LEARNING WITH INSTANCE-DEPENDENT LABEL NOISE: BALANCING ACCURACY AND FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Incorrect labels hurt model performance when the model overfits to noise. Many state-of-the-art approaches that address label noise assume that label noise is independent from the input features. In practice, however, label noise is often feature or instance-*dependent*, and therefore is biased (i.e., some instances are more likely to be mislabeled than others). Approaches that ignore this dependence can produce models with poor discriminative performance, and depending on the task, can exacerbate issues around fairness. In light of these limitations, we propose a two-stage approach to learn from datasets with instance-dependent label noise. Our approach utilizes *anchor points*, a small subset of data for which we know the ground truth labels. On many tasks, our approach leads to consistent improvements over the state-of-the-art in discriminative performance (AUROC) while balancing model fairness (area under the equalized odds curve, AUEOC). For example, when predicting acute respiratory failure onset on the MIMIC-III dataset, the harmonic mean of the AUROC and AUEOC of our approach is 0.84 (SD 0.01) while that of the next best baseline is 0.81 (SD 0.01). Overall, our approach leads to learning more accurate and fair models compared to existing approaches in the presence of instance-dependent label noise.

1 INTRODUCTION

Datasets used to train machine learning models can contain incorrect labels (i.e., label noise). While label noise is widely studied, the majority of past work focuses on when the noise is independent from an instance's features (i.e., instance-independent label noise) [Song et al. (2020)]. However, label noise is sometimes biased and depends on an instance's features (i.e., instance-dependent) [Wei et al. (2022b)], leading to different noise rates within subsets of the data. This results in model overfitting, and in tasks where the dataset contains instances from different groups corresponding to some sensitive attribute, this can also lead to disparities in performance [Liu (2021)]. For example, consider the task of predicting cardiovascular disease among patients admitted to a hospital. Compared to male patients, female patients may be more likely to be misdiagnosed [Maserejian et al. (2009)] and thus mislabeled, potentially leading to worse predictions for female patients. Although instance-dependent label noise has recently received more attention [Cheng et al. (2020b); Xia et al. (2020); Wang et al. (2021a)], the effect of these approaches on model fairness has been relatively understudied [Liu (2021)]. Here, we address the limitations of current approaches and propose a novel method for learning with instance-dependent label noise, specifically examining how modeling assumptions affect existing issues around model fairness.

Broadly, current work addressing instance-dependent label noise falls into one of two categories: 1) that which learns to identify mislabeled instances [Cheng et al. (2020a); Xia et al. (2022); Zhu et al. (2022a)], and 2) that which learns to optimize a noise-robust objective function [Feng et al. (2020); Wei et al. (2022a)]. In the first category, instances identified as mislabeled are either filtered out [Kim et al. (2021)] or relabeled [Berthon et al. (2021)]. In some settings, this approach can have a negative effect on model fairness. For example, when instances represent individuals belonging to subgroups defined by a sensitive attribute, approaches that filter out mislabeled individuals could ignore a disproportionately higher number of individuals from subgroups with more label noise. While relabeling approaches use all available data, they can be sensitive to assumptions around the noise distribution [Ladouceur et al. (2007)]. In the second category, current approaches rely on objective functions that are less prone to overfitting to the noise, while using all of the data and

Table 1: Notation. We summarize our notation, with the notation appearing in the left column and a description in the right column. Superscripts in parentheses represent specific instances (e.g., $\mathbf{x}^{(i)}$).

Notation	Description	Notation	Description
n	Number of instances	d	Number of features
c	Number of classes	G	Number of groups
$\mathbf{x} \in \mathbb{R}^d$	Feature vector	$\tilde{y} \in \{1, 2, ..., c\}$	Observed label
$\hat{y} \in [0, 1]^c$	Predicted class probabilities	$y \in \{1, 2, ..., c\}$	Ground truth label
$\hat{\beta}_\phi$	$P(y == \tilde{y} \mid \tilde{y}, \mathbf{x})$	A	Set of anchor points
θ	Main model parameters	ϕ	Auxiliary model parameters

observed labels [Chen et al. (2021)]. However, like the first category, many rely on assumptions like the memorization effect, and thus, potentially suffer from the same limitations [Wang et al. (2021a)].

In light of these limitations, we propose an approach that addresses instance-dependent label noise, makes no assumptions about the noise distribution, and utilizes all data during training. We leverage a set of anchor points (i.e., a subset of data for which we know the observed *and* ground truth labels [Xia et al. (2019)]) that is representative of the target population. Generally, these can be obtained during data collection via manual review by an expert of a random subset. Anchor points allow one to initialize the model's decision boundary and learn the underlying pattern of label noise. In our approach, we do both of these in a pre-training step. We then add the rest of the data and fine tune the model's decision boundary, minimizing a cross-entropy loss that is weighted based on the learned pattern of label noise.

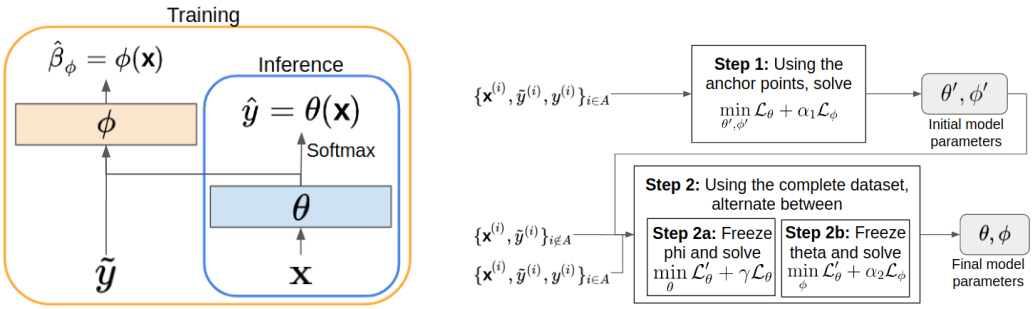
On synthetic and real data, we evaluate our approach in terms of discriminative performance and model fairness, measured using the area under the receiver operator curve (AUROC) and area under the equalized odds curve (AUEOC). We demonstrate that our approach improves on state-of-the-art baselines from the noisy labels and fairness literature, such as stochastic label noise [Chen et al. (2021)] and group-based peer loss [Wang et al. (2021b)]. Overall, our contributions are: 1) a novel approach to learn from datasets with instance-dependent noise that utilizes anchor points; 2) a systematic examination of different settings of label noise, showing where the baselines fail with respect to discriminative performance and fairness; 3) empirical results showing that the proposed approach is robust to both to the noise rate and amount of noise disparity between subgroups, reporting the model's ability to balance discriminative performance and fairness; 4) a demonstration of how performance of the proposed approach changes when assumptions about the anchor set are violated.

2 METHODS

We introduce a two-stage approach for learning with instance-dependent label noise that leverages a small set of anchor points for which we have both observed and ground truth labels. Here, we describe our notation and proposed approach, including the training procedure and objective function.

Notation and Problem Setup Our main notation is in **Table 1**. We aim to learn a model using dataset $D = \{\mathbf{x}^{(i)}, \tilde{y}^{(i)}\}_{i=1}^n$ that classifies each instance into one of c classes based on $\mathbf{x}$. In the presence of noisy labels, $\tilde{y}$ is not always equal to y. For each anchor point, we know both $\tilde{y}$ and y. This is in contrast to the rest of the data, where we only know $\tilde{y}$. We aim to learn model parameters, θ, such that $\theta(\mathbf{x})$ represents the predicted class probabilities. Using $\theta(\mathbf{x})$, we obtain prediction $\hat{y}$.

For the rest of this paper, we focus on the following case of instance-dependent noise. Let f be the function used to generate the ground truth labels (i.e., $f(\mathbf{x}) = y$), and let $\tilde{m}$ be the function used to generate an instance's risk of being mislabeled (i.e., $y = \tilde{y}$ if $\tilde{m}(\mathbf{x})$ is above some threshold). For example, suppose $f(\mathbf{x}) = 1$ for true positive instances, and $\tilde{m}(\mathbf{x}) = 1$ if $x_1 > 0.5$ and 0 otherwise. In this case, instances where $x_1 > 0.5$ have noisy labels. Although f and $\tilde{m}$ were deterministic in this example for simplicity, they can also be based on probabilistic functions. We denote the set of anchor points (i.e., the anchor set) as set A consisting of instances i such that both $\tilde{y}^{(i)}$ and $y^{(i)}$ are known and assuming that it includes instances where $\tilde{y}^{(i)} \neq y^{(i)}$. Although it is not always possible to obtain anchor points, the setting above does occur in practice. For example, in healthcare, automated labeling tools based on the structured contents of the electronic health record are often



(a) Proposed network. The model makes class predictions, $\hat{y}$, at training and inference time using the portion parameterized by θ. At training time, it also predicts whether the observed label is correct using the portion parameterized by ϕ. θ and ϕ are pre-trained using A and then fine tuned with the complete dataset.

(b) Proposed training pipeline. We begin by pretraining the model using the anchor points (i.e., data with observed and ground truth labels), then train on the noisy data. $\mathcal{L}_\theta$ is the objective function used to train on the anchor points. $\mathcal{L}_\phi$ is the objective function used to learn the label confidence score ($\hat{\beta}_\phi$). $\mathcal{L}'_\theta$ is the objective function used to train the noisy data. α_1, α_2, and γ are scalar hyperparameters.

Figure 1: Proposed Approach

applied to identify cohorts of interest or outcomes of interest [Upadhyaya et al. (2017); Norton et al. (2019); Tjandra et al. (2020); Yoo et al. (2020)]. However, such practical definitions are not always reflective of ground truth. Thus, these labeling tools require validation by comparing to a subset of the data that was chart reviewed by an expert. As a result, f is whether the condition is present, $\tilde{m}$ is the automatic labeling tool, and the anchor set is the expert-labeled subset.

Proposed Approach We describe the proposed network and training procedure below, including how we learn to predict both class probabilities and label confidence (i.e., $P(y == \tilde{y}|\tilde{y}, \mathbf{x})$).

Proposed Network. Our proposed network (**Figure 1a**) consists of two components. The first, θ, is a feed-forward network that uses feature vector $\mathbf{x}$ to predict the class probabilities, $\hat{y} = \theta(\mathbf{x})$. The second component, ϕ, is an auxiliary feed-forward network that uses observed label $\tilde{y}$ and the pre-softmax output of θ to compute $\hat{\beta}_\phi = P(y == \tilde{y}|\tilde{y}, \mathbf{x})$, an instance-dependent prediction for whether the observed label is correct. $\hat{\beta}_\phi$ can be considered as a confidence score for the observed label, with higher values indicating higher confidence. Learning $\hat{\beta}_\phi$ learns the underlying pattern of label noise by forcing the model to distinguish between which examples are correctly or incorrectly labeled. This will be helpful during training, which we describe below. By including the observed label as input to ϕ, our approach also applies to instance-independent label noise by accounting for the case when the underlying pattern of label noise cannot be learned from the features alone. At training time, we compute the loss using the outputs from both θ and ϕ. At inference time, we compute the class predictions from θ only since $\tilde{y}$ is not available.

Training Procedure. Our overall procedure is summarized in **Figure 1b**. In Step 1, we pre-train the θ and ϕ networks using only the anchor points, minimizing a standard objective function based on cross entropy, where $\alpha_1 \in \mathbb{R}^+$ is a scalar hyperparameter: $\min_{\theta', \phi'} \mathcal{L}_\theta + \alpha_1 \mathcal{L}_\phi$. $\mathcal{L}_\theta$, captures the cross entropy loss between the the class predictions and the ground truth labels. It learns the weights for θ and in turn, the model's decision boundary. $\mathbb{I}$ is an indicator function.

$$\mathcal{L}_\theta = \frac{-1}{|A|} \sum_{i \in A} \sum_{j=1}^{c} \mathbb{I}\left(y^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)$$

$\mathcal{L}_\phi$ captures the cross entropy loss between the predicted confidence score $\hat{\beta}_\phi^{(i)}$ and the actual agreement between $\tilde{y}^{(i)}$ and $y^{(i)}$. It learns the weights for ϕ and in turn, the underlying label noise pattern.

$$\mathcal{L}_\phi = \frac{-1}{|A|} \sum_{i \in A} \mathbb{I}\left(\tilde{y}^{(i)} == y^{(i)}\right) log\left(\hat{\beta}_\phi^{(i)}\right) + \mathbb{I}\left(\tilde{y}^{(i)} \neq y^{(i)}\right) log\left(1 - \hat{\beta}_\phi^{(i)}\right)$$

In Step 2, we initialize θ and ϕ using the network weights from Step 1 and fine tune them using the complete dataset, which potentially contains noisy labels. Step 2 can be further divided into two

parts, Step 2a and Step 2b. Each part aims to improve a specific component of the network (e.g., θ) using another component of the network (e.g., ϕ). Throughout Step 2, we alternate between Step 2a and Step 2b in a manner similar to expectation maximization so that we continually improve both θ and ϕ. In Step 2a, we freeze ϕ and train θ by minimizing the objective $\min_\theta \mathcal{L}'_\theta + \gamma\mathcal{L}_\theta$. $\gamma \in \mathbb{R}^+$ is a scalar hyperparameter. In Step 2b, we freeze θ and train ϕ using the objective $\min_\phi \mathcal{L}'_\theta + \alpha_2\mathcal{L}_\phi$. $\alpha_2 \in \mathbb{R}^+$ is a scalar hyperparameter. $\mathcal{L}'_\theta$ computes the cross entropy loss over the potentially noisy, non-anchor points. Each instance is weighted by the model's confidence in whether the observed label is correct via $\hat{\beta}_\phi^{(i)}$, taking advantage of the model's learned noise pattern.

$$\mathcal{L}'_\theta = \frac{-1}{|\overline{A}|}\sum_{k=1}^{G}\frac{1}{1-\hat{r}_k}\sum_{i\in\overline{A}\cap g_k}\sum_{j=1}^{c}\hat{\beta}_\phi^{(i)}\mathbb{I}\left(\tilde{y}^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)$$

The quantity $1-\hat{r}_k$ is defined as follows. We introduce the sets g_k for $k = 1, 2, ..., G$ to represent disjoint subgroups of interest in the dataset, where $g_a \cap g_b = \emptyset$ for all $a = 1, 2, ..., G$, $b = 1, 2, ..., G$ with $a \neq b$ and $\cup_{k=1}^{G} g_k = D$. Each group g_k is then associated with estimated noise rate $\hat{r}_k = \frac{1}{|g_k|}\sum_{i\in g_k} 1-\hat{\beta}_\phi^{(i)}$ and estimated clean (i.e., correct) rate $1-\hat{r}_k = \frac{1}{|g_k|}\sum_{i\in g_k}\hat{\beta}_\phi^{(i)}$. Although weighting each instance by $\hat{\beta}_\phi$ is a form of soft filtering, weighting each group by the inverse of its overall clean rate avoids the effect of de-emphasizing groups with higher predicted noise rates. As a result, the expected value of $\mathcal{L}'_\theta$ with respect to $\hat{\beta}_\phi$ is equal to the cross entropy loss between the model's predictions and ground truth labels (see **Appendix A** for proof). However, this assumes accurate estimates of $\hat{\beta}_\phi$. Thus, we expect that the proposed approach will perform best when the anchor set is representative of the target population (i.e., there is no dataset shift from the anchor set to the rest of the dataset), since this will produce the best estimate of $\hat{\beta}_\phi$. In **Section 4**, we test this.

During Step 2a, $\mathcal{L}'_\theta$ is used to train θ by learning to predict $\hat{y}$ such that it matches observed label $\tilde{y}$ on instances that are predicted to be correctly labeled. During Step 2b, $\mathcal{L}'_\theta$ is used to train ϕ. Here, since θ is frozen and ϕ is not, the network learns to predict the optimal $\hat{\beta}_\phi$. Based on $\mathcal{L}'_\theta$ alone, there are two possible ways to learn $\hat{\beta}_\phi$. The first is to consistently make $\hat{\beta}_\phi$ close to 0. The second is to learn $\hat{\beta}_\phi$ such that it is close to 1 when $\hat{y}$ matches $\tilde{y}$ and close to 0 when $\hat{y}$ does not match $\tilde{y}$. Since $\tilde{y}$ is used as a proxy for y in this step, the second way aligns with what we want $\hat{\beta}_\phi$ to represent. To encourage this over the first way (i.e., consistently predicting 0 for $\hat{\beta}_\phi$), we include $\mathcal{L}_\phi$ in Step 2b, which is *not* minimized by consistently predicting 0 for $\hat{\beta}_\phi$. Note that, in order to expect a benefit by including the noisy data in Step 2b, we rely on the cluster assumption [Singh et al. (2008)] from semi-supervised learning, which broadly states that labeled data fall into clusters and that unlabeled data aid in defining these clusters. In the context of Step 2b, 'labeled' and 'unlabeled' are analogous to whether we know if the ground truth and observed labels match (i.e., anchor point versus non-anchor point). As a result, we rely on the anchor set being representative of the target population here as well to avoid dataset shift.

In contrast to previous filtering approaches, our approach utilizes all data during training. Moreover, it does not require a specialized architecture beyond the auxiliary network to compute $\hat{\beta}_\phi$. Thus, it can be used to augment existing architectures in a straightforward manner.

3 Experimental Setup

We empirically explore the performance of our proposed approach relative to state-of-the-art baselines on five benchmark prediction tasks with two different label noise settings. Implementation details are provided in **Appendices C** and **D**.

Datasets We consider five different binary prediction tasks on four datasets from several domains on synthetic and real datasets. While work in noisy-label learning often focuses on image datasets (e.g., MNIST or CIFAR 10/100), given our emphasis on fairness, we selected datasets reflective of tasks in which harmful biases can arise (e.g., predicting clinical outcomes, predicting whether a criminal defendant is likely to re-offend, and predicting income).

Synthetic: We generated a dataset containing 5,000 instances as described below. Each instance had 30 features, and features for each instance were drawn from a standard normal distribution.

We defined the feature at index 0 to be a synthetic sensitive attribute. Instances with values below the 20th percentile for this feature were considered as the 'minority', and the rest were considered as the 'majority'. Features 10-19 for the majority instances and features 20-29 for the minority instances were set to 0 to provide more contrast between the two groups. Coefficients describing the contribution of each feature to the outcome, $\mathbf{w}$, were also drawn from a standard normal distribution, and an outcome z was computed by taking the dot product of the features and coefficients and passing it through a sigmoid. A binary ground truth label, y, for each instance was assigned, based on z. Instances whose value for z fell above the 50th percentile were given the label 1. All other instances were labeled 0. The positive rates for the synthetic majority and minority groups were 37.5% and 32.3%, respectively.

MIMIC-III: Within the healthcare domain, we leverage a publicly available dataset of electronic health record data [Johnson et al. (2016)]. We considered two separate prediction tasks: onset of 1) acute respiratory failure (ARF) and 2) shock in the ICU (intensive care unit) [Oh et al. (2019)]. MIMIC-III includes data pertaining to vital signs, medications, diagnostic and procedure codes, and laboratory measurements. We considered the four hour prediction setup for both tasks as described by [Tang et al. (2020)]. This resulted in 15,873 and 19,342 ICU encounters for each task, respectively. After preprocessing (see **Appendix B**), each encounter had 16,278 and 18,186 features, for each task respectively. We used race as a sensitive attribute, with about 70% of patients being White (positive rate 4.5% [ARF], 4.1% [shock]) and 30% being non-White (4.4% [ARF], 3.7% [shock]).

Adult: a publicly available dataset of census data [Dua & Graf (2017)]. We considered the task of predicting whether an individual's income is over $50,000. This dataset includes data pertaining to age, education, work type, work sector, race, sex, marital status, and country. Its training and test sets contain 32,561 and 16,281 individuals, respectively. We used a pre-processed version of this dataset[1]. We only used 1,000 randomly selected individuals out of 32,561 during training and only included features pertaining to age, education, work type, marital status, work sector, and sex to make the task more difficult (see **Appendix B**). We also normalized each feature to have a range of 0-1. This resulted in 56 features per individual. We used sex as a sensitive attribute, with 67.5% of individuals being male (positive rate 30.9%) and 32.5% being female (positive rate 11.3%).

COMPAS: a publicly available dataset collected by ProPublica from Broward County, Florida, USA [Angwin et al. (2016)]. We considered the task of predicting recidivism within two years, i.e., whether a criminal defendant is likely to re-offend. COMPAS includes data pertaining to age, race, sex, and criminal history. We used a pre-processed version of this dataset[2] and also normalized each feature to have a range of 0-1 (see **Appendix B**). This resulted in 6,172 individuals with 11 features per individual. We used race as a sensitive attribute, with 65.8% of individuals being White (positive rate 39.1%) and 34.2% being non-White (positive rate 44.5%).

Introducing Label Noise To test the robustness of our approach in different settings of label noise, we introduce noise in two ways. Like past work [Song et al. (2020)], our setup is limited in that our noise is synthetic and we use the labels in the real datasets as ground truth, since we do not have access to actual ground truth labels on these public datasets. For uniform random noise, each instance had its label flipped with a fixed probability (e.g., 0.3). In this setting, noise rates were the same between subgroups. For instance-dependent noise, mislabeling was a function of the features. Let $\mathbf{w}_m$ ($\mathbf{w}_m \sim N(0, 0.33)^D$) and z_m ($z_m = \sigma(\mathbf{x} \cdot \mathbf{w}_m)$) denote the coefficients describing the contribution of each feature to mislabeling and the risk of mislabeling, respectively. Whether an instance was mislabeled was based on z_m and the desired noise rate ($\tilde{m}(\mathbf{x}) = y \; if \; z_m > 1 - noise \; rate, \; 1 - y \; otherwise$). For example, for a noise rate of 30%, instances whose value for z_m was above the 70th percentile had their labels flipped. This setup allowed us to vary the overall noise rate and noise rate within subgroups in a straightforward manner. Across datasets, the noise rate in the 'minority' population was always greater than or equal to that of the 'majority' group.

Evaluation Metrics We evaluate our proposed approach in terms of both discriminative performance and model fairness. For discriminative performance, we evaluate using the area under the receiver operating characteristic curve (AUROC) (higher is better). For model fairness, we measure the area under the equalized odds curve (AUEOC) [de Freitas Pereira & Marcel (2020)] (higher is better). We use the AUEOC rather than equalized odds since equalized odds requires choosing

[1] https://github.com/AissatouPaye/Fairness-in-Classification-and-Representation-Learning

[2] https://www.kaggle.com/danofer/compass

a classification threshold to measure the true and false positive rates. For classification threshold τ, we calculate the equalized odds (EO(τ)) between two groups, called 1 and 2, as shown below. $TP_a(\tau)$ and $FP_a(\tau)$ denote true and false positive rates for group a at threshold τ, respectively. The AUEOC is obtained by plotting the EO against all possible values of τ and calculating the area under the curve. We also compute the harmonic mean (HM) between the AUROC and AUEOC to highlight how the approaches balanced discriminative performance and fairness. We compute the harmonic mean so that the overall measure is more heavily weighted by the worse performing metric. For example, if a classifier has AUROC=0.5 and AUEOC=1.0, then we place greater emphasis on the poor discriminative performance.

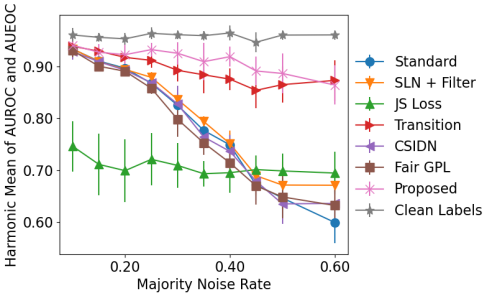
$$EO(\tau) = \frac{2 - |TP_1(\tau) - TP_2(\tau)| - |FP_1(\tau) - FP_2(\tau)|}{2}$$

Baselines We evaluate our proposed approach in the context of several baselines to test different hypotheses. Standard does not account for label noise and assumes that $\tilde{y} = y$ is always true. SLN + Filter [Chen et al. (2021)] combines filtering [Arpit et al. (2017)] and SLN [Chen et al. (2021)] and was shown to outperform state-of-the-art approaches such as Co-Teaching [Han et al. (2018)] and DivideMix [Li et al. (2020)]. It relies on filtering heuristics, which indirectly rely on uniform random label noise to maintain discriminative performance and model fairness. JS (Jensen-Shannon) Loss [Englesson & Azizpour (2021)] builds from semi-supervised learning and encourages model consistency when predicting on perturbations of the input features and was shown to be competitive with other state-of-the-art noise-robust loss functions [Ma et al. (2020)]. It was proposed in the context of instance-independent label noise. Transition [Xia et al. (2020)] learns to correct for noisy labels by learning a transition function and was shown outperform state-of-the-art approaches such as MentorNet [Jiang et al. (2018)]. It applies to instance-dependent label noise, but it assumes that the contributions of each feature to mislabeling and input reconstruction are identical. CSIDN (confidence-scored instance-dependent noise) [Berthon et al. (2021)] also learns a transition function and was shown outperform state-of-the-art approaches such as forward correction [Patrini et al. (2017)]. Like our approach, CSIDN uses the concept of 'confidence' in the observed label to help with training. Unlike our approach, CSIDN uses the model's class predictions directly as confidence scores (instead predicting them via an auxiliary network) and uses them to learn the transition function (as opposed to re-weighting the loss). Fair GPL [Wang et al. (2021b)] builds on work addressing uniform random label noise [Jiang & Nachum (2020)] and uses peer loss (i.e., data augmentation that reduces the correlation between the observed label and model's predictions) within subgroups defined by a sensitive attribute [Wang et al. (2021b)]. It assumes that label noise depends on only the sensitive attribute. We also train a model using the ground truth labels (referred to as Clean Labels) as an empirical upper bound for discriminative performance.

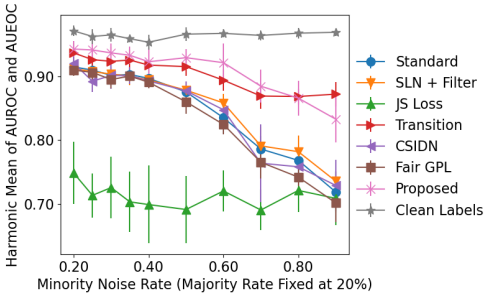
4 Results and Discussion

Is the Approach Robust to Varying Noise Levels? Here, we investigated how robust the proposed approach and baselines were to varying amounts of instance-dependent label noise (**Figure 2a**). On the synthetic dataset, we considered noise rates between 10% and 60% in the majority group. For the minority group, we considered noise rates of that were consistently 20% higher than that of the majority group, to keep the noise disparity level (i.e., the difference in noise rates between subgroups) constant. As expected, all approaches degrade as the noise rate increases, with the proposed approach experiencing the least degradation up to a majority noise rate of 60%. The baseline Transition was also robust to the noise rate, although overall performance was generally worse than the proposed approach. At a noise rate of 60%, it performed similarly to the proposed approach. The other approaches experienced more degradation, likely because they were for instance-independent label noise, highlighting the need for approaches that are designed for instance-dependent noise.

Is the Approach Robust to Varying Noise Disparities? Here we investigated how robust the proposed approach and baselines were to varying levels of noise disparity (**Figure 2b**). With a majority noise rate of 20% throughout, we evaluated the approaches with minority noise rates between 20% and 90% (i.e., from no disparity to a disparity of 70%) on the synthetic dataset. In line with our expectations, the proposed approach was robust to changes in the noise disparity. This is likely because the objective function $\mathcal{L}'_\theta$ from Step 2 of training accounts for disparities by scaling each instance-specific loss term with the reciprocal of its estimated group clean rate (i.e., 1 - the group noise rate).

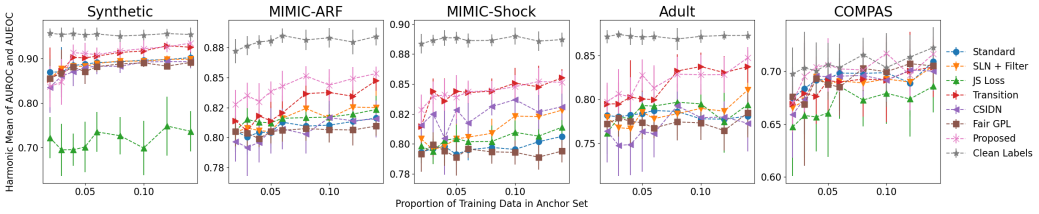


(a) As the noise rate increases, the proposed approach generally shows the least degradation.

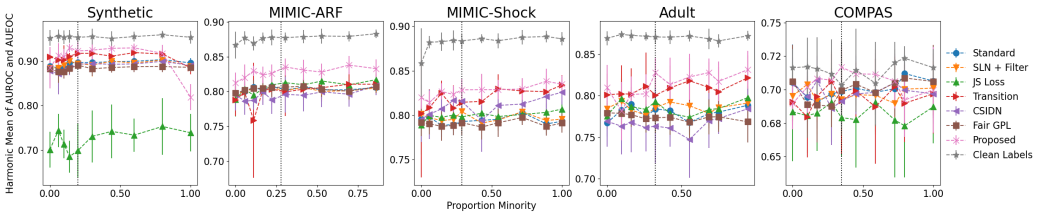


(b) As the noise disparity increases, the proposed approach generally shows the least degradation.

Figure 2: Robustness to noise rate and noise disparity in an instance dependent setting. We plot the mean and standard deviation for 10 random seeds.



(a) As we decrease the anchor set size (proportion of training data) performance decreases. Still, at an anchor set size of 5%, the proposed approach generally outperforms the baselines.



(b) As we vary the anchor set bias (proportion of minority instances) performance varies. The proposed approach is generally robust to changes in the bias of the anchor set. The dashed vertical black line shows the proportion at which the minority group occurs in the dataset (i.e., an unbiased anchor set).

Figure 3: Robustness to anchor set composition. Mean and standard deviation for 10 random seeds.

Most of the baselines, such as SLN + Filter, which learn to filter out noisy instances, experienced a larger drop in performance as the disparity increased. This aligns with our hypothesis. Although the baseline Transition also demonstrated robustness to the noise disparity and outperformed the proposed approach at minority noise rates higher than 80%, this setting is unlikely to be realistic.

Does Anchor Set Composition Matter? Our next set of experiments tests the proposed approach in settings where we relax key assumptions about the anchor set. Here, we consider all datasets in the instance-dependent noise setting. The majority/minority noise rates were 20%/40%, respectively.

Part 1: Anchor set size. We varied the size of the anchor set, testing at sizes between 1% and 15%, while keeping the anchor points representative of the population (**Figure 3a**). As expected, when the anchor set is small, performance worsens. However, the proposed method generally improves over the baselines at an anchor set size of as small as 3% of the training data. On the real datasets, the proposed approach had a higher HM compared to the baselines at an anchor set size of $< 5\%$, showing greater balance between discriminative performance and fairness. As the size of the anchor set increased, some of the baseline approaches were able to match that of the proposed method (e.g., for MIMIC-Shock, the baseline Transition had similar performance to the proposed method when the anchor set was larger than 5%). This is in line with expectations, as this makes the overall training set cleaner, and thus easier to learn from.

Part 2: Biased anchor set. Our proposed approach assumes that the anchor points are representative of the population. Here, we test how the proposed approach performs when this assumption is violated. We varied the amount of bias in the anchor set by changing the proportion at which the subgroups were present. We kept the size of the anchor set constant at 10% of the training data (2% for MIMIC-III on both tasks). We observe that the proposed approach generally provides the most benefit in situations where the anchor set is relatively balanced, i.e., when the minority proportion is 20%-80% (**Figure 3b**). In scenarios where the proposed approach is less beneficial compared to the baselines, one subgroup dominated the anchor set. This is shown in **Figure 3b** on the extremes of the x-axis, which correspond to an anchor set that is heavily over-represented for one subgroup and heavily under-represented for the other. Our approach relied on having a relatively unbiased anchor set for estimating $\hat{\beta}_\phi$ in order to avoid introducing dataset shift between the two steps of our training pipeline. Thus, these results are in line with our expectations and highlight a limitation of our approach. However, we observe that the anchor set does not need to be perfectly unbiased for the proposed approach to be beneficial for some datasets, such as MIMIC-III.

Additional Experiments: We include results for experiments where we vary the proportion at which the minority group appears in the entire dataset (with instance-dependent noise) as opposed to only the anchor set and results for the uniform random noise setting in **Appendix E**. Overall, we find that our approach is effective when the minority group is small, though the added benefit decreases at minority proportions below 10%. Our approach is also robust to random noise.

5 Related Work

Label noise is a well-studied problem in machine learning [Song et al. (2020)] and may be instance-independent or instance-dependent. However, regardless of the setting, past approaches fall into two broad categories. The first aims to identify mislabeled instances via filtering heuristics or by estimating the probability of mislabeling [Patrini et al. (2017); Li et al. (2020); Cheng et al. (2020b)]. The second adapts regularization techniques, such as label smoothing, peer loss, and contrastive learning to the label noise setting [Manwani & Sastry (2013); Menon et al. (2019); Xue et al. (2022); Liu et al. (2022)]. We also examine how label noise has been tackled more recently in the fairness literature and how past work has utilized anchor points.

Identifying Mislabeled Data Approaches that learn to identify mislabeled instances fall into two sub-categories: 1) filtering approaches and 2) relabeling approaches. Filtering approaches use heuristics to identify mislabeled instances (e.g., MentorNet [Jiang et al. (2018)], Co-teaching [Han et al. (2018)], FINE [Kim et al. (2021)]). Many are based on the idea that correctly labeled instances are easier to classify than mislabeled instances (i.e., the memorization effect) [Arpit et al. (2017)]. For example, mislabeled instances could be those that the model incorrectly classifies [Verbaeten (2002); Khoshgoftaar & Rebours (2004); Thongkam et al. (2008); Chen et al. (2019)], have a high loss value [Yao et al. (2020a)], or significantly increase the complexity of the model [Gamberger et al. (1996)]. Given the identified mislabeled instances, these approaches either ignore them during training [Zhang et al. (2020)] or treat them as 'unlabeled' and apply techniques from semi-supervised learning (e.g., DivideMix [Li et al. (2020)], SELF [Nguyen et al. (2020)]). Overall, these heuristics have been shown to improve discriminative performance. However, depending on the setting, their application can result in disproportionately discarding subsets of data, which in turn could hurt model fairness.

For binary classification, some approaches 'correct' (i.e., switch) the observed label for such instances [Han et al. (2020); Zheng et al. (2020)]. Building on this idea, others make use of a transition function that estimates the probability of the observed label being correct. Model predictions can then be adjusted by applying the function to the classifier's predictions for each class. Some works manually construct the transition function from expert knowledge [Patrini et al. (2017)], while others learn it [Xiao et al. (2015); Xu et al. (2019); Yao et al. (2020b); Xia et al. (2020); Zheng et al. (2021); Berthon et al. (2021); Jiang et al. (2022); Bae et al. (2022)]. However, such approaches often make assumptions on the form of the noise distribution, and past work has shown that results are sensitive to the choice of distribution [Ladouceur et al. (2007)].

To date, much of the work described above assumes instance-independent label noise (i.e., the mislabeling is independent of the features). However, when this assumption is violated, the model may overfit to label noise [Lukasik et al. (2020)]. From an emerging body of work in instance-dependent

label noise [Cheng et al. (2020b); Xia et al. (2020); Wang et al. (2021c); Zhu et al. (2022b)], current approaches remain limited in that they still rely on filtering heuristics. Our approach applies in the instance-dependent setting and does not rely on filtering heuristics. While similar to a transition function in some aspects, our approach requires fewer probability estimates on label correctness (two estimates compared to c^2 for a transition function) while achieving state-of-the-art performance.

Noise-Robust Loss Functions Prior work examines how regularization techniques can be adapted to the noisy labels setting, addressing issues related to overfitting on noisy data [Menon et al. (2019); Lukasik et al. (2020); Englesson & Azizpour (2021)]. Recently, label smoothing, and in some cases negative label smoothing, were found to improve the accuracy on both correctly labeled and mislabeled data [Lukasik et al. (2020); Wei et al. (2022a)]. With this approach, the observed labels are perturbed by a small, pre-determined value, with all labels receiving the same perturbation at every training epoch. Follow-up work found that, instead of applying the same perturbation at each epoch, adding a small amount of Gaussian stochastic label noise (SLN) at each epoch resulted in further improvements, as it helped to escape from local optima [Chen et al. (2021)]. However, these approaches were most beneficial in the context of augmenting existing methods that identify mislabeled instances (e.g., stochastic label noise is applied to instances that are identified as correctly labeled by filtering approaches), and thus, potentially suffer from the same limitations. Alternatively, recent work has also proposed perturbing the features instead of the labels to encourage consistency in the model's predictions [Englesson & Azizpour (2021)], though mainly in the context of instance-independent label noise. Others have proposed noise-robust variations of cross entropy loss [Feng et al. (2020); Wang et al. (2021a)] that generally relied on assumptions like the memorization effect.

Label Noise and Fairness Label noise has also been addressed within the fairness literature, where it is particularly relevant to tasks that involve predicting outcomes for people (e.g., predicting cardiovascular disease for male and female patients). When the frequencies at which subgroups (defined by a sensitive attribute) appear are different within a dataset, past work has shown that common approaches addressing label noise can increase the prediction error for minority groups (i.e., rarer subgroups) [Liu (2021)]. To date, recent work proposed to re-weight instances from subgroups during training where model performance is poorer [Jiang & Nachum (2020)]. Other lines of work use peer loss [Liu & Guo (2020)] within subgroups [Wang et al. (2021b)]. However, they assume that label noise is random or that label noise depends on only the sensitive attribute. Our proposed approach applies in the more general case in that it allows for label noise to be affected by multiple features, instead of only the sensitive attribute. We train with weighted cross-entropy loss, where weights are based on predicted label correctness rather than performance on the observed labels.

Anchor Points for Addressing Label Noise Much also remains to be explored in utilizing anchor points. To date, anchor points are generally used to learn a transition function [Xia et al. (2019; 2020); Berthon et al. (2021)] or for label correction directly [Wu et al. (2021)]. Here, we use them to 1) pre-train the model, and 2) predict label correctness. The first part builds from work in semi-supervised learning [Cascante-Bonilla et al. (2021)], which has shown improvements from pre-training on labeled data. The second part is similar to a transition function, but differs in that we use the correctness predictions to re-weight the loss rather than adjust the predictions. Generally, anchor-based approaches maintain model fairness by implicitly assuming that the anchor points are representative of the target population. Our approach also uses this assumption, and we empirically explore how model performance changes when the anchor points are biased (i.e., not representative), since it may be easier to obtain correct labels for specific subgroups.

6 Conclusion

We introduced a novel approach for learning with instance-dependent noisy labels. Our two-stage approach utilizes the complete dataset without assuming independence between the features and label noise while taking advantage of a small set of anchor points. On synthetic and real datasets, we showed that the proposed approach led to improvements over state-of-the-art baselines with respect to balancing discriminative performance and model fairness. Our approach is not without limitations. In our experiments, we demonstrated that the success of the approach depends on the size and representativeness of the anchor set. Nonetheless this work brings together and builds on key aspects from the instance-dependent noisy labels literature, highlighting issues related to fairness.

REFERENCES

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

Devansh Arpit, Stanislaw K Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron C Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, 2017.

HeeSun Bae, Seungjae Shin, Byeonghu Na, JoonHo Jang, Kyungwoo Song, and Il-Chul Moon. From noisy prediction to true label: Noisy prediction calibration via generative model. In *International Conference on Machine Learning*, pp. 1277–1297. PMLR, 2022.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, pp. 825–836. PMLR, 2021.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pp. 1062–1070, 2019.

Pengfei Chen, Guangyong Chen, Junjie Ye, Jingwei Zhao, and Pheng-Ann Heng. Noise against noise: stochastic label noise helps combat inherent label noise. In *International Conference on Learning Representations*, 2021.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2020a.

Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International Conference on Machine Learning*, pp. 1789–1799. PMLR, 2020b.

Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems. *arXiv e-prints*, pp. arXiv–2011, 2020.

Dheeru Dua and Casey Graf. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2017.

Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.

Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *International Joint Conferences on Artificial Intelligence*, pp. 2206–2212, 2020.

Dragan Gamberger, Nada Lavrač, and Sašo Džeroski. Noise elimination in inductive concept learning: A case study in medical diagnosis. In *International Workshop on Algorithmic Learning Theory*, pp. 199–212. Springer, 1996.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pp. 4006–4016. PMLR, 2020.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Zhimeng Jiang, Kaixiong Zhou, Zirui Liu, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. An information fusion approach to learning with instance-dependent label noise. In *International Conference on Learning Representations*, 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Taghi M Khoshgoftaar and Pierre Rebours. Generating multiple noise elimination filters with the ensemble-partitioning filter. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004.*, pp. 369–375. IEEE, 2004.

Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Martin Ladouceur, Elham Rahme, Christian A Pineau, and Lawrence Joseph. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics*, 63(1): 272–279, 2007.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*. PMLR, 2022.

Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pp. 6725–6735. PMLR, 2021.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pp. 6226–6236. PMLR, 2020.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*. PMLR, 2020.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pp. 6543–6553. PMLR, 2020.

Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013.

Nancy N Maserejian, Carol L Link, Karen L Lutfey, Lisa D Marceau, and John B McKinlay. Disparities in physicians' interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment. *Journal of women's health*, 18(10):1661–1667, 2009.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2019.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*, 2020.

Jenna M Norton, Kaltun Ali, Claudine T Jurkovitz, Krzysztof Kiryluk, Meyeon Park, Kensaku Kawamoto, Ning Shang, Sankar D Navaneethan, Andrew S Narva, and Paul Drawz. Development and validation of a pragmatic electronic phenotype for ckd. *Clinical Journal of the American Society of Nephrology*, 14(9):1306–1314, 2019.

Jeeheh Oh, Jiaxuan Wang, Shengpu Tang, Michael W. Sjoding, and Jenna Wiens. Relaxed parameter sharing: Effectively modeling time-varying relationships in clinical time-series. In *MLHC*, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21, 2008.

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020.

Jaree Thongkam, Guandong Xu, Yanchun Zhang, and Fuchun Huang. Support vector machine for outlier detection in breast cancer survivability prediction. In *Asia-Pacific Web Conference*, pp. 99–109. Springer, 2008.

Donna Tjandra, Raymond Q Migrino, Bruno Giordani, and Jenna Wiens. Cohort discovery and risk stratification for alzheimer's disease: an electronic health record-based approach. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1):e12035, 2020.

Sudhi G Upadhyaya, Dennis H Murphree Jr, Che G Ngufor, Alison M Knight, Daniel J Cronk, Robert R Cima, Timothy B Curry, Jyotishman Pathak, Rickey E Carter, and Daryl J Kor. Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 1(1):100–110, 2017.

Sofie Verbaeten. Identifying mislabeled training examples in ilp classification problems. In *Proceedings of twelfth Belgian-Dutch conference on machine learning*, pp. 1–8, 2002.

Deng-Bao Wang, Yong Wen, Lujia Pan, and Min-Ling Zhang. Learning from noisy labels with complementary loss functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10111–10119, 2021a.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 526–536, 2021b.

Qizhou Wang, Bo Han, Tongliang Liu, Gang Niu, Jian Yang, and Chen Gong. Tackling instance-dependent label noise via a universal probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10183–10191, 2021c.

Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning*. PMLR, 2022a.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022b.

Yichen Wu, Jun Shu, Qi Xie, Qian Zhao, and Deyu Meng. Learning to purify noisy labels via meta soft label corrector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10388–10396, 2021.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in neural information processing systems*, pp. 6838–6849, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33, 2020.

Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*, 2022.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: An information-theoretic noise-robust loss function. *Advances in Neural Information Processing Systems*, 32, 2019.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*. PMLR, 2022.

Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pp. 10789–10798. PMLR, 2020a.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in Neural Information Processing Systems*, 33, 2020b.

Jung Eun Yoo, Dong Wook Shin, Kyungdo Han, Dahye Kim, Seung-Pyo Lee, Su-Min Jeong, Jinkook Lee, and SangYun Kim. Blood pressure variability and the risk of dementia: a nationwide cohort study. *Hypertension*, 75(4):982–990, 2020.

Xuchao Zhang, Xian Wu, Fanglan Chen, Liang Zhao, and Chang-Tien Lu. Self-paced robust learning for leveraging clean labels in noisy data. In *AAAI*, pp. 6853–6860, 2020.

Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pp. 11447–11457. PMLR, 2020.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning*, pp. 27412–27427. PMLR, 2022a.

Zhaowei Zhu, Jialu Wang, and Yang Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning*. PMLR, 2022b.

APPENDIX

A PROPOSED APPROACH: PROPOSED AND CLEAN LABEL LOSS

Here, we show that minimizing the proposed loss $\mathcal{L}'_\theta$ from step 2 of the proposed method is equal to minimizing cross entropy on the clean labels in expectation.

$$\mathcal{L}'_\theta = \frac{-1}{|\overline{A}|} \sum_{k=1}^{G} \sum_{i \in \overline{A} \cap g_k} \frac{1}{1-\hat{r}_k} \sum_{j=1}^{c} \hat{\beta}_\phi^{(i)} \mathbb{I}\left(\tilde{y}^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)$$

$$\mathbb{E}\left[\sum_{k=1}^{G} \sum_{i \in \overline{A} \cap g_k} \frac{1}{1-\hat{r}_k} \sum_{j=1}^{c} \hat{\beta}_\phi^{(i)} \mathbb{I}\left(\tilde{y}^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)\right]$$

$$= \sum_{k=1}^{G} \sum_{i \in \overline{A} \cap g_k} \frac{1}{1-\hat{r}_k} \sum_{j=1}^{c} \mathbb{E}\left[\hat{\beta}_\phi^{(i)} \mathbb{I}\left(\tilde{y}^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)\right]$$

$$= \sum_{k=1}^{G} \sum_{i \in \overline{A} \cap g_k} \frac{1}{1-\hat{r}_k} \sum_{j=1}^{c} (1-\hat{r}_k) \mathbb{I}\left(y^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)$$

$$= \sum_{k=1}^{G} \sum_{i \in \overline{A} \cap g_k} \sum_{j=1}^{c} \mathbb{I}\left(y^{(i)} == j\right) log\left(\hat{y}_j^{(i)}\right)$$

As a reminder, each group g_k is then associated with estimated noise rate $\hat{r}_k = \frac{1}{|g_k|} \sum_{i \in g_k} 1 - \hat{\beta}_\phi^{(i)}$ and estimated clean (i.e., correct) rate $1 - \hat{r}_k = \frac{1}{|g_k|} \sum_{i \in g_k} \hat{\beta}_\phi^{(i)}$. We can express the noise and clean rates in terms of $\hat{\beta}_\phi^{(i)}$ since

$$1 - r_k = \frac{1}{|g_k|} \sum_{i \in g_k} \mathbb{I}\left(\tilde{y}^{(i)} == y^{(i)}\right)$$

$$= P(y == \tilde{y} | \tilde{y}, \mathbf{x}) \; for \; a \; randomly \; chosen \; instance \; in \; g_k$$

$$= \frac{1}{|g_k|} \sum_{i \in g_k} P(y^{(i)} == \tilde{y}^{(i)} | \tilde{y}^{(i)}, \mathbf{x}^{(i)})$$

where r_k and $1 - r_k$ are the actual noise and clean rates within group k, respectively. Therefore, since $\hat{\beta}_\phi$ is trained to predict $P(y == \tilde{y} | \tilde{y}, \mathbf{x})$, we estimate the noise and clean rates using $\hat{\beta}_\phi$.

B PREPROCESSING DETAILS

Here, we provide more detail on real dataset pre-processing.

B.1 MIMIC-III

Data were processed using the FlexIble Data Driven pipeLinE (FIDDLE), [Tang et al. (2020)], a publicly available pre-processing tool for electronic health record data. We used the same features as [Tang et al. (2020)] for our tasks. More information can be found at https://physionet.org/content/mimic-eicu-fiddle-feature/1.0.0/.

B.2 ADULT

Although, we used a pre-processed version of this dataset, we omitted features pertaining to education, work type, and work sector to make the task more difficult. More specifically, in the file 'headers.txt' at the repository mentioned in Footnote 1, we kept all features beginning with 'age', 'workclass', 'education', 'marital status', and 'occupation'. We also kept the 'Sex_Female' feature. Values were normalized for each feature to have a range of 0-1 by subtracting by the minimum value observed among all individuals and dividing by the range.

Table 2: For each dataset, we list the range of hyperparameters considered for each dataset. For each hyperparameter, the lower bound is shown in the top row, and the upper bound is shown in the bottom row. For hyperparameters we did not tune, only one row is shown.

Hyperparameter	Synthetic	MIMIC-ARF	MIMIC-Shock	Adult	COMPAS
Learning Rate	0.00001	0.00001	0.000001	0.00001	0.0001
	0.01	0.001	0.001	0.01	0.05
L2 Constant	0.0001	0.000001	0.0001	0.0001	0.0001
	0.1	0.01	0.1	0.1	0.01
Layer Size	10	500	500	100	10
Filter Threshold	0.40	0.50	0.50	0.50	0.50
	1.00	1.00	1.00	0.90	0.90
Noise Added	0.00001	0.00001	0.00001	0.0001	0.0001
	0.01	0.001	0.001	0.001	0.01
Number of Parts	1	1	1	1	1
	10	10	10	10	10
α_{GPL}	0.01	0.1	0.001	0.01	0.01
	1.0	1.0	1.0	1.0	1.0
$\alpha_{1Proposed}$	0.1	0.1	0.01	0.1	0.01
	10.0	10.0	10.0	10.0	10.0
$\gamma_{Proposed}$	0.1	0.1	0.01	0.1	0.01
	10.0	10.0	10.0	10.0	10.0
$\alpha_{2Proposed}$	0.1	0.1	0.01	0.1	0.01
	10.0	10.0	10.0	10.0	10.0

B.3 COMPAS

Although, we used a pre-processed version of this dataset, we omitted the feature 'score_factor' (i.e., the risk score for recidivism from the ProPublica model) to make the task more difficult. Values were normalized for each feature to have a range of 0-1 by subtracting by the minimum value observed among all individuals and dividing by the range.

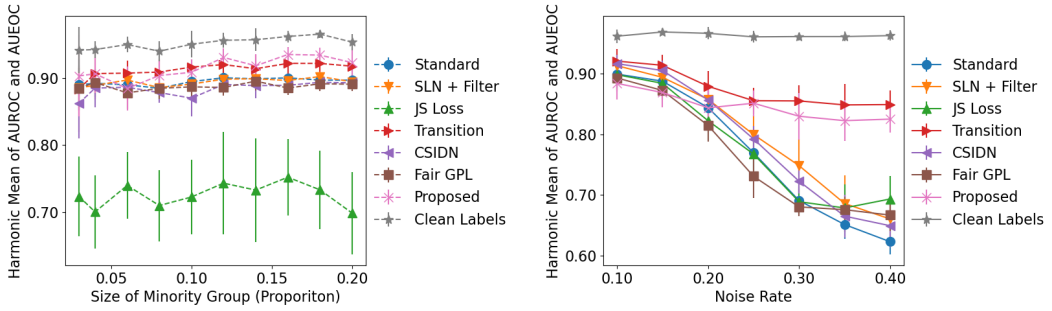
C Network and Training Details

Here, our ranges of hyperparameters and implementation choices for the proposed network. All networks were trained on Intel(R) Xeon(R) CPUs, E7-4850 v3 @ 2.20GHz and Nvidia GeForce GTX 1080 GPUs. All layers were initialized with He initialization from a uniform distribution. We also include are source code as part of the submission in the supplemental attachment.

C.1 Hyperparameter Values Considered

Here, we show the range of values we considered for our random search. More details are provided in **Table 2**. For any hyperparameters associated with the Adam optimizer not mentioned above, we used the default values. We divide our training data into five batches during training. Fandom seeds (for Pytorch, numpy, and Python's random) were initialized with 123456789.

Not all hyperparameters were used with each approach. 'Filter Threshold' and 'Noise Added' were only used with the baseline SLN + Filter. Here, Filter Threshold refers to the minimum value of the predicted probability of the observed label for an instance to be considered 'correctly labeled'. For example, if Filter Threshold=0.5, then all examples whose predicted probability for the observed label is at least 0.5 are considered 'correct' and used during training. 'Number of Parts' was only used with the baseline Transition. 'α_{GPL}' was only used with the baseline Fair GPL, and '$\alpha_{Proposed}$' was only used with the proposed method. Here, $\alpha_{Proposed}$ corresponds to the term α that was used to balance $\mathcal{L}_\theta$ and $\mathcal{L}_\phi$. We refer to it as $\alpha_{Proposed}$ in this section to distinguish it from the α value used by the baseline Fair GPL.



(a) Comparison of the proposed approach to baselines when varying the size of the minority set in the whole dataset. The proposed approach outperforms the baselines on minority sizes of 10% or higher.

(b) Comparison of the proposed approach to baselines with uniform random noise. Overall, the proposed approach is robust to the noise rate and is competitive with the baselines.

Figure 4: Extra experiments.

C.2 Network Details

For the overall architecture, we used a feed forward network with two hidden layers. The auxiliary β prediction component was also implemented with two feed forward layers. All layer sizes are as described in **Table 2**. In addition, we used the ReLU activation function. The complete implementation can be found in the attached code.

C.3 Source Code

See the attached zip file. We will make the code public upon publication.

D Implementation Details

For each dataset, we randomly split the data into 80/20% training/test, ensuring that data from the same individual did not appear across splits. For the Adult dataset, we used the test set provided and randomly selected 1,000 individuals from the training set. We then randomly selected 10% of the training data for all datasets except MIMIC-III from each subgroup to be anchor points, thereby ensuring that they were representative of the overall population. For the MIMIC-III dataset, 2% from each subgroup were selected as anchor points due to the larger size of the dataset. Then, for all datasets, half of the anchor points were then set aside as a validation set to use during training for early stopping and hyperparameter selection, while the other half remained in the training set. Later, in our experiments, we evaluate when the anchor set size varies and when the anchor set is biased. **All approaches (i.e., baselines and proposed) were given access to ground truth labels for data in the anchor set so that some approaches did not have an unfair advantage.**

All models were trained in Python3.7 and Pytorch1.7.1 [Paszke et al. (2017)], using Adam [Kingma & Ba (2014)]. Hyperparameters, including the learning rate, L2 regularization constant, and objective function scalars (e.g., α), were tuned using random search, with a budget of 20. We used early stopping (patience=10) based on validation set performance, where we aimed to maximize the HM. We report results on the held-out test set, showing the mean and standard deviation over 10 replications.

E Results: Additional Experiments (Synthetic Data)

We examined performance when we varied the size of the minority group (**Figure 4a**) and found that our approach is no longer beneficial at minority groups smaller than 10% of the dataset. This presents a limitation for future work. We also examined uniform random noise (**Figure 4b**). At low noise rates, all approaches achieved similar performance. At a noise rate of 40%, all approaches

outperformed the standard approach, which did not account for label noise. Our approach was still robust to the noise rate and competitive with the baselines.