Detecting noisy labels with repeated cross-validations

Jianan Chen[1,2,*] ✉, Vishwesh Ramanathan[1,2,*], Tony Xu[1,2], and Anne L. Martel[1,2] ✉

[1] Department of Medical Biophysics, University of Toronto, Toronto, ON, CA
chenjn2010@gmail.com; a.martel@utoronto.ca
[2] Sunnybrook Research Institute, Toronto, ON, CA

Abstract. Machine learning models experience deteriorated performance when trained in the presence of noisy labels. This is particularly problematic for medical tasks, such as survival prediction, which typically face high label noise complexity with few clear-cut solutions. Inspired by the large fluctuations across folds in the cross-validation performance of survival analyses, we design Monte-Carlo experiments to show that such fluctuation could be caused by label noise. We propose two novel and straightforward label noise detection algorithms that effectively identify noisy examples by pinpointing the samples that more frequently contribute to inferior cross-validation results. We first introduce Repeated Cross-Validation (ReCoV), a parameter-free label noise detection algorithm that is robust to model choice. We further develop fastReCoV, a less robust but more tractable and efficient variant of ReCoV suitable for deep learning applications. Through extensive experiments, we show that ReCoV and fastReCoV achieve state-of-the-art label noise detection performance in a wide range of modalities, models and tasks, including survival analysis, which has yet to be addressed in the literature. Our code and data are publicly available at https://github.com/GJiananChen/ReCoV.

1 Introduction

Label noise, where labels inaccurately represent the data, is ubiquitous in real-world machine learning datasets. Training machine learning models in the presence of noisy labels may lead to deteriorated performance and inaccurate conclusions. Mislabelling can arise from various sources such as human error, inadequate data quality, and encoding errors [9]. In medical data, attaining a definitive gold standard label is often intractable, and is further complicated by inter-observer variability. Consequently, label noise in medical datasets is in general more complex and more difficult to address [12, 13].

A substantial body of research has been dedicated to the problem of label noise in classification tasks. A major branch of such methods involves detecting and removing noisy labels based on the metrics, predictions and uncertainties

[*] Equal contributions, alphabetical order

produced by the classification model itself. However, the performance of these approaches is then limited by model robustness, and may suffer from over- or under-detection of noisy labels [9]. Another avenue focuses on increasing model reliability despite the known presence of label noise, with techniques such as robust loss functions and network structures, semi-supervised data selection and data re-weighting [17,25,26]. These algorithms have also shown promise but face the risk of over-fitting to specific datasets. In our work, we choose to focus on the explicit detection of noisy labels as it offers an opportunity for experts to review and potentially re-evaluate ambiguous and challenging cases [16].

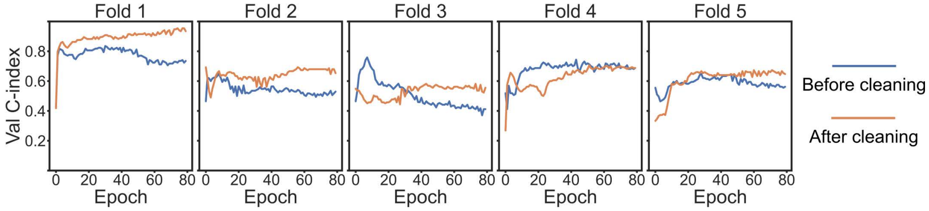

Fig. 1. Validation performance (concordance-index) on the TCIA Head-Neck-PET-CT dataset, comparing performance pre- and post-cleaning of noisy samples.

The differences in model performance across cross-validation folds are usually perceived as completely random or as “normal” fluctuations caused by different seeds and software implementations [15, 20]. Nevertheless, inspired by the frequent observation of large performance gaps between different cross-validation folds in outcome prediction tasks, we hypothesize that such differences may not be entirely random but could be influenced by label noise, and if so, such influence may be leveraged for identifying noisy samples (**Fig. 2**). We thereby propose Repeated Cross-Validations (ReCoV), a model-agnostic and parameter-free workflow for label noise detection, and fastReCoV, a less powerful but significantly more efficient variant of ReCoV tailored to resource-intensive modern deep learning frameworks. With extensive experiments in a large variety of datasets, models, and tasks, we empirically corroborate the influence of label noise on fold-specific validation performance and demonstrate that ReCoV and fastReCoV achieve state-of-the-art performance in noisy label detection in medical imaging datasets with real-world label noise.

2 Methods

Assumptions Similar to existing research in this field [10, 18], we assume that a class-conditional noise process maps the true label y to the observed label $\tilde{y}$. Specifically, for each label index j, there exists an independent mislabeling probability $p(\tilde{y} = i \mid y = j)$, where i, j belong to a label set $[m] = \{1, 2, ..., m\}$.

Modeling cross-validations in classification task In a dataset with N samples, among which ϵN samples are noisy, where $\epsilon \in (0, 0.5)$ is the noise ratio. For a k-fold cross-validation, the probability of each sample landing in a specific fold follows Bernoulli($\frac{1}{k}$). The expected number of noisy samples in each ordered fold (*i.e.* folds ordered from the most number of noisy samples to the least number of noisy samples) follows Hypergeometric($\mathrm{N}, \epsilon\mathrm{N}, \mathrm{N}/k$) [11]. The expected number of noisy samples in the fold with the highest number of noisy samples $\mathbb{E}(n_{\text{most}})$ can be calculated by finding the maximum of the probability mass function of the hyper-geometric distribution [7], while the expected number of noisy samples in each fold $\mathbb{E}(n_{\text{mean}}) = \frac{\epsilon \mathrm{N}}{k}$. In **each iteration** of cross-validation with a completely random split, $\mathbb{E}(n_{\text{diff}}) = \mathbb{E}(n_{\text{most}}) - \mathbb{E}(n_{\text{mean}})$ **more noisy samples** are in the **fold with the highest number of noisy samples** compared to the average. It also follows that the distribution of clean and noisy samples in the fold with the most noisy samples (coined as the occurrence distributions) follow two Binomial distributions $\mathrm{B}(n_{\text{fold}}, p_{\text{clean}})$ and $\mathrm{B}(n_{\text{fold}}, p_{\text{noisy}})$, where

$$n_{\text{fold}} = \frac{\mathrm{N}}{k} \tag{1}$$

$$p_{\text{noisy}} = \frac{\mathbb{E}(n_{\text{most}})}{\mathrm{N}/k} \tag{2}$$

$$p_{\text{clean}} = 1 - p_{\text{noisy}} \tag{3}$$

Repeated Cross-Validations (ReCoV) for noise detection We propose ReCoV (**Algorithm S1**) to identify noisy examples by pinpointing the samples that more frequently contribute to inferior cross-validation results. In each repeated run of cross-validations with a different data split, the indices of samples in the worst-performing fold are appended to a list, referred to as the candidate pool. As more runs are recorded, the standard deviations of the occurrence distributions of clean and noisy samples increase at a lower rate compared to the increase in the difference of their means in relation to N_{runs}. As a result, the distributions become more separated with the increase of N_{runs}. With a large number of runs, the occurrence distributions will be clearly separable, preventing over-/under- detection. The effect of other factors such as model strength or data imbalance will be alleviated with repeating multiple independent runs.

FastReCoV In order to make ReCoV computationally feasible for deep learning models and high-dimensional inputs, we introduce fastReCoV (**Algorithm 1**). Based on the framework of ReCoV, fastReCoV achieves much improved efficiency with a few modifications.

In fastReCoV, we employ a genetic-algorithm-inspired approach to solve the optimization problem of making noisy samples accumulate in the worst fold faster. Specifically, we construct a memory bank that stores the probability of each sample being noisy, where a lower memory value means a higher chance

Algorithm 1: Pseudocode of fastReCoV in a Python-like style

```
# N_runs: number of runs
# k: number of folds
# $p$: a list of sampling probabilities for each sample
# $\tau$: temperature parameter controlling sampling process

memory = Zeros(len(data)) # initialize memory as 0's
identified = [] # initialize identified noisy samples as empty
p = Ones(len(data))/len(data) # initialize p to be uniform

for run in range(N_runs): # repeat for N_runs times
    # split data by weighted sampling w/o replacement
    train_sets, val_sets = WeightedSample(data, k, p, replace=False)
    # Randomly drop $\alpha$ identified noisy samples from training
    train_sets = DropNoisy(train_sets, identified, $\beta$)
    models.train(train_sets) # train k models with k train_sets
    sample_val_metric = models.test(val_sets) # calculate metrics

    # Update memory with exponential moving average across runs
    memory = $\alpha$ * memory + (1-$\alpha$) * sample_val_metric
    identified = data[memory<T]) # identify with threshold T

    p = Softmax(memory/$\tau$)] # update sampling probabilities
```

for the sample to be noisy. The memory bank is updated in each run using exponential average with a task-specific ranking metric (**Table S1**). A weighted sampling function based on memory values is incorporated to give low memory value samples higher chance to be assigned to the worse folds. The probabilities used in the weighted sampling involves a temperature parameter τ that balances the trade-off between efficiency and the potential to find better local minimums. Bigger τ induce a more uniform distribution and relatively random splits, and smaller τ leads to a sharper distribution, creating more greedy searching of noisy samples. The memory value for each sample is also used for segregating noisy samples from clean samples. The segregation threshold can be a predefined threshold, a predefined quantile, or based on Gaussian mixture models. Additionally, we implement a procedure in which a certain percentage β of the most noisy samples are randomly excluded from the training process in each run, to improve the robustness of trained models.

3 Experiments

We performed experiments in four public datasets with various tasks, models, and types of noises. Specifically, we used Mushroom and CIFAR10N as sandboxes to lay out mathematical foundations and quantitatively evaluate noisy label detection performance. We then included PANDA and HECKTOR to ensure that our method is practical and robust in large-scale real-world medical imaging datasets. 5-fold cross-validation is used in all ReCoV experiments. For detailed descriptions about model structures please refer to the corresponding references and our Github repository. Hyperparameters of fastReCoV for different experiments are summarized in **Tabel S1**.

Binary classification on Mushroom with ReCoV The Mushroom dataset (n=8124, https://archive.ics.uci.edu/ml/datasets/mushroom) is a widely used noisy label detection benchmark from the UCI data repository. Mushroom comprises 22 categorical features (converted to 117 dummy variables) used to predict whether a mushroom is poisonous).

Experimental design: First, we simulate class-conditional label noise by randomly flipping a proportion of labels ($\epsilon = 0.1$). Next, we run a Monte-Carlo simulation to simulate how clean and noisy samples are sampled in the most noisy fold (fold with the highest number of noisy samples) and accumulate in the candidate pool across runs. The Monte-Carlo simulation only considers random sampling in cross-validations and does not take model training or evaluation into account. We then run ReCoV on the same noisy dataset for the same number of runs as the Monte-Carlo simulation, recording samples in the fold with the worst validation performance to generate experimental occurrence distributions. The purpose of this experiment is to evaluate the similarities between the theoretical and experimental occurrence distributions, where a high match would suggest that the fold with the highest number of noisy samples is most often the fold with the worst validation performance, thereby suggesting an association between label noise and fold-specific validation performance.

A logistic regression model is trained to perform binary classification. We compare ReCoV with two state-of-the-art algorithms, namely Confident Learning [18], and Clustering TRaining Losses (CTRL) [27] by comparing their accuracy in noise detection, as well as the accuracy of the retrained models after removing identified noisy samples.

Multi-class classification on CIFAR-10N with fastReCoV The CIFAR-10 dataset is a classic computer vision dataset that consists of 50000 training images and 10000 test images from 10 classes [14]. The CIFAR-10N dataset is a variant of CIFAR-10 with multiple sets of human-annotated real-world label noise introduced by crowd-sourcing annotations from Amazon Mechanical Turk [24]. We selected the “aggre” (9.03% noise ratio) and “worst” (40.21% noise ratio) label sets to reflect different noise level settings. This dataset is used to investigate the performance and behavior of fastReCoV in a modest-sized dataset with real-world label noise.

Experimental design: We compared two different feature extractors: a ViT-base-patch16 model pretrained on ImageNet-21k in a supervised fashion ($n_{dim} = 768$) [8], and a ViT-S/14 model pretrained on 142 million images with DINOv2 in a self-supervised fashion ($n_{dim} = 384$, stronger features) [19]. A logistic regression is trained based on the extracted features to predict the class of images.

We evaluate our method with the accuracy of the retrained model after removing identified noisy samples.

Survival prediction on HECKTOR with ReCoV and fastReCoV For noise detection in the medical domain, we first evaluate fastReCoV on the training set of HECKTOR2022 (the test set was not released), which contains 524 (390 for training and 134 for held-out test) PET-CT images of head and neck cancer patients from multiple institutes for a survival prediction challenge [1]. Survival prediction datasets are naturally noisy, especially with noise from additional sources such as loss of follow-up and disease-irrelevant mortality. In this experiment, we evaluate the ability of ReCoV and fastReCoV for detecting noisy samples in outcome prediction, a relatively unexplored field in noise detection due to the difficulty in dealing with the time-to-event labels. We evaluate our method with the concordance-index of the retrained model after removing identified noisy samples.
Experimental design: Radiomic features of the CT images are extracted using pyradiomics based on the provided individual tumor segmentations [22]. A multiple instance survival neural network is trained to predict 2-year survival outcomes of patients based on multifocal tumor features [4, 5]. 4000 runs of ReCoV is performed.

Regression on PANDA with fastReCoV The Prostate Cancer Grade Assessment (PANDA) dataset, is a challenge dataset focused on predicting ISUP grade 1-5 (or 0 for benign) from Hematoxylin & Eosin stained whole-slide images of prostate biopsies [3]. The dataset comprises 11,000 whole-slide images for training from Radboud University Medical Center and Kariolinsa Institute. Two external datasets with 844 and 4,675 images, respectively, are available for evaluation. The noise for this dataset mostly comes from different levels of annotator expertise in different centers, and the inter-observer variability for a subjective grading task.
Experimental design: For the image regression task on the PANDA dataset, we first extract patches at a regular stride, at 1 μm/px, and derived feature vectors using the pretrained CTransPath model [23]. These feature vectors were then combined using Trans-MIL [21], a multiple instance learning classification model. The model was trained using cross-entropy of the predicted probabilities and the annotated grade. We compare fastReCoV with the noise detection strategy of the winning team for the PANDA challenge, and their approach is to remove samples based on the gap between the prediction of their model and the ground truth. The approaches are evaluated using the quadratic weighted kappa metric of the retrained models after removing identified noisy samples.

4 Results

The occurrence distributions from running ReCoV match the Monte Carlo simulations strikingly well (**Fig. 2**). Our results on **Mushroom** empirically corrobo-

Table 1. Quantitative comparison of ReCoV, Confident learning (CL) and Clustering TRaining Losses for label error detection (CTRL) in three repeated trials, accuracy values are presented in the format of mean ± standard deviation. For performance of CL and CTRL, we took results reported in the CTRL paper.

Noise ratio	Method	Mask accuracy	Retrained model accuracy
10	No clean	90.0 ± 0.0	94.7 ± 0.8
	CL	97.1 ± 0.3	99.3 ± 0.3
	CTRL	99.9 ± 0.1	99.9 ± 0.1
	ReCoV	**100 ± 0.0**	**100 ± 0.0**
20	No clean	80 ± 0.0	86.1 ± 1.3
	CL	89.2 ± 0.7	95.5 ± 0.7
	CTRL	98.9 ± 0.2	99.2 ± 0.2
	ReCoV	**100 ± 0.0**	**100 ± 0.0**

Table 2. Performance of label noise detection algorithms on CIFAR-10N. Accuracy of retrained models on the test set are reported.

Features	Clean	Noisy	fastReCoV	Naive	Random
CIFAR-10N-aggre, Noise Ratio=9.03%					
ImageNet	94.54	92.97	**94.84**	94.33	92.51
DINOv2	96.46	95.23	**96.63**	96.19	95.07
CIFAR-10N-worst, Noise Ratio=40.21%					
ImageNet	94.54	86.69	**89.58**	88.99	86.46
DINOv2	96.46	90.26	**92.25**	90.58	90.02

rate the association between cross-validation fold-specific validation performance with the prevalence of label noise in the folds. In fact, as the distributions are highly similar, it becomes straightforward to calculate the number of runs required to achieving any given percentage of overlap between noisy and clean sample distributions. In (**Fig. 2**) we show results for 4.5% (2σ), 0.3% (3σ) and ~0% overlap of noisy and clean sample distributions. This result suggests that with a large number of runs ReCoV can accurately separate clean and noisy samples without the need for finding a separation threshold. When compared to existing noise detection methods, ReCoV outperforms Confident learning (CL) [18] and Clustering TRaining Losses [27] for label error detection (CTRL) by achieving a perfect separation between clean and noisy samples and perfect retrained model accuracy in both 10% and 20% noise ratio ($p<0.001$, DeLong's test, **Table 1**).

In **CIFAR-10N**, fastReCoV consistently achieves the best noise detection performance with both sets of features and both noise ratios (**Table 2**). Two approaches are compared with fastReCoV. "Naive" stands for the process of identifying noisy cases according to the disagreement between model predictions and ground truth. Though it is a simple method, Naive is easy to implement, and relatively generalizable. Naive is often the first choice when tackling complex medical data and was the approach adopted by the top performing team for the PANDA challenge [3]. Random noise detection, which means randomly removing samples from the training set, serves the purpose of an ablation test. FastReCoV

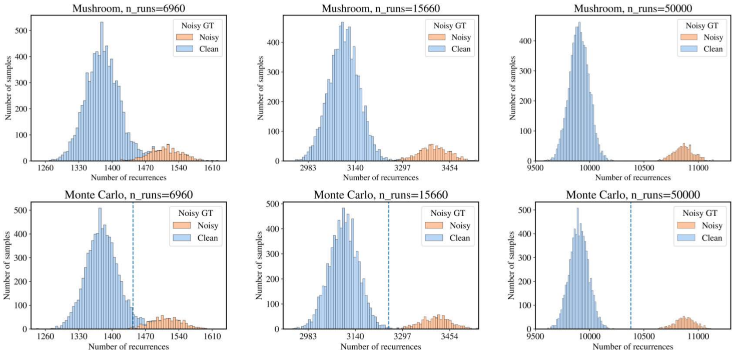

Fig. 2. Results on the Mushroom dataset (N=8124, noise ratio=10%) matches Monte Carlo simulations. Number of runs are selected to show 4.5%, 0.3% and ∼0 overlap of noisy and clean sample distribution. Dashed lines in the Monte Carlo plots refer to the theoretical separation thresholds.

detected the real-world annotation errors with sensitivity of 93% and specificity of 99% for the "aggre" noise level and sensitivity of 92% and specificity of 98% for the "worst" noise level.

The experiments in CIFAR-10N highlights the robustness of fastReCoV for detecting real-world annotation noise with different noise levels and models of different strengths. It's worth noting that the retrained models with fastReCoV achieved better performance compared to training on clean data in the "aggre" noise setting. Since the feature extractors were frozen during training and inference, this result suggests that fastReCoV removed some hidden noisy samples or confusing samples in the training set and improved model generalizability.

To the best of our knowledge, we have proposed the first explicit label noise detection algorithm for survival analysis. In **HECTOR**, ReCoV and fastReCoV improved the concordance-index of the multiple instance survival prediction model from 0.550 to 0.635 and 0.624, respectively (**Table 3**), which are considerable improvements in survival analysis without modifications to model structure and learning procedures. This task illustrates the trade-off between ReCoV and fastReCov: ReCoV required much higher runtime, but produced superior performance in the end.

In **PANDA**, fastReCoV again achieved the best retrained model QWK score in both held-out test sets, while Naive failed to improve model performance in test set 2 (**Table 4**). A few grade 5 and benign samples with the lowest weights (*i.e.* most noisy) are displayed in **Fig. S2** for further examinations.

FastReCoV (all required runs) took 3.5mins for CIFAR10N, 3.5hrs for HECKTOR and 40.6hrs for PANDA on one Nvidia Titan Xp GPU. The algorithms are fold-independent and can be further accelerated by parallelizing multiple GPUs.

Table 3. Performance of label noise detection algorithms on HECKTOR. Concordance-index and standard deviation of retrained models in 100 repeats are reported.

HECKTOR2022	**Baseline**	**ReCoV**	**fastReCoV**	**Random**
Held-out test (n=134)	0.550 ± 0.031	**0.635 ± 0.030**	0.624 ± 0.023	0.560 ± 0.045

Table 4. Performance of label noise detection algorithms on PANDA. Quadratic Weighted Kappa of retrained models are reported.

	Baseline	**Naive**	**fastReCoV**	**Random**
Test set 1 (n=844)	0.886	0.898	**0.908**	0.880
Test set 2 (n=4675)	0.866	0.836	**0.874**	0.845

5 Conclusions and Discussion

In conclusion, we discovered that the fluctuation of validation performance across cross-validation folds contains information on the distribution of noisy labels. Through extensive experiments, we show that this information can be leveraged to detect label noise in repeated cross-validations. ReCoV and fastReCoV are powerful and plug-and-play label noise detection algorithms that can be applied to a variety of machine learning models in a variety of supervised learning tasks. In our experiments we used pretrained models as feature extractors instead of training end-to-end deep models. With the advances in medical foundation models [2, 6], we believe ReCoV-based approaches can be a practical choice.

Acknowledgments. The authors would like to thank The Natural Sciences and Engineering Research Council of Canada (NSERC) for funding.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andrearczyk, V., Oreiller, V., Boughdad, S., Rest, C.C.L., Elhalawani, H., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., Hatt, M., et al.: Overview of the hecktor challenge at miccai 2021: automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In: Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings, pp. 1–37. Springer (2022)
2. Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al.: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering **7**(6), 756–779 (2023)
3. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)

4. Chen, J., Cheung, H.M., Milot, L., Martel, A.L.: Aminn: Autoencoder-based multiple instance neural network improves outcome prediction in multifocal liver metastases. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 752–761. Springer (2021)
5. Chen, J., Martel, A.L.: Metastatic cancer outcome prediction with injective multiple instance pooling. arXiv preprint arXiv:2203.04964 (2022)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
7. David, H.A., Nagaraja, H.N.: Order statistics. John Wiley & Sons (2004)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Frénay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE transactions on neural networks and learning systems **25**(5), 845–869 (2013)
10. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer. In: International conference on learning representations (2016)
11. Harkness, W.L.: Properties of the extended hypergeometric distribution. The Annals of Mathematical Statistics **36**(3), 938–945 (1965)
12. Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z.: Improving medical images classification with label noise using dual-uncertainty estimation. IEEE transactions on medical imaging **41**(6), 1533–1546 (2022)
13. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical image analysis **65**, 101759 (2020)
14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
15. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. Advances in neural information processing systems **7** (1994)
16. Matic, N., Guyon, I., Bottou, L., Denker, J., Vapnik, V.: Computer aided cleaning of large databases for character recognition. In: 11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems. vol. 1, pp. 330–331. IEEE Computer Society (1992)
17. Matuszewski, D.J., Sintorn, I.M.: Minimal annotation training for segmentation of microscopy images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 387–390. IEEE (2018)
18. Northcutt, C., Jiang, L., Chuang, I.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Reinke, A., Tizabi, M.D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A.E., Rädsch, T., Sudre, C.H., Acion, L., Antonelli, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nature methods pp. 1–13 (2024)
21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)

22. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77**(21), e104–e107 (2017)
23. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis (2022)
24. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: A study using real-world human annotations. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=TBWA6PLJZQm
25. Xiao, R., Dong, Y., Wang, H., Feng, L., Wu, R., Chen, G., Zhao, J.: Promix: Combating label noise via maximizing clean sample utility. In: Elkind, E. (ed.) Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. pp. 4442–4450. International Joint Conferences on Artificial Intelligence Organization (8 2023). https://doi.org/10.24963/ijcai.2023/494, https://doi.org/10.24963/ijcai.2023/494, main Track
26. Xu, Z., Lu, D., Wang, Y., Luo, J., Jayender, J., Ma, K., Zheng, Y., Li, X.: Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 3–13. Springer (2021)
27. Yue, C., Jha, N.K.: Ctrl: Clustering training losses for label error detection. arXiv preprint arXiv:2208.08464 (2022)