**Problem-1 : Robust & Generalizable Models**

To design **hybrid 2D/3D deep learning and ensemble frameworks** that mitigate **data scarcity, imbalance, and domain shift**, thereby improving the **generalization and stability** of COVID-19 CT classification across diverse datasets and institutions.

Step:1

to design a **hybrid 2D/3D deep-learning + ensemble** system that tackles **data scarcity, imbalance, and domain shift** for COVID-19 CT classification across institutions.

**1) Data pipeline & harmonization (domain shift starts here)**

* **Site-aware splits:** Do *leave-one-center-out (LOCO)* CV: train on N−1 hospitals, test on the held-out hospital. Also keep an *internal* validation per site.
* **Voxel spacing & slice thickness:** Resample to a common in-plane spacing (e.g., 1.0–1.25 mm) and cap the through-plane spacing by slice selection (e.g., 1.5–2.5 mm effective).
* **Lung & lesion masks:** Use a robust pre-trained U-Net to segment lungs; optionally segment lesions (weak labels OK). Crop to lung bounding box to reduce scanner bias.
* **Intensity standardization:** Z-score within lung mask + fixed windowing (e.g., −1,000 to 400 HU) and histogram matching to a *reference site* (optional).
* **Quality control:** Flag outliers (truncated FOV, metal artifacts); route to a noisy-label mitigation step (below).

**2) Label noise mitigation (stability under scarcity)**

* **Co-teaching or MentorNet:** Train two 2D slice models; each filters the other’s likely noisy samples (small-loss trick).
* **Consensus relabeling:** For cases where RT-PCR/clinical labels disagree with imaging phenotype, use a teacher ensemble to generate soft labels; mix with hard labels (α≈0.7).

**3) Hybrid 2D/3D model design (core idea)**

**A. Per-slice 2D encoder (rich texture)**

* Backbone: ResNet-50/101 or small ViT (DeiT-S), **pretrained with medical SSL** (e.g., MoCoD, SimCLR-Med, masked autoencoder on chest CT slices).
* Output: slice embeddings ei∈Rde\_i \in \mathbb{R}^dei​∈Rd + slice logits.

**B. 3D context encoder (shape & volumetrics)**

* Backbone: Lightweight 3D ResNet-18/RegNet-3D; input = downsampled volume or fixed-depth cubes (e.g., 64 slices with stride).
* Output: volume embedding v∈Rdv \in \mathbb{R}^dv∈Rd.

**C. Slice-to-study aggregation (MIL)**

* Use **Attention-based MIL** or a **Transformer** over ordered slice embeddings {ei}\{e\_i\}{ei​} (axial “sequence”). Add **positional encodings** for slice index and **site tokens** for scanner/domain metadata (if available).

**D. Fusion (2.5D/3D)**

* Concatenate [AttnPool({ei}),v]→f∈R2d[ \text{AttnPool}(\{e\_i\}), v ] \rightarrow f \in \mathbb{R}^{2d}[AttnPool({ei​}),v]→f∈R2d and pass through a small MLP for study-level logits.
* Optional **cross-attention**: Query = 3D embedding, Keys/Values = slice embeddings (lets 3D context select informative slices).

**4) Losses to fight imbalance & shift**

* **Primary:** Focal loss (γ=2) or **Class-balanced (effective number)** loss.
* **Calibration:** Add **Label Smoothing** (ε=0.05) and **Temperature Scaling** post-hoc.
* **Domain-generalization (DG) regularizers (no target data needed):**
  + **MixStyle** in feature space (randomize instance statistics) and/or **AugMix**/**RandAugment** in image space.
  + **CORAL / MMD loss** across mini-batches stratified by site (align second-order stats).
  + **IRM / GroupDRO** using site as environment to avoid learning site-specific shortcuts.
* **Consistency:** Mean-Teacher or EMA teacher with **consistency loss** under strong augmentations (stabilizes low-data regimes).

**5) Data scarcity remedies (use them together)**

* **Self-supervised pretraining:** Pretrain the 2D and 3D encoders on *unlabeled* CTs (even non-COVID) with masked modeling or contrastive instance discrimination.
* **Weak/semi-supervision:** Pseudo-label unlabeled studies using the teacher ensemble; only accept high-confidence cases (p>0.9) with entropy thresholding.
* **Generative augmentation (optional but powerful):** Class-conditioned diffusion to synthesize lesion-augmented patches; gate with a realism classifier to avoid drift.
* **Radiomics+DL hybrid:** Concatenate compact radiomics (e.g., 64–128 features from lesion masks) with learned embeddings; this often boosts generalization with little data.

**6) Robust ensembling (diversity by design)**

Build diversity along four axes and **stack** them:

1. **Architectures:** ResNet vs ConvNeXt vs ViT (2D); 3D-ResNet vs 3D-DenseNet (3D).
2. **Input views:** Different HU windows (lung/mediastinal), slice strides, and crop paddings.
3. **Training criteria:** Focal vs class-balanced; with/without DG regularizers.
4. **Seeds / snapshots:** Snapshot ensembles or SWA for cheap diversity.

* **Level-1:** Average logits within each family (2D-MIL family, 3D family).
* **Level-2 Stacking:** Train a **meta-learner** (ridge-logistic) on held-out folds using Level-1 outputs + simple metadata (age/sex/site) if allowed.

**7) Domain adaptation (if you can see target site unlabeled data)**

* **UDA (unsupervised DA):**
  + **SHOT / AdaBN:** Freeze classifier head; adapt batch-norm stats to target site and optimize information maximization on target.
  + **DANN / CDAN:** Adversarial aligner head to make features site-invariant.
  + **Self-training:** EMA teacher generates target pseudo-labels with confidence ramp-up; apply **class-balanced sampling** to avoid collapse.
* **Test-time adaptation:** TENT (entropy minimization) on BN affine parameters per test batch; cache per-site stats.

**8) Training recipe (what to actually run)**

* **Curriculum:** Start with 2D slice pretraining (SSL), then supervised on slices; next add MIL/Transformer aggregator; finally add 3D stream and fusion.
* **Balanced mini-batches:** Site-aware, class-aware batching (e.g., 4 studies × 2 sites × balanced class if possible). If not, use **Deferred Re-weighting** (increase minority weight after warmup).
* **Augmentations that don’t break anatomy:** Random crop within lung box, rotation ≤10°, elastic small deforms, Gaussian noise, blur, intensity jitter; avoid flips that change laterality meaning if laterality is used.
* **Optimization:** AdamW, lr 3e-4 (2D) / 1e-4 (3D), cosine decay, warmup 5 epochs. Mixed precision, gradient clipping (1.0).
* **Early stopping:** Monitor **macro-AUC** on the *held-out site* (not the internal val) to prevent overfitting to source domains.

**9) Evaluation for generalization & stability**

* **Primary metrics:** AUROC (macro), AUPRC, sensitivity@fixed specificity (0.90), **ECE** (calibration), **Brier score**.
* **Across-site fairness:** Report per-site AUROC, and the **range / std across sites** (lower is better stability).
* **Robustness checks:**
  + Common corruptions (noise/blur/compression) at multiple severities.
  + **OOD detection**: energy score or Mahalanobis on penultimate features; report FPR@95%TPR for OOD vs in-dist.
* **Ablations (make them small but decisive):**
  + 2D only vs 3D only vs **2D+3D hybrid**.
  + No DG vs CORAL vs MixStyle vs both.
  + Loss variants (CE vs Focal vs Class-Balanced).
  + With vs without radiomics fusion.
  + With vs without test-time adaptation.

**10) Deployment hygiene (so it’s clinically usable)**

* **Uncertainty-aware outputs:** Use **deep ensembles** (K=5) or MC-Dropout at test; with **temperature scaling**; expose confidence + calibrated probabilities.
* **Case-level explanations:** Attention heatmaps over slices (MIL attention), Grad-CAM for decisive slices; store with predictions for audit.
* **Monitoring:** Per-site drift via BN stats or feature embeddings; trigger TTA or scheduled re-calibration.
* **Governance:** Follow **TRIPOD-AI / CONSORT-AI** reporting; document site distributions, scanners, and all preprocessing.

**Assignment : Modules to Students**

1. **Preprocess & Segment** lungs; resample; normalize; dataset cards per site.
2. **SSL Pretrain** (2D & 3D) on unlabeled CTs.
3. **Supervised 2D Slice Model** (+co-teaching for noisy labels).
4. **MIL Aggregator** (Transformer with site tokens).
5. **3D Stream** (light 3D ResNet) + **Fusion**.
6. **DG Training** (MixStyle + CORAL + Focal/Class-balanced).
7. **Ensemble & Stacking** across architectures & windows.
8. **UDA/TTA** when target site unlabeled data appear.
9. **Evaluation** (LOCO, calibration, robustness, OOD).
10. **Docs & Model Card** (TRIPOD-AI checklist, ablations).

**Problem-2 : Explainability & Clinical Trust**

To integrate **explainable AI (XAI) methods** (e.g., Grad-CAM, SHAP) with classification models, enabling **transparent, interpretable, and clinically verifiable outputs** that support radiologists in decision-making and improve trust in AI-assisted diagnosis.

# 1) What you’ll explain (and how)

* **Study-level class decision** → heatmaps on representative slices/3D volume + a short textual rationale (e.g., “bilateral peripheral GGOs drove +COVID score”).
* **Slice/region contributions** → localized saliency (Grad-CAM/Grad-CAM++/LayerCAM) + **lesion-overlap stats** against masks (if available).
* **Feature-level contributions** (for hybrid DL+radiomics or metadata) → **SHAP/TreeSHAP** bars/tables.
* **Model components** → MIL/Transformer **attention rollout** to show which slices/instances were decisive; prototype nearest neighbors (if using ProtoPNet-style heads).

# 2) XAI toolbox (choose per module)

**2D slice encoders**

* Grad-CAM/Grad-CAM++ on last conv block; Guided-Grad-CAM for sharper edges.
* Integrated Gradients for sanity (robust to gradient saturation).

**3D encoders**

* **3D-CAM**: Grad-CAM in 3D then **MIP** to coronal/sagittal or per-slice reformat; optionally 3D connected-component pruning to remove noise.

**MIL/Transformer aggregators**

* **Attention maps** over slice tokens (show top-k slices + their CAMs).
* **Attention rollout** across layers to avoid single-head bias.

**Fusion with tabular/radiomics**

* **SHAP** for tree/linear meta-learners; KernelSHAP (small curated background set) for neural heads.

**Concept-level**

* **TCAV** concepts (e.g., “peripheral GGO”, “crazy-paving”): build concept datasets with radiologist help; report **TCAV scores** per class.

**Counterfactuals (optional but powerful)**

* Diffusion/VAEs to generate minimally edited volumes/slices that flip the prediction; constrain edits to lung mask to keep anatomy plausible.

# 3) Rendering pipeline for clinicians

1. **Select slices**: top-k attention slices + a coverage set (apices/base/mid-lung).
2. **Overlay**:
   * CAM heatmap on windowed CT (alpha-blended, fixed colorbar).
   * Contours of **lesion masks** (if available) to visually assess overlap.
3. **Panel outputs**:
   * Study-level class prob + **calibrated confidence** (temperature scaling).
   * **Top features (SHAP)**: small bar chart (e.g., Age↑, LesionVolume↑, SpO₂↓).
   * **Evidence tiles**: (slice thumbnail, CAM, short caption).
4. **Export**:
   * DICOM **Seg** (binarized salient regions) + **DICOM SR** (quant tables: %lung covered, lesion overlap, TCAV, attention-slice indices) for PACS.
   * PDF one-pager for tumor board / audit trail.

# 4) Faithfulness & robustness checks (make explanations trustworthy)

* **Sanity checks (Adebayo)**: randomize weights → CAMs must degrade.
* **Sufficiency/Comprehensiveness**:
  + Mask-IN salient regions ⇒ prob should stay high (sufficiency).
  + Mask-OUT salient regions ⇒ prob should drop (comprehensiveness).
* **Insertion/Deletion curves**: progressively reveal/remove pixels by saliency; AUC↑ is better.
* **Pointing-game/IoU** with radiologist lesion masks: % heatmap max inside lesion; IoU@τ.
* **Stability**: explanation similarity (SSIM/IoU) across augmentations and across ensemble members.
* **Uncertainty-aware XAI**: aggregate CAMs over a **deep ensemble** or MC-Dropout; show **variance map** (shaded overlay) to flag unreliable explanations.

# 5) Clinical verification protocol (radiologist-in-the-loop)

* **Blinded review**: 50–100 studies/site. Radiologists mark whether highlighted regions are clinically plausible (Likert 1–5).
* **Signal vs artifact audit**: count explanations dominated by vessels, ribs, bed artifacts → rate as “spurious”.
* **Decision support utility**: measure Δ in **reader study** (AUC, time-to-decision) with/without XAI.
* **Discrepancy workflow**: button to flag “misleading heatmap”; flagged cases feed a **hard-example buffer** for periodic retraining or explanation penalty.

# 6) Training-time tweaks that improve explainability

* **CAM-aware regularization**: constrain saliency to lung mask (total variation loss + outside-lung penalty).
* **Prototype heads**: add a small **prototype layer** (ProtoPNet-style) whose prototypes are real slices/patches; show nearest training patches as “this looks like…”.
* **Attention guidance**: if lesion masks available, weakly supervise attention (KL to soft lesion maps).

# 7) Governance & reporting (for TRIPOD-AI/CONSORT-AI)

* Document: models, layers used for CAM, background set for SHAP, thresholds, calibration method.
* Log with each prediction: model version, confidence, explanation hashes, metrics (sufficiency/comprehensiveness), and data provenance (site/scanner).
* **Bias checks**: per-site/per-scanner explanation overlap with lesions; alert if a site shows off-lesion saliency drift.

# 8) Minimal implementation plan (do this in order)

1. **Hooks**: register Grad-CAM/IG hooks on 2D & 3D backbones; wrap MIL attention extraction.
2. **Visualizer**: consistent colorbar, windowing presets (lung: −1000→400 HU), slice grid.
3. **Metrics**: implement pointing-game, sufficiency/comprehensiveness, insertion/deletion curves; add to validation step.
4. **SHAP**: compute once per study for fusion features; cache summaries.
5. **Exporter**: DICOM Seg/SR writer + PDF report generator.
6. **Clinician UI**: compact viewer with (a) predictions, (b) evidence tiles, (c) SHAP bar, (d) download DICOM SR/Seg.

**M1. Data governance & curation**

* **Objective:** Build de-identified, site-aware datasets and splits.
* **Tasks:** DICOM→NIfTI, de-ID checks, site stratification, LOCO splits, dataset cards.
* **I/O:** In: raw CTs + labels; Out: curated dataset + JSON metadata.
* **Skills:** DICOM, Python, ethics/compliance.
* **AC:** Reproducible splits; dataset card with site counts, slice thickness histograms.

**M2. Preprocessing & harmonization**

* **Objective:** Standardize voxel spacing, windowing, lung ROI.
* **Tasks:** Resample (e.g., 1.25×1.25×2.0 mm), lung segmentation (pretrained U-Net), crop to lung bbox, HU window presets.
* **I/O:** In: curated scans; Out: preprocessed volumes + lung masks.
* **Skills:** SimpleITK/Monai, NumPy.
* **AC:** <2% spacing deviation; QC notebook with failure flags.

**M3. Baseline 2D slice classifier**

* **Objective:** Strong slice model to feed MIL.
* **Tasks:** ResNet/ConvNeXt training, focal/class-balanced loss, label smoothing.
* **I/O:** In: slices; Out: slice logits + embeddings.
* **Skills:** PyTorch/Lightning.
* **AC:** AUROC↑ vs CE baseline; training script with config file.

**M4. MIL/Transformer study-level aggregator**

* **Objective:** Convert slice evidence to study decision.
* **Tasks:** Attention MIL / small ViT over slice embeddings, positional encodings, site token.
* **I/O:** In: slice embeddings; Out: study logits + attention weights.
* **Skills:** Transformers, batching.
* **AC:** Gains on LOCO validation; attention heatmaps for top-k slices.

**M5. Lightweight 3D encoder**

* **Objective:** Capture volumetric context.
* **Tasks:** 3D-ResNet-18 on downsampled stacks; mixed precision; gradient clipping.
* **I/O:** In: 3D volumes; Out: 3D embedding + logits.
* **Skills:** 3D convs, memory management.
* **AC:** Adds ≥0.02 AUROC when fused with M4.

**M6. Hybrid fusion (2D+3D)**

* **Objective:** Fuse MIL (2D) + 3D streams.
* **Tasks:** Concat/cross-attention fusion; small MLP head; calibration (temperature scaling).
* **I/O:** In: MIL + 3D embeddings; Out: fused prediction.
* **Skills:** Model wiring, calibration.
* **AC:** Best macro-AUROC; ECE ≤ 0.05 on held-out site.

**M7. XAI—saliency & attention (Grad-CAM family)**

* **Objective:** Clinically useful heatmaps.
* **Tasks:** Grad-CAM/++ hooks for 2D & 3D; attention rollout; variance maps over ensemble.
* **I/O:** In: trained models; Out: heatmaps per slice/volume + attention indices.
* **Skills:** Backprop hooks.
* **AC:** Sanity-check pass (Adebayo randomization); visual panel with fixed colorbar.

**M8. XAI—feature attribution (SHAP) & prototypes**

* **Objective:** Explain tabular/radiomics/meta features and show “this looks like…” examples.
* **Tasks:** TreeSHAP for meta-learner; KernelSHAP with curated background; ProtoPNet-style nearest-neighbor tiles.
* **I/O:** In: features/embeddings; Out: SHAP bars + prototype tiles.
* **Skills:** SHAP, metric learning.
* **AC:** Stable SHAP ordering across bootstraps; prototype gallery with indices.

**M9. Explanation validity metrics**

* **Objective:** Quantify trustworthiness.
* **Tasks:** Comprehensiveness/sufficiency, insertion/deletion curves, pointing-game/IoU vs lesion masks.
* **I/O:** In: heatmaps + labels/masks; Out: metric reports.
* **Skills:** Evaluation design.
* **AC:** ≥25% prob drop (comp), ≥80% prob retain (suff), pointing-game ≥85%.

**M10. Robustness & domain-generalization**

* **Objective:** Improve cross-site stability.
* **Tasks:** MixStyle/AugMix, CORAL loss, GroupDRO; stress tests (noise/blur/compression).
* **I/O:** In: training loops; Out: DG-trained checkpoints + robustness curves.
* **Skills:** Regularizers, data aug.
* **AC:** Reduced per-site AUROC std; robustness curve AUC↑ vs baseline.

**M11. Clinician-facing report & DICOM export**

* **Objective:** Make results PACS-friendly.
* **Tasks:** Generate PDF one-pager (probabilities, evidence tiles, SHAP bar), DICOM-Seg (salient regions), DICOM-SR (tables).
* **I/O:** In: predictions + XAI; Out: PDF, DICOM Seg/SR.
* **Skills:** pydicom, reporting.
* **AC:** Radiologist can open Seg/SR; report prints on A4 cleanly.

**M12. Reproducibility, model cards & governance**

* **Objective:** Make it auditable and publishable.
* **Tasks:** Hydra/ConfigArgParse, seed control, wandb logs, TRIPOD-AI checklist, model card, data provenance.
* **I/O:** In: project repo; Out: docs + “reproduce.sh”.
* **Skills:** MLOps, documentation.
* **AC:** Fresh clone reproduces metrics within ±1% AUROC.