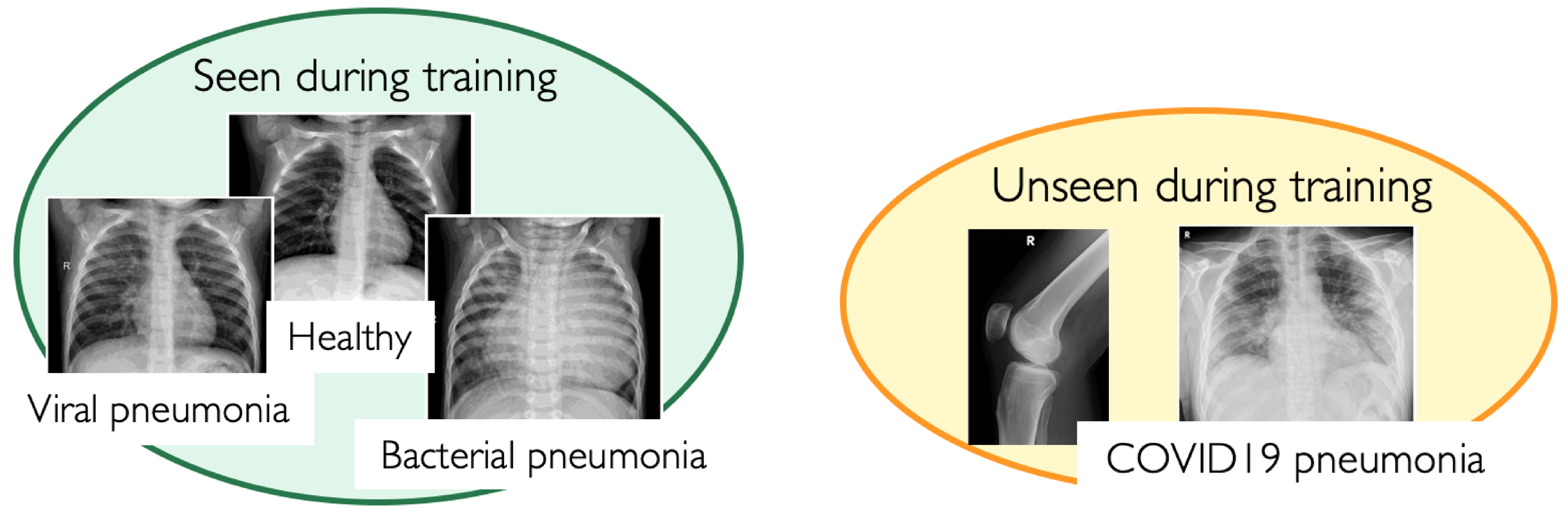
Novelty detection using ensembles with regularized disagreement

Alexandru Ţifrea, Eric Stavarache, Fanny Yang

Department of Computer Science, ETH Zurich

Novel classes as OOD data

Problem: Classifier predictions are incorrect on novel classes.
→ Flag data from unseen classes as out-of-distribution (OOD).



→ Novel classes are often similar to in-distribution (ID) classes
⇒ difficult to distinguish ID and OOD data.

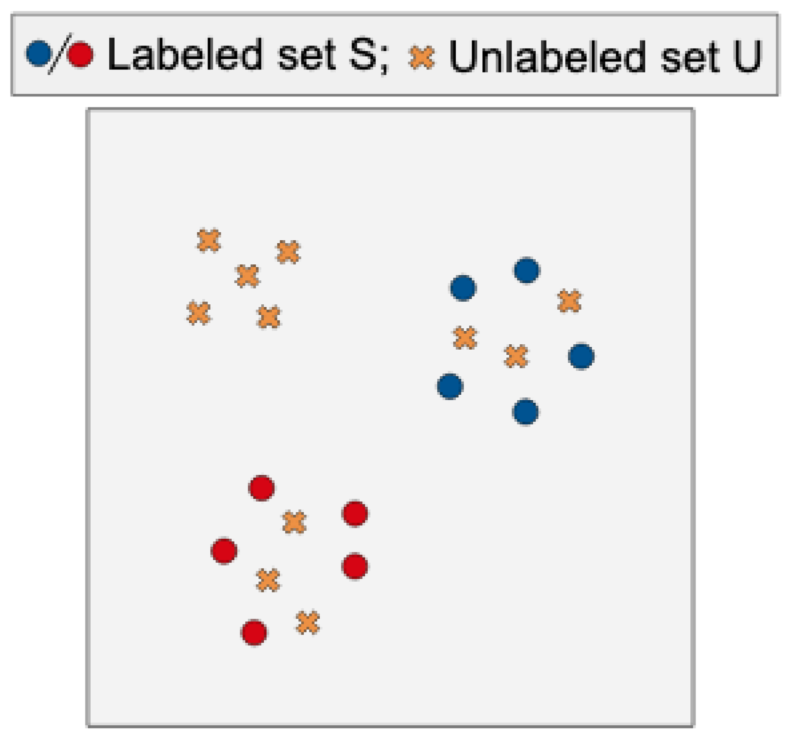
Existing OOD detection methods (assuming different access to OOD data) **perform poorly on novel-class detection**.

Our setting

Available data:

- Labeled set with ID samples.
  → e.g. the training set for the prediction task.
- Unlabeled set with unknown mixture of ID and OOD data.
  → e.g. hospital collects all X-rays performed during the day.

Unknown OOD setting:



Unknown OOD setting

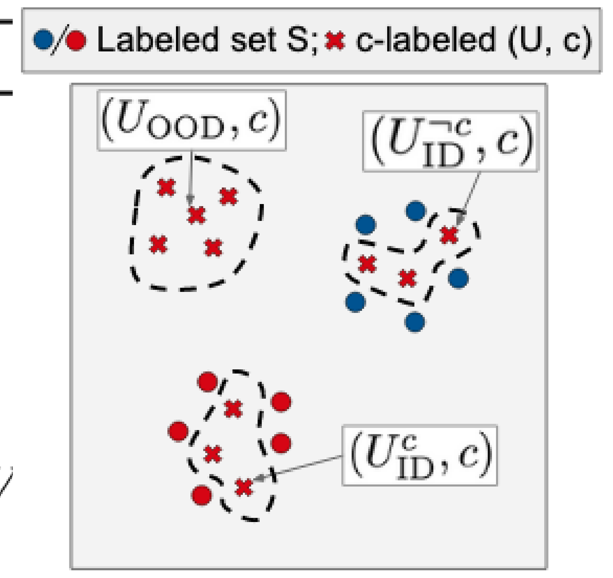
Previous methods that employ the Unknown OOD setting (e.g. nnPU, MCD) **fail to leverage unlabeled data effectively**.

Our approach

Idea: Train an **Ensemble w/ Regularized Disagreement**.

Algorithm 1: Fine-tuning the ERD ensemble

Input : Train set S, Validation set V, Unlabeled set U, Weights W pretrained on S, Ensemble size K
Result: ERD ensemble $\{f_{y_i}\}_{i=1}^K$
Sample K different labels $\{y_1, \ldots, y_K\}$ from $\mathcal{Y}$
for $c \leftarrow \{y_1, \ldots, y_K\}$ **do** `// fine-tune` K `models`
  $f_c \leftarrow \mathit{Initialize}(W)$
  $(U, c) \leftarrow \{(x, c) : x \in U\}$
  $f_c \leftarrow \mathit{FinetuneWithEarlyStopping}\,(f_c, S \cup (U, c); V$
return $\{f_{y_i}\}_{i=1}^K$



Subsets of the c-labeled U

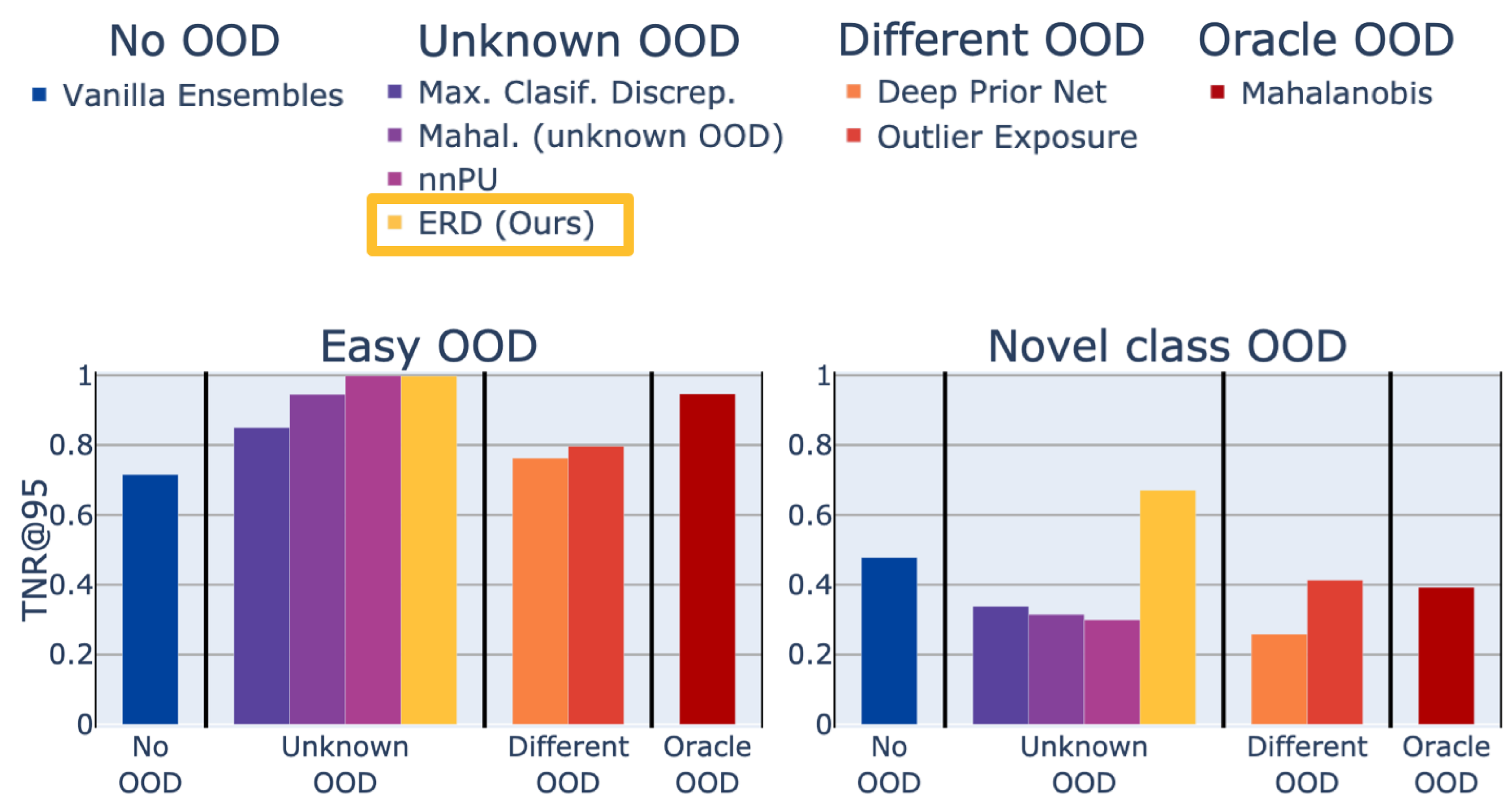
At test time:

- For a test sample x, use outputs $f_1(x), \ldots, f_k(x)$ to compute the **average pairwise disagreement score** (details later).
  → Flag as OOD samples with score larger than threshold τ.

Experiments

Easy OOD: SVHN vs CIFAR10, CIFAR10 vs SVHN etc
Novel class OOD: CIFAR100[0-49] vs CIFAR100[50-99] etc

Evaluation metric: TNR at a TPR of 95%.
→ TNR = correctly identified ID; TPR = correctly flagged OOD.

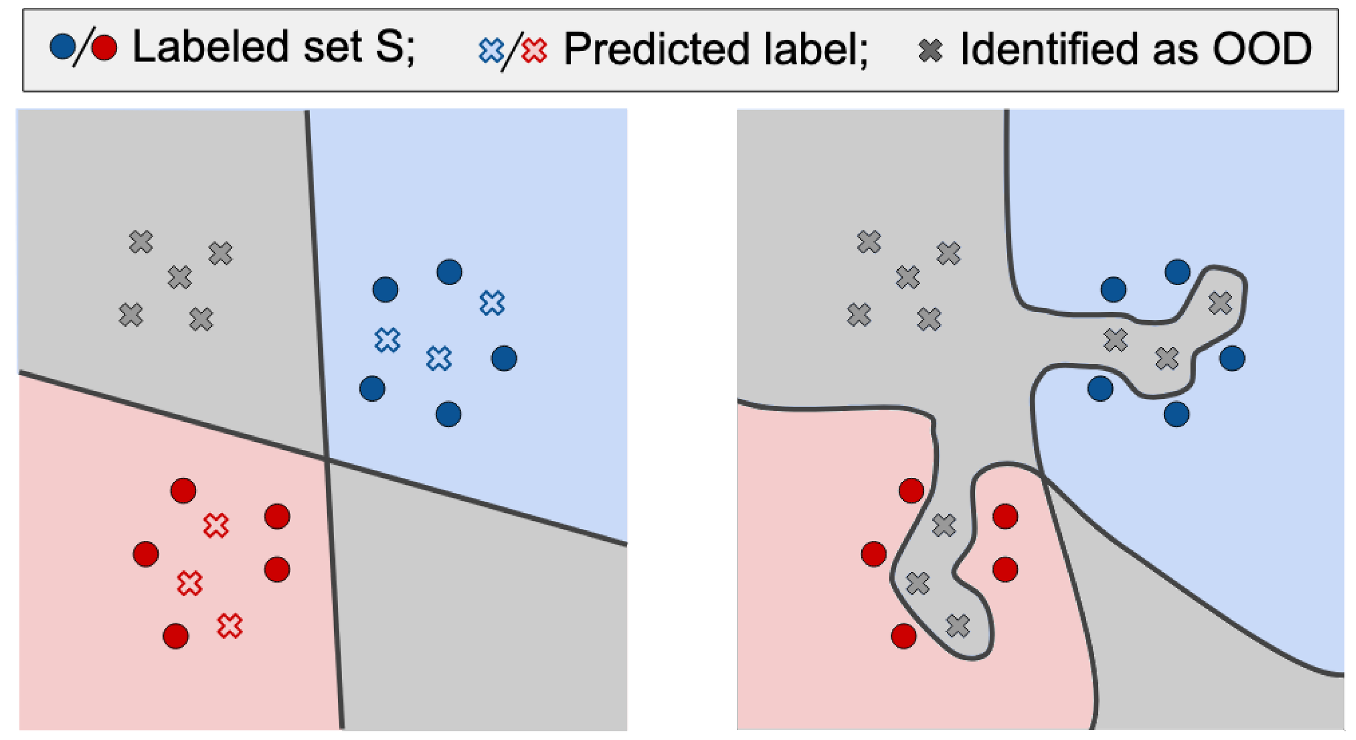


Our approach makes use of two key ingredients:

1. regularization
2. a suitable score for OOD detection.

Key 1: Role of regularization

Goal: Prevent complex models from interpolating on $S \cup (U, c)$.



Advantages of early stopping:

- We prove that there exists an **optimal stopping time** at which every model predicts: (1) the correct label on ID data; and (2) the arbitrary label on the OOD unlabeled data.
- Efficient model selection (requires only one training run).

Key 2: Ensemble disagreement score

Prior work: Entropy of average predictor (H ∘ Avg).
Our average pairwise disagreement score:

$$(\mathrm{Avg} \circ \rho)(f_1(x), \ldots, f_K(x)) := \frac{2}{K(K-1)} \sum_{i \neq j} \rho\left(f_i(x), f_j(x)\right)$$

→ e.g. ρ = total variation distance

- Unlike (H ∘ Avg), our score exploits ensemble diversity.
  ⇒ lower FPR at the same TPR