Improving Medical Image Classification in Noisy Labels Using only Self-supervised Pretraining

Bidur Khanal[1(✉)], Binod Bhattarai[4], Bishesh Khanal[3], and Cristian A. Linte[1,2]

[1] Center for Imaging Science, RIT, Rochester, NY, USA
bk9618@rit.edu
[2] Biomedical Engineering, RIT, Rochester, NY, USA
[3] NepAl Applied Mathematics and Informatics Institute for Research (NAAMII), Patan, Nepal
[4] University of Aberdeen, Aberdeen, UK

Abstract. Noisy labels hurt deep learning-based supervised image classification performance as the models may overfit the noise and learn corrupted feature extractors. For natural image classification training with noisy labeled data, model initialization with contrastive self-supervised pretrained weights has shown to reduce feature corruption and improve classification performance. However, no works have explored: i) how other self-supervised approaches, such as pretext task-based pretraining, impact the learning with noisy label, and ii) any self-supervised pretraining methods alone for medical images in noisy label settings. Medical images often feature smaller datasets and subtle inter-class variations, requiring human expertise to ensure correct classification. Thus, it is not clear if the methods improving learning with noisy labels in natural image datasets such as CIFAR would also help with medical images. In this work, we explore contrastive and pretext task-based self-supervised pretraining to initialize the weights of a deep learning classification model for two medical datasets with self-induced noisy labels—*NCT-CRC-HE-100K* tissue histological images and *COVID-QU-Ex* chest X-ray images. Our results show that models initialized with pretrained weights obtained from self-supervised learning can effectively learn better features and improve robustness against noisy labels.

Keywords: medical image classification · label noise · learning with noisy labels · self-supervised pretraining · warm-up obstacle · feature extraction

1 Introduction

Medical image classification using supervised learning relies on large amounts of representative data with accurately annotated labels to achieve good gener-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-44992-5_8.

B. Bhattarai et al. (Eds.): DEMI 2023, LNCS 14314, pp. 78–90, 2023.
https://doi.org/10.1007/978-3-031-44992-5_8

alization. However, recent practices of crowd-sourcing for data labeling or automatically generating labels from patients' medical reports using algorithms, and the high variability among expert annotators introduce higher levels of label noise in medical datasets. Moreover, supervised deep learning is highly susceptible to label noise as the models can easily overfit the noisy labels, leading to corrupt representation learning and compromising generalizability [18,19,24,39]. Correcting label noise in large medical image datasets is expensive and requires extensive human resources and time-consuming protocols. Several methods for learning with noisy labels (LNL) have been introduced in natural image datasets to minimize the influence of label noise on the training [2,8,12,25,32,34]. Similar methods, with adjustments, have also been applied to medical image classification [15,26,36,44].

Many LNL methods rely on a warm-up phase, a small number of initial epochs during which the model is trained directly using all the noisy training data [8,15,25,27]. While the warm-up phase is important to kickstart the model and learn basic features important for proper separation of noisy labels from clean labels at a later phase [41,42], the high noise rate makes it challenging to avoid memorizing wrong labels and learning poor feature extractors. Zheltonozhskii et al. [42] referred to this issue as the "warm-up obstacle". One may use supervised pretraining to learn good feature extractors and train with noisy labels to mitigate the warm-up obstacle. However, this approach presents challenges in medical datasets due to the limited availability of large labeled datasets that closely align with the given new dataset. Alternatively, if existing medical datasets already contain valuable metadata such as gender and age information, one may pretrain to predict such auxiliary information before proceeding with training on the main task involving noisy labels. Such an approach could minimize feature corruption, as the auxiliary tasks are relatively straightforward and less likely to contain label noise. However, if the datasets lack such metadata, another approach is to use self-supervised learning techniques for pretraining to learn feature extractors, without relying on any labels.

Some studies have demonstrated the benefits of contrastive learning-based self-supervised pretraining to improve robustness against noisy labels in natural image datasets [41,42]. However, no extensive study has been conducted to investigate which self-supervised pretraining is suitable for a specific scenario, therefore providing no such prior knowledge that can be adapted and used in medical image classification. Additionally, medical images come with some caveats that make it challenging to apply various self-supervised techniques in medical datasets (discussed in Sect. 2.2).

In this work, we investigate contrastive learning and propose simple and intuitive pretext task-based self-supervised pretraining approaches to improve robustness against noisy labels in the medical image classification problem. We show that self-supervised pretraining can significantly improve the robustness against noisy labels in the existing classification framework. Furthermore, we explored the implications of this pretraining approach on existing LNL methods by pretraining the models before the warm-up phase.

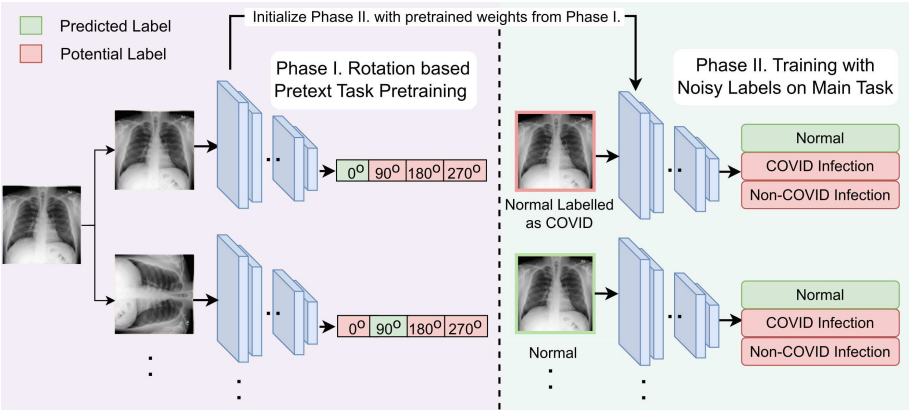


Fig. 1. Our approach involves two phases: I. Pretrain the model with pretext task-based self-supervised technique (left), and II. Retrain the pretrained model on medical image classification with noisy labels using LNL approaches (right).

Our contributions can be summarized as follows: **1)** To our knowledge, we are the first to investigate the use of only self-supervised pretraining to improve robustness in the presence of noisy labels for medical image classification; **2)** We propose the use of pretext task-based self-supervised pretraining in classification with noisy labels, which hasn't been studied even with natural image datasets; **3)** Using two representative datasets, namely X-ray and histopathology images, induced with label noise at various rates, we show that self-supervised pretraining alone improves the feature extractor, thus helping overcome the warm-up obstacle in LNL methods, yielding significantly improved performance while reducing the label memorization.

2 Related Works

2.1 Learning with Noisy Labels in Medical Images

Several methods have been proposed to robustly train medical image classifiers with noisy labels [16]. Pham et al. [30] used label smoothing to reduce the impact of noisy labels in thoracic disease classification. Dgani et al. [4] introduced a noise layer and modified the network architecture to address unreliable labels in breast classification. Le et al. [22] used a sample reweighting technique to robustly train a pancreatic cancer detection model with noisy labels, while Xue et al. [35] used a similar reweighting technique for skin lesion classification with noisy datasets.

Ju et al. [15] used dual-uncertainty estimation to tackle two cases: label noise due to disagreement among experts and single-target label noise, in skin lesions, prostate cancer, and retinal disease. Ying et al. [37] improved COVID-19 chest X-ray classification through techniques like PCA, low-rank representation, neighborhood graph regularization, and k-nearest neighbor. Similarly, Zhou et al.

[44] employed consistency regularization and disentangled distribution learning for multi-label disease classification and severity grading in chest X-rays and diabetic retinopathy. Xue et al. [36] combined a student-teacher network with a co-training strategy to improve prostrate cancer grading, skin classification, chest X-ray classification, and histopathology cancer detection, in different label noise settings. Liu et al. [26] proposed co-correcting, a curriculum learning-based label correction strategy, for robust training with noisy labels in metastatic tissue classification and melanoma classification.

Despite incorporating some concepts from self-supervised learning, no research has explored the impact of just self-supervised pretraining on enhancing robustness against noisy labels.

2.2 Self-supervised Pretraining

Several self-supervised techniques have emerged recently [7], encompassing simple pretext task-solving approaches [5,6,40], contrastive learning methods [3,10,38], and generative approaches [9,28,43]. Generative approaches have shown promise but face challenges due to training instability and high computational resource requirements. Additionally, the majority of recent generative approaches necessitate a Transformer as the backbone, and investigating the robustness of a Transformer-based architecture against label noise, in comparison to a CNN, is a distinct topic of discussion. Furthermore, recent mask image modeling-based generative approaches that learn by randomly masking a certain portion of the image may inadvertently miss crucial features. This issue could be problematic for medical datasets that rely on subtle image cues [13] and requires a separate investigation. Therefore, for this work, we considered focusing solely on contrastive learning, which is widely used, and the pretext task-based approach, which is simple but unexplored, leaving generative approaches for future investigation.

In this study, we chose three pretext tasks: *Rotation prediction*, *Jigsaw puzzle*, and *Jigmag puzzle*, and a contrastive approach: *SimCLR*. *Rotation prediction* [6] trains a model to predict the rotation degree of an image in various orientations. *Jigsaw puzzle* [29] requires training a model to learn to predict the arrangement of shuffled, non-overlapping patches in an image. *Jigmag puzzle* [20], originally proposed for histopathology images, learns to predict the arrangement of patches obtained from magnifying an image at various factors. *SimCLR* utilizes a contrastive loss to compare the representations of different augmented views of the same input, aiming to bring closer the augmented views (positive pairs) of the same image while keeping the augmented views of other images (negative pairs) far apart in the representation space. The benefit of pretraining depends on the suitability of the pretraining task with the main task [23].

We selected these pretext tasks because *Rotation prediction* and *Jigsaw puzzle* are commonly used in the literature, while *Jigmag puzzle* was specifically proposed to address the subtlety of medical images. Other pretext tasks, such as *Colorization* [21], were deemed unsuitable for our grayscale X-ray image and stained histopathology image datasets. *SimCLR* was chosen for the contrastive

approach due to its simplified framework, which eliminates the need for special memory buffers or specialized architectures.

3 Datasets

3.1 COVID-QU-Ex

This dataset is a collection of chest X-ray images obtained from various patients [33] categorized into three groups: COVID infection, Non-COVID infection, and Normal. The dataset consists of a total of 27,132 training images, with 8,561 classified as Normal, 9,010 as Non-COVID-19, and 9,561 as COVID-19 cases. Additionally, there is an exclusive test set containing 6,788 images for evaluation.

3.2 NCT-CRC-HE-100K

This dataset has 100,000 histopathological image patches of size 224×224 extracted from stained tissue slides [17] featuring nine classes, such as adipose, lymphocytes, mucus, etc. The test set uses a different CRC-VAL-HE-7K dataset consisting of 7,180 images, featuring all nine classes of the training set.

4 Experimental Setup

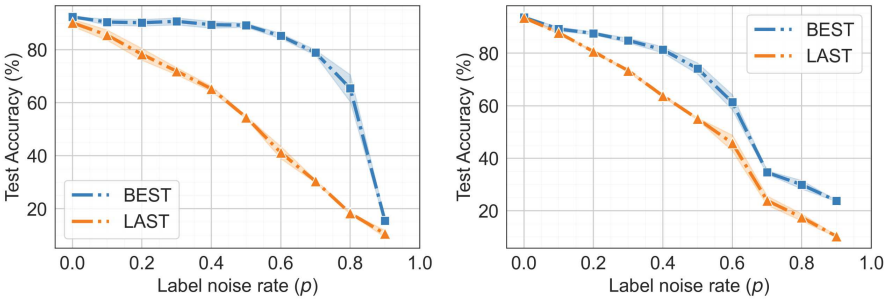
4.1 Random Label Noise

To evaluate how a deep learning classifier performs on high label noise, we randomly flipped all the labels in the training set such that the original labels are assigned to any other labels within the close-set with some probability [14]. Assuming a training dataset $\{(\mathbf{x}_i, y_i)\}_i^n \in \mathcal{D}$, which contains n samples, x_i is a data point belonging to the set $\mathcal{X} \in \mathbb{R}^d$, and y_i is its corresponding class label from a close-set classes $C = \{c_0, c_1, .., c_4\}$. For any sample $(\mathbf{x}_i, y_i)$, we change its label y_i to $\hat{y}_i \overset{p}{\sim} C \setminus y_i$, where p is the noise probability and $C \setminus y_i$ denotes any label of close-set classes other than the true label. The label noise is symmetrical for all the classes within the close set. We conducted experiments using four different noise rates $p \in \{0.5, 0.6, 0.7, 0.8\}$.

As depicted in Fig. 2, the impact of noisy labels on test performance varies across datasets; *NCT-CRC-HE-100K* remains robust to noise below 0.5, whereas *COVID-QU-Ex* is affected at lower rates also.

4.2 Methodology

Our approach involves two stages: i) pretrain a model using self-supervised learning on the given dataset to learn meaningful feature extractors, and ii) train the pretrained model for medical image classification on the same dataset with noisy labels (Fig. 1). We primarily focus on the first stage, experimenting with four self-supervised tasks. In the second stage, we experimented with cross-entropy alone,



(a) NCT-CRC-HE-100K (b) COVID-QU-Ex

Fig. 2. Test performance as a function of training label noise rate (noise probability p) ranging from scale 0 to 1, with the shaded region indicating variability across three experimental trials. BEST indicates the highest test accuracy achieved, while LAST denotes the average test accuracy achieved in the last five epochs.

followed by two state-of-the-art LNL approaches, Co-teaching [8] and DivideMix [25].

Co-teaching selectively samples clean examples by ranking the training loss, while DivideMix utilizes a Gaussian Mixture Model (GMM) to categorize examples into clean and noisy groups based on the training loss of each sample. DivideMix also applies the MixMatch [1] semi-supervised learning approach by treating noisy labels as unlabeled examples. Notably, Co-teaching focuses on clean sample selection, while DivideMix does both clean sample selection and noisy label correction, but both use dual networks and utilize a warm-up phase.

Evaluation: Following [8,25], we evaluate the best test classification accuracy (BEST) and the average test accuracy of the last five epochs (LAST). The test set serves as a pseudo-test set, accessing the model's maximum performance with BEST, while LAST measures if the model has overfitted to noisy labels (see Fig. 2).

4.3 Implementation Details

Self-supervised Pretraining: We utilized the ResNet18 architecture for all experiments. For Rotation prediction, images were resized and underwent strong data augmentations: random horizontal flips, small rotations (10°), sharpness adjustment, equalization, and auto contrast. The model had to predict the rotation angle from four possible angles $(0°, 90°, 180°, 270°)$.

For Jigsaw puzzle solving, we performed similar strong augmentations and divided the resized input image into a 3×3 grid of patches. The patches were resized to 64×64 pixels, normalized with patch mean and standard deviation, and randomly shuffled to create one of the 1000 chosen permutations[1]. Then, they

[1] https://github.com/bbrattoli/JigsawPuzzlePytorch.

were passed through the ResNet18 feature extractor and concatenated before being fed into a fully connected output layer that predicted the input permutation.

For Jigmag puzzle solving, we applied the aforementioned augmentations and randomly magnified the input image at different locations using nine magnification factors ranging from 1 to 5. The magnified patches were resized, normalized, and randomly rearranged into one of the 1000 chosen permutations similar to the Jigsaw puzzle. A fully connected softmax layer after the ResNet feature extractor predicted the input permutation. We used SimCLR [3] implemented in[2], setting the default parameters. The input images were resized and augmented with random horizontal flips, color jitter, and Gaussian blur. Table 1 summarizes all training hyperparameter settings. All the methods were trained until convergence, with the number of epochs and learning rates chosen accordingly. For instance, Rotation prediction performed best with an SGD learning rate of 0.01 compared to other settings, while Jigsaw and Jigmag converged effectively using a batch size of 128.

Table 1. Hyperparameters used for training various self-supervised methods.

Datasets	Method	Input size	Batch	Epochs	Wt decay	Lr	Optim	Sheduler
NCT-CRC-HE-100K	Rotation	224×224	256	70	10^{-4}	0.01	SGD	Cosine Annealing
	Jigsaw	64×64	128	50	10^{-4}	0.001	Adam	Cosine Annealing
	Jigmag	64×64	128	50	10^{-4}	0.001	Adam	Cosine Annealing
	SimCLR	224×224	256	100	10^{-4}	0.001	Adam	Cosine Annealing
COVID-QU-Ex	Rotation	224×224	256	70	10^{-4}	0.01	SGD	Cosine Annealing
	Jigsaw	64×64	128	60	10^{-4}	0.001	Adam	Cosine Annealing
	Jigmag	64×64	128	60	10^{-4}	0.001	Adam	Cosine Annealing
	SimCLR	224×224	256	200	10^{-4}	0.001	Adam	Cosine Annealing

Learning with Noisy Labels: In this stage, we took the ResNet18 feature extractor initialized with from self-supervised training, added a fully connected output layer, and retrained the entire model with noisy labels. In the first set of experiments, we trained the model using standard cross-entropy loss without any modifications. The training process involved a batch size of 256, an SGD optimizer with a momentum of 0.9, weight decay of 10^{-4}, an initial learning rate of 0.01, and 50 training epochs. In the second set of experiments, we used two LNL methods.

[2] https://github.com/sthalles/SimCLR.

For Co-teaching, we followed the original paper's [8] recommendations and set the warm-up epochs to 10, $\tau = p$ and $c = 1$, where p is the label noise rate in data. As for DivideMix, we slightly adjusted the original hyperparameters [25], setting the warm-up epochs to 10, $M = 2$, $T = 0.2$, $\alpha = 4$, $\tau = 0.2$, and $\lambda_u = 0$ for $p = \{0.5, 0.6, 0.7\}$, while λ_u was changed to 0.25 for $p = 0.8$. Both methods maintained other training hyperparameters the same as the standard cross-entropy approach, except for DivideMix, where a batch size of 128 was used. To avoid confirmation bias, both Co-teaching and DivideMix original implementations initialize the dual networks with different weights. Similarly, in our approach, we adopt this strategy by initializing the dual networks with two distinct pretrained weights obtained from separate self-supervised training under the same settings.

All our experiments were implemented in Python 3.8 using the PyTorch 12.1.1 framework and trained on an A100 GPU (40 GB). We ran 3 experimental trials for each case to report the mean and standard deviation.

5 Results

Self-supervised Pretraining Improves Robustness Against Noisy Labels: In Fig. 3, we compared models trained with standard cross-entropy (CE) loss, using weights initialized from the self-supervised pretraining against PyTorch's default randomized He initialization [11].

The results demonstrate that self-supervised pretrained models significantly improve in terms of the BEST and LAST accuracy, particularly at high noise rates in the *NCT-CRC-HE-100K*. Specifically, SimCLR achieved better performance in both BEST and LAST accuracy at all noise rates, while Jigsaw and Jigmag also notably improve the LAST accuracy at $p = \{0.6, 0.7, 0.8\}$.

Similar trends are observed in the *COVID-QU-Ex*, where SimCLR, Jigmag, and Jigsaw outperform others significantly at $p = \{0.5, 0.6\}$ in terms of both BEST and LAST accuracy. At $p = \{0.7, 0.8\}$, rotation performs better, but the improvements are not as good as those observed with SimCLR, Jigsaw, and Jigmag in the $p = \{0.5, 0.6\}$ range. However, SimCLR, Jigmag, and Jigsaw performed worst than the cross entropy in the range $p = \{0.7, 0.8\}$.

The choice of the best self-supervised task varies based on the dataset, noise rate, and evaluation criteria, but the results achieved using self-supervised pretraining consistently show better performance compared to directly starting training from PyTorch's default randomized He initialization.

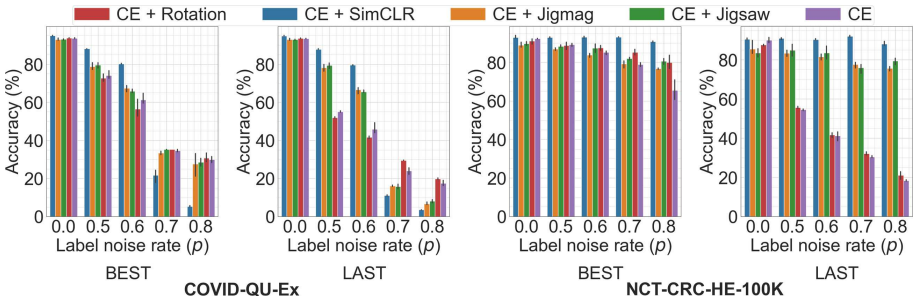
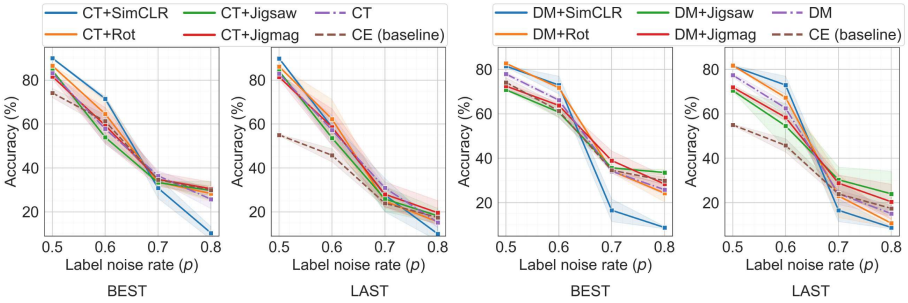


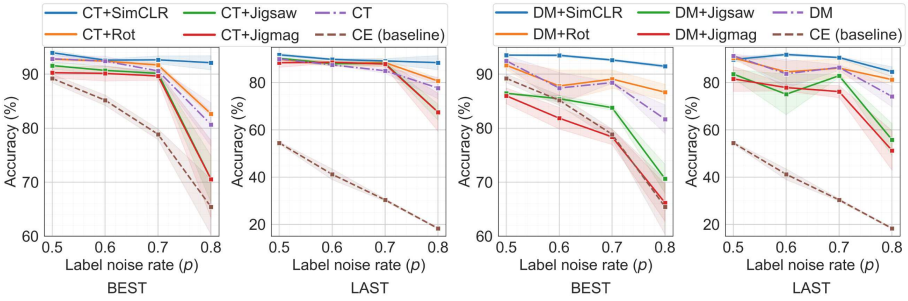
Fig. 3. Performance comparison of models trained starting from default weights vs. trained from weights initialized from self-supervised pretraining, when using standard cross entropy (CE) loss at different training label noise rates (p), in *COVID-QU-Ex* and *NCT-CRC-HE-100K*. BEST denotes the best test accuracy, while LAST denotes the average test accuracy achieved of the last five epochs. The experiments were run for three experimental trials to report the error bar.

Self-supervised Pretraining for LNL Methods: We compared LNL methods trained from PyTorch's default He initialization with those trained using weights initialized from a self-supervised pretraining in Fig. 4. The results show that the LNL methods already achieved good performance in the *NCT-CRC-HE-100K*, with only a small room for improvement at a lower noise rate. At higher noise rates, SimCLR achieved the highest BEST and LAST accuracy, followed by Rotation prediction. Initializing LNL with Jigsaw and Jigmag didn't improve the performance, but rather degraded it.

In the *COVID-QU-Ex*, we observed an improvement in classification performance in the noise range $p = \{0.5, 0.6\}$ for both Coteaching and DivideMix when using pretrained weights from the SimCLR and Rotation prediction task. However, beyond $p = \{0.7, 0.8\}$, the scores exhibited high variability, making it difficult to identify the best performance. SimCLR struggled with high noise rates, possibly due to excessive contrastive learning augmentations that overlooked vital subtle features valuable for discerning classes at high label noise. But, this speculation needs further study. Additionally, in *COVID-QU-Ex*, we noticed that the clean samples selected by LNL methods were biased towards one class and ignored the other classes, particularly at high noise levels, making it intriguing to investigate the cause behind this phenomenon in the future.



(a) COVID-QU-Ex



(b) NCT-CRC-HE-100K

Fig. 4. Performance comparison of existing LNL methods when initialized with self-supervised pretraining against baselines at different training label noise rates (p), in (a) *COVID-QU-Ex* and (b) *NCT-CRC-HE-100K*. Cross entropy (CE) denotes the actual baseline, Coteaching (CT) and Dividemix (DM) are existing LNL methods, and the term after + represents various self-supervised pretraining methods. BEST denotes the best test accuracy, while LAST denotes the average test accuracy achieved of the last five epochs. The shaded region along the line indicates the variability across three experimental trials.

6 Conclusion

We examined the effectiveness of utilizing self-supervised pretraining alone to improve the model's robustness against noisy labels in medical image classification. The choice of self-supervised task varied depending on the dataset, noise rate, and, evaluation criteria, with SimCLR consistently yielding the best results in most cases with LNL.

This study addresses a gap in the current research by serving as a first demonstration of the benefits of various self-supervised pretraining for medical image classification with noisy labels and offering valuable insights to mitigate the impact of high label noise. In the future, we plan to investigate additional self-supervised baselines and further explore how the nature and size of the dataset

influence the improvements offered by self-supervised pretraining in terms of robustness against noisy labels.

Additionally, in this work, we have limited the investigation to CNN-based architecture. It would be interesting to investigate how recent Transformer-based architectures behave under various levels of label noise, whether off-the-shelf LNL methods work with Transformer architecture, and how Transformer-based self-supervised techniques improve robustness against noisy labels.

Acknowledgments. Research reported in this publication was supported by the National Institute of General Medical Sciences Award No. R35GM128877 of the National Institutes of Health, the Office of Advanced Cyber Infrastructure Award No. 1808530 of the National Science Foundation, and the Division Of Chemistry, Bioengineering, Environmental, and Transport Systems Award No. 2245152 of the National Science Foundation. We would like to thank the Research Computing team [31] at the Rochester Institute of Technology for proving computing resources for this research.

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems 32 (2019)
2. Chen, P., Liao, B.B., Chen, G., Zhang, S.: Understanding and utilizing deep neural networks trained with noisy labels. In: International Conference on Machine Learning, pp. 1062–1070. PMLR (2019)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Dgani, Y., Greenspan, H., Goldberger, J.: Training a neural network based on unreliable human annotation of medical images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 39–42. IEEE (2018)
5. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430 (2015)
6. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
7. Gui, J., Chen, T., Cao, Q., Sun, Z., Luo, H., Tao, D.: A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. arXiv preprint arXiv:2301.05712 (2023)
8. Han, B., et al.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances In Neural Information Processing Systems 31 (2018)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)

11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
12. Hu, W., Li, Z., Yu, D.: Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. arXiv preprint arXiv:1905.11368 (2019)
13. Huang, S.-C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Med. **6**(1), 74 (2023)
14. Jiang, L., Zhou, Z., Leung, T., Li, L.-J., Fei-Fei, L.: Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning, pp. 2304–2313. PMLR (2018)
15. Lie, J., et al.: Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Trans. Med. Imaging **41**(6), 1533–1546 (2022)
16. Karimi, D., Dou, H., Warfield, K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Medical Image Anal. **65**, 101759 (2020)
17. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. **16**(1), e1002730 (2019)
18. Khanal, B., Kamrul Hasan, S.M., Khanal, B., Linte, C.A.: Investigating the impact of class-dependent label noise in medical image classification. In: Medical Imaging 2023: Image Processing, vol. 12464, pp. 728–733. SPIE (2023)
19. Khanal, B., Kanan, C.: How does heterogeneous label noise impact generalization in neural nets? In: Bebis, G., et al. (eds.) ISVC 2021. LNCS, vol. 13018, pp. 229–241. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-90436-4_18
20. Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging **40**(10), 2845–2856 (2021)
21. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6874–6883 (2017)
22. Le, H., Samaras, D., Kurc, T., Gupta, R., Shroyer, K., Saltz, J.: Pancreatic cancer detection in whole slide images using noisy label annotations. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 541–549. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_60
23. Lee, J.D., Lei, Q., Saunshi, N., Zhuo, J.: Provable self-supervised learning: Predicting what you already know helps. Adv. Neural. Inf. Process. Syst. **34**, 309–323 (2021)
24. Lee, K., Yun, S., Lee, K., Lee, H., Li, B., Shin, J.: Robust inference via generative classifiers for handling noisy labels. In: International Conference on Machine Learning, pp. 3763–3772. PMLR (2019)
25. Li, J., Socher, R., Hoi, S.C.H.: Dividemix: learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations (2020)
26. Liu, J., Li, R., Sun, C.: Co-correcting: noise-tolerant medical image classification via mutual label correction. IEEE Trans. Med. Imaging **40**(12), 3580–3592 (2021)
27. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Adv. Neural. Inf. Process. Syst. **33**, 20331–20342 (2020)

28. Liu, X., et al.: Self-supervised learning: generative or contrastive. IEEE Trans. Knowl. Data Eng. **35**(1), 857–876 (2021)
29. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
30. 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)
31. Rochester Institute of Technology. Research computing services (2022)
32. Song, H., Kim, M., Park, D., Lee, J.-G.: How does early stopping help generalization against label noise? arXiv preprint arXiv:1911.08059 (2019)
33. Tahir, A.M et al.: COVID-QU-Ex Dataset (2022)
34. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: a joint training method with co-regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
35. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.-A.: Robust learning at noisy labeled medical images: Applied to skin lesion classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1280–1283. IEEE (2019)
36. Xue, C., Lequan, Yu., Chen, P., Dou, Q., Heng, P.-A.: Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE Trans. Med. Imaging **41**(6), 1371–1382 (2022)
37. Ying, X., Liu, H., Huang, R.: Covid-19 chest x-ray image classification in the presence of noisy labels. Displays, 102370 (2023)
38. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. In: International Conference on Machine Learning, pp. 12310–12320. PMLR (2021)
39. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021)
40. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
41. Zhang, X.: Codim: learning with noisy labels via contrastive semi-supervised learning. arXiv preprint arXiv:2111.11652 (2021)
42. Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A.M., Litany, O.: Contrast to divide: Self-supervised pre-training for learning with noisy labels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1657–1667 (2022)
43. Zhou, J.: ibot: image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
44. Zhou, Y., Huang, L., Zhou, T., Sun, H.: Combating medical noisy labels by disentangled distribution learning and consistency regularization. Futur. Gener. Comput. Syst. **141**, 567–576 (2023)