Minimax Risk Classifiers for Mislabeled Data: a Study on Patient Outcome Prediction Tasks

Lucia Filippozzi LUCIA.FILIPPOZZI@UNITN.IT
Department of Mathematics
University of Trento
Trento, Italy

Santiago Mazuelas SMAZUELAS@BCAMATH.ORG
BCAM — Basque Center for Applied Mathematics
IKERBASQUE, Basque Foundation for Science
Bilbao, Spain

Iñigo Urteaga IURTEAGA@BCAMATH.ORG
BCAM — Basque Center for Applied Mathematics
IKERBASQUE, Basque Foundation for Science
Bilbao, Spain

Abstract

Healthcare datasets are often impacted by incorrect or mislabeled data, due to imperfect annotations, data collection problems, ambiguity, and subjective interpretations. Incorrectly classified data, referred to as "*noisy labels*," can significantly degrade the performance of supervised learning models. Namely, noisy labels hinder the algorithm's ability to accurately capture the true underlying patterns from observed data. More importantly, evaluating the performance of a classifier when only noisy test labels are available is a significant complication. We hereby tackle the challenge of trusting the labeling process both in training and testing, as noisy patient outcome labels in healthcare raise methodological and ethical considerations. We propose a novel adaptation of Minimax Risk Classifiers (MRCs) for data subject to noisy labels, both in training and evaluation. We show that the upper bound of the MRC's expected loss can serve as a useful estimator for the classifier's performance, especially in situations where clean test data is not available. We demonstrate the benefits of the proposed methodology in healthcare tasks where patient outcomes are predicted from mislabeled data. The proposed technique is accurate and stable, avoiding overly optimistic assessments of prediction error, a significantly harmful burden in patient outcome prediction tasks in healthcare.

1. Introduction

Healthcare datasets are often affected by noisy and mislabeled data, due to practical challenges ranging from the complexities of daily clinical practice, to the intricacies of the healthcare process. Data collection issues, ambiguity and subjective interpretations, as well as annotations and codes driven by billing purposes —instead of clinical judgements— are amongst the most significant reasons for low data quality in healthcare (Kompa et al., 2021).

This work addresses the particular healthcare challenge of dealing with *noisy labels for patient outcome prediction*, which directly hinders the success of *training* and *evaluating* machine learning (ML) models in the clinical practice, and the promises of precision medicine.

Addressing the issue of noisy outcomes in medical ML applications is crucial, as incorrect or biased predictions can lead to misdiagnoses, improper treatment decisions, and compromised patient care. We motivate and illustrate the challenges of noisy labels in healthcare with two well-studied ML for healthcare tasks: Intensive Care Unit (ICU) mortality prediction and the use of mammography for early detection of breast cancer.

The sensitivity of mammography for early cancer detection varies widely, with up to 3 out of 4 recommended biopsies later being judged unnecessary (Elter et al., 2007). Amongst the factors for this low predictive value are the wide variation in breast tissue density among subjects, as well as the intra- and inter-radiologist labelling variability (Baker et al., 1996; Elmore et al., 2003). Hence, the difficulty in assessing mammographies for tumor diagnosis (Baker et al., 2004), and the need to carefully consider the uncertain nature of mammography labels: not all positively labeled mammographies unequivocally indicate malign tumors, i.e., these are often noisy labels.

The use of ML for the ICU mortality prediction problem is also a task prone to mislabeling, as a consequence of the complex medical conditions and interventions involved (Choi et al., 2022). Due to the increased mortality rate among ICU patients, predicting patient outcomes in the ICU is a topic of continuous investigation, from diverse severity index definitions (Patel and Grant, 1999) to the development of a variety of predictive ML models (Silva et al., 2006; Kim et al., 2011; Taylor et al., 2015; Kang et al., 2020; Chiew et al., 2020; Abad and Lee, 2021; Cohen et al., 2021; Choi et al., 2022). Accurate outcome prediction in the ICU is crucial not only for informed clinical decision-making, but also for guiding the allocation of healthcare resources, such as ICU beds, in an optimal and ethically responsible manner (Chiew et al., 2020). Predictive models in the ICU aim not only at identifying patients at high risk of mortality, but also those who can get discharged. Early ICU discharge is desirable because it shortens the time spent in it (reducing the likelihood of infections and healthcare costs), but often increases the probability of ICU readmissions and post-discharge unanticipated death (Chrusch et al., 2009; Niven et al., 2014; Pilcher et al., 2007). Blindly trusting the labelling process within a single ICU stay window is one of the factors causing uncertain mortality labels: patient-discharges do not unequivocally indicate survival of the patient. Current ICU mortality predictive efforts mostly focus on disentangling survival from readmission, e.g., by (re-)training ML models to predict a distinct set of classes: death, readmitted, survived (Campbell et al., 2008; Badawi and Breslow, 2012; de Hond et al., 2023). These approaches require extending the data collection process beyond a single ICU stay and continuous relabelling —procedures costly in time and resources. More importantly, they are always subject to the unavoidable uncertainty of when the survival label must be reconsidered.

We recall that the uncertainty on healthcare labels (e.g., ICU mortality and mammographic outcomes) affects both the training and the evaluation process of ML pipelines with pre-collected healthcare datasets. Contrary to existing approaches that ignore potential mislabeling in the data or resort to its continuous relabelling, we fully embrace the labelling uncertainty in patient outcome predictions, and propose a solution to help close the gap in training and evaluating ML predictive models based on noisy labels.

The general ML literature contains numerous supervised classification methods targeted to effectively manage noisy labels (Frenay and Verleysen, 2014), which we survey and summarize, along those specific to healthcare, in Section 2.2.

Many of these methods tackle the issue by adapting the learning process to account for noise, such as modeling label noise during training (Frenay and Verleysen, 2014). Among them, Natarajan et al. (2013); Patrini (2016); Patrini et al. (2017) have designed a corrected loss function that demonstrates robustness to label noise. Additionally, approaches like the one introduced by Northcutt et al. (2021) aim to purify the data, i.e., noisy labels are identified and managed (e.g., re-classified or removed) through Confident Learning (CL).

While considerable effort has been dedicated to the development of techniques to learn from noisy labels, the *evaluation of learning methods on noisy labels* appears to be under-estimated and under-explored, both in the general and healthcare specific ML literatures. Common ML practice operates by training algorithms on noisy data, subsequently conducting its performance evaluation on clean data (van den Hout and van der Heijden, 2002; Stempfel and Ralaivola, 2009; Natarajan et al., 2013; Patrini et al., 2017; Natarajan et al., 2018; Tripathi and Hemachandra, 2018).

This is a particularly significant gap, as it is common that the only samples available in practice —for training and evaluation— are all affected by noise, with such scenarios posing considerable challenges in healthcare applications. To the best of our knowledge, the only proposed technique for ML evaluation on noisy data relies on the construction of an unbiased error estimator, as briefly mentioned by Patrini et al. (2017, Section 5.1); however, no experimental evaluation was provided. An alternative to accommodating noisy labels in evaluation is to first cleanse the data, e.g., via methods as in Northcutt et al. (2021), and then to evaluate the ML model with the common error estimator using these (supposedly) clean labels.

In this work, we propose a robust supervised learning solution to learn predictive models of patient outcomes that *accommodate noisy labels in training and in evaluation*. We cast the learning task as a robust optimization problem, aimed at maximizing/minimizing a given objective (e.g., the classification loss) subject to certain constraints, defined via an uncertainty set. For distributionally robust approaches to standard supervised classification, the uncertainty set is related to the limited number of available training samples. For cases with noisy labels, such uncertainty set is impacted also by the additional lack of knowledge induced by unreliable supervision, i.e., the noisy labels.

We hereby devise a novel adaptation of the Minimax Risk Classifier (MRC) method to data subject to noisy labels, and empirically validate and evaluate its practical properties and theoretical guarantees. The proposed algorithm not only enables effective learning from noisy label data, but it also provides worst-case error probabilities. These probabilities serve as error estimates for the algorithm's predictive performance evaluation, particularly useful in scenarios where access to clean test data remains unattainable. To the best of our knowledge, this is the first robust predictive solution that can be trained and evaluated using noisy labels, of critical importance in healthcare.

The main contributions of this work are as follow:

- We present a new learning algorithm for MRCs, based on a bias-correction procedure, that effectively learns from *noisy training data*.
- We propose to use the worst-case probability of error —a byproduct of the learning process— to assess the performance of our algorithm on *noisy test data*.
- We show that our proposed estimator exhibits stability and accuracy, avoiding overly optimistic assessments that can be harmful, especially in medical applications.

Generalizable Insights about Machine Learning in the Context of Healthcare

The contribution and significance of this work in the context of Machine Learning for Healthcare are two-pronged. On the one hand, we bring the attention of ML practitioners onto the challenge of dealing with noisy labels, both in training and testing. Providing evaluation robustness in the presence of noisy labels is an often overlooked methodological challenge that, nonetheless, is of undeniable practical significance. In healthcare, it is at least very costly, if not impossible, to attain fully clean data for evaluation of real-world ML deployments. On the other hand, we present a novel MRC-based supervised learning method that accommodates noisy labels in training and testing. We provide empirical evidence that MRCs are not only an accurate methodology for classification tasks in healthcare, but a robust tool for learning from noisy label data that provides worst-case error probabilities, trustworthy even for data with noisy labels. The robustness of the proposed methodology and its performance guarantees are particularly useful in healthcare, where access to clean test data is challenging, costly and often unattainable. Resolving the issue of noisy outcomes in medical ML applications is crucial, because incorrect or biased predictions can lead to misdiagnoses, improper treatment decisions, and compromised patient care.

2. Preliminaries

Supervised classification uses instance-label pairs to determine a classification rule to assign a label for each new instance. Here we denote with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \dots |\mathcal{Y}|\}$ the set of *instances* and the set of *labels*, respectively; and with $\Delta(\mathcal{X} \times \mathcal{Y})$ the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$. We represent instance-label pairs as real vectors based on a given *feature mapping* $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$. We define such a mapping via vector representations $\Psi : \mathcal{X} \to \mathbb{R}^D$ of instances and one-hot encoding of the labels, i.e.,

$$\Phi(\mathbf{x}, y) = \mathbf{e}_y \otimes \Psi(\mathbf{x}) \;=\; \begin{bmatrix} \mathbb{1}\,(y = 1)\,\Psi(\mathbf{x}) \\ \vdots \\ \mathbb{1}\,(y = |\mathcal{Y}|)\,\Psi(\mathbf{x}) \end{bmatrix}, \tag{1}$$

where $\mathbf{e}_i$ denotes the i-th vector of the canonical basis of $\mathbb{R}^{|\mathcal{Y}|}$. Usual choices for the instance mapping function are the identity $\Psi(\mathbf{x}) = \mathbf{x}$, or more complex feature representations, like Random Fourier Features (RFFs) (Rahimi and Recht, 2007).

We use $T(\mathcal{X}, \mathcal{Y})$ to denote the set of classification rules $h : \mathcal{X} \to \Delta(\mathcal{Y})$, and $h(y|\mathbf{x})$ for the probability assigned to label $y \in \mathcal{Y}$ for instance $\mathbf{x} \in \mathcal{X}$. For every probabilistic classification rule h, a deterministic version can be defined as

$$h_d(y|\mathbf{x}) = \mathbb{1}\left(y = \arg\max_y h(\cdot|\mathbf{x})\right) . \tag{2}$$

As a consequence, it holds that $h_d(y|\mathbf{x}) \in \{0,\ 1\}$ and $h_d(\mathbf{x}) := h_d(\cdot|\mathbf{x}) = \mathbf{e}_{\overline{y}}$ for certain $\overline{y}$.

We denote with $\ell(h, (\mathbf{x}, y))$ the loss of rule h for data pair $(\mathbf{x}, y)$. In the remaining of this work, we assume ℓ to be the 01-loss, defined using the deterministic rule h_d as

$$\ell_{01}(h, (\mathbf{x}, y)) = 1 - h_d(y|\mathbf{x}) . \tag{3}$$

We denote with $\ell(h, p)$ the *expected loss* of the classification rule h with respect to $p \in \Delta(\mathcal{X} \times \mathcal{Y})$:

$$\ell(h, p) := \mathbb{E}_p \left[\ell(h, (\mathbf{X}, Y))\right] . \tag{4}$$

If $p^* \in \Delta(\mathcal{X} \times \mathcal{Y})$ is the true underlying distribution of the instance-labels pairs, then we denote $\ell_{01}(h, p^*) := \mathbb{E}_{p^*} \left[\ell_{01}(h, (\mathbf{X}, Y))\right]$. In particular, for a deterministic classifier h_d, this indicates its probability of error:

$$\ell_{01}(h_d, p^*) = \mathbb{P}(h_d(\mathbf{X}) \neq \mathbf{e}_Y) . \tag{5}$$

2.1. Noisy Labels

When label noise is present, training samples $\{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ follow a noisy-distribution $\tilde{p}$, typically referred to as the *corrupted distribution*, that differs from the true, noiseless, underlying distribution p^*. In the following, we utilize the common assumption of instance-independent noise and known label flipping probabilities (van den Hout and van der Heijden, 2002; Frenay and Verleysen, 2014; Natarajan et al., 2013; Patrini et al., 2017; Abad and Lee, 2021). In practice, these probabilities can be estimated using multiple methods (Liu et al., 2023; Northcutt et al., 2021; Liu and Tao, 2015; Xia et al., 2019; Li et al., 2021). We assume instance-independent noise in the labelling process —a strong assumption that may not always hold true in healthcare— to make the learning problem tractable, and as a starting point for our research.

Although a limiting assumption, instance-independent noise is reasonable in certain healthcare contexts (Abad and Lee, 2021). For instance, if classification labels indicate different health conditions, some classes may be harder to label correctly; e.g., an illness that is hard to diagnose or that can be easily confused with another. In these cases, we may know that in 10% of instances with true disease $Y = i$, experts incorrectly assess it with $Y = j$ —note that such prior knowledge can also provide estimates for the label noise probabilities. Besides, in the context of mortality prediction (Abad and Lee, 2021), it is often assumed that "*label uncertainty is class-conditional, and it can be identified based on the class labels, not the data [14]–[16]*". For more realistic healthcare scenarios where certain patient characteristics make the classification task inherently more challenging, one must accommodate instance-dependent noise —see the Limitations section for future directions on how to extend this work to instance-dependent noise settings.

In the instance-independent noise scenario, each original label y_i may be flipped to a different label category $\tilde{y}_i$ with some probability, determined by the transition matrix

$$T = \begin{bmatrix} \rho_{1,1} & \cdots & \rho_{1,|\mathcal{Y}|} \\ \vdots & \ddots & \vdots \\ \rho_{|\mathcal{Y}|,1} & \cdots & \rho_{|\mathcal{Y}|,|\mathcal{Y}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|} , \tag{6}$$

where $\rho_{i,j} := \mathbb{P}(\tilde{Y} = i | Y = j)$ denotes the probability of flipping the true label $y = j$ to $\tilde{y} = i$. For instance, in binary classification problems, we have transition matrix

$$T = \begin{bmatrix} 1 - \rho_1 & \rho_2 \\ \rho_1 & 1 - \rho_2 \end{bmatrix} \in \mathbb{R}^{2 \times 2}, \text{ with } \rho_1, \rho_2 \in [0, 1/2) . \tag{7}$$

2.2. Learning a classifier with Noisy Labels

We provide here an overview of existing techniques designed to learn classifiers with noisy labels. A first line of work relies on algorithms that are shown to be noise-robust, as surveyed by Frenay and Verleysen (2014). In these, label noise is not modeled, nor cleaned, before learning. For instance, Pechenizkiy et al. (2006) showed that feature extraction can help in obtaining classifiers that are somehow robust to the presence of noise. A different solution consists in cleansing the data: i.e., noisy labels are identified and directly dealt with (re-classified/removed), before the learning stage (Brodley and Friedl, 1999; Sáez et al., 2016; Northcutt et al., 2021). However, both noise-robust algorithms and filtering methods become, in general, inadequate when dealing with more complex cases of label noise.

There also exist algorithms that directly embed the noise process into the learning algorithm —in a similar vein as we do. These revolve mostly around the principle of bias correction, to mitigate the effects of corrupted labels. In van den Hout and van der Heijden (2002), misclassification was addressed in the context of Randomized Responses, where they employ matrix T in Equation (6) to correct noise-distorted responses. Both Natarajan et al. (2013) and Patrini et al. (2017) constructed unbiased estimators for the classification loss function adapted to the presence of noisy labels, to be used in an empirical risk minimization procedure. Their corrected loss is a linear combination of the loss values for each label, with coefficients derived from the terms in T^{-1}. Similar ideas have been developed by Stempfel and Ralaivola (2009), proving that it is possible to estimate noise-free slack errors using a modified version of the hinge loss; and by Liu and Tao (2015), who proved that any loss can be used for noisy-label classification if one leverages importance re-weighting. Further approaches include methods that do not assume knowledge of noise rates, such as loss-based methods (Zhang and Sabuncu, 2018; Englesson and Azizpour, 2021), and in particular, peer-loss methods (Liu and Guo, 2020).

In healthcare, where labeling requires domain expertise and suffers from high inter- and intra-observer variability, learning with noisy labels is an important field of research. There are significant efforts based on the aforementioned theory of noise-tolerant learning (Aslam and Decatur, 1996), which have been used, for example, to learn phenotypes using noisy training data (Agarwal et al., 2016). Beyond clinical phenotyping, medical imaging is another field in which labeling errors have been mitigated via a variety of techniques (Karimi et al., 2020). As images are often labeled by multiple experts, disagreement and inconsistent labeling occurs, for which many techniques have been proposed, see (Ju et al., 2022; Karimi et al., 2020) and references therein. Solutions to train models with access to both clean and noisy labels have also been proposed, such as Alternating Loss Correction (Boughorbel et al., 2018) or meta-learning-based ones (Ren et al., 2018). However, for these solutions to work, a separate dataset with clean labels is required, either for label corruption estimation (Boughorbel et al., 2018) or loss minimization (Ren et al., 2018).

On the contrary, we here target the more common healthcare scenario where clean labels are hardly attainable, and the data volume in itself will not compensate errors nor reduce disagreement, yet require robust learning and performance evaluation with noisy labels.

2.3. Evaluating a classifier with Noisy Labels

The natural choice to evaluate the performance of a classifier *on clean labels* is to estimate its probability of error, i.e., $\ell_{01}(h, p^*)$ defined in Equation (5).

When a test set with clean labels is available, $\ell_{01}(h, p^*)$ can be estimated with the empirical average over the available N_{te} test-samples

$$\text{CE} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \ell_{01}(h, (\mathbf{x}_i, y_i)) \ . \tag{8}$$

We will refer to this as the Classification Error (CE). Nonetheless, when dealing with noisy data where clean test data is practically unattainable (e.g., noisy outcome predictions in healthcare), label corruption biases the sample average above, resulting in an unreliable estimator for the probability of error. We will denote this biased estimator of $\ell_{01}(h, p^*)$ as the Biased Loss Estimator (BLE), computed as follows

$$\text{BLE} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \ell_{01}(h, (\mathbf{x}_i, \tilde{y}_i)) \ . \tag{9}$$

To the best of our knowledge, the only existing alternative for evaluation on noisy data relies on the construction of an unbiased error estimator (Patrini et al., 2017, Section 5.1) based on a corrected loss function that takes into account the characteristics of the labelling noise, which we denote as Unbiased Loss Estimator (ULE):

$$\text{ULE} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \tilde{\ell}_{01}(h, (\mathbf{x}_i, \tilde{y}_i)) \ , \tag{10}$$

where the modified loss function can be derived (see Appendix A), and it is defined as:

$$\tilde{\ell}_{01}(h, (\mathbf{x}_i, y_i)) = (T^{-1})_{1,y_i} \ell_{01}(h, (\mathbf{x}_i, 1)) + (T^{-1})_{2,y_i} \ell_{01}(h, (\mathbf{x}_i, 2)) \ . \tag{11}$$

This loss coincides exactly with the corrected loss by Natarajan et al. (2013); Patrini (2016); Patrini et al. (2017), discussed in Section 2.2.

Although ULE serves as an unbiased estimator of the actual error probability, making it a preferable choice over BLE, it is susceptible to significant variability, particularly when dealing with high levels of noise. This volatility is due to the presence of the inverse of the noise transition matrix in Equation (11): as the noise rates in T approach values nearing 0.5, the determinant of T decreases significantly, approaching zero. Consequently, the values of the inverse matrix become extremely sensitive to noise rate estimates.

An alternative strategy for evaluating ML performance with noisy labels involves utilizing label-cleansing tools to correct the corrupted labels, before evaluating the classifier; e.g., employing the library presented by Northcutt et al. (2021). These new, *assumed to be* cleansed labels, that we represent with $\hat{y}_i$, are then used to calculate the conventional probability of error, which we denote as the Loss Estimator (LE)

$$\text{LE} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} \ell_{01}(h, (\mathbf{x}_i, \hat{y}_i)) \ . \tag{12}$$

It's crucial to recognize that accurately learning the noise matrix, and effectively cleaning the labels can be very challenging, especially when dealing with particularly difficult datasets or high noise rates. We deem a dataset to be "*difficult*" when it is hard to obtain clean labels. This occurs not only when the noise-label rates are high, but also when it is challenging to develop accurate classification rules, even without label noise: i.e., for an inherently challenging classification task, due to limitations in observed data and the underlying class boundaries. In the latter scenario, it becomes almost impossible to distinguish a noisy label from a clean one, because predicting the actual label of each example is in itself very difficult. These difficulties may lead to overly optimistic assessments and, consequently, a possibly harmful error estimate.

In healthcare, where precision is of utmost importance, such assessment variability falls short of the desired standards, as it is imperative to maintain a reasonable level of confidence in the potential errors of the techniques used. We therefore propose a robust and stable method to assess ML performance evaluation with noisy labels.

3. A novel adaptation of MRCs for Noisy Labels

In this section, we present how to learn MRCs from data subject to noisy labels. After introducing the principles of MRCs, we describe how the proposed MRC-based solution provides robust estimate of the classifier's error probability, suitable for evaluating the algorithm's performance, even in scenarios where clean test data are not available.

MRCs are classification rules that minimize the worst-case expected loss, with respect to distributions in uncertainty sets that contain the true underlying distribution with high probability (Mazuelas et al., 2020, 2022, 2023).

We say that $h^{\mathcal{U}}$ is an MRC for the set $\mathcal{U}$ if it is a solution of the following minimax risk problem

$$h^{\mathcal{U}} \in \arg \min_{h \in T(\mathcal{X},\mathcal{Y})} \max_{p \in \mathcal{U}} \ \ell(h,p) \ , \tag{13}$$

with $\mathcal{U}$ defined as an uncertainty set of distributions given by expectation constraints

$$\mathcal{U} = \{p \in \Delta(\mathcal{X},\mathcal{Y}) \ : \ |\mathbb{E}_p\left[\boldsymbol{\Phi}(\mathbf{X},Y)\right] - \boldsymbol{\tau}| \preccurlyeq \boldsymbol{\lambda}\} \ , \tag{14}$$

where $\boldsymbol{\tau}$ is an estimator of $\mathbb{E}_{p^*}\left[\boldsymbol{\Phi}(\mathbf{X},Y)\right]$, and $\boldsymbol{\lambda} \succcurlyeq \mathbf{0}$ is a confidence vector that quantifies the mean vector component-wise error $|\mathbb{E}_{p^*}\left[\boldsymbol{\Phi}\right] - \boldsymbol{\tau}|$. The mean and confidence vectors can be obtained from the training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ as

$$\boldsymbol{\tau} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Phi}(\mathbf{x}_i, y_i) \ \text{ and } \ \boldsymbol{\lambda} = \lambda_0 \frac{\mathbf{s}}{\sqrt{n}} \ , \tag{15}$$

where $\mathbf{s}$ is the vector of the component-wise sample standard deviations of $\{\boldsymbol{\Phi}(\mathbf{x}_i, y_i)\}_{i=1}^n$, and hyperparameter $\lambda_0 \in (0,1]$ controls the size of the set $\mathcal{U}$.

We denote with $R(\mathcal{U})$ the *Minimax Risk* against $\mathcal{U}$,

$$R(\mathcal{U}) := \min_{h \in T(\mathcal{X},\mathcal{Y})} \max_{p \in \mathcal{U}} \ \ell(h,p) \ . \tag{16}$$

Mazuelas et al. (2020, 2022, 2023) proved that the MRC $h^{\mathcal{U}}(\cdot|\mathbf{X})$ that solves Equation (16) can be obtained as a linear combination of the feature mapping, i.e., $h^{\mathcal{U}}(y|\mathbf{x}) = \boldsymbol{\Phi}(\mathbf{x}, y)^\top \boldsymbol{\mu}^*$. Without loss of generality, we hereafter use the deterministic MRC $h_d^{\mathcal{U}}(\mathbf{x}) = \mathbf{e}_{y^*}$, which classifies each instance $\mathbf{x}$ with the label y^* maximizing the probability $h^{\mathcal{U}}(\cdot|\mathbf{x})$,

$$y^* := \arg\max_{y \in \mathcal{Y}} h^{\mathcal{U}}(y|\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \boldsymbol{\Phi}(\mathbf{x}, y)^\top \boldsymbol{\mu}^* \ . \tag{17}$$

The coefficients $\boldsymbol{\mu}^*$ of the MRC's linear combination are learned as the solution to the convex optimization problem

$$\mathcal{P}_{\boldsymbol{\tau},\boldsymbol{\lambda}}: \qquad \min_{\boldsymbol{\mu}} \ \ 1 - \boldsymbol{\tau}^\top \boldsymbol{\mu} + \boldsymbol{\lambda}^\top |\boldsymbol{\mu}| + \varphi(\boldsymbol{\mu}) \ , \tag{18}$$

$$\text{with} \ \ \varphi(\boldsymbol{\mu}) = \max_{\mathbf{x} \in \mathcal{X}, \mathcal{C} \subseteq \mathcal{Y}} \frac{\sum_{y \in \mathcal{C}} \boldsymbol{\Phi}(\mathbf{x}, y)^\top \boldsymbol{\mu} - 1}{|\mathcal{C}|} \ ,$$

and the Minimax Risk $R(\mathcal{U})$ of the learned solution is given by

$$R(\mathcal{U}) = 1 - \boldsymbol{\tau}^\top \boldsymbol{\mu}^* + \boldsymbol{\lambda}^\top |\boldsymbol{\mu}^*| + \varphi(\boldsymbol{\mu}^*) \ . \tag{19}$$

We refer the interested reader to Theorem 2 in Section 2.2 of Mazuelas et al. (2023) for a complete proof of these results.

Noisy labels and MRCs. For training data $\{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ corrupted with T as described in Section 2.1, the sample average is a biased estimator of the target expectation $\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)]$. Hence, feeding the MRC with such biased estimator leads to unreliable classifiers: it results in a predictor for the noisy labels, not for the noise-free labels we are interested in.

We address the challenge of learning classifiers with noisy labels by proposing a *bias-correction* procedure for MRCs. Specifically, we compute an unbiased estimator of the true, target expectation $\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)]$ and its corresponding mean-vector accuracy estimator $\boldsymbol{\lambda}$, based on the label's noise characteristics given by T. Leveraging these, we devise a novel adaptation of MRCs that computes corrected estimators for $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ under mislabeling, to obtain *a reliable classifier with performance guarantees* for noise-free labels.

3.1. Learning MRCs with Noisy Labels

We present below the mathematical details of the proposed framework to learn MRCs for misslabeled data. We start with the definition of a correction matrix, of use for the derivation and presentation of our subsequent Theorem 1.

Let $\{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ be training samples corrupted via a non-singular noise matrix T; we define the *correction matrix* $(T^{-1} \otimes I_D)$ as

$$T^{-1} \otimes I_D \ = \ \begin{bmatrix} r_{11} I_D & \dots & r_{1|\mathcal{Y}|} I_D \\ r_{21} I_D & \dots & r_{2|\mathcal{Y}|} I_D \\ \vdots & & \vdots \\ r_{|\mathcal{Y}|1} I_D & \dots & r_{|\mathcal{Y}||\mathcal{Y}|} I_D \end{bmatrix} \ , \tag{20}$$

where r_{ij} denotes the component of matrix T^{-1} in the i-th row and j-th column, $\otimes$ is the Kronecker product, and I_D an identity matrix of the dimension of the vectors $\boldsymbol{\Psi}(\cdot)$ used to define Φ in Equation (1).

Theorem 1 *The estimator*

$$\boldsymbol{\tau} \;:=\; (T^{-1} \otimes I_D) \cdot \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Phi}(\mathbf{x}_i, \tilde{y}_i) \tag{21}$$

is unbiased for the true (noiseless) expectation $\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)]$*, i.e.* $\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)] = \mathbb{E}_{\tilde{p}}[\boldsymbol{\tau}]$.

In addition, its sample variance equals

$$\mathbf{V} := \frac{(T^{-1} \otimes I_D)\Sigma(T^{-1} \otimes I_D)^\top}{n} \;, \tag{22}$$

with Σ *the sample variance matrix of* $\{\boldsymbol{\Phi}(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$.

The estimator in Equation (21) is a weighted average of the feature mappings, based on the *correction matrix* defined by the Kronecker product $(T^{-1} \otimes I_D)$.

That is, the estimator can be rewritten as

$$\boldsymbol{\tau} \;=\; \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i, \tilde{y}_i) \quad \text{with} \quad \mathbf{z}(\mathbf{x}, y) = \sum_{i=1}^{|\mathcal{Y}|} (T^{-1})_{jy} \boldsymbol{\Phi}(\mathbf{x}, j) \;. \tag{23}$$

Hence, the unbiased estimator $\boldsymbol{\tau}$ is computed as the sample average of $\mathbf{z}(\mathbf{x}, y)$, which is a linear combination of the feature mapping $\boldsymbol{\Phi}(\mathbf{x}, \cdot)$ over all the possible observable labels with coefficients given by T^{-1}. This linear combination transforms the sum of the feature-label mappings $\sum_{i=1}^n \boldsymbol{\Phi}(\mathbf{x}_i, \tilde{y}_i)$ into an unbiased estimate of $\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)]$, effectively adjusting for label noise in the training data.

Proof Define the quantity $\mathbf{z}(\mathbf{x}, y) := \; ((T^{-1})^\top)_{y\cdot} \boldsymbol{\Phi}(\mathbf{x}, \cdot) \;=\; \sum_{j=1}^{|\mathcal{Y}|} (T^{-1})_{jy} \boldsymbol{\Phi}(\mathbf{x}, j)$.

It holds:

$$\begin{aligned}
\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)] \;&=\; \sum_{\mathbf{x},y} \boldsymbol{\Phi}(\mathbf{x}, y) p^*(\mathbf{x}, y) = \sum_{\mathbf{x}} \left(\sum_{y} \boldsymbol{\Phi}(\mathbf{x}, y) p^*(\mathbf{x}, y) \right) = \sum_{\mathbf{x}} \mathbf{p}^*(\mathbf{x}, \cdot)^\top \boldsymbol{\Phi}(\mathbf{x}, \cdot) \\
&\overset{(*)}{=} \sum_{\mathbf{x}} \left(T^{-1} \tilde{\mathbf{p}}(\mathbf{x}, \cdot) \right)^\top \boldsymbol{\Phi}(\mathbf{x}, \cdot) = \sum_{\mathbf{x}} \tilde{\mathbf{p}}(\mathbf{x}, \cdot)^\top \underbrace{(T^{-1})^\top \boldsymbol{\Phi}(\mathbf{x}, \cdot)}_{=:\; \mathbf{z}(\mathbf{x}, \cdot)} \\
&= \sum_{\mathbf{x},y} \mathbf{z}(\mathbf{x}, y) \tilde{p}(\mathbf{x}, y) \;=\; \mathbb{E}_{\tilde{p}}[\mathbf{z}(\mathbf{X}, Y)] \;,
\end{aligned}$$

where $(*)$ follows from the fact that $\tilde{\mathbf{p}}(\mathbf{x}, \cdot) \;=\; T \; \mathbf{p}^*(\mathbf{x}, \cdot)$ —see Lemma 4 in Appendix A.

As a direct consequence, we get that an unbiased estimator of $\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)]$ is given by

$$\boldsymbol{\tau} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i, \tilde{y}_i) \;, \text{ with } \; \mathbf{z}(\mathbf{x}, y) = \sum_{j=1}^{|\mathcal{Y}|} ((T^{-1})^\top)_{yj} \boldsymbol{\Phi}(\mathbf{x}, j) = \sum_{j=1}^{|\mathcal{Y}|} (T^{-1})_{jy} \boldsymbol{\Phi}(\mathbf{x}, j) \;. \tag{24}$$

Transforming Equation (24) into matrix form, we get Equation (21).

Let us now denote with Σ the variance matrix of data $\Phi(\mathbf{x}_i, \tilde{y}_i)$. By basic properties of variance-covariance matrices (see e.g., Petersen and Pedersen (2012)), it follows that the variance of $\boldsymbol{\tau}$ is

$$\mathbf{V} = \frac{(T^{-1} \otimes I_D)\Sigma(T^{-1} \otimes I_D)^\top}{n} \ .$$

■

Since $\boldsymbol{\lambda}$ in Equation (14) measures the mean vector's component-wise accuracy $|\mathbb{E}_{p^*}[\Phi] - \boldsymbol{\tau}|$, a natural choice for its estimator is the standard deviation estimate of $\boldsymbol{\tau}$

$$\boldsymbol{\lambda} = \lambda_0 \sqrt{diag(\mathbf{V})} \ , \tag{25}$$

where $\mathbf{V}$ is given by Equation (22).

Relation to other methods. The proposed MRC-based algorithm shares a common objective with the methods of Natarajan et al. (2013) and Patrini (2016); Patrini et al. (2017): the development of a bias-correction procedure for learning with noisy labels. However, they diverge in their design and execution. While our method devises an unbiased estimator for the feature mapping's expectation $\boldsymbol{\tau}$ as in Equation (21), Natarajan's and Patrini's approaches construct an unbiased estimator for their training losses, used for empirical risk minimization. The latter approaches result in a corrected loss function as in Equation (11), which is essentially a linear combination (with coefficients weighted according to label-noise probabilities) of the loss values associated with each observed (noisy) label. Instead, we incorporate noisy label information into the feature mapping's sufficient statistics, and keep the loss function intact —which is dealt with by the usual MRC optimization procedure.

3.1.1. Learning MRCs with Noisy Labels using estimated noise rates

The proposed MRC for noisy labels relies on the assumption of known noise rates, i.e., the matrix T is known. However, noise rates must often be estimated in practice, see discussion on potential methods for this in Section 2.1. We hereby study the impact of using an approximated $\widehat{T}$, and show that the proposed method is not severely affected, with corresponding empirical results provided in Section 5.3.

When using an approximate matrix $\widehat{T}$ instead of the true, yet unknown matrix T, the resulting estimate

$$\widehat{\boldsymbol{\tau}} = (\widehat{T}^{-1} \otimes I_D)\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Phi}(\mathbf{x}_i, \tilde{y}_i) \tag{26}$$

is biased. The next result bounds the relative error introduced by such a biased estimator under an approximated noise matrix $\widehat{T}$.

Theorem 2 *Let $\xi_{min}(T) \geq 0$ be the smallest singular value of matrix T. If matrix $\widehat{T}$ satisfies that $\|T - \widehat{T}\|_2 \leq \varepsilon\, \xi_{min}(T)/2$ for $0 < \varepsilon < 1$. Then, the estimator $\widehat{\boldsymbol{\tau}}$ in Equation (26) satisfies*

$$\frac{\|\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)] - \mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}]\,\|_2}{\|\mathbb{E}_{p^*}[\boldsymbol{\Phi}(\mathbf{X}, Y)\|_2} \leq \varepsilon \quad .$$

Proof Let A and B be matrices defined as $A = T \otimes I_D$ and $B = \widehat{T} \otimes I_D$, and m be the vector $m = \mathbb{E}_{p^*}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{\Phi}(\mathbf{x}_i, \tilde{y}_i)\right]$. Then, we have that $B\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] = m$ and, using Theorem 1, we also have that $A\mathbb{E}_{p^*}[\Phi(\mathbf{X}, Y)] = m$. Therefore,

$$
\begin{aligned}
& B\left(\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)] + \mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\right) \;=\; m \;=\; A\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)] \\
\Rightarrow\quad & (B-A)\,\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)] + (B-A+A)\left(\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\right) \;=\; 0 \\
\Rightarrow\quad & A\left(\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\right) = (A-B)\left(\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)] + \mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\right) \\
\Rightarrow\quad & \mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)] = A^{-1}(A-B)\left(\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)] + \mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\right) \;.
\end{aligned}
$$

Taking the norm-2 over the above expression, we have:

$$
\begin{aligned}
& \|\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2 \le \|A^{-1}\|_2\|A-B\|_2\left(\|\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2 + \|\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2\right) \\
\Rightarrow\quad & \|\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2 \;\le\; \frac{\|A^{-1}\|_2\|A-B\|_2}{1 - \|A^{-1}\|_2\|A-B\|_2}\|\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2 \;.
\end{aligned}
$$

Since the norm-2 of a matrix is given by its largest singular value, and the singular values do not change by doing the Kronecker product with an identity matrix, the following holds:

$$
\|A^{-1}\|_2 = \|T^{-1}\|_2 \le (\xi_{\min}(T))^{-1} \qquad \text{and} \qquad \|A-B\|_2 = \|T-\widehat{T}\|_2 \;.
$$

Hence,

$$
\frac{\|\;\mathbb{E}_{p^*}[\widehat{\boldsymbol{\tau}}] - \mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2}{\|\;\mathbb{E}_{p^*}[\Phi(\mathbf{X},Y)]\,\|_2} \le \frac{\varepsilon/2}{1-\varepsilon/2} \le \varepsilon \;,
$$

because the function $f(t) = t/(1-t)$ is increasing for $t \in (0,1)$, $(\xi_{\min}(T))^{-1}\|T-\widehat{T}\|_2 \le \varepsilon/2$, and $\varepsilon \le 1$ by assumption. ■

Theorem 2 shows that the error due to the usage of a matrix $\widehat{T} \neq T$ does not significantly increase with the approximation error $T-\widehat{T}$, as long as the transition matrix T is not nearly singular; i.e., the smallest singular value of T is not close zero. Such condition is satisfied as long as the noise rates are not very high. For instance, in the binary classification case, the singular values of T are 1 and $1-\rho_1-\rho_2$ so that the smallest singular value of T is close to zero only if both ρ_1 and ρ_2 are near 0.5 —a scenario with little practicality, with noise rates so high that the labeling process is dominated by random noise.

3.2. Evaluating MRCs with Noisy Labels

Given a learned classifier, evaluating it with noisy labels poses a significant challenge, as discussed in Section 2.3. Here, we highlight a useful property of the proposed algorithm, by showing how the worst-case error probability (i.e., the Minimax Risk) of an MRC is a valuable estimator of the classifier's performance —even when clean, noiseless test data is unavailable.

Let $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ be defined respectively as in Equations (21) and (25). Consider the corresponding $\mathcal{U}$ as defined in Equation (14). If $\mathcal{U}$ is not empty and $h^{\mathcal{U}}$ is a 01-MRC for the uncertainty set $\mathcal{U}$, it holds that

$$
\ell(h^{\mathcal{U}}, p^*) \le R(\mathcal{U}) + \left(\;|\boldsymbol{\tau}^* - \boldsymbol{\tau}|^\top - \boldsymbol{\lambda}^\top\;\right)|\boldsymbol{\mu}^*| \;, \tag{27}
$$

where $\boldsymbol{\tau}^* := \mathbb{E}_{p^*}[\mathbf{\Phi}(\mathbf{X}, Y)]$ denotes the *true* expectation, and $R(\mathcal{U})$ is the Minimax Risk in Equation (19). See a proof in Mazuelas et al. (2023).

Remark 3 *Equation (27) implies that, if* $|\boldsymbol{\tau}^* - \boldsymbol{\tau}| \prec \boldsymbol{\lambda}$*, or equivalently, if the true distribution* $p^* \in \mathcal{U}$ *(i.e., the uncertainty set* $\mathcal{U}$ *is not empty), then*

$$\ell(h^{\mathcal{U}}, p^*) \leq R(\mathcal{U}) \ .$$

Hence the Minimax Risk $R(\mathcal{U})$ *bounds the classifier's expected loss. Note that* $\boldsymbol{\lambda}$ *encodes a dependency with hyperparameter* λ_0*, as illustrated in Equations (15) and (25), that trade-offs the generality of the uncertainty set, with how tight the bound* $R(\mathcal{U})$ *is. Namely,* λ_0 *controls the width of the uncertainty set, which when it contains the underlying distribution, ensures that the worst-case expected loss upper bound holds. As one reduces* λ_0[1]*, the uncertainty set is shrinked, enabling a tighter upper bound as the minimum value of* $R(\mathcal{U})$ *in Equation (16) is similar to the minimum value* $\min_{h \in T(\mathcal{X},\mathcal{Y})} \ \ell(h, p^*)$*.*

Recalling that the expected 01-loss coincides with the probability of error in Equation (5), we can state that the Minimax Risk (worst-case error probability) obtained at learning with Equation (19) provides an upper bound for the probability of error. As a consequence, a learned MRC's worst-case error probability —computable as a byproduct of the MRC's learning algorithm itself— does not require access to clean test data. Hence, it can be used to assess the true error probability of a MRC in scenarios with only noisy data, useful in practice as demonstrated experimentally in Section 5.

4. Study Design

In order to evaluate the accuracy and robustness of the proposed methodology, we design an experimental study that combines real-world healthcare datasets with synthetic mislabeling[2]. To carefully investigate the impact of noisy labels on patient outcome prediction, we assume that labels in the datasets we describe in Section 4.1 are clean[3], and simulate noisy labels as described in Section 4.3. By simulating diverse label noise patterns, we have access to multiple mislabeled outputs and to the ground-truth labeling captured in real-world healthcare tasks. This enables us to assess how the proposed Noisy MRC generalizes across datasets and mislabelling regimes.

4.1. Real-world healthcare data

We evaluate the applicability of the presented method in the patient outcome prediction tasks discussed as motivating this work: i.e., an *ICU Mortality dataset* and a *Mammographic Mass* predictive dataset, which we describe below. Additional experiments on several UCI datasets (Kelly et al.) can be found in Appendices C and D.

4.1.1. ICU Mortality dataset

We tackle the challenge of predicting patient survival under random label noise, using data from the first 24 hours of an ICU, as provided by the MIT Global Open Source Severity of Illness Score (GOSSIS), made public in the context of the 2020 WiDS Datathon[4].

1. In the numerical experiments we present, we set $\lambda_0 = 1$.
2. We are not aware of publicly available real-world datasets with both true and noisy labels.
3. We acknowledge that these datasets may as well be subject to mislabeling. However, it is impossible to have access to their corresponding ground truth.
4. Available at https://www.kaggle.com/competitions/widsdatathon2020/data.

This *ICU Mortality* dataset is comprised of more than 130,000 hospital ICU patient visits, spanning a one-year timeframe. It contains 185 features ($\mathbf{X}$) —demographic data (e.g. gender, ethnicity, age, height, ...), lab results, and various medical measures— and `hospital_death` as the target variable (Y), describing the patients' outcome in the ICU: $Y = 0$ indicates survival; $Y = 1$, death. The dataset is unbalanced in its outcomes, with only 8.63% of deaths. Summary statistics of the dataset are described in Table 1.

	Original	Processed
N. patients	91.713	15.802
N. total features	185	148
N. continuous features	168	126
N. categorical (binary) features	23 (15)	22 (14)

Table 1: Characteristics of the original and pre-processed *ICU Mortality* dataset.

Guided by the results of Cohen et al. (2021), we preprocess the dataset as follows:

1. drop features that contain a percentage of missing values greater or equal to 80%;
2. drop features with zero standard deviation;
3. drop features with limited informative value (`hospital_id`, `encounter_id`, `patient_id`), as per the correlation of these features and the target variable;
4. drop data instances that have more than 70% of missing values in their features;
5. substitute the remaining missing data values with the median of each feature;
6. normalize each feature across patients;
7. perform one-hot encoding of categorical variables.

4.1.2. Mammographic Mass

We tackle the task of discriminating benign and malignant mammographic masses based on patient's age and BI-RADS attributes, using UCI's *Mammographic Mass*[5] dataset. The data consists of 961 patient records collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. It contains 5 features ($\mathbf{X}$) —age, BI-RADS assessment, three BI-RADS attributes— together with the target variable `severity` (Y) indicating whether the cancer is benign ($Y = 0$) or malignant ($Y = 1$).

4.2. Feature Choices

The methodology presented in Section 3 is applicable for any feature embedding. In this study, our goal is not to find the best features set for each healthcare task, but to demonstrate how the proposed methodology performs across different noisy-label problems, independently of the feature mapping. Hence, our feature choice is the same across datasets.

Instead of resorting to the basic bias and identity map $\Psi(\mathbf{x}) = [1,\ \mathbf{x}]^\top$, we embed the datasets with RFF mappings (Rahimi and Recht, 2007), defined as

$$\Psi(\mathbf{x}) = \left[\cos(\mathbf{w}_1^\top \mathbf{x}),\ \ldots,\ \cos(\mathbf{w}_K^\top \mathbf{x}),\ \sin(\mathbf{w}_1^\top \mathbf{x}),\ \ldots,\ \sin(\mathbf{w}_K^\top \mathbf{x})\right]^\top ,$$

5. Available at https://archive.ics.uci.edu/dataset/161/mammographic+mass.

with $\{\mathbf{w}_i\}_{i=1}^{K} \sim \mathcal{N}_d\left(\mathbf{0}, \frac{1}{\sigma^2} I\right)$, $K = 300$, $\sigma = \sqrt{d/2}$, for d the dimension of the original feature space $\mathcal{X}$. Classifiers based on RFF embeddings can provide highly nonlinear classification rules that approximate the solutions in a reproducing kernel Hilbert space (RKHS). In particular, the features described above correspond to a Gaussian kernel with scaling parameter $\sigma = \sqrt{d/2}$.

We note that alternative mapping functions could also be explored within the proposed method. For instance, if a practitioner wants to apply deep-learning models within the MRC-based framework (e.g., leveraging deep representation learning), one can readily use a pre-trained model for feature extraction, by using the last layer of the network as input-features to the MRC algorithm.

We chose RFF maps primarily because they do not require an additional learning stage, making the process more straightforward and efficient. Deep-learning-based feature representations might offer comparability to existing work, yet they introduce additional complexity and computational overhead, which we avoid in this study.

4.3. Noisy Labels

Real-world healthcare data may contain label imperfections due to diverse factors. For instance, in ICU readmission, post-discharge unanticipated deaths are mislabeled as survival (early ICU discharge). In mammography-based breast cancer screening, many of the malignant mammography labels are subsequently proven with biopsies to be actually benign. Since it is difficult to find real-world dataset with both true and noisy labels, we here create synthetic mislabeling of the datasets described in Section 4.1. We follow the most common label corruption model studied in the literature, i.e., *instance-independent* noise.

For results presented below, we randomly switch ground-truth labels based on two parameters, $\rho_1 = \mathbb{P}(\tilde{Y} = 0|Y = 1)$ and $\rho_2 = \mathbb{P}(\tilde{Y} = 1|Y = 0)$, as described in Equation (7). Namely, in mortality prediction tasks, ρ_1 represents the probability of erroneously labeling a patient as surviving ($y = 0$) if it later deceases ($y = 1$), e.g., post-discharge death; while ρ_2 represents the probability of erroneously labeling a patient as deceased ($y = 1$) when they survive ($y = 0$). We study larger values of ρ_1, and smaller values of ρ_2, reflecting that patient discharges do not unequivocally indicate long-term survival —a common challenge in ICUs as explained in the Introduction Section 1.

5. Experiments

We present below how different mislabelling patterns affect patient outcome predictions on the *ICU mortality* and *Mammographic mass* datasets described above.

We compare the proposed methodology to state-of-the-art baselines described in Section 5.1. We assess in Section 5.2 their predictive accuracy under a known noise matrix T assumption, to then scrutinize the more realistic scenario of an unknown noise matrix in Section 5.3.

5.1. Baselines and training set-up

We implement[6] and evaluate a set of variations of the proposed MRC and these baselines:

6. The developed codebase is provided at https://github.com/lucia2p2z/NoisyMRC.

- `Noisy MRC`: the proposed adaptation of MRC for *noisy* labels;
- `Naive MRC`: a naive MRC trained directly on *noisy* labels;
- `Oracle MRC`: an *oracle* MRC that is trained with *clean* labels, i.e., the ground truth —only available in a simulated scenario;
- `Noisy LR`: a Logistic Regression (LR) classifier, adapted to noisy labels by training with the loss proposed by Natarajan et al. (2013);
- `Naive LR`: a naive LR, trained on *noisy* labels;
- `Oracle LR`: an *oracle* LR, trained with *clean* labels, i.e., the ground truth —only available in a simulated scenario;
- `CleanLearning`: the learning method proposed by Northcutt et al. (2021), implemented using the `cleanlab.classification.CleanLearning` method in their Python library, `cleanlab`[7], which cleans out the errors in the labels while training;
- `Cleansed MRC`: a naive MRC trained on labels that have been previously cleansed using the method `find_label_issues` in `cleanlab`'s Python library;
- `Cleansed LR`: a naive LR trained on labels that have been previously cleansed using the method `find_label_issues` in `cleanlab`'s Python library.

Training and evaluation. We perform random train-test ($80\% - 20\%$) splits for each dataset, with label noise added according to specific ρ_1 and ρ_2 values for each healthcare dataset described in Section 4.1. We compute and present average classifier performances over 100 folds, as well as their variability across folds as standard deviations over runs.

We examine the influence of different training sizes, from smaller to larger (consecutive) portions of data, per-fold. After training, the learnt classifiers are evaluated on test data.

5.2. Results with T known

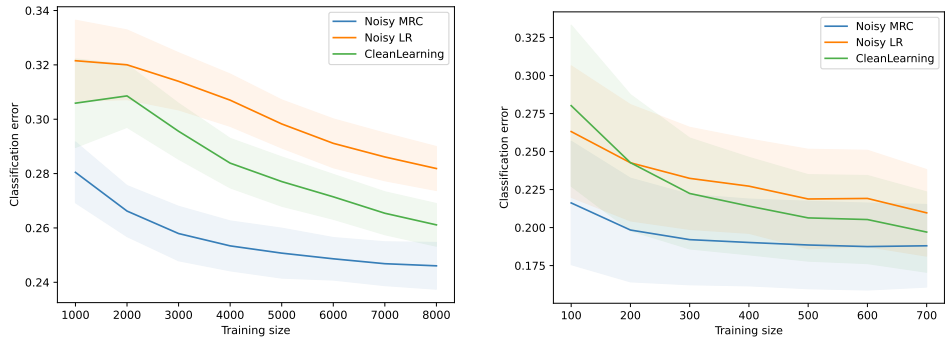
5.2.1. Evaluation on clean labels.

We present results comparing the proposed MRC's performance with other learning methods, trained on *noisy* labels and tested on *clean* labels. Note that this evaluation is only possible in our simulated noisy labeling scenario, where access to ground truth is possible.

Figure 1 illustrates the impact of different training sample sizes in predictive performance, for specific values of $\rho_1 = 0.25$, $\rho_2 = 0.10$. Table 2 displays the classification error of different techniques for fixed training sizes, with $\rho_2 = 0.10$ and different noise rates ρ_1. Additional results, with different noise rates and comparisons across methods, can be found in Appendix C.1.

From these results, we conclude that the proposed `Noisy MRC` outperforms other classifiers (namely, `CleanLearning` and `Noisy LR`) in terms of classification error, as it achieves the lowest error rate when learning from noisy labels across experiments.

7. Code available at https://pypi.org/project/cleanlab/ – Version downloaded in July 2023.



(*a*) *ICU Mortality* dataset (*b*) *Mammographic Mass* dataset

Figure 1: Experiment with T known and evaluation on clean labels. Classification error of classifiers for $\rho_1 = 0.25$, $\rho_2 = 0.10$. Our method (blue) outperforms baselines.

Datasets	Methods	$(\rho_1, \rho_2) = (0.1, 0.1)$	$(\rho_1, \rho_2) = (0.25, 0.1)$	$(\rho_1, \rho_2) = (0.40, 0.1)$
Mortality	`Oracle MRC`	0.244 ± 0.008	0.242 ± 0.008	0.242 ± 0.008
	`Oracle LR`	0.228 ± 0.007	0.227 ± 0.006	0.228 ± 0.007
	`Naive MRC`	0.247 ± 0.008	0.255 ± 0.008	0.502 ± 0.008
	`Noisy MRC`	0.246 ± 0.008	**0.246 ± 0.009**	**0.250 ± 0.008**
	`Naive LR`	**0.242 ± 0.008**	0.266 ± 0.007	0.325 ± 0.008
	`Noisy LR`	0.264 ± 0.008	0.282 ± 0.008	0.303 ± 0.010
	`CleanLearning`	0.243 ± 0.007	0.261 ± 0.008	0.300 ± 0.008
Mammogr.	`Oracle MRC`	0.187 ± 0.0256	0.183 ± 0.0269	0.188 ± 0.0272
	`Oracle LR`	0.195 ± 0.0247	0.194 ± 0.0257	0.194 ± 0.0266
	`Naive MRC`	0.188 ± 0.0259	0.197 ± 0.0292	0.410 ± 0.0603
	`Noisy MRC`	**0.186 ± 0.0264**	**0.188 ± 0.0271**	**0.198 ± 0.0275**
	`Naive LR`	0.207 ± 0.0230	0.213 ± 0.0289	0.369 ± 0.0339
	`Noisy LR`	0.206 ± 0.0238	0.209 ± 0.0285	0.262 ± 0.0287
	`CleanLearning`	0.199 ± 0.0229	0.197 ± 0.0264	0.289 ± 0.0364

Table 2: Experiments with T known and evaluation on clean labels. Average and standard deviation of classification errors for a fixed training size $n_{\text{train}} = 8.000$ (for *Mortality* dataset), $n_{\text{train}} = 700$ (for *Mammographic* dataset), different ρ_1 and $\rho_2 = 0.1$. Among the methods trained on noisy labels, `Noisy MRC` can more adequately adapt to noise in the labels.

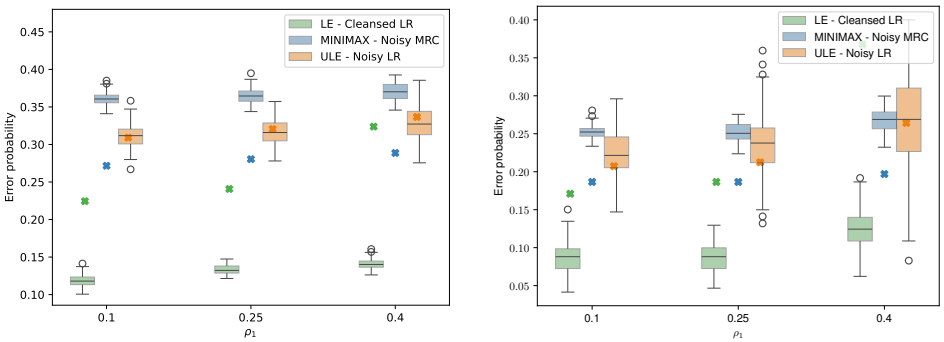
5.2.2. Evaluation on noisy labels.

We now report results when evaluating the performance of the classifiers in the most realistic case, where only noisy labels are available in both training and testing.

We compare `Cleansed LR`, `Noisy MRC` and `Noisy LR`, evaluated respectively with these error estimates:

- LE, computed as in Equation (12), to evaluate `Cleansed LR`;
- MINIMAX, as in Equation (19), to upper bound the classification error of `Noisy MRC`;
- ULE, computed as in Equation (10), to estimate the classification error of `Noisy LR`.

With these, we assess the effectiveness of our proposed error estimator; i.e., the Minimax Risk, as a robust upper bound of the classification error probability, described in Remark 3. We additionally assess the CE, as defined in Equation (8), which is only possible due to our synthetic misslabeling procedure, where we have access to the original, clean ground truth. Figure 2 shows the variability and accuracy of these error estimators, for fixed values of ρ_2 and training sizes (additional results can be found in Appendix C.2).



(*a*) *ICU Mortality* dataset (*b*) *Mammographic Mass* dataset

Figure 2: Experiment with T known and evaluation on noisy labels. Error measures of classifiers for fixed $n_{\text{train}} = 1.000$ (for *Mortality* dataset), $n_{\text{train}} = 700$ (for *Mammographic* dataset), $\rho_2 = 0.10$ and varying ρ_1. `Noisy MRC` (blue) avoids overly optimistic assessments when evaluated on noisy test data.

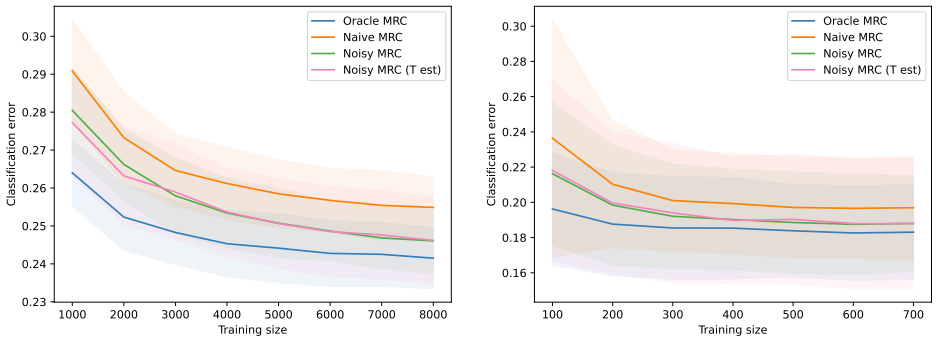
The boxplots in Figure 2 illustrate the distribution of the three error metrics —LE, MINIMAX, and ULE— across 100 repetitions, as ρ_1 (x-axis) increases. The **x** markers represent the mean value of the true classification error (CE) computed across all 100-folds, showcasing the accuracy and validity of the error measures. We observe that LE boxplots are always significantly lower than the true classifier error. Despite often being more accurate, ULE exhibits high estimation variability, particularly with larger noisy rates. In contrast, the proposed MINIMAX metric offers stability and prevents overly optimistic assessments.

5.3. Results with T unknown

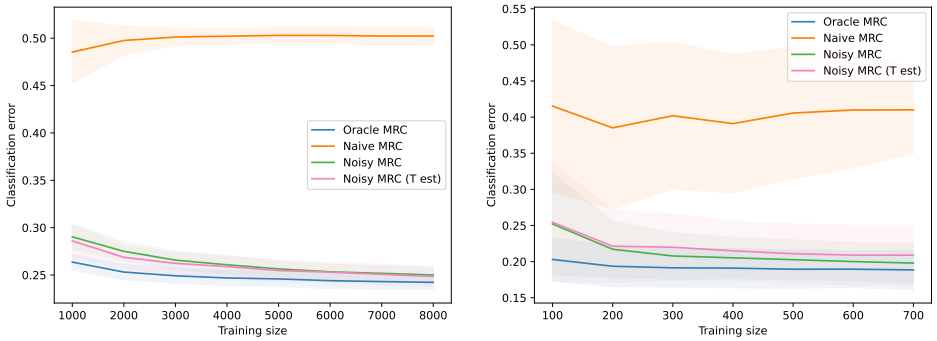
In real-life healthcare scenarios, one does not have knowledge of the noise matrix T corrupting the labels. In such case, one can either (**a**) estimate the matrix T, e.g., by employing Confident Learning (Northcutt et al., 2021), and then applying any of the baselines with the estimated $\hat{T}$; or (**b**) cleanse the labels with a method of choice, e.g., by using the method of Northcutt et al. (2021), and then applying the naive version of the classifiers.

We here provide results for both approaches when training on noisy data and evaluating in ground truth datasets. We leverage the `cleanlab` Python Library to either estimate T, or to cleanse the labels before the learning process. More precisely, we use the library's `estimate_noise_matrices` method for results in Figure 3, and the `find_label_issues` method for results in Figure 4, respectively.

Results in Figure 3 demonstrate that `Noisy MRC` exhibits accurate and robust performance even when trained with estimated noise matrix $\hat{T}$; i.e., it is a robust learning method even under a potentially misspecified noise matrix, as theoretically analyzed in Theorem 2.



(a) *ICU Mortality*: $\rho_1 = 0.25,\ \rho_2 = 0.10$ (b) *Mammographic*: $\rho_1 = 0.25,\ \rho_2 = 0.10$



(c) *ICU Mortality*: $\rho_1 = 0.40,\ \rho_2 = 0.10$ (d) *Mammographic*:$\rho_1 = 0.40,\ \rho_2 = 0.10$

Figure 3: Experiment with T unknown and evaluation on clean labels. Classification error of MRCs for different $\rho_1,\ \rho_2$. MRC performances trained with $\hat{T}$ estimated (pink) and with T known (green) are consistent.

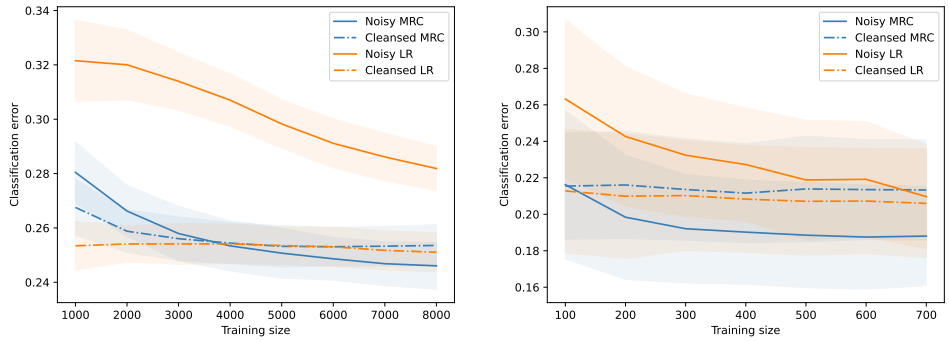
On the contrary, `Naive MRC` is noticeably unsuccessful under high noise rates, as shown in Figures 3(c) and 3(d). In Appendices D.1.1 and D.1.2 we provide results on the accuracy of the estimated transition matrix $\widehat{T}$, compared to the ground truth T, used in these experiments.

We additionally observe in Figure 4 that the performance of different algorithms when relying on `cleanlab` for label cleansing is very dependent on the corruption noise value ρ. When the noise rates are high (Figures 4(c) and 4(d)), it becomes challenging —if not impossible— for the `find_label_issues` method to identify and rectify incorrect labels appropriately —notice how both `Cleansed LR` and `Cleansed MRC` are unsuccessful.

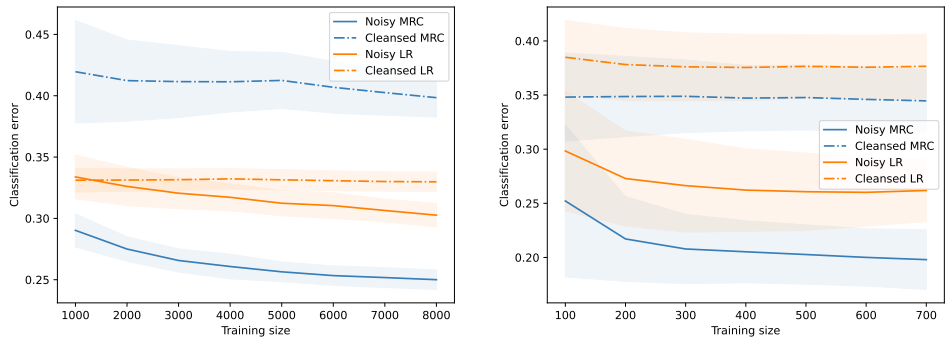
Additional plots for a wider range of misslabeling probabilities over datasets can be found in Appendices D.1 and D.2.

6. Discussion

Despite significant efforts devoted to developing ML techniques for learning from noisy labels, existing approaches lack a robust evaluation procedure in the presence of misslabeling, both in the general ML and healthcare-specific literature.



(*a*) *ICU Mortality*: $\rho_1 = 0.25,\ \rho_2 = 0.10$ (*b*) *Mammographic*: $\rho_1 = 0.25,\ \rho_2 = 0.10$



(*c*) *ICU Mortality*: $\rho_1 = 0.40,\ \rho_2 = 0.10$ (*d*) *Mammographic*: $\rho_1 = 0.40,\ \rho_2 = 0.10$

Figure 4: Experiment with T unknown and evaluation on clean labels. Classification error of methods trained on *cleansed* labels, for different $\rho_1,\ \rho_2$. Performance gap when using `cleanlab` is very dependent on the noise rates.

To address this gap, we introduced a robust MRC-based supervised learning approach capable of handling noisy labels, in both training and evaluation. We proposed to use worst-case error probabilities —a byproduct of learning MRCs— to assess the performance of the MRC algorithm under noisy testset labels. Our novel methodology extends MRCs to noisy label scenarios, offering practical performance improvements and theoretical assurances.

With a thorough empirical study based on real-world healthcare datasets subject to misslabeling, we demonstrated that the proposed algorithm is robust and accurate in the presence of noisy labels, and that the Minimax Risk estimator is useful in scenarios where access to clean test data is unattainable.

Presented results show the deficiencies of existing methods and the advantages of the proposed methodology: (*i*) LE heavily underestimates the true classification error —which is far from ideal, particularly in medical contexts; (*ii*) ULE is more accurate than LE, yet incurs in high estimation variability —especially for larger noisy rates; and (*iii*) the proposed MINIMAX metric not only provides more stable results than ULE, but it also prevents overly optimistic assessments —this overconfidence issue arises significantly when relying on label cleansing approaches, as measured via the LE.

Notably, Theorem 2 and results in Section 5.3 demonstrate that `Noisy MRC` exhibits great performance even when trained with an estimated noise matrix $\hat{T}$ (Figure 3): i.e., it is a robust learning method even under a potentially misspecified noise matrix.

We additionally observed that the performance of different algorithms when relying on a noise-correction method to identify and rectify incorrect labels is very dependent on the corruption noise value ρ (Figure 4). Hence, an approach relying on label cleansing is likely to fail under high noise rates.

In contrast, we emphasize the significant performance benefits provided by the proposed `Noisy MRC` —even with estimated $\hat{T}$— when compared to the `Cleansed MRC`. The proposed `Noisy MRC` not only provides accurate predictions that are robust to misslabeling, but it also prevents overly optimistic assessments.

To the best of our knowledge, this is the first robust predictive learning solution that can be trained and evaluated using noisy labels, of critical importance in patient outcome predictions in healthcare, very often subject to misslabeling.

Limitations. The `Noisy MRC` learning procedure is based on a given, fixed set of features, which requires defining and learning a set of features prior to the classification task. On the one hand, this is a benefit of the methodology in the healthcare context, as it enables feature design and selection, of importance for the method's explainability. On the other, it implies that it is not an end-to-end learning framework. However, we note that features can be learned in a separate portion of a dataset, using any feature learning methodology, e.g., deep representation learning.

A potential end-to-end implementation and training of `Noisy MRCs` (e.g., using neural networks) is feasible, if a loss function corresponding to Equation (18) is derived that enables automatic differentiation. However, in such cases, the learned features would depend on the training samples, leading to a challenging assessment of the error in the MRC expectation estimates. In addition, it will be even more difficult to ensure the necessary conditions described in Remark 3 that enable performance guarantees at learning, a key contribution of this work.

We acknowledge that the proposed work is limited to the instance-independent label noise assumption, as in (Natarajan et al., 2013; Frenay and Verleysen, 2014; Abad and Lee, 2021). Label noise in practical healthcare scenarios may exhibit instance-dependence, as certain patient groups may be more susceptible to label corruption than others.

To the best of our knowledge, only few works study the case of instance-dependent noise, e.g., (Liu et al., 2023; Menon et al., 2018), due to the increased complexity associated to the instance-dependent case. These works theoretically analyze the feasibility of learning in such scenarios, while algorithmic developments for these cases largely remain an open research area.

We leave extending the presented MRC-based framework to handle instance-dependent label noise as future work. Towards such goal, we need to construct different estimators $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ in Equations (21) and (22) that account for instance-dependent noise. Precisely, we need to rederive Equation (28) and its corresponding theory (i.e., an equivalent to Lemma 4 in Appendix A) following $\tilde{\mathbf{p}}(\mathbf{x}, \cdot) = T_r \mathbf{p}^*(\mathbf{x}, \cdot)$, where the noise matrix T_r would now depend on the region r of the feature space $\mathcal{X}$.

However, identifying the relevant features and patient groups (regions of the feature space $\mathcal{X}$) susceptible to label corruption is a challenging task, particularly given the complexity of healthcare data and the multitude of features involved in clinical practice. Identification and estimation of per-region dependent noise matrices is a research question in itself in the context of healthcare, as it requires not only acknowledging a corruption in labeling, but a deep understanding of how the misslabeling occurs (as a function of patient features) in healthcare practice.

Acknowledgments

The first author is currently affiliated with the Fondazione Bruno Kessler (FBK), Trento, Italy, and the University of Trento, Italy. The research presented here was conducted while the first author was at the Basque Center for Applied Mathematics (BCAM), Bilbao, Spain. Iñigo Urteaga is supported by "la Caixa" foundation's LCF/BQ/PI22/11910028 award. Additional funding in direct support of this work has been provided by projects PID2022-137063NB-I00, CNS2022-135203, and CEX2021-001142-S funded by MICIU/AEI /10.13039/501100011033 and by the European Union NextGeneration EU / PRTR, as well as through the BERC 2022-2025 program funded by the Basque Government.

References

Z. S. H. Abad and J. Lee. Detecting uncertainty of mortality prediction using confident learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1719–1722, 2021.

V. Agarwal, T. Podchiyska, J.M. Banda, V. Goel, Tiffany I. L., E.P. Minty, T.E. Sweeney, E. Gyang, and N.H. Shah. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6):1166–1173, 2016.

J.A. Aslam and S.E. Decatur. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

O. Badawi and M.J. Breslow. Readmissions and Death after ICU Discharge: Development and Validation of Two Predictive Models. *PLOS ONE*, 7(11):e48758, 2012.

J. A. Baker, P. J. Kornguth, and C. E. Floyd. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *American Journal of Roentgenology*, 166(4):773–778, 1996. doi: 10.2214/ajr.166.4.8610547. URL https://doi.org/10.2214/ajr.166.4.8610547. PMID: 8610547.

J. A. Baker, J. Y. Lo, D. M. Delong, and C. E. Floyd. Computer-aided detection in screening mammography: Variability in cues. *Radiology*, 233(2):411–417, 2004. doi: 10.1148/radiol.2332031200. URL https://doi.org/10.1148/radiol.2332031200. PMID: 15358850.

S. Boughorbel, F. Jarray, N. Venugopal, and H. Elhadi. Alternating Loss Correction for Preterm-Birth Prediction from EHR Data with Noisy Labels. *arXiv preprint arXiv:1811.09782*, 2018.

C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

A. J. Campbell, J. A. Cook, G. Adey, and B. H. Cuthbertson. Predicting death and readmission after intensive care discharge. *British Journal of Anaesthesia*, 100(5):656–662, 2008.

C. Chiew, N. Liu, T. H. Wong, Y. Sim, and H. R. Abdullah. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Annals of surgery*, 272(6):1133, 2020.

M. K. Choi, D. Kim, E. J. Choi, Y. J. Jung, Y. J. Choi, J. H. Cho, and S. H. Jeong. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Scientific reports*, 12(1):7180, 2022.

C.A. Chrusch, K.P. Olafson, P.M. McMillan, D.E. Roberts, and P.R. Gray. High occupancy increases the risk of early death or readmission after transfer from intensive care. *Critical Care Medicine*, 37(10):2753, 2009.

S. Cohen, N. Dagan, N. Cohen-Inger, D. Ofer, and L. Rokach. ICU survival prediction incorporating test-time augmentation to improve the accuracy of ensemble-based models. *IEEE Access*, 9:91584–91592, 2021.

A.A.H. de Hond, I.M.J. Kant, M. Fornasa, G. Cinà, P.W.G. Elbers, P.J. Thoral, M. Sesmu Arbous, and E. W. Steyerberg. Predicting Readmission or Death After Discharge From the ICU: External Validation and Retraining of a Machine Learning Model. *Critical Care Medicine*, 51(2):291, 2023.

J. G. Elmore, C. Y. Nakano, T. D. Koepsell, L. M. Desnick, C. J. D'Orsi, and D. F. Ransohoff. International Variation in Screening Mammography Interpretations in Community-Based Programs. *JNCI: Journal of the National Cancer Institute*, 95(18):1384–1393, 09 2003. ISSN 0027-8874. doi: 10.1093/jnci/djg048. URL https://doi.org/10.1093/jnci/djg048.

M. Elter, R. Schulz-Wendtland, and T. Wittenberg. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11):4164–4172, 2007. doi: https://doi.org/10.1118/1.2786864. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.2786864.

E. Englesson and H. Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 34, pages 30284–30297. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fe2d010308a6b3799a3d9c728ee74244-Paper.pdf.

B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

L. Ju, X. Wang, L. Wang, D. Mahapatra, X. Zhao, Q. Zhou, T. Liu, and Z. Ge. Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation. *IEEE Transactions on Medical Imaging*, 41(6):1533–1546, 2022.

M.W. Kang, J. Kim, D.K. Kim, K.H. Oh, K.W. Joo, Y.S. Kim, and S.S. Han. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. *Critical Care*, 24(1):1–9, 2020.

D. Karimi, H. Dou, S.K. Warfield, and A. Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

M. Kelly, R. Longjohn, and K. Nottingham. Uci machine learning repository. URL http://archive.ics.uci.edu.

S. Kim, W. Kim, and R.W. Park. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research*, 17 (4):232–243, 2011.

B. Kompa, J. Snoek, and A.L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

X. Li, T. Liu, B. Han, G. Niu, and M. Sugiyama. Provably end-to-end label-noise learning without anchor points. In *International conference on machine learning*, volume 139, pages 6403–6413. PMLR, 2021.

T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Y. Liu and H. Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6226–6236. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/liu20e.html.

Y. Liu, H. Cheng, and K. Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pages 21475–21496. PMLR, 2023.

S. Mazuelas, A. Zanoni, and A. Pérez. Minimax classification with 0-1 loss and performance guarantees. In *Advances in Neural Information Processing Systems*, pages 302–312, 2020.

S. Mazuelas, Y. Shen, and A. Pérez. Generalized maximum entropy for supervised classification. *IEEE Transactions on Information Theory*, 68(4):2530–2550, 2022.

S. Mazuelas, M. Romero, and P. Grünwald. Minimax risk classifiers with 0-1 loss. *Journal of Machine Learning Research*, 24(208):1–48, 2023.

A.K. Menon, B. van Rooyen, and N. Natarajan. Learning from binary labels with instance-dependent corruption. *Machine Learning*, 107, 2018.

N. Natarajan, I. S Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Cost-sensitive learning with noisy lables. *Journal of Machine Learning Research*, 18:1–33, 2018.

D.J. Niven, J.F. Bastos, and H.T. Stelfox. Critical Care Transition Programs and the Risk of Readmission or Death After Discharge From an ICU: A Systematic Review and Meta-Analysis. *Critical Care Medicine*, 42(1):179, 2014.

C. Northcutt, L. Jiang, and I. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

P.A. Patel and B.J.B. Grant. Application of mortality prediction systems to individual intensive care units. *Intensive care medicine*, 25:977–982, 1999.

G. Patrini. *Weakly Supervised Learning via Statistical Sufficiency*. PhD thesis, Australian National University, 2016.

G. Patrini, A. Rozza, A.K. Menon, R. Nock, and L. 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*, pages 1944–1952, 2017.

M. Pechenizkiy, A. Tsymbal, S. Puuronen, and O. Pechenizkiy. Class noise and supervised learning in medical domains: The effect of feature extraction. In *19th IEEE symposium on computer-based medical systems*, pages 708–713, 2006.

K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2012.

D. V. Pilcher, G. J. Duke, C. George, M. J. Bailey, and G. Hart. After-Hours Discharge from Intensive Care Increases the Risk of Readmission and Death. *Anaesthesia and Intensive Care*, 35(4):477–485, 2007.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, page 1177–1184, 2007.

B. Ramana, M. Surendra, P. Babu, and N. Bala Venkateswarlu. A critical comparative study of liver patients from usa and india: An exploratory analysis. *International Journal of Computer Science*, 9, 2012.

M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to Reweight Examples for Robust Deep Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018.

J. A. Sáez, B. Krawczyk, and M. Woźniak. On the influence of class noise in medical data classification: Treatment using noise filtering methods. *Applied Artificial Intelligence*, 30 (6):590–609, 2016.

A. Silva, P. Cortez, M.F.e Santos, L. Gomes, and J. Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial intelligence in medicine*, 36(3):223–234, 2006.

G. Stempfel and L. Ralaivola. Learning SVMs from sloppily labeled data. In *Proceedings of the 19th International Conference on Artificial Neural Networks*, page 884–893, 2009.

R. Taylor, J. Pare, A. Venkatesh, H. Mowafi, E. Melnick, W. Fleischman, and M. Hall. Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data-driven, machine learning approach. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 23, 2015.

S. Tripathi and N. Hemachandra. Cost sensitive learning in the presence of symmetric label noise. *Journal of Machine Learning Research*, 18(155):1–33, 2018.

A. van den Hout and P. G. M. van der Heijden. Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review*, 70(2):269–288, 2002.

X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Appendix A. Proofs of Section 2

Lemma 4 *Let p^* be the true distribution and $\tilde{p}$ be the observed-corrupted distribution. Then it holds,*

$$\tilde{\mathbf{p}}(\mathbf{x},\cdot) \;=\; T\;\mathbf{p}^*(\mathbf{x},\cdot)\;. \tag{28}$$

Proof of Lemma 4 Using directly the factorization of probability distributions, we get that:

$$\begin{aligned}\tilde{p}(\mathbf{x},y) \;&=\; p(\mathbf{x},\tilde{y}) \;=\; \sum_{y} p(\mathbf{x},\tilde{y},y) \;=\; \sum_{y\in\mathcal{Y}} p(\tilde{y}|\mathbf{x},y)p^*(\mathbf{x},y) = \\ &= \sum_{y\in\mathcal{Y}} p(\tilde{y}|y)p^*(\mathbf{x},y)\;, \end{aligned}\tag{29}$$

where Equation (29) is justified by the fact that the probability of flipping y to $\tilde{y}$ is assumed to be independent from $\mathbf{x}$. The thesis follows by simply considering the Equation (29) in matrix form, that is:

$$\tilde{\mathbf{p}}(\mathbf{x},\cdot) \;=\; T\;\mathbf{p}^*(\mathbf{x},\cdot)$$

where $\tilde{\mathbf{p}}(\mathbf{x},\cdot)$ and $\mathbf{p}^*(\mathbf{x},\cdot)$ are probability column vectors in $\mathbb{R}^{|\mathcal{Y}|}_{\geq 0}$. ■

Lemma 5 *In the notation of Section 2, let ℓ denote a generic loss function. Then ULE defined in Equation (10) as*

$$ULE = \frac{1}{N}\sum_{i=1}^{N} \tilde{\ell}(h,(\mathbf{x}_i,\tilde{y}_i))\;,$$

with

$$\tilde{\ell}(h,(\mathbf{x}_i,y_i)) = (T^{-1})_{1,y_i}\ell(h,(\mathbf{x}_i,1)) \;+\; (T^{-1})_{2,y_i}\ell(h,(\mathbf{x}_i,2))$$

is an unbiased estimator of $\ell(h,p^*)$.

Proof of Lemma 5 . Without loss of generality, we will encode the labels as $\mathcal{Y}=\{1,2\}$. Then it holds:

$$\ell(h,p^*) := \mathbb{E}_{p^*}\left[\ell(h,(\mathbf{x},y))\right] = \sum_{\mathbf{x},y}\ell(h,(\mathbf{x},y))p^*(\mathbf{x},y) = \sum_{\mathbf{x}}\mathbf{p}^*(\mathbf{x},\cdot)^\top\begin{bmatrix}\ell(h,(\mathbf{x},1))\\ \ell(h,(\mathbf{x},2))\end{bmatrix}$$

Using Lemma 4:

$$\ell(h,p^*) = \sum_{\mathbf{x}}\tilde{\mathbf{p}}(\mathbf{x},\cdot)^\top(T^{-1})^\top\begin{bmatrix}\ell(h,(\mathbf{x},1))\\ \ell(h,(\mathbf{x},2))\end{bmatrix} = \sum_{\mathbf{x}}\tilde{\mathbf{p}}(\mathbf{x},\cdot)^\top\begin{bmatrix}(T^{-1})_{1,1} & (T^{-1})_{2,1}\\ (T^{-1})_{1,2} & (T^{-1})_{2,2}\end{bmatrix}\begin{bmatrix}\ell(h,(\mathbf{x},1))\\ \ell(h,(\mathbf{x},2))\end{bmatrix}$$

$$= \sum_{\mathbf{x}}\tilde{\mathbf{p}}(\mathbf{x},\cdot)^\top\begin{bmatrix}(T^{-1})_{1,1}\ell(h,(\mathbf{x},1)) + (T^{-1})_{2,1}\ell(h,(\mathbf{x},2))\\ (T^{-1})_{1,2}\ell(h,(\mathbf{x},1)) + (T^{-1})_{2,2}\ell(h,(\mathbf{x},2))\end{bmatrix} =$$

$$= \sum_{\mathbf{x}}\sum_{y}\tilde{p}(\mathbf{x},y)\left[(T^{-1})_{1,y}\ell(h,(\mathbf{x},1)) + (T^{-1})_{2,y}\ell(h,(\mathbf{x},2)\right] = \sum_{\mathbf{x}}\sum_{y}\tilde{p}(\mathbf{x},y)\tilde{\ell}(h,(\mathbf{x},y))$$

This implies that:

$$\ell(h,p^*) = \mathbb{E}_{\tilde{p}}\left[\tilde{\ell}(h,(\mathbf{x},y))\right]\;,$$

hence, ULE defined in Equation (10) is an unbiased estimator of $\ell(h,p^*)$. ■

Appendix B. Description of additional healthcare datasets

This section is devoted to the accurate description of the additional datasets we used to evaluate the proposed methods.

B.0.1. Haberman's Survival:

We here tackle again the challenge of predicting patients' survival, using an UCI's dataset[8] (smaller than the one proposed in Section 4.1.1), that collects cases from a study conducted at the University of Chicago's Billings Hospital between 1958 and 1970.

It focuses on patient survival cases who underwent breast cancer surgery. In particular, it consists of 306 instances with 3 continuous features (X) — age, the year of the surgery and the number of positive axillary nodes detected — and `survival_status` as target variable Y indicating whether the patient survived five years or longer ($Y = 1$) or died within five years ($Y = 2$) after the surgery.

Summary statistics of the dataset are described in Table 3.

	Original
N. patients	306
N. total features	3
N. continuous features	3
N. categorical (binary) features	0 (0)

Table 3: Characteristics of the original (and already polished) *Haberman's Survival* dataset.

B.0.2. Indian Liver Patient dataset:

We here deal with the prediction task to determine whether a patient suffers from liver disease or not, using the UCI's dataset *Indian Liver Patient dataset*[9]. The original dataset was first proposed by Ramana et al. (2012) as a critical comparison of patients across USA and India.

This dataset consists of 583 patients records among which 416 with liver disease and 167 without liver disease, collected from the north-east region of Andhra Pradesh, India. It includes 10 variables (X) — demographic data (age and gender), lab results and several biochemical markers — and `Selector` as target variable (Y), categorizing patients into groups based on their liver condition.

Summary statistics of the dataset are described in Table 4.

B.0.3. Pima Indians Diabetes dataset:

We here take the challenge of predicting whether or not an individual has been diagnosed with diabetes, using the dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases[10]. This dataset collects data from 768 women at least 21 years old of

8. Available at https://archive.ics.uci.edu/dataset/43/haberman+s+survival.
9. Available at https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset.
10. Available at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

	Original
N. patients	583
N. total features	10
N. continuous features	9
N. categorical (binary) features	1 (1)

Table 4: Characteristics of the original (and already polished) *Indian Liver Patient* dataset.

Pima Indian heritage. It contains 8 features (X) — including the number of pregnancies the patient has had, their BMI, insulin level, age — and `outcome` as the target variable (Y), describing whether an individual has been diagnosed with diabetes ($Y = 1$) or not ($Y = 0$).

Summary statistics of the dataset are described in Table 5.

	Original
N. patients	768
N. total features	8
N. continuous features	8
N. categorical (binary) features	0 (0)

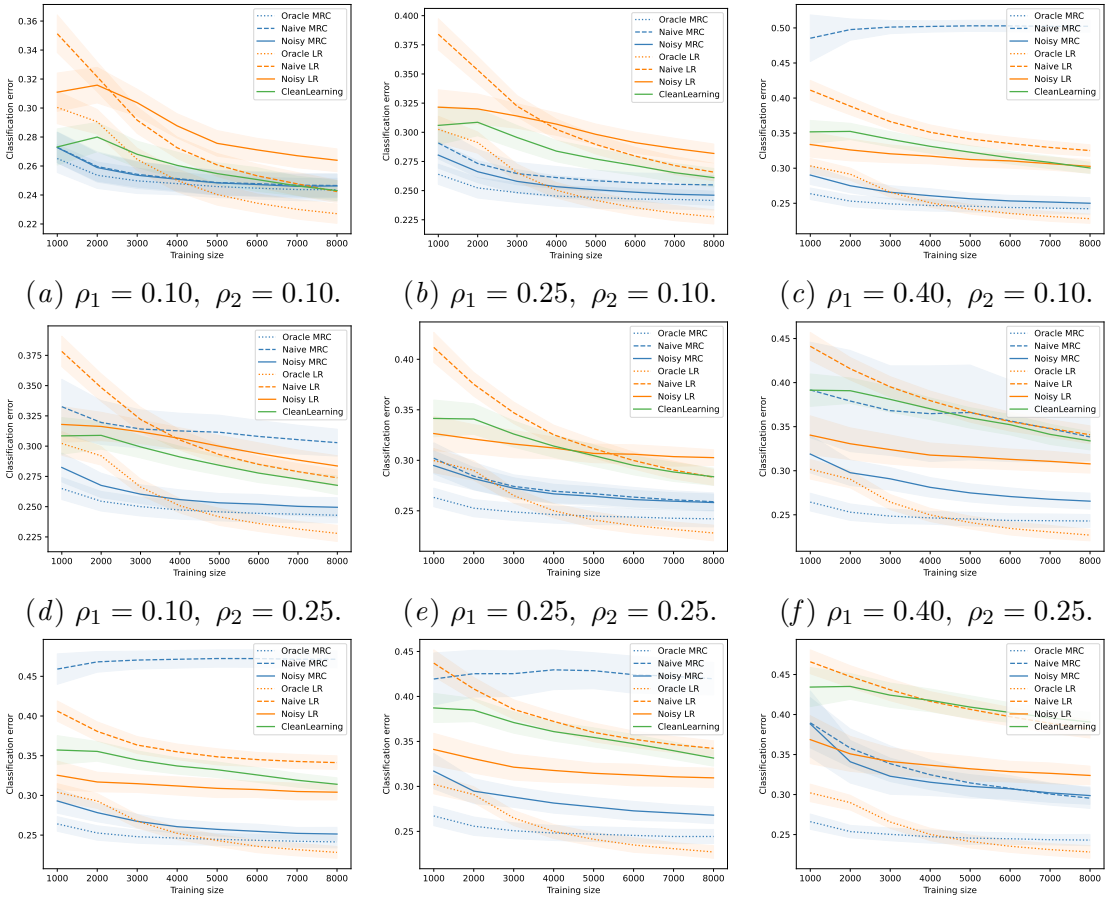
Table 5: Characteristics of the original (and already polished) *Pima Indians Diabetes* dataset.

Appendix C. Additional numerical results with T known

C.1. Evaluation on clean labels

Here we present additional numerical results on methods evaluated on clean labels.

C.1.1. ICU Mortality Dataset:



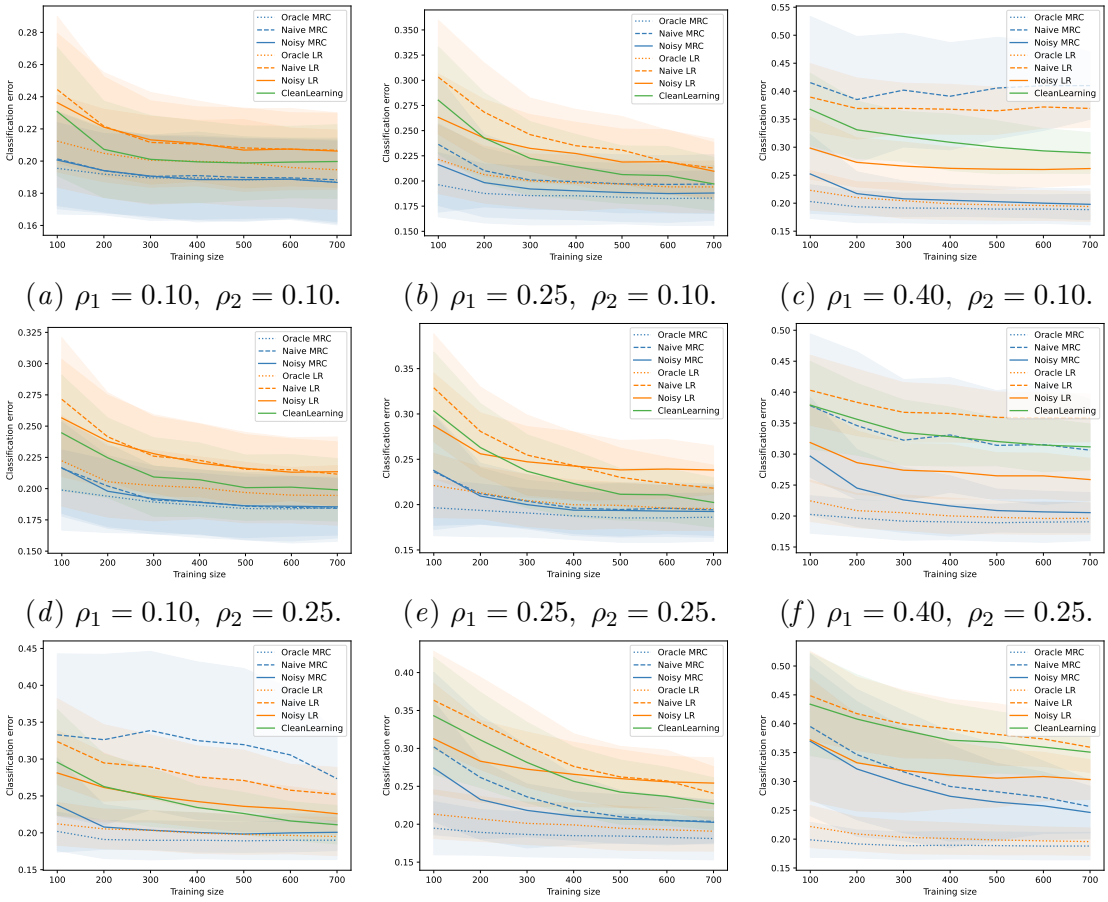
(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 5: Experiment with T known and evaluation on clean labels. Classification error for different ρ_1, $\rho_2 = 0.10$. `Noisy MRC` (solid blue) outperforms `Noisy LR` (solid orange) and `CleanLearning` (solid green).

C.1.2. Mammographic Mass Dataset:



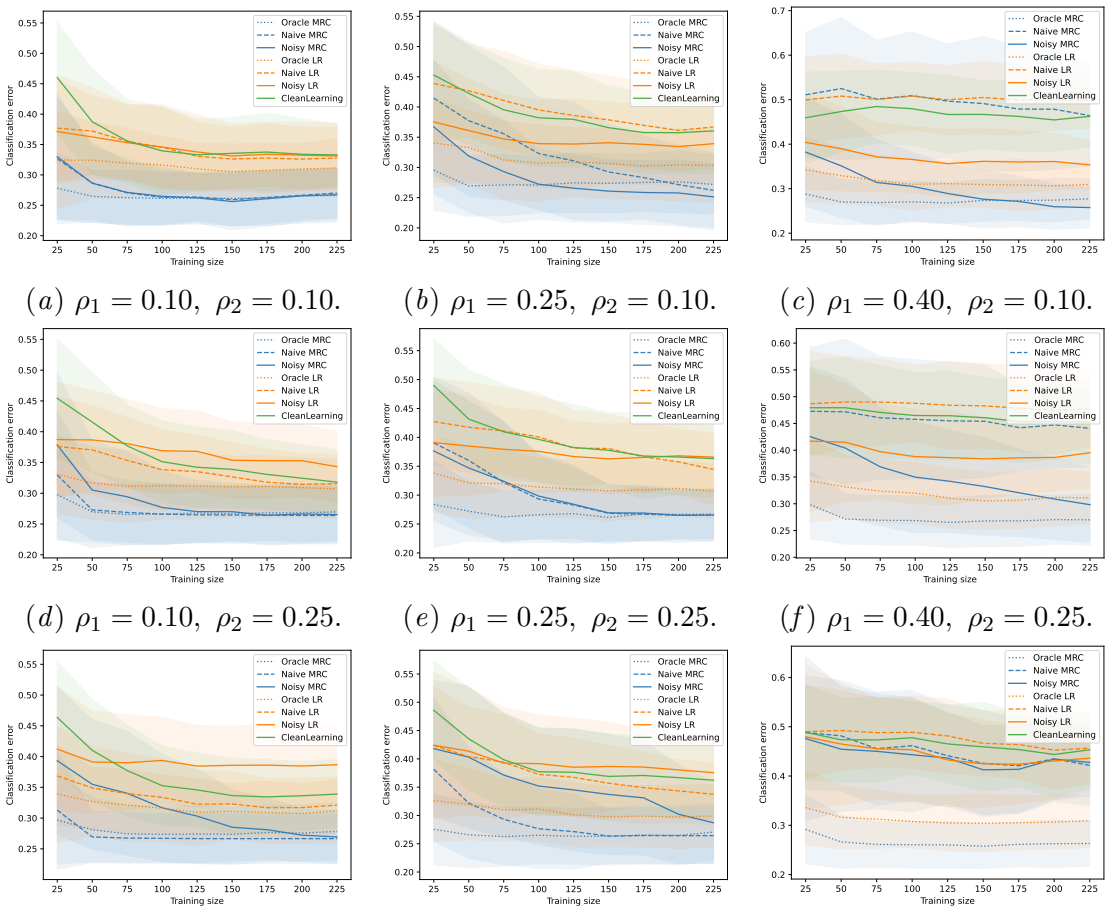
(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 6: Experiment with T known and evaluation on clean labels. Classification error for different ρ_1, $\rho_2 = 0.10$. `Noisy MRC` (solid blue) outperforms `Noisy LR` (solid orange) and `CleanLearning` (solid green).

C.1.3. Haberman's Survival Dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

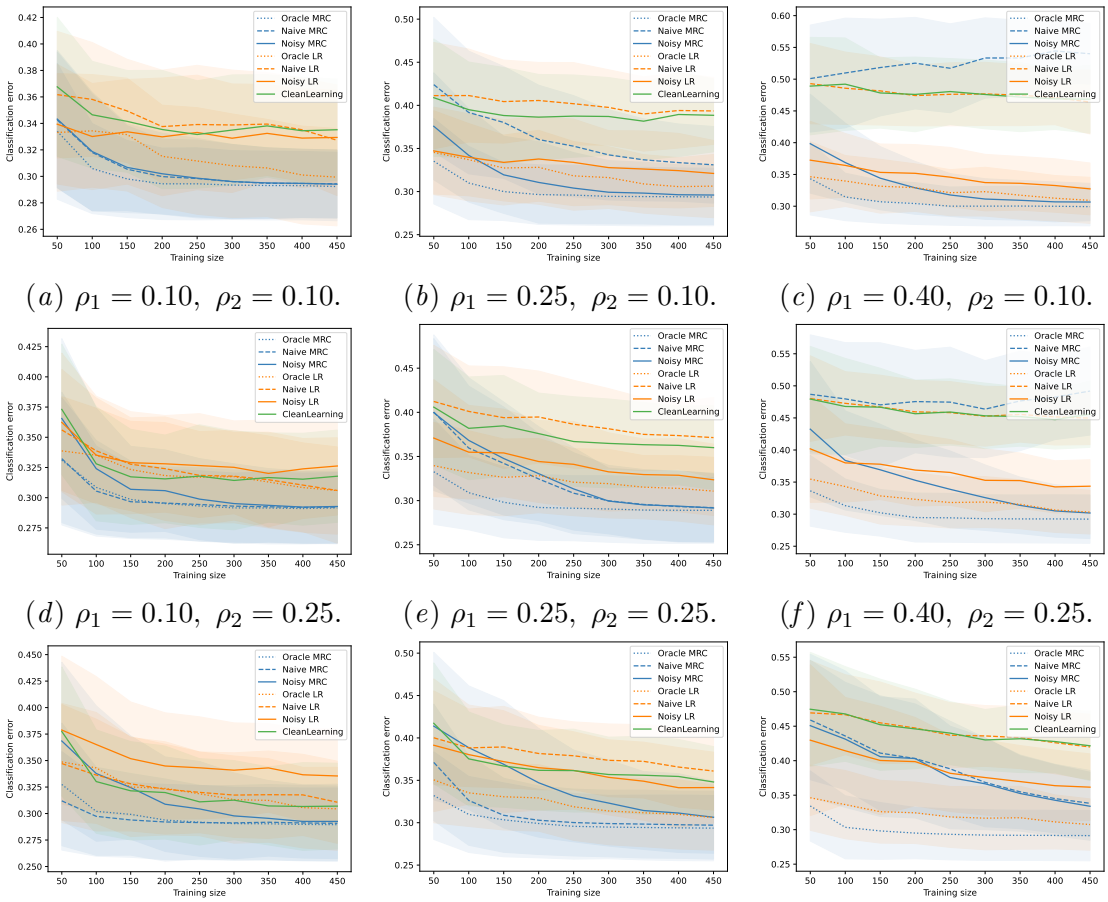
(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 7: Experiment with T known and evaluation on clean labels. Classification error for different ρ_1, $\rho_2 = 0.10$. `Noisy MRC` (solid blue) outperforms `Noisy LR` (solid orange) and `CleanLearning` (solid green).

Notice that for this dataset, as the noise rates rise, the classification errors of all the methods notably increase. This phenomenon can be attributed to the limited size of the dataset. With a small dataset like this one, the classifiers struggle to accurately account for noise, resulting in increasingly significant classification errors, and leading to classifiers that are essentially rendered useless (i.e., with classification error around 0.5, see e.g., Figure 7(i) with $\rho_1 = \rho_2 = 0.40$).

C.1.4. Indian Liver Patient dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

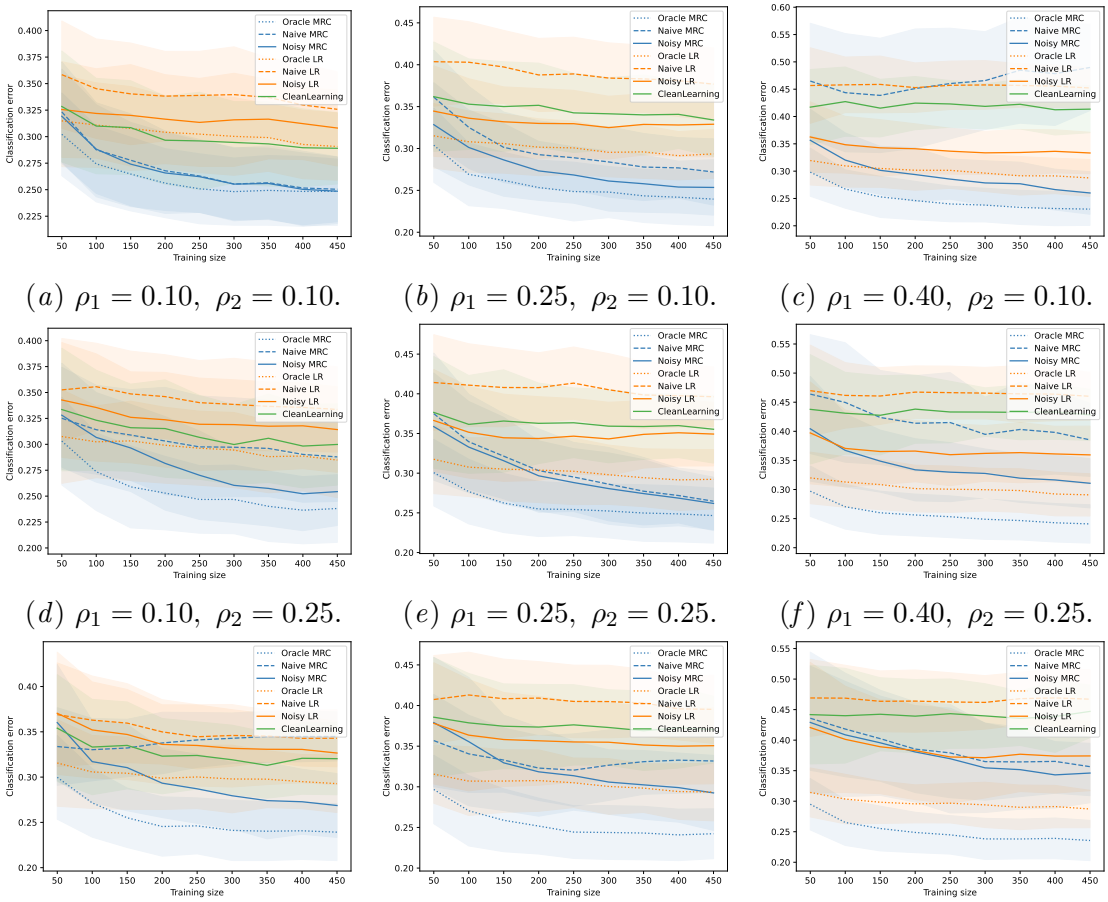
(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 8: Experiment with T known and evaluation on clean labels. Classification error for different ρ_1, $\rho_2 = 0.10$. `Noisy MRC` (solid blue) outperforms `Noisy LR` (solid orange) and `CleanLearning` (solid green).

As we observed in Appendix C.1.3, we also encounter a similar issue of increased classification error with larger noise rates, attributable to the restricted sample size. Moreover, notice how the method `CleanLearning` encounters significant challenges in effectively addressing noise, particularly evident when $\rho_1 = 0.4$.

C.1.5. Pima Indians Diabetes dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 9: Experiment with T known and evaluation on clean labels. Classification error for different ρ_1, $\rho_2 = 0.10$. `Noisy MRC` (solid blue) outperforms `Noisy LR` (solid orange) and `CleanLearning` (solid green).

C.1.6. Summary Results across all healthcare datasets:

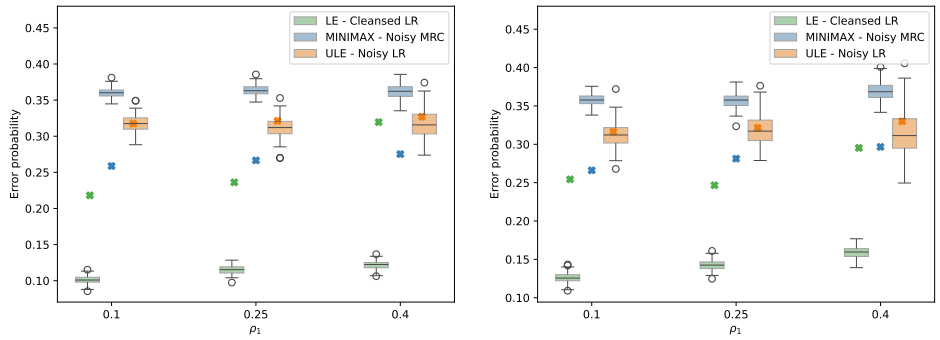
Datasets	Methods	$(\rho_1, \rho_2) = (0.1, 0.1)$	$(\rho_1, \rho_2) = (0.25, 0.1)$	$(\rho_1, \rho_2) = (0.40, 0.1)$
Mortality	`Oracle MRC`	0.244 ± 0.008	0.242 ± 0.008	0.242 ± 0.008
$n_{\text{train}} = 8000$	`Oracle LR`	0.228 ± 0.007	0.227 ± 0.006	0.228 ± 0.007
	`Naive MRC`	0.247 ± 0.008	0.255 ± 0.008	0.502 ± 0.008
	`Noisy MRC`	0.246 ± 0.008	**0.246 ± 0.009**	**0.250 ± 0.008**
	`Naive LR`	**0.242 ± 0.008**	0.266 ± 0.007	0.325 ± 0.008
	`Noisy LR`	0.264 ± 0.008	0.282 ± 0.008	0.303 ± 0.010
	`CleanLearning`	0.243 ± 0.007	0.261 ± 0.008	0.300 ± 0.008
Mammogr.	`Oracle MRC`	0.187 ± 0.0256	0.183 ± 0.0269	0.188 ± 0.0272
$n_{\text{train}} = 700$	`Oracle LR`	0.195 ± 0.0247	0.194 ± 0.0257	0.194 ± 0.0266
	`Naive MRC`	0.188 ± 0.0259	0.197 ± 0.0292	0.410 ± 0.0603
	`Noisy MRC`	**0.186 ± 0.0264**	**0.188 ± 0.0271**	**0.198 ± 0.0275**
	`Naive LR`	0.207 ± 0.0230	0.213 ± 0.0289	0.369 ± 0.0339
	`Noisy LR`	0.206 ± 0.0238	0.209 ± 0.0285	0.262 ± 0.0287
	`CleanLearning`	0.199 ± 0.0229	0.197 ± 0.0264	0.289 ± 0.0364
Haberman	`Oracle MRC`	0.266 ± 0.045	0.272 ± 0.049	0.277 ± 0.045
$n_{\text{train}} = 225$	`Oracle LR`	0.311 ± 0.048	0.303 ± 0.061	0.310 ± 0.049
	`Naive MRC`	0.270 ± 0.040	0.262 ± 0.062	0.464 ± 0.113
	`Noisy MRC`	**0.267 ± 0.041**	**0.251 ± 0.055**	**0.258 ± 0.046**
	`Naive LR`	0.328 ± 0.051	0.367 ± 0.067	0.499 ± 0.064
	`Noisy LR`	0.331 ± 0.054	0.339 ± 0.066	0.354 ± 0.055
	`CleanLearning`	0.333 ± 0.051	0.360 ± 0.069	0.463 ± 0.070
Liver	`Oracle MRC`	0.292 ± 0.026	0.294 ± 0.033	0.299 ± 0.030
$n_{\text{train}} = 450$	`Oracle LR`	0.299 ± 0.037	0.306 ± 0.037	0.309 ± 0.036
	`Naive MRC`	**0.294 ± 0.026**	0.331 ± 0.044	0.540 ± 0.047
	`Noisy MRC`	**0.294 ± 0.026**	**0.296 ± 0.034**	**0.307 ± 0.029**
	`Naive LR`	0.327 ± 0.042	0.394 ± 0.039	0.464 ± 0.050
	`Noisy LR`	0.329 ± 0.036	0.321 ± 0.038	0.327 ± 0.041
	`CleanLearning`	0.335 ± 0.038	0.388 ± 0.042	0.468 ± 0.054
Diabetes	`Oracle MRC`	0.248 ± 0.030	0.240 ± 0.032	0.231 ± 0.030
$n_{\text{train}} = 450$	`Oracle LR`	0.291 ± 0.032	0.294 ± 0.029	0.288 ± 0.034
	`Naive MRC`	**0.250 ± 0.031**	0.272 ± 0.040	0.490 ± 0.079
	`Noisy MRC`	**0.249 ± 0.032**	**0.254 ± 0.033**	**0.260 ± 0.039**
	`Naive LR`	0.326 ± 0.035	0.377 ± 0.043	0.452 ± 0.037
	`Noisy LR`	0.308 ± 0.036	0.329 ± 0.039	0.333 ± 0.038
	`CleanLearning`	0.289 ± 0.034	0.334 ± 0.044	0.414 ± 0.043

Table 6: Experiments with T known and evaluation on clean labels, on different datasets. Average and standard deviation of classification errors for different ρ_1, fixed $\rho_2 = 0.1$, and fixed training size n_{train}. Among the methods trained on noisy labels, our method `Noisy MRC` can more adequately adapt to noise in the labels.

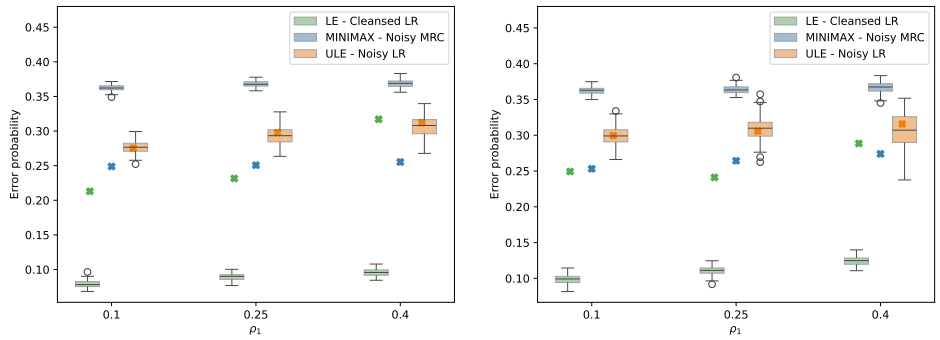
C.2. Evaluation on noisy labels

We present additional numerical results to give a complete comparison of our `Noisy MRC` and the two baselines, `Cleansed LR` and `Noisy LR`, when evaluated on noisy labels. Specifically, we presents boxplots for fixed values of ρ_2 and fixed training size, for different datasets.

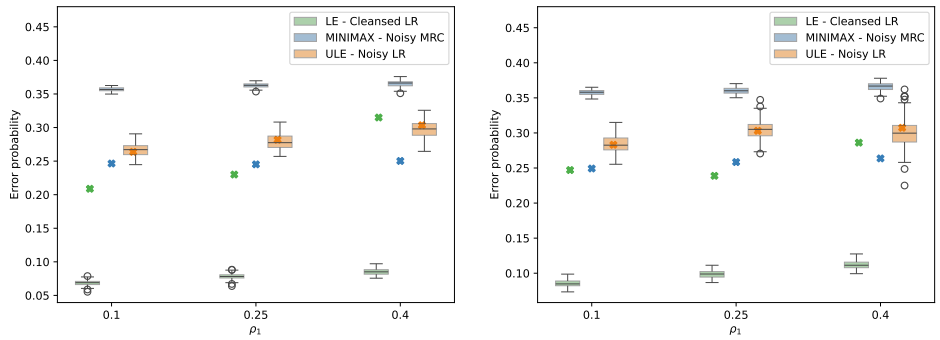
C.2.1. ICU Mortality Dataset:



(a) $n_{\text{train}} = 2.000,\ \rho_2 = 0.10.$ (b) $n_{\text{train}} = 2.000,\ \rho_2 = 0.25.$



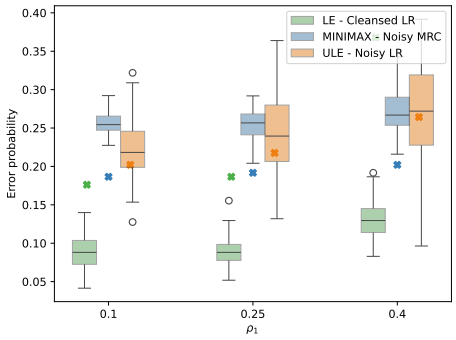
(c) $n_{\text{train}} = 5.000,\ \rho_2 = 0.10.$ (d) $n_{\text{train}} = 5.000,\ \rho_2 = 0.25.$



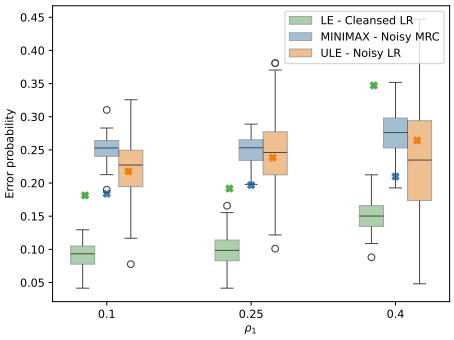
(e) $n_{\text{train}} = 8.000,\ \rho_2 = 0.10.$ (f) $n_{\text{train}} = 8.000,\ \rho_2 = 0.25.$

Figure 10: Experiment with T known and evaluation on noisy labels. Error measures of methods for $\rho_2 = 0.10$ (left column), $\rho_2 = 0.25$ (right column) and and different training sizes.

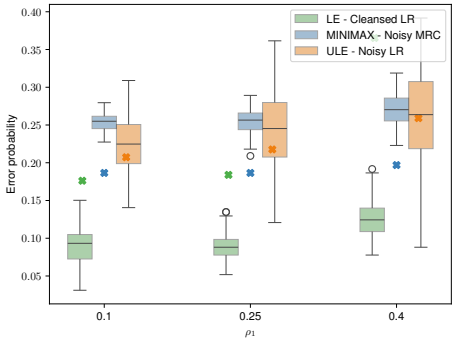
C.2.2. Mammographic Mass Dataset:



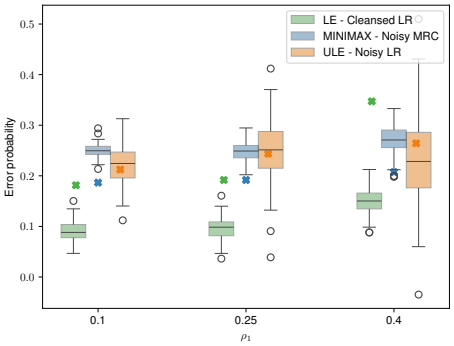
(a) $n_{\text{train}} = 500,\ \rho_2 = 0.10.$



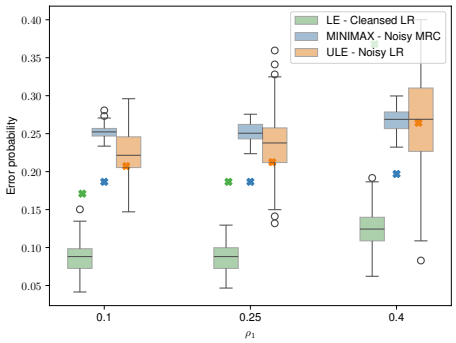
(b) $n_{\text{train}} = 500,\ \rho_2 = 0.25.$



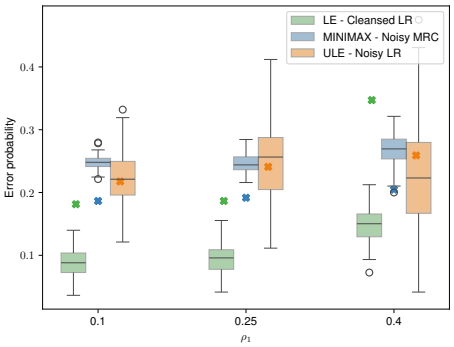
(c) $n_{\text{train}} = 600,\ \rho_2 = 0.10.$



(d) $n_{\text{train}} = 600,\ \rho_2 = 0.25.$



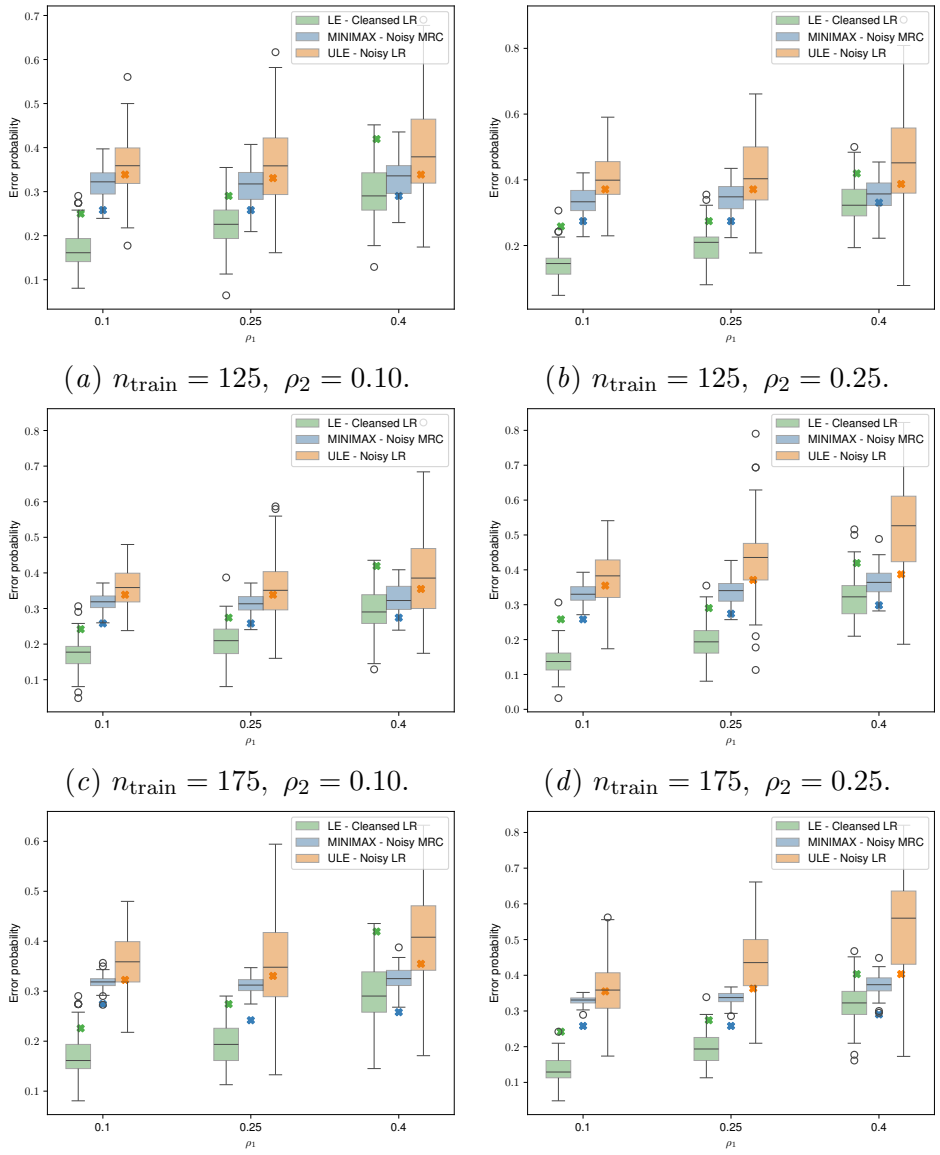
(e) $n_{\text{train}} = 700,\ \rho_2 = 0.10.$



(f) $n_{\text{train}} = 700,\ \rho_2 = 0.25.$

Figure 11: Experiment with T known and evaluation on noisy labels. Error measures of methods for $\rho_2 = 0.10$ (left column), $\rho_2 = 0.25$ (right column) and and different training sizes.

C.2.3. Haberman's Survival Dataset:



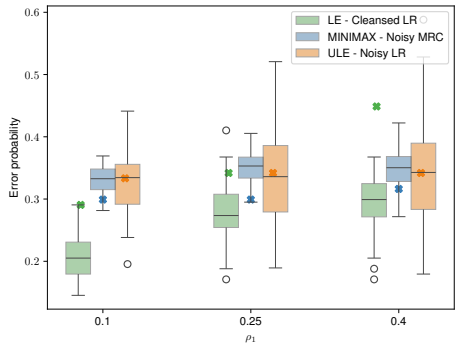
(a) $n_{\text{train}} = 125,\ \rho_2 = 0.10.$ (b) $n_{\text{train}} = 125,\ \rho_2 = 0.25.$

(c) $n_{\text{train}} = 175,\ \rho_2 = 0.10.$ (d) $n_{\text{train}} = 175,\ \rho_2 = 0.25.$

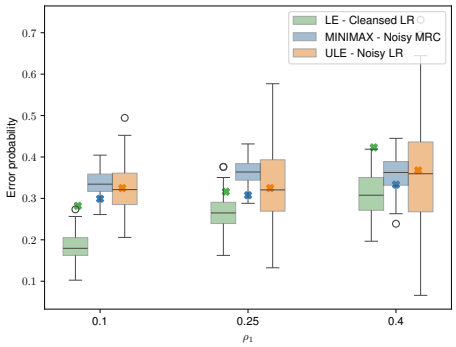
(e) $n_{\text{train}} = 225,\ \rho_2 = 0.10.$ (f) $n_{\text{train}} = 225,\ \rho_2 = 0.25.$

Figure 12: Experiment with T known and evaluation on noisy labels. Error measures of methods for $\rho_2 = 0.10$ (left column), $\rho_2 = 0.25$ (right column) and and different training sizes.

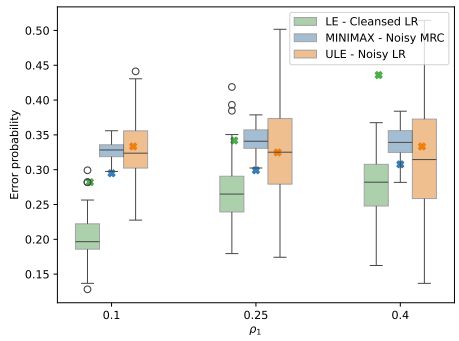
C.2.4. Indian Liver Patient dataset:



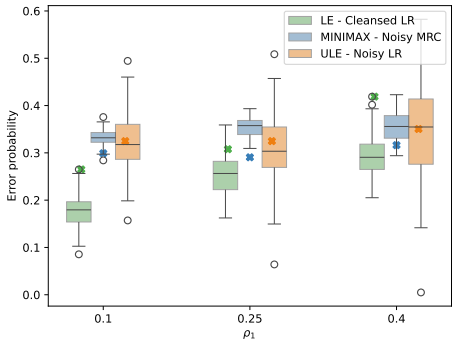
(a) $n_{\text{train}} = 250,\ \rho_2 = 0.10.$



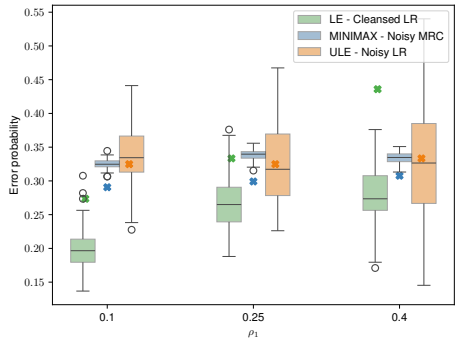
(b) $n_{\text{train}} = 250,\ \rho_2 = 0.25.$



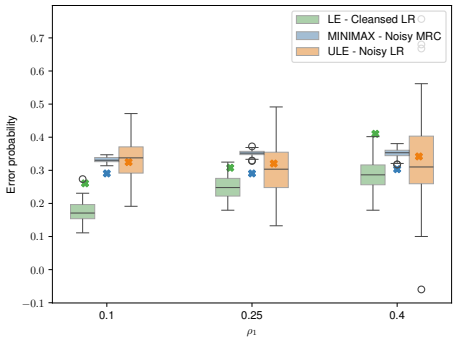
(c) $n_{\text{train}} = 350,\ \rho_2 = 0.10.$



(d) $n_{\text{train}} = 350,\ \rho_2 = 0.25.$



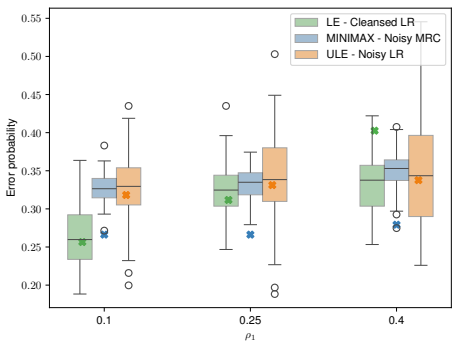
(e) $n_{\text{train}} = 450,\ \rho_2 = 0.10.$



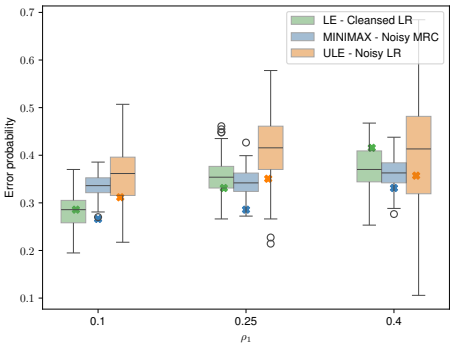
(f) $n_{\text{train}} = 450,\ \rho_2 = 0.25.$

Figure 13: Experiment with T known and evaluation on noisy labels. Error measures of methods for $\rho_2 = 0.10$ (left column), $\rho_2 = 0.25$ (right column) and and different training sizes.

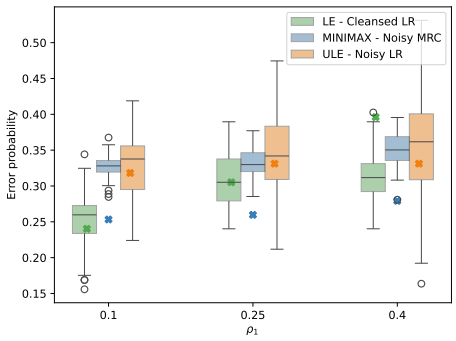
C.2.5. Pima Indians Diabetes dataset:



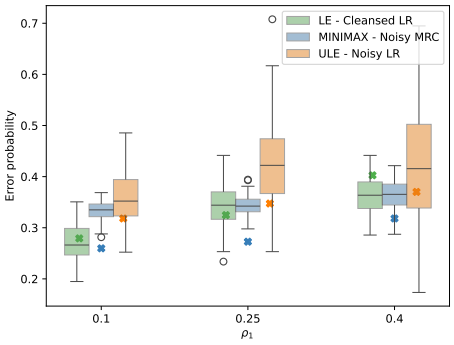
(a) $n_{\text{train}} = 250,\ \rho_2 = 0.10.$



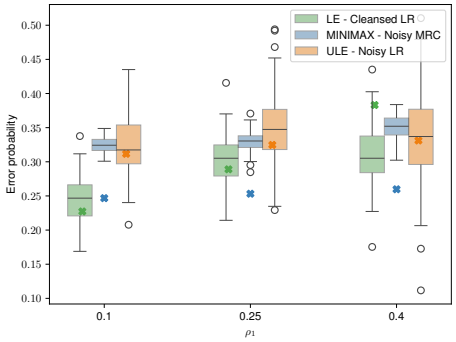
(b) $n_{\text{train}} = 250,\ \rho_2 = 0.25.$



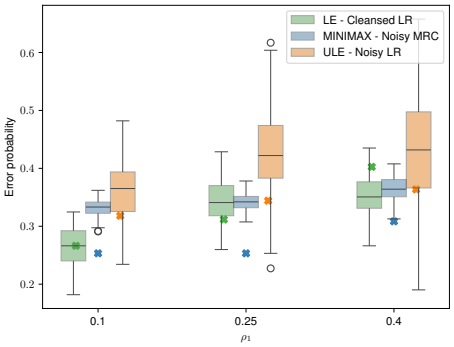
(c) $n_{\text{train}} = 350,\ \rho_2 = 0.10.$



(d) $n_{\text{train}} = 350,\ \rho_2 = 0.25.$



(e) $n_{\text{train}} = 450,\ \rho_2 = 0.10.$



(f) $n_{\text{train}} = 450,\ \rho_2 = 0.25.$

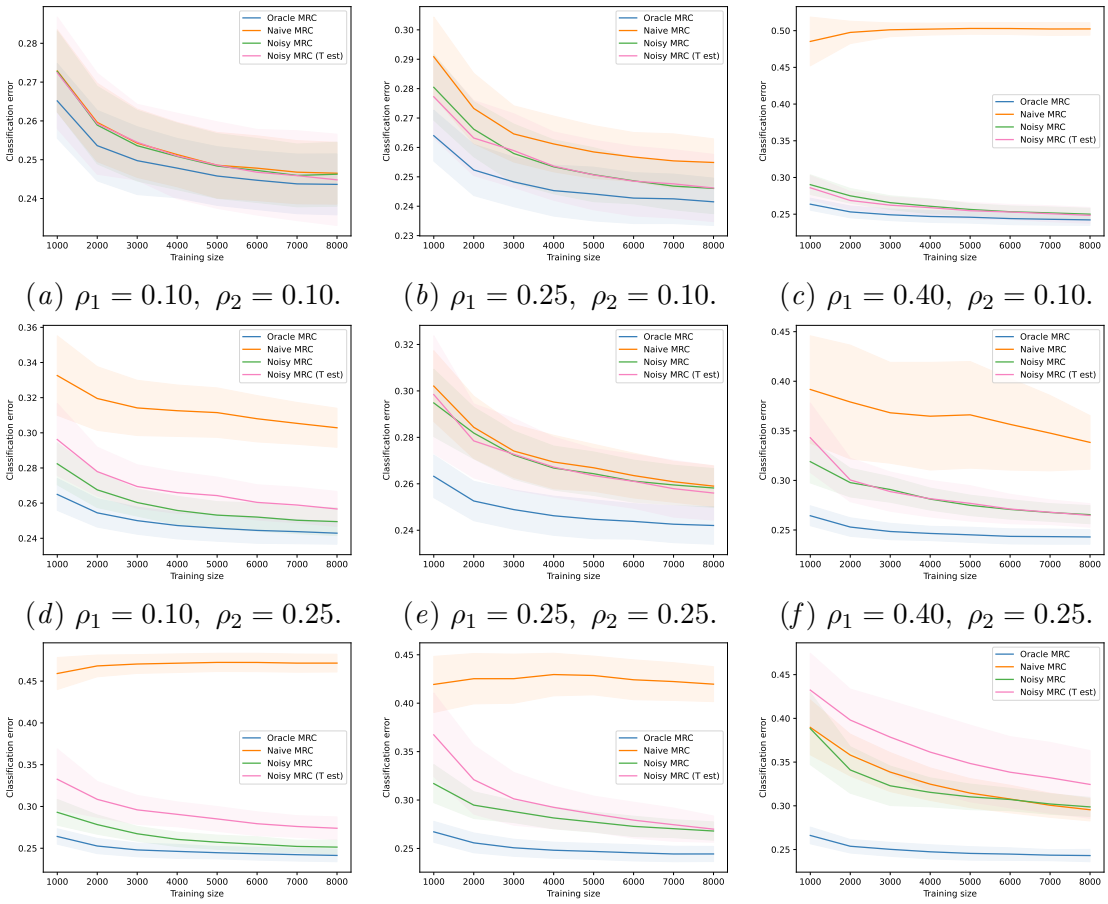
Figure 14: Experiment with T known and evaluation on noisy labels. Error measures of methods for $\rho_2 = 0.10$ (left column), $\rho_2 = 0.25$ (right column) and and different training sizes.

Appendix D. Additional numerical results with T unknown

D.1. Learning with T estimated

We present additional results for experiments with T estimated that reinforce the findings of Section 5.3.

D.1.1. ICU Mortality Dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 15: Experiment with T unknown and evaluation on clean labels. Classification error of MRCs for different $\rho_1,\ \rho_2$. Performances of our method trained with $\hat{T}$ estimated (pink) and with T known (green) are consistent

We present below the estimation accuracies for the estimated $\hat{T}$ on the *ICU Mortality* dataset. We have computed the average relative error of the estimated transition matrix (as well as the minimum and maximum estimated values) across 100-folds for $ntrain = 8.000$ for different values of ρ_1 and ρ_2.

Table 7: $\rho_1 = 0.10,\ \rho_2 = 0.10$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	1.249 ± 0.081	0.207	0.244
$\hat{\rho}_2$	1.101 ± 0.076	0.190	0.226

Table 8: $\rho_1 = 0.1,\ \rho_2 = 0.25$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	1.363 ± 0.079	0.221	0.254
$\hat{\rho}_2$	0.353 ± 0.039	0.316	0.363

Table 9: $\rho_1 = 0.10,\ \rho_2 = 0.40$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	1.289 ± 0.064	0.215	0.243
$\hat{\rho}_2$	0.215 ± 0.027	0.456	0.510

Table 10: $\rho_1 = 0.25,\ \rho_2 = 0.10$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.401 ± 0.040	0.329	0.382
$\hat{\rho}_2$	1.202 ± 0.075	0.204	0.247

Table 11: $\rho_1 = 0.25,\ \rho_2 = 0.25$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.443 ± 0.033	0.342	0.383
$\hat{\rho}_2$	0.382 ± 0.035	0.322	0.371

Table 12: $\rho_1 = 0.25,\ \rho_2 = 0.40$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.391 ± 0.034	0.331	0.373
$\hat{\rho}_2$	0.199 ± 0.027	0.458	0.512

Table 13: $\rho_1 = 0.40,\ \rho_2 = 0.10$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.205 ± 0.026	0.454	0.509
$\hat{\rho}_2$	1.152 ± 0.069	0.195	0.231

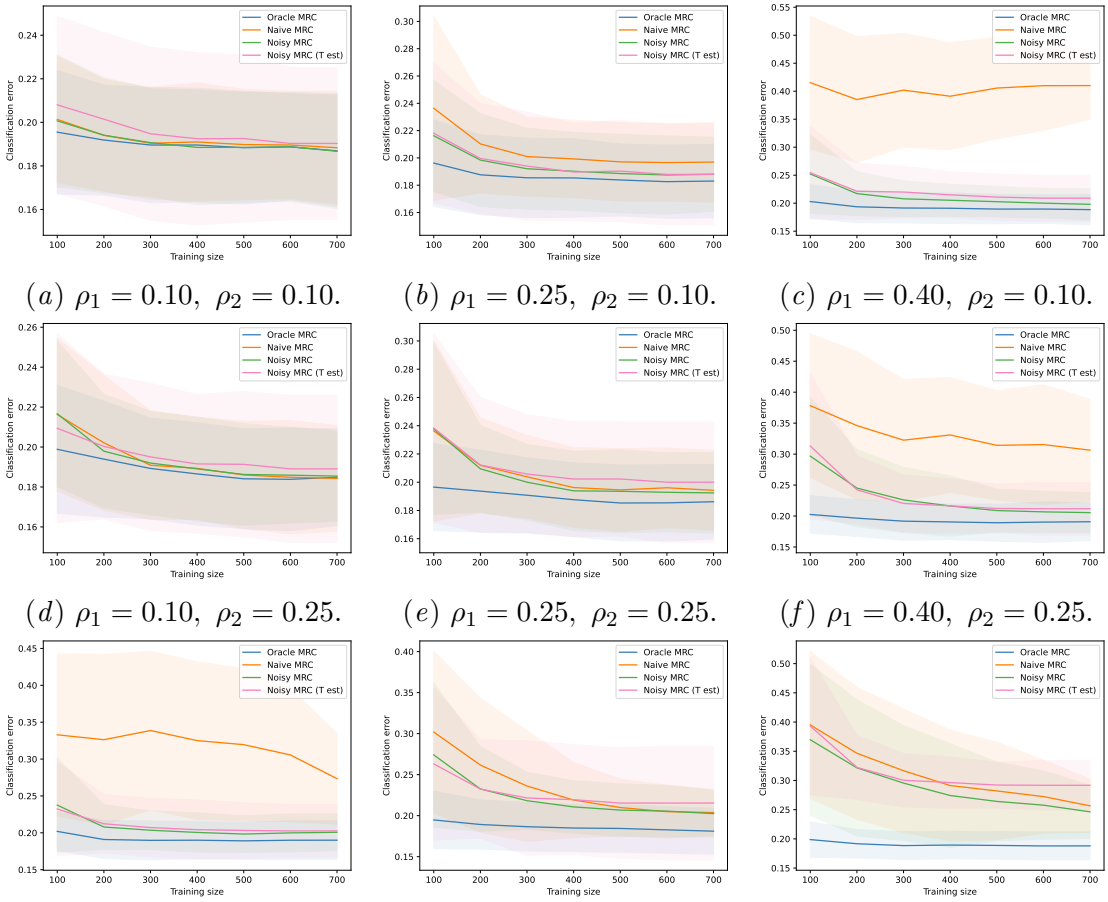
Table 14: $\rho_1 = 0.40,\ \rho_2 = 0.25$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.212 ± 0.021	0.466	0.503
$\hat{\rho}_2$	0.343 ± 0.032	0.315	0.353

Table 15: $\rho_1 = 0.4,\ \rho_2 = 0.4$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.167 ± 0.023	0.443	0.487
$\hat{\rho}_2$	0.149 ± 0.026	0.432	0.482

D.1.2. Mammographic Mass Dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 16: Experiment with T unknown and evaluation on clean labels. Classification error of MRCs for different $\rho_1,\ \rho_2$. Performances of our method trained with $\hat{T}$ estimated (pink) and with T known (green) are consistent

We present below the estimation accuracies for the estimated $\hat{T}$ on the *Mammographic Mass* dataset. We have computed the average relative error of the estimated transition matrix (as well as the minimum and maximum estimated values) across 100-folds for $ntrain = 700$ for different values of ρ_1 and ρ_2.

Table 16: $\rho_1 = 0.10,\ \rho_2 = 0.10$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.952 ± 0.232	0.147	0.255
$\hat{\rho}_2$	1.182 ± 0.338	0.124	0.291

Table 17: $\rho_1 = 0.10,\ \rho_2 = 0.25$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.989 ± 0.233	0.137	0.262
$\hat{\rho}_2$	0.376 ± 0.138	0.256	0.434

Table 18: $\rho_1 = 0.10,\ \rho_2 = 0.40$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.891 ± 0.224	0.132	0.241
$\hat{\rho}_2$	0.147 ± 0.091	0.372	0.538

Table 19: $\rho_1 = 0.25,\ \rho_2 = 0.10$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.309 ± 0.127	0.264	0.406
$\hat{\rho}_2$	1.253 ± 0.277	0.167	0.293

Table 20: $\rho_1 = 0.25,\ \rho_2 = 0.25$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.301 ± 0.122	0.241	0.397
$\hat{\rho}_2$	0.334 ± 0.126	0.255	0.411

Table 21: $\rho_1 = 0.25,\ \rho_2 = 0.40$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.294 ± 0.125	0.235	0.458
$\hat{\rho}_2$	0.126 ± 0.081	0.380	0.585

Table 22: $\rho_1 = 0.40,\ \rho_2 = 0.10$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.174 ± 0.091	0.350	0.586
$\hat{\rho}_2$	1.150 ± 0.242	0.143	0.287

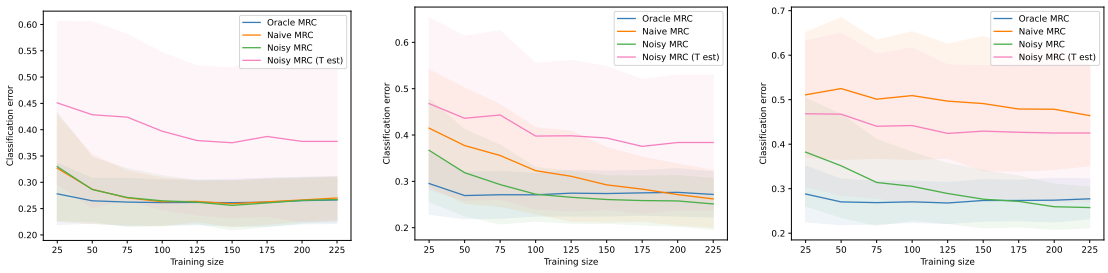
Table 23: $\rho_1 = 0.40,\ \rho_2 = 0.25$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.178 ± 0.088	0.382	0.552
$\hat{\rho}_2$	0.275 ± 0.124	0.200	0.390

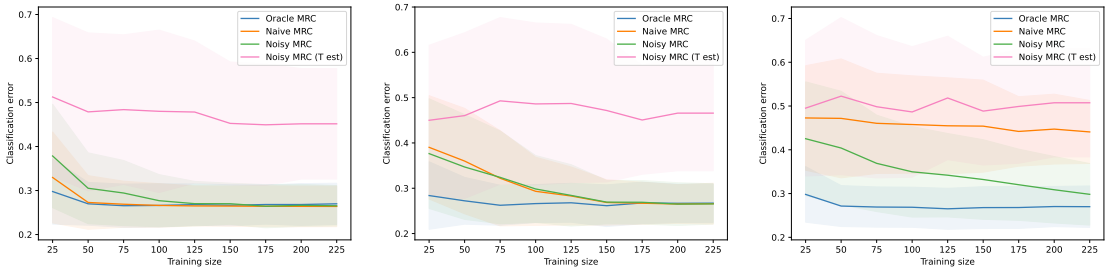
Table 24: $\rho_1 = 0.40,\ \rho_2 = 0.40$

Setting	Average Relative Error ± Standard deviation	$\hat{\rho}^{min}$	$\hat{\rho}^{max}$
$\hat{\rho}_1$	0.179 ± 0.089	0.382	0.553
$\hat{\rho}_2$	0.097 ± 0.078	0.320	0.510

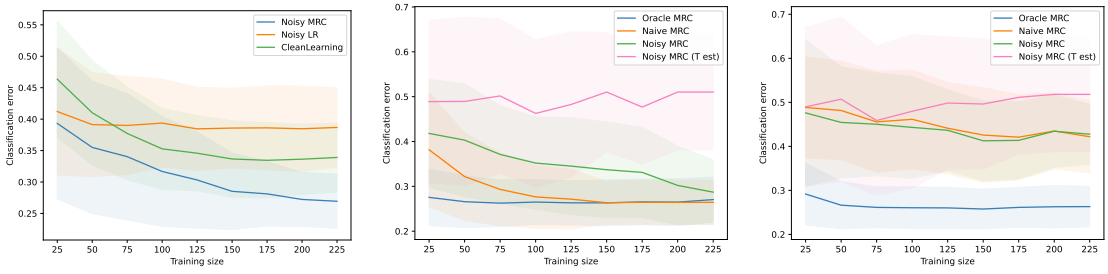
D.1.3. Haberman's Survival Dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$



(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

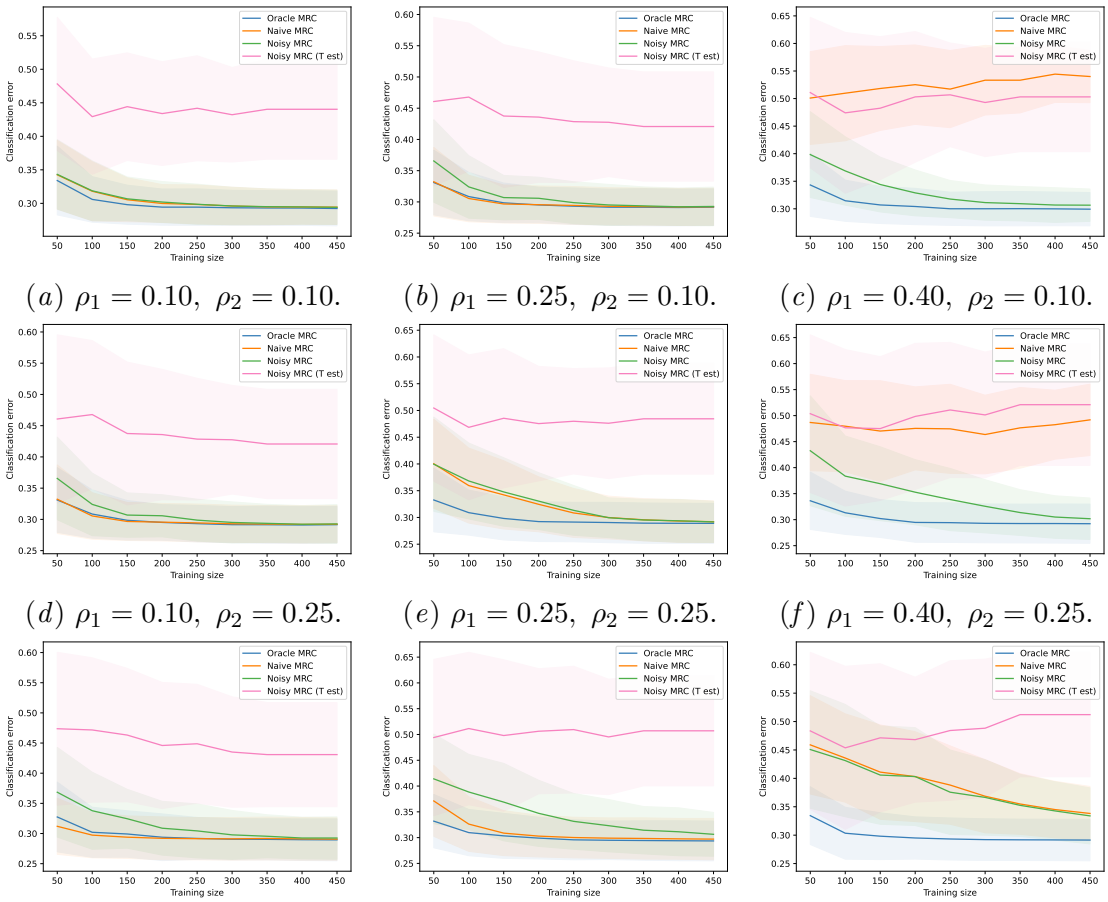


(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 17: Experiment with T unknown and evaluation on clean labels. Classification error of MRCs for different $\rho_1,\ \rho_2$. Performances of `Noisy MRC` trained with $\hat{T}$ estimated (pink) are poorer than with T known (green).

The performances of `Noisy MRC` trained with the estimated matrix $\hat{T}$ (pink) exhibit significantly poorer results compared to those trained with the true matrix T (green). This discrepancy can be attributed to the fact that the method employed for estimating the matrix T does not work well with a training dataset of such small size.

D.1.4. Indian Liver Patient dataset:



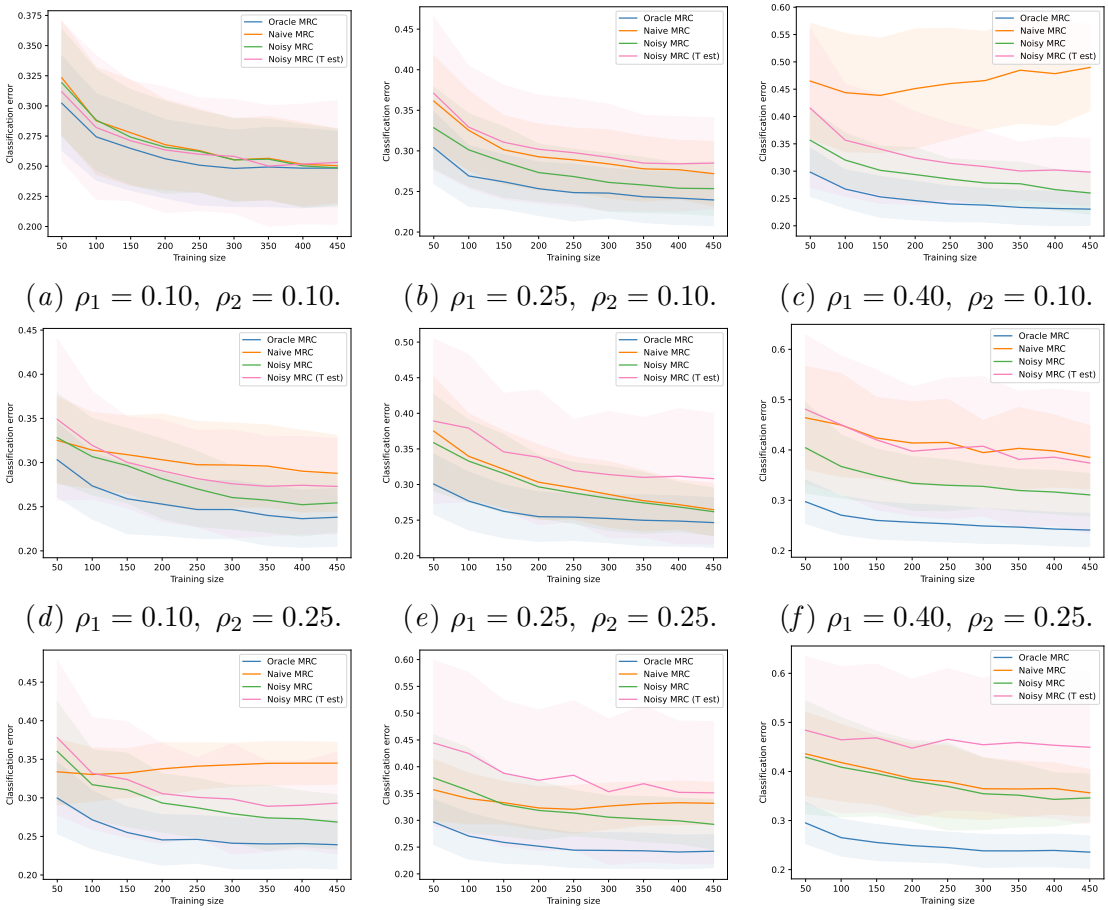
(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 18: Experiment with T unknown and evaluation on clean labels. Classification error of MRCs for different $\rho_1,\ \rho_2$. Performances of `Noisy MRC` trained with $\hat{T}$ estimated (pink) are poorer than with T known (green) due to limited size of the datasets.

D.1.5. Pima Indians Diabetes dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

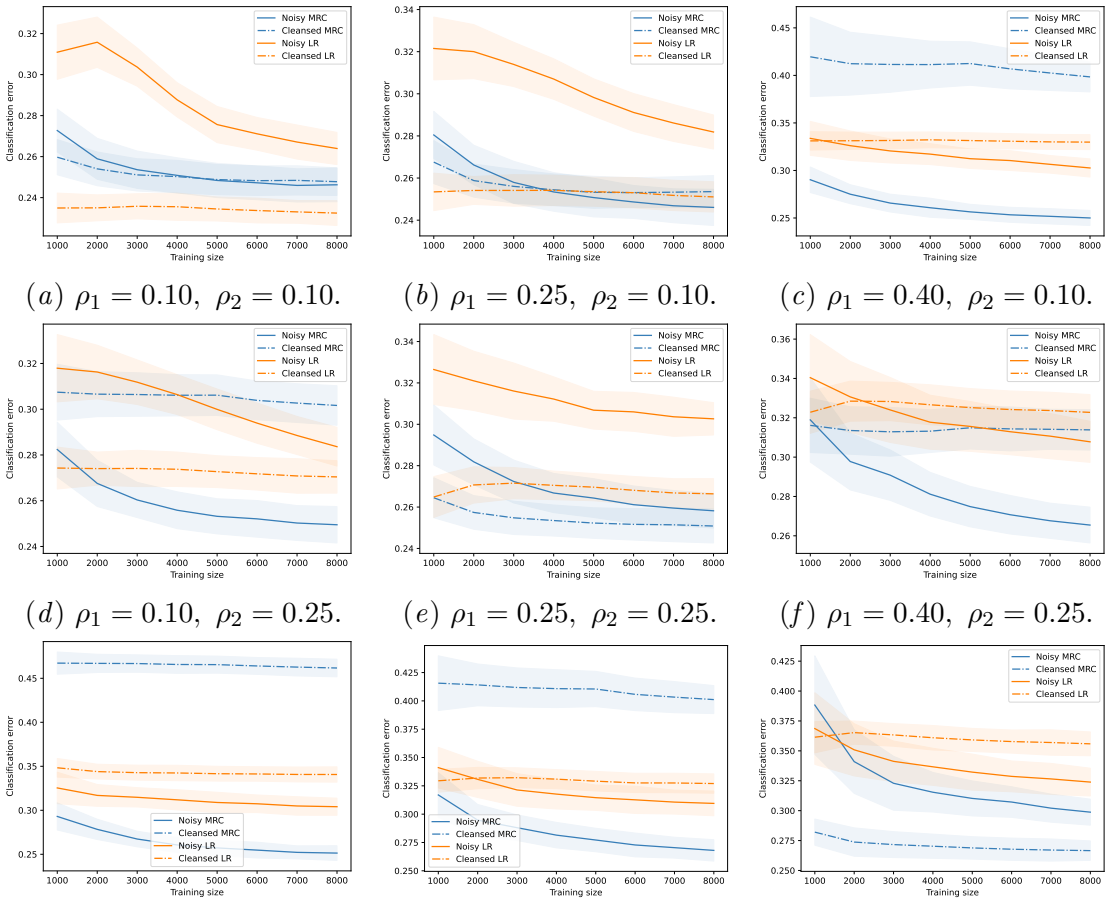
(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 19: Experiment with T unknown and evaluation on clean labels. Classification error of MRCs for different $\rho_1,\ \rho_2$. Performances of `Noisy MRC` trained with $\hat{T}$ estimated (pink) and with T known (green) are in general consistent, although subject to higher performance variability.

D.2. Learning with cleansed labels

Here we present additional results regarding the scenario where the matrix T is unknown and we cleanse the labels with `cleanlab` before applying a naive version of the classifiers.

D.2.1. ICU Mortality Dataset:



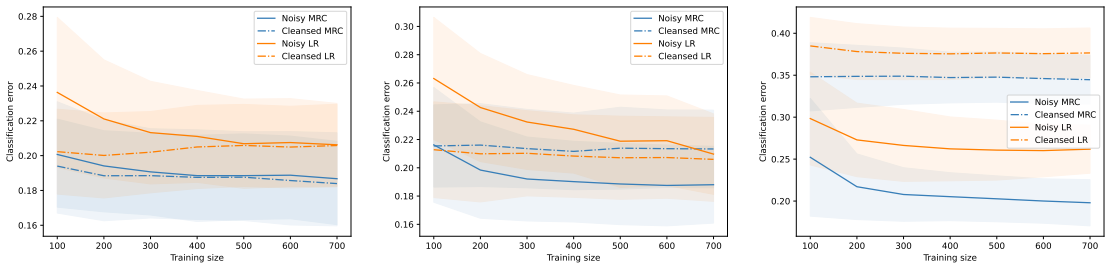
(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

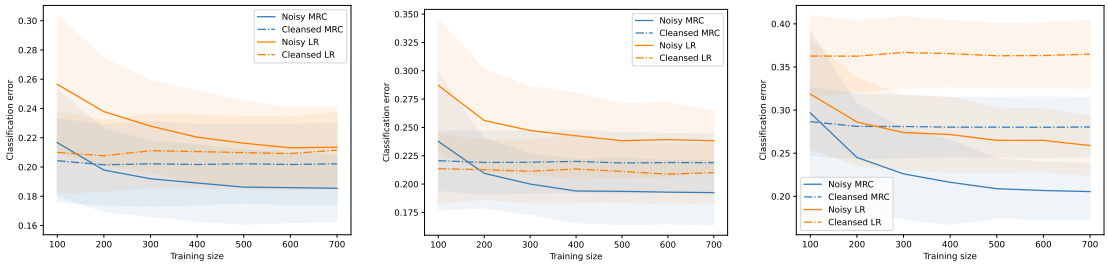
(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 20: Experiment with T unknown and evaluation on clean labels. Classification error of methods trained on *cleansed* labels, for different ρ_1, ρ_2. Performance gap when using `cleanlab` is very dependent on the noise rates.

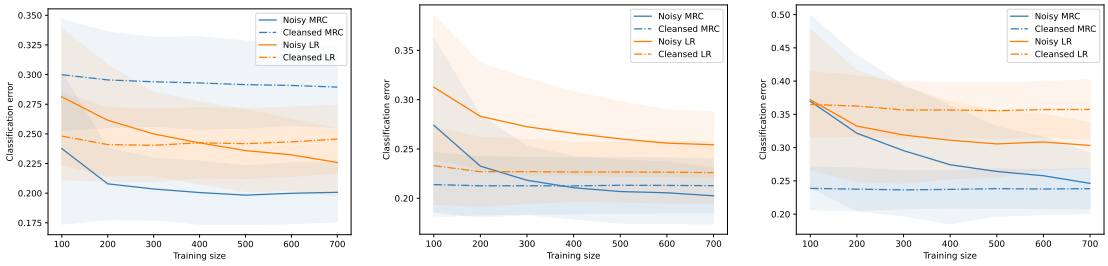
D.2.2. Mammographic Mass Dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$



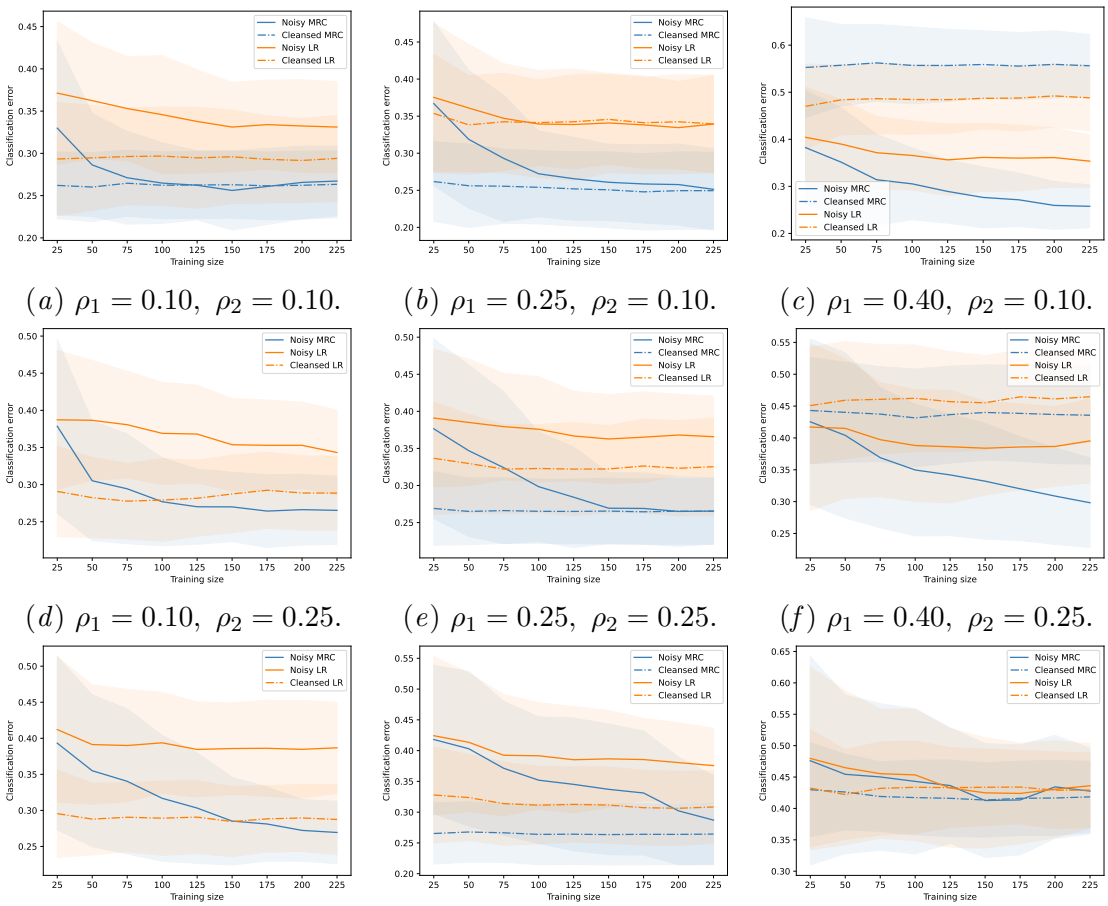
(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$



(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 21: Experiment with T unknown and evaluation on clean labels. Classification error of methods trained on *cleansed* labels, for different $\rho_1,\ \rho_2$. Performance gap when using `cleanlab` is very dependent on the noise rates.

D.2.3. Haberman's Survival Dataset:



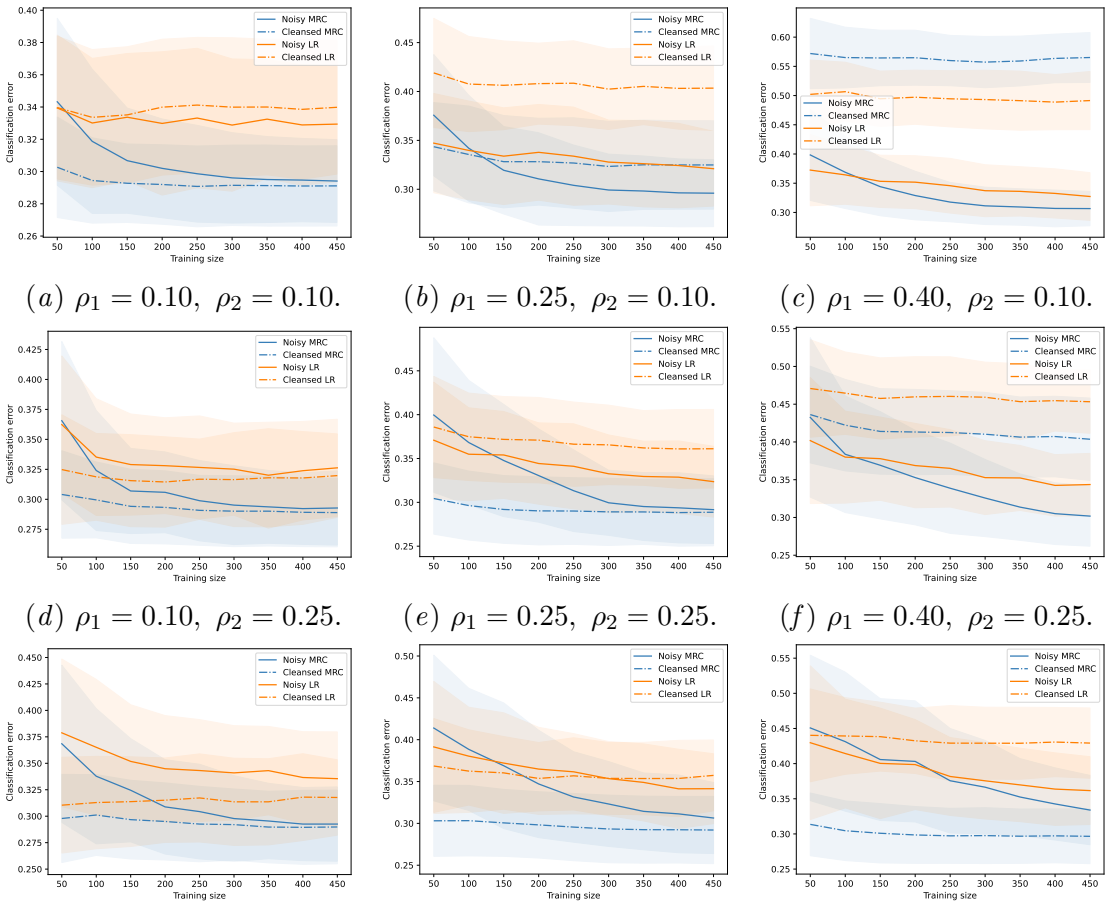
(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 22: Experiment with T unknown and evaluation on clean labels. Classification error of methods trained on *cleansed* labels, for different ρ_1, ρ_2. Performance gap when using `cleanlab` is very dependent on the noise rates.

D.2.4. Indian Liver Patient dataset:



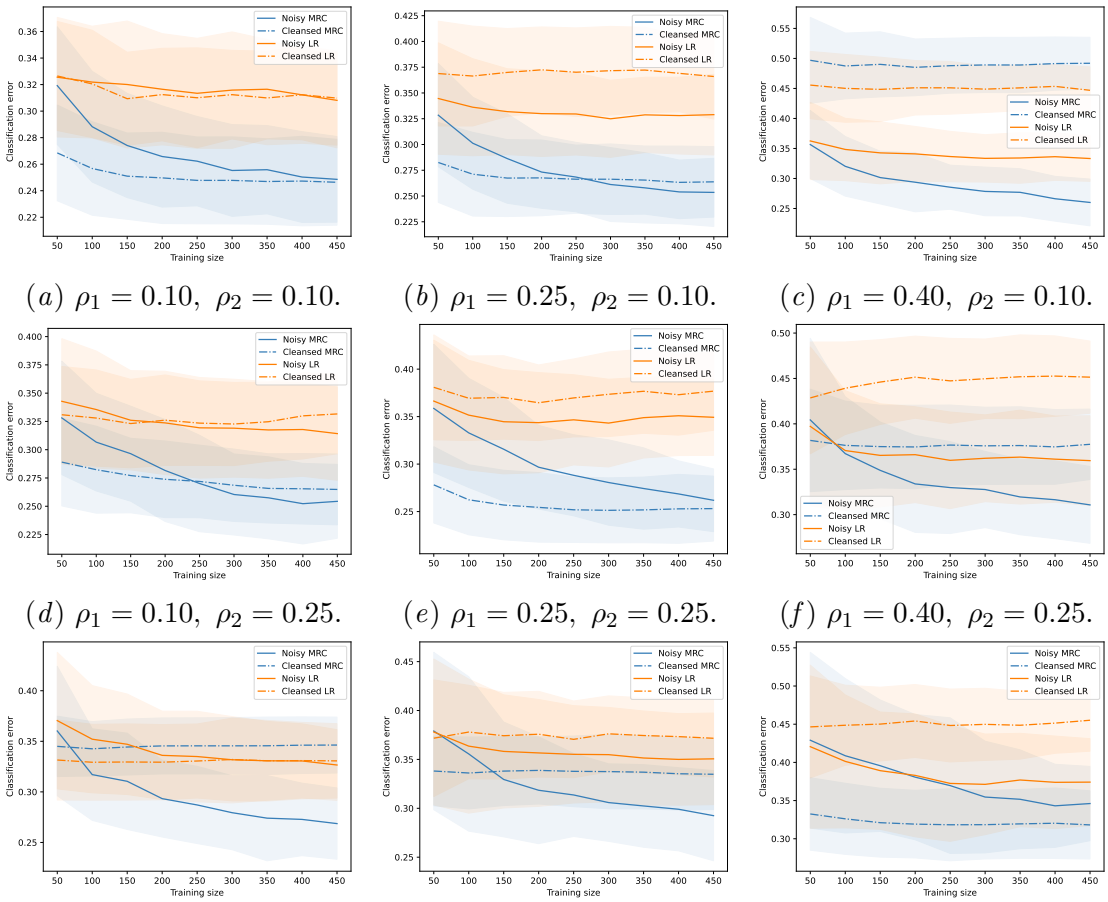
(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 23: Experiment with T unknown and evaluation on clean labels. Classification error of methods trained on *cleansed* labels, for different $\rho_1,\ \rho_2$. Performance gap when using `cleanlab` is very dependent on the noise rates.

D.2.5. Pima Indians Diabetes dataset:



(a) $\rho_1 = 0.10,\ \rho_2 = 0.10.$ (b) $\rho_1 = 0.25,\ \rho_2 = 0.10.$ (c) $\rho_1 = 0.40,\ \rho_2 = 0.10.$

(d) $\rho_1 = 0.10,\ \rho_2 = 0.25.$ (e) $\rho_1 = 0.25,\ \rho_2 = 0.25.$ (f) $\rho_1 = 0.40,\ \rho_2 = 0.25.$

(g) $\rho_1 = 0.10,\ \rho_2 = 0.40.$ (h) $\rho_1 = 0.25,\ \rho_2 = 0.40.$ (i) $\rho_1 = 0.40,\ \rho_2 = 0.40.$

Figure 24: Experiment with T unknown and evaluation on clean labels. Classification error of methods trained on *cleansed* labels, for different $\rho_1,\ \rho_2$. Performance gap when using `cleanlab` is very dependent on the noise rates.