IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received xxx xx, xxxx, accepted xxx xx, xxxx, date of publication xxx xx, xxxx, date of current version xxx xx, xxxx.

Digital Object Identifier xx.xxxx/ACCESS.xxxx.xxxxxxx

Learning from Multiple Expert Annotators for Enhancing Anomaly Detection in Medical Image Analysis

KHIEM H. LE[2,3,†], TUAN V. TRAN[2,3,†], HIEU H. PHAM [2,3], HIEU T. NGUYEN[1], TUNG T. LE[1], and HA Q. NGUYEN[1],

[1]Smart Health Center, VinBigData JSC, Hanoi, Vietnam
[2]VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[3]College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam

Corresponding author: Hieu H. Pham (e-mail: hieu.ph@vinuni.edu.vn)

ABSTRACT Recent years have experienced phenomenal growth in computer-aided diagnosis systems based on machine learning algorithms for anomaly detection tasks in the medical image domain. However, the performance of these algorithms greatly depends on the quality of labels since the subjectivity of a single annotator might decline the certainty of medical image datasets. In order to alleviate this problem, aggregating labels from multiple radiologists with different levels of expertise has been established. In particular, under the reliance on their own biases and proficiency levels, different qualified experts provide their estimations of the "true" bounding boxes representing the anomaly observations. Learning from these nonstandard labels exerts negative effects on the performance of machine learning networks. In this paper, we propose a simple yet effective approach for the enhancement of neural networks' efficiency in abnormal detection tasks by estimating the actually hidden labels from multiple ones. A re-weighted loss function is also used to improve the detection capacity of the networks. We conduct an extensive experimental evaluation of our proposed approach on both simulated and real-world medical imaging datasets, MED-MNIST and VinDr-CXR. The experimental results show that our approach is able to capture the reliability of different annotators and outperform relevant baselines that do not consider the disagreements among annotators. Our code is available at https://github.com/huyhieupham/learning-from-multiple-annotators.

I. INTRODUCTION

The recent success of computer-aided diagnosis (CAD) systems can be attributed to the emergence of supervised learning algorithms [1, 2, 3, 4, 5, 6] and the availability of large-scale human-labeled datasets [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. These systems have been playing a significant role as clinicians' assistants in their decision-making process when analyzing medical images or making an assessment of the patient's condition [22, 23, 24]. One of the indispensable factors contributing to this accomplishment is the need for high-quality labeled datasets. In fact, in order to possess these datasets with gold standard labels for training, there are various costs associated with the collecting procedures, including costs of cleaning data, diversifying data, obtaining expert labeling of data, and so on, which seems to be infeasible and economically unjustified to apply those in many healthcare-related tasks. Additionally, if only one expert makes annotations for medical datasets, this is more susceptible to subjectivity, which might increase the uncertainty of those datasets. Instead, in order to alleviate the subjective criteria in the medical-imaging labeling process, we may have multiple labels provided by different clinicians or professional annotators, and this is considered *repeated labeling* [25, 8, 26]. In practice, there might exist a substantial amount of disagreement among the clinicians since each of them visually examines the medical images and provides a subjective version of the gold standard labels. This problem causes high inter-reader variability [27, 28, 29], which could come at the significant expense of machine-learning-based models' performance if those annotations are used as ground truths blindly.

Many previous approaches have been done to mitigate the effects of inter-observer variations on annotation procedures, which can be divided into two main groups: (i) simultaneous

†Equal contribution.

approach and (ii) two-stage approach. In the first category, models aim to curate labels and learn a supervised network jointly in an end-to-end fashion. In contrast, in the second category, the actual estimated labels are first curated from multiple ones [30], which is known as *truth inference*, and subsequently, a supervised model is trained on these labels. All of those approaches have achieved impressive results on both classification and segmentation tasks [31, 32]. However, to the best of our knowledge, very little attention has been drawn to the same problem in detection tasks. In this paper, the key consideration of our work is to propose the two-stage approach to addressing such a problem in supervised abnormal detection tasks, where we have multiple annotators providing a set of possibly nonstandard labels, but no absolute gold standard. In particular, the key to our proposed method is to allow deep learning-based detectors to give an estimate of the actual hidden labels with the aim of improving their detection capacity of abnormalities from chest X-ray scans.

As outlined in Figure 1, the proposed detection method encompasses two stages that estimate the reliability of multiple annotators and allows a deep learning network to learn from these curated labels. In the first stage (on the left-handed side), given a medical image and a set of expert annotations, a Weighted Boxes Fusion (WBF) algorithm [33] is leveraged to estimate the true labels and their confidence scores, which is generally regarded as the *Truth Inference* step. The second stage, as shown in the right-handed side of Figure 1, is to train an object detector on estimated labels with a re-weighted loss function using implicit annotators' agreement, which is represented by the estimated confidence scores. For evaluation, we first simulate and test the proposed approach on a multiple-experts-detection dataset from MNIST [34], called MED-MNIST. After that, we demonstrate the potential of a real-world chest X-ray (CXR) dataset with radiologists' annotations, namely VinDr-CXR. In comparison to two baselines: (**i**) treating all of the nonstandard labels blindly as the ground truth, and (**ii**) ensembling the models supervised by individual expert annotations, our proposed method provides better detection performance in terms of mAP scores.

To summarize, the main contributions of the article are as follows:

- We introduce a simple yet effective method that allows a deep-learning network to learn from multiple annotators. Specifically, the proposed approach aims at estimating the true annotations with confidence scores from multiple ones and then using those to train a deep learning-based detector with a re-weighted loss function. This helps eliminate uncertainty in the learning process and provides higher label quality to train predictive models.
- We evaluate the proposed approach on both simulated and real medical imaging datasets and find significant performance improvements compared to the baseline approaches. We also release the used CXR dataset, which is available at https://vindr.ai/datasets/cxr.

The rest of this paper is organized as follows. Related works on learning from multiple annotators and weighted training techniques are reviewed in Section II. Section III presents the details of the proposed method with a focus on how to estimate the ground truth annotations. Section IV provides extensive experiments on a simulated object detection dataset and a real-world chest X-ray dataset. Finally, section V discusses the experimental results, some key findings, and limitations of this work, and concludes the paper.

II. RELATED WORKS

In this section, we investigate and discuss some research directions and existing works that are highly related to our work, including learning from multiple annotators and weighted training techniques.

Learning from multiple annotators. In classification and segmentation problems, a plethora of different works has shown the potential in reducing the degree of annotator disagreement by estimating the actual labels. Such approaches can be categorized into two groups: two-stage approaches [35, 30, 36] and one-stage approaches [32, 37, 38]. Two-stage approaches first infer the true labels from various ones, then train a model using the curated ones. The most basic solution for integrating the information from multiple labels is based on majority voting (MV) [39], in which the majority annotations are treated as the ground truth. However, such approaches have a potentially serious drawback: MV natively eliminates the uncertainty of the majority labels, and as a consequence, the generated single label would be suboptimal because of its bias. Other approaches incorporate additional information into the truth inference procedure, such as the annotators' proficiency [40], the annotators' confusion matrix [41, 42], or the difficulty of each sample [43]. Two-stage approaches have the advantage of simplicity due to the single-task manner of each stage, but they might not fully exploit the raw annotations. One-stage approaches or jointly learning approaches simultaneously estimate the hidden true labels and learn the desired model from possibly noisy labels of multiple annotators by formulating a multi-task problem. Earlier approaches [44, 45] explore the Expectation Maximization (EM) algorithm for jointly modeling the annotators' ability and the latent ground truth. Several recent approaches employ end-to-end frameworks which enable the neural networks to learn directly from the noisy labels by using a [46], and further developed by incorporating annotators' confusion matrix [31, 32], or instance features [37]. However, to the best of our knowledge, our proposed method is the first one that aims to handle the detection task.
Weighted training examples. Previous works on the use of weighted training examples can be divided into two groups: (i) emphasize hard examples and (ii) emphasize easy examples. Methods in group (i) include hard-example

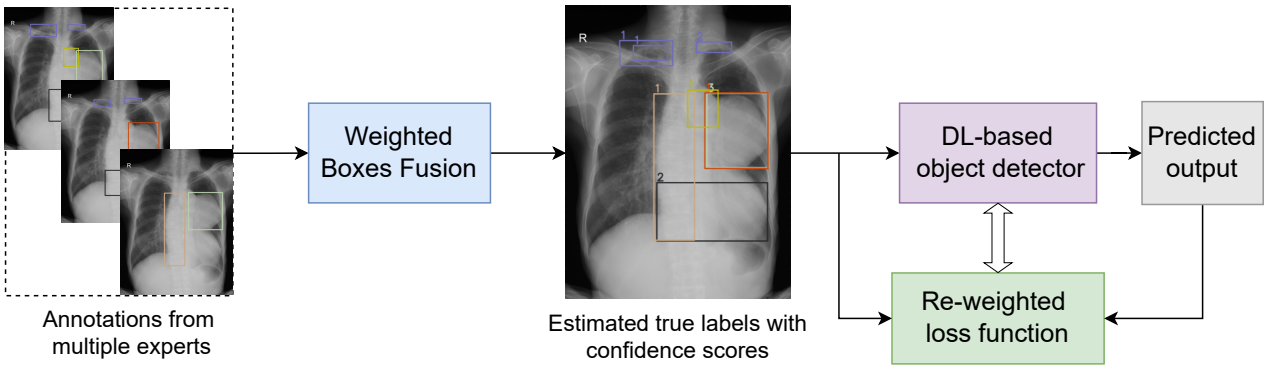


FIGURE 1: Illustration of the proposed approach that aims to build a deep learning system for abnormal detection on medical scans from multiple expert annotators. The training process contains two stages. The first stage focuses on truth inference, in which it estimates the true labels using the WBF algorithm [33] with the implicit annotator's agreement as confidence scores. The second uses the estimated confidence scores to train a deep learning-based detector using a re-weighted object detection loss function. To provide abnormality analysis during the testing phase, only the fully trained image detector is required.

mining [47, 48], which is a bootstrapping technique over the difficult examples; boosting algorithms [49], where the misclassified examples in preceding weak classifiers are assigned with higher weights; and focal loss [50] that addresses class imbalance problems by adding a regulator to the cross-entropy loss for focusing on hard negative examples. Approaches in group (ii) are instances of broader topics such as curriculum learning [51], which is biologically inspired by gradual human learning, with easier examples preferred in early training stages; learning with noisy labels [52, 53], which prefers examples with smaller training losses as they are more likely to be clean.

Unlike any approaches above, we propose in this paper a re-weighted loss function that assigns more weights to more confident examples that determine by the consensus of multiple annotators. Our experimental results validate the correctness of this hypothesis.

III. PROPOSED METHOD

We describe in this section our main contribution which is a framework to enhance anomaly detection from medical images via multiple annotators. After estimating the hidden actual labels, the framework allows the object detector to supervisely learn from these estimated labels.

A. PROBLEM FORMULATION

Given a set of N training images $\mathbf{X} = \{x_i\}_{i=1}^{N}$ with corresponding bounding box annotations $\mathbf{y} = \left\{\tilde{y}_i^{(r)}\right\}_{i=1}^{N}$, where $\tilde{y}_i^{(r)}$ representing the bounding box label for the example $\mathbf{x}_i$ provided by r^{th} annotator in a set of R multiple annotators. In this work, we use labels $\left\{\tilde{y}_i^{(r)}\right\}_{i=1}^{N}$ to estimate a single set of actual labels with corresponding confidence scores $\{\mathbf{y}_i; c_i\}_{i=1}^{N}$, then a supervised object detector is trained with these estimated labels by using the proposed re-weighted loss function. In order to evaluate the effectiveness of the proposed method, we use a gold-standard test set containing M examples $\mathcal{T} = \left\{\left(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\right)\right\}_{j=1}^{M}$. In medical imaging scenarios, where the true labels are not available, we obtain the gold-standard test labels $\mathbf{y}^{(j)}$ from the consensus of a group of competent radiologists. Figure 1 shows an overview of our method.

B. ESTIMATING THE ACTUAL LABELS

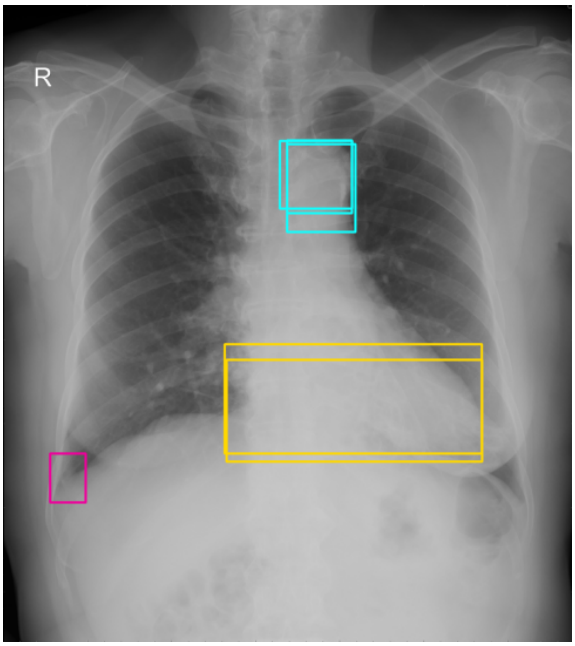
We estimate the actual labels using Weighted Boxes Fusion (WBF) algorithm [33]. This technique is originally used for combining predictions from multiple sources, e.g., predictions from different detection models. We modify the WBF algorithm and describe it in Algorithm 1, where the bounding box labels of different annotators are combined into a single set of bounding boxes with corresponding confidence scores. They are then used to train the image detectors with a re-weighted loss function. The qualitative results of Algorithm 1 on the VinDr-CXR dataset are illustrated in Figure 2. The greater agreement between annotators (two or three annotators have the same diagnosis for a finding on the image) reach, the more correct the fused bounding box is.

C. TRAINING METHODOLOGY

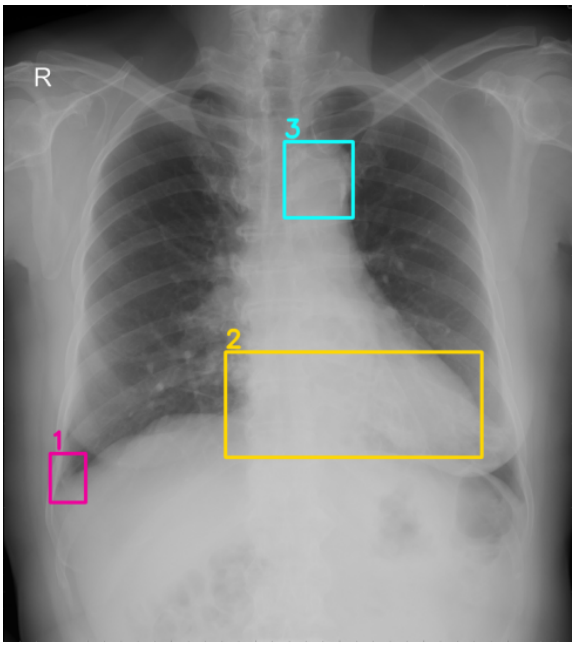
Object detection is a multi-task problem where two loss functions are used: (1) the localization loss $\mathcal{L}_{\text{loc}}$ for predicting bounding box offsets and (2) the classification loss $\mathcal{L}_{\text{cls}}$ for predicting conditional class probabilities. In this work, we focus on one-stage anchor-based detectors. A general form of the loss function for those detectors can be written as

$$\begin{aligned} \mathcal{L}\left(p, p^*, t, t^*\right) &= \mathcal{L}_{cls}\left(p, p^*\right) + \beta I(t) \mathcal{L}_{loc}\left(t, t^*\right) \\ I(t) &= \begin{cases} 1 & \text{if IoU}\left\{a, a^*\right\} > \eta \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

where t and t^* are the predicted and ground truth box coordinates, p and p^* are the class category probabilities, respectively; $\text{IoU}\left\{a, a^*\right\}$ denotes the Intersection over Union (IoU) between the anchor a and its ground truth a^*; η is



(a) The original annotations provided by multiple radiology experts. The same abnormal finding is represented by the sample color.



(b) Fused boxes with corresponding confidence scores after applied the WBF algorithm.

FIGURE 2: (a) Visualization of multiple expert annotations on a chest X-ray example from the VinDr-CXR dataset [26] and (b) the fused boxes with confidence scores obtained by the WBF algorithm.

Algorithm 1: The WBF algorithm applied for multiple expert annotations

Input: An image $\mathbf{x}$ with a list of annotations $\tilde{y}$ given by a set $\mathbb{S}(R)$ of R experts. The expert $r \in \mathbb{S}(R)$ with proficiency p_r provides the annotations including r_x boxes, $A_r = [\text{box}_1, \ldots, \text{box}_{r_x}]$. All of the experts' annotations being merged into a list A.
Output: A list of k fused boxes $F = [\text{box}_1, \ldots, \text{box}_k]$.

1 Declare empty lists L and F for boxes clusters and fused boxes, respectively. Each position in the list L can have a cluster of boxes or a single box. Each position in F has only one box, which is the fused box from the corresponding cluster in L.

2 Iterate through all boxes in A in a cycle and attempt to find a matching box in the list F. Two boxes are defined matched if they have a high degree of overlap (e.g. IoU > 0.4). If there are more than one matching boxes in F, the one with the highest IoU will be chosen.

3 If the matching box is not found in step 1, add the current box to L and F as new entry for the new cluster before moving on to the next box in the list A.

4 If the match is found in step 1, add this box to the list L at the position pos which corresponds to the matching box in the list F

5 Set the fused box's coordinates $F[pos]$ to be the weighted average of T boxes accumulated in cluster $L[pos]$ with the following formulas:

$$x_{1,2} \coloneqq \frac{\sum_{i=1}^{T} p_i x_{1,2}}{\sum_{i=1}^{T} p_i}$$

$$y_{1,2} \coloneqq \frac{\sum_{i=1}^{T} p_i y_{1,2}}{\sum_{i=1}^{T} p_i}$$

6 Set the the fused boxes' confidence scores in F to the number of boxes in the corresponding cluster in L once all boxes in A have been processed.

$$c \coloneqq c \min(T, N)$$

The fused boxes with confidence scores now represent the annotators' level of agreement.

an IoU threshold for objectness, i.e. the confidence score of whether there is an object or not; β is a constant for balancing two loss terms $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ [54].

We use fused boxes confidence scores c_k^i obtained from Algorithm 1 to get a re-weighted loss function that emphasizes boxes with high annotators agreement. The new loss function, which we name Experts Agreement Re-weighted Loss (EARL) can now be written as

$$\mathcal{L}(p, p^*, t, t^*) = c\mathcal{L}_{cls}(p, p^*) + c\beta I(t)\mathcal{L}_{loc}(t, t^*), \quad (2)$$

IV. EXPERIMENTS

We validate the proposed method in both synthetic and real-world scenarios: (1) the MED-MNIST, an object

detection dataset, which was simulated from MNIST [34] with multiple annotations; (2) VinDr-CXR [26], a chest X-ray dataset with labels provided by multiple radiologists. In the following sections, we describe those two datasets and our experiment setup, as well as experimental results.

A. DATASETS

1) MED-MNIST Dataset

Based on MNIST [34] – a database of handwritten digits, we synthesize a multiple-experts-detection dataset, called MED-MNIST, in two steps: (1) in order to generate a dataset for the detection task, we randomly merge various digits into black-background images, where each digit regarded as an object with a corresponding bounding box (as visualized in Figure 3a), (2) multiple annotations are assigned to each object representing the different opinions of experts. In the case of this simulation dataset, we make an assumption that those experts have the same proficiency p. These annotations are generated with two values: (i) class labels and (ii) object coordinates. In order to synthesize the expert annotations on class labels, a unique transition matrix $A_k(k \in \{1, \dots, R\})$ is generated for each expert E_k to represent the expert misclassification through probability distributions. We further use an additional class, namely `no_obj`, for simulating the false negative mistakes. The exemplars of a transition matrix are visualized in Figure 4. Regarding the object coordinate perturbations, we replicate the bounding box annotations by randomizing the bounding boxes that are highly overlapping with the given true bounding box. Both factors (i) and (ii) are controlled by proficiency p. Specifically, A_k are diagonally dominant ($a_{ii} > a_{ij}$ for all $i \neq j$), and $a_{ii} = \min(\max(\zeta, \alpha), 1)$ with $\alpha \sim \mathcal{N}(p, \sigma)$, ζ and σ are hyperparameters and set to 0.05 and 0.5, respectively. The simulated bounding boxes are subject to *IoU* with the true bounding box being larger than p. In particular, the number of expert annotations per sample R is 3, and the proficiency p is 0.8. The simulated MED-MNIST dataset consists of 5,000 samples for training, 1,000 for hold-out validation, and 1,000 for testing.

2) VinDr-CXR Dataset

VinDr-CXR [26] is the largest public chest X-ray database with radiologist-generated annotations. It consists of 18,000 chest X-ray scans, with 15,000 for training and 3,000 for testing sets, all of which have labels of both the localization of abnormal areas and the classification of common thoracic diseases. In practice, the annotations were obtained by a group of 17 radiologists who have at least eight years of experience. Each image in the training group was independently labeled by three radiologists, while that in the testing set were meticulously treated and obtained by the consensus of 5 radiologists. Several samples from the VinDr-CXR dataset are shown in Figure 5.

3) Rads-VinDr-CXR Dataset

One idiosyncratic characteristic of the VinDr-CXR dataset [26] is that 94.28% of the abnormal scans in the training set (3,315 out of 3,516) were annotated by a group of three radiologists with their correspondence IDs being *R8*, *R9* and *R10*. As a result, we generate Rads-VinDr-CXR, a sub-dataset that includes only samples annotated by those three radiologists. The Rads-VinDr-CXR is appropriate to validate the proposed approach.

B. EXPERIMENTAL SETTINGS

1) Evaluation metric

For all experiments, we validate the detection performance using the standard mean average precision metric at a threshold of 0.4 (mAP@0.4) [55]. Specifically, a predicted object is a true positive if it has an IoU of at least 0.4 with a ground truth bounding box. The average precision (AP) is the mean of 101 precision values, corresponding to recall values ranging from 0 to 1 with a step size of 0.01. The final metric is the mean of AP overall lesion categories. We also employ mAP@[0.5:0.95:0.05] as an additional metric to assess the model's performance on different IoU thresholds, ranging from 0.5 to 0.95 with a step size of 0.05.

2) Implementation Details

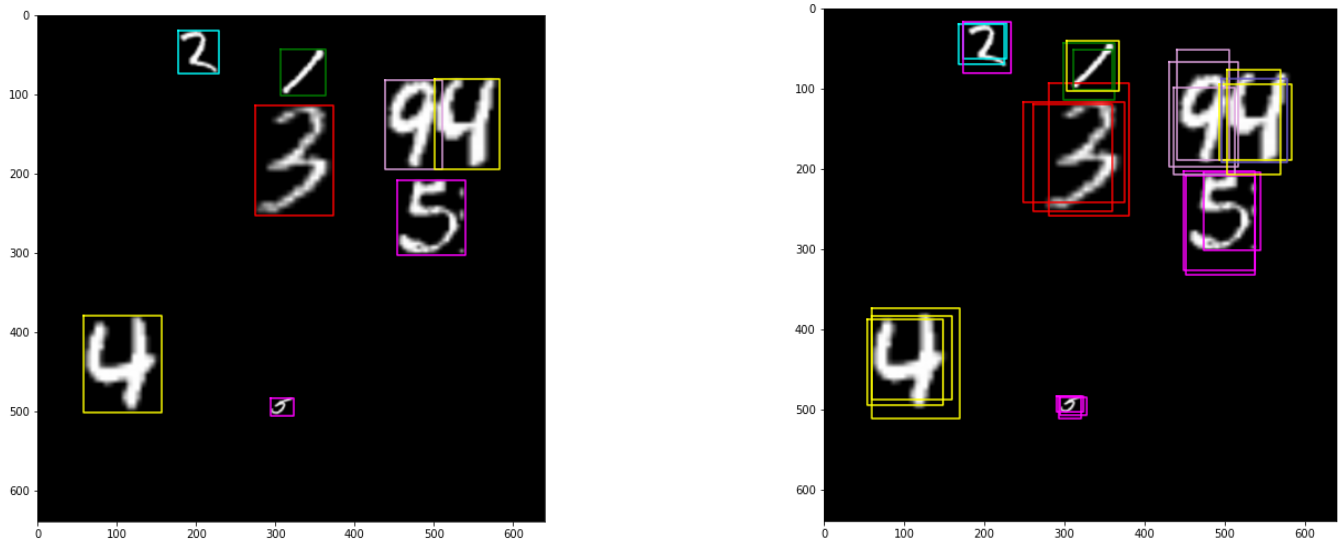
The main detector used in our experiments is YOLOv5-S [56]. The network is built with PyTorch 1.7.1 and trained on two NVIDIA RTX 2080 Ti GPUs. All training and testing images are resized to the dimension of 640×640 pixels. The detector is trained for 50 epochs with 1 cycle learning rate decay [57] using the SGD optimizer [58]. The initial learning rate is set to 1e-3. To validate the robustness of the proposed approach across different deep learning detectors, we further train and evaluate EfficientDet [59] with sizes D3 and D4. Specifically, all images are resized to 640×640 pixels, and the model is trained for 30 epochs with a constant learning rate 3e-4 using the AdamW optimizer [60].

3) Comparison with state-of-the-art methods

To the best of our knowledge, there is no existing multiple-annotators model for object detection tasks in the literature. We compare the performance of our proposed method against two baselines: a) assuming all of the annotators' opinions as the ground truth; b) an ensemble of independent models trained on separate radiologists' annotation sets [61]. On the Rads-VinDr-CXR dataset, we further compare our method with the Rads-ensemble, which is the ensemble of independent models trained on separate radiologists' annotation sets. In this case, the WBF algorithm is used to combine the predictions of those models.

C. EXPERIMENTAL RESULTS

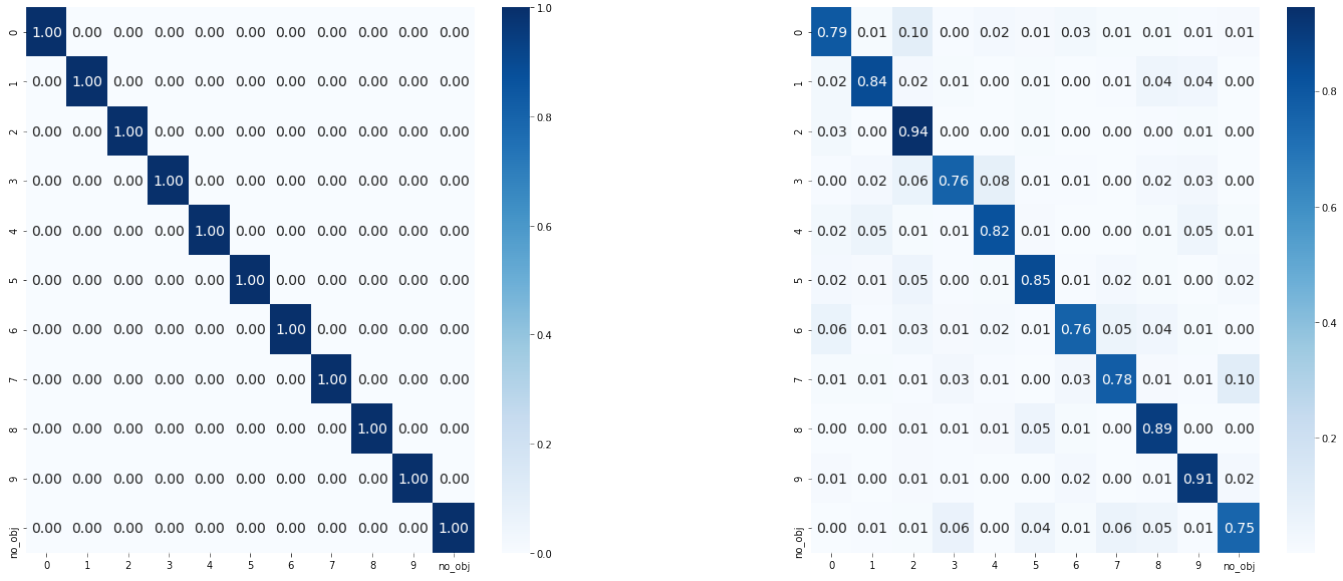
Table 1 and Table 2 report the experimental results of the YOLOv5-S detector on MED-MNIST and VinDr-CXR datasets, respectively. On both synthetic and real-world



(a) MNIST Detection

(b) Simulated expert annotations

FIGURE 3: The MED-MNIST dataset with multiple expert annotations, obtained by perturbing boxes and classes from the MNIST dataset [34].



(a) Original transition matrix

(b) Simulated expert transition matrix

FIGURE 4: Visualization of the original and synthesized transition matrices. To simulate the false negative scenario, we use an additional class called `no_obj`.

datasets, the proposed approach outperforms the baselines, even with the ensemble of individual experts' models. Specifically, on the test set of the MED-MNIST dataset, our method reports an overall mAP@0.4 of 0.980 and an overall mAP@[0.5:0.95:0.05] of 0.849. These results are much higher than the performance of the baseline with mAP@0.4 = 0.975 and mAP@[0.5:0.95:0.05] = 0.815, boosting the mAP scores of the baseline by 0.51% and 4.2%, respectively. Experimental results on the VinDr-CXR and Rads-VinDr-CXR datasets also validate the effectiveness of the proposed method. We achieve an overall mAP@0.4 of 0.200 on the VinDr-CXR dataset and an overall mAP@0.4 of 0.158 on the Rads-VinDr-CXR dataset. We emphasize that these results outperform both the baseline model, the individual model trained on the label provided by the individual annotator (i.e. *R8*, *R9*, *R10*), as well as the ensemble model.

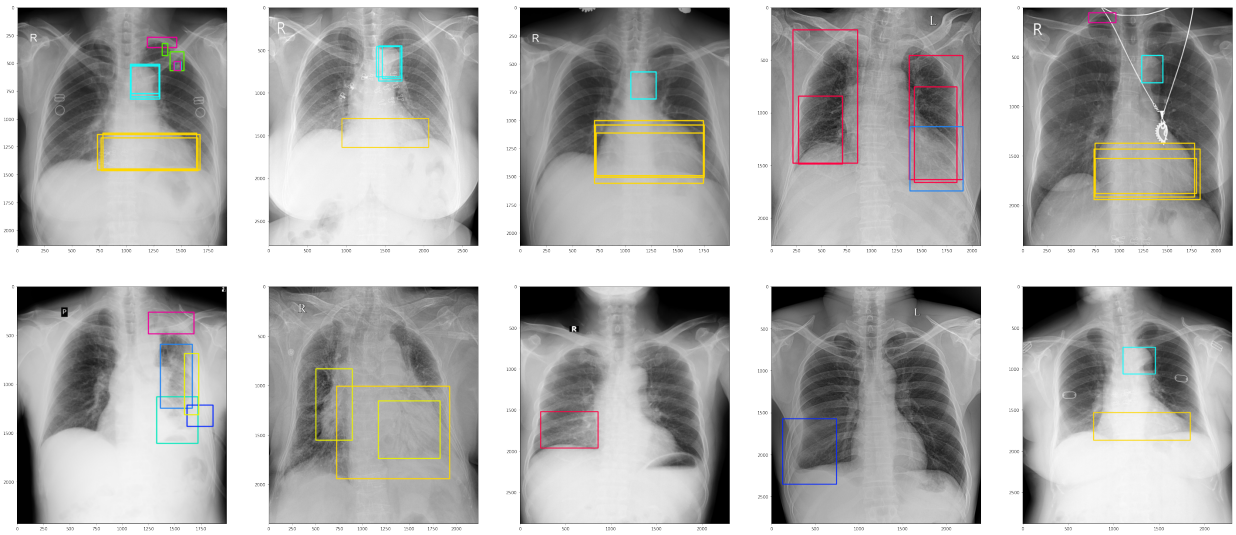


FIGURE 5: Visualization of abnormal findings (different bounding box colors represent different findings) from the VinDr-CXR dataset: (top) Each scan in the training set was annotated by three different radiologists; (bottom) Test set annotations were obtained from the consensus of five radiologists.

The experimental results with the EfficientDet detector are provided in Table 3. We found that better detection performances compared to the baseline have been reported. This evidence confirms the robustness of the proposed approach across deep learning detectors.

TABLE 1: Experimental results on the MED-MNIST dataset. The highest scores are highlighted in red.

Method	mAP@0.4	mAP@[0.5:0.95:0.05]
Baseline	0.975	0.815
WBF+EARL (ours)	0.980	0.849

TABLE 2: Experimental results on the VinDr-CXR and Rads-VinDr-CXR datasets with the YOLOv5-S detector. The highest scores are highlighted in red.

Dataset	Method	mAP@0.4
VinDr-CXR	Baseline	0.190
	WBF+EARL (ours)	0.200
Rads-VinDr-CXR	Baseline	0.148
	R8	0.121
	R9	0.132
	R10	0.124
	Rads-ensemble [61]	0.154
	WBF+EARL (ours)	0.158

TABLE 3: Experimental results on the VinDr-CXR dataset while EfficientDet is used as the detector. The scores are measured in mAP@[0.5:0.95:0.05], with highest values highlighted in red.

	Baseline	WBF+EARL
EfficientDet-D3	0.1142	0.1353
EfficientDet-D4	0.1223	0.1431

V. DISCUSSIONS

To the best of our knowledge, the proposed method is the first effort to train an image detector from labels provided by multiple annotators, which is crucial in constructing high-quality CAD systems for medical imaging analysis. In particular, we empirically showed a notable improvement in terms of mAP scores by estimating the true labels and then integrating the implicit annotators' agreement into the loss function to emphasize the accurate bounding boxes over the imprecise ones. The idea is simple but effective, allowing the overall framework can be applied in training any machine learning-based detectors.

Despite the fact that the proposed method has a higher predictive performance than the relevant baselines, we acknowledge that the proposed method has some limitations. First, the overall architecture is not end-to-end. It may not fully exploit the benefits of combining truth inference and training the desired image detector. Second, applying the WBF algorithm to annotation sets with a high level of noise may produce low-quality training data. This case is quite impractical in the medical imaging field when the annotators are competent clinical experts, but it frequently occurs in the general *learning from crowds* problems.

We have planned to advance the current research by exploring several potential directions. Firstly, the two stages of the proposed method can be made interactive by replacing the averaging operator in the WBF algorithm with the weighted average, where the weights can be learned under supervision. In this way, the first stage can directly receive feedback signals from the second stage. Furthermore, the proposed method's robustness to noisy annotations and generalization to out-of-sample distributions will be thoroughly examined.

VI. CONCLUSION

We introduced a framework for supervised object detection models to learn from multiple annotators by estimating the actual labels beforehand. We leveraged the Weighted Boxes Fusion (WBF) algorithm to obtain the aggregated annotations with the implicit annotators' agreement as confidence scores. The estimated annotations are then used to train a deep learning detector with a re-weighted loss function that incorporates the confidence scores for localizing abnormal findings more accurately. We demonstrated that the proposed approach outperforms current state-of-the-art baselines in both synthetic and real-world scenarios.

VII. ACKNOWLEDGEMENTS

This work was supported by Smart Health Center at VinBigData JSC. The authors gratefully acknowledge all anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. Medical Image Analysis, 42:60–88, 2017.

[2] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. Medical Physics, 47(5):e218–e227, 2020.

[3] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. CheXnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225, 2017.

[4] F Pasa, V Golkov, F Pfeiffer, D Cremers, and D Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. Scientific reports, 9(1):1–9, 2019.

[5] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology, 292(1):60–66, 2019.

[6] Hieu H Pham, Ha Q Nguyen, Hieu T Nguyen, Linh T Le, and Lam Khanh. An accurate and explainable deep learning system improves interobserver agreement in the interpretation of chest radiograph. arXiv preprint arXiv:2208.03545, 2022.

[7] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042, 2019.

[8] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 590–597, 2019.

[9] Hieu T Nguyen, Hieu H Pham, Nghia T Nguyen, Ha Q Nguyen, Thang Q Huynh, Minh Dao, and Van Vu. Vindr-spinexr: A deep learning framework for spinal lesions detection and classification from radiographs. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 291–301. Springer, 2021.

[10] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. Radiology: Artificial Intelligence, 2(3):e190211, 2020.

[11] Ngoc H Nguyen, Hieu H Pham, Thanh T Tran, Tuan NM Nguyen, and Ha Q Nguyen. Vindr-pcxr: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. arXiv preprint arXiv:2203.10612, 2022.

[12] Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865, 2020.

[13] H. Hieu Pham, T. Thanh Tran, and H. Quy Nguyen. Vindr-pcxr: An open, large-scale pediatric chest x-ray dataset for interpretation of common thoracic diseases (version 1.0.0). PhysioNet. https://doi.org/10.13026/k8qc-na36., 2022.

[14] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. MedRxiv, 2020.

[15] Hieu Trung Nguyen, Ha Quy Nguyen, Hieu Huy Pham, Khanh Lam, Linh Tuan Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. medRxiv, 2022.

[16] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M Nasrallah, Theodore D Satterthwaite, Yong Fan, et al. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage, 208:116450, 2020.

[17] Hieu Huy Pham, Hieu Nguyen Trung, and Ha Quy Nguyen. Vindr-mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography (version 1.0.0). PhysioNet. https://doi.org/10.13026/br2v-7517., 2022.

[18] Hieu H Pham, Dung V Do, and Ha Q Nguyen. Dicom imaging router: An open deep learning framework for

classification of body parts from dicom x-ray scans. medRxiv, 2021.

[19] David Wallis and Irène Buvat. Clever hans effect found in a widely used brain tumour mri dataset. Medical Image Analysis, 77:102368, 2022.

[20] Hoang C Nguyen, Tung T Le, Hieu H Pham, and Ha Q Nguyen. Vindr-ribcxr: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest x-rays. arXiv preprint arXiv:2107.01327, 2021.

[21] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering, 63(7):1455–1462, 2015.

[22] Garry Choy, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E Samir, Oleg S Pianykh, J Raymond Geis, Pari V Pandharipande, James A Brink, and Keith J Dreyer. Current applications and future impact of machine learning in radiology. Radiology, 288(2):318, 2018.

[23] Martina Gurgitano, Salvatore Alessio Angileri, Giovanni Maria Rodà, Alessandro Liguori, Marco Pandolfi, Anna Maria Ierardi, Bradford J Wood, and Gianpaolo Carrafiello. Interventional radiology ex-machina: Impact of artificial intelligence on practice. La radiologia medica, 126(7):998–1006, 2021.

[24] Ngoc Huy Nguyen, Ha Quy Nguyen, Nghia Trung Nguyen, Thang Viet Nguyen, Hieu Huy Pham, and Tuan Ngoc-Minh Nguyen. Deployment and validation of an ai system for detecting abnormal chest radiographs in clinical settings. Frontiers in Digital Health, page 130, 2022.

[25] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps, pages 323–350, 2018.

[26] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. arXiv preprint arXiv:2012.15029, 2020.

[27] T. Watadani, F. Sakai, T. Johkoh, S. Noma, M. Akira, K. Fujimoto, A. A. Bankier, K. S. Lee, N. L. Müller, J. W. Song, J. S. Park, D. A. Lynch, D. M. Hansell, M. Remy-Jardin, T. Franquet, and Y. Sugiyama. Interobserver variability in the CT assessment of honeycombing in the lungs. Radiology, 266(3):936–944, Mar 2013.

[28] A. B. Rosenkrantz, R. P. Lim, M. Haghighi, M. B. Somberg, J. S. Babb, and S. S. Taneja. Comparison of interreader reproducibility of the prostate imaging reporting and data system and likert scales for evaluation of multiparametric prostate MRI. American Journal of Roentgenology, 201(4):W612–618, Oct 2013.

[29] E. Lazarus, M. B. Mainiero, B. Schepps, S. L. Koelliker, and L. S. Livingston. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology, 239(2):385–391, May 2006.

[30] Victor S. Sheng and Jing Zhang. Machine learning with crowdsourcing: A brief summary of the past research and future directions. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):9837–9843, Jul. 2019.

[31] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 15750–15762. Curran Associates, Inc., 2020.

[32] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11244–11253, 2019.

[33] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing, 107:104117, Mar 2021.

[34] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

[35] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? Proc. VLDB Endow., 10(5):541–552, January 2017.

[36] Yuan Jin, Mark Carman, Ye Zhu, and Yong Xiang. A technical survey on statistical modelling and design methods for crowdsourcing quality control. Artificial Intelligence, page 103351, 2020.

[37] Jingzheng Li, Hailong Sun, Jiyi Li, Zhijun Chen, Renshuai Tao, and Yufei Ge. Learning from multiple annotators by incorporating instance features, 2021.

[38] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 15750–15762. Curran Associates, Inc., 2020.

[39] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, page 614–622, New York, NY, USA, 2008. Association for Computing Machinery.

[40] David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems.

Advances in Neural Information Processing Systems, 24, 2011.

[41] A. Dawid and A. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of The Royal Statistical Society Series C-applied Statistics, 28:20–28, 1979.

[42] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. In Proceedings of the 7th International Conference on Neural Information Processing Systems, NIPS'94, page 1085–1092, Cambridge, MA, USA, 1994. MIT Press.

[43] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in Neural Information Processing Systems, 22:2035–2043, 2009.

[44] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, page 889–896, New York, NY, USA, 2009. Association for Computing Machinery.

[45] Anil Ramakrishna, Rahul Gupta, Ruth B Grossman, and Shrikanth S Narayanan. An expectation maximization approach to joint modeling of multidimensional ratings derived from multiple annotators. In Interspeech, pages 1555–1559, 2016.

[46] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, volume 32, 2018.

[47] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages 886–893 vol. 1, 2005.

[48] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 761–769, 2016.

[49] Robert E Schapire. Explaining AdaBoost. In Empirical Inference, pages 37–52. Springer, 2013.

[50] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 2980–2988, 2017.

[51] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[52] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3):107–115, 2021.

[53] Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Proceedings of the 34th International Conference on Machine Learning, ICML'17, page 233–242. JMLR.org, 2017.

[54] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055, 2019.

[55] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 88(2):303–338, 2010.

[56] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. arXiv, April 2021.

[57] Leslie N Smith. Cyclical learning rates for training neural networks. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 464–472, 2017.

[58] Sebastian Ruder. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747, 2016.

[59] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10781–10790.

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[61] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. Proceedings of the AAAI Conference on Artificial Intelligence, 32, 03 2017.

KHIEM H. LE Khiem H. Le got the Honors Bachelor Degree in Mathematics and Computer Science from the Vietnam National University, Ho Chi Minh City. He was a Research Intern at Smart Health Center, VinBigData JSC, and currently is a Research Assistant at College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam. His research interests include Computer Vision, Deep Learning and AI for Biomedical Applications.

TUAN V. TRAN Tuan V. Tran got the Bachelor Degree of Engineering in Telecommunications Engineering from Vietnam National University, Ho Chi Minh City. He was a Research Intern at Smart Health Center, VinBigData JSC, and currently is a Research Assistant at VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam. His research interests include Deep Learning Applications for Computer Vision, AI Secure and Federated Learning.

HIEU H. PHAM Dr. Pham Huy Hieu is an Assistant Professor at the College of Engineering and Computer Science (CECS), VinUniversity, and serves as Associate Director at VinUni-Illinois Smart Health Center. He received his Ph.D. in Computer Science from the Toulouse Computer Science Research Institute (IRIT), University of Toulouse, France, in 2019. Previously, he earned the Degree of Engineer in Industrial Informatics from Hanoi University of Science and Technology (HUST), Vietnam, in 2016. His research interests include Computer Vision, Machine Learning, Medical Image Analysis, and their applications in Smart Healthcare. He is the author, co-author of 30 scientific articles appeared in about 20 conferences and journals such as Computer Vision and Image Understanding, Neurocomputing, Scientific Data (Nature), Frontiers in Digital Health, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Medical Imaging with Deep Learning (MIDL), IEEE International Conference on Image Processing (ICIP), Asian Conference on Computer Vision (ACCV), and IEEE International Conference on Computer Vision (ICCV). He is also currently serving as Reviewers for MICCAI, ICCV, CVPR, IET Computer Vision Journal (IET-CVI), IEEE Journal of Biomedical and Health Informatics (JBHI), and Nature Scientific Reports. Before joining VinUniversity, Dr. Hieu worked at Vingroup Big Data Institute (VinBigData) as a Research Scientist and Head of the Fundamental Research Team. With this position, he led several research projects on Medical AI, including collecting various types of medical data, managing and annotating data, and developing new AI solutions for medical analysis.

HA Q. NGUYEN Dr. Ha Q. Nguyen was born in Hai Phong, Vietnam, in 1983. He received the B.S. degree in mathematics from the Hanoi National University of Education, Hanoi, Vietnam, the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2005, 2009, and 2014, respectively. During 2009–2011, he was a Lecturer in electrical engineering with the International University, Vietnam National University, Ho Chi Minh City, Vietnam, during 2014–2017, he was a Postdoctoral Research Associate with the Biomedical Imaging Group, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, and during 2017–2018, he was a Signal Processing Engineer with the Viettel Research and Development Institute, Hanoi, Vietnam. He is currently the Head of Medical Imaging Department at the Vingroup Big Data Institute, Hanoi, Vietnam. His research interests include medical image analysis, machine learning, computational imaging, and data compression. Dr. Nguyen was a Fellow of Vietnam Education Foundation, cohort 2007. He was the recipient of the Best Student Paper Award (second prize) of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2014 for his paper (with P.A. Chou and Y. Chen) on the compression of human body sequences using graph wavelet filter banks.