; outage-recovery verified.

---

1) System blueprint (what you will build)

1.1 Pipeline overview (CT only)

- **DICOM ingest → pre-proc** (spacing normalize, lung windowing, slice packing 2.5D).
- **Nodule locator (detector):** lightweight 3D/2.5D candidate generator (e.g., Mobile-UNet-RPN or DeepSE-lite).
- **Segmenter:** 2.5D **Mobile-UNet / Half-UNet** with **light channel+spatial attention** for per-nodule masks.
- **Risk head:** small context-aware classifier (deep features + a few radiomics like sphericity, texture) → benign/malignant likelihood; optional **clinical fields** (age/sex/smoking) if

available.
- **Explainability:** Grad-CAM++ (risk head) + mask overlays + per-nodule thumbnails; **uncertainty** via MC-dropout / test-time augmentation (flag low-confidence).
- **Grounded reporting:** generate a **structured JSON** of findings (per-nodule size, lobe, type, risk, Lung-RADS).
- **LLM reporter:** from JSON → (a) **clinician report**; (b) **patient summary** in **English + Hindi** (and add 1–2 regional languages later).
- **ABDM/FHIR exporter:** ImagingStudy + DiagnosticReport bundles; consent artifacts; audit log.

1.2 Reference architecture (software)

- **Backend**: Python (FastAPI) serving models; async job queue (RQ/Celery) for study-level tasks; local SQLite/Postgres for metadata & audit.
- **Model runtime**: PyTorch (primary), ONNXRuntime (fallback); optional INT8 quantization for speed.
- **Frontend**: lightweight web UI (React or Streamlit for speed) with **offline-first** cache, large tap targets, bilingual text, print/PDF export.
- **MLOps & Repro**: DVC (data/versioning), MLflow or Weights&Biases (experiments), GitHub Actions (unit tests/CI).
- **Security/Privacy**: local processing by default; reversible anonymization; encryption at rest; consent capture.

---

2) Datasets & splits (reporting + detection/segmentation)

All open-source; curated for your 3 objectives.

Report generation (objective 3):

- **CT-RATE** (CT volumes **with** radiology reports) → train/validate clinician-style report generation.
- **LNDb v4** (CT + reports + nodule attributes alignment) → **factuality checks** (counts/sizes/sites).
- **No lay summaries exist:** create **templated bilingual** (EN↔Hindi) lay summaries from clinician reports, review 200 cases with a clinician for quality.

Detection/segmentation & risk (objective 1):

- **LIDC-IDRI** (multi-reader nodules); **LUNA16** (detection splits) → baseline & comparability.
- **Histopath-labeled subset** (where available) → hard endpoints for malignancy.
- **External holdout**: reserve 10–15% (different centers/scanners) for **final external test**.

Splits strategy

- Train/val/test per dataset; also **cross-dataset** test (e.g., train LIDC+LUNA → test LNDb subset).
- **Study-level** splitting (no patient leakage).
- Document seeds & version hashes with DVC.

---

3) Modeling plan (objective 1)

3.1 Baselines → finalists (week-by-week)

- **W1–2:** 2.5D **Mobile-UNet** seg baseline (Dice target ⩾0.75 on LIDC).
- **W2–3:** Detector: Mobile-RPN or CenterNet-lite; measure **FROC/CPM**; FP/scan goal ⩽4 (pre-refine).
- **W3–4:** Risk head (deep + radiomics) AUC target ⩾0.85 → iterate to ⩾0.90; add **uncertainty**.
- **W4–5:** Joint training (det → seg → risk) or staged with hard-negative mining.
- **W6+:** Compression (pruning/INT8 optional), batch-inference pipeline, latency profiling on RTX 4060.

3.2 Key tricks to hit targets

- **Augmentations** tuned for CT: elastic, blur, noise, motion, FOV shifts; HU clipping in [-1000, 400]; spacing normalize (1–1.25 mm).
- **Small-nodule sensitivity**: multi-scale features; attention gates kept **light**; increased positive sampling for <6 mm nodules.
- **False positives reduction**: per-candidate verifier (shallow 3D CNN) + rule-outs (vessels/pleura heuristics).
- **Generalization**: mix datasets; keep an **external site** test; calibrate thresholds per site if needed.

3.3 Metrics to report

- **Detection**: FROC curve; **CPM**; Sens@1/2 FP/scan; per-scan sensitivity.
- **Segmentation**: Dice, 95% Hausdorff, volumetric error.
- **Risk**: ROC-AUC; PR-AUC; calibration (ECE).
- **Latency**: end-to-end time per study; peak RAM/VRAM.

---

4) Explainability & safety (objective 2)

- **Visual XAI**: Grad-CAM++ on risk head; overlay with **precise masks** from segmenter; per-nodule mini-gallery (axial/coronal/sagittal).
- **Text rationales**: templated one-liners ("6.8 mm solid nodule in RUL; benign features absent; follow Lung-RADS-3").
- **Uncertainty policy**: if (entropy high OR disagreement high) ⇒ label **"Needs radiologist review"**; block lay summary alarming language.
- **Human-in-loop**: quick-edit for report sections; **red-flag** banner for urgent findings.
- **Trust study**: 2 radiologists rate 50 cases for **usefulness & correctness** of explanations.

---

5) Grounded multilingual reporting (objective 3)

5.1 Structured schema (your single source of truth)

```
json                                                    Copy  Edit
```

```
{
  "study_uid": "...",
  "nodules": [
    {"id":"N1","lobe":"RUL","type":"solid","long_axis_mm":7.2,"vol_mm3":190,
     "spiculation":false,"calcification":false,"risk":0.72,"risk_ci":[0.60,0.82],
     "uncertain": false}
  ],
  "summary": {"nodule_count":1, "largest_mm":7.2, "suspicious": true, "lung_rads":"3"},
  "quality": {"motion":"mild","noise":"moderate"},
  "recommendation": "CT follow-up in 6 months"
}
```


5.2 Reporter design

- **Clinician report (EN):** sectioned (Technique/Findings/Impression/Recommendation) **fully derived** from schema; strict templates (no LLM free-write).
- **Patient summary (EN+Hi):** simple sentences + icons; **reading level ⩽ 8th grade**; **TTS** optional.
- **Additional languages:** plug IndicTrans2/Marian + glossary; human review of 100 samples/language.
- **Factuality checks:** rule-based validators (counts/sizes/lobe consistency); unit tests on 200 cases.
- **Latency budget:** ⩽ **2 s** for generation (schema→text).

---

6) Interoperability: ABDM/FHIR

- **Bundles**: `ImagingStudy` (series/instances), `Observation` (per-nodule), `DiagnosticReport` (conclusion + references).
- **Identifiers**: placeholder ABHA/OTP flow (sandbox); consent status; provenance/audit.
- **Exporter**: one-click **JSON** + printable **A4 PDF**; offline store-and-forward queue.

---

7) Product & UX (rural-ready)

- **Offline-first:** local job queue; resumable after power/net loss; **USB export** fallback.
- **Kiosk UI:** 4-step wizard (Identify → Load CT → Analyze → Share); large buttons; bilingual toggle; high-contrast theme.
- **Operator aids:** progress bar; typical 30–60 s per study; clear errors; "send to specialist" with status.
- **Privacy & DPDP:** on-device by default; encryption at rest; consent capture with plain-language notice.

---

8) Execution plan to Jan 2026 (6 sprints, 4 people)

Team (fixed roles):

- **ML-A (Vision lead)** – detector/segmenter/risk, latency.

- **ML-B (Reporting/NLP)** – schema, reporter, multilingual, validators.
- **Data/MLOps** – DICOM, preproc, splits, DVC/MLflow, evaluation harness, FHIR export.
- **Frontend** – kiosk UI, viewer/overlays, offline queue, PDF/print.

Sprint 1 (Aug 19 – Sep 15, 2025): Foundations

Goals: data pipeline, baselines, metrics, schema.

- Data/MLOps: DICOM→NIfTI; spacing normalize; slice packing; DVC repo; MLflow set-up; draft FHIR mapping.
- ML-A: Mobile-UNet 2.5D seg baseline (Dice ⩾0.75); candidate detector skeleton; augmentations library.
- ML-B: lock **schema**; clinician report templates; draft lay-summary style guide; 50 bilingual samples.
- Frontend: Figma flows; FastAPI boilerplate; basic web viewer (axial scroll, window/level).
  Milestone: first end-to-end dry run: **1 study → schema → draft report** (no risk yet).

Sprint 2 (Sep 16 – Oct 15): Core models v1

Goals: reliable detection/seg; initial risk; XAI.

- ML-A: detector v1 (Sens@2 FP ⩾85%); seg Dice ⩾0.80; per-candidate risk head v0 (AUC ⩾0.85).
- ML-B: Grad-CAM++ over risk head; clinician report v1 (template-filled); factuality validator v0.
- Data/MLOps: eval harness (FROC/CPM, Dice, AUC); external holdout carved; latency profiler on 4060.
- Frontend: overlays (mask + heatmap); bilingual switch; offline job queue skeleton.
  Gate: end-to-end latency **⩽90 s**/study; factuality errors ⩽10% on 50-case set.

Sprint 3 (Oct 16 – Nov 15): Quality & multilingual

Goals: lift accuracy; reduce FP; Hindi live.

- ML-A: hard-negative mining; small-nodule boosts; FP/scan down to **⩽2** at target sens.
- ML-B: Hindi reporter v1; glossary; **validator v1** (counts/size/site/side); TTS PoC.
- Data/MLOps: cross-dataset test (train LIDC+LUNA → test LNDb subset); metrics dashboard.
- Frontend: kiosk wizard; PDF/print; **ABDM/FHIR exporter** (sandbox JSON).
  Gate: Sens (⩾6 mm) **⩾95%** @ ⩽2 FP/scan; factuality **⩽5%**; latency **⩽60 s**.

Sprint 4 (Nov 16 – Dec 10): Hardening & usability

Goals: trust & safety; clinician UX.

- ML-A: uncertainty flags; calibrate thresholds; segmenter refinement; save time by skipping seg for obvious negatives.
- ML-B: patient summary tone/guardrails; red-flag phrasing; regenerate/quick-edit tools.
- Data/MLOps: ABDM sandbox validation (bundle acceptance); consent/audit logging.
- Frontend: outage-recovery; big tap targets; print layouts; feedback button.
  Gate: XAI user study (2 rads, 50 cases) ⩾80% helpfulness; ABDM bundles validate.

Sprint 5 (Dec 11 – Jan 05): Pilot prep & documentation

Goals: stability; test packs; docs.

- Data/MLOps: freeze datasets/splits; versioned models; reproducible training script; benchmark sheet.
- ML-A/ML-B: error triage; finalize thresholds; save calibrated model.
- Frontend: installation script; crash-safe local store; operator quick-guide (Hindi/EN).
- All: draft paper (methods, metrics, usability), proposal/presentation.
  Gate: release candidate v1.0; install + run on clean machine in **<30 min**.

Sprint 6 (Jan 06 – Jan 31): Field validation & submission

Goals: mini-pilot, final report, demo.

- Run **prospective** mini-pilot (even 10–20 retrospective clinic studies if live access not possible).
- Collect operator/patient feedback; finalize metrics; polish paper/manuscript.
- Deliverables: final demo, code repo (clean), dataset cards, model card, compliance notes.

---

9) Engineering checklists

9.1 Data governance

- Strip PHI; hash IDs; document licenses; dataset cards (source, label quality, splits).
- DVC tracked raw→processed transforms; reproducible notebooks.
- Bias notes: scanner types, slice thickness, demographics if available.

9.2 Code quality

- Unit tests for: preprocessors, schema validators, reporter templates, FHIR exporter.
- Lint/format; type hints; CI on push; deterministic seeds.

9.3 Performance budget (RTX 4060)

- Target **⩽60 s**/study end-to-end.
- Profile: IO (DICOM read) ⩽10 s; model stack ⩽40 s; report/exports ⩽10 s.
- If needed: mixed precision; smaller crops; batch slices; async IO.

---

10) Evaluation protocol (publication-ready)

- **Primary**: FROC/CPM; Sens@1 & @2 FP/scan; Dice; risk AUC; calibration; latency.
- **External test**: different center/scanner subset.
- **Ablations**: (a) no attention; (b) no radiomics; (c) single-scale vs multi-scale; (d) uncertainty off.
- **User studies**:
  - Radiologist panel: 50 cases, **helpfulness** of XAI (Likert) + factuality audit.
  - Patient/technician: 20 people, **readability** & **comprehension** of summary (Hindi/EN).
- **Stats**: 95% CIs via bootstrapping; McNemar for paired error comparisons.

---

11) Novelty vs existing systems (your positioning)

- **Software-only, rural-ready**: offline queue, outage recovery, low-touch install—rare in hospital-centric products.
- **Grounded bilingual reporting**: strict schema → clinician & lay outputs + TTS; measurable **factuality**.
- **Transparent trust**: masks + Grad-CAM++ + uncertainty → **human-in-loop** policy.
- **ABDM/FHIR by default**: export bundles that plug into India's rails.
- **Evidence package**: external dataset test + usability studies—often missing in prior papers.

---

12) Risk register (with mitigations)

Risk	Impact	Mitigation
Small-nodule sensitivity <95%	Missed cancers	Multi-scale features; raise native resolution for candidates; threshold calibration; prioritize recall then filter FP.
FP/scan too high	Workflow overload	Secondary verifier; vessel suppression; site-specific thresholds; uncertainty gating.
LLM hallucination in reports	Unsafe text	**Template-first** generation; validators; blocklist; keep free-text minimal.
Latency >60 s	Poor UX	Mixed precision; crop-then-segment; parallel slice batches; optional INT8; skip seg for obvious negatives.
Language quality	Miscommunication	Human review loop; glossary; back-translation checks; add audio TTS.
Data heterogeneity	Gen-gap	Cross-dataset training; external holdout; calibration per site.

---

13) Publication & dissemination plan

- **Venue strategy** (pick 1–2): IEEE JBHI / Scientific Reports (application-oriented), EJR-Open / BMC Med Imaging (clinical imaging), or a MICCAI workshop (methods).
- **Paper outline**: Introduction (rural need) → Methods (edge-lean stack, schema-grounded reporting) → Results (FROC/Dice/AUC, latency, XAI usability, bilingual comprehension) → ABDM/FHIR appendix + code links.
- **Artifacts**: anonymized test splits, config files, model card, demo video.

---

14) Work breakdown (RACI quick view)

Workstream	ML-A	ML-B	Data/MLOps	Frontend
Preprocess & splits	C	C	**R**	I
Detector/Seg/Risk	**R**	C	C	I
Explainability & uncertainty	**R**	C	C	I
Schema & clinician report	C	**R**	C	I
Patient summary (multilingual)	I	**R**	C	C
Eval harness & dashboards	C	C	**R**	I
FHIR/ABDM exporter	C	C	**R**	C
UI & viewer	I	C	C	**R**
Offline queue/print	I	I	C	**R**
Docs & paper	C	**R**	R	C

(R=Responsible, A=Accountable, C=Consulted, I=Informed)

---

15) What to do this week (to start strong)

1. **Create repos** (app, models) + enable DVC & MLflow; set code style & CI.
2. **Lock dataset plan**; download LIDC/LUNA, LNDb v4, CT-RATE; define splits; write dataset cards.
3. **Implement preprocess** pipeline; push a first "study→slices" notebook with hashes.
4. **Train Mobile-UNet seg baseline** for 3 days; log Dice; save weights.
5. **Draft schema** + doctor template + 10 bilingual lay summaries; write validator v0.
6. **Skeleton UI**: viewer with axial scroll; load dummy overlays; define routes.
7. **Schedule weekly demo**; open a public roadmap doc shared by the team.

If you want, I can also generate:

- a **Gantt** (Aug→Jan) and a printable **RACI** PDF,
- the **FHIR JSON** stubs and a **report template** pack,
- or a ready-to-run **cookiecutter repo** (FastAPI + PyTorch + Streamlit + DVC/MLflow) so you can code from hour 1.