CALIBRATION-ENHANCED APPROACH TO BIASED POSITIVE-UNLABELED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Positive-Unlabeled (PU) learning addresses the problem of training a binary classifier using an incomplete training dataset consisting of a subset of labeled positive instances and a set of unlabeled instances, which contains both positive and negative examples. A practically relevant and recently intensively studied scenario is biased PU learning, where the labeling mechanism determining which positive examples receive labels may depend on the feature values of the instances. A naive method in PU learning, known as the non-traditional classifier (NTC), treats all unlabeled instances as negative. However, this leads to a biased estimation of the true class posterior probability. More effective learning from biased PU data is possible but challenging, as it typically requires estimating the propensity score, defined as the instance-specific probability that a positive example is labeled. Existing propensity score estimation methods are often computationally expensive, since they rely on iterative procedures that alternately update the classifier and the propensity score model. In this work, we show that under relatively weak functional assumption on the propensity score, it is possible to accurately estimate the class posterior without explicitly modeling the former, provided that a small calibration set is available. The proposed approach (called PUCal) is based on an appropriate calibration of probabilities produced by the NTC on a small set of fully observable observations. We show that such a procedure leads to a more accurate classifier compared to existing methods for biased PU learning even when only a very limited number of observations are available in the calibration set. We also provide estimation error bounds for the proposed method.

1 INTRODUCTION

1.1 MOTIVATION

Positive-unlabeled (PU) learning is a partially supervised learning framework in which the goal is to train a binary classifier using labeled data that are known to be positive and unlabeled data that may contain both positive and negative instances (Bekker and Davis, 2020). In recent years, PU learning has received considerable attention, motivated by the increasing number of practical applications in which such data structures naturally occur.

More specifically, PU data arise in situations where a set of confirmed positive cases is available, together with a potentially large number of unlabeled observations. For example, in social network analysis, labeled positive instances may correspond to user profiles reported as fake, whereas unlabeled instances may correspond to profiles that have not been reported as fake (Yuan et al., 2019). Importantly, the unlabeled group commonly includes both positive cases (i.e., fake profiles that remain undetected or unreported) and negative cases (i.e., genuine user profiles). Similar situations occur in disease reporting, adverse drug reactions (Cao et al., 2017; Lee and Chen, 2021), insurance claim analysis (Frees et al., 2014), fake reviews detection (Li et al., 2014) and many other applications (Li et al., 2021).

Learning from PU data is challenging because the true class labels are only partially observed. A naive approach treats all unlabeled observations as negative and uses the labeling indicator as the response variable (Elkan and Noto, 2008). This approach, commonly referred to as the non-traditional classifier (NTC), constitutes one of the baseline methods in PU learning. However, as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

expected, NTC provides biased estimates of the posterior probability of the true class, which often results in limited predictive performance. This issue becomes particularly severe when the labeling frequency, defined as the proportion of positive instances that receive labels, is low.

To address this problem, numerous dedicated PU learning methods have been proposed (Kiryo et al., 2017; Chen et al., 2020b; Sevetlidis et al., 2024; Chen et al., 2020a; Zhao et al., 2022; Dai et al., 2025; Gong et al., 2021; 2022; Bekker et al., 2019; Kato et al., 2019; Gerych et al., 2022; Coudray et al., 2023; Rejchel et al., 2024), based on different methodological principles and data generating mechanisms (Elkan and Noto, 2008; Bekker and Davis, 2020; Coudray et al., 2023; Mielniczuk and Wawrzeńczyk, 2024; Wang et al., 2026). A key challenge in PU learning is modeling the labeling mechanism, which is commonly characterized by the propensity score function, defined as the instance-specific probability that a positive observation is labeled. In general and realistic settings, the propensity score depends on the observed features, corresponding to the Selected At Random (SAR) assumption (Bekker and Davis, 2020). For example, the probability that a disease is reported for an affected individual may depend on various factors, such as age, the presence of comorbidities, or access to healthcare services. Consequently, the labeled observations constitute a biased sample of the positive population, giving rise to the biased PU learning problem (Gong et al., 2021; 2022; Bekker et al., 2019; Kato et al., 2019; Gerych et al., 2022; Coudray et al., 2023; Rejchel et al., 2024).

Biased PU learning is challenging, both conceptually and computationally, because it requires estimation of the propensity score. The main difficulty arises from the interplay between the two components: estimation of the propensity score necessitates knowledge of the true class, whereas prediction of the true class probability requires knowledge of the propensity score. This circular dependency has motivated iterative approaches, in which models for the propensity score and the posterior probability of the true class are alternately estimated and updated (Bekker et al., 2019; Gong et al., 2021; Rejchel et al., 2024; Tang et al., 2025). Although such methods can provide effective solutions, they are computationally demanding. An alternative strategy relies on imposing restrictive assumptions on the form of the propensity score (Gerych et al., 2022), which may limit the applicability of the resulting models.

1.2 Contribution

In this work, we propose a different approach from listed above. We consider a general assumption regarding the labeling mechanism, referred to as the Function of Posterior (FOP) assumption (Teisseyre et al., 2025), which states that the propensity score is a function of the posterior probability of the true class. This assumption generalizes several assumptions considered in the PU learning literature, including SCAR (Selected Completely at Random), IOO (Invariance of Order), PG (Probabilistic Gap), LIN (Linear), and SIGMOID (Sigmoid). The last two correspond to the functional forms of the propensity score, namely the linear and sigmoidal forms, respectively. The FOP assumption is natural, as the probability of being labeled and the probability of belonging to the positive class are often expected to be determined by the same underlying features. For example, profiles whose characteristics and content suggest violations of the law or the presence of misinformation are likely to be reported more frequently as inappropriate.

Our approach is based on a simple observation with important practical implications. We show that, under FOP assumption, the posterior probability of the positive class can be expressed as a transformation of the posterior probability of the labeled class. This transformation can be estimated using a relatively small fully labeled calibration dataset. We further show that state-of-the-art calibration methods, such as Beta Calibration and Isotonic Regression, can effectively learn this transformation. This setting arises naturally in applications where large amounts of historical PU data are readily available, whereas obtaining fully labeled data is possible, but substantially more costly. For instance, positive instances may be identified through user reports or other naturally occurring signals indicating potentially violating webpages, profiles, or content. In contrast, obtaining reliable negative labels typically requires systematic manual or expert verification, making the construction of a fully labeled dataset expensive. The resulting setup reflects a natural separation of roles: the large historical PU dataset provides the main source of information for learning, while the small fully labeled sample is used exclusively for post-hoc calibration.

The proposed method achieves higher predictive power compared with traditional biased PU learning approaches that require direct estimation of the propensity score. A limitation of our method is that it relies on the FOP assumption, which may be violated in certain specific scenarios, for example, when the factors governing the labeling mechanism are different from those determining class membership. Nevertheless, our experimental results demonstrate that the proposed method remains effective even in such cases.

Furthermore, we provide theoretical analysis of the proposed framework, including an estimation error bound. These analyses show, as expected, that the accuracy of posterior probability estimation depends on both the quality of the posterior probability estimates for the class indicator provided by NTC and the accuracy of the estimated transformation function.

2 BACKGROUND

2.1 BIASED PU LEARNING

Let $\mathbf{x} \in \mathcal{X}$ be the feature vector and $y \in \{0, 1\}$ the true class label (with $y = 1$ indicating the positive class). Moreover, we consider label indicator $s \in \{0, 1\}$, which indicates whether the instance is labeled ($s = 1$) or unlabeled ($s = 0$). In Positive-Unlabeled (PU) learning, each instance is represented by the triple $(\mathbf{x}, y, s)$, generated from some unknown density $p(\mathbf{x}, y, s)$ and it is assumed that negative examples cannot be labeled, i.e., $p(s = 1|y = 0, \mathbf{x}) = 0$. Therefore, $s = 1$ denotes a labeled positive instance, and $s = 0$ denotes an unlabeled instance, which could either be positive or negative.

We assume that independent and identically distributed (iid) random vectors $(\mathbf{x}_i, y_i, s_i)$ for $i = 1, \ldots, n$ are drawn from an unknown distribution $p(\mathbf{x}, y, s)$. The PU training data is given by $\mathcal{D}_{PU} = \{(x_i, s_i)\}_{i=1}^{n}$, since y_i is not directly observed. In the context of classification, the objective is to train a classifier that predicts y for a new instance $\mathbf{x}$, using only the incompletely labeled training set $\mathcal{D}_{PU}$. Adopting a probabilistic perspective, the goal in PU learning is to estimate the posterior probability $y(\mathbf{x}) = p(y = 1|\mathbf{x})$, which allows us to define an appropriate decision rule. We assume that $y(\mathbf{x}) \in (0, 1)$. Note that the NTC classifier estimates $s(\mathbf{x}) = p(s = 1|\mathbf{x})$, which differs from $y(\mathbf{x})$, or more precisely, $s(\mathbf{x})$ underestimates $y(\mathbf{x})$.

A key element necessary for accurately estimating $y(\mathbf{x})$ is the propensity score function. The propensity score function, defined as

$$e(x) = p(s = 1|\mathbf{x}, y = 1),$$

is the instance-specific probability that a positive example's label will be observed. It models the labeling mechanism underlying PU data. A high value of the propensity score indicates that a positive observation described by vector $\mathbf{x}$ will be assigned a label with a high probability.

A common approach to estimate $y(\mathbf{x})$, used in classical PU learning, is to apply an Empirical Risk Minimization (ERM) approach. Let us consider the most commonly used loss function, namely the cross-entropy loss. In the ERM approach, the goal is to minimize the empirical risk function, which for the cross-entropy corresponds to the theoretical risk

$$\mathcal{R}(y_\theta) = -\mathbb{E}_{\mathbf{x},y}[y \log y_\theta(\mathbf{x}) + (1 - y) \log(1 - y_\theta(\mathbf{x}))], \tag{1}$$

where $y_\theta(\mathbf{x})$ is the output of some parametric model characterized by parameter vector θ. Direct optimization of the empirical equivalent of (1) is not possible, because we do not have access to the true values of y. However, the risk function (1) can be written as a weighted risk function (Bekker and Davis, 2020)

$$\mathcal{R}(y_\theta) = p(s = 1)E_{\mathbf{x}|s=1}\left[\frac{1}{e(\mathbf{x})}\big(l(1, y_\theta(\mathbf{x})) - l(0, y_\theta(\mathbf{x}))\big)\right] + E_{\mathbf{x}} l(0, y_\theta(\mathbf{x})).$$

Importantly, the empirical counterpart of the above risk, can be optimised, provided that $e(\mathbf{x})$ is known and the final estimator is $\widehat{y}(\mathbf{x}) = \arg\min_{y_\theta \in \mathcal{G}} \widehat{\mathcal{R}}(y_\theta)$. In practice, the propensity score is not observed and must be estimated from the data, which poses a substantial challenge, as we have already noted in the Introduction. The objective of this work is to develop a method for estimating $y(\mathbf{x})$ without requiring the computationally expensive estimation of $e(\mathbf{x})$.

2.2 ASSUMPTIONS ON LABELING MECHANISM

In the PU learning literature, various assumptions have been considered regarding the labeling mechanism, i.e., the form of the propensity score, that enable classifier learning. Early works (Elkan, 2001) adopted the highly restrictive SCAR assumption, under which the propensity score is constant $e(\mathbf{x}) = c$ that is, the labeling mechanism is independent of the feature variables. Another studied assumption is that the propensity score is linearly related to the posterior probability (Gerych et al., 2022). Throughout this paper, we will refer to this assumption as the LIN assumption. It states that $e(\mathbf{x}) = k \cdot y(\mathbf{x})$. We note that LIN does not include SCAR as a special case. It therefore seems more reasonable to assume e.g. that $e(\mathbf{x}) = \sigma(a + k \cdot y(\mathbf{x}))$, for some parameters a, k. We will refer to this assumption as the SIGMOID assumption. It reduces to SCAR for $k = 0$.

Kato et al. (2019) introduced the Invariance of Order (IOO) assumption, which stipulates that $s(\mathbf{x})$ and $y(\mathbf{x})$ induce the same order on the feature space: $s(\mathbf{x}_1) < s(\mathbf{x}_2) \iff y(\mathbf{x}_1) < y(\mathbf{x}_2)$. The Probabilistic Gap (PG) (He et al., 2018) assumption requires that $e(\mathbf{x})$ is a monotonically increasing function of $y(\mathbf{x})$. In this work, we consider the following more general assumption.

Assumption (Function of Posterior (FOP)). *The propensity score function is given by* $e(\mathbf{x}) = g(y(\mathbf{x}))$*, for some function* $g : (0, 1) \to (0, 1)$.

The FOP assumption has previously been considered in PU learning, although in a different context from the one studied in this work. Namely, it has been shown (Płatek and Mielniczuk, 2023; Teisseyre et al., 2025) that when the true classification function is linear and the feature distribution is elliptically contoured, the NTC classifier enables the recovery of the direction of the linear decision boundary separating the classes, with only the intercept requiring estimation. In the present work, we demonstrate novel and interesting implications of this assumption.

The FOP assumption encompasses various popular assumptions used in PU learning (SCAR, LIN, SIGMOID, IOO, PG). For example, the IOO assumption constitutes a special case of FOP, obtained by selecting the function g such that $w(t) = t \cdot g(t)$ is monotonically increasing. The FOP assumption implies that the features governing the posterior probability of the positive class also influence the labeling mechanism. This assumption is natural in many real-world applications. For instance, in medical diagnosis, a high probability of disease is often associated with observable symptoms, which subsequently increases the likelihood that the disease is reported, detected, or diagnosed. Naturally, one can also consider settings in which the FOP assumption does not hold. It implies that observations with identical posterior probabilities, i.e., $y(\mathbf{x}_1) = y(\mathbf{x}_2)$, must have identical propensity scores, i.e., $e(\mathbf{x}_1) = e(\mathbf{x}_2)$. This requirement may be violated in scenarios where the labeling mechanism depends on additional features that affect the probability of being labeled but do not influence the target variable. A simple parametric example illustrating both cases, where the FOP assumption holds and where it does not, is given in the Appendix A.4.

3 LEARNING VIA CALIBRATION

3.1 CALIBRATING NON-TRADITIONAL CLASSIFIER

The following theorem provides the theoretical foundation for the approach proposed in this work.

Theorem 1. *Let us assume that the FOP assumption is satisfied for such function* $g : (0, 1) \to (0, 1)$*, that* $w(t) = t \cdot g(t)$ *is strictly monotonic. Then, there exists function* $h : (0, 1) \to (0, 1)$ *such that* $y(\mathbf{x}) = h(s(\mathbf{x}))$.

Proof. Using the law of total probability, the fact that in PU learning negative instances cannot be labeled, and the FOP assumption, we can write:

$$s(\mathbf{x}) = p(s = 1|\mathbf{x}) = y(\mathbf{x})e(\mathbf{x}) = y(\mathbf{x})g(y(\mathbf{x})) = w(y(\mathbf{x})).$$

By assumption, the function $w(t)$ is invertible. Defining $h(t) = w^{-1}(t)$ immediately yields the assertion of the theorem. □

As the case when $w(t) = t \cdot g(t)$ is strictly increasing is more common, we note that it is weaker than the analogous property that $g(t)$ is strictly increasing. Indeed, the last amounts to $tw'(t) - w(t) > 0$

for $t \in (0,1)$, whereas the first to $w'(t) > 0$. Further on, we assume that Theorem 1 holds.
In this work, we assume that, in addition to the primary PU dataset, a small fully labeled calibration dataset $\mathcal{D}_{cal} = \{(\mathbf{x}_i^{cal}, y_i^{cal})\}_{i=1}^{n_{cal}}$ is available. However, we assume that this dataset is not accessible during the training of the PU classifier and can only be utilized post hoc. This setting reflects a realistic scenario in which the primary model is trained on a large historical PU dataset, whereas the calibration data may, for example, originate from a pilot study and are available only in limited quantities. The proposed procedure (called PUCal) consists of three steps.

Step 1. We estimate $s(\mathbf{x})$ using the PU dataset $\mathcal{D}_{PU}$ and the NTC (naive method). More precisely, we define the estimator $\widehat{s}(\mathbf{x}) = \arg\min_{s_\theta \in \mathcal{F}} \widehat{\mathcal{R}}_{ntc}(s_\theta)$, which is the mimimizer of the empirical risk $\widehat{\mathcal{R}}_{ntc}(s_\theta)$ corresponding to the theoretical risk of the form

$$\mathcal{R}_{ntc}(s_\theta) = -E_{\mathbf{x},s}[s \log s_\theta(\mathbf{x}) + (1-s)\log(1-s_\theta(\mathbf{x}))]$$

Step 2. Given set $\mathcal{D}_{cal}$ and $\widehat{s}(\mathbf{x})$, we define set $\widetilde{\mathcal{D}}_{cal} = \{(\widehat{s}(\mathbf{x}_i^{cal}), y_i^{cal})\}_{i=1}^{n_{cal}}$. Subsequently, we estimate function h using data $\widetilde{\mathcal{D}}_{cal}$. Specifically, we define the estimator $\widehat{h}(\mathbf{x}) = \arg\min_{h_\theta \in \mathcal{H}} \widehat{\mathcal{R}}_{cal}(h_\theta)$, which is the mimimizer of the empirical risk $\widehat{\mathcal{R}}_{cal}(h_\theta)$ corresponding to the theoretical risk of the form

$$\mathcal{R}_{cal}(h_\theta) = -E_{\mathbf{x},y}[y \log h_\theta(\widehat{s}(\mathbf{x})) + (1-y)\log(1-h_\theta(\widehat{s}(\mathbf{x}))]$$

Step 3. Given $\widehat{s}(\cdot)$ from Step 1 and $\widehat{h}(\cdot)$ from Step 2, we define a calibrated estimator $\widehat{y}(\mathbf{x}) := \widehat{h}(\widehat{s}(\mathbf{x}))$. The specific methods for estimating the function h are described in Section 3.3.

Note that this procedure is particularly attractive for n_{cal} much smaller than n when low dimensional parametric family is used for calibration, see Remark 1 below Theorem 4. The additional justification of the above algorithm is provided by the following result, which shows that calibration of the posterior probability $s(\mathbf{x})$ yields the unobservable posterior probability $y(\mathbf{x})$. Thus, $\widehat{s}(\mathbf{x})$ can be regarded as an uncalibrated estimator of $y(\mathbf{x})$, which is subsequently calibrated using an auxiliary calibration set $\mathcal{D}_{cal}$.

Theorem 2. *Assume that Theorem 1 is valid and let $m(x)$ be a function calibrating $s(x)$ at theoretical level i.e. $p(y=1|s(\mathbf{x})) = m(s(\mathbf{x}))$. Then $m(s(\mathbf{x})) = y(\mathbf{x})$. Moreover, if the image $s((0,1)) = (0,1)$ then $m \equiv h$.*

Proof. The first statement follows from the following sequence of equalities

$$m(s(\mathbf{x})) = p(y=1|s(\mathbf{x})) = p(y=1|h(s(\mathbf{x}))) = p(y=1|y(\mathbf{x})) = y(\mathbf{x}),$$

where the first equality is definition of function m, the second follows from strict monotonicity of h, the third is due to Theorem 1 and the last one to the property that posterior is perfectly calibrated i.e. $p(y=1|y(\mathbf{x})=p) = p$ (see e.g. (Sun et al., 2023), eq. (7)). The second claim follows from the first after noting that $s(\cdot)$ takes any possible value in $(0,1)$. □

3.2 Estimation Error Bounds

We now state theoretical results for estimators $\widehat{y}(\mathbf{x})$ of $y(\mathbf{x})$ based on Empirical Risk minimizers of s and h. As our main aim here is estimation of the true posterior for the PU data, they concern expected absolute difference between $\widehat{y}(\mathbf{x})$ and $y(\mathbf{x})$ i.e. $E_{\mathbf{x}}|y(\mathbf{x}) - \widehat{y}(\mathbf{x})|$. Note that this is a random quantity as a function of $\widehat{y}$. Let

$$r_n^{\mathcal{F}}(\widehat{s}) := \mathcal{R}_{ntc}(\widehat{s}) - \min_{s_\theta \in \mathcal{F}} \mathcal{R}_{ntc}(s_\theta)$$

be regret of the risk of $\widehat{s}$ pertaining to $\mathcal{F}$. Regret is widely used predictive measure of performance of ERM classifiers, see e.g. Wainwright (2019). $\mathcal{F}$ is defined as a family of functions s_θ from $\mathcal{X}$ with values in $(0,1)$, where $\theta \in \Theta$ may be finite-dimensional (parametric case) or infinite-dimensional (nonparametric case). Analogously

$$r_{n_{cal}}^{\mathcal{H}}(\widehat{h}) := \mathcal{R}_{cal}(\widehat{h}) - \min_{h_\theta \in \mathcal{H}} \mathcal{R}_{cal}(h_\theta)$$

be the regret of $\widehat{h}$ corresponding to $\mathcal{H}$, where $\mathcal{H} = \{h_\theta\}$ is parametric family of functions from (0,1) to (0,1). Let $\widetilde{h}(\widehat{s}(\mathbf{x})) = p(y=1|\widehat{s}(\mathbf{x}))$ and $\widetilde{h}(z) = \widetilde{h}(\widehat{s}(\mathbf{x}))$ for $z = \widehat{s}(\mathbf{x})$. The following theorem holds. Its proof and the proofs of the following results involve careful examination of the fact that estimator $\widehat{s}$ is $\mathcal{D}_{PU}$-measurable random variable.

Theorem 3. *Assume that h is differentiable with bounded derivative h', moreover, $s \in \mathcal{F}$ and $\widetilde{h} \in \mathcal{H}$. Then conditionally on $\mathcal{D} = \mathcal{D}_{PU} \cup \widetilde{\mathcal{D}}_{cal}$ we have*

$$E_{\mathbf{x}}|y(\mathbf{x}) - \widehat{y}(\mathbf{x})| \leq ||h'||_{\infty}\sqrt{2 \times r_n^{\mathcal{F}}(\widehat{s})} + \sqrt{r_{n_{cal}}^{\mathcal{H}}(\widehat{h})/2},$$

where $||h'||_{\infty} = sup_{s\in(0,1)}|h'(s)|$.

The proof of the Theorem is relegated to the appendix. Here we prove only the crucial Lemma showing that recalibration can not amplify the distortion introduced by replacing s by its estimator $\widehat{s}$.

Lemma 1. *We have that*

$$E_{\mathbf{x}}|\widetilde{h}(\widehat{s}(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))| \leq E_{\mathbf{x}}|h(s(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))|. \tag{2}$$

Proof. Note that conditionally on $\mathcal{D}_{PU}$ we have

$$\widetilde{h}(\widehat{s}(\mathbf{x})) = E(y|\widehat{s}(\mathbf{x})) = E[E(y|\mathbf{x})|\widehat{s}(\mathbf{x})] = E(h(s(\mathbf{x}))|\widehat{s}(\mathbf{x})), \tag{3}$$

where the second equality follows from the tower property, since given $\mathcal{D}_{PU}$, $\widehat{s}(\mathbf{x})$ is a deterministic function of $\mathbf{x}$, and the third follows from Theorem 1. Thus we have that

$$\Big|\widetilde{h}(\widehat{s}(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))\Big| = \Big|E[h(s(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))|\widehat{s}(\mathbf{x})]\Big| \leq E\Big[\Big|h(s(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))\Big||\widehat{s}(\mathbf{x})\Big]$$

and thus taking expectations wrt to $\mathbf{x}$ on both sides we have

$$E_{\mathbf{x}}|\widetilde{h}(\widehat{s}(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))| \leq E_{\mathbf{x}}|h(s(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))|.$$

□

The following Theorem proved in the Appendix is a consequence of the above result and, e.g. Theorem 4.10 in (Wainwright, 2019).

Theorem 4. *Let $\mathrm{Rad}_n(\mathcal{F})$ be Rademacher complexity of $\mathcal{F}$ of order n for distribution $p(\mathbf{x}, s)$ and $\mathrm{Rad}_{n_{cal}}(\mathcal{H})$ Rademacher complexity of $\mathcal{H}$ of order n_{cal} for distribution of $p(\widehat{s}(\mathbf{x}), y)$ with fixed $\widehat{s}$. Moreover, assume that conditions of Theorem 3 are satisfied. Then*

$$E_{\mathbf{x}}|\widehat{y}(\mathbf{x}) - y(\mathbf{x})| = \mathcal{O}\big(\sqrt{\mathrm{Rad}_n(\mathcal{F})}||h'||_{\infty} + \sqrt{\mathrm{Rad}_{n_{cal}}(\mathcal{H})}\big) \tag{4}$$

almost surely.

We also note that if $\mathcal{H}$ is of polynomial discrimination or has finite Vapnik-Chervonenkis (VC) dimension, its Rademacher complexity does not depend on underlying distribution $p(\widehat{s}(\mathbf{x}), y)$ and satisfies $\mathrm{Rad}_{n_{cal}}(\mathcal{H}) \leq C(\log n/n)^{1/2}$ for some absolute constant $C > 0$. This easily follows from lemma 4.41 in (Wainwright, 2019). Thus we have

Corollary 1. *Let $\mathcal{F}$ and $\mathcal{H}$ be classes of polynomial discrimination. Then the order of $E_{\mathbf{x}}|\widehat{y}(\mathbf{x}) - y(\mathbf{x})|$ is $\mathcal{O}\big(\log n/n)^{1/4}||h'||_{\infty} + (\log n_{cal}/n_{cal})^{1/4}\big)$ almost surely.*

Remark 1. *For parametric classes $\mathcal{H}$, such as those considered here (e.g., beta calibration or Platt scaling), we have $\mathrm{Rad}_{n_{cal}}(\mathcal{H}) = \mathcal{O}(n_{cal}^{-1})$. Moreover, in view of M-estimation approach (see e.g. (van der Vaart, 1998), Chapter 5), in the case of parametric $\mathcal{H}$ and assuming the boundedness of the Hessian of the risk, it is conjectured that the second term in (4) can be replaced by $\mathcal{O}(1/n_{cal})$. Thus, estimation within $\mathcal{H}$ typically constitutes a simpler problem, resulting in a faster convergence rate to zero. This advantage is, however, counterbalanced by the smaller sample size, since, commonly, $n_{cal} < n$.*

The above theorems involve a constant $||h'||_{\infty}$, which can be explicitly calculated under certain assumptions regarding the labeling mechanism, such as SCAR or SIGMOID. Indeed, under SIGMOID assumption $e(\mathbf{x}) = \sigma(a + ky(\mathbf{x}))$ discussed in Section 2.2, we have that $w(s) = s \cdot g(s) = s\sigma(a + k \cdot s)$ and $w'(s) = \sigma(a + k \cdot s)(1 + k \cdot s(1 - \sigma(a + k \cdot s))$. Thus, for $k \geq 0$ the minimal value of $w'(s)$ on $(0, 1)$ is $\sigma(a)$ and it follows that $||h'||_{\infty} = 1/w'(0) = 1/\sigma(a)$.

3.3 BASELINE METHODS FOR ESTIMATING THE FUNCTION h

As indicated in the previous section, the effectiveness of the proposed method depends on the quality of the estimation of $s(\cdot)$ using NTC, as well as on the quality of the estimation of the function $h(\cdot)$. Below, we discuss possible approaches for estimating the calibrating function $h(\cdot)$. A natural solution, commonly used in calibration, is the beta calibration method (Kull et al., 2017), in which we assume that

$$y(\mathbf{x}) = h(s(\mathbf{x})) = \sigma[a \log(s(\mathbf{x})) + b \log(1 - s(\mathbf{x})) + c],$$

where $\sigma(\cdot)$ is sigmoid activation function and the parameters a, b, c are estimated using $\tilde{\mathcal{D}}_{cal}$. We note that h is increasing when $a > 0$ and $b < 0$, thus in this case assumptions of Theorem 1 are satisfied. The method reduces to other popular calibration methods under additional constraints on the parameters. Notice that for $b = -a$, we obtain Platt scaling $h(s(\mathbf{x})) = \sigma[a \cdot \text{logit}(s(\mathbf{x})) + c]$ (Platt, 1999), while imposing the additional constraint $c = 0$ results in temperature scaling $h(s(\mathbf{x})) = \sigma[\text{logit}(s(\mathbf{x}))/T]$, with $T = a^{-1}$ (Guo et al., 2017). For Platt scaling, h is an increasing function. We also include Isotonic Regression as one of the state-of-the-art baseline methods for which monotonicity of the solution is imposed (de Leeuw et al., 2009). Figure 1 illustrates the performance of the proposed PUCal method in combination with beta calibration and isotonic regression.

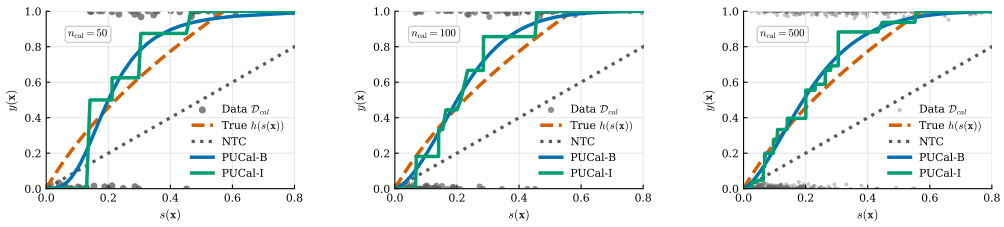

Figure 1: Illustration of the goodness of approximation of the proposed PUCal method on synthetic data with known posterior probability $y(\mathbf{x})$ and propensity score $e(\mathbf{x}) = \sigma(a + y(\mathbf{x}))$, where a is chosen to ensure that $p(s = 1|y = 1) = 0.5$. Results are shown for three calibration set sizes $\mathcal{D}_{\text{cal}}$: 50, 100, and 500. Two variants are considered: PUCal-B (Beta calibration) and PUCal-I (isotonic regression).

4 EXPERIMENTS

In the experimental study, we evaluate the effectiveness of the proposed PUCal method. Within PUCal, we investigate four calibration methods for estimating the function h: Beta Calibration (PUCal-B), Isotonic Regression (PUCal-I), Platt Scaling (PUCal-P), and Temperature Scaling (PUCal-T).

Our main objective was to verify how PUCal performs compared with methods based on propensity score estimation. Therefore, as competing approaches, we consider state-of-the-art methods designed for biased PU learning that rely on propensity score estimation, namely SAREM (Bekker et al., 2019), LBE (Gong et al., 2021; 2022), and JERM (Rejchel et al., 2024). In addition, we include the NTC method as a baseline, as well as an ORACLE classifier, which assumes access to the true class labels and therefore represents an upper performance bound.

Since SAREM, JERM, and LBE do not make use of calibration data, we also evaluate their calibrated variants to ensure a fair comparison. For example, LBE-B and LBE-I denote the LBE method combined with Beta Calibration and Isotonic Regression, respectively, where the calibration model is fitted using the calibration set.

4.1 EXPERIMENTAL FRAMEWORK

Training (45%) and test (45%) subsets are drawn from each considered dataset. From the remaining 10% of observations we draw a calibration set of size n_{cal}. The description of the datasets and their preprocessing details are provided in the Appendix A.2. Following the common practice in

biased PU learning studies, we generate PU data from the fully labeled training set using a predefined labeling strategy. Specifically, positive instances are selected according to the probability distribution defined by the propensity score function. We consider six different labeling strategies: **Strategy S1**: $e(\mathbf{x}) = a$; **Strategy S2**: $e(\mathbf{x}) = \sigma(a + y(\mathbf{x}))$; **Strategy S3**: $e(\mathbf{x}) = \sigma(a - y(\mathbf{x}))$; **Strategy S4**: $e(\mathbf{x}) = a \cdot y(\mathbf{x})$; **Strategy S5**: $e(\mathbf{x}) = \sigma(a + x_{i^*})$, where i^* is the index of the feature least correlated with y, according to mutual information (Cover and Thomas, 2006); **Strategy S6**: $e(\mathbf{x}) = \sigma(a + \mathbf{x}^T \mathbf{1}_p)$, where $\mathbf{1}_p = (1, 1, \ldots, 1)^\top$.

In the above labeling scenarios, the parameter a is selected to control the labeling frequency $c = p(s = 1|y = 1)$. We report the results for the labeling frequency $c = 0.5$. It is easy to check that strategies S1-S4 satisfy assumptions of Theorem 1. In particular in case of S3, $e(\mathbf{x})$ is *decreasing* function of $y(\mathbf{x})$ but $w(t)$ defined in Theorem 1 is strictly *increasing* and thus invertible. In contrast, strategies S5 and S6 do not satisfy the FOP assumption. Consequently, scenarios S5 and S6 provide a useful framework for evaluating the robustness of the proposed method to violations of the FOP assumption.

The individual methods are evaluated using three evaluation metrics: Balanced Accuracy (BACC), Expected Calibration Error (ECE), and Approximation Error (AE) for posterior probability, computed on the test set. Appendix A.3 provides the definitions and discussion of these metrics.

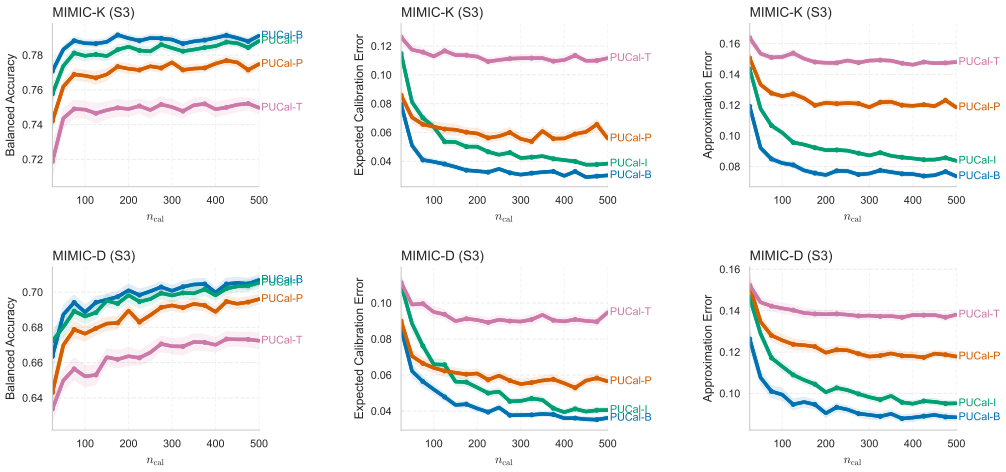

Figure 2: Evaluation metrics (BACC, ECE and AE) of the PUCal method as a function of the calibration set size n_{cal}. Results are reported for the MIMIC-K and MIMIC-D datasets and strategy S3. Results for other strategies are provided in the Appendix.

4.2 Discussion

In the main body of the paper, we present the results for the medical MIMIC datasets (MIMIC-K and MIMIC-D), which are representative of the overall findings. The results for the remaining datasets are provided in the Appendix.

Choice of Calibration Method. Among the four calibration methods considered (Beta, Platt, Temperature Scaling and Isotonic Regression), Beta Calibration achieves the best performance, followed closely by Isotonic Regression, whereas Platt Scaling and Temperature Scaling perform substantially worse (Figure 2 and 4-5 in the Appendix). This suggests that more flexible methods are better able to estimate the transformation function h, which is consistent with intuition. Furthermore, more pronounced differences between Beta Calibration and Isotonic Regression can be observed for the ECE and AE metrics, whereas for BACC, the differences are negligible.

Effect of Calibration Set Size. As expected, we observe an improvement in the performance metrics as the size of the calibration set increases (Figure 2 and 4-5 in the Appendix). This improvement is more pronounced for the AE and ECE metrics in the case of two best-performing methods. However, it is worth noting that the curves stabilize quickly, and even for relatively small calibration set sizes (on the order of 100 samples), we achieve values close to the optimum.

Performance of PUCal Compared with Classical Biased PU Methods. Our goal was not only to

determine whether PUCal achieves better values of the performance metrics, but also whether it does so more efficiently than methods based on propensity score estimation (SAREM, LBE and JERM). Figure 3 (as well as the figures 6, 7, 8 in the Appendix) illustrating the overall relationship between performance and computational efficiency, aggregated across all simulations and labeling strategies, shows that PUCal achieves performance comparable to or better than that of the best-performing biased PU method, while requiring substantially lower computational cost.

Performance for various labeling strategies and when the FOP assumption is violated. Detailed results for the individual strategies are presented in the plots 9-10, 11-12 and 13-14 in the Appendix. Although the results vary across datasets and labeling strategies, several consistent patterns can be observed. PUCal clearly outperforms NTC, generally achieves performance comparable to the best-performing biased PU method, and often performs only marginally worse than the ORACLE method. It is also worth noting that, for the SAREM, LBE, and JERM methods, their calibrated counterparts generally achieve better performance. For most datasets, PUCal exhibits high performance under the S5–S6 strategies, for which the FOP assumption does not hold. This demonstrates a certain degree of robustness of the proposed method to violations of this assumption.

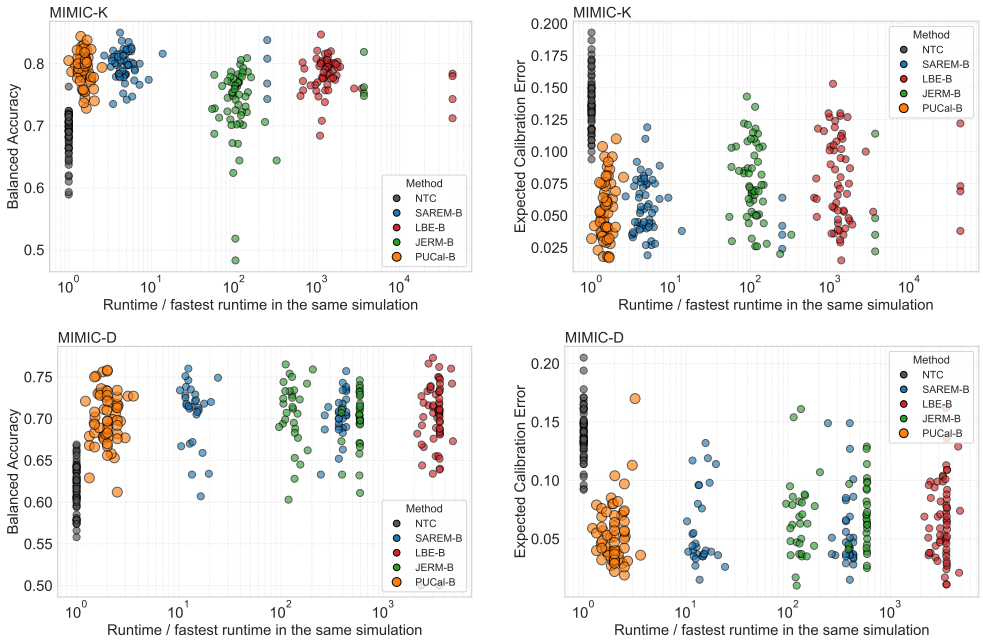

Figure 3: Comparison of evaluation metrics (BACC and ECE) versus computational efficiency for MIMIC datasets. Each dot corresponds to an individual simulation. The runtime is normalized with respect to the fastest method in the same simulation and displayed on a logarithmic scale.

5 Conclusions

The main finding of this work is that, under the FOP assumption, effective learning from biased PU data can be achieved by replacing the computationally expensive estimation of the propensity score with the calibration of the posterior probability estimated by the NTC classifier. Our experiments demonstrate that, under this assumption, the posterior probability of the true class can indeed be estimated effectively. Moreover, the proposed PUCal method shows robustness to deviations from the FOP assumption. The main cost of avoiding explicit propensity score estimation is the need for a calibration dataset, which, as discussed earlier, is typically more expensive to obtain than PU data. We have established error bounds on the expected absolute error of PUCal and shown that error induced by calibration is not larger that the one due to estimation of $s(\mathbf{x})$. This work also opens up several interesting directions for future research, including methods for verifying whether the FOP assumption holds in a given application. Such verification appears particularly challenging without resorting to the estimation of the propensity score itself.

AI USE STATEMENT

In this work, we did not use generative AI tools for any purpose, including text generation, literature review, or source code generation.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or the collection of new data. We use publicly available datasets for our experiments and comply with the applicable terms of use and licensing requirements. The proposed method does not introduce additional ethical risks beyond those associated with the standard use of the datasets and machine learning methods considered in this work.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our results, we provide detailed descriptions of the proposed method and experimental setup in the paper and supplementary material. We make all source code available, enabling the reproduction of all experiments and the generation of the figures presented in the paper. We also provide scripts for downloading and preprocessing the image data. In addition, we provide ready-to-use tabular representations of the image datasets, prepared for analysis. For the MIMIC dataset, we provide a script demonstrating how the dataset was prepared for analysis. Appendix A.2 provides a detailed description of the datasets, while Appendix A.3 provides a detailed description of the evaluation metrics used in the experiments. The proofs of all theorems are provided either in the main body of the paper or in the Supplementary Material.

REFERENCES

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. *Proceedings of Machine Learning Research*, 2021.

J. Bekker and J. Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109:719–760, 2020.

J. Bekker, P. Robberechts, and J. Davis. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data. In *Proceedings of the 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ECML'19, pages 71–85, 2019.

S. Bromuri, D. Zufferey, J. Hennebert, and M. Schumacher. Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. *Journal of Biomedical Informatics*, 51:165–175, 2014.

X. Cao, P. Zhang, W. Wang, W. Chaovalitwongse, J. Hu, and F. Wang. Adverse drug reaction prediction with symbolic latent dirichlet allocation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31 of *AAAI'17*, 2017.

H. Chen, F. Liu, Y. Wang, L. Zhao, and H. Wu. A variational approach for learning from positive and unlabeled data. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS'20, pages 14844–14854, 2020a.

X. Chen, W. Chen, T. Chen, Y. Yuan, C. Gong, K. Chen, and Z. Wang. Self-PU: Self boosted and calibrated positive-unlabeled training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020b.

O. Coudray, Ch. Keribin, P. Massart, and P. Pamphile. Risk bounds for positive-unlabeled learning under the selected at random ssumption. *Journal of Machine Learning Research*, pages 1–31, 2023.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

Y. Dai, Z. Hou, C. Li, Y. Xu, E. Wang, and X. Li. A closer look to positive-unlabeled learning from fine-grained perspectives: An empirical study. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

J. de Leeuw, K. Hornik, and P. Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, 2008.

Ch. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the seventeenth international joint conference on artificial intelligence*, volume 17, pages 973–978, 2001.

E. W. Frees, R. A. Derrig, and G. Meyers. *Predictive Modeling Applications in Actuarial Science*. Cambridge University Press, 2014.

F. Futami and M. Fujisawa. Information-theoretic generalisation analysis for expected calibration error. In *Proceedings of the 38th Conference on Neural information Processing Systems*, Neurips 2024, 2024.

W. Gerych, T. Hartvigsen, L. Buquicchio, E. Agu, and E. Rundensteiner. Recovering the propensity score from biased positive unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'22, pages 6694–6702, 2022.

C. Gong, Q. Wang, T. Liu, B. Han, J. You, J. Yang, and D. Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Trans Pattern Anal Mach Intell*, pages 1–16, 2021.

C. Gong, M. I. Zulfiqar, C. Zhang, S. Mahmood, and J. Yang. A recent survey on instance-dependent positive and unlabeled learning. *Fundamental Research*, 2022.

C. Guo, Y. Sun, and Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on machine learning*, PMLR, 2017.

F. He, T. Liu, J. Webb, and D. Tao. Instance-dependent PU learning by Bayesian optimal relabeling. manuscript, 2018. URL `arXiv:180802180`.

A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, Anthony C. L., and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:1–9, 2016.

M. Kato, T. Teshima, and J. Honda. Learning from positive and unlabeled data with a selection bias. In *Proceedings of the 7th International Conference on Learning Representations*, pages 1–12, 2019.

R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS'17, pages 1674–1684, 2017.

M. Kull, T. S. Filho, and P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *AISTATS'17*, pages 623–631, 2017.

C. Y. Lee and Y. P. Chen. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Briefings in Bioinformatics*, 22(2):1884–1901, 2021.

F. Li, S. Dong, A. Leier, M. Han, X. Guo, J. Xu, X. Wang, S. Pan, C. Jia, Y. Zhang, G. Webb, L. J. M. Coin, C. Li, and J. Song. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics*, 23(1), 2021.

H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao. Spotting fake reviews via collective positive-unlabeled learning. In *2014 IEEE International Conference on Data Mining*, ICDM'14, pages 899–904, 2014.

J. Mielniczuk and A. Wawrzeńczyk. Single-sample versus case-control sampling scheme for Positive Unlabeled data: the story of two scenarios. *Fundamenta Informaticae*, 191:1–17, 2024.

M. Płatek and J. Mielniczuk. Enhancing naive classifier for positive unlabeled data based on logistic regression approach. In *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, page 225–233, 2023.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularised likelihood methos. *Advances in Large Margin Claasifiers*, pages 61–74, 1999.

W. Rejchel, P. Teisseyre, and J. Mielniczuk. Joint empirical risk minimization for instance-dependent positive-unlabeled data. *Knowledge-Based Systems*, 304:1–21, 2024.

V. Sevetlidis, G. Pavlidis, S. Mouroutsos, and A. Gasteratos. Dense-PU: Learning a Density-Based Boundary for Positive and Unlabeled Learning. *IEEE Access*, 12:1–12, 2024.

Z. Sun, D. Song, and A. Hero. Minimum-risk recalibration of classifiers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS' 23, 2023.

X. Tang, C. Xu, H. Tao, X. Ma, and C. Hou. Confidence-based PU learning with instance-dependent label noise. *IEEE Transactions on Neural Networks and Learning Systems*, 36:14283–14297, 03 2025.

P. Teisseyre, T. Martens, J. Bekker, and J. Davis. Learning from biased positive-unlabeled data via threshold calibration. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, AISTATS'25, pages 2314–2322, 2025.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

M. Wainwright. *High-dimensional Statistics*. Cambridge University Press, 2019.

W. Wang, D.-D. Wu, M. Li, J. Zhang, G. Niu, and M. Sugiyama. Accessible, realistic, and fair evaluation of positive-unlabeled learning algorithms. In *Proceedings of the International Conference on Learning Representations*, ICRL'26, 2026.

D. Yuan, Y. Miao, N. Z. Gong, Z. Yang, Q. Li, D. Song, Q. Wang, and X. Liang. Detecting fake accounts in online social networks at the time of registrations. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, CCS'19, pages 1423–1438, 2019.

Y. Zhao, Q. Xu, Y. Jiang, P. Wen, and Q. Huang. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, CVPR'22, pages 14461–14470, 2022.

D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri. Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Computers in Biology and Medicine*, 65:34–43, 2015.

A SUPPLEMENTARY MATERIAL

A.1 PROOF OF THEOREM 3

Proof. We have

$$E_{\mathbf{x}}|y(\mathbf{x}) - \widehat{y}(\mathbf{x})| = E_{\mathbf{x}}|h(s(\mathbf{x})) - \widehat{h}(\widehat{s}(\mathbf{x}))| \leq$$
$$E_{\mathbf{x}}|h(s(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))| + E_{\mathbf{x}}|h(\widehat{s}(\mathbf{x})) - \widehat{h}(\widehat{s}(\mathbf{x}))| := I + II$$

Define $D_{KL}\big(\mathbf{s}(\mathbf{x})\big|\big|\widehat{\mathbf{s}}(\mathbf{x})\big)$ to be Kullback-Leibler (KL) divergence (Cover and Thomas, 2006) between binary measures $\mathbf{s}(\mathbf{x}) = (s(\mathbf{x}), 1 - s(\mathbf{x}))$ and $\widehat{\mathbf{s}}(\mathbf{x}) = (\widehat{s}(\mathbf{x}), 1 - \widehat{s}(\mathbf{x}))$. Observe that the term I can be bounded by

$$E_{\mathbf{x}}||h'||_\infty|\widehat{s}(\mathbf{x}) - s(\mathbf{x})| \leq ||h'||_\infty|\sqrt{\frac{1}{2}[\mathcal{R}_{ntc}(\widehat{s}) - \mathcal{R}_{ntc}(s)]} = ||h'||_\infty\sqrt{\frac{1}{2}r_n^{\mathcal{F}}(\widehat{s})},$$

due to Pinsker inequality (cf (Wainwright, 2019), inequality (3.59)):

$$2(\widehat{s}(\mathbf{x}) - s(\mathbf{x}))^2 = 2||\widehat{\mathbf{s}}(\mathbf{x}) - \mathbf{s}(\mathbf{x}))||_{TV}^2 \leq D_{KL}(\mathbf{s}(\mathbf{x})||\widehat{\mathbf{s}}(\mathbf{x}))$$

Indeed we have

$$E_{\mathbf{x}}|\widehat{s}(\mathbf{x}) - s(\mathbf{x})| \leq \sqrt{E_{\mathbf{x}}(\widehat{s}(\mathbf{x}) - s(\mathbf{x}))^2} \leq \sqrt{\frac{1}{2}E_{\mathbf{x}}[D_{KL}(\mathbf{s}(\mathbf{x})||\widehat{\mathbf{s}}(\mathbf{x}))]} = \sqrt{\frac{1}{2}(\mathcal{R}_{ntc}(\widehat{s}) - \mathcal{R}_{ntc}(s))},$$

where the last equality follows from definition of KL divergence. Since $s \in \mathcal{F}$, we have that minimiser of $\mathcal{R}_{ntc}$ in $\mathcal{F}$ is s, as loglikelihood defines proper loss, and the bound above is $\sqrt{r_n^{\mathcal{F}}(\widehat{s})/2}$. Note that as $\widetilde{h} \in \mathcal{H}$ we have that $\widetilde{h} = \min_{h_\theta \in \mathcal{H}} \mathcal{R}_{cal}(h_\theta)$. We can write

$$II \leq E_{\mathbf{x}}|\widetilde{h}(\widehat{s}(\mathbf{x})) - \widehat{h}(\widehat{s}(\mathbf{x}))| + E_{\mathbf{x}}|\widetilde{h}(\widehat{s}(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))|$$

The first term is bounded as before by

$$E_{\mathbf{x}}|\widetilde{h}(\widehat{s}(\mathbf{x})) - \widehat{h}(\widehat{s}(\mathbf{x}))| \leq \sqrt{\frac{1}{2}[\mathcal{R}_{cal}(\widehat{h}) - \mathcal{R}_{cal}(\widetilde{h})]}\Big) = \sqrt{r_{n_{cal}}^{\mathcal{H}}(\widehat{h})/2}. \tag{5}$$

For the second term we apply Lemma 1

$$E_{\mathbf{x}}|\widetilde{h}(\widehat{s}(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))| \leq E_{\mathbf{x}}|h(s(\mathbf{x})) - h(\widehat{s}(\mathbf{x}))| = I, \tag{6}$$

which ends the proof. □

PROOF OF THEOREM 4

Denote by $\omega = (\omega_1, \omega_2)$ elementary event of the sample space $\Omega = \Omega_1 \times \Omega_2$, where ω_1 corresponds to $\mathcal{D}_{PU}$ and ω_2 to $\mathcal{D}_{cal}$. In view of Theorem 4.10 in (Wainwright, 2019), together with derivation on p. 104 there, we have that

$$r_n^{\mathcal{F}}(\widehat{s}) \leq 2\,\mathrm{Rad}_n(\mathcal{F}) + \delta$$

with probability at least $1 - \exp(-n\delta^2/2)$. This implies, using Borel-Cantelli lemma, that

$$r_n^{\mathcal{F}}(\widehat{s}) = \mathcal{O}(\mathrm{Rad}_n(\mathcal{F}))$$

$p(\mathbf{x}, s)$-almost surely. Denote by $A \subset \Omega_1$ the corresponding set of probability 1. For fixed $\omega_1 \in A$ we have, using the above argument again, that

$$r_{n_{cal}}^{\mathcal{F}}(\widehat{h}) = \mathcal{O}(\mathrm{Rad}_{n_{cal}}(\mathcal{H}))$$

$p(\widehat{s}(\mathbf{x}), y)$-almost surely. Denote by B_{ω_1} the corresponding set of probability 1 and let $W = \{(\omega_1, \omega_2) : \omega_1 \in A, \omega_2 \in B_{\omega_1}\}$. Then Fubini theorem immediately implies that $P(W) = 1$ and on W the conclusion of Theorem holds in view of Theorem 3.

We also note that if $\mathcal{H}$ is of polynomial discrimination, in particular it has finite Vapnik-Chervonenkis dimension, its Rademacher complexity does not depend on underlying distribution $p(\widehat{s}(\mathbf{x}), y)$ and satisfies $\mathrm{Rad}_{n_{cal}}(\mathcal{H})) \leq C(\log n/n)^{1/2}$ for some absolute constant $C > 0$. This easily follows from lemma 4.41 in (Wainwright, 2019).

A.2 DATASETS AND BASE CLASSIFIERS

We consider the following widely used benchmark datasets: MNIST, CIFAR10, USPS and Fashion. Our binary class definitions were based on experimental setups commonly adopted in previous studies on PU learning (Kiryo et al., 2017; Chen et al., 2020b; Gong et al., 2021). For MNIST, the dataset was divided according to digit parity, with even digits assigned to the positive class and odd digits to the negative class (Kiryo et al., 2017; Chen et al., 2020b; Gong et al., 2021). In CIFAR10, images of vehicles were treated as positive examples, whereas images of animals constituted the negative class (Kiryo et al., 2017; Chen et al., 2020b). For USPS, digits smaller than five were designated as positive, while all remaining digits were assigned to the negative class. In Fashion, upper-body clothing categories were considered positive examples, with all other clothing categories grouped into the negative class (Sevetlidis et al., 2024). For the image datasets, we employ a pre-trained ResNet18 model to obtain feature representations. Specifically, the 512-dimensional feature vector corresponding to each image is extracted from the average pooling layer of the network. These representations are subsequently provided as input to the PU learning models.

Our experiments also include the MIMIC clinical database (Johnson et al., 2016), which comprises medical information for $33{,}166$ patients admitted to various intensive care units. Patient's diagnoses are represented according to the ICD9 coding system. In our analysis, we consider only adult patients aged over 16 years. We use two disease groups as class variables (Bromuri et al., 2014): KIDNEY (Kidney disease; ICD-9 codes 580–589) and DIABETES (Diabetes mellitus; ICD-9 codes 249–250). Throughout the experiments, we refer to the setting with KIDNEY as the target class as MIMIC-K, and to the setting with DIABETES as the target class as MIMIC-D. The dataset provides 308 numerical features describing different aspects of the patients' clinical and demographic characteristics. These variables include laboratory measurements, medical scores, and administrative information, such as age, gender, and marital status. Before conducting the experiments, the data were pre-processed by normalizing the variables and removing missing values. We followed the preprocessing procedure adopted in previous studies using the MIMIC dataset; further details are available in Bromuri et al. (2014); Zufferey et al. (2015).

For both the tabular MIMIC data and the vector representations of the image data, we employ logistic regression with L2 regularization as the base classifier for all considered methods (NTC (Elkan and Noto, 2008), SAREM (Bekker et al., 2019), LBE (Gong et al., 2021), and JERM (Rejchel et al., 2024)). This choice is consistent with the original formulations of SAREM, LBE, and JERM, in which logistic regression was used as the default base classifier. Moreover, in the case of JERM, logistic regression is an intrinsic component of the method, and replacing it with a different classifier is not possible within its original formulation.

A.3 EVALUATION MEASURES

To evaluate the performance of the methods, we compute three metrics: Balanced Accuracy (BACC), Expected Calibration Error (ECE), and Approximation Error (AE). BACC is a standard measure of the predictive performance of classification models, particularly useful for imbalanced datasets. In addition to the accuracy of the constructed classifiers, one is often concerned with the confidence of their predictions, i.e., how accurately the corresponding posterior estimators reflect the probabilities of the predicted events. Classifiers with accurate predictions are considered well-calibrated, with the gold standard being perfect calibration, which corresponds to

$$p(y = 1|\widehat{y}(\mathbf{x}) = q) = q,$$

for any $q \in (0, 1)$. Since perfect calibration holds for the posterior probability $\widehat{y}(\mathbf{x}) = y(\mathbf{x})$ (Sun et al., 2023), it is intuitive that deviations from perfect calibration may be due to estimation errors, but also to overfitting and, typically, to the high dimensionality of the problem (Bai et al., 2021).

The usual calibration metric is the estimated value of the True Calibration Error,

$$TCE = E_{\mathbf{x}}|p(y = 1|\widehat{y}(\mathbf{x})) - \widehat{y}(\mathbf{x})|,$$

based on binning, called the Expected Calibration Error (ECE) (Futami and Fujisawa, 2024).

Finally, to investigate the accuracy of posterior probability estimation, it is also important to analyze the estimation error. However, for real-world datasets, the true posterior probability $y(\mathbf{x}) = p(y =$

$1|\mathbf{x})$ is unknown. To overcome this limitation, we define the approximation error as

$$AE = n_{test}^{-1} \sum_{i=1}^{n_{test}} |\tilde{y}(\mathbf{x}_i) - \widehat{y}(\mathbf{x}_i)|,$$

where $\tilde{y}(\mathbf{x}_i)$ denotes the estimator obtained using the ORACLE method (assuming access to the true labels), whereas $\widehat{y}(\mathbf{x}_i)$ denotes the estimator obtained using one of the considered methods. Therefore, the approximation error serves as a practical approximation of the estimation error that can be computed for real-world datasets. This metric is important due to the theoretical results concerning the estimation of posterior probabilities.

A.4 An example illustrating the FOP assumption

A simple parametric example illustrating cases, where the FOP assumption holds and where it does not, is given by a double logistic model, frequently used in the PU learning literature (Gong et al., 2021; Rejchel et al., 2024). Suppose that the posterior probability and the propensity score are modeled as

$$y(\mathbf{x}) = \sigma(\beta^\top \mathbf{x}), \qquad e(\mathbf{x}) = \sigma(\gamma^\top \mathbf{x}),$$

where $\sigma(\cdot)$ denotes the logistic function. The FOP assumption is satisfied only if the vectors β and γ are linearly dependent, i.e., $\gamma = a\beta$ for some constant a. In such a case,

$$e(\mathbf{x}) = \sigma(a\beta^\top \mathbf{x}) = \sigma(a \cdot \mathrm{logit}(y(\mathbf{x}))),$$

and therefore the propensity score depends only on the posterior probability. However, in general, the parameters β and γ may not be proportional and then FOP condition is not satisfied. Then, two observations may have the same posterior probability but different propensity scores.

A.5 Additional Experimental Results

The appendix below contains the full results of the experiments.

- Figures 4 and 5 show how the considered evaluation metrics (BACC, ECE and AE) depend on the size of the calibration set for the MIMIC-K and MIMIC-D datasets. We consider four variants of the PUCal method: PUCal-B, PUCal-I, PUCal-P and PUCal-T described in Section 4.
- Figures 6, 7 and 8 show a comparison of the predictive performance (BACC, ECE and AE) and computational efficiency of the considered methods. The following methods are considered: NTC, SAREM-B, JERM-B, LBE-B and PUCal-B.
- Figures 9-10 illustrate the distributions of BACC for the following methods: NTC, ORACLE, PUCal-B, SAREM, SAREM-B, JERM, JERM-B, LBE, LBE-B.
- Figures 11-12 illustrate the distributions of ECE for the following methods: NTC, ORACLE, PUCal-B, SAREM, SAREM-B, JERM, JERM-B, LBE, LBE-B.
- Figures 13-14 illustrate the distributions of AE for the following methods: NTC, ORACLE, PUCal-B, SAREM, SAREM-B, JERM, JERM-B, LBE, LBE-B.

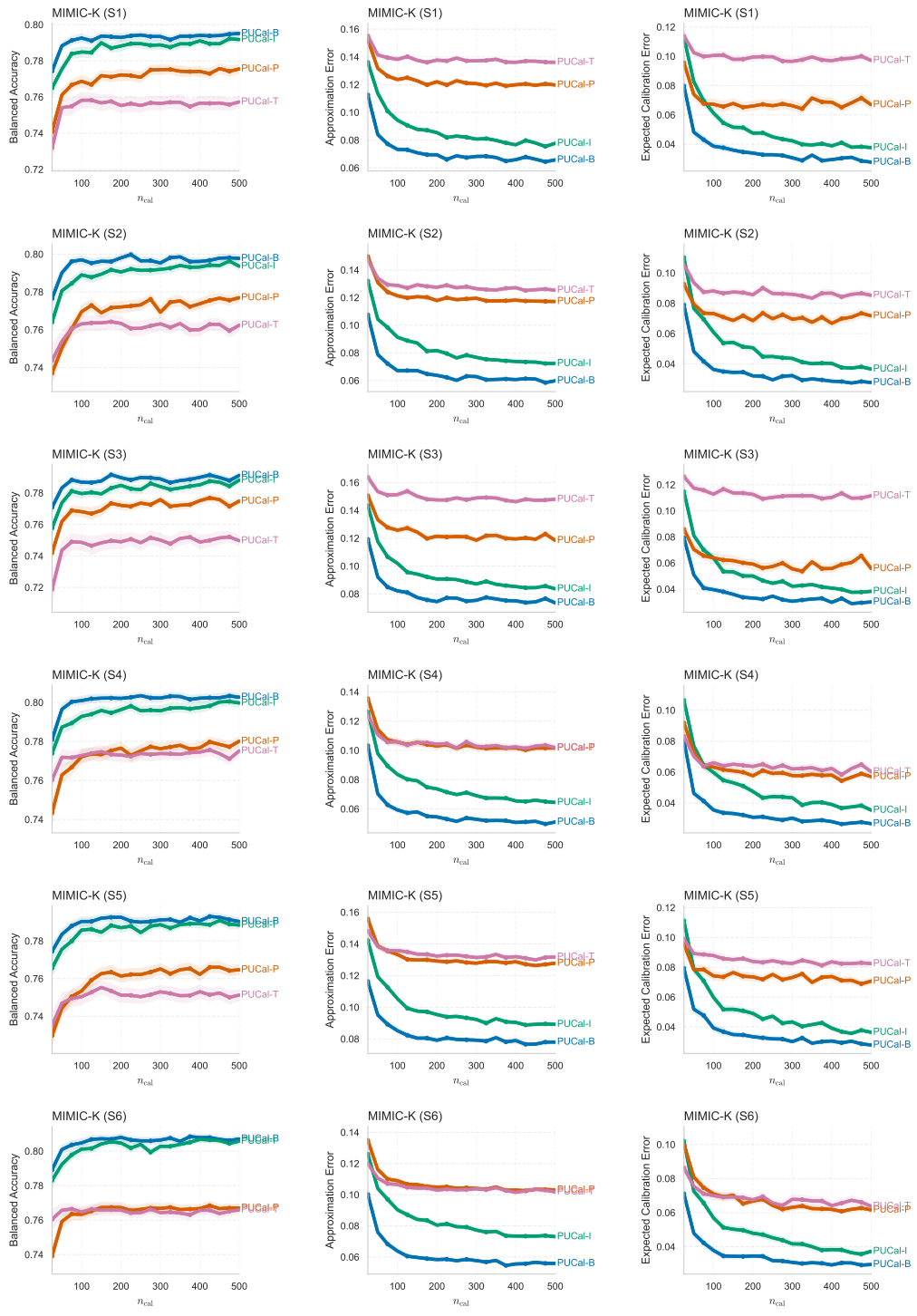

Figure 4: Evaluation metrics (BACC, ECE and AE) of the PUCal method as a function of the calibration set size. Results are reported for the MIMIC-K dataset.

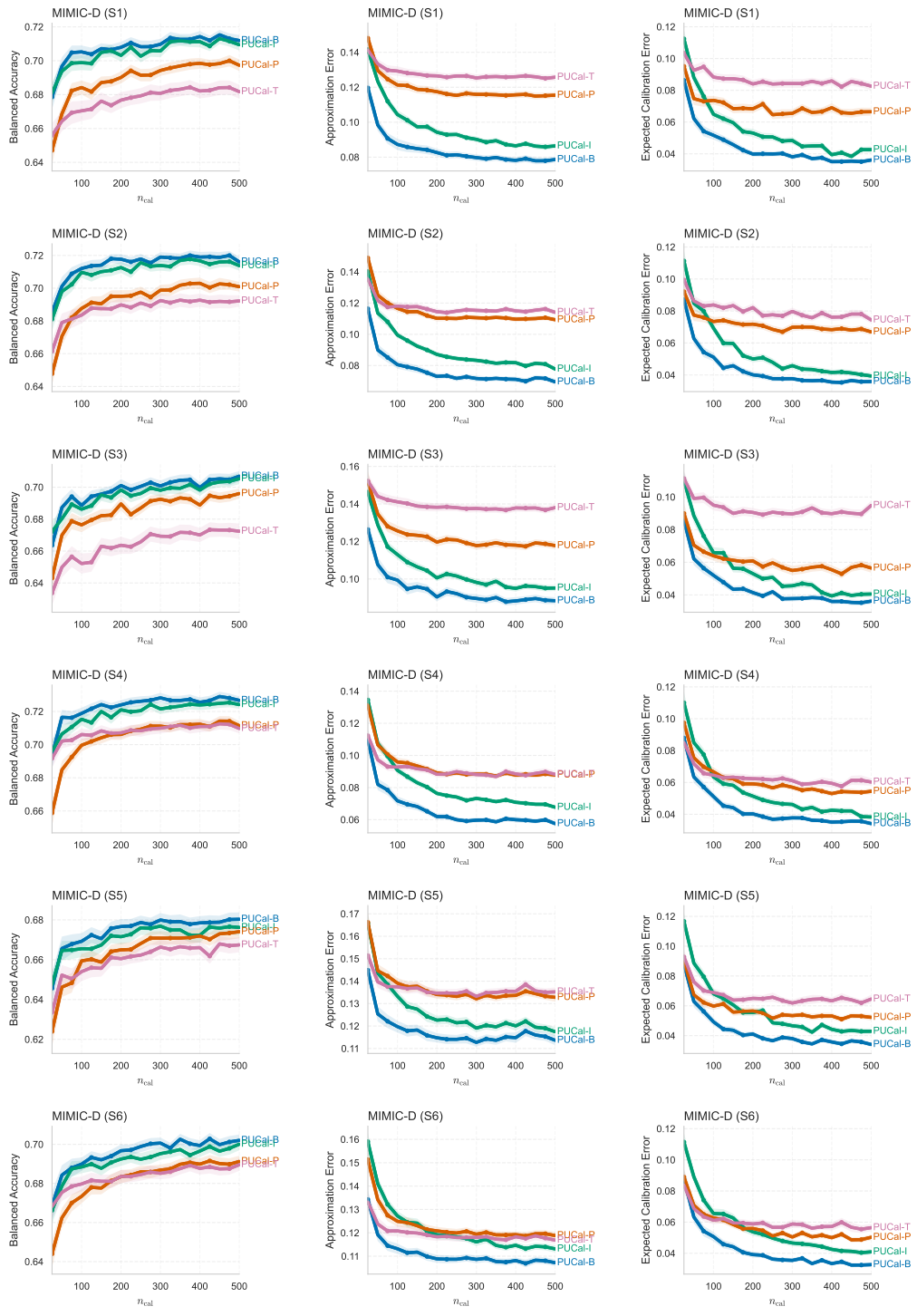

Figure 5: Evaluation metrics (BACC, ECE and AE) of the PUCal method as a function of the calibration set size. Results are reported for the MIMIC-D dataset.

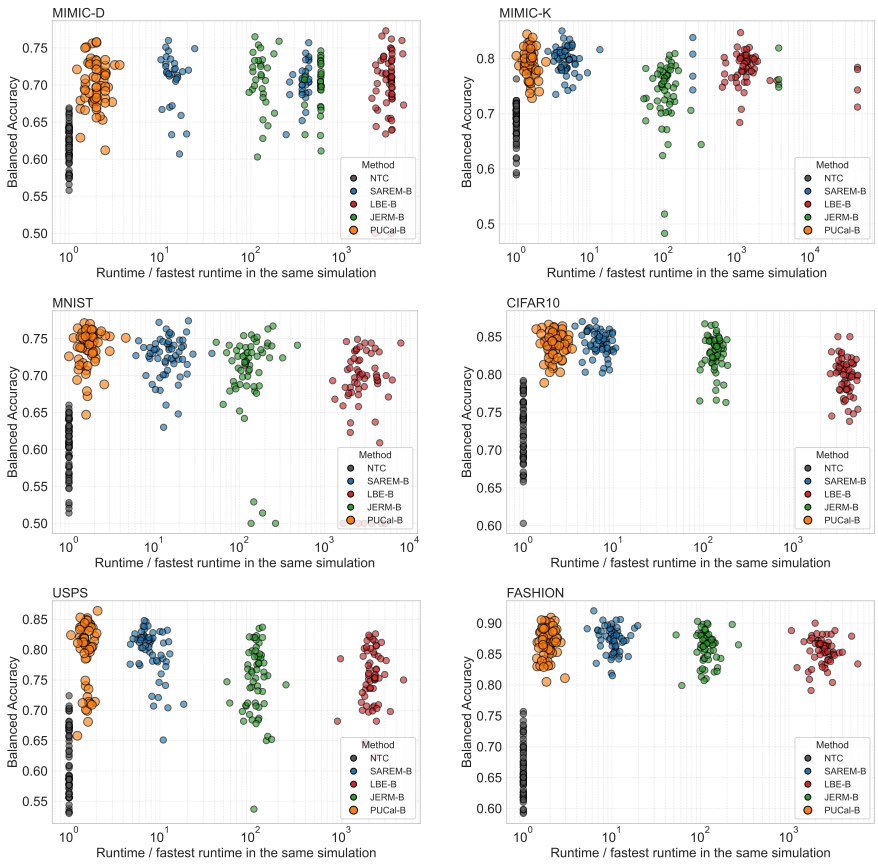

Figure 6: Comparison of predictive performance (Balanced Accuracy) and computational efficiency of the considered methods. Each dot corresponds to an individual simulation. The runtime is normalized with respect to the fastest method in the same simulation and displayed on a logarithmic scale.

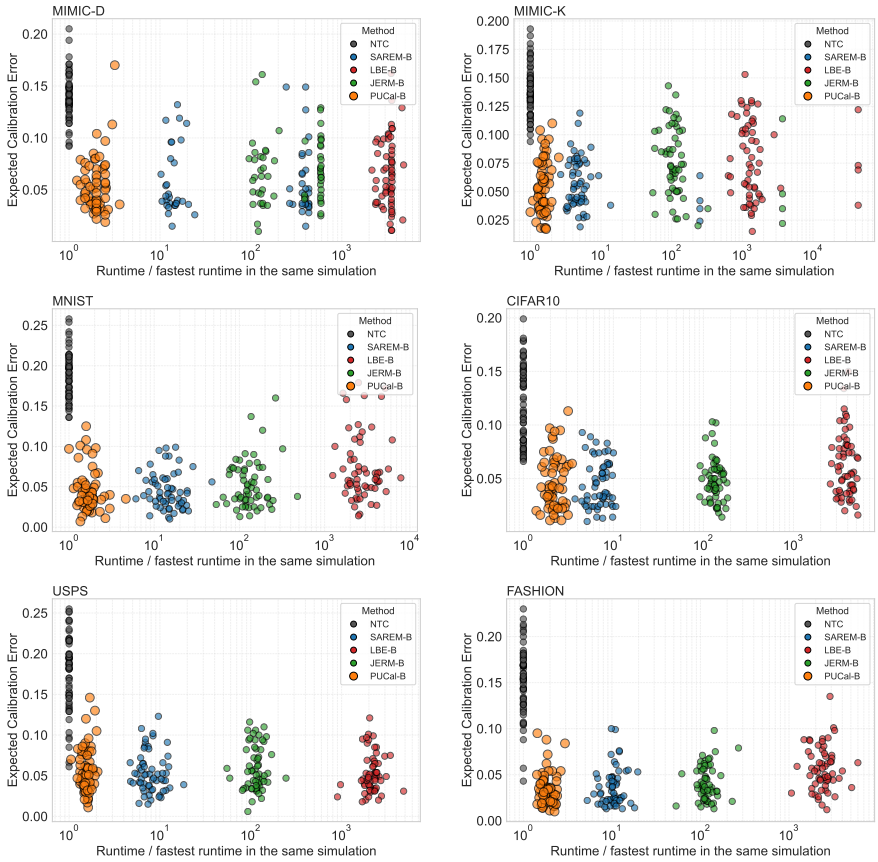

Figure 7: Comparison of Expected Calibration Errors (ECE) and computational efficiency of the considered methods. Each dot corresponds to an individual simulation. The runtime is normalized with respect to the fastest method in the same simulation and displayed on a logarithmic scale.

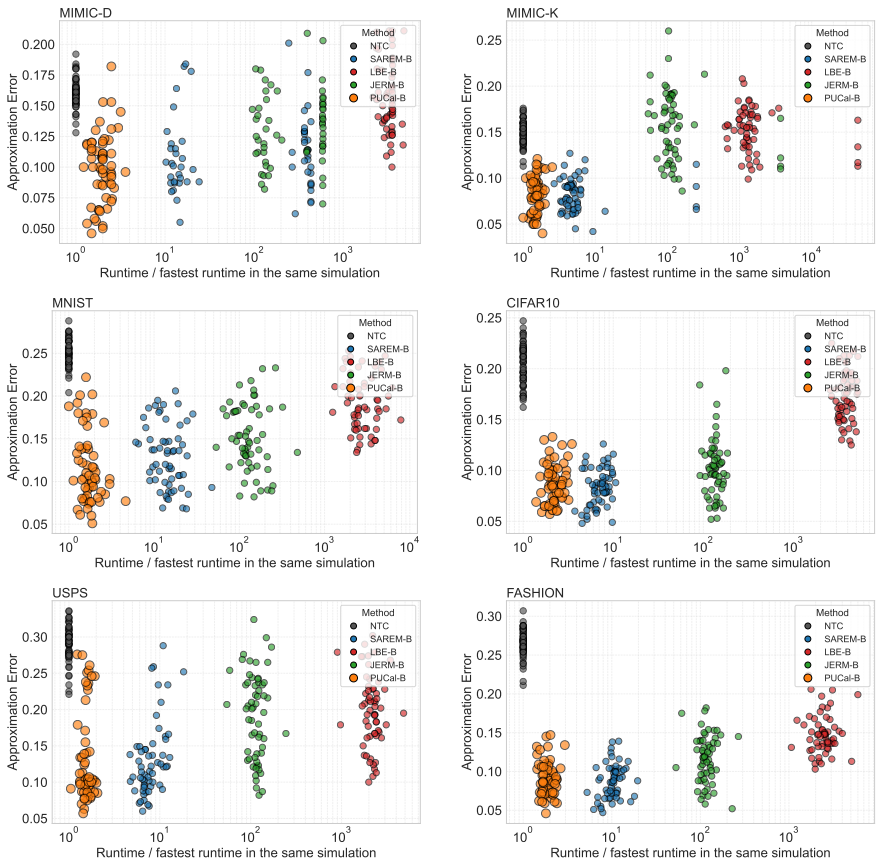

Figure 8: Comparison of Approximation Errors (AE) and computational efficiency of the considered methods. Each dot corresponds to an individual simulation. The runtime is normalized with respect to the fastest method in the same simulation and displayed on a logarithmic scale.

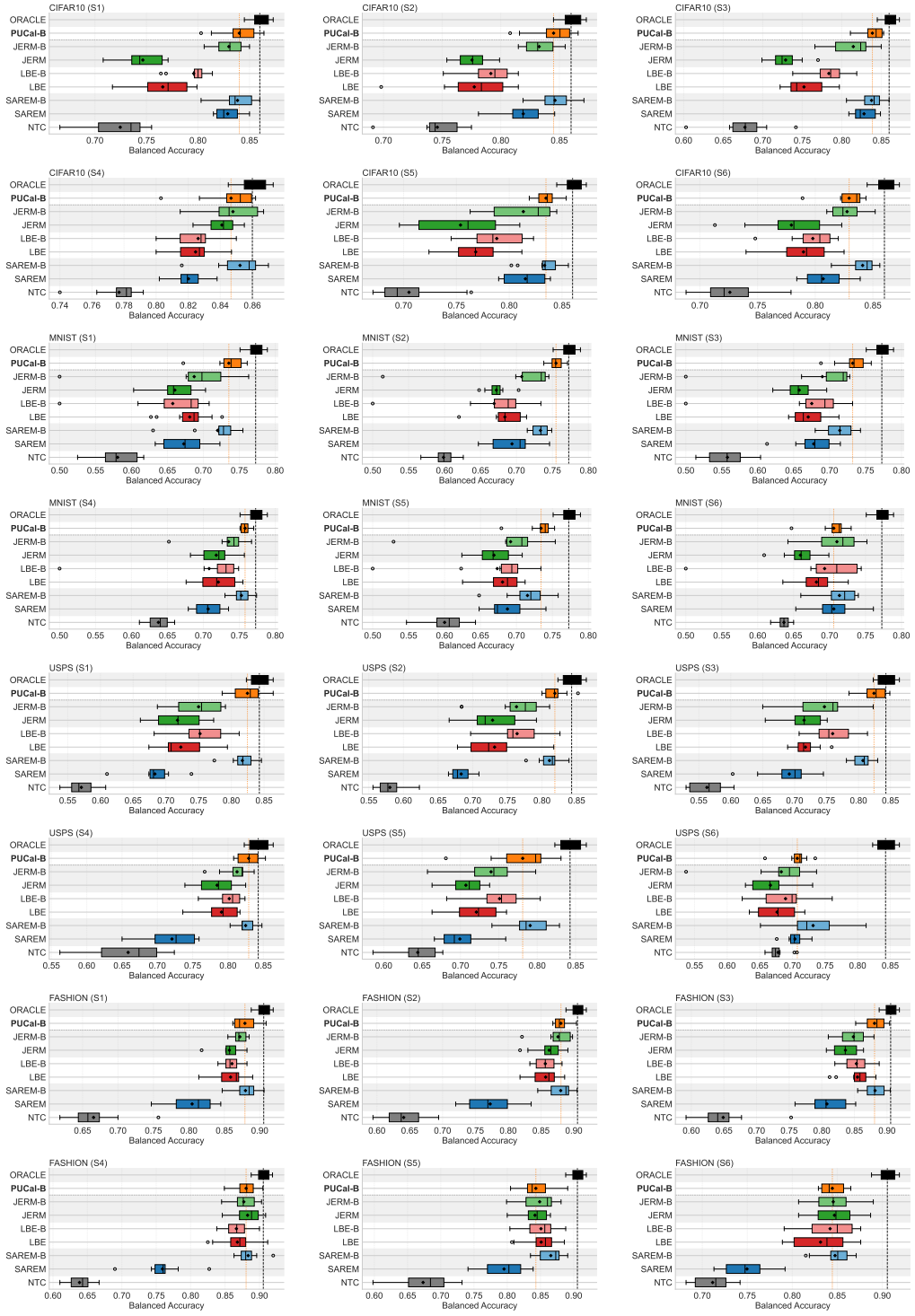

Figure 9: Balanced accuracy for the CIFAR10, MNIST, USPS and Fashion datasets using labeling strategies S1–S6.

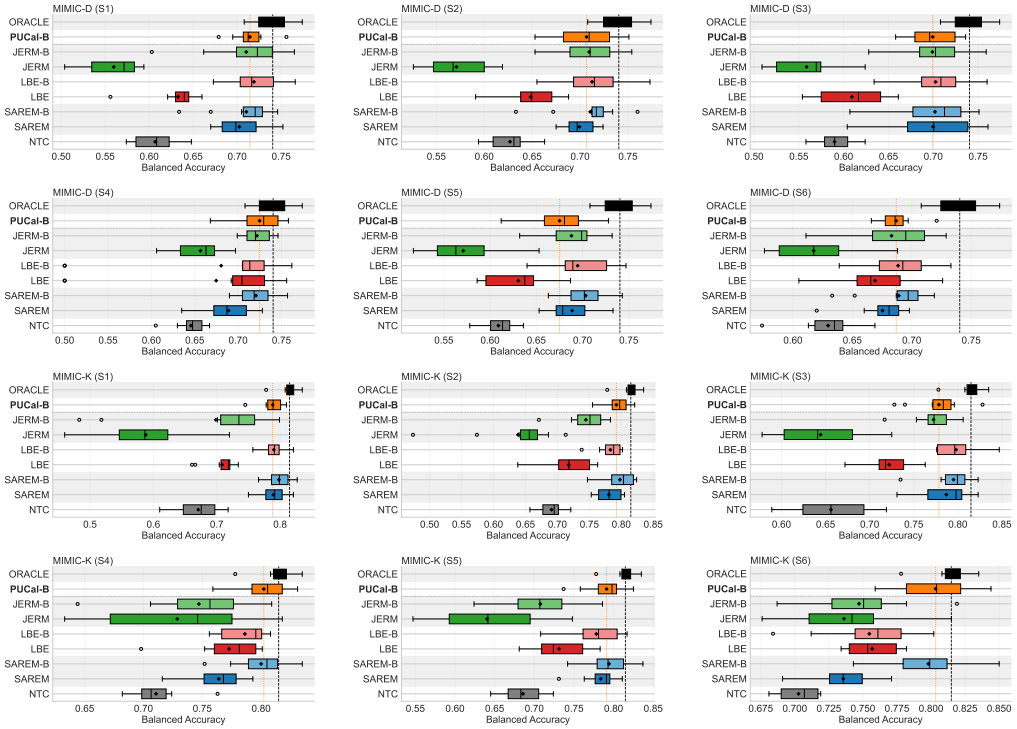

Figure 10: Balanced accuracy for the MIMIC-D and MIMIC-K datasets using labeling strategies S1–S6.

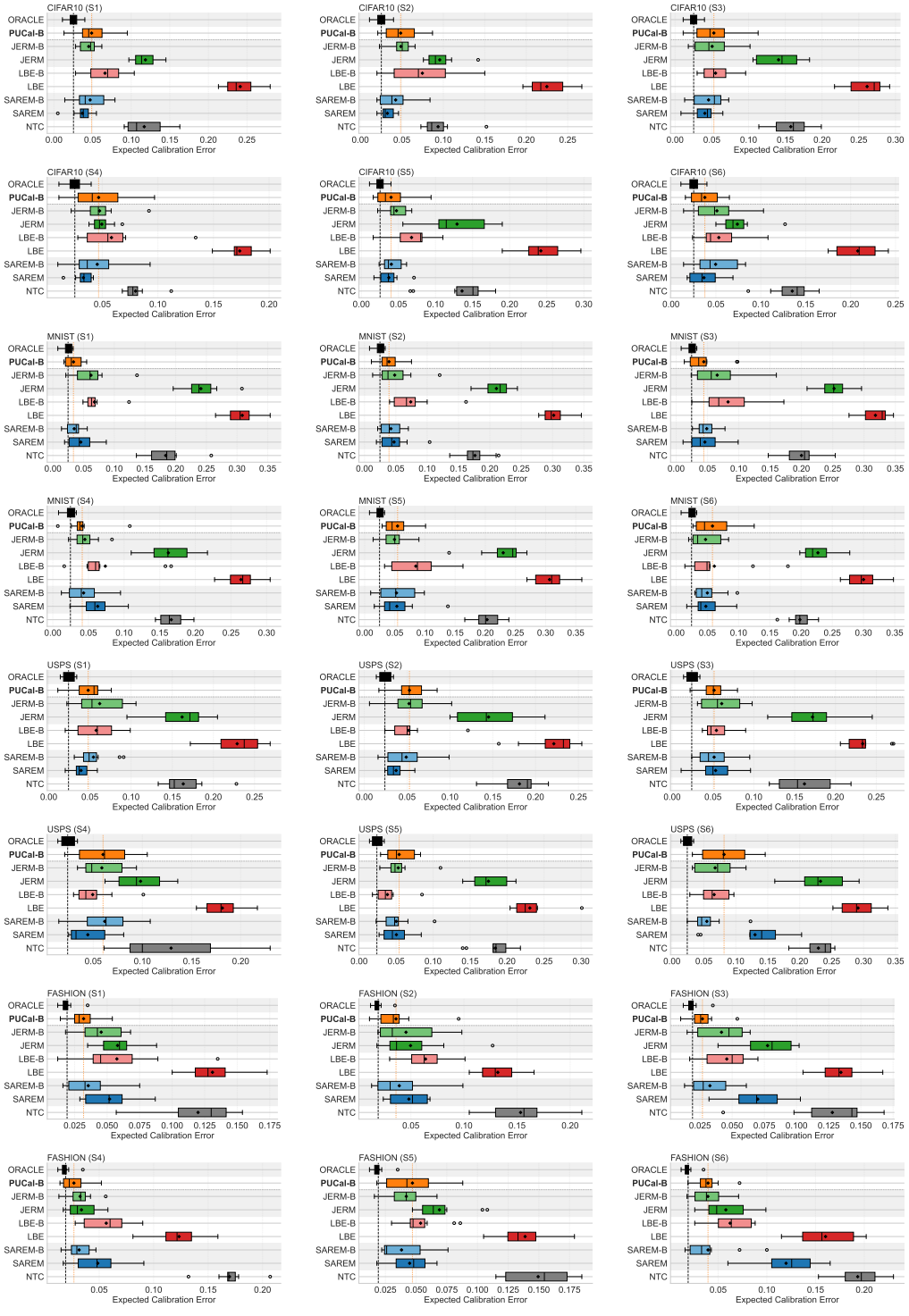

Figure 11: Expected Calibration Errors for the CIFAR10, MNIST, USPS and Fashion datasets using labeling strategies S1–S6.

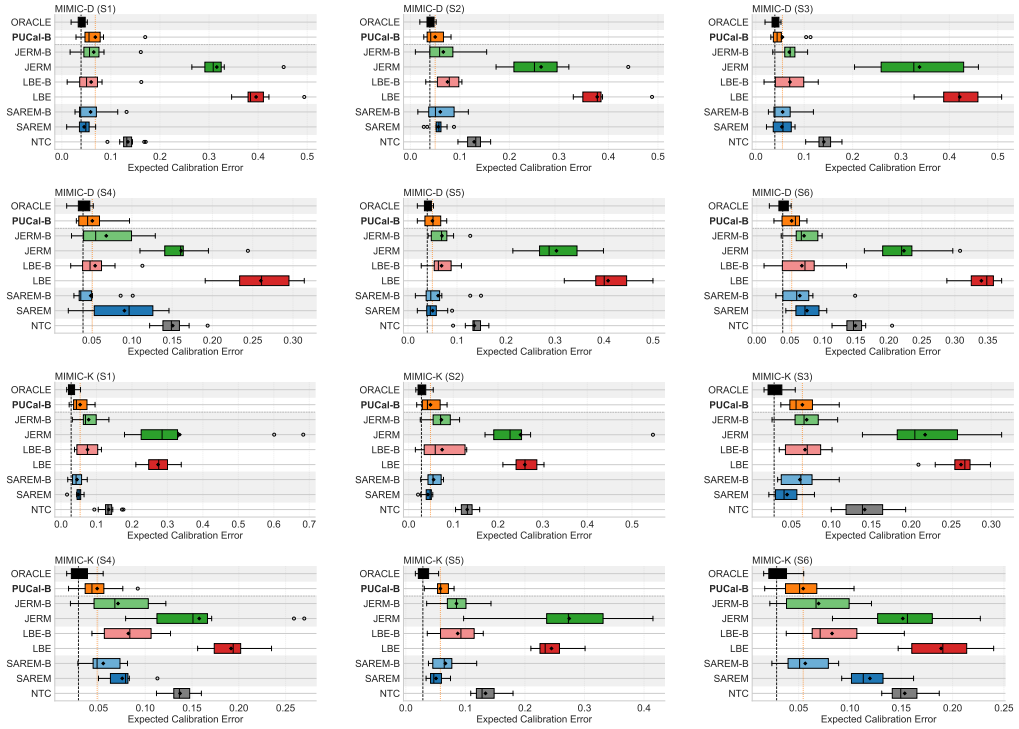

Figure 12: Expected Calibration Errors for the MIMIC-D and MIMIC-K datasets using labeling strategies S1–S6.

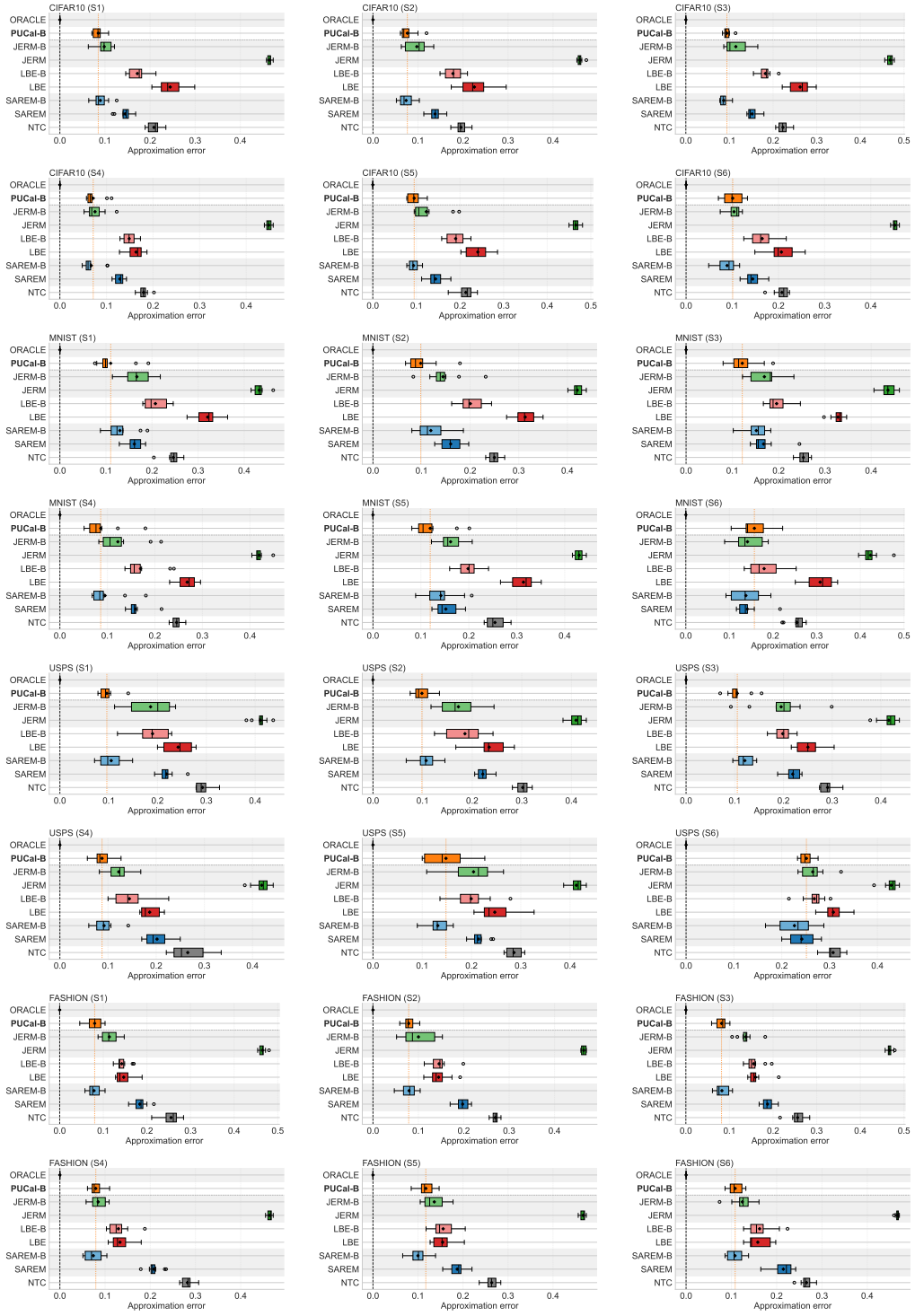

Figure 13: Approximation errors for the CIFAR10, MNIST, USPS and Fashion datasets using labeling strategies S1–S6.

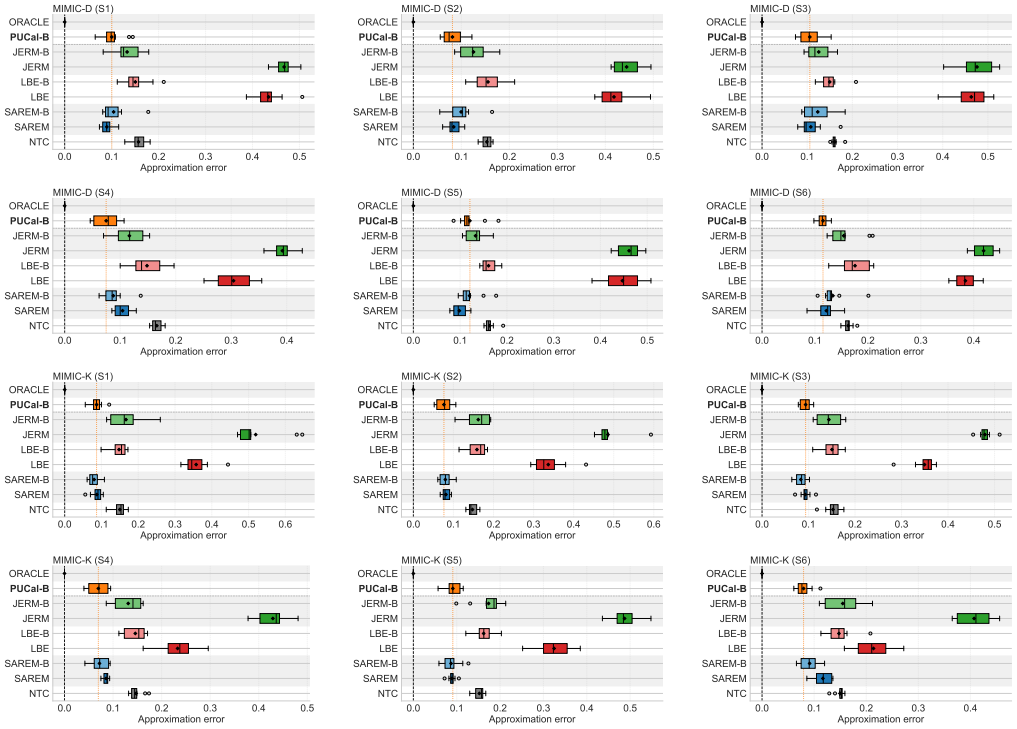

Figure 14: Approximation errors for the MIMIC-D and MIMIC-K datasets using labeling strategies S1–S6.