Learning from Disagreement:
Curating Multi- Annotators Labels into Probabilistic Supervision

Haaroon Afroz Ognawala [1], David Tschirschwitz [1], Claudio R. Jung [2], Volker Rodehorst [1]

Bauhaus University, Weimar, Germany [1] | Universidade Federal do Rio Grande do Su, Brazil [2]

ECCV 2026 | MALMÖ | SEPT 8–12

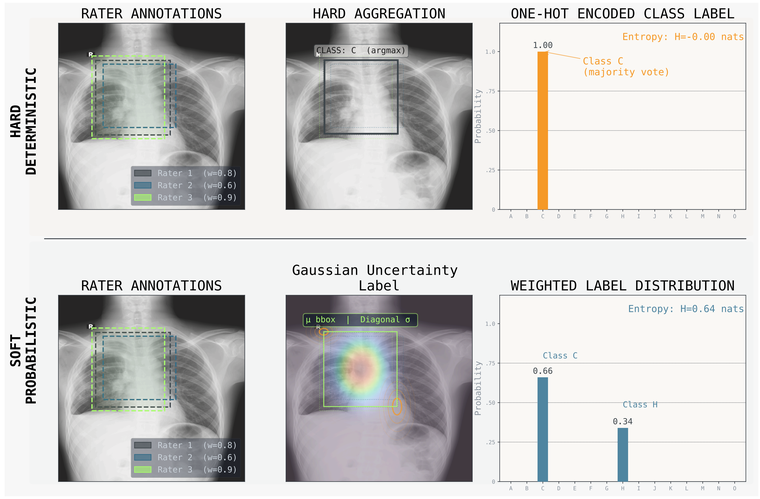
Bauhaus-Universität Weimar

1. Motivation

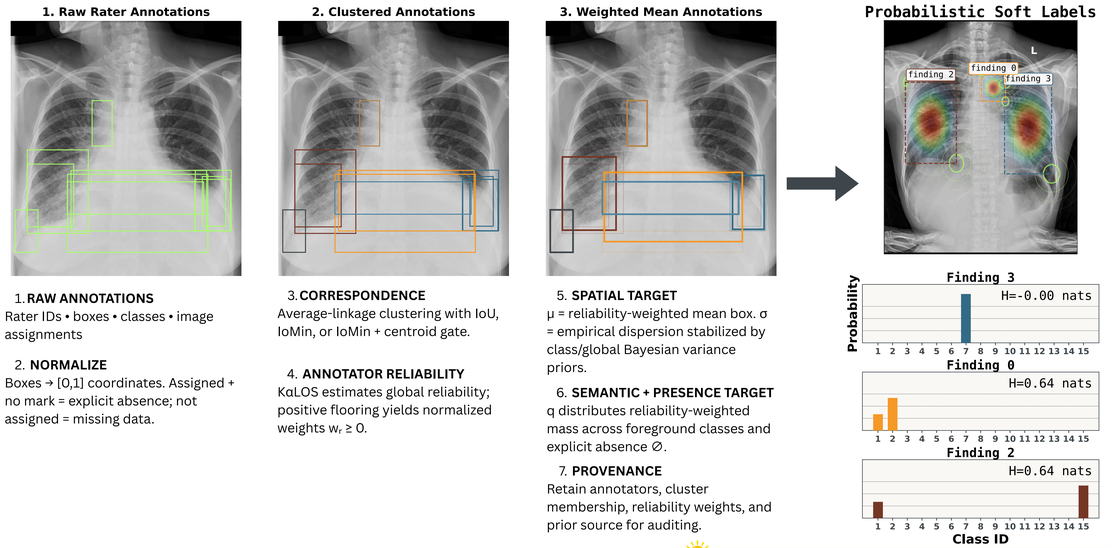
- Multi-annotator datasets inherently contain disagreement.
- Experts may disagree on object location, class, or even presence.
- Conventional consensus aggregation collapses repeated annotations into a single hard label, throwing away nuance, ambiguity, and subjective expertise.
- Our goal is to preserve disagreement during dataset curation and use it as supervision.
- The curated target is probabilistic and retains uncertainty.

$$z = (\mu, \sigma, q)$$

μ= mean box,
σ = spatial uncertainty,
q= probabilistic distribution over classes



2. From Multi-Annotator Labels to Probabilistic Labels (Data Curation)



1. **RAW ANNOTATIONS**
Rater IDs • boxes • classes • image assignments

2. **NORMALIZE**
Boxes → [0,1] coordinates. Assigned + no mark = explicit absence; not assigned = missing data.

3. **CORRESPONDENCE**
Average-linkage clustering with IoU, IoMin, or IoMin + centroid gate.

4. **ANNOTATOR RELIABILITY**
KαLOS estimates global reliability; positive flooring yields normalized weights $w_r \geq 0$.

5. **SPATIAL TARGET**
μ = reliability-weighted mean box. σ = empirical dispersion stabilized by class/global Bayesian variance priors.

6. **SEMANTIC + PRESENCE TARGET**
q distributes reliability-weighted mass across foreground classes and explicit absence ∅.

7. **PROVENANCE**
Retain annotators, cluster membership, reliability weights, and prior source for auditing.

Preserves Localization, Semantic, and Presence Disagreement.

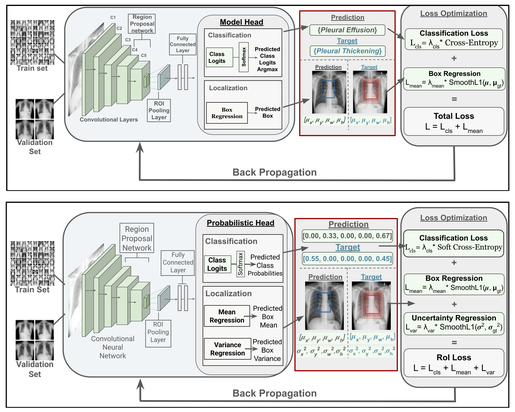
3. Probabilistic Detector (Prob-Faster-RCNN)

Standard Two-Stage Faster-RCNN
Input:
1. GT Box position [x,y,w,h],
2. Single Majority-Vote Class Label [C_i]

Output:
1. Predicted Box Position,
2. Softmax Class Label

Probabilistic Two-Stage Faster-RCNN
Input:
1. Mean Box Position [μ_x, μ_y, μ_w, μ_h]
2. Spatial Uncertainty [σ_x, σ_y, σ_w, σ_h]
3. Weighted Class Probablities [Q]

Output:
1. Predicted Box Position
2. Predicted Spatial Uncertainty
3. Predicted Probability Distribution



Shared Experimental Setup:
- Same Dataset split and annotation-correspondence strategy.
- Same ResNet-50-FPN backbone, RPN, RoIAlign, anchors, and RoI sampling.
- Same SGD optimizer, 30 epochs, warm-up → cosine annealing, and batch size of 4.
- Same repeat-factor sampling and data augmentations.
- Only the supervision and second-stage prediction differ: hard labels supervise one class + mean box; probabilistic labels supervise (**μ, σ, q**).

4. Evaluations Metrics

Conventional detection accuracy based on correct class and box overlap.

Combined error from localization, false positives, and missed detections.

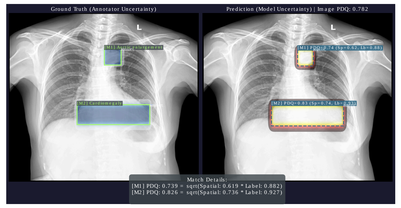
Conventional detection accuracy based on correct class and box overlap.

Joint quality of probabilistic spatial and class predictions.

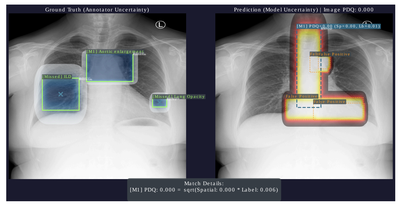
Difference between predicted and annotator-derived class distributions.

5. Results

Sample Inferences with PDQ Scores



Best PDQ: VinDr-CXR - Resnet50



Worst PDQ: VinDr-CXR - Resnet50

	Deterministic Metrics			Probabilistic Metrics				
Configuration	**mAP@40↑**	**mAP@50↑**	**moLRP↓**	**PDQ↑**	**Spatial PDQ↑**	**Label PDQ↑**	**Expected IoU↑**	**JS-Divergence↓**
H-R50-IoU	0.207	0.176	0.870	-	-	-	-	-
S-R50-IoU	0.172	0.149	0.940	0.0605	0.299	0.477	0.6161	0.1755
H-R50-IoMin	0.174	0.144	0.882	-	-	-	-	-
S-R50-IoMin	0.248	0.200	0.923	0.0532	0.278	0.540	0.4482	0.1845
H-R50-IominCG	0.181	0.151	0.881	-	-	-	-	-
S-R50-IoMinCG	0.182	0.151	0.940	0.0485	0.304	0.510	0.5352	0.1888

Table 1 : Evaluation Results on Publically available VinDR-CXR Dataset

Configuration Legend:
H = Majority-Vote Hard Labels
S = Probabilistic Soft Labels
R50 = Resnet50 backbone
IoU = Intersection over Union
IoMin = Intersection over Minimum
IoMinCG = IoMin with Centroid Gate

Dataset (Domain)	Hard Label- mAP@50↑	Soft Label- mAP@50↑
TexBig2023 (Document Layout)	0.720	0.739
NuCLS (Histopathology)	0.213	0.215
LVIS (Natural Setting)	0.013	0.026
MultiOrg (Biomedical)	0.599	0.641

Table 2 : Cross-Dataset Portability

Findings:
- IoMin gives the highest mAP; IoU gives the best PDQ, Expected IoU, and JSD.
- Deterministic and probabilistic metrics rank curation strategies differently.
- Soft-Label curation and training process can be translated to other multi-annotator datasets.

6. Conclusion & Takeaways

Key Findings:
- **Disagreement is a usable supervision signal**. Repeated annotations can be curated into probabilistic targets (**μ,σ,q**) that preserve localization, semantic, and presence disagreement rather than collapsing them into consensus labels.
- **The probabilistic targets are learnable**. They successfully train a probabilistic Faster R-CNN under all three VinDr-CXR correspondence settings.
- **Curation choices matter**. Annotation correspondence changes the generated targets and downstream model ranking; IoMin gives the highest observed mAP, while IoU gives the strongest PDQ, Expected IoU, and JSD.
- **mAP does not tell the whole story**. Probabilistic metrics expose information invisible to deterministic detection scores; learned uncertainty also tracks localization error better than the hard-model prior heuristic (**ρ=0.48 vs 0.23**).
- **Cross-dataset feasibility**: The same probabilistic representation and learning formulation were configured and trained on four additional heterogeneous multi-annotator datasets.

Limitations
- **Diagonal Gaussian uncertainty** — does not model correlations or multimodal spatial disagreement.
- **Global Annotator Reliability** — one reliability weight per annotator rather than case- or class-dependent reliability.
- **Complete Reporting Assumption** — an assigned annotator's omission is interpreted as evidence of absence.
- **Probabilistic Evaluation remains unresolved** — existing metrics measure selected properties, but there is no established protocol for validating probabilistic ground truth as a whole.

Annotator disagreement need not be discarded as noise—it can serve as structured, learnable supervision for object detection.

Check out the GitHub Repository here!

www.uni-weimar.de