Investigating the impact of class-dependent label noise in medical image classification

Bidur Khanal, S.M. Kamrul Hasan, Bishesh Khanal, Cristian A. Linte

Bidur Khanal, S.M. Kamrul Hasan, Bishesh Khanal, Cristian A. Linte, "Investigating the impact of class-dependent label noise in medical image classification," Proc. SPIE 12464, Medical Imaging 2023: Image Processing, 1246437 (3 April 2023); doi: 10.1117/12.2654420

SPIE.

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

Investigating the impact of class-dependent label noise in medical image classification

Bidur Khanal[a] (✉), S. M. Kamrul Hasan[a] (✉), Bishesh Khanal[c], and Cristian A. Linte[a,b]

[a]Center for Imaging Science, Rochester Institute of Technology, NY, USA
[b]Biomedical Engineering, Rochester Institute of Technology, NY, USA
[c]NepAl Applied Mathematics and Informatics Institute for research (NAAMII), KTM, Nepal

ABSTRACT

Label noise is inevitable in medical image databases developed for deep learning due to the inter-observer variability caused by the different levels of expertise of the experts annotating the images, and, in some cases, the automated methods that generate labels from medical reports. It is known that incorrect annotations or label noise can degrade the actual performance of supervised deep learning models and can bias the model's evaluation. Existing literature show that noise in one class has minimal impact on the model's performance for another class in natural image classification problems where different target classes have a relatively distinct shape and share minimal visual cues for knowledge transfer among the classes. However, it is not clear how class-dependent label noise affects the model's performance when operating on medical images, for which different output classes can be difficult to distinguish even for experts, and there is a high possibility of knowledge transfer across classes during the training period. We hypothesize that for medical image classification tasks where the different classes share a very similar shape with differences only in texture, the noisy label for one class might affect the performance across other classes, unlike the case when the target classes have different shapes and are visually distinct. In this paper, we study this hypothesis using two publicly available datasets: a 2D organ classification dataset with target organ classes being visually distinct, and a histopathology image classification dataset where the target classes look very similar visually. Our results show that the label noise in one class has a much higher impact on the model's performance on other classes for the histopathology dataset compared to the organ dataset.

Keywords: Label noise, image classification, class-dependent label noise

1. INTRODUCTION

Supervised deep learning methods for medical image classification problems have shown great promise in diverse healthcare applications in real clinical settings, such as patient triage based on from X-ray images of Tuberculosis [1], view classification in echocardiography [2], and cancer screening [3]. However, building large annotated medical image databases required for training such models is very expensive and can often contain label noise due to the inter-observer variability and different levels of expertise of the annotators. The label noise can be more severe when efforts are made to reduce the cost of annotation, such as using non-experts or automated methods to generate labels from medical reports [4,5]. Although several recent efforts have focused on "cleaning" noisy labels with minimal budget [6], it is still not clear how the intricate relationships between different factors such as class-specific noise levels and the visual similarity between the target classes affect the model's performance on individual class predictions. Understanding such interdependence could help develop better methods that are more robust to label noise, that can clean noisy labels, or even utilize the information available on the label noise.

(✉)
Further author information:
Bidur Khanal (E-mail: bk9618@rit.edu)
S. M. Kamrul Hasan (E-mail: sh3190@rit.edu)
Bishesh Khanal (E-mail: bishesh.khanal@naamii.org.np)
Cristian A. Linte (E-mail: calbme@rit.edu)

Medical Imaging 2023: Image Processing, edited by Olivier Colliot, Ivana Išgum,
Proc. of SPIE Vol. 12464, 1246437 · © 2023 SPIE
1605-7422 · doi: 10.1117/12.2654420

Recently, *Khanal et al.* used two common computer vision datasets to explore the effects of class-dependent label noise, i.e., label noise that depends on class.[7] They showed that with limited knowledge transfer from one class to another, label noise on one class did not affect the performance of the model on the classes that did not have label noise. We argue that this result may not always translate in cases when the target classes have very similar visual appearance that are not easily distinguishable. Such scenarios are fairly common in medical image classification problems and hence require further attention, especially because medical images exhibit several differences in the way they are annotated and pose some distinctive characteristics, such as the requirement of experts, class imbalance for normal vs abnormal classes, same anatomy or shape with differences in texture for different classes, etc. To the best of our knowledge, the impact of such class-dependent label noise on medical images has not been explored yet.

We hypothesize that when the target classes have visually similar shapes and hence a stronger likelihood for knowledge transfer across the classes, the impact of label noise in one class might have a much more pronounced impact on the model's performance on other classes. To investigate this hypothesis, we studied two different types of publicly available datasets: i) a 2D organs classification dataset[8] with the axial view of 11 organs where the target organ classes are visually distinct even for a layman, and ii) a histopathology image classification dataset[8] with 9 classes where the target classes look very similar (visually) for a layman.

2. OVERVIEW OF THE PROBLEM SETUP

In this section, we describe how the class-dependent label noise is introduced into the dataset: given a data distribution $p(x_i, y_i)$ over images and class labels, we have a source domain having a training set, $\{(\mathbf{x}_i, y_i)\}_i^n \in \mathcal{D}$, with n number of samples, where $x_i \in \mathcal{X} \in \mathbb{R}^d$ is the data point, $y_i \in C$ is the corresponding class label, and $C = \{c_0, c_1, .., c_4\}$ denotes the close-set of classes. Within the close-set, certain classes are chosen to be corrupted, while others are unchanged. For any given sample $(\mathbf{x}_i, y_i)$, if y_i belongs to the list of classes to-be-corrupted, it is replaced with another class label from the to-be-corrupted class list. The replacement is accomplished using random sampling with a probability. While the training set is corrupted by a probability, the test set is kept unaltered. The validation set, which is evaluated to select the best model from training, can either be corrupted or kept unaltered. The unaltered validation set resembles the case where a small, accurately-labeled set is known to the prior. Therefore, we didn't corrupt the validation set with any label noise.

A neural network $f : \mathcal{X} \longrightarrow C$ is trained to categorize the classes after injecting label noise in the training set. The purpose is to investigate how the network behaves when class-dependent label noise is introduced in other classes while classifying the noise-less classes. 1 shows the pseudocode for producing class-dependent label noise.

Pseudocode 1 Strategy to Induce Class-dependent Label Noise

Input: Dataset $\{(\mathbf{x}_i, y_i)\}_i^n \in \mathcal{D}$, where $x_i \in \mathcal{X} \in \mathbb{R}^d$ and $y_i \in C$, Classes $C = \{c_0, c_1, .., c_4\}$
Select : Uncorrupt classes list $C_{uncorrupt} = \{c_0, \}$,
Corrupt classes list $C_{corrupt} = C \setminus C_{uncorrupt}$, ▷ Subtract uncorrupt subset from all classes
Corrupt probability p

for $i = 1$ to n **do**
  if $y_i \in C_{corrupt}$ **then**
    Corrupt target list $C_{target} = C_{corrupt} \setminus y_i$ ▷ remove ground-truth class from target list
    $y_i \overset{p}{\sim} C_{target}$ ▷ replace with a class from target list with probability p

3. DATASETS

To further validate our proposed approach, we evaluate our method on two publicly available medical image classification benchmarks. We choose two datasets, OrganAMNIST,[9] and PathMNIST,[10] from a larger-scale MNIST-like collection of standardized lightweight medical images, MedMNIST.[8] We chose one dataset (3.1) with images that have distinct visual shapes and appearance across classes, and another (3.2) with images that

look visually similar with small differences across most classes. This is based on the premise that similar-looking classes transfer knowledge across them, therefore label noise in one class is likely to influence other clean classes. Hence, we wanted to investigate both scenarios where there may or may not be sufficient transfer across classes.

3.1 OrganAMNIST

The OrganAMNIST[8] dataset was created from axial plane views of 3D computed tomography (CT) images of Liver Tumor Segmentation Benchmark[10] by downsampling the original image to 28×28 resolution gray-scale images. The dataset contains 34,581 training, 6,491 validation, and 17,778 test images, each belonging to only one of 11 body organs such as bladder, heart, kidney, etc.

3.2 PathMNIST

The PathMNIST[8] dataset was created by down-sampling image patches from stained histological images[9] to 28×28 resolution. Each patch represents only one of 9 classes such as adipose, lymphocytes, mucus, etc. There are 89,996 training images and 10,004 validation images. A separate distinct test set contains 7,180 images.

3.3 EXPERIMENTS

For all our experiments, we used the naïve ResNet-18 architecture[11] that takes in 28×28 resolution RGB or grayscale input. We experimented with the two datasets separately following the strategy discussed in section 1 to corrupt the labels to induce label noise. For OrganAMNIST, out of 11 classes, we corrupted all the classes of the training set except one (referred to as an uncorrupted class). The label corruption was done randomly with eight different probabilities 0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 1.0 each. We made sure that there are at least 3 trials for each corruption probability, obtaining mean and standard deviation of test per-class accuracy. We conducted the same set of experiments with three randomly chosen uncorrupted classes: bladder, heart, and left lung. In total, we ran 8×3 experiments for all three classes resulting in total $8 \times 3 \times 3$ experiments. For each experiment, the network was initialized randomly and trained for 100 epochs with batch size of 128 using an initial learning rate of 0.1, SGD with 0.9 momentum and $1 \times e^{-4}$ decay rate, and cosine annealing learning scheduler. We selected the best model based on the mean per-class accuracy on the validation set instead of using overall accuracy, to prevent biasness towards a class as the dataset is imbalanced. The validation set wasn't corrupted assuming the case that a small, correctly labeled set is known to the prior.

For the PathMNIST dataset, we used the same strategy for all experiments. The only difference was the total number of classes and selected uncorrupted classes: adipose, mucus, and smooth muscle. Same as before, there were $8 \times 3 \times 3$ experiment trials in total. Moreover, all hyper-parameters were maintained the same as in OrganAMNIST, except for the initial learning rate which was changed to 0.01.

3.4 EVALUATION METRICS

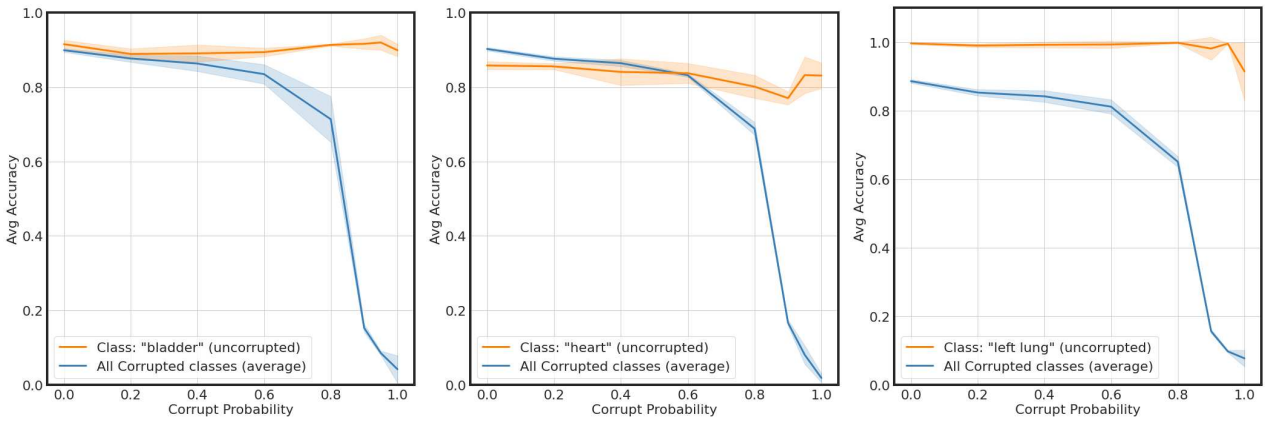
We used per-class accuracy as the main evaluation metric to measure the performance in each class, computed as: $\frac{\text{number of correct prediction of class } c}{\text{total number samples of class } c}$, which is essentially equivalent to the recall rate. We computed the average per-class accuracy of all the corrupted classes, reported as corrupted-class accuracy, while we computed accuracy for uncorrupted-class separately. Both values were compared to see the impact of corrupted classes on uncorrupted classes at a given probability.

4. RESULTS AND DISCUSSION

We compare the average per-class accuracy and accuracy of uncorrupted class at each corruption probability using two line graphs as illustrated in Fig. 1 and Fig. 2. As expected, the curve representing the uncorrupted class drops with the increase in corruption strength in the training set. But, the performance drop is not linear. Initially, the performance drop in the corrupted classes is gradual up to 0.8 corruption probability; however, after 0.8 corruption probability, there is a drastic drop in the test performance. This depicts that the impact of label noise can be reduced up to some extent if a clean validation set is used to evaluate and select the best model while training.[12]

When there is no label noise, i.e when corruption probability is 0.0, it is similar to a naïve classification model and the average accuracy is maximum ($\sim$ 90%). At 100% label noise, i.e., when the corruption probability is 1, the average accuracy drops sharply, demonstrating the detrimental effect of label noise. For both datasets, we see identical behavior in the corrupted class. However, the observation is slightly different in uncorrupted classes, when comparing the two datasets. The accuracy of the uncorrupted class (orange line) in OrganAMNIST (Fig. 1) is consistent with less variance, while PathMNIST shows high variance (Fig. 2). As *Khanal et al* discussed, in Fig. 1 and 2, we see that the accuracy of the uncorrupted class (orange line) doesn't drop as that of the corrupted classes, demonstrating that the neural network can still capture relevant information to correctly classify the clean class even in the presence of other noisy classes.

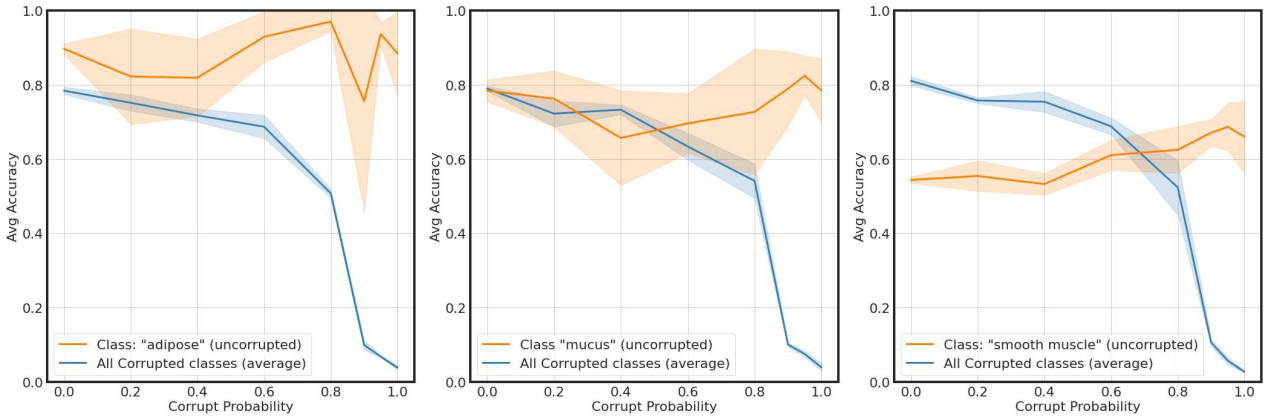
In OrganAMNIST, the performance in the uncorrupted class does not drop with increasing label noise in other classes for visually distinct class (Fig 1). For instance, when "left-lung" is chosen as the uncorrupted class, its performance does not drop, even when other classes are fully corrupted. The per-class accuracy curve of other uncorrupted classes "bladder" and "heart" isn't perfectly straight, but still consistent compared to PathMNIST. For PathMNIST, the variance in the per-class accuracy of uncorrupted classes is high across all corruption probabilities and does not follow any definite pattern. It is interesting to see in Fig. 2 (c) that for the class "smooth muscle", its per-class accuracy improves with the increase in label noise in other corrupted classes. Intuitively, the label noise in other classes negatively interfered with the uncorrupted class, thus acting as a regularizer and improving the performance in the uncorrupted class. The experimental results support our hypothesis that similar-looking classes transfer knowledge across them, therefore label noise in one class is likely to influence other clean classes.



(a) (b) (c)

Figure 1: Results for OrganAMNIST: Mean per-class accuracy of classes with noisy labels (corrupted classes) vs. class accuracy of a noise-less class (uncorrupted class) at each corruption probability. The plots show how the performance in noise-less class is impacted by noisy labels from other classes at various strength of corruption. Class accuracy of clean class does not drop with the drop in per-class accuracy of noisy classes. Panels (a), (b), and (c) show similar patterns when different classes are chosen as uncorrupted classes. The shaded region in plots indicate the variance across experiment trials.

To the best of our knowledge, we are the first to analyze the impact of class-dependent label noise in medical image dataset. Although previous works[13–15] have studied label noise in medical images and adapted various strategies to improve robustness in its presence, the challenge that is posed when some classes encounter more noisy labels than others has yet to be explored, especially when the visual cues for identifying the specific medical condition are so subtle that even expert annotators can get confused. Depending upon how classes share the joint knowledge, they can either positively or negatively interfere during training. Such interference is likely to also transfer the effect of the label, thus, either degrading or boosting performance in class with clean labels.



(a) (b) (c)

Figure 2: Results for PathMNIST: Mean per-class accuracy of classes with noisy labels (corrupted classes) vs. class accuracy of a clean class (uncorrupted class) at each corruption probability. The plots show how the performance in noise-less class is impacted by noisy labels from other classes at various strength of corruption. Class accuracy of clean class is highly variable and does not show a consistent pattern. Panels (a), (b), and (c) shows the similar pattern when different classes are chosen as uncorrupted classes.The shaded region in plots indicate the variance across experiment trials.

5. CONCLUSION AND FUTURE WORKS

In this work, we studied the impact of class-dependent label noise in medical image classification using two datasets: OrganAMNIST and PathMNIST. Depending upon the dataset and how distinct classes are, the noisy labels in some classes may or may not impact the performance in clean classes. Using PathMNIST dataset, we highlighted that label noise can impact the performance in a correctly labeled class, if other classes contain label noise. In this study, we limited our work to a small-resolution dataset because even at low input resolution, the classification performance is fairly good for clean labels in both datasets. However, we intend to investigate this matter further in the future using additional high-resolution datasets. Additionally, in our experiments, we considered a small, clean validation set is known to the prior, which is a fair assumption in a real-world setting, as a small subset can be carefully labeled. However, we are also interested in studying the case where the clean validation set isn't available and the model is allowed to fit with noisy labels of the training set. Ultimately, with further insights, we intend to build a method to learn with noisy labels in medical images, reducing the impact of class-dependent 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, and the Office of Advanced Cyber infrastructure Award No. 1808530 of the National Science Foundation.

REFERENCES

[1] Khan, F. A., Majidulla, A., Tavaziva, G., Nazish, A., Abidi, S. K., Benedetti, A., Menzies, D., Johnston, J. C., Khan, A. J., and Saeed, S., "Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease," *The Lancet Digital Health* **2**(11), e573–e581 (2020).

[2] Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J. H., Harrington, R. A., Liang, D. H., Ashley, E. A., and Zou, J. Y., "Deep learning interpretation of echocardiograms," *NPJ digital medicine* **3**(1), 1–10 (2020).

[3] Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., and Umutlu, L., "Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis," *The Lancet Digital Health* **4**(7), e507–e519 (2022).

[4] This, H., "Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation," *Radiology* **294**, 421–431 (2020).

[5] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in [*Proceedings of the AAAI conference on artificial intelligence*], **33**(01), 590–597 (2019).

[6] Bernhardt, M., Castro, D. C., Tanno, R., Schwaighofer, A., Tezcan, K. C., Monteiro, M., Bannur, S., Lungren, M. P., Nori, A., Glocker, B., et al., "Active label cleaning for improved dataset quality under resource constraints," *Nature communications* **13**(1), 1–11 (2022).

[7] Khanal, B. and Kanan, C., "How does heterogeneous label noise impact generalization in neural nets?," in [*International Symposium on Visual Computing*], 229–241, Springer (2021).

[8] Yang, J., Shi, R., and Ni, B., "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in [*IEEE 18th International Symposium on Biomedical Imaging (ISBI)*], 191–195 (2021).

[9] Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., et al., "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS Medicine* **16**(1), e1002730 (2019).

[10] Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.-W., Han, X., Heng, P.-A., Hesser, J., et al., "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056* (2019).

[11] He, K., Zhang, X., Ren, S., and Sun, J., "Identity mappings in deep residual networks," in [*European conference on computer vision*], 630–645, Springer (2016).

[12] Song, H., Kim, M., Park, D., and Lee, J.-G., "How does early stopping help generalization against label noise?," *arXiv preprint arXiv:1911.08059* (2019).

[13] Xue, C., Dou, Q., Shi, X., Chen, H., and 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)*], 1280–1283, IEEE (2019).

[14] Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., and Ge, Z., "Improving medical images classification with label noise using dual-uncertainty estimation," *IEEE Transactions on Medical Imaging* (2022).

[15] Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A., "Deep learning with noisy labels: exploring techniques and remedies in medical image analysis," *Medical Image Analysis* **65**, 101759 (2020).