# Learning to Defer with Limited Expert Predictions

**Patrick Hemmer*, Lukas Thede*, Michael Vössing, Johannes Jakubik, Niklas Kühl**

Karlsruhe Institute of Technology
{patrick.hemmer, michael.voessing, johannes.jakubik, niklas.kuehl}@kit.edu, lukas.thede@alumni.kit.edu

## Appendix

We provide further information about our proposed approach concerning related work, a formalization as an algorithm, and additional implementation details. Moreover, we present additional analyses for the synthetic experts $H_{60}$ and $H_{90}$ on CIFAR-100 and provide experimental results for a second radiologist. Lastly, we report further information on the bias analysis for the NIH dataset.

## Extended Approach

We elaborate on how the setting in this work differs from the standard semi-supervised few-shot learning problem, followed by a formalization of the proposed approach.

### Differentiation from Semi-Supervised Few-Shot Learning

In recent years, several works have proposed approaches that combine semi-supervised and few-shot learning to improve the performance of few-shot learning tasks. Ren et al. (2018) extend prototypical networks (Snell, Swersky, and Zemel 2017) to incorporate unlabeled data. Liu et al. (2019) obtain labels for the unlabeled data by using a graph constructed between labeled and unlabeled data in combination with label propagation. Li et al. (2019) apply self-training in each optimization round to generate labels for the unlabeled instances. Yu et al. (2020) propose a transfer-learning approach that uses prior information in the form of a pretrained feature extractor to imprint the weights of a classifier which are then updated using semi-supervised learning. Even though these approaches are related to our work, they are not applicable to the problem of learning artificial expert predictions as in our setting base and novel task consist of the same instances. In detail, approaches pursuing the idea that for a novel task, the query instances are closer to the corresponding support instances of the same class in the embedding space than to the support instances from other classes turn out not to be able to learn the human expert's capabilities. The reason for this observation is that the embedding model trained on the base task clusters the same instances within the novel task in the embedding space according to the ground truth labels. As in our setting the instances of the novel task do not differ from the instances of the base task, the clusters within the embedding space of the novel task are still determined by the labels of the base task rather than by the labels of the novel task.

### Formalization of the Approach

Algorithm A1: Approach for generating artificial expert predictions for learning to defer algorithms

**Require:** $D^l = \{(x_i, y_i, h_i)\}_{i=1}^{l}, D^u = \{(x_j, y_j)\}_{j=1}^{u}$
**for** number of iterations **do**
  Sample a minibatch $B \subseteq D = D^l \cup D^u$
  **for** $i \in B$ **do**
    <Train embedding model with classifier $\Omega(\Phi_{emb}(\cdot))$ using $(x_i, y_i)$>
  **end for**
  <Backpropagate $\mathcal{L}$ using Eq.(1)>
**end for**

**for** $i \in L$ **do**
  $h_i^{bin} = \Psi(h_i, y_i)$
**end for**

**for** number of iterations **do**
  Sample minibatch $B^l \subseteq D^l$ and $B^u \subseteq D^u$
  **for** $i \in B^l$ and $j \in B^u$ **do**
    $f_i = \Phi_{emb}(x_i)$
    $f_j = \Phi_{emb}(x_j)$
    <Train expertise predictor model $\Phi_{ex}(\cdot)$ using $(f_i, h_i^{bin})$ and $f_j$ with a semi-supervised algorithm, e.g., FixMatch or CoMatch>
  **end for**
  <Backpropagate $\mathcal{L}$ using Eq.(3)>
**end for**

**for** $j$ in $U$ **do**
  $\hat{h}_j^{bin} = \arg\max \Phi_{ex}(\Phi_{emb}(x_j))$
  $\hat{h}_j = \Theta(\hat{h}_j^{bin}, y_j)$
**end for**
**Output:** $D^l = \{(x_i, y_i, h_i)\}_{i=1}^{l}, D^u = \{(x_j, y_j, \hat{h}_j)\}_{j=1}^{u}$

*These authors contributed equally.

Algorithm A1 provides a formalization of the training procedure of our proposed approach as a prior step to the training of learning to defer algorithms (Algorithm A2) requiring both a ground truth label and human expert prediction for each instance in the training dataset.

Algorithm A2: Training of learning to defer algorithms with human and artificial expert predictions

**Require:** $D^l = \{(x_i, y_i, h_i)\}_{i=1}^{l}, D^u = \{(x_j, y_j, \hat{h}_j)\}_{j=1}^{u}$

**for** number of iterations **do**
  Sample minibatch $B^l \subseteq D^l$ and $B^u \subseteq D^u$
  **for** $i \in B^l$ and $j \in B^u$ **do**
    <Train learning to defer algorithm using $(x_i, y_i, h_i)$ and $(x_j, y_j, \hat{h}_j)$>
  **end for**
  <Backpropagate $\mathcal{L}$>
**end for**

## Extended Implementation Details

We provide additional information about the implementation details for the experiments. All experiments were conducted on a Red Hat Enterprise Linux 8.2 system and a NVIDIA A100-PCIE with 40GB RAM. We used PyTorch 1.8.0 for the implementation.

### Hyperparameters of the Approach and the Baselines

This section contains additional implementation details for the experiments on the CIFAR-100 and NIH datasets.

Table B1 presents the hyperparameters of the embedding models using an EfficientNet-B1 (Tan and Le 2019) for CIFAR-100 and a ResNet18 (He et al. 2016) for the NIH dataset. A single fully connected layer with softmax activation is used as the classifier to predict the ground truth labels.

Table B2 displays the hyperparameters of the expertise predictor models, including our proposed approaches and the considered baselines. The classifier for predicting the artificial binary expert predictions consists of a single fully connected layer with a softmax activation. For the *Embedding-SVM* baseline, we use an RBF kernel and a regularization parameter of 5.

### Choice of Metric for the Evaluation of Artificial Expert Predictions

As the transformation to a binary problem results in an imbalanced class distribution, we select a performance metric that accurately represents the performance of the artificial expert predictions compared with the human expert predictions. For this reason, we need to understand the meaning of both recall and precision for generating artificial expert predictions: Recall states how many of all existing correct expert predictions were predicted as correct by the expertise predictor model. In contrast, precision indicates how often an expert prediction is correct when predicted as correct by the expertise predictor model. Experiments with different learning to defer algorithms have revealed that it turns out to be less of an issue for the downstream performance of the learning to defer algorithms when the expertise predictor model mistakenly does not identify instances as expert strengths. For learning to compensate for the expert's weaknesses, it rather turns out to be more important that the strengths denoted by the artificial expert predictions actually contain the human expert strengths. Therefore, for evaluating the performance of the artificial expert predictions, the importance of missing correct expert predictions can be weighted less. Instead, we put more weight on the ability of the expertise predictor model that the instances for which the artificial expert predictions suggest the expert to be correct indeed turn out to be correct. Therefore, we chose the $F_{0.5}$-score as a metric for evaluating the performance of the artificial expert predictions.

### Hyperparameters of the Learning to Defer Algorithms

Table B3 and Table B4 present the hyperparameters of the learning to defer algorithms for the three considered approaches of Mozannar and Sontag (2020), Raghu et al. (2019), and Okati, De, and Rodriguez (2021) for the experiments on the CIFAR-100 and NIH datasets respectively. For the learning to defer algorithms, no hyperparameter finetuning was conducted as our work focuses on the relative evaluation of the approaches trained on human and artificial expert predictions compared to an approach trained on a dataset with complete human expert predictions.

In their implementation Okati, De, and Rodriguez (2021) designed their approach to work with average human experts providing a class-wise prediction probability instead of categorical labels. We generate one-hot encodings of all expert predictions to allow the approach to be trained with categorical expert predictions and artificial expert predictions. To enable the calculation of the NLL loss, we subtract an $\varepsilon$ of 5e-4 from the predicted class within the one-hot encoding and distribute it over the remaining other classes generating soft labels for all instances. Moreover, Okati, De, and Rodriguez (2021) propose to train the classifier model to complement the human expert's weaknesses by using only the loss of those instances assigned to the classifier for backpropagation. However, for the experiments on the NIH dataset, this complementary training approach leads to the classifier focusing solely on one of the two classes, predicting it without regard to the input. To account for this problem, we include a general NLL loss calculated on all instances into the classifier's training and add it to the loss calculated only on those instances assigned to the classifier with a factor of 0.1. This step ensures a certain level of generalization of the classification model while still allowing it to be trained to complement the human expert.

## Extended Results

In this section, we report additional analyses for the CIFAR-100 dataset (Krizhevsky 2009) and additional experimental results for the NIH dataset (Majkowska et al. 2020; Wang

| Embedding Model (Feature Extractor) | Loss-Function | Optimizer (lr, weight-decay, momentum, Nesterov) | LR-Scheduler (gamma, milestones) | Epochs (CIFAR-100, NIH) |
|---|---|---|---|---|
| EfficientNet-B1 (Tan and Le 2019) | Cross-Entropy Loss | SGD (1e-1, 5e-4, 0.9, True) | Multistep-Scheduler (0.2, [60, 120, 160]) | 200, - |
| ResNet18 (He et al. 2016) | Cross-Entropy Loss | SGD (1e-3, 5e-4, 0.9, True) | Multistep-Scheduler (0.2, [60, 120, 160]) | -, 200 |

Table B1: Hyperparameters of the trained embedding models for the experiments conducted on CIFAR-100 and the NIH dataset.

| Expertise Predictor Model | Loss-Function | Optimizer (lr, weight-decay, momentum, Nesterov) | LR-Scheduler (gamma, milestones) | Epochs (CIFAR-100, NIH) |
|---|---|---|---|---|
| Embedding-NN | Cross-Entropy Loss | SGD (4e-3, 0, 0.9, True) | Multistep-Scheduler (0.2, [50, 70, 90]) | 150, 150 |
| CoMatch | CoMatch Loss | SGD (3e-2, 5e-4, 0.9, False) | Cosine-Scheduler (-, -) | 150, 25 |
| FixMatch | FixMatch Loss | SGD (3e-2, 5e-4, 0.9, False) | Cosine-Scheduler (-, -) | 150, 25 |
| Embedding-CoMatch | CoMatch Loss | SGD (3e-2, 5e-4, 0.9, False) | Cosine-Scheduler (-, -) | 50, 25 |
| Embedding-FixMatch | FixMatch Loss | SGD (3e-2, 5e-4, 0.9, False) | Cosine-Scheduler (-, -) | 50, 25 |

Table B2: Hyperparameters of the trained expertise predictor models for the experiments conducted on CIFAR-100 and the NIH dataset.

et al. 2017) for another radiologist. Moreover, for the NIH dataset, we present further analyses with regard to whether the artificial expert predictions induce performance disparities across existing groups.

## CIFAR-100

Figure C1 to Figure C7 present the classifier coverages of the learning to defer algorithms of Mozannar and Sontag (2020), Raghu et al. (2019), and Okati, De, and Rodriguez (2021) on the CIFAR-100 dataset. The classifier coverages represent the percentage of instances assigned to the classifier and are visualized per superclass, each sorted by decreasing expert strength. Expert strength denotes the number of subclasses within a superclass that the human expert predicts correctly. Generally, the classifier coverages are negatively correlated with the expert's strength as the learning to defer algorithms assign each instance to the team member for which the probability of a correct prediction is highest. In order to evaluate the coverages, we include the classifier coverages from the respective learning to defer algorithm trained on a complete set of human expert predictions. In this context, it is desired that the coverages resulting from the learning to defer algorithms trained with human and artificial expert predictions match their coverages when trained on a complete set of human expert predictions as closely as possible. Due to the aforementioned correlation between the classifier coverages and the expert's strengths, the coverages can serve as an indicator of the expert's strength. Subsequently, the comparison of the classifier coverages for the human and artificial expert predictions with the coverages for a complete set of expert predictions allows gaining further insights into the ability of the proposed approach to mimic the true expert's strengths and weaknesses through the artificial expert predictions.

Figure C1 visualizes the classifier coverages of the learning to defer algorithms trained on human and artificial expert predictions for the synthetic expert $H_{60}$. The artificial expert predictions are generated by the *Embedding-SSL* approaches, respectively. In general, these coverages match well with the coverages of the approaches trained on a complete set of expert predictions. This demonstrates the ability of the *Embedding-SSL* approaches to reflect the expert's strengths and weaknesses even from small sets of expert predictions. Furthermore, the coverages match especially well for superclasses of high expert strength (superclasses 1, 2, 7, 8, 12, 15, 16) and low expert strength (superclasses 19, 5, 8). In contrast, the coverages tend to diverge more for classes of medium expert strengths. This indicates that learning the expert's capabilities works best for superclasses that unambiguously belong to the expert's strengths or weaknesses. In contrast, it turns out to be more difficult to learn the capabilities in cases where the expert is neither particularly strong nor weak. In this context, it can be observed that for a higher number of human expert predictions available for training, the differences between the approaches trained on a complete set of human expert predictions decrease compared with the coverages of the approaches trained on human and artificial expert predictions.

In addition, we present the coverages for the more accurate synthetic expert $H_{90}$ in Figure C2. The results substantiate the previous observation that the capabilities of the expert can be learned best for the superclasses that clearly belong to the expert's strengths or weaknesses. As the synthetic expert $H_{90}$ is overall more accurate, the number of superclasses unambiguously belonging to the expert's strengths is considerably higher, explaining why the capabilities of the more accurate synthetic expert $H_{90}$ are easier to learn for all approaches.

Next, we present the classifier coverages of the *Embedding-SSL* approaches alongside the *Embedding-SL* and *SSL* approaches for $l = 120$ human expert predictions in Figure C3. Compared to the *Embedding-SL* and *SSL* approaches, the coverages of the *Embedding-SSL* approaches match, on average, the coverages with a complete set of expert predictions closest. This substantiates the benefit of 1) the embedding model and 2) the use of semi-supervised learning for training the expertise predictor model to learn the expert's capabilities from a small set of only six expert predictions per superclass.

Figure C4 and Figure C5 display the classifier coverages of the learning to defer algorithms trained on human and artificial expert predictions for the synthetic experts $H_{60}$ and $H_{90}$, respectively. The artificial expert predictions are

| Learning to Defer Algorithm | Backbone (depth, widen-factor) | Loss-Function | Optimizer (LR, weight-decay, momentum, nesterov) | LR-Scheduler | Epochs |
|---|---|---|---|---|---|
| Mozannar and Sontag (2020) | WideResNet (28, 4) | Surrogate Loss | SGD (1e-1, 5e-4, 0.9, True) | Cosine-Scheduler | 50 |
| Raghu et al. (2019) Classifier | WideResNet (28, 4) | Cross-Entropy Loss | SGD (1e-1, 5e-4, 0.9, True) | Cosine-Scheduler | 200 |
| Raghu et al. (2019) Human Error Model | WideResNet (10, 4) | Cross-Entropy Loss | SGD (1e-1, 5e-4, 0.9, True) | Cosine-Scheduler | 100 |
| Okati, De, and Rodriguez (2021) Classifier | WideResNet (28, 4) | NLL Loss | SGD (1e-3, 5e-4, 0.9, True) | Cosine-Scheduler | 100 |
| Okati, De, and Rodriguez (2021) Human Error Model | WideResNet (10, 4) | NLL Loss | SGD (1e-1, 5e-4, 0.9, True) | Cosine-Scheduler | 100 |

Table B3: Hyperparameters of the utilized learning to defer algorithms for the experiments conducted on the CIFAR-100 dataset.

| Learning to Defer Algorithm | Backbone | Loss-Function | Optimizer (LR, weight-decay, momentum, nesterov) | LR-Scheduler | Epochs |
|---|---|---|---|---|---|
| Mozannar and Sontag (2020) | ResNet18 | Surrogate Loss | SGD (1e-4, 5e-4, 0.9, True) | Cosine-Scheduler | 50 |
| Raghu et al. (2019) Classifier | ResNet18 | Cross-Entropy Loss | SGD (1e-3, 5e-4, 0.9, True) | Cosine-Scheduler | 200 |
| Raghu et al. (2019) Human Error Model | ResNet18 | Cross-Entropy Loss | SGD (5e-5, 5e-4, 0.9, True) | Cosine-Scheduler | 100 |
| Okati, De, and Rodriguez (2021) Classifier | ResNet18 | NLL Loss | SGD (1e-5, 5e-4, 0.9, True) | Cosine-Scheduler | 100 |
| Okati, De, and Rodriguez (2021) Human Error Model | ResNet18 | NLL Loss | SGD (1e-3, 5e-4, 0.9, True) | Cosine-Scheduler | 100 |

Table B4: Hyperparameters of the three utilized learning to defer algorithms for the experiments conducted on the NIH dataset.

generated by the *Embedding-SL* approaches. The coverages tend to have a larger deviation from the coverages of the learning to defer algorithms trained on a complete set of expert predictions compared with the *Embedding-SSL* approaches. Furthermore, the results indicate a tendency of the *Embedding-SL* approaches to overestimate the expert's strength, especially for $l < 120$, as the classifier coverages are lower than the classifier coverages of the learning to defer algorithms trained on complete sets of expert predictions. Lower classifier coverages subsequently represent higher human expert coverages, thus indicating a larger number of instances being assigned to the expert.

Lastly, Figure C6 and fig. C7 visualize the classifier coverages of the learning to defer algorithms trained on human and artificial expert predictions for both synthetic experts $H_{60}$ and $H_{90}$, respectively. The artificial expert predictions are generated by the *SSL* approaches. These coverages visualize the limited ability of the *SSL* approaches to learn the expert's capabilities from small sets of expert predictions, particularly for $l < 400$, as the coverages differ considerably from those trained on a complete set of expert predictions.

## NIH Chest X-rays

Table C1 shows the experimental results for another radiologist (labeler-id: 4295194124) on the NIH dataset, using the same hyperparameters as for the radiologist with labeler-id 4295342357. From the results in Table C1, we observe that the *Embedding-SSL* approaches outperform the *SSL* baselines, on average, by 1.60% across all $l$ values. Moreover, for $l < 100$, they exhibit an average performance improvement of 1.80% compared to the *Embedding-SL* approaches. These findings provide further confirmation of the feasibility of the approach for generating artificial expert predictions.

Figure C8 displays the results of the learning to defer algorithms trained on human and artificial expert predictions for different numbers of $l$ available human expert predictions. The performance of the respective learning to defer algorithm is shown relative to its performance resulting from the training on a complete dataset of expert predictions. The results reinforce the previous findings, indicating that the learning to defer algorithms are capable of surpassing both the performances of the classifier and the expert conducting the task in isolation. In addition, for the algorithm by Mozannar and Sontag (2020), the *Embedding-SSL* approaches outperform the *SSL* baselines, on average, by 3.16% across all $l$ values. In comparison to the *Embedding-SL* baselines they achieve an average improvement of 1.30% for $l < 100$. Regarding the algorithm presented by Raghu et al. (2019), the *Embedding-SSL* approaches outperform the *SSL* baselines on average by 1.94% across all $l$ values and perform similar to the *Embedding-SL* baselines. For the algorithm introduced by Okati, De, and Rodriguez (2021), the *Embedding-SSL* approaches slightly outperform the *SSL* baselines by 0.88% for $l < 100$. Regarding the *Embedding-SL* baselines, the *Embedding-SSL* approaches exhibit a comparable performance across all $l$ values.

In addition to the evaluation of the learning to defer algorithms, we also analyze whether the proposed approach amplifies performance disparities regarding patients' gender and age for this radiologist. The algorithm by Mozannar and Sontag (2020) exhibits an average decrease in bias between the *Embedding-SSL* approaches and the upper boundary (*Complete Expert Predictions*) across all human expert predictions $l$ of 2.38 percentage points (pp) for gender and a decrease of 0.94 pp for age. When examining the algorithm by Raghu et al. (2019), we observe an increase of 2.09 pp for gender and an increase of 2.43 pp for age. The algorithm by Okati, De, and Rodriguez (2021) shows a 5.15 pp increase in bias for gender and a 2.16 pp increase in bias for age.

Lastly, we visualize in Figure C9 and Figure C10 the differences in performance disparities for gender and age considering different numbers of human expert predictions $l$ for radiologists 4295342357 and 4295194124. For different values of $l$ and both gender and age, we do not find that the artificial expert predictions systematically amplify the performance disparities of the learning to defer algorithms.

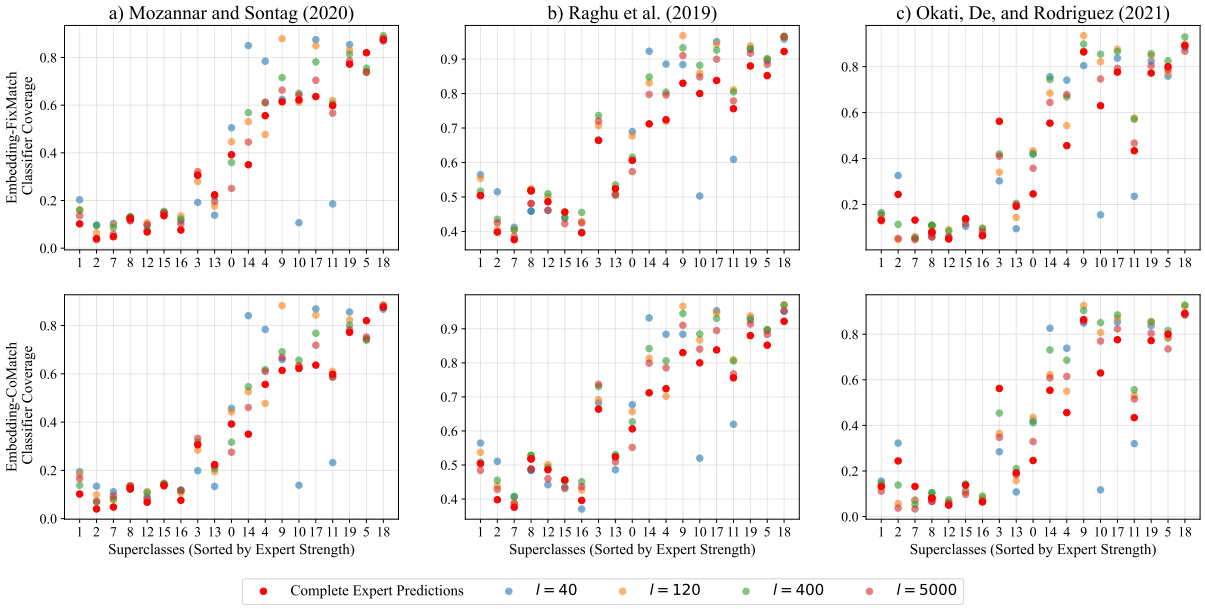


Figure C1: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *Embedding-FixMatch* and *Embedding-CoMatch* approaches, respectively, for the synthetic expert $H_{60}$ based on different numbers of $l$ available human expert predictions.

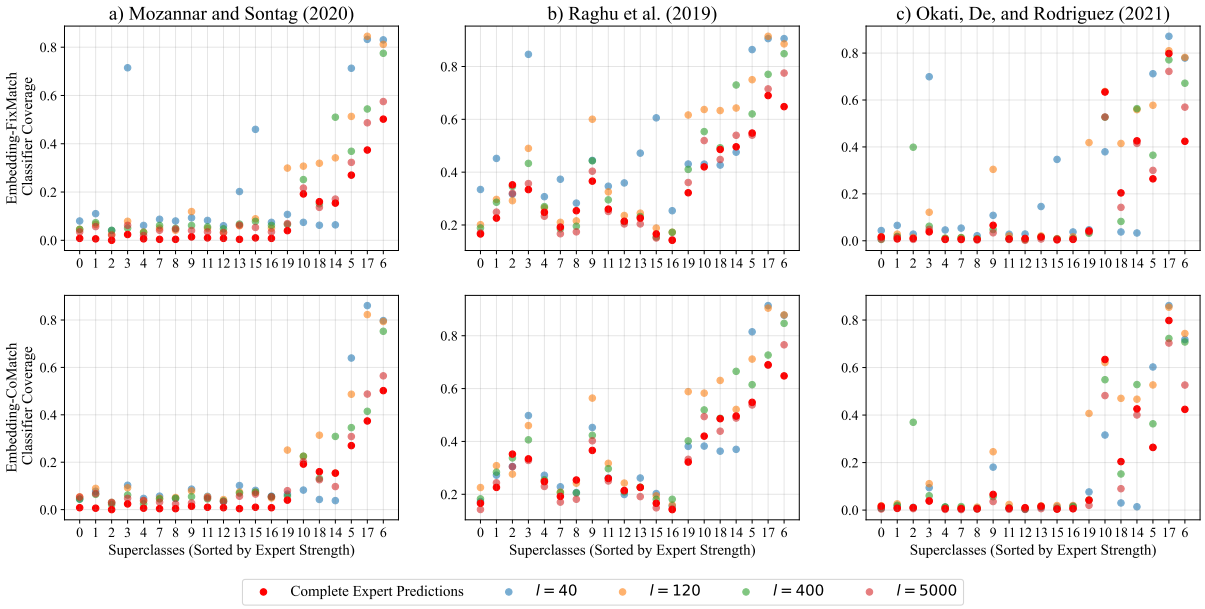


Figure C2: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *Embedding-FixMatch* and *Embedding-CoMatch* approaches, respectively, for the synthetic expert $H_{90}$ based on different numbers of $l$ available human expert predictions.

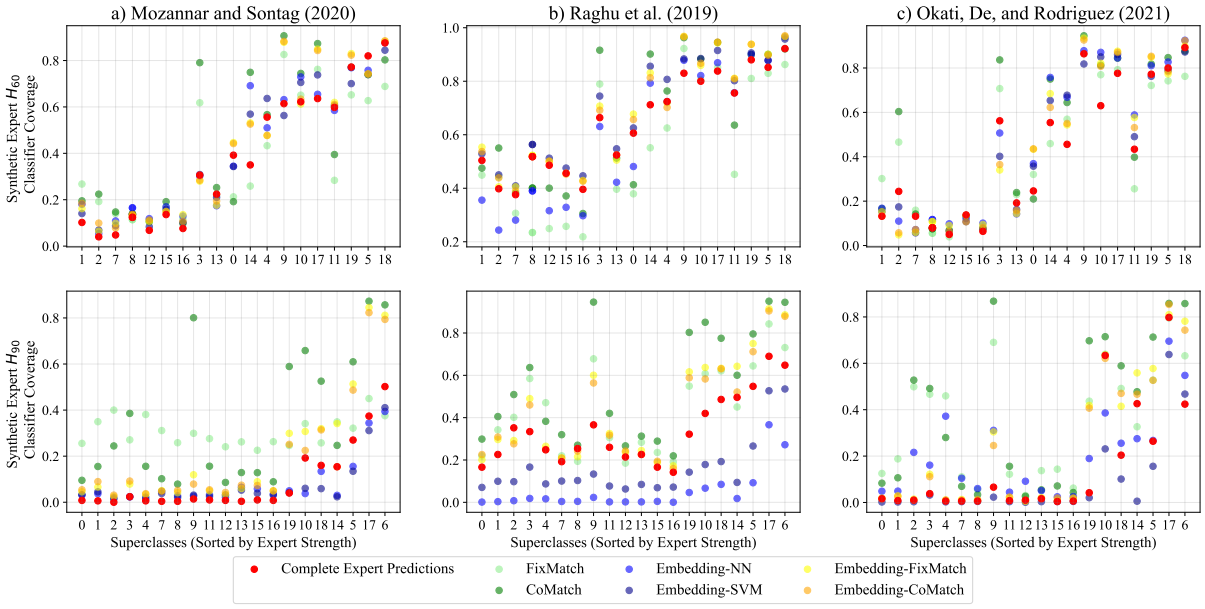


Figure C3: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *Embedding-SSL*, *Embedding-SL*, and *SSL* approaches, respectively, based on $l = 120$ available human expert predictions.

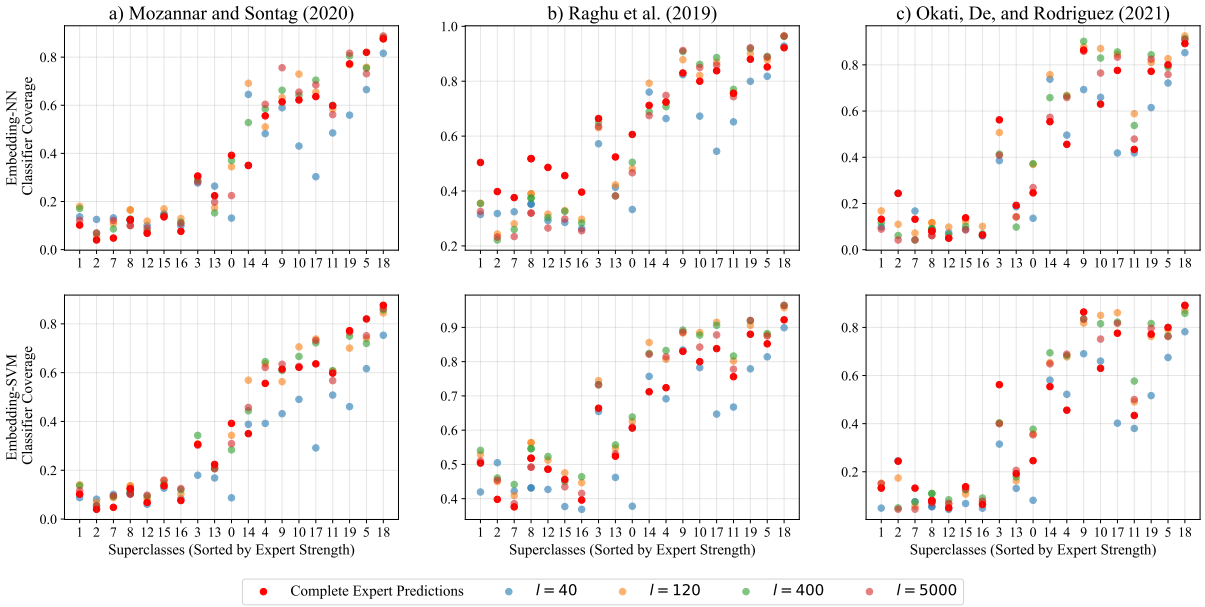


Figure C4: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *Embedding-NN* and *Embedding-SVM* approaches, respectively, for the synthetic expert $H_{60}$ based on different numbers of $l$ available human expert predictions.

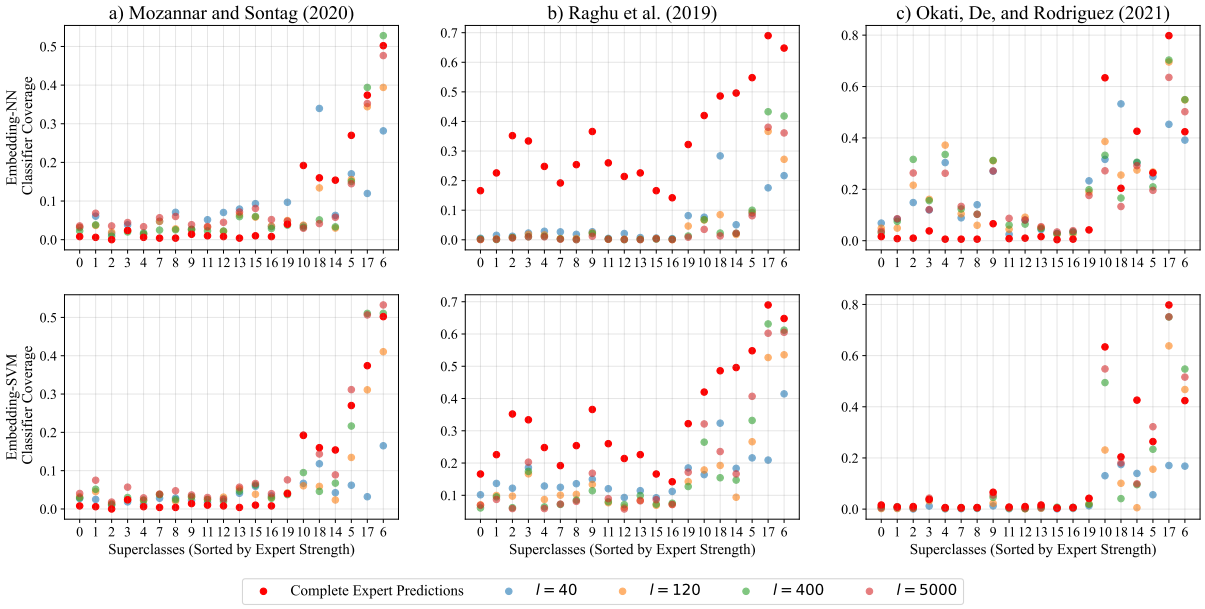


Figure C5: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *Embedding-NN* and *Embedding-SVM* approaches, respectively, for the synthetic expert $H_{90}$ based on different numbers of $l$ available human expert predictions.

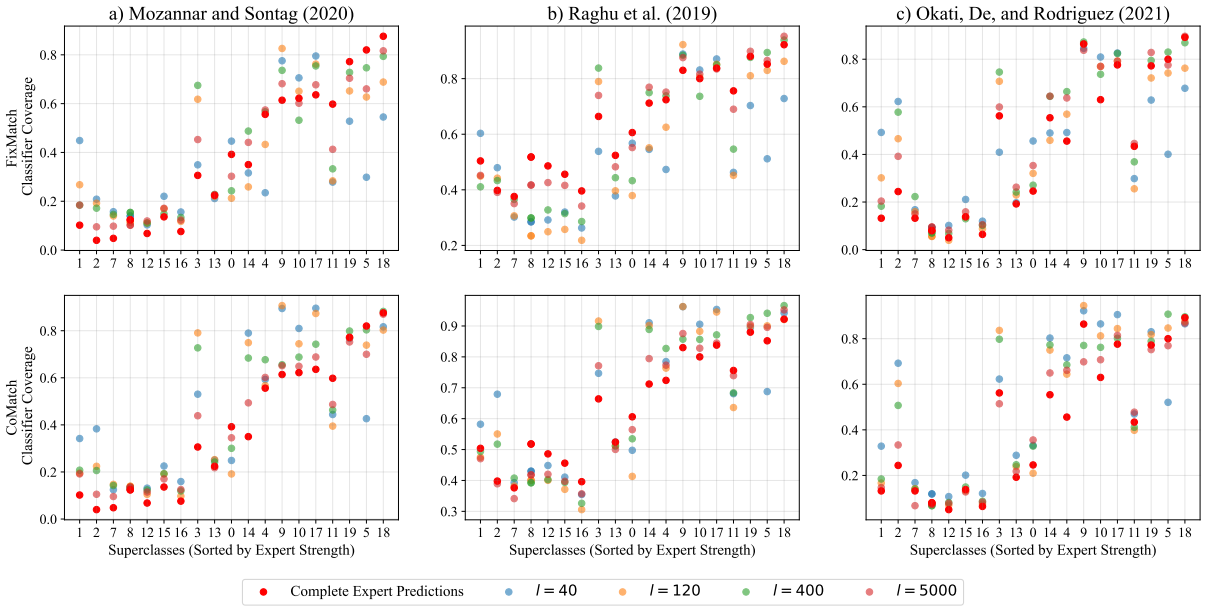


Figure C6: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *FixMatch* and *CoMatch* approaches, respectively, for the synthetic expert $H_{60}$ based on different numbers of $l$ available human expert predictions.

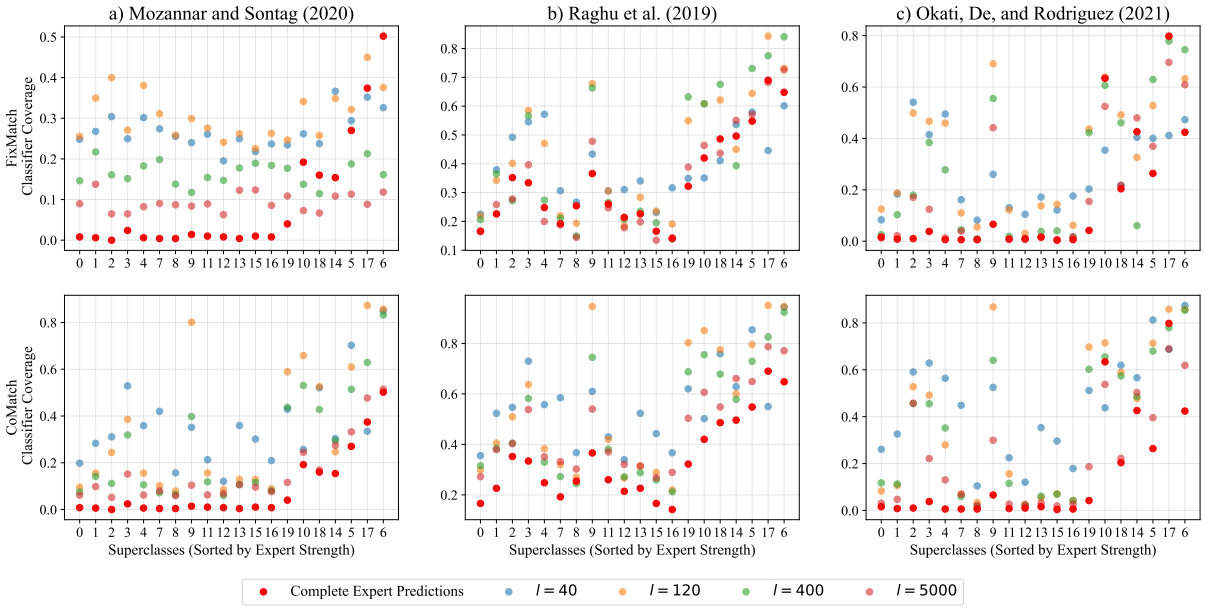


Figure C7: Classifier coverages for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on human and artificial expert predictions. The artificial expert predictions are generated by the *FixMatch* and *CoMatch* approaches, respectively, for the synthetic expert $H_{90}$ based on different numbers of $l$ available human expert predictions.

| $l$ | 4 | 8 | 12 | 20 | 40 | 100 | 500 |
|---|---|---|---|---|---|---|---|
| FixMatch | 82.37 (±1.29) | 83.92 (±0.69) | 84.96 (±0.68) | 84.23 (±0.69) | 85.63 (±0.71) | 86.43 (±0.42) | 93.3 (±0.53) |
| CoMatch | 83.77 (±0.38) | 85.31 (±0.7) | 85.38 (±0.37) | 82.98 (±0.42) | **87.24** (±0.08) | 86.67 (±1.37) | 94.35 (±0.16) |
| EmbeddingNN | 82.24 (±4.73) | 85.15 (±1.72) | 85.57 (±0.48) | 85.25 (±0.91) | 86.39 (±1.32) | 87.61 (±1.32) | **94.87** (±0.53) |
| EmbeddingSVM | 82.23 (±5.79) | 83.42 (±4.74) | 85.61 (±1.55) | 85.6 (±0.76) | 86.51 (±0.84) | 87.94 (±0.71) | 94.4 (±0.81) |
| EmbeddingFM | **87.5** (±0.2) | **86.44** (±0.36) | 87.3 (±0.25) | 85.84 (±0.3) | 86.04 (±0.25) | 87.48 (±0.67) | 93.36 (±0.17) |
| EmbeddingCM | 84.56 (±0.58) | 85.26 (±1.43) | **87.45** (±0.47) | **86.31** (±0.39) | 86.27 (±0.56) | **88.01** (±0.5) | 93.42 (±0.13) |

Table C1: $F_{0.5}$ mean and standard deviation of the artificial expert predictions generated by the expertise predictor model using different numbers of $l$ available human expert predictions of radiologist 4295194124 on the NIH dataset.

| $l$ | 4 | 8 | 12 | 20 | 40 | 100 | 500 |
|---|---|---|---|---|---|---|---|
| FixMatch | 78.15 (±1.29) | 83.96 (±0.98) | 80.35 (±1.31) | 77.66 (±4.26) | 74.97 (±0.56) | 84.68 (±0.33) | 93.13 (±0.25) |
| CoMatch | 81.56 (±0.64) | 83.78 (±1.08) | 83.19 (±0.75) | 81.56 (±0.65) | 82.7 (±0.36) | 85.66 (±0.72) | 92.71 (±0.33) |
| EmbeddingNN | 86.76 (±3.99) | 90.17 (±2.29) | 90.67 (±1.67) | **92.09** (±1.01) | 91.96 (±0.92) | 92.43 (±0.85) | 94.23 (±0.29) |
| EmbeddingSVM | 71.71 (±18.73) | 87.71 (±2.56) | 88.03 (±2.91) | 91.56 (±1.28) | **92.29** (±0.63) | **93.07** (±0.55) | **95.1** (±0.4) |
| EmbeddingFM | **90.2** (±0.15) | 91.36 (±0.0) | **91.42** (±0.0) | 91.03 (±0.18) | 91.49 (±0.04) | 91.42 (±0.04) | 93.24 (±0.1) |
| EmbeddingCM | 89.56 (±1.44) | **91.43** (±0.13) | 91.39 (±0.0) | 91.1 (±0.07) | 90.98 (±0.09) | 91.44 (±0.09) | 93.27 (±0.14) |

Table C1: $F_{0.5}$ mean and standard deviation of the artificial expert predictions generated by the expertise predictor model using different numbers of $l$ available human expert predictions of radiologist 4295342357 on the NIH dataset.

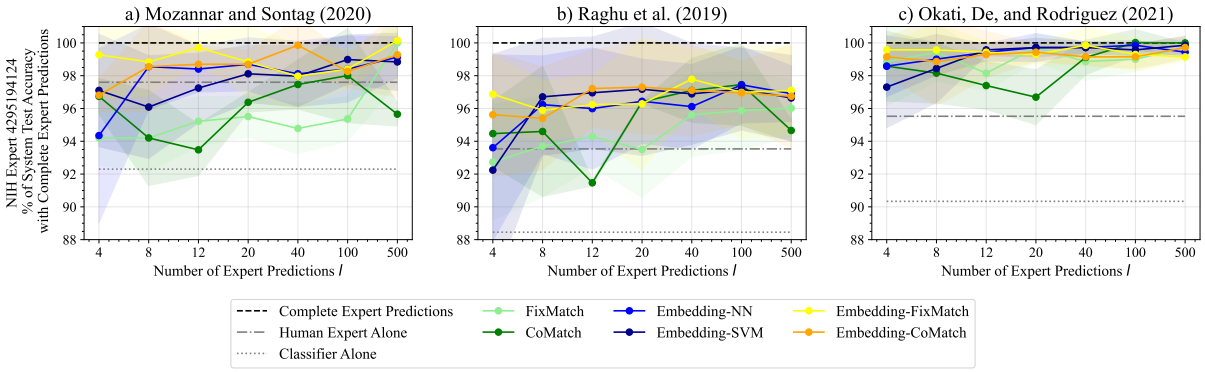


Figure C8: Experimental results for the learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on different numbers of human expert predictions $l$ and artificial expert predictions for radiologist 4295194124. Artificial expert predictions are generated using different numbers of human expert predictions $l$. We report mean and standard deviation over five seeds.

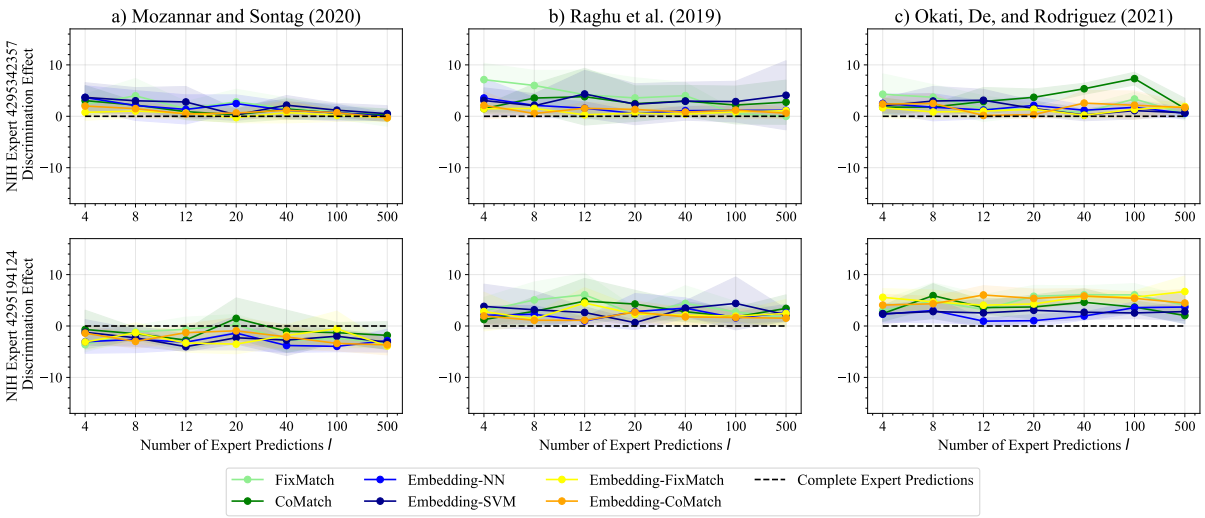


Figure C9: Differences in the discrimination effect for patients' gender (absolute difference in the accuracy between male and female patients) between our approach and the upper boundary. The learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on different numbers of human expert predictions $l$ and artificial expert predictions are displayed for radiologists 4295342357 and 4295194124. We report mean and standard deviation over five seeds.

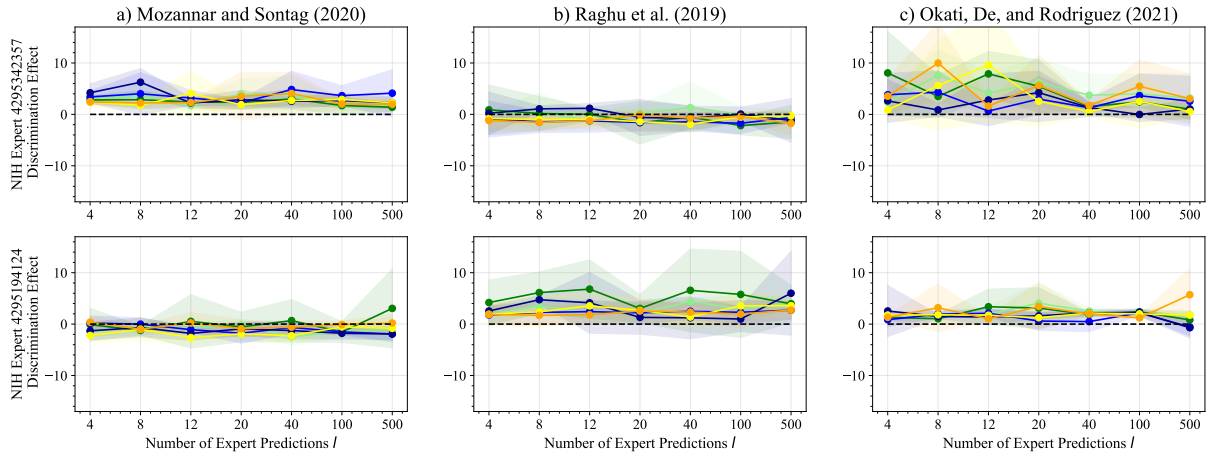


Figure C10: Differences in the discrimination effect for patients' age (mean absolute deviation of the accuracy of each age bin from the overall accuracy) between our approach and the upper boundary. The learning to defer algorithms of a) Mozannar and Sontag (2020), b) Raghu et al. (2019), and c) Okati, De, and Rodriguez (2021) trained on different numbers of human expert predictions $l$ and artificial expert predictions are displayed for radiologists 4295342357 and 4295194124. We report mean and standard deviation over five seeds.

## References


He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 770–778.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report.

Li, X.; Sun, Q.; Liu, Y.; Zhou, Q.; Zheng, S.; Chua, T.-S.; and Schiele, B. 2019. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*.

Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S. J.; and Yang, Y. 2019. Learning to propagate labels: transductive propagation network for few-shot learning. In *International Conference on Learning Representations*.

Majkowska, A.; Mittal, S.; Steiner, D. F.; Reicher, J. J.; McKinney, S. M.; Duggan, G. E.; Eswaran, K.; Cameron Chen, P.-H.; Liu, Y.; Kalidindi, S. R.; et al. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2): 421–431.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 7076–7087.

Okati, N.; De, A.; and Rodriguez, M. 2021. Differentiable learning under triage. *Advances in Neural Information Processing Systems*.

Raghu, M.; Blumer, K.; Corrado, G.; Kleinberg, J.; Obermeyer, Z.; and Mullainathan, S. 2019. The algorithmic automation problem: prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*.

Tan, M.; and Le, Q. 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Conference on Computer Vision and Pattern Recognition*, 2097–2106.

Yu, Z.; Chen, L.; Cheng, Z.; and Luo, J. 2020. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Conference on Computer Vision and Pattern Recognition*, 12856–12864.