GLANCE: Combating Label Noise using Global and Local Noise Correction for Multi-Label Chest X-ray Classification

Xianze Ai[1,2], Zehui Liao[1,2], and Yong Xia[1,2](✉)

[1] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xian 710072, China
yxia@nwpu.edu.cn

Abstract. Chest X-ray imaging is essential for diagnosing thoracic diseases, with multi-label classification playing a critical role in identifying multiple conditions from a single image. Despite deep neural networks significantly advancing this field, noisy labels extracted from clinical reports pose a significant challenge, undermining the performance of deep models. Several research attempts have been made to address this issue but fail to consider the critical inter-class correlations prevalent in chest X-ray diagnostics. To this end, we propose a Global and Local Noise Correction framework. Our framework comprises a classification backbone and two primary components: a global noise correction module and a local noise correction module. The global noise correction module calculates the noise transition matrix based on the label co-occurrence frequencies and uses the estimated noise transition matrix to reduce the impact of the noisy labels. The local noise correction module treats the temporal ensembling of samples historical predictions as the instance-specific pseudo labels, which also serve as the supervision. The proposed framework addresses the shortcomings of existing techniques, *i.e.*, the unreliability of noise transition matrices in the presence of class imbalances and zero co-occurrence frequencies. Comprehensive experimental results demonstrate that our framework surpasses competing methods, showcasing its superior ability to combat label noise and improve multi-label chest X-ray classification accuracy.

Keywords: Chest X-ray classification · Label noise · Label co-occurrence

1 Introduction

Chest X-Ray (CXR) imaging is the most common screening technique, effectively assisting in the clinical diagnosis and treatment of various thoracic diseases [10]. Multi-label CXR classification involves simultaneously identifying multiple conditions or abnormalities in a single X-ray image [1]. The advent of deep neural networks (DNNs) has significantly advanced this field [16,22,2]. The success of

DNNs heavily relies on accurately labeled training data. However, clinical data often contain label noise because natural language processing (NLP) techniques are used to automatically extract labels from diagnostic reports, and these extracted labels lack verification by professional physicians [4]. Although this solution reduces annotation costs, the noisy labels negatively impact DNNs' performance and generalization ability [24]. Therefore, it is crucial to develop robust multi-label CXR classification methods that can handle label noise [26,15,11,18].

In recent years, an increasing number of researchers have focused on this issue. The MODL-KNNS method [26] adopts a model ensemble strategy and averages all models predictions as the pseudo label, which is then refined by the k-nearest neighbor smoothing strategy. The LSR method [15] employs label smoothing regularization to mitigate the impact of noisy labels. The NVUM method [11] maintains a non-volatile running average of model logits as training targets. The SNEL method [18] learns from noisy labels by performing model ensemble and designing noise-robust loss functions. Despite their robustness, these methods do not account for inter-class correlations, which are prevalent in thoracic diseases [3,15]. For instance, emphysema and pneumothorax often co-occur, as do effusion and atelectasis. This co-occurrence suggests potential label correlations in multi-label learning, which have been explored through methods such as graph convolution [5,23] and label co-occurrence analysis [9,13]. Specifically, the method proposed by Chen et al. [9] utilizes label co-occurrence to estimate the noise transition matrix, leveraging label correlation to address label noise. However, this method encounters challenges when applied to chest X-ray images, where the noise transition matrix may not exhibit diagonal dominance due to the imbalanced class distribution. This can render the estimated matrix unreliable for model training. Furthermore, in some cases, the co-occurrence frequency between classes may be zero, leading to an invalid noise transition matrix.

To address these issues, we propose a **G**lobal and **L**oc**A**l **N**oise **C**orr**E**ction (GLANCE) framework for chest X-ray classification tasks with noisy labels. Our GLANCE framework comprises a classification network that includes an encoder followed by a fully connected (FC) layer, a global noise correction (GNC) module, and a local noise correction (LNC) module. Given an input image, the classification network outputs the probabilistic prediction, whose supervision is obtained from the GNC and LNC modules. The GNC module computes the global noise transition matrix based on label co-occurrence frequencies to rectify the noisy observed labels. Here the temporal ensembling is adopted to boost the calculation of the noise transition matrix. Concurrently, the LNC module generates instance-specific pseudo labels through temporal ensembling of samples historical predictions, leveraging the DNNs tendency to fit clean data first. Both the globally corrected labels and the instance-specific pseudo labels are utilized to supervise the predictions, thereby optimizing the model.

The main contributions of this work are as follows: (1) We introduce the GLANCE framework, designed to mitigate label noise in multi-label CXR classification tasks. Our framework demonstrates the efficacy of leveraging interclass correlations to address label noise within this task. (2) We leverage label co-occurrence frequencies to calculate the noise transition matrix. By incorporating temporal ensembling and local noise correction, we mitigate issues caused by imbalanced class distributions and zero co-occurrence frequencies, resulting in more robust matrix estimation. (3) Experimental results show that our GLANCE framework outperforms six competing methods in combating label noise in chest X-ray classification.

2 Method

2.1 Problem Definition and Overview

Given a noisy multi-label classification dataset $D = \{(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i \in \mathbb{R}^{H\times W\times R}$ represents the i-th CXR image size $H \times W$ with R colour, $\bar{\boldsymbol{y}}_i \in \{0,1\}^C$ is the label of $\boldsymbol{x}_i$, N is the total number of samples, C denotes the number of classes. The presence of the j-th disease in the CXR image is indicated by a '1' at the j-th position in $\bar{\boldsymbol{y}}_i$, a '0' signifies absence. However, it is important to note that some of the indications of disease presence or absence in the labels of dataset D may be erroneous due to noise. Our goal is to develop a robust multi-label classification model for CXR images by training on this noisy dataset D.

As depicted in Fig. 1, our GLANCE framework consists of a classification backbone, a GNC module, and an LNC module. The classification backbone undergoes a warm-up phase for E_{warm} epochs and then is trained intercorporated with the GNC and LNC modules. The GNC module computes the global noise transition matrix to rectify the noisy observed labels. Concurrently, the LNC module generates instance-specific pseudo labels through temporal ensembling of samples' historical predictions. Both the globally corrected labels and the instance-specific pseudo labels are utilized to supervise the predictions, thereby optimizing the model.

2.2 Classification Backbone

DenseNet [7] serves as the foundational architecture $\mathcal{F}(\Theta)$, incorporating an encoder and an FC layer, where Θ denote its parameters. The encoder contains several dense blocks followed by an average pooling layer. Given an input image $\boldsymbol{x}_i$, the encoder derives the image feature, which is then fed into the FC layer and a sigmoid function S, yielding the probabilistic output $\boldsymbol{p}_i = S(\mathcal{F}(\boldsymbol{x}_i; \Theta))$. During the warm-up phase, the backbone is optimized using the binary cross entropy (BCE) loss calculated between the observed label $\bar{\boldsymbol{y}}_i$ and its prediction $\boldsymbol{p}_i$. After the warm-up phase, the backbone is optimized by integrating the GNC and LNC modules. We now delve into the details of the GNC and LNC modules.

2.3 Global Noise Correction Module

In each epoch, we first identify samples with low training loss values and form them as a reliable set D_r [27]. Specifically, we calculate the standard multi-label

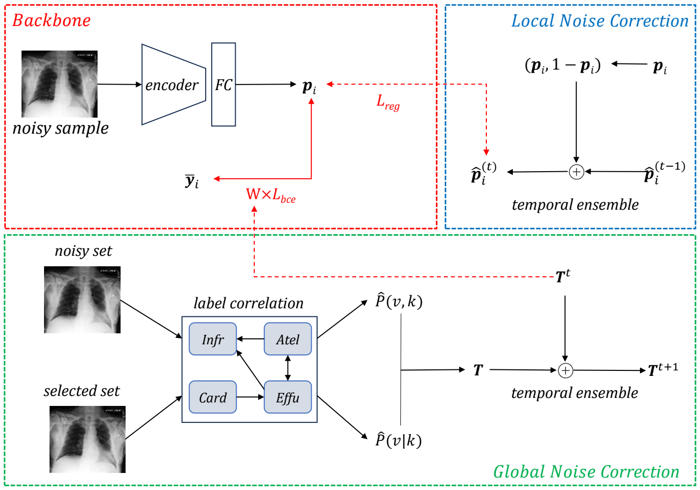

Fig. 1. Overview of our GLANCE framework. $\mathbf{y}_i$ represents the label of the training set, $\boldsymbol{p}_i$ represents the model prediction, and $\hat{\boldsymbol{p}}_i^{(t)}$ represents the time-integrated pseudo-label of the sample at the t th epoch. $\mathbf{T}$ represents the estimated noise transition matrix, where $\mathbf{T}^t$ represents the weighted noise transition matrix obtained at the t th epoch. W represents the weight matrix for weighting the cross-entropy loss, which is calculated by $\mathbf{T}$ and $\boldsymbol{p}_i$.

classification loss for noisy training examples and then set a pre-defined threshold to select samples with lower training loss values. Subsequently, we consider the relationship between co-occurrence probability and noise transition probability, established through four equations (see Eq. 1). By solving this bilinear decomposition problem, the noise transition matrix is calculated. The noise transition matrix for this task is denoted as $\boldsymbol{T} \in [0,1]^{C\times 2\times 2}$. For class $j \in \{1,2,...,C\}$, four elements of its noise transition matrix $\boldsymbol{T}^j \in [0,1]^{2\times 2}$ can be calculated using the following system of equations:

$$
\begin{aligned}
P(\overline{Y}_j\!=\!0,\overline{Y}_i\!=\!0)&=P(Y_j\!=\!0)P(\overline{Y}_i\!=\!0|Y_j\!=\!0)\boldsymbol{T}^j_{00}+P(Y_j\!=\!1)P(\overline{Y}_i\!=\!0|Y_j\!=\!1)\boldsymbol{T}^j_{10},\\
P(\overline{Y}_j\!=\!0,\overline{Y}_i\!=\!1)&=P(Y_j\!=\!0)P(\overline{Y}_i\!=\!1|Y_j\!=\!0)\boldsymbol{T}^j_{00}+P(Y_j\!=\!1)P(\overline{Y}_i\!=\!1|Y_j\!=\!1)\boldsymbol{T}^j_{10},\\
P(\overline{Y}_j\!=\!1,\overline{Y}_i\!=\!0)&=P(Y_j\!=\!0)P(\overline{Y}_i\!=\!0|Y_j\!=\!0)\boldsymbol{T}^j_{01}+P(Y_j\!=\!1)P(\overline{Y}_i\!=\!0|Y_j\!=\!1)\boldsymbol{T}^j_{11},\\
P(\overline{Y}_j\!=\!1,\overline{Y}_i\!=\!1)&=P(Y_j\!=\!0)P(\overline{Y}_i\!=\!1|Y_j\!=\!0)\boldsymbol{T}^j_{01}+P(Y_j\!=\!1)P(\overline{Y}_i\!=\!1|Y_j\!=\!1)\boldsymbol{T}^j_{11}.
\end{aligned}
\tag{1}
$$

where $\boldsymbol{T}^j_{00}+\boldsymbol{T}^j_{01}=1, \boldsymbol{T}^j_{10}+\boldsymbol{T}^j_{11}=1$, $P(Y_i\!=\!1)$ represents the probability that the sample is accurately labeled as 1 in the dataset D. The co-occurrence frequency $\hat{P}\left(\overline{Y}_j=v,\overline{Y}_i=k\right), v,k \in \{0,1\}$ is calculated through frequency counting using the reliable set D_r (see Eq. 2) and treated as co-occurrence probability $P(\overline{Y}_j=$

$v, \overline{Y}_i = k)$ in Eq. 1.

$$\hat{P}\left(\overline{Y}_j = v, \overline{Y}_i = k\right) = \frac{1}{n} \sum_{D} \mathbb{1}\left[\overline{y}_j = v, \overline{y}_i = k\right] \tag{2}$$

The indicator $\mathbb{1}[A]$ represents whether the condition A is met. Similarly, the conditional probability $P(\overline{Y}_i = v \mid Y_j = k)$ in Eq. 1 is approximated as frequency by

$$\hat{P}\left(\overline{Y}_i = v \mid Y_j = k\right) = \frac{\sum_{D_r} \mathbb{1}\left[\overline{y}_i = v, \overline{y}_j = k\right]}{\sum_{D_r} \mathbb{1}\left[\overline{y}_j = k\right]} \tag{3}$$

As the co-occurrence probabilities and conditional probabilities are known, we can obtain $\boldsymbol{T}^j$, $P(Y_j = 0)$ and $P(Y_j = 1)$.

For class j, we can derive C–1 noise transition matrices, denoted as $\boldsymbol{T}_r^j$ where $r \in \{1, 2, ..., C-1\}$. These noise transition matrices are different and we select one by Eq. 4 for model training.

$$\boldsymbol{T}_j = \arg\min_{\boldsymbol{T}_{jr}} \sum_{i=1}^{C-1} \left\|\boldsymbol{T}_{jr} - \boldsymbol{T}_{ji}\right\|_1, \tag{4}$$

where $|| \bullet ||$ represents the L1-norm.

Inspired by early learning and memorization phenomena [12], temporal ensembling is used for boosting noise transition matrix estimation. Firstly, initialize the noise transition matrices of C classes: $\boldsymbol{T}_j^{(0)} = [[1, 0], [0, 1]]$. Each epoch we can solve the $\boldsymbol{T}_j^{(t)}$ and update them by Eq. 5.

$$\boldsymbol{T}_j^{(t+1)} = \beta \boldsymbol{T}_j^{(t)} + (1-\beta)\boldsymbol{T}_j, \tag{5}$$

In cases where the estimated transition matrices exhibit illegal, they are replaced by the initial matrix to reduce the impact of the wrong estimation. Finally, the noise transition matrix $\boldsymbol{T}^{(t)}$ is used to calculate the weight W of the loss function using the risk consistent algorithm. The global weighted BCE loss is shown as 6:

$$L_{global} = -W\left[\bar{\boldsymbol{y}} log(\boldsymbol{p}) + (1 - \bar{\boldsymbol{y}}) log(1 - \boldsymbol{p})\right]. \tag{6}$$

2.4 Local Noise Correction Module

As previously discussed, the estimation of the noise transition matrix through the analysis of label co-occurrence can sometimes be unreliable. It is noteworthy that certain classes may not have any actual associations with other classes. Therefore, a local noise correction module is introduced. We extend early learning regularization [12] to the multi-label classification for this local noise correction. First, we define the binary prediction for sample i as $\boldsymbol{p}_i^{'} = (1 - \boldsymbol{p}_i, \boldsymbol{p}_i)$. For each sample i, a history vector $\hat{\boldsymbol{p}}_i$ is saved, which is calculated using the historical prediction of the model. Early learning regularization aims to maximize the inner product of the model's current output and the history vector. The local regularization can be calculated by Eq. 7:

$$L_{local} = \log(1 - \langle \hat{\boldsymbol{p}}_i^{(t)'}, \boldsymbol{p}_i^{'} \rangle). \tag{7}$$

where $\hat{\boldsymbol{p}}_i^{(t)'}$ represents the historical record vector of sample i at the t th epoch. The historical record vector contains C classes, with each class featuring a binary predictive outcome. $\langle\rangle$ represents the inner product. For $\hat{\boldsymbol{p}}_i^{(t)'}$, we initialize $\hat{\boldsymbol{p}}_i^{(0)'} = 0$ and calculate the historical record value of each epoch according to Eq. 8:

$$\hat{\boldsymbol{p}}_i^{(t)'} = \gamma \hat{\boldsymbol{p}}_i^{(t-1)'} + (1 - \gamma) \boldsymbol{p}_i^{'}. \tag{8}$$

where γ is a parameter that controls the influence of early memory.

2.5 Loss

Finally, the total loss is:

$$L = L_{global} + \alpha L_{local}. \tag{9}$$

where α is a hyperparameter that controls the weight of the L_{local}.

3 Experiments and Results

3.1 Dataset and Experimental Setup

Dataset. The NIH Chest X-ray14 Dataset [19] comprises 112,120 X-ray images with disease labels from 30,805 unique patients. The dataset includes 15 classes: 14 diseases and one No findings class. Images can be classified as No findings or one or more disease classes: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, and Hernia. The authors used NLP to extract disease labels from the associated radiological reports. These labels are expected to be over 90% accurate and suitable for noisy label learning. As the official data split, the data is divided into a training set of 86,524 samples and a test set of 25,596 samples at the patient level. Additionally, 10% of the training data is randomly selected as a validation set. An additional test set, annotated by five radiologists, provides more reliable data for model evaluation [14].

Implementation Details. DenseNet [7], pre-trained on the ImageNet dataset [6], is adopted as the backbone network. The model is trained for 30 epochs. The initial learning rate lr is set to 0.0001 and decays to 0.00001 when the training epoch reaches 20. The Adam optimizer [8] is used, with a batch size of 32. The dropout ratio is set to 0.2. During training, all images are resized to 224×224. Data augmentation strategies include random affine transformations and random flipping. We set hyperparameters in GLANCE to $E_{warm} = 5, \alpha = 2.5, \beta = \gamma = 0.9$ for the NIH dataset and introduce a mean teacher network [17] to further enhance the models robustness. The experiments are conducted on a workstation with one NVIDIA GTX 1080Ti GPU using the PyTorch framework.

Table 1. Test accuracy (%) of our GLANCE, baseline, and five competing methods on the official test set of the NIH Chest X-ray14 dataset. The best and second-best results in each row are highlighted in **bold** and underline, respectively.

Class	Method						
	BCE	GCE	GLS	MLT	USDNL	NVUM	Ours
Atelectasis	0.7672	0.7761	0.7804	0.7816	0.7792	0.7838	**0.7849**
Cardiomegaly	0.8752	0.8914	0.8866	0.8813	0.8841	0.8923	**0.8929**
Effusion	0.8084	0.8282	0.8284	0.8245	0.8269	0.8338	**0.8353**
Infiltration	0.6934	0.6987	0.7050	0.7042	0.6962	**0.7062**	0.7046
Mass	0.8239	0.8264	0.8267	0.8305	0.8241	0.8317	**0.8325**
Nodule	0.7569	0.7682	0.7700	0.7718	0.7718	**0.7770**	0.7726
Pneumonia	0.7149	0.7288	0.7261	0.7224	0.7276	0.7309	**0.7321**
Pneumothorax	0.8612	0.8596	0.8547	0.8544	0.8539	0.8665	**0.8677**
Consolidation	0.7442	0.7543	0.7491	0.7478	0.7546	0.7539	**0.7632**
Edema	0.8352	0.8464	0.8446	0.8410	0.8412	**0.8495**	0.8477
Emphysema	0.8984	0.9044	0.9045	0.8973	0.8971	**0.9061**	**0.9061**
Fibrosis	0.7998	0.8145	**0.8240**	0.8160	0.8142	0.8216	0.8208
Pleural Thicken	0.7726	0.7840	0.7842	0.7878	0.7811	**0.7885**	0.7877
Hernia	0.8922	0.8793	0.9205	0.9185	0.9240	0.9148	**0.9270**
Average	0.8031	0.8114	0.8146	0.8128	0.8126	0.8183	**0.8197**
No Findings	0.7298	0.7293	0.7332	0.7332	0.7299	0.7395	**0.7436**

Table 2. Test accuracy (%) of our GLANCE, baseline, and five competing methods on the additional test set of the NIH Chest X-ray14 dataset. The best and second-best results in each row are highlighted in **bold** and underline, respectively.

Class	Method						
	BCE	GCE	GLS	MLT	USDNL	NVUM	Ours
Average	0.8697	0.8732	0.8818	0.8739	0.8791	0.8878	**0.8907**
No Findings	0.9364	0.9331	0.9380	0.9393	0.9409	0.9414	**0.9457**

3.2 Comparative Experiments

The proposed GLANCE framework is compared with a baseline method and five recent methods: (1) Binary Cross-Entropy (BCE) loss, which is the baseline method, (2) Generalized Cross-Entropy (GCE) loss [25], which combines the standard cross-entropy loss and the mean absolute error loss, (3) Generalized Label Smoothing (GLS) [20], which uses the positively or negatively weighted average of both the hard observed labels and uniformly distributed soft labels as the target labels, (4) Multi-label Transition Matrix (MLT) [9], which estimates the noise transition matrix through label co-occurrence. (5) Uncertainty-Based Single Dropout (USDNL) [21], which performs sample selection based on the uncertainty estimated using a single dropout after early training, (6) Non-volatile unbiased memory (NVUM) [11], which stores a non-volatile running average of model logits as the training targets. It also uses Mixup and the mean teacher network to improve the models robustness. All competing methods were re-

Table 3. Test accuracy (%) of our GLANCE and its four variants on the official test set of the NIH Chest X-ray14 dataset. The best result in each column is highlighted in **bold**.

BCE	GNC	LNC	MT	Average	No Findings
✓				0.8031	0.7298
✓	✓			0.8158	0.7375
✓		✓		0.8165	0.7413
✓	✓	✓		0.8185	0.7427
✓	✓	✓	✓	**0.8197**	**0.7436**

Table 4. Test accuracy (%) of our GLANCE and its four variants on the additional test set of the NIH Chest X-ray14 dataset. The best result in each column is highlighted in **bold**.

BCE	GNC	LNC	MT	Average	No Findings
✓				0.8697	0.9364
✓	✓			0.8857	0.9411
✓		✓		0.8838	0.9439
✓	✓	✓		0.8879	0.9436
✓	✓	✓	✓	**0.8907**	**0.9457**

implemented using their released codes. The learning rate, batch size, weight decay, optimizer, epochs number, and backbone are kept the same as those in our GLANCE for a fair comparison. Table 1 and Table 2 show the comparison results on the official and additional test sets, respectively. Experimental results indicate that our GLANCE achieves the best result on both the official test set and the additional test set.

3.3 Ablation Analysis

Both the GNC and LNC modules play an essential role in the proposed GLANCE framework. We conducted ablation studies on the NIH Chest X-ray14 dataset to investigate the effectiveness of these two modules individually. The ablation results on the official and additional test sets are shown in Table 3 and Table 4, respectively. The results indicate that incorporating either the GNC module or the LNC module with the baseline model (*i.e.*, BCE) improves the performance on both diseases and No findings identification. Additionally, using the mean teacher (MT) mechanism further enhances the models performance.

4 Conclusion

In this paper, we propose to model the label correlation for facilitating the label-noise-robust learning for chest X-ray classification. To achieve this, we propose the GLANCE framework to model the global label correlation and the local sample historical prediction. We perform global correction by estimating the noise transition matrix through label co-occurrence, and perform sample-local correction based on the historical prediction of each sample. We conducted experiments on the NIH Chest X-ray14 dataset and the results show that our GLANCE framework performs better than other competing methods significantly. Ablation studies demonstrate the contribution of the global noise correction module and the local noise correction module.

Acknowledgement. This work was supported in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), in part by the National Natural Science Foundation of China under Grants 62171377, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2022056.

References

1. Agrawal, T., Choudhary, P.: Segmentation and classification on chest radiography: a systematic survey. The Visual Computer **39**(3), 875–913 (2023)
2. Bhosale, Y.H., Patnaik, K.S.: Iot deployable lightweight deep learning application for covid-19 detection with lung diseases using raspberrypi. In: 2022 International conference on IoT and blockchain technology (ICIBT). pp. 1–6. IEEE (2022)
3. Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P.: Deep hierarchical multi-label classification of chest x-ray images. In: International conference on medical imaging with deep learning. pp. 109–120. PMLR (2019)
4. Chen, P., Ye, J., Chen, G., Zhao, J., Heng, P.A.: Robustness of accuracy metric and its inspirations in learning with noisy labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11451–11461 (2021)
5. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5177–5186 (2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Li, S., Xia, X., Zhang, H., Zhan, Y., Ge, S., Liu, T.: Estimating noise transition matrix with label correlations for noisy multi-label learning. Advances in Neural Information Processing Systems **35**, 24184–24198 (2022)
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
11. Liu, F., Chen, Y., Tian, Y., Liu, Y., Wang, C., Belagiannis, V., Carneiro, G.: Nvum: Non-volatile unbiased memory for robust medical image classification. In:

International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 544–553. Springer (2022)

12. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems **33**, 20331–20342 (2020)
13. Liu, Z., Cheng, Y., Tamura, S.: Multi-label local to global learning: a novel learning paradigm for chest x-ray abnormality classification. IEEE Journal of Biomedical and Health Informatics (2023)
14. Nabulsi, Z., Sellergren, A., Jamshy, S., Lau, C., Santos, E., Kiraly, A.P., Ye, W., Yang, J., Pilgrim, R., Kazemzadeh, S., et al.: Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and covid-19. Scientific reports **11**(1), 15523 (2021)
15. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
16. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
17. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
18. Wang, H., He, J., Cui, H., Yuan, B., Xia, Y.: Robust stochastic neural ensemble learning with noisy labels for thoracic disease classification. IEEE Transactions on Medical Imaging (2024)
19. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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. pp. 2097–2106 (2017)
20. Wei, J., Liu, H., Liu, T., Niu, G., Sugiyama, M., Liu, Y.: To smooth or not? when label smoothing meets noisy labels. arXiv preprint arXiv:2106.04149 (2021)
21. Xu, Y., Niu, X., Yang, J., Drew, S., Zhou, J., Chen, R.: Usdnl: uncertainty-based single dropout in noisy label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10648–10656 (2023)
22. Yan, C., Yao, J., Li, R., Xu, Z., Huang, J.: Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. pp. 103–110 (2018)
23. Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 649–665. Springer (2020)
24. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
25. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
26. Zhou, Y., Huang, L., Zhou, T., Shao, L.: Many-to-one distribution learning and k-nearest neighbor smoothing for thoracic disease identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 768–776 (2021)

27. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE transactions on medical imaging **41**(4), 881–894 (2021)