Active label cleaning: Improving dataset quality under resource constraints

Mélanie Bernhardt[1,*], Daniel C. Castro[1,*], Ryutaro Tanno[1], Anton Schwaighofer[1], Kerem C. Tezcan[1], Miguel Monteiro[1], Shruthi Bannur[1], Matthew Lungren[2], Aditya Nori[1], Ben Glocker[1], Javier Alvarez-Valle[1], and Ozan Oktay[1,†]

[1] *Health Intelligence, Microsoft Research Cambridge, Cambridge, CB1 2FB, UK*
[2] *Department of Radiology, Stanford University, Palo Alto, CA 94304, USA*

[*] *These authors contributed equally to this work.*
[†] *Corresponding author: ozan.oktay@microsoft.com*

Imperfections in data annotation, known as label noise, are detrimental to the training of machine learning models and have an often-overlooked confounding effect on the assessment of model performance. Nevertheless, employing experts to remove label noise by fully re-annotating large datasets is infeasible in resource-constrained settings, such as healthcare. This work advocates for a data-driven approach to prioritising samples for re-annotation—which we term "active label cleaning". We propose to rank instances according to estimated label correctness and labelling difficulty of each sample, and introduce a simulation framework to evaluate relabelling efficacy. Our experiments on natural images and on a new medical imaging benchmark show that cleaning noisy labels mitigates their negative impact on model training, evaluation, and selection. Crucially, the proposed active label cleaning enables correcting labels up to 4× more effectively than typical random selection in realistic conditions, making better use of experts' valuable time for improving dataset quality.

Introduction

The success of supervised machine learning primarily relies on the availability of large datasets with high-quality annotations. However, in practice, labelling processes are prone to errors, almost inevitably leading to noisy datasets—as seen in ML benchmark datasets[1]. Labelling errors can occur due to automated label extraction[2,3], ambiguities in input and output spaces[4], or human errors[5] (e.g. lack of expertise). At training time, incorrect labels hamper the generalisation of predictive models, as labelling errors may be memorised by the model resulting in undesired biases[6,7]. At test time, mislabelled data can have detrimental effects on the validity of model evaluation, potentially leading to incorrect model selection for deployment as the true performance may not be faithfully reflected on noisy data. Label cleaning is therefore crucial to improve both model training and evaluation.

Relabelling a dataset involves a laborious manual reviewing process and in many cases the identification of individual labelling errors can be challenging. It is typically not feasible to review every sample in large datasets. Consider for example the NIH ChestXray dataset[3], containing 112k chest radiographs depicting various diseases. Diagnostic labels were extracted from the radiology reports via an error-prone automated process[8]. Later, a subset of images (4.5k) from this dataset were manually selected and their labels were reviewed by expert radiologists in an effort driven by Google Health[2]. Similarly, 30k randomly selected images from the same dataset were relabelled for the RSNA Kaggle challenge[9]. Such relabelling initiatives are extremely resource-intensive, particularly in the absence of a data-driven prioritisation strategy to help focusing on the subset of the data that most likely contains errors.

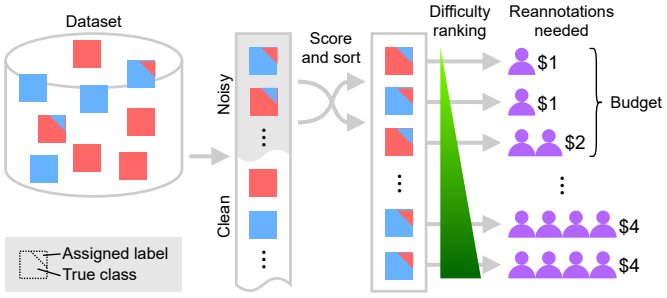


Fig. 1 Overview of the proposed active label cleaning. A dataset with noisy labels is sorted to prioritise clearly mislabelled samples, maximising the number of corrected samples given a fixed relabelling budget.

Due to the practical constraints on the total number of re-annotations, samples often need to be prioritised to maximise the benefits of relabelling efforts (see Fig. 1), as the difficulty of reviewing labelling errors can vary across samples. Some cases are easy to assess and correct, others may be inherently ambiguous even for expert annotators (Fig. 2). For such difficult cases, several annotations (i.e. expert opinions) may be needed to form a ground-truth consensus[2,10], which comes with increasing relabelling "cost". Hence, there is a need for relabelling strategies that consider both resource constraints and individual

sample difficulty—especially in healthcare, where availability of experts is limited and variability of annotations is typically high due to the difficulty of the tasks[11].

While there are learning approaches designed specifically to handle label noise during training, we consider these strategies to be insufficient for two reasons. First and most importantly, in situations with noisy evaluation data, there is no reliable method to determine whether robust learning is effective, as the performance metrics are directly compromised by label noise. Second, noise-robust learning attempts to train a maximally accurate model in the presence of incorrect labels, often by disregarding samples that could otherwise be highly informative, some of which may even be correctly labelled. In contrast, label cleaning aims to improve the quality and usefulness of a dataset to preserve as many cases as possible by fixing incorrect labels. This is imperative in safety-critical domains such as healthcare, as model robustness must be validated on clean labels.

Prioritising samples for labelling also underpins the paradigm of active learning, whose goal is to select unlabelled samples that would be most beneficial for training in order to improve the performance of a predictive model on a downstream task. The key difference here for the proposed approach is that our goal is not only to improve model performance but also to maximise the quality of labels given limited resources, which makes it valuable for both training and evaluation of predictive models. In more detail, we demonstrate how active learning and noise-robust learning (NRL) can play complementary roles in coping with label noise.

In this work, we begin by defining the active label cleaning setting in precise terms, along with the proposed relabelling priority score. Using datasets of natural images and of chest radiographs, we then demonstrate experimentally the negative impacts of label noise on training and evaluating predictive models, and how cleaning the noisy labels can mitigate those effects. Third, we show via simulations that the proposed active label cleaning framework can effectively prioritise samples to re-annotate under resource constraints, with substantial savings over naive random selection. Fourth, we analyse how robust-learning[12] and self-supervision[13] techniques can further improve label cleaning performance. Lastly, we validate our choice of scoring function, which accounts for sample difficulty and noise level, comparing with an active learning baseline.

Results

Active label cleaning. In this work, we introduce a sequential label cleaning procedure that maximises the number of corrected samples under a total resource budget $B \in \mathbb{N}$:

$$\max \underbrace{\frac{1}{N}\sum_{i=1}^{N} 1[\hat{y}_i = y_i]}_{\text{correctness of majority labels}} \quad \text{s.t.} \underbrace{\sum_{i=1}^{N} \|\hat{\ell}_i\|_1 \leq B}_{\text{budget constraint}}, \qquad (1)$$

where we assume access to a dataset $\mathcal{D} = \{(\mathbf{x}_i, \hat{\ell}_i)\}_{i=1}^{N}$, with the i^{th} image $\mathbf{x}_i$, label counts vector $\hat{\ell}_i \in \mathbb{N}^C$ with C classes, and corresponding majority label $\hat{y}_i = \arg\max_{c \in \{1,\ldots,C\}} \hat{\ell}_i^c$.

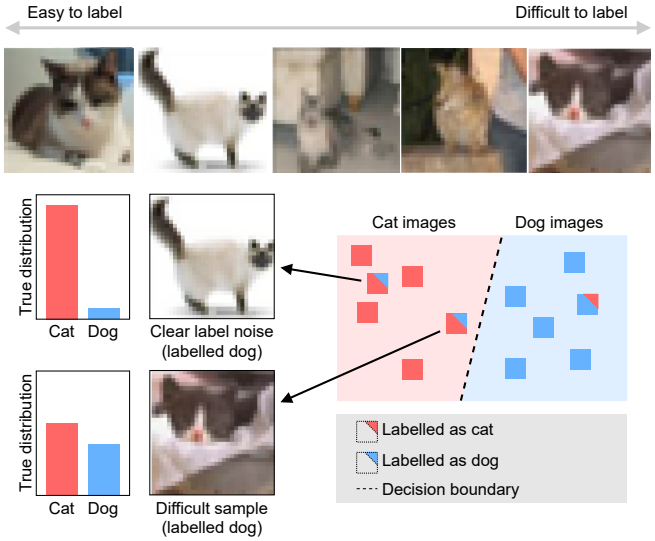


Fig. 2 Image labelling can become difficult due to ambiguity in input space[14]. Top row shows the spectrum of ambiguity for cat images sampled from CIFAR10H dataset. The 2D plot illustrates different types of mislabelled samples: clear noise and difficult cases. We expect the former to be adjacent to semantically similar samples with a different label, and the latter to be closer to the optimal decision boundary.

All samples are assumed to initially contain at least one label ($\|\hat{\ell}_i\|_1 = \sum_{c=1}^{C} \hat{\ell}_i^c \geq 1$), and some instances are mislabelled, i.e. their initial (majority) label $\hat{y}$ deviates from the true class y. Unlike the traditional active learning objective, we aim to achieve a clean set of labels that can later be used not only for model training but also for benchmarking purposes. In that regard, our setting differs from active learning by allowing for re-annotation of already labelled instances as in 're-active' learning[15]. This is an important consideration as there is often a trade-off between label quality and downstream model performance[15,16]. Yet, conventional approaches[17,18] may not be practically feasible with modern deep learning, as they often entail repeated retraining of models to measure the impact of each sample.

The proposed framework (see Table 1) determines relabelling priority based on predicted posteriors from a trained classification model, $p_\theta(\hat{y}|\mathbf{x})$, parametrised by θ. Label cleaning is performed over multiple iterations and at each iteration either a single or a batch of samples are relabelled. Within an iteration, samples are first ranked according to predicted label correctness and ambiguity (i.e. annotation difficulty). Then, each prioritised sample is reviewed sequentially by different annotators until a majority is formed among all the collected labels $\hat{\ell}_i$. In the next iteration, the remaining samples are re-prioritised and the process repeats until the relabelling budget (B) is exhausted. Assuming that each annotation has a fixed cost, the more annotations a sample requires until a majority is reached, the more expensive its relabelling. Hence, to achieve the objective in Eq. (1), clearly mislabelled samples need to be prioritised over difficult cases, and correctly labelled samples should have the lowest relabelling priority.

To this end, we propose to rank available samples by the

following scoring function Φ:

$$\Phi(\mathbf{x}, \hat{\ell}; \theta) = \underbrace{CE(\hat{\ell}, p_\theta)}_{\text{noisiness} \uparrow} - \underbrace{H(p_\theta)}_{\text{ambiguity} \downarrow} . \quad (2)$$

The first term, defined as the cross-entropy from the normalised label counts to the predicted posteriors,

$$CE(\hat{\ell}, p_\theta) = -\mathbb{E}_{\hat{\ell}/\|\hat{\ell}\|_1}[\log p_\theta(\hat{y}|\mathbf{x})] , \quad (3)$$

corresponds to the estimated noisiness (i.e. negative log-likelihood) of the given labels. On the other hand, obtaining a majority label vote requires different numbers of re-annotations for different images, depending on their difficulty. This is quantified by the entropy term $H(p_\theta)$ defined over posteriors,

$$H(p_\theta) = -\mathbb{E}_{p_\theta(\hat{y}|\mathbf{x})}[\log p_\theta(\hat{y}|\mathbf{x})] , \quad (4)$$

which penalises ambiguous cases in the ranking. Similar objectives have been used in the contexts of semi-supervised learning[19] and entropy regularisation[20] to penalise over-confident posterior predictions from models. In Methods section, we present different options for the predictive model p_θ to be used in computing these quantities.

Datasets used in the experiments. To analyse label noise scenarios, we experiment with two imaging datasets, namely CIFAR10H and NoisyCXR, containing multiple annotations per data point, which helps us to model the true label distributions and associated labelling cost. In CIFAR10H[5,14], on average 51 manual annotations were collected for each image in the test set of CIFAR10[21], a collection of 10k natural images grouped in 10 classes. The experiments on CIFAR10H are intended to understand the challenges associated with different noise rates, noise models, and active relabelling scenarios. We also set up a new benchmark for label cleaning on medical images, NoisyCXR, comprising 26.6k chest radiographs and multiple labels from clinical datasets indicating the presence of pneumonia-like opacities[3,9]. This serves as a more challenging imaging benchmark, where modelling difficulties are coupled with the challenges associated with noisy labels and sample prioritisation.

Label noise undermines both model training and evaluation. In particular, training with noisy labels is known to impair the performance of predictive models[6,7], as the latter may be forced to memorise wrong labels instead of learning generalisable patterns. As an example, a vanilla convolutional neural network (CNN) classifier trained on CIFAR10H with clean labels achieves 73.6% accuracy on a clean test set of 50k images. However, if the same model is trained on a version of this dataset with 30% label noise, its test accuracy degrades to 64.1%, a substantial drop of 8.5%. Even a noise-robust classifier, initialised with self-supervision (i.e. pre-trained with images only), drops from 80.7% to 78.7% in accuracy when fine-tuned on the same noisy training labels as opposed to clean labels.

We also highlight the adverse implications of validation label noise on model evaluation. Its most evident impact is that true model performance can be underestimated. For instance, consider CIFAR10H with 40% noise rate in both training and validation labels (instance-dependent noise[22]; see Methods). A self-supervised classifier, as above, is estimated to be only 46.5% accurate on these noisy validation labels, compared to 65.8% on the true, clean labels. Additionally, model rankings based on noisy validation metrics may be unreliable, leading to misinformed model selection or design decisions. For example, a second classifier, trained on the same training set via co-teaching[12], achieves a lower 45.9% accuracy on the same noisy validation set, while its real accuracy (unobservable in practice) on clean labels is 69.9%. In other words, the worst performing model (here, self-supervised) would have been chosen for deployment because of misleading validation metrics obtained on noisy labels.

In summary, label noise can negatively affect not only model building, but also validation. The latter is especially relevant in high-risk applications, e.g. for the regulation of models in healthcare settings. Our results in later sections demonstrate how active cleaning of noisy training and evaluation labels can help mitigate such issues.

Table 1 Active label cleaning

Given: $Y = \{\ell_i\}_{i=1}^N$: True label distributions
Input: $\mathcal{D} = \{(\mathbf{x}_i, \hat{\ell}_i)\}_{i=1}^N$: Dataset with noisy labels
$B \in \mathbb{N}$: Relabelling budget
$b \in \mathbb{N}$: Update frequency

1: $\theta \leftarrow$ TRAINROBUSTMODEL($\mathcal{D}$)
2: $\mathcal{I}_{\text{avail}} \leftarrow \{1, \ldots, N\}, \quad \mathcal{I}_{\text{cleaned}} \leftarrow \emptyset$
3: count $\leftarrow 0$
4: **while** count $< B$ **do** ▷ If budget remains
5: $\quad j \leftarrow \arg\max_{i \in \mathcal{I}_{\text{avail}}} \Phi(\mathbf{x}_i, \hat{\ell}_i; \theta)$ ▷ Rank (Eq. (2))
6: $\quad$ **repeat**
7: $\quad\quad \hat{\ell}_j \leftarrow \hat{\ell}_j +$ SAMPLE(ℓ_j) ▷ Acquire one-hot label
8: $\quad\quad$ count $\leftarrow$ count + 1
9: $\quad$ **until** majority formed in $\hat{\ell}_j$
10: $\quad \mathcal{I}_{\text{avail}} \leftarrow \mathcal{I}_{\text{avail}} \setminus \{j\}, \quad \mathcal{I}_{\text{cleaned}} \leftarrow \mathcal{I}_{\text{cleaned}} \cup \{j\}$
11: $\quad \mathcal{D} \leftarrow \{(\mathbf{x}_i, \hat{\ell}_i) : i \in \mathcal{I}_{\text{avail}} \cup \mathcal{I}_{\text{cleaned}}\}$
12: $\quad$ **if** count divisible by b **then**
13: $\quad\quad \theta \leftarrow$ UPDATE($\theta, \mathcal{D}$) ▷ Fine-tune model
14: $\quad$ **end if**
15: **end while**
16: **return** $\mathcal{D}$

Simulation of sequential relabelling. A traditional way to evaluate a label cleaning algorithm is via its ability to separate noisy from clean labels (e.g. detection rate)[17,18]. However, such an evaluation does not necessarily take into account the sequential nature of sample selection, label acquisition, and model updates. Further, it typically assumes all new labels are correct, neglecting the effect of sample ambiguity on how many annotations are required. Both are crucial factors in measuring the effectiveness of label cleaning efforts in terms of resource usage and final outcome.

To account for these factors, we propose a realistic simulation of the entire relabelling process that leverages the true label distribution of each sample. As shown in Table 1, the simulation proceeds sequentially. Once a sample has been selected for relabelling, new labels are drawn from its true distribution (Table 1, Line 7), which reflects the probabilities of real annotators as-

Table 2 Classification accuracy (%) before and after label cleaning.

	Selector	Scoring	Classifier	Before cleaning	After cleaning
(1)	Vanilla	Eq. (2)	Vanilla	64.1	68.4
(2)	Vanilla	BALD[23]	Vanilla	64.1	68.3
(3)	SSL	Eq. (2)	Vanilla	64.1	70.9
(4)	SSL	Eq. (2)	SSL	78.7	80.3
(5)	Co-teaching	Eq. (2)	Co-teaching	66.5	68.8
(6)	–	–	ELR[24]	67.0	–
(7)	(Clean training)		Vanilla	73.6	–
(8)	(Clean training)		SSL	80.7	–

Models are evaluated on a clean test set (N = 50k) before and after relabelling 32.7% of samples in the training set (CIFAR10H, N = 5k, η = 30%).

signing each class to a given sample. This enables more faithful modelling of the expected relabelling effort for each sample and of the likely label errors. Through this simulation, the performance of sample selection algorithms can be measured in terms of percentage of corrected labels in the dataset as a function of the number of re-annotations.

The relabelling simulation is used as a testbed to evaluate three different sample ranking algorithms: standard CNN (vanilla), co-teaching[12], and self-supervised pre-training with fine-tuning[13]. All approaches employed the same image encoder (ResNet-50), augmentations, and optimiser type (see supplement for implementation details). At each iteration, new labels are sampled from the true label distribution, and experiments are repeated with 5 independent random seeds. Performance of selection algorithms is compared to two baselines: the oracle and random selector, which determine respectively the upper and lower bounds. The random selector chooses the next sample to be annotated uniformly at random as done in previous studies[2,9], while the oracle simulates the ideal ranking method by having access to true label distribution. Knowledge of the true distribution enables the oracle to prioritise least difficult noisy samples first, hence maximising the number of corrected samples for a given resource constraint. The maximum relabelling budget (B) in the simulation is set to the expected number of re-annotations required by the oracle to clean all mislabelled samples.

Cleaning noisy training labels improves predictive accuracy. All label cleaning models are first trained on their respective training datasets (CIFAR10H and NoisyCXR), and subsequently used to rank samples within the same set for relabelling. In the rest of the manuscript, the term "selector" refers to the models used for label cleaning. Table 2 shows that all training methods benefit from newly acquired labels regardless of their underlying weight initialisation (e.g. self-supervised learning – SSL) and noise-robust learning (NRL) strategy, as in ELR[24]. In that regard, label cleaning serves as a complementary approach to NRL by recycling data samples containing noisy labels. Additionally, the comparison between rows 3 and 6 show that acquiring a new set of labels can yield significantly better results than state-of-the-art NRL models trained on noisy labels, which performs closer to the upper-bound (row 7) where training is done with all true labels. Class imbalance, inherent noise model, and difficulty of samples are expected to introduce challenges to most NRL methods.

Label cleaning can be done in a resource-effective manner. In Table 2, we also observe that sample ranking methods (selectors) can influence the outcome of label cleaning efforts in terms of predictive accuracy (see rows 1–3). To further investigate the impact of selector types, we carried out label cleaning simulations (see Fig. 3) on both CIFAR10H and NoisyCXR datasets for initial noise rates of $\eta = 15\%$ and 12.7%, respectively. The methods are compared quantitatively regarding fraction of corrected labels (i.e. reduction in noise rate) for varying number of re-annotations. Additionally, the sample ranking is visually illustrated in Fig. 4 along with the predicted label correctness and ambiguity scores.

It is apparent that, for a fixed improvement in number of corrected labels, all proposed selectors require much fewer re-annotations than random. For instance, the vanilla selector can reach 90% correct labels with 3× (NoisyCXR) or 2.5× (CIFAR10H) fewer re-annotations than random. More specifically, we can see that noise-robust learning (here, co-teaching) benefits selection algorithms in prioritising noisy labels. Additionally, SSL yields further performance improvements by learning generic semantic representations without requiring labels, which in principle reduces chances of memorising noise. In that regard, it is a complementary feature to traditional noise-robust learning approaches as long as there is enough data for training. In particular, SSL pre-training yields improved noisy sample detection on NoisyCXR dataset in comparison to initialisation from scratch or pre-trained ImageNet weights. Figure 5 exemplifies clear noisy and difficult cases for NoisyCXR.

The experiments are also repeated for larger noise rates (see Supplementary Figure 2) to identify the limits beyond which sample selection algorithms deteriorate towards random sampling. The results show similar performance as long as diagonal dominance[25] holds on average in class confusion matrices (see Fig. 6b), which is verified by experimenting with uniform and class-dependent noise models[26]. Moreover, graph-based selectors outperform the ones with linear head on higher noise rates, experimentally suggesting that transductive approaches may be more resilient against noise in these regimes.

Sample scoring function impacts cleaning performance. We study the influence of the self-entropy term ($H(p_\theta)$) used in Eq. (2) in an ablation study. We see that selectors prioritising clear label noise cases yield higher numbers of corrected label noise, and this effect is further pronounced by including the entropy term in the scoring function (see Supplementary Figure 3).

Additionally, ranking samples for labelling is related to active learning, whereby one annotates unlabelled instances in order to maximise model performance metrics. Typical active-learning methods attempt to identify the most informative examples[27], relying on e.g. core-sets[28] or high predictive uncertainty[29]. A representative example is Bayesian active learning by disagreement (BALD)[23,29], which quantifies the mutual information between each sample's unknown label and the model parameters.

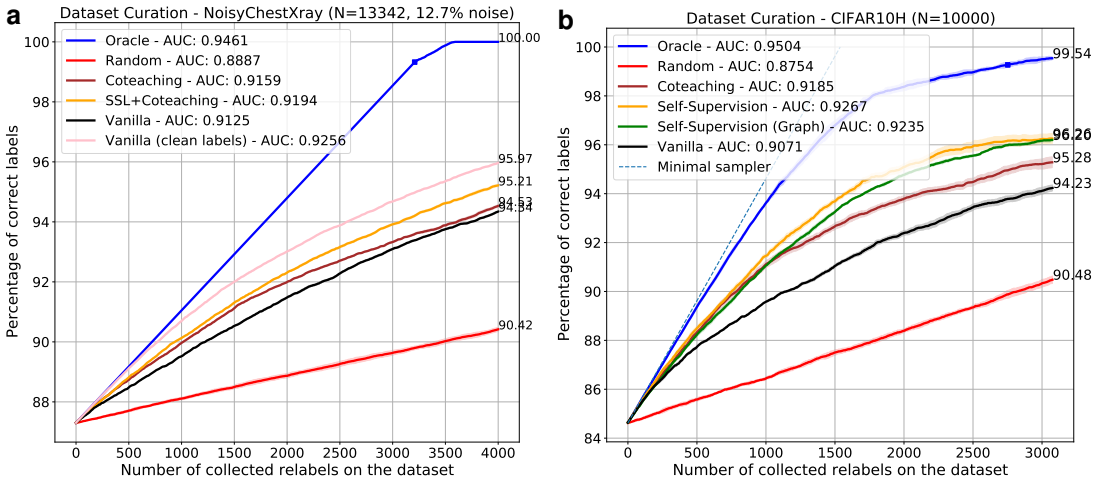


Fig. 3 Results of the label cleaning simulation on training datasets. a NoisyCXR (η = 12.7%); **b** CIFAR10H (η = 15%). For a given number of collected labels (x-axis), a cost-efficient algorithm should maximise the number of samples that are now correctly labelled (y-axis). The correctness of acquired labels is measured in terms of accuracy. The area-under-the-curve (AUC) is reported as a summary of cleaning efficiency of each selector across different relabelling budgets. The upper and lower bounds are set by oracle (blue) and random sampling (red) strategies. The pink curve (**a**) illustrates the practical upper bound of cleaning performance when the selector model is trained solely on clean labels.

However, these criteria may simply prioritise under-represented parts of the data space rather than samples with noisy labels.

To analyse the suitability for label cleaning of such an approach, we repeat these experiments with the BALD score in replacement of the function in Eq. (2), using the same relabelling budget. Rows 1–2 in Table 2 show that classification accuracy gains on the test set are comparable between the two. However, the label quality improvement is significantly inferior for BALD (Supplementary Figure 2), as it favours samples with the highest disagreement—which may not correspond to label noise. Thus, in addition to improving model accuracy, active label cleaning offers the flexibility to improve the quality of noisy datasets.

Cleaning noisy evaluation labels mitigates confounding in model selection. In noise-robust learning[12,30–32], validation data is traditionally assumed to be perfectly labelled and can be relied on for model benchmarking and hyperparameter tuning. Furthermore, classification accuracy has been shown to be robust against noise[25,33] assuming that labelling errors and inputs are conditionally independent, $P(\hat{y}|y,\mathbf{x}) = P(\hat{y}|y)$. While these assumptions may be valid in certain applications, they often do not hold for real-world datasets, wherein the labelling process clearly relies on examining the inputs. Thus, models can learn biases from the training data that would correlate with labelling errors in the evaluation set (see Fig. 6a).

To illustrate the pitfalls of evaluating models on noisy labels, we have conducted experiments on CIFAR10H using symmetric (*SYM*) and instance-dependent noise (*IDN*) models[22] (see supplement for details). In these experiments, diagonal dominance[25] is preserved for all confusion matrices, and the statistical dependence between labelling errors, input, and true label are the same in both training and validation sets. Furthermore, all the available relabelling resources are used for cleaning the noisy labels in the evaluation set instead of training data to obtain an unbiased estimate of model ranking and performance.

Table 3 Classification accuracy under noisy evaluation labels.

	Model	True	Noisy (40%)	Cleaned (10%)	Cleaned (20%)
SYM	M1	73.32	45.19 (.21)	54.37 (.13)	62.53 (.09)
	M2	**80.16**	**48.93** (.23)	**58.45** (.10)	**67.42 (.10)**
IDN	M1	**69.91**	45.93 (.10)	49.97 (.06)	**55.87** (.06)
	M2	65.76	**46.50** (.13)	49.90 (.12)	55.28 (.10)

Co-teaching (M1) and SSL (M2) models are compared on a noisy CIFAR10H validation set $\mathcal{D}_{\text{eval}}$ over three runs using different label initialisations. Both approaches use ResNet-50 with different weight initialisation and regularisation. We compare classification accuracy on true, noisy, and cleaned labels. M2 is deliberately less regularised (i.e. weight decay) and is expected to perform worse on a more challenging *IDN* noise model.

ResNet-50 models[34] with different regularisation settings are trained with noisy training labels ($|\mathcal{D}_{\text{train}}| = 10k$) and are later used to clean labels on evaluation set ($|\mathcal{D}_{\text{eval}}| = 50k$). Labels for sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{eval}}$ are determined for each experiment separately using *SYM* and *IDN* noise models with 40% noise rate on average. The experimental results (Table 3) show that standard performance metrics and model ranking strongly depend on the underlying conditions leading to labelling errors, which shows the need for clean validation sets in practical applications for model selection.

To tackle these challenges, we applied the active relabelling procedure (SSL-Linear) on this noisy validation set by using the same corresponding training data ($\mathcal{D}_{\text{train}}$). Through sample prioritisation and relabelling of 10% of the entire set ($\mathcal{D}_{\text{eval}}$), the bias in model selection can be alleviated as shown in Table 3. At the end of the label cleaning procedure, the noise rates are reduced to 31.32% (*IDN*) and 30.25% (*SYM*); in other words, 86.8% (*IDN*) and 97.5% (*SYM*) of the selected images were initially mislabelled and then corrected. Here, *IDN* model is observed to be more challenging for identifying noisy labels.

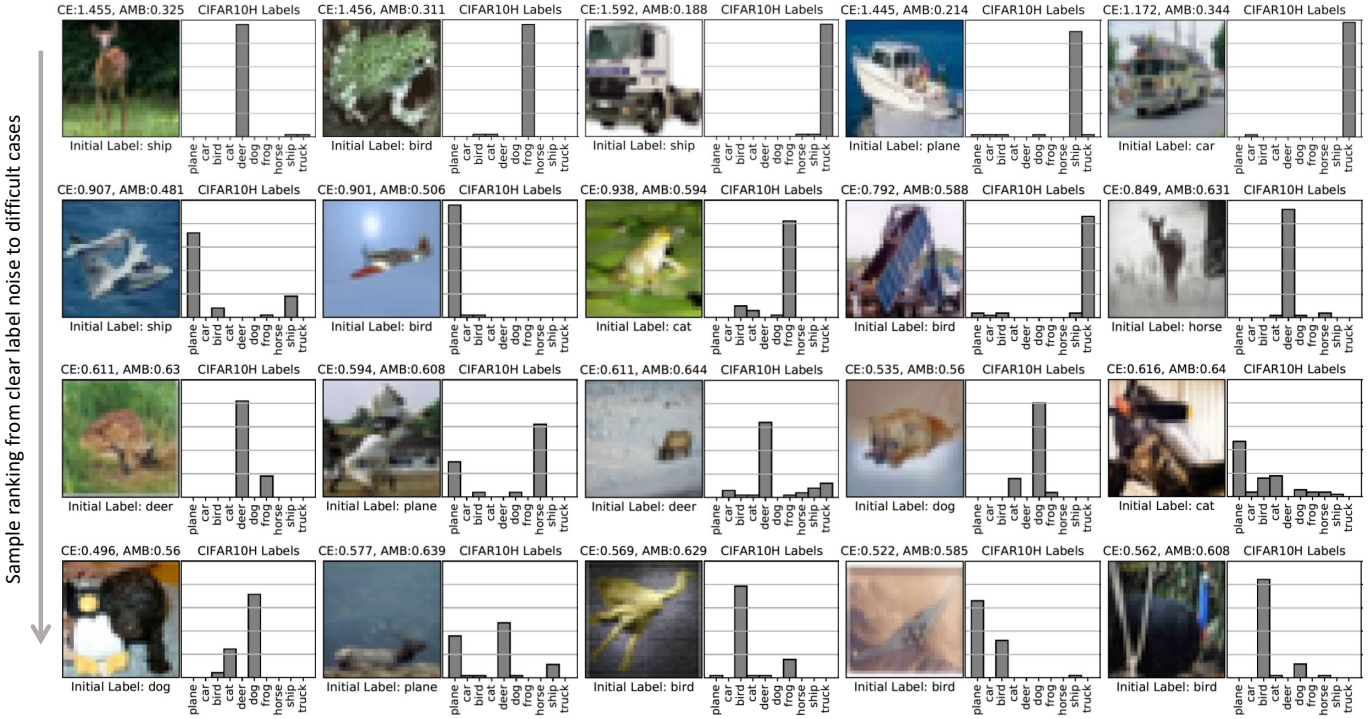


Fig. 4 Ranking of CIFAR10H samples (15% initial noise rate) by the *SSL-Linear* algorithm. The top row illustrates a representative subset of images ranked at the top-10 percentile with the highest priority for relabelling. Similarly, the second and third rows correspond to 25–50 and 50–75 percentiles, respectively. At the bottom, ambiguous examples that fall into the bottom 10% of the list ($N = 2241$) are shown. Each example is shown together with its true label distribution to highlight the associated labelling difficulty. This can be compared against the label noisiness (cross-entropy; CE) and sample ambiguity (entropy; AMB) scores predicted by the algorithm (see Eq. (2)), shown above each image. As pointed out earlier, adjudication of samples provided at the bottom does require a large number of re-annotations to form a consensus. The authors in[14] explore the causes of ambiguity observed in these samples.

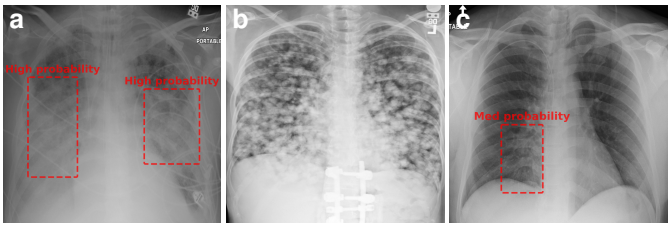


Fig. 5 Chest X-ray images selected from NoisyCXR dataset, which do not contain "pneumonia" label in the NIH dataset[3]. a Correctly identified noise case with pneumonia-like opacities shown with bounding boxes. **b** Wrongly flagged sample with a correct label; here the model confuses lung nodules with pneumonia-like opacities. **c** A difficult case with subtle abnormality where radiologists indicated medium-confidence in their diagnosis as shown by the highlighted region (RSNA study[9]).

Discussion

This work investigated the impact of label noise on model training and evaluation procedures by assessing its impact in terms of (I) predictive performance drop, (II) model evaluation results, and (III) model selection choices. As potential ways to mitigate this problem can be resource-demanding depending on the application area, we defined cost-effective relabelling strategies to improve the quality of datasets with noisy class labels. These solutions are benchmarked in a newly proposed simulation framework to quantify potential resource savings and improvement in downstream use-cases.

In particular, we highlight the importance of cleaning labels in a noisy evaluation set. We showed that neglecting this step may yield misleading performance metrics and model rankings that do not generalise to the test environment. This can, in turn, lead to overoptimistic design decisions with negative consequences in high-stakes applications such as healthcare. One of our main findings is that the patterns of label error in the data (i.e., structural assumptions about label errors shown in Fig. 6a) can have as large an impact on the efficacy of label cleaning and robust-learning methods as the average noise rates, as evidenced by the results obtained on both training and validation sets. We therefore recommend carefully considering the underlying mechanisms of label noise when attempting to compare possible solutions.

The results also suggest that even robust-learning approaches may not fully recover predictive performance under high noise rates. In such cases, SSL pre-training is experimentally shown to be a reliable alternative, outperforming noise-robust models trained from scratch, even more so with the increasing availability of unlabelled datasets. Lastly, we show that acquiring new labels can complement noise-robust learning by recycling data samples even if their labels are noisy, and can also handle biased labels. Thus, the two domains can be combined to obtain not only a better model, but also clean data labels for downstream applications.

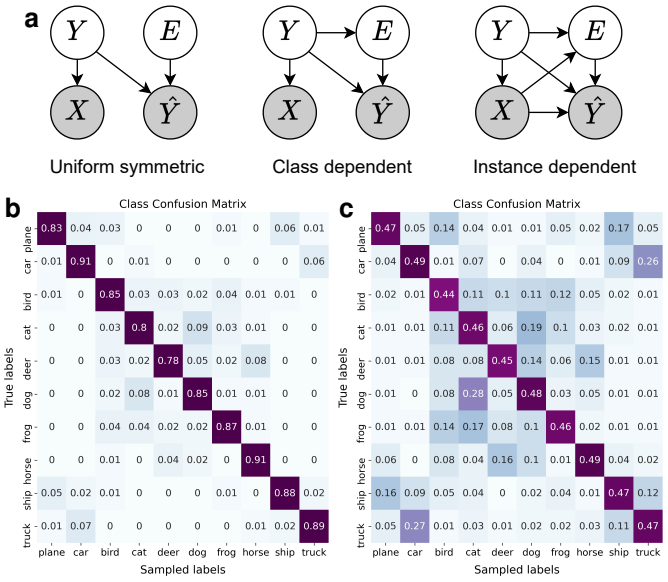


Fig. 6 Understanding label noise patterns. **a** Different label noise models used in robust learning. The statistical dependence between input image (X), true label (Y), observed label ($\hat{Y}$), and error occurrence (E) is shown with arrows (adapted from Frénay et al.[39]). **b–c** CIFAR10H class confusion matrices (temperature $\tau = 2$) for all samples (**b**) and difficult samples only (**c**).

Future work directions. Although the present study focused on imaging, the proposed methodology is not limited to this data modality, and empirical validation with other input types is left for future work. It will also be valuable to explore, for example, having the option to also annotate unlabelled samples, or actively choosing the next annotator to label a selected instance. Such extensions to active cleaning will significantly broaden its application scope, enabling more reliable deployment of machine learning systems in resource-constrained settings.

Similarly, multi-label fusion techniques[35–38] can be used within the proposed label cleaning procedure to restore true label distribution by modelling labelling process and aggregating multiple noisy annotations. Such approaches critically rely on the availability of multiple labels for each sample—which can be realised towards the end of relabelling efforts.

Methods

CIFAR10H dataset. CIFAR10H[5,14] is an image dataset comprising the test set of CIFAR10 (10k images) and multiple labels collected from crowdsourcing. Each image on average contains 51.1 labels distributed over 10 class categories. In our study, the true label distributions were calculated by normalising the histogram of these labels on a per image basis. In the experiments, we assume that our initial dataset contains labelling errors. These partially noisy labels were obtained by sampling the initial label by taking into account the label distribution for each sample. The noise rate in experiments is controlled by applying temperature scaling[40] (τ) to each distribution, which preserves the statistical dependence between input images and their corresponding labels. The higher the temperature, the closer to a uniform distribution the label distribution becomes, hence the higher the probability of sampling a noisy label for a given sample.

Samples in this dataset are not all equally difficult to label or classify[14]. For instance, for some images, all annotators agreed on the same label, for some others collected labels were more diverse as the image was harder to identify. Hence, we defined sample difficulty as a function of the normalised Shannon entropy of its target distribution, $\sum_{c=1}^{C} p_c \log_C p_c$. If the entropy associated to this distribution was higher than a certain threshold (0.3 in our experiments), the sample was classified as "difficult". The entropy of the target distribution can also be related to the expected number of relabellings required for this sample (see problem definition in Results). In Fig. 6b, we show the average class confusion matrices over the CIFAR10H dataset for $\tau = 2$. In Fig. 6c, the confusion matrix of only the difficult samples is shown to highlight the classes that are confused the most in a standard labelling process.

Table 4 Correspondence of chest radiography labels from the RSNA challenge[9], NIH[3], and NoisyCXR datasets.

	RSNA labels[9]	
	Pneumonia-like opacity	No pneumonia-like opacity
NIH labels[3]:		
Pneumonia	367	**441**
Consolidation/infiltration (not pneumonia)	*3,988*	*8,551*
Other diseases only	**1,101**	5,567
No finding	**556**	6,113
NoisyCXR labels:		
Pneumonia-like opacity	3,956	**1,296**
No pneumonia-like opacity	**2,056**	19,376
Total	6,012	20,672

The disagreements between two label sets are highlighted in bold font. Fields in italic indicate an unclear agreement between both sets of labels. NoisyCXR labels are collected from the original NIH labels where possible, and the remainder are uniformly sampled (10%) from the "Consolidation/infiltration" category to increase the noise rate further in the experiments.

NoisyCXR: a medical benchmark dataset for label cleaning. In the experiments, we used a subset of the medical image dataset released by the NIH[3]. The original dataset contains 112k chest radiographs with labels automatically extracted from medical records using a natural language processing algorithm. In the original data release, labels are grouped into 14 classes (non-mutually exclusive) indicating the presence of various diseases or no finding at all. However, there have been studies[8] showing errors in these auto-extracted image labels. Later, the RSNA released the Pneumonia Detection Challenge[9] aiming to detect lung opacities indicative of pneumonia. The challenge dataset is comprised of 30k images randomly sampled from the NIH ChestXray dataset mentioned above. To adjudicate the original labels released by the NIH, a board of radiologists reviewed and re-assessed each image on this dataset and delimited the presence of lung opacities associated to pneumonia. Samples are hence classified into two categories: "pneumonia-like lung opacity" and "no pneumonia-like lung opacity". Here we aim to utilise a realistically noisy version of this RSNA dataset encountered in real-world application scenarios. To this end, the original NIH labels are leveraged to construct a noisy version of the true labels released by the RSNA. In more detail, we analysed the data in four categories (see rows in Table 4) based on their original NIH labels and the class taxonomy[41] in order to map this original label to a binary label indicating presence or absence of pneumonia-like opacities:

- Cases where the NIH label contains "Pneumonia". All these samples were assigned to the "pneumonia-like opacity" category in NoisyCXR.
- Cases associated to opacities potentially linked to pneumonia: NIH label contains "consolidation" or "infiltration" (but not "pneumonia"). For these cases, it is unclear whether the original NIH label should be mapped to "pneumonia-like opacity" or not[41,42]. Hence, for these cases, in NoisyCXR, we attributed the correct (final RSNA) label to 90% of the cases and flipped the label of the remaining 10%.
- Cases tagged as "no finding", all assigned to the "no pneumonia-like opacity" category.
- Cases linked to other pathology (i.e., original label only contains pathology unrelated to pneumonia e.g. pleural effusion, nodules, fluid[3]), all assigned to the "no pneumonia-like opacity" category.

The obtained binary mapping hence gives us a set of noisy labels for the "pneumonia-like opacity" classification task whereas the labels released as part

of the challenge are considered as the correct labels for the data cleaning experiment. The final noisy dataset contains 26,684 images with a noise rate of 12.6% (see Table 4). As in the case of CIFAR10H, the uncertainty in the final labels permits to separate ambiguous from easy cases. In particular, using the confidence associated to each pneumonia-like lung opacity bounding box[43] we distinguished cases as follows:

- If the final label was "Pneumonia-like opacity" but all bounding boxes were of low confidence, we considered the case as difficult and define the label distribution for relabel sampling to $\mathbb{P}_{\text{label}}(\text{Pneumonia opacity}) = 0.66$. If at least one of the boxes had a medium or high confidence score, the case is considered easy and we set $\mathbb{P}_{\text{label}}(\text{Pneumonia opacity}) = 1$.
- If the final label was "No Pneumonia-like opacity" but some readers delineated opacities, the sample is considered ambiguous and it's class label is set to $\mathbb{P}_{\text{label}}(\text{Pneumonia opacity}) = 0.33$, Otherwise, it is considered easy and $\mathbb{P}_{\text{label}}(\text{Pneumonia opacity}) = 0$.

Learning noise-robust image representations. Noise-robust learning (NRL) is essential for handling noise and inconsistencies in labels which can severely impact the predictive performance of models[6,7]. Although, this type of machine learning is most commonly employed to improve model's predictive performance, in our setting, we use NRL methods to obtain an unbiased sample scoring function in active label cleaning. Interested readers should refer to the extended literature review[44] for further information on this topic.

The sample selection algorithms presented in the next section heavily rely on models trained with NRL methods; in particular, we are utilising methods from two broad categories: model regularisation[45–47] and sample exclusion approaches[12,48,49]. The former set of methods have been shown to be effective against memorising noisy labels[6], by favouring simpler decision boundaries. Sample exclusion approaches, on the other hand, aim to identify noisy samples during training by tracking loss values of individual samples[12,49] or gradient vector distributions[24,50]. Alternative meta-model based approaches are not considered in this study as unbiased clean validation set may not always be available in real-world scenarios.

Sample selection algorithms. Here we propose three selection algorithms for accurately estimating class posteriors, $p_\theta(\hat{y}|\mathbf{x})$, to prioritise mislabelled data points in the scoring function Φ (see Eq. (2)). In detail, we first train deep neural networks with noise robustness, then use them to identify the corrupted labels. The methods differ in how the robustness is introduced and how the networks are used afterwards. These methods were proposed and studied to take into account different noise cleaning setups, including the number of labelled and unlabelled samples, and prior knowledge on the expected noise rate.

As a baseline approach, networks are trained with noisy labels and augmented images by minimising a negative log-likelihood loss term, which is referred to as vanilla. This approach is expected to be biased by the noisy labels and to perform sub-optimally in prioritising samples for relabelling.

Supervised co-teaching training. To cope with noisy labels, we first propose to use co-teaching[12] in a supervised learning setting, which has shown promising performance in classification accuracy and robustness. In the co-teaching scheme, two identical networks f_{θ_1} and f_{θ_2} with different initialisation are trained simultaneously, but the batch of images at each training step for f_{θ_1} is selected by f_{θ_2} and vice versa. The rationale is that images with high loss values during early training are ones that are difficult to learn, because their labels disagree with what the network has learned from the easy cases—indicating these labels might be corrupted. Also, using two networks rather than one prevents a self-confirmation loop as in[51]. Co-teaching thus provides a method for robust training with noisy labels that can be employed to identify corrupted labels. We ensemble the predictions of both trained networks as $\frac{1}{2}\sum_m f_{\theta_m}(\mathbf{x}_i) = \hat{\boldsymbol{\pi}}_i$, where $\hat{\boldsymbol{\pi}}_i \in [0,1]^C$ and $\sum_{c=1}^{C} \hat{\pi}_i^c = 1$, and use the predicted distribution, $\hat{\boldsymbol{\pi}}_i$, in the scoring function Φ to rank the samples.

Self-supervision for robust learning. Noise-robust learning can still be influenced by label noise and exclude difficult samples, as they yield large loss values. Self-supervised learning (SSL) is proposed as a second approach to address this and further improve the sample ranking performance whenever an auxiliary task can be defined to pre-train models. As the approach does not require target labels and only extracts information from the images, it is an ideal way to learn domain-specific image encoders that are unbiased by the label noise. It can also be used in conjunction with the co-teaching algorithm in an end-to-end manner.

To this end we use BYOL[13] to learn an embedding for the images, i.e., a non-linear transformation $\mathbf{x}_i \mapsto g_\theta(\mathbf{x}_i) = \mathbf{h}_i$, where $\mathbf{h}_i \in \mathbb{R}^L$ and L is the size of the embedding space. As SSL is agnostic to the labels, the embedding will be formed based only on the image content, such that similar images will be close in the embedding space and dissimilar images will be further apart. The learnt encoder can later be used to build noise-robust models. One way to achieve this is by fixing the weights of g_θ and train a linear classifier on top of the learned embedding as $\mathbf{h}_i \mapsto f_\varphi(\mathbf{h}_i) = \hat{\boldsymbol{\pi}}_i$, which is referred to as "SSL-Linear". To obtain better calibrated posteriors, large logits are penalised at training time with function $\gamma(x) = \alpha \cdot \tanh(x/\alpha)$[13] and optimised with smoothed labels[52]. The linear classifier can only introduce a linear separation of the images according to their given labels. Assuming images with similar contents (i.e. nearby in embedding space) should have similar labels, the simple linear decision boundary introduces robustness to mislabels.

Self-supervision with graph diffusion. As an alternative to the linear classifier, we employ a graph classifier to embody the idea that the label for each sample should be consistent with its neighbourhood. We construct a graph based on the SSL embeddings $\mathbf{h}_i$ (see above), as g_θ already optimises for their cosine similarities during training. We thus define the graph's affinity matrix $\mathbf{W}$ as $W_{ij} = \frac{1}{2}(1 + \mathbf{h}_i^\mathsf{T}\mathbf{h}_j / \|\mathbf{h}_i\|\|\mathbf{h}_j\|)$ for $i \neq j$ and $W_{ii} = 0$. We additionally set $W_{ij} = 0$ for all but the K nearest neighbours j of each node i. We then follow the regularised label spreading approach[53], which minimises

$$Q(\mathbf{F}) = \frac{1}{2}\Big(\underbrace{\operatorname{tr}[\mathbf{F}^\mathsf{T}(\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}})\mathbf{F}]}_{\sum_{i,j} W_{ij}\left\|\mathbf{f}_i/D_{ii}^{1/2} - \mathbf{f}_j/D_{jj}^{1/2}\right\|^2} + \mu\|\mathbf{F} - \mathbf{Y}\|_F^2\Big), \tag{5}$$

where $\mathbf{f}_i \in [0,1]^C$ are the soft labels of each node (i^{th} row of $\mathbf{F}$), $\mathbf{D}$ is the degree matrix, $\mathbf{Y}$ are the given labels of all nodes, $\mathbf{I}$ is the identity matrix, and $\mu > 0$ is a regularisation coefficient. The first term in Eq. (5) measures label consistency between neighbours, while the second is the discrepancy between predicted and given labels. The optimal spread labels are given in closed form as $\mathbf{F}^* = \frac{\mu}{1+\mu}(\mathbf{I} - \frac{1}{1+\mu}\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}})^{-1}\mathbf{Y}$.

Data availability

The CIFAR10H dataset is available at https://github.com/jcpeterson/cifar-10h. The raw data for NoisyCXR can be downloaded from https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data, and the labels used in our experiments are processed in the linked code repository.

Code availability

All the code for our label cleaning benchmarks, robust learning models, and experiments will be released in an open-source repository upon publication.

References

1. Northcutt, C. G., Athalye, A. & Lin, J. Pervasive label errors in ML benchmark test sets, consequences, and benefits. In *NeurIPS 2020 Workshop on Security and Data Curation Workshop* (2020).
2. Majkowska, A. *et al.* Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* **294**, 421–431 (2020).
3. Wang, X. *et al.* ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471 (2017).
4. Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X. & van den Oord, A. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159* (2020).
5. Peterson, J. C., Battleday, R. M., Griffiths, T. L. & Russakovsky, O. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, 9617–9626 (2019).
6. Arpit, D. *et al.* A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *PMLR*, 233–242 (PMLR, 2017). arXiv:1706.05394.
7. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations* (2017). URL https://openreview.net/forum?id=Sy8gdB9xx.

8. Oakden-Rayner, L. Exploring large-scale public medical image datasets. *Academic Radiology* **27**, 106 – 112 (2020). URL http://www.sciencedirect.com/science/article/pii/S107663321930488X. Special Issue: Artificial Intelligence.
9. RSNA pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge (2018).
10. Krause, J. *et al.* Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).
11. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
12. Han, B. *et al.* Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 8527–8537 (2018).
13. Grill, J.-B. *et al.* Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, vol. 33, 21271–21284 (2020).
14. Battleday, R. M., Peterson, J. C. & Griffiths, T. L. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications* **11**, 1–14 (2020).
15. Lin, C. H., Mausam & Weld, D. S. Re-active learning: Active learning with relabeling. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 1845–1852 (AAAI Press, 2016).
16. Ipeirotis, P. G., Provost, F., Sheng, V. S. & Wang, J. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* **28**, 402–441 (2014).
17. Ekambaram, R. *et al.* Active cleaning of label noise. *Pattern Recognition* **51**, 463–480 (2016).
18. Northcutt, C. G., Jiang, L. & Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. *arXiv preprint arXiv:1911.00068* (2019).
19. Grandvalet, Y. & Bengio, Y. Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, 529–536 (2004).
20. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł. & Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548* (2017).
21. Krizhevsky, A. Learning multiple layers of features from tiny images. Tech. Rep., University of Toronto (2009).
22. Xia, X. *et al.* Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, vol. 33 (2020).
23. Gal, Y., Islam, R. & Ghahramani, Z. Deep Bayesian active learning with image data. In *International Conference on Machine Learning*, 1183–1192 (PMLR, 2017).
24. 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** (2020).
25. 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, 11451–11461 (2021).
26. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R. & Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1944–1952 (2017).
27. Toneva, M. *et al.* An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations* (2018).
28. Sener, O. & Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations* (2018).
29. Houlsby, N., Huszár, F., Ghahramani, Z. & Lengyel, M. Bayesian active learning for classification and preference learning (2011). URL https://arxiv.org/abs/1112.5745.
30. Junnan Li, S. C. H. H., Richard Socher. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations* (2020).
31. Nguyen, D. T. *et al.* SELF: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations* (2020).
32. Zhang, Z. & Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31*, 8778–8788 (2018).
33. Lam, C. P. & Stork, D. G. Evaluating classifiers by means of test data with noisy labels. In *IJCAI*, vol. 3, 513–518 (2003).
34. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
35. Khetan, A., Lipton, Z. C. & Anandkumar, A. Learning from noisy singly-labeled data. In *International Conference on Learning Representations* (2018). arXiv:1712.04577.
36. Rodrigues, F., Pereira, F. & Ribeiro, B. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters* **34**, 1428–1436 (2013).
37. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C. & Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11244–11253 (2019).
38. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging* **23**, 903–921 (2004).
39. Frénay, B. & Verleysen, M. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems* **25**, 845–869 (2014).
40. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330 (2017).
41. Irvin, J. *et al.* CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 590–597 (2019).
42. Oakden-Rayner, L. Quick thoughts on ChestXray14, performance claims, and clinical tasks (2017). URL https://lukeoakdenrayner.wordpress.com/2017/11/18/quick-thoughts-on-chestxray14-performance-claims-and-clinical-tasks/.
43. Stein, A. Pneumonia boxes likelihood data – RSNA pneumonia detection challenge (2020). URL https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/discussion/150643.
44. Song, H., Kim, M., Park, D. & Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199* (2020).
45. Krogh, A. & Hertz, J. A. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems 4*, 950–957 (Morgan-Kaufmann, 1992).
46. Lukasik, M., Bhojanapalli, S., Menon, A. & Kumar, S. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning*, 6448–6458 (2020).
47. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).
48. Huang, J., Qu, L., Jia, R. & Zhao, B. O2U-Net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3326–3334 (2019).
49. Yu, X. *et al.* How does disagreement help generalization against label corruption? In *International Conference on Machine Learning* (2019).
50. Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L. & Schölkopf, B. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329* (2020).
51. 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*, 2309–2318 (2018).
52. Müller, R., Kornblith, S. & Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, vol. 32 (2019).
53. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, 321–328 (2004).

Active label cleaning: Improving dataset quality under resource constraints

Supplementary information

Mélanie Bernhardt[1,*], Daniel C. Castro[1,*], Ryutaro Tanno[1], Anton Schwaighofer[1], Kerem C. Tezcan[1], Miguel Monteiro[1], Shruthi Bannur[1], Matthew Lungren[2], Aditya Nori[1], Ben Glocker[1], Javier Alvarez-Valle[1], and Ozan Oktay[1,†]

[1] *Health Intelligence, Microsoft Research Cambridge, Cambridge, CB1 2FB, UK*
[2] *Department of Radiology, Stanford University, Palo Alto, CA 94304, USA*

[*] *These authors contributed equally to this work.*
[†] *Corresponding author: ozan.oktay@microsoft.com*

Additional experiments

Classification performance of noise-robust models. We test the robustness of the proposed approaches to the corrupted labels in the training set and measure how well they generalise to an unseen set with clean labels. To this end, all networks are trained on a dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ containing both mislabelled and correctly labelled data points. To assess the robustness to the noisy labels, we separate $\mathcal{D}_{\text{train}}$ into disjoint subsets of correctly labelled ($\mathcal{D}_{\text{corr}}$) and mislabelled cases ($\mathcal{D}_{\text{misl}}$). Similarly, the robustness is measured on a hold-out evaluation set ($\mathcal{D}_{\text{eval}}$) with clean labels.

We analyse the classification metrics for these three groups separately throughout training for the different proposed methods. The results are reported in Table 1 in terms of multi-class classification accuracy (CIFAR10H) and ROC-AUC values for binary classification (NoisyCXR). In this analysis setup, robustness means that the network should learn to ignore the incorrect labels of the mislabelled cases and yield an accuracy close to chance for these. On the other hand, this should not come at a price of reduced performance for the correctly labelled cases and the network should yield a high accuracy for these. Hence the ideal case would be a large gap between the accuracy of these two groups ($\mathcal{D}_{\text{corr}}$, $\mathcal{D}_{\text{misl}}$) throughout training and the gap is a good proxy measure of sample selectors' performance in the relabelling procedure.

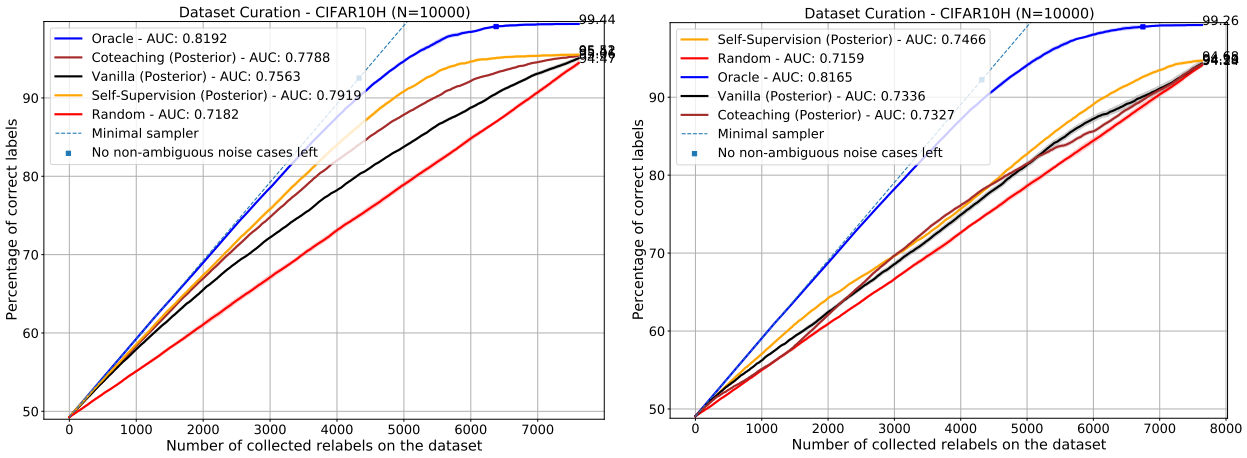
Table 1 The models used in the simulation experiments are also evaluated in terms of their classification performance on images with clean and noisy labels. In more detail, the networks are trained on a dataset containing partly mislabelled images: CIFAR10H: $|\mathcal{D}_{\text{train}}| = 10k$, $\eta = 15\%$; NoisyCXR: $|\mathcal{D}_{\text{train}}| = 13.3k$, $\eta = 12.7\%$, where η is the noise rate. Their performance is evaluated on an evaluation set $\mathcal{D}_{\text{eval}}$ with clean labels, and training set with clean $\mathcal{D}_{\text{corr}}$ and noisy labels $\mathcal{D}_{\text{misl}}$ ($\mathcal{D}_{\text{misl}} \cap \mathcal{D}_{\text{corr}} = \emptyset$). The size of the evaluation sets is $|\mathcal{D}_{\text{eval}}| = 50k$ for CIFAR10H and $|\mathcal{D}_{\text{eval}}| = 13.3k$ for NoisyCXR. The results are aggregated over three separate runs with different seeds (std. deviation in parentheses) and reported in terms of classification accuracy (Acc) and ROC-AUC (AUC).

Dataset	Model	Acc on $\mathcal{D}_{\text{eval}}$	Acc on $\mathcal{D}_{\text{corr}}$	Acc on $\mathcal{D}_{\text{misl}}$
CIFAR10H	Vanilla	77.22 (0.29)	92.38 (0.20)	45.50 (2.78)
	Co-teaching	79.19 (0.22)	88.38 (0.26)	26.79 (1.64)
	SSL + Linear head	80.56 (0.11)	85.37 (0.20)	12.62 (0.44)
		AUC on $\mathcal{D}_{\text{eval}}$	AUC on $\mathcal{D}_{\text{corr}}$	AUC on $\mathcal{D}_{\text{misl}}$
NoisyCXR	Vanilla	83.52 (0.85)	89.87 (0.42)	32.71 (1.86)
	Co-teaching	84.91 (0.31)	90.17 (0.26)	28.10 (0.29)
	SSL + Co-teaching	87.45 (0.02)	91.84 (0.16)	25.92 (0.19)

Simulation experiments with larger noise rates.

CIFAR10H Label cleaning simulations are repeated for larger noise rates to understand the sensitivity of selection algorithms to noise rate. This way we also explore the limits at which data-driven label cleaning algorithms may become unreliable. For this purpose, we use the same set of CIFAR10H images ($|\mathcal{D}| = 10k$) but with a different initial set of noisy labels sampled from the label distribution using a larger temperature parameter ($\tau > 2$). In this way, we are able to control the initial noise rate up to 34%. For noise values beyond this rate, a second parameter is introduced as an additive bias term applied to the confusion matrix, which can serve as a uniform symmetric or class-dependent noise model. At the end, we construct three different sets of initial labels with the following average noise rates and noise models: (I) 30% noise (CIFAR10H distribution), (II) 50% noise (CIFAR10H + symmetric), and (III) 50% (CIFAR10H + class-dependent).

For larger noise rates, the cost-effectiveness of sample selection algorithms are experimentally observed to be preserved as shown in Figs. 1a and 1b. It is important to note that in 30% noise case the breakdown of clear and difficult noisy cases are 23.24% and 6.66% respectively. On the other hand, when we switch to class-dependent noise model with more than 50% noise rate, the sample ranking quality begins to degrade as diagonal dominance is no longer preserved for some certain classes (e.g. dog, cat, deer, horse, and automobile). We expect the algorithms to degrade towards random sampling as we introduce further



a CIFAR10H – Noise rate 50% Symmetric Noise

b CIFAR10H – Noise rate 50% Asymmetric Noise

Fig. 1 (b) Label cleaning performance for noise rate of 30% (6.66% are difficult and incorrectly labelled); and (c, d) for noise rate of 50%, using symmetric noise and a more realistic, asymmetric noise model.

bias and noise into the initial set of labels. Lastly, the experimental results show that the graph classifier performs slightly better than linear head in the larger noise regime when both rely on the same set of fixed *SSL* embeddings.

NoisyCXR In this section, we investigate the impact of higher noise rates on label cleaning simulations for NoisyCXR. To construct datasets with more noise, we followed the same procedure as described in Methods but instead of sampling 10% of noise within the NIH "Consolidation/Infiltration" category we sampled 25% noise within this category. Including both the NIH noise and the sampled noise, the dataset contained in total 20% noise. On this dataset, note that 43% of positive initial (noisy) labels are incorrectly labelled, making it inherently challenging to learn an accurate classifier on this noisy training dataset. With this higher noise rate, we observe that the sample selection quality of *co-teaching* degrade towards the *vanilla* model. On the other hand, we still observe that combining co-teaching with self-supervised pre-training improves the quality of the label cleaning procedure, as the model is able to separate noisy from clean cases meaningfully earlier in training, hence improving co-teaching.

Further analysis on the scoring function. To better understand the impact of each term in the scoring function Φ, we conducted simulation experiments by dropping the self-entropy term that captures the sample ambiguity. We hypothesise that this term encourages selectors to give a higher priority to clear label noise cases over the samples with higher labelling difficulty. To experimentally verify this, we first split the noisy samples into two disjoint subsets: clear noisy and difficult noisy. This is done by thresholding the normalised entropy of each sample's true label distribution at 0.3. Then we tracked the size of both noisy sets throughout the active label cleaning procedure to understand the sample ranking pattern. The experiments are conducted on the CIFAR10H dataset with 30% and 50% noise rates. The corresponding results are shown in Fig. 4. We observe that by including the self-entropy term, the *SSL* selector first prioritises the clear label noise cases as highlighted by the difference between green and orange curves in the plot (middle and right), which also yields a slight performance improvement in terms of mislabel correction rate. Moreover, we observe that the best performing sample selectors favour correcting first the clear label noise cases as it can be seen by the difference between the *SSL* and *Vanilla* approaches.

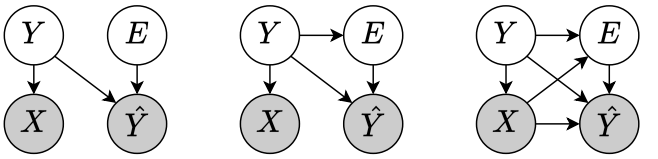


Fig. 2 Different label noise models used in robust learning. The statistical dependence between input image (X), true label (Y), observed label ($\hat{Y}$), and error occurrence (E) is shown with arrows. Adapted from Frénay et al.[1].

Impact of model update frequency: An ablation study is conducted to understand the influence of updating model's beliefs during the relabelling process (see Table 2). In that regard, experiments are repeated for different b values which determines the frequency of model updates using the collected labels. With these updates, an improved performance is observed on all model based approaches, which converge towards the upper-bound set by a model trained with all clean labels. The graph based classifier, on the other hand, does not require such updates as collected labels can be simply integrated in the ranking procedure at any time point without the need for fine-tuning.

Instance-dependent noise model. Uniform and class-dependent noise models, illustrated in Fig. 2, do not have a direct statistical dependency on the image content. This may lead researchers to design and experiment with artificial noise models that may not

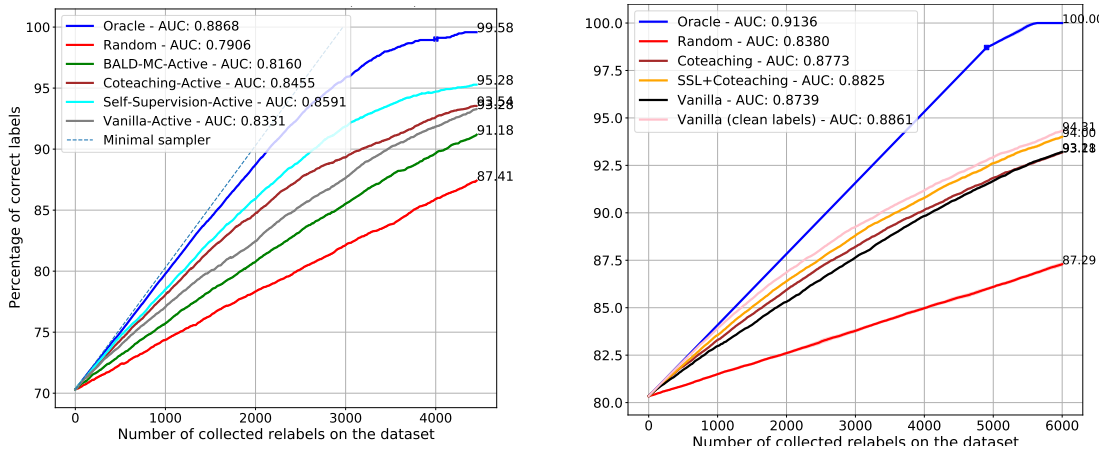


Fig. 3 Correctness of labels (y-axis) with respect to re-labelling budget (x-axis): Results are obtained on both CIFAR10H and NoisyCXR datasets with $\eta = 30\%$ and $\eta = 19.7\%$ initial noise rates respectively: Standard active learning based sample scoring functions, such as BALD[2] (in green), do not necessarily prioritise noisy labels in the relabelling procedure.

Table 2 Ablation study on the frequency of model updates (b) in active relabelling procedure. Its impact on the number of corrected labels is assessed on CIFAR10H (30% initial noise, B = 4.5k) and NoisyCXR (12.7% initial noise, B = 4k) for *vanilla* and *SSL-Linear*. The simulations are run over 5 different seeds of model fine-tuning and label sampling.

		% of Correct Labels (std) and AUC			
	Model	b = 500	b = 1000	b = 2000	$b = \infty$
NoisyCXR	Oracle	-	-	-	100.0 (0), .946
	Vanilla	94.81 (.08) .916	94.82 (.04) .915	94.75 (.05) .914	94.34 (.03) .913
	SSL	95.12 (.04) .919	95.04 (.04) .919	95.02 (.04) .918	95.03 (.02) .918
CIFAR10H	Oracle	-	-	-	99.38 (.19), .887
	Vanilla	93.23 (.11) .836	93.14 (.25) .836	92.72 (.36) .834	91.27 (.34) .829
	SSL	95.21 (.20) .861	95.12 (.19) .860	94.97 (.19) .858	94.94 (.17) .858

always hold true in practical scenarios. For instance, it would be more difficult to label low-resolution images of brown horses in comparison to white ones, since class confusion probability between horse and deer is likely to be higher.

To model such dependencies, recent work has proposed the use of instance-dependent noise (IDN) models[3,4]. In our study, we re-implemented the IDN model proposed in[4] (Algorithm 2), where instance-specific flip rates $\mathbf{p}_i \in \mathbb{R}^C$ are computed for each instance i using a set of weights $\{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_C\}$ as $\mathbf{p}_i = \mathbf{W}_{y_i} \text{vec}(\mathbf{x}_i)$. The weights $\mathbf{W}_c \in \mathbb{R}^{C \times L}$ are drawn from a normal distribution, $\text{vec}(\mathbf{W}_c) \sim \mathcal{N}(0, \mathbf{I})$. The individual class flip probabilities are later softmax-normalised across all classes and scaled with the desired noise rate to generate instance-specific class confusion matrices. In contrast to[4], here we utilise image embeddings generated with a pretrained image model (ResNet-50, ImageNet) instead of raw pixel values $\mathbf{x}_i$. This is mainly intended to associate the generated class transition models with higher-level image semantics.

Implementation Details

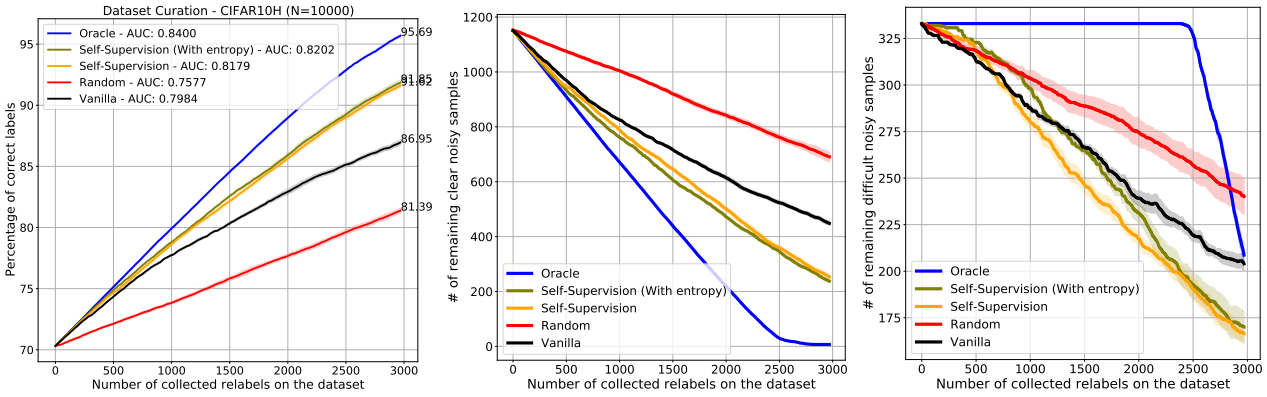
CIFAR10H.

Supervised models: For all experiments, we trained a ResNet-50 classifier with a bottleneck unit containing convolution layers with batch-normalisation. The models are trained for 160 epochs with the SGD optimiser and Nesterov momentum (0.9) using a base learning rate of 0.1 and weight decay of 10^{-4}. The learning rate is reduced by a factor of 0.1 at epochs 80 and 120. Each gradient update is computed over a mini-batch of 256 input images by computing cross-entropy with respect to smoothed target labels ($\epsilon = 0.1$). Images are augmented prior to model forward pass by applying standard natural image transformations (e.g. colour jitter, affine transformation, and horizontal flips).

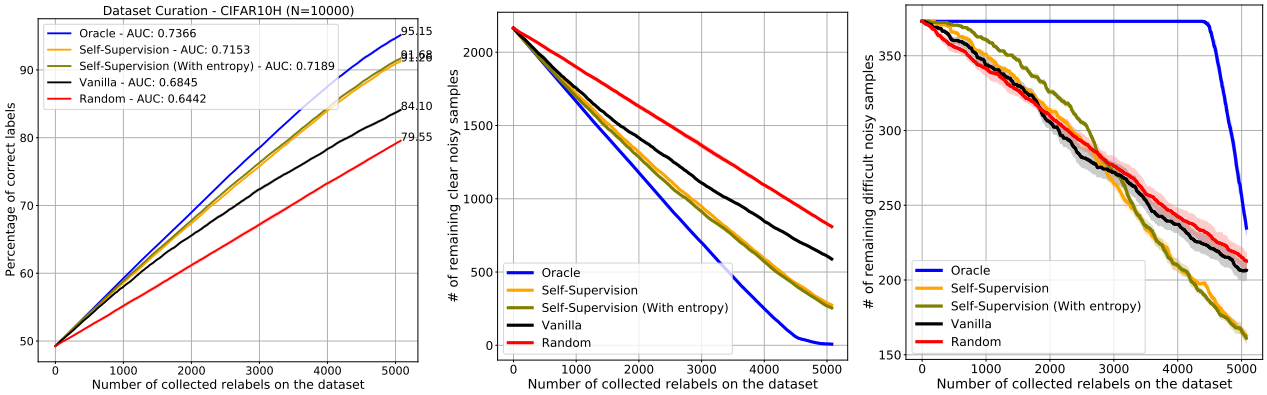
For the training of co-teaching models[5], the drop-rate is set to match the expected noise rate in the datasets, which can be further tuned on a separate validation set for potentially improved performance. In the initial warm-up phase (up to epoch 10), all samples are used in SGD updates, including the ones with large loss value.

SSL training: For the training of BYOL[6] self-supervision models, CIFAR10H dataset ($10k$) is used as a training set. For each true image pair, the standard image augmentations are applied, which are outlined in detail in the SimCLR study[7]. The encoder models are trained with Adam[8] optimiser for 2500 epochs with the base learning rate of 10^{-3} and batch size of 512. As a learning rate scheduler, a cosine decay function is utilised. More importantly, in order not to bias the model comparison, the same ResNet-50 architecture is used as the base encoder for the BYOL models.

To build an image classifier on top of *SSL* embeddings, a single linear layer is trained for 120 epochs using both weight decay (10^{-3}) and tanh logit regularisation terms ($\alpha = 20$, defined in Methods) in order to avoid over-fitting on the noisy labels. Similarly to the NoisyCXR experiments, co-teaching could potentially be utilised together with *SSL* weights without freezing the encoder parameters, which is omitted in CIFAR10H



a CIFAR10H – Noise Rate 30%



b CIFAR10H – Noise Rate 50%

Fig. 4 Results of noisy label cleaning on CIFAR10H dataset with different noise rates. The figures on the left illustrate the percentage of correct labels with respect to the number of relabels collected in the simulation. Here we see that by including the self-entropy term in *SSL* selector (green), the final outcome can be slightly improved (91.85 vs 91.02). Similarly, the number of remaining clear and difficult label noisy cases are shown in the middle and right, respectively. Best performing methods prioritise clear label noise cases over the ones with more labelling difficulty. Lastly, we see that this behaviour is further emphasised by including the ambiguity term in the sample scoring function (green vs orange curves).

experiments as the linear head experimentally proved to be sufficient in identifying noisy labels.

NoisyCXR.

Supervised models: For all experiments, a ResNet-50 encoder is trained with the Adam optimiser[8] with cross-entropy loss, using a learning rate of 10^{-5} for models trained from scratch (resp. 10^{-6} for fine-tuning), weight decay of 10^{-4} and batch size 32. For all models, we use 50% of the data for training and 50% for validation. At training time, the under-represented class ("Pneumonia-like opacity") is over-sampled to handle the class-imbalance according to the class distribution observed in the training set. In terms of preprocessing, images are first augmented in native resolution (if applicable), then resized to 256×256, and finally they are centre-cropped to 224×224. At training time, the following augmentations are used: random horizontal flipping ($p = 0.5$), random rotation ($[-30°, 30°]$), random shear transform ($[-15, 15]$), random contrast, random brightness, and random crop ($[80\%, 100\%]$). Models trained from random initialisation are trained for 150 epochs; models initialised from a pre-trained checkpoint are fine-tuned for 100 epochs. For "SSL-Linear", we built a linear image classifier on top of the frozen SSL encoder, trained for 100 epochs with a starting learning rate of 10^{-4} decreased to 10^{-5} after 10 epochs; we used tanh logits regularization ($\alpha = 10$) and reduced the set of augmentations to horizontal flips.

For the training of co-teaching models[5], the drop-rate is set to match the expected noise rate in the datasets. The warm-up phase was set to 20 epochs during which no samples are dropped (resp. 10 epochs when initialised from SSL checkpoint).

SSL training: The self-supervised models have been trained with BYOL[6] using a ResNet-50 encoder. The final model is trained using the NIH dataset[9] (80% training, 20% validation), using an effective batch size of 4800 (batches of 600 pairs of images of size 224×224 on 8 GPUs) for 1000 epochs. The momentum parameter τ used in BYOL teacher encoder was set to 0.99. As augmentations, we use random horizontal flipping (with probability $p = 0.5$), random rotation and shear (drawn from $[-180°, 180°]$ and $[-40, 40]$), random crop ($[40\%, 100\%]$), CutOut (masking out rectangles of varying size, size covering 15–40% of the image), elastic transforms[10] ($a = 34$, $s = 4$, applied with probability 0.5), random contrast and

Table 3 The impact of image augmentations on the quality of learnt BYOL embeddings, when applied to medical images. The embedding quality is assessed in terms of linear separation of classes by training a linear-head on top of BYOL encoder output. At each experiment, one data augmentation is omitted to measure its impact, and classification results are obtained on the validation set using ROC-AUC.

Augmentations	ROC-AUC
All w/out cropping	0.853
All w/out brightness and contrast	0.866
All w/out rotation and shear	0.868
All augmentations (baseline)	0.871

brightness changes, and adding Gaussian noise ($\sigma = 0.01$, applied with probability 0.5).

We ran experiments on the NoisyCXR dataset to determine the impact of each of those augmentations on the quality of the resulting embedding. To this end, a linear classifier is learnt on top of image embeddings during BYOL training and its performance is monitored with ROC-AUC metric on the validation set. By disabling one augmentation at a time and comparing to the baseline with all augmentations, we see that image cropping contributes most to the results, followed by contrast and brightness augmentations, then rotation and shear, results are provided in Table 3. Using the full set of augmentations often meant that performance kept increasing when training for longer, whereas using fewer augmentations led to performance plateauing earlier in training. Histogram normalisation had a noticeable impact on the result. Using the native images without histogram normalisation led to an AUC roughly 0.6% higher than with histogram normalisation.

References

1. Frénay, B. & Verleysen, M. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems* **25**, 845–869 (2014).
2. Gal, Y., Islam, R. & Ghahramani, Z. Deep Bayesian active learning with image data. In *International Conference on Machine Learning*, 1183–1192 (PMLR, 2017).
3. Chen, P., Ye, J., Chen, G., Zhao, J. & Heng, P.-A. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35.
4. Xia, X. *et al.* Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, vol. 33 (2020).
5. Han, B. *et al.* Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 8527–8537 (2018).
6. Grill, J.-B. *et al.* Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, vol. 33, 21271–21284 (2020).
7. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607 (PMLR, 2020).
8. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations* (2015). arXiv: 1412.6980.
9. Wang, X. *et al.* ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471 (2017).
10. Simard, P. Y., Steinkraus, D. & Platt, J. C. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)*, 958–963 (2003).