A Comparative Study of Noisy Label Detection Techniques in a Thai Hospital's Chest X-ray Database

1st Sorasit Tatitaisakul
Institute of Field Robotics
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
sorasit.phun@mail.kmutt.ac.th

2nd Suesarn Wilainuch
Perceptra Co., Ltd.
Bangkok, Thailand
suesarn@perceptra.tech

3rd Isarun Chamveha
Perceptra Co., Ltd.
Bangkok, Thailand
isarun@perceptra.tech

4th Warasinee Chaisangmongkon
Institute of Field Robotics
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
warasinee.cha@kmutt.ac.th

***Abstract*—This paper addresses the problem of noisy labels in chest X-ray datasets, which significantly impact the training of deep neural network models. Noisy labels often occur due to errors in reports from experts or the use of algorithms to extract labels from medical reports written in natural language. To tackle this issue, we compared the effectiveness of O2U-net, a state-of-the-art noisy label detection method, and NVUM, a noise-resistant model training technique in identifying noisy samples. We contrasted these methods with a heuristic approach which uses a simple classification model to flag samples with large differences between predicted and actual labels as noisy. Our findings indicated that NVUM outperformed the other methods in identifying noisy labels, providing a promising solution to the challenge of noisy labels in medical image analysis.**

***Index Terms*—Noisy label detection, Chest X-ray image, noise robust training, multi-label classification**

I. Introduction

Noisy labels, referring to samples in a dataset with incorrect labels, pose a significant challenge when training deep neural network models for medical image analysis [9]. The acquisition of high-quality manually labeled images from experts requires substantial investments of time and funds. While natural language processing (NLP) tools are commonly used to extract information from radiologists' reports in a cost-effective manner, this approach tends to introduce significant noise into the dataset [6], [13]. Furthermore, the task of identifying lesions in medical images is inherently challenging and even reports from experts can contain errors [9], further complicating the model training process.

To address the problem of noisy labels, two main approaches are commonly used: training a noise-robust model and employing a noisy label detection method. The first approach focuses on preventing overfitting on samples with noisy labels during the model training process [10], while the latter aims to identify samples with noisy labels, enabling re-labeling or removal from the dataset [5].

In this study, our primary focus was on finding the most effective way to identify noisy label samples, which can then be re-labeled by experts, resulting in a cleaner dataset for deep learning model training. The key contribution of this paper lies in the comparative analysis of noisy label detection methods within the context of the Thai hospital's chest X-ray database. To this end, we utilized O2U-net, a state-of-the-art noisy label detection method, which uses a cyclical learning rate schedule and tracks each sample's loss throughout training to find noisy labels [5]. We also adapted a noise-resistant model training technique, NVUM, which stores the model's initial state during training where it fits with easy samples and uses the memory to calculate loss regularization to prevent the model from overfitting with noisy labels. We identified noisy samples using NVUM by noting samples where model predictions significantly differed from their labels or by tracking each sample's NVUM loss across all epochs. We compared these methods to a heuristic approach, where a simple image classification model flags samples with significant probability differences as noisy labels. Our results demonstrated that NVUM was more effective at identifying noisy labels compared to other methods.

II. Related work

Numerous strategies have been proposed to address the noisy label issue. Han et al. [3] incorporated two deep neural networks (DNNs) in the training procedure, using disagreements between them to identify and reject noisy labels. Similarly, Jiang et al. [7] employed two DNNs, bifurcated into MentorNet and StudentNet. The MentorNet dynamically adjusts the sample weight for the StudentNet during training, whilst the StudentNet serves as the primary classification

network. Li et al. [8] utilizes semi-supervised learning and treating noisy labels as unlabeled samples.

Among these methods, Huang et al.'s [5] O2U-net demonstrates outstanding performance in addressing noisy labels. The approach relies on a simple concept that during the early stage of training, or while the model is underfitting, it predominantly aligns with clean samples, and at the later stage of the training, the model tends to overfit the noisy label samples. From this concept, they propose a cyclical learning rate schedule that begins with a high learning rate, linearly declining to a low rate, and abruptly reverts to a high rate when it hits the minimum desired learning rate. This schedule cyclically shifts the model state from overfitting to underfitting when learning rate suddenly changes from minimum to maximum. During training, the loss of each sample is collected and normalized in each epoch using that epoch's average. The cumulative loss, or the sum of sample losses from all epochs, is then calculated. Samples with high cumulative loss are flagged as noisy labels.

The Non-Volatile Unbiased Memory (NVUM) approach proposed by Liu et al. [10] relies on the same concept of O2U-net that the model initially fits with clean samples during the early training stages. To take advantage of this concept, the NVUM method introduces a memory matrix, denoted as t, which retains the sample logits from the earlier epochs. This memory is then used in loss regularization computation to prevent the model from overfitting with noisy labels. Furthermore, NVUM addresses the imbalanced data issue, which is prevalent in medical image datasets, by combining a class distribution, denoted as π, with the stored logits to act as class weight factors during training.

III. Methodology

A. Dataset

We employed two benchmark datasets in this study. The first dataset, referred to as the COVID-19 dataset, was curated by a team of researchers from Qatar University [2], [11], which is accessible through the Kaggle database. This dataset comprises frontal-view chest X-ray images obtained from various online databases, specifically focusing on COVID-19 cases. The COVID-19 dataset consists of 10,192 normal images and 10,973 abnormal images. The abnormal images are annotated based on 3,616 COVID-19 positive cases, 6,012 non-COVID-19 lung infections, and 1,345 cases of viral pneumonia. The original authors achieved a classification accuracy of 0.99 using the DenseNet201 model for the task of normal-abnormal classification [2]. In our implementation, we employed the DenseNet121 model and obtained an accuracy score of 0.948. Given its high level of classification feasibility, we selected this dataset as a preliminary testing dataset to establish a baseline comparison among all the noise detection methods we studied.

The second dataset used in this study was the Siriraj30k dataset, which comprises 29,849 high-quality chest X-ray images and corresponding radiologist reports from Siriraj Hospital in Bangkok, Thailand. We randomly selected images of individuals aged 15 years and older from the database and excluded any low-quality images. To annotate the images in this dataset, a two-step process was followed. First, NLP tools were used to label the images based on the radiologist reports. Then, trained annotators reviewed the labeler's results, along with the radiologist reports and images, to assign 7 classes to each image. These classes included Cardiomegaly, Edema, Pleural Effusion, Atelectasis, Mass, Nodule, and Lung Opacity Group (Infiltration, Consolidation, and Lung Opacity).

The Siriraj30k dataset contains realistic noisy labels originating from multiple sources, including instances where annotators missed keywords mentioned in radiologist reports (report annotation noise), as well as cases where the radiologist reports themselves overlooked lesions present in the chest X-ray images (report noise). Additionally, the dataset includes challenging cases with small lesions that even radiologists find difficult to identify (hard case noise) which may confuse classification models.

The two datasets were split into a 90% training set and a 10% validation set. The training set was used in noisy label detection experiments, while the validation set was used to determine the appropriate stopping epoch for training models. The distribution of annotations in both datasets can be seen in Table I and Table II.

TABLE I
COVID-19 Dataset Distribution

	Training set	Validation set
Abnormal	9,837	1,136
Normal	9,211	981
Total	19,048	2,117

TABLE II
Siriraj30k Dataset Distribution

	Training Set	Validation Set
Cardiomegaly	4,815	478
Edema	249	14
Pleural Effusion	847	88
Atelectasis	322	32
Mass	494	54
Nodule	1,508	158
Lung Opacity Group	4,472	473
Normal	17,340	1,873
Total	30,047	3,170

B. Noisy label detection model development

This study examines four different strategies for finding samples with incorrect or 'noisy' labels. The first method, named **Probability Difference**, uses a deep learning model trained to detect chest X-ray abnormalities to calculate the absolute difference between the predicted label probabilities and the actual labels. Samples with high differences are flagged as noisy labels. This method serves as a heuristic baseline. The second method, termed **O2U-net Cumulative Loss**, uses the O2U-net training method and calculates a

'noise score' for each sample based on the cumulative loss over training epochs. The final two methods use the NVUM noise-robust model to calculate noise scores. One variant, the **NVUM Cumulative Loss**, mirrors the O2U-net method by collecting the loss of each sample throughout training to be used as a noise score. The other, the **NVUM Probability Difference**, uses a noise-robust model already trained with a noisy dataset to calculate a probability difference score for each sample, which is then used as a noise score.

The O2U-net approach, with its simplicity and effectiveness, is ideal for our study on noisy labels in chest X-ray datasets. Our implementation largely follows the original O2U-net approach as described by Huang et al. [5]. However, to better suit multi-label classification tasks, such as in chest X-ray datasets, we have made a few modifications. Specifically, we have adjusted the loss function to binary cross-entropy loss and the activation function at the final model layer to a sigmoid function. The end result of the O2U-net training loop is a collection of cumulative sample loss. These losses are then sorted in descending order, and the top k samples with the highest loss are identified as noisy label samples. Here, k represents the number of noisy label samples anticipated to be captured.

For NVUM, our implementation closely aligns with the original one presented in [10] using suggested hyperparameter values proposed in the paper. We incorporated the parameter β with a value of 0.9, where $\beta \in [0, 1]$ serves as a parameter that controls the influence of the memory matrix (t). This implies that we assigned a weight of 0.9 to the previous logits and 0.1 to the current logits when computing the loss regularization term.

In model training, we resized the images to 224 x 224 and augmented them using horizontal flipping, linear and gamma contrast, brightness enhancement, and affine transformation. For the DNN model, we used the ImageNet [12] pre-trained DenseNet121 [4] and Adam optimizer for all approaches. We used binary cross entropy loss (BCE loss) as loss function in model training and noise score for ranking noisy labels. Each approach varies in technique and the training process, which means batch sizes, learning rates, and epochs are not uniform across all methods. Their specific settings are displayed in Table III.

TABLE III
IMPLEMENTATION DETAIL IN EACH APPROACH.

Parameters	Baseline (Probability Difference)	O2U-net (Cumulative Loss)	NVUM (Probability Difference)	NVUM (Cumulative Loss)
Batch size	16	64	64	64
Learning rate	1e-04	[a]Cyclical linear	1e-04	1e-04
Epoch	Early stop	40	Early stop	40

[a]Maximum learning rate=1e-3, minimum learning rate = 1e-5, cycle = 20 epoch

We used an early stop in Baseline (Probability Difference) and NVUM (Probability Difference) approaches as both use a fine-tuned model to determine the predicted probability. We trained the model until the validation accuracy no longer improved, then chose the model that demonstrated the highest accuracy on the validation set.

C. Evaluation metric

As outlined in Section III-B, each noisy label detection approach yields a sorted list of all samples, ranked by noise score. To identify potential noisy labels, we designated a parameter k and selected the top k samples from this sorted list. We used *Precision@k*, which is the number of correctly identified noisy labels (true positives) divided by k, as the main evaluation metric.

For COVID-19 dataset, we injected the noisy label into this relatively clean dataset by randomly flipping the labels. In this study, we used 5%, 10%, 20%, and 40% noise proportion for evaluation, with k equating to the respective noise proportion.

For Siriraj30k dataset, a multi-label classification dataset where multiple classes in the same image can be flagged as noisy, additional steps are required to select noisy labels. First, we assigned each class in a sample its ranking noise score, selecting the class with the highest noise score among seven to represent that sample's noise score. Then we sorted all samples by this score and rejected the top k samples to the noisy label group. Among the noisy label group, the minimum sample's noise score will serve as a threshold value. We use this threshold value to recheck all other classes in the noisy label group to determine if that class is a noisy label or not. As Siriraj30K dataset is a real-world noisy dataset where we do not know how many samples contain noisy labels, we set k to 1,000 for fair comparison across methods.

TABLE IV
PRECISION OF NOISE DETECTION MODELS ON COVID-19 DATASET

Noise proportion	Baseline (Probability Difference)	O2U-net (Cumulative Loss)	NVUM (Probability Difference)	NVUM (Cumulative Loss)
952 (5%) samples	0.689	0.816	0.789	**0.843**
1,904 (10%) samples	0.790	0.812	0.724	**0.821**
3,809 (20%) samples	0.794	0.814	0.808	**0.838**
7,619 (40%) samples	0.663	0.701	0.771	**0.799**

IV. RESULTS

A. Results from COVID-19 Dataset

As described in III, the COVID-19 dataset is used to evaluate how different methodologies perform under conditions where noisy and clean instances are clearly discernible. The original dataset was divided into a training set of 19,048 examples and a validation set of 2,117 examples. Synthetic noise was introduced into the training set via random label inversion at varying fractions of 5%, 10%, 20%, and 40% of

the total training set. For instance, in the 10% scenario, this implies the inclusion of 1,904 noisy label instances and 17,144 original instances within the training set.

As shown in Table IV, the results obtained across all noise proportion settings consistently demonstrated high precision in all our noisy label detection approaches. This indicates that our developed approaches are correctly implemented and they can identify approximately 80% of the noisy labels in a simple task. Notably, the NVUM (Cumulative Loss) approach, which combines elements from the O2U-net and NVUM methods, emerged as the most effective among all approaches.

B. Results from Siriraj30k Dataset

Unlike the COVID-19 dataset, the Siriraj30K dataset posed a more realistic challenge in identifying noisy labels. To analyze the results, we picked 1,000 samples with the highest noise scores in each approach and investigated them in detail. We reviewed chest X-ray images, radiologist reports, labels from annotators, and heatmaps from in-house chest X-ray image classifier [1] in detail to categorize the selected potentially noisy samples into 4 groups: (1) Annotation Noise Group, where our review shows that annotators have made a mistake in their interpretation of radiologist reports, (2) Report Noise Group, where our examination reveals missing lesions in the diagnostic reports, (3) Hard Case Noise Group, which includes instances where radiologists have mentioned in the reports that lesions are small or unclear, suggesting their uncertainty in the presence of the lesions, (4) "Non-noisy Group" are samples that are already correctly labeled.

We anticipate that the Probability Difference method will effectively detect Hard Case Noise, where lesions are ambiguous or unclear as indicated by radiologists in their reports. However, Hard Case Noise samples are easy to detect using a keyword detection model applied to the radiologist reports, and therefore they are not the primary focus of our study. We are more concerned about the Annotation Noise and Report Noise which are more subtle and difficult to detect.

As depicted in the Table V, when considering only the Annotation Noise Group and Report Noise Group, it is evident that the approaches incorporating NVUM surpass the baseline and O2U-net in performance. Specifically, the total precision (excluding hard cases) is approximately 16% superior to the other two approaches. Surprisingly, when considering total precision, O2U-net was the least efficient in detecting noisy samples, even less effective than the heuristic approach.

In the COVID-19 dataset, the limitations of O2U-net might not be readily apparent because the noise in the data was artificially introduced. Artificial noise tends to be more random and less correlated compared to real-world noise. If the model fits the systematic errors in the dataset, it could be hard to distinguish between correct and incorrect signals, especially in imbalance environments where the majority class dominates the loss calculation.

NVUM's strength lies in using an adaptive loss regularization term while training. On noisy samples where the model is less sure, the regularization term increases, actively preventing the model from overfitting noisy labels. NVUM can maintain a high loss for noisy samples, aiding in noisy-label detection. Additionally, it uses class distribution which could improve its detection of noisy labels in a data imbalance situation, which is common in medical image analysis.

In a comparison between NVUM (Probability Difference) and NVUM (Cumulative Loss), the findings are consistent with the COVID-19 dataset. It is evident that the use of cumulative loss from all epochs is more effective than predicted probability. Predicted probability solely utilizes the final state of the model post-training, whereas cumulative loss captures all states of the model throughout the training process, thereby enhancing the detection of noisy label samples.

To better understand the advantages of our approach, we split noisy samples into negative (normal) and positive (abnormal) based on its label. As shown in the Table VI, our approach is particularly helpful in cases where radiologists may have under-reported abnormal features in the chest X-ray images. Upon closer inspection, we found that the majority of cases in the Report Noise Group belong to the Cardiomegaly class, which necessitates a time-consuming cardiothoracic ratio (CTR) measurement for disease identification. Consequently, the primary radiologist protocol typically prioritizes on the other diseases, potentially resulting in a higher likelihood of overlooking Cardiomegaly compared to other conditions.

V. Conclusion

Our study demonstrated NVUM's superiority over O2U-net and the baseline approach in handling real-world noise in chest X-ray datasets. In the real-world dataset, incorporating terms such as loss regularization and class distribution within NVUM proves more beneficial than modifying the learning rate schedule, as observed in O2U-net. Furthermore, our analysis of the Thai hospital's database revealed that the predominant noisy cases are associated with the Report Noise Group. Incorporating an AI diagnostic assistant into diagnostic and annotation workflows could reduce errors in future batches of data. Lastly, It's important to note that the aspiration of 100% clean data is practically unattainable in the realm of medical imaging, as image interpretation is often subjective and can vary between observers. Consequently, the true effectiveness of noisy label detectors is hard to measure, as there's no perfect ground truth. Nevertheless, our findings underscore the potential of NVUM in handling noisy labels in medical imagery, providing valuable guidance to practitioners in the field.

Acknowledgment

We would like to express our gratitude to the Department of Radiology, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand, for their support in providing the chest X-ray images and reports dataset. Additionally, we thank Perceptra Co., Ltd. for contribution of the Inspectra Labeler and Inspectra model. Approval of all ethical and experimental procedures for this work was granted by KMUTT-IRB (Cert.

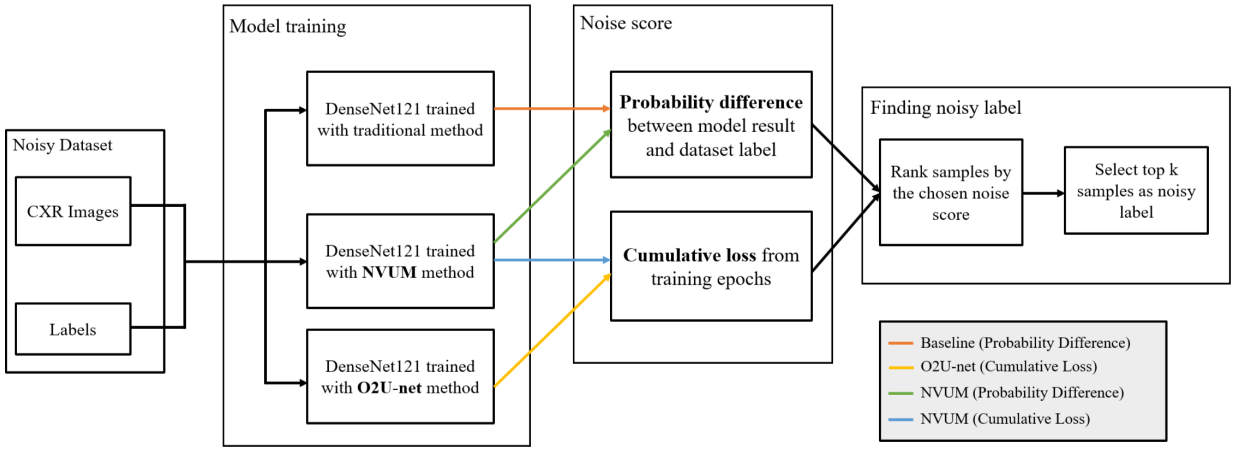

Fig. 1. Overall procedure of all noisy label detection approaches

TABLE V
BREAKDOWN OF NOISE TYPES IN THE TOP 1,000 DETECTED NOISY CASES FOR EACH MODEL

Scenarios	Baseline (Probability Difference)	O2U-net (Cumulative Loss)	NVUM (Probability Difference)	NVUM (Cumulative Loss)
Annotation Noise Group	28 (0.028)	30 (0.030)	33 (0.033)	35 (0.035)
Report Noise Group	77 (0.077)	71 (0.071)	151 (0.151)	154 (0.154)
Hard Case Noise Group	579 (0.579)	567 (0.567)	512 (0.512)	538 (0.538)
Non-noisy Noise Group	316 (0.316)	332 (0.332)	304 (0.304)	273 (0.273)
Total precision	684 / 1000 (0.684)	668 / 1000 (0.668)	696 / 1000 (0.696)	**727 / 1000 (0.727)**
Total precision (w/o hard cases)	105 / 421 (0.249)	101 / 433 (0.233)	184 / 488 (0.377)	**189 / 462 (0.409)**

TABLE VI
BREAKDOWN OF POSITIVE AND NEGATIVE CASES AMONG THE TOP 1,000 NOISY CASES, CATEGORIZED BY NOISE TYPES

Scenarios	Baseline (Probability Difference)		O2U-net (Cumulative Loss)		NVUM (Probability Difference)		NVUM (Cumulative Loss)	
	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos
Annotation Noise Group	4	24	3	27	6	27	9	26
Report Noise Group	55	22	45	26	**135**	16	**129**	25
Hard Case Noise Group	0	589	0	579	0	524	0	550
Non-noisy Noise Group	15	316	5	339	5	314	6	280
Total precision	0.797	0.668	0.906	0.651	0.966	0.644	0.958	0.682

No. KMUTT-IRB-COE-2021-043) performed in line with the international guidelines of human research protection.

REFERENCES

[1] Isarun Chamveha, Trongtum Tongdee, Pairash Saiviroonporn, and Warasinee Chaisangmongkon. Local adaptation improves accuracy of deep learning model for automated x-ray thoracic disease detection: A thai study. *arXiv preprint arXiv:2004.10975*, 2020.

[2] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020.

[3] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q

Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[5] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3326–3334, 2019.

[6] 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.

[7] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018.

[8] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[9] 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.

[10] Fengbei Liu, Yuanhong Chen, Yu Tian, Yuyuan Liu, Chong Wang, Vasileios Belagiannis, and Gustavo Carneiro. Nvum: Non-volatile unbiased memory for robust medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 544–553. Springer, 2022.

[11] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[13] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.