MULTI-MODEL EVALUATION WITH LABELED AND UNLABELED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

It remains difficult to select a machine learning model from a set of candidates in the absence of a large, labeled dataset. To address this challenge, we propose a framework to compare multiple models that leverages three aspects of modern machine learning settings: multiple machine learning classifiers, continuous predictions on all examples, and abundant unlabeled data. The key idea is to estimate the joint distribution of classifier predictions using a mixture model, where each component corresponds to a different class. We present preliminary experiments on a large health dataset and conclude with future directions.

1 INTRODUCTION

The machine learning community has developed an immense number of trained models, spanning a wide variety of modalities and tasks (You et al., 2021; Arango et al., 2023). Faced with these options, practitioners must determine which model is most appropriate for their needs. Historically, comparing machine learning classifiers requires access to labeled data, but high-quality labeled data is often prohibitively expensive or impossible to obtain. The challenge of deciding between several options is not new: doctors often face the choice between similar drugs to prescribe or treatments to recommend. Unlike prescription medicine, however, machine learning models do not undergo rigorous and well-documented clinical trials to assess their strengths and weaknesses. Choosing between this plethora of models is thus not straightforward.

Addressing this challenge, we propose a framework for model comparison that makes use of *unlabeled* data. The framework captures three key properties of machine learning ecosystems: multiple machine learning models, continuous predictions, and abundant unlabeled data. While past work has addressed some of these properties (Welinder et al., 2013; Ji et al., 2020; Chouldechova et al., 2022; Bommasani et al., 2022), no existing framework accommodates all three.

Concretely, we introduce a general mixture model framework to estimate the class-conditional joint distribution of multiple models' predictions. The resulting mixture model is well-suited to estimate a broad class of performance metrics, both for individual models (e.g. calibration error or accuracy) and for the set of candidate models as a whole (e.g. relative accuracy and systemic failures (Kleinberg & Raghavan, 2021)). We instantiate this mixture model with a semi-supervised kernel density estimator (KDE).

We ground our experiments in two domains: (1) healthcare and (2) content moderation. Both of these domains naturally have more plentiful unlabeled data relative to labeled data. Our results are encouraging and show that incorporating unlabeled data into model evaluations can improve the quality of those evaluations.

2 METHODS

Problem Setting We consider a setting in which a model consumer (for example, a hospital system) must choose between several classification models. Formally, there are M classification models $[f_1, f_2, \ldots, f_M]$ designed for the same prediction task, so $f_i : X \rightarrow [0, 1]$, where X is the domain of the input (X could, for example, correspond to EHR data, and in principle could be any modality). Models in the set may differ by function class, training data, or training hyperparameters. We assume the practitioner has access to the *continuous* predictions, or scores, corresponding to the

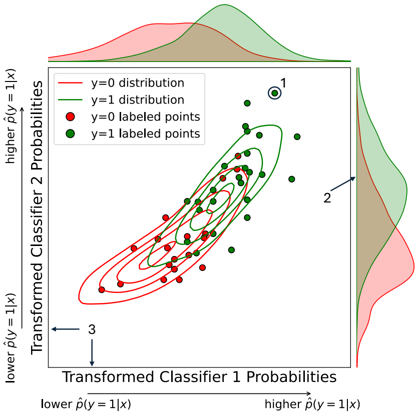

Figure 1: Joint distribution of model predictions. There are three key features we leverage: (1) labeled and unlabeled data, (2) continuous predictions, and (3) predictions from multiple models.

class probabilities, so for any given x, let $\mathbf{s} = [s_1, \dots, s_M] = [f_1(x), \dots, f_M(x)] \in [0,1]^M$ be the concatenated predictions of all models.

During evaluation, we have access to two datasets: (1) an small labeled dataset, $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{\ell}$ and (2) a larger unlabeled dataset $\mathcal{D}_U = \{(x_i)\}_{i=1}^{u}$. Both the labeled and unlabeled examples are drawn i.i.d. from the same distribution. The goal of the practitioner is to use these datasets to estimate and compare model performance according to metrics she may care about, such as expected calibration error (ECE) or AUROC. To this end, if she had access to the true y for each point, this task would be simple. Despite not receiving ground truth access in $\mathcal{D}_U$, she has access to several *proxies* through $\mathrm{s}_i = [f_1(x_i), \dots, f_M(x_i)]$.

Proposal Our aim is to develop an approach that exploits three common properties of modern multi-model settings: (1) multiple available trained models, (2) abundance of unlabeled data, relative to labeled data, (3) access to each model's predicted probabilities on every input. Prior work often exploits one or two of these properties but not all three; for example, annotator aggregation methods such as Dawid-Skene assume multiple discrete label proxies and accomodate unlabeled data, but do not incorporate continuous predicted probabilities.

To leverage these properties, we introduce a general mixture model framework. In particular, for each point in our dataset, we wish to estimate both $P(y_i|\mathrm{s}_i)$, the true label, and $P(\mathbf{s_i}|y_i)$, the distribution of predictions conditional on the true label. Together, we can write $P(y_i|\mathrm{s}_i) \propto P(\mathbf{s_i}|y_i)P(y_i)$, which provides a posterior on y_i given predictions from each model $\mathbf{f}$. When y_i is unobserved (i.e. in the unlabeled dataset), we treat y_i as a latent variable; otherwise, y_i is revealed and we update the likelihood calculations accordingly. When we sum these likelihoods over the entire dataset we get:

$$\begin{aligned} P(\mathcal{D}_U, \mathcal{D}_L) &\propto \sum P(\mathbf{s_i}|y_i)P(y_i) \\ &= \lambda_L \sum_{i:i\in\mathcal{D}_L} P(\mathbf{s_i}|y_i) \cdot \sum_{i:i\in\mathcal{D}_U} P(\mathbf{s_i}|y_i)P(y_i) \end{aligned}$$

where λ_L modulates the relative weight of the labeled data in the likelihood. To parameterize $P(\mathbf{s_i}|y_i)$, we utilize *compositional data transforms*, or one-to-one mappings between $[0,1]$ and $\mathbb{R}$ (Aitchison, 1982). These invertible transformations enable us to map predicted probabilities to real space without sacrificing any information. We will refer to model predictions when transformed as *scores*.

Parametrization The class-conditional distribution of scores $P(\mathbf{s_i}|y_i)$ is flexible to any parameterization which can be reliably learned in the semi-supervised setting described above. We note the parameterized distribution as $P_\theta(\mathbf{s}|y)$.

Here, we use a kernel density estimator to parameterize $P_\theta(\mathbf{s})|y)$ and leave a detailed exploration of different mixture model parametrizations for future work. Kernel density estimators are particularly well-suited for the task because they place no parametric assumptions on the functional form of the distribution they are used to estimate; this is useful for distributions of predictions, which can vary widely across outcomes and models. Traditionally, mixture models are fit through expectation-maximization (EM), and alternate between estimating cluster assignments and estimating clusters based on those assignments. We take a similar approach, where we alternate between estimating the true label for a given example, and estimating the class-conditional distribution of model predictions based on these estimated labels. The estimated clusters are then used to infer labels for the predictions on the unlabeled dataset, and so on. For all experiments, we fit each kernel density estimator with a Gaussian kernel and a bandwidth of 0.5 and optimize the parameters using EM over 50 epochs. We initialize cluster assignments through a KNN classifier, where $K = 5$. We estimate cluster assignments using stochastic classification, where we draw the true label according to the estimated $P(y_i|\mathbf{s}_i)$. Note that this differs from *maximum a posteriori* classification, where we estimate cluster assignment based on the highest probability cluster assignment; this approach would not properly reflect the overlap of the class-conditional distributions of model predictions.

3 Experiments

Data We validate the proposed framework on two datasets. The first is MIMIC-IV Johnson et al. (2023), a large dataset of electronic health records describing 418K patient visits to a Boston-area emergency department. For this dataset, we focus on three clinically relevant tasks: **hospitalization** (predicting hospital admission based on features available during triage, $p(y = 1) = .45$), **critical outcomes** (predicting inpatient mortality or a transfer to the ICU within 12 hours, $p(y = 1) = .06$), and **emergency department revisits** (predicting a patient's return to the emergency department within 3 days, $p(y = 1) = .03$). We split and preprocess data according to prior work Xie et al. (2022); Movva et al. (2023). The second is CivilComments, a large dataset of comments collected from online news fora which have been annotated for **toxicity** Koh et al. (2021). Toxicity detection is a popular binary classification task in natural language processing, motivated by the increasing demands of moderating online platforms (Gillespie, 2020). Model comparison is particularly salient for this task; as of this writing, HuggingFace hosts hundreds of machine learning models trained to detect abusive language.

We divide the available data into three splits (where, for MIMIC-IV, no patient appears in more than one split). A portion of each dataset is used to train models in the model set; each model sees the same training data, and we define the train data splits according to prior work Movva et al. (2023); Koh et al. (2021). Of the data not used to train the models, we reserve 50% for estimating performance, and 50% for estimating ground truth for each performance metric. No method sees data from the test split, which is used to estimate ground truth performance. Additionally, no method sees data from the train split, which is assumed to be unavailable (and even if it were available, could not be used to evaluate the set of models).

Models Our set of candidate models for MIMIC-IV contains three clinical risk scores. Each are generated by different machine learning models (a logistic regression, multi-layer perceptron, and a decision tree) fit to data in the train split. For a visualization of model score distributions, see Figure 1. The models achieve similar performance in terms of ECE and AUC, with slight differences in AUPRC. For models trained to predict a critical outcome, ECEs range from .003 to .014, AUCs range from .881 to .884 and AUPRCs range from .37 to .44. Results on this set of models are meant to reflect performance when the set of models is similar along many metrics.

Our set of candidate models for CivilComments contains three predictive models, which each use the same architecture (DistilBERT Sanh et al. (2019)), paired with a different loss function. The three loss functions are empirical risk minimization, invariant risk minimization, and the CORAL algorithm, which attempts to match summary statistics—for example, means and covariances—of network activations across domains (where here, a domain refers to comments targeting different demographic groups) Sun & Saenko (2016). Models in this set exhibit greater variability in performance, with ECEs ranging from 0.05 to 0.10, AUCs ranging from 0.86 to 0.94, and AUPRCs ranging from 0.39 to 0.71.

Baselines We compare against three baselines. The first, *Labeled*, represents the standard approach to model evaluation and makes use of the available labeled data. We also compare against *Pseudo-Labeled*, which trains a classifier to predict the true label from the model predictions—directly estimating $P(y_i|\mathbf{s}_i)$—and then labels the unlabeled examples using this classifier. We additionally compare to a standard baseline drawn from the noisy annotator literature, *Dawid-Skene*, which uses multiple potentially noisy binary annotations to estimate the latent true label of each example.

Evaluation During evaluation, we wish to estimate performance metrics a practitioner comparing models may care about. We evaluate the following metrics on (details of each metric are provided in the Appendix):

1. **Expected Calibration Error (ECE)** measures the alignment between predicted probabilities and observed outcomes in a classification model. Formally, for a binary classification model f, ECE is defined as $\int_0^1 |P(y=1|f(x)=p)-p|dp$ and in practice model evaluators bin predictions based on p. We evaluate calibration with *quantile* binning across ten bins. Intuitively, calibration quantifies the average discrepancy between a model's confidence and realized outcomes.
2. **AUROC** measures the separability between two classes for a particular model's predictions, focusing on the tradeoff between true positive rate and false positive rate. A large value indicates that the model is able to sharply separate the two classes.
3. **AUPRC** measures the tradeoff between precision and recall. A large value indicates that the model is able to simultaneously achieve high recall and precision. AUPRC is often useful in label imbalanced settings.

In our experiments, we measure our framework's ability to recover the ground truth for the three metrics above. We measure the mean absolute difference between our estimated ECE with limited labeled data and the "true" ECE if every point were labeled. Smaller values in $|\text{True ECE} - \text{Estimated ECE}|$ indicate better estimates of ECE.

4 Results

Figure 2 plots the accuracy of performance estimates using different methods, across four outcomes; lower is better in all graphs.

Performance across metrics We observe the greatest benefits relative to labeled data alone when measuring expected calibration error. We hypothesize that this is because estimating ECE requires *binning* and then averaging calibration error across bins. This process tends to yield greater variability when the number of labeled points per bin is small. In contrast, metrics like AUROC and AUPRC do not bin predictions and thus are more competitive with labeled data alone.

Comparison to baselines Dawid-Skene, as expected, performs poorly across metrics and outcomes. This is unsurprising because the method restrictively assumes that errors between models are independent, conditional on class; this is not true across the four outcomes we consider, and is unlikely to be true of machine learning models in general. In addition, Dawid-Skene operates on binary predictions on each example, and is unable to exploit the information contained in the continuous distribution of model predictions. There are cases where pseudo-labeling using logistic regression (red) can outperform the KDE (for example, estimating AUC for the ED revisit outcome or ECE for the hospitalization outcome), but this behavior is not consistent (the KDE outperforms pseudo-labeling in label-constrained settings in the remaining 7 metric-outcome combinations).

Improvements in model rankings In many cases, we might hope to recover *relative* estimates of performance metrics (e.g. ranking) instead of the absolute metrics. Figure 3 evaluates our ability to select the best model (according to a pre-specified metric). Our results suggest that the KDE method is able to substantially improve our ability to identify the best model compared to the use of labeled data alone (with 10 and 50 labeled examples, plotted in different shades). Concretely, the use of 10 labeled examples identifies the best model according to ECE a third of the time,

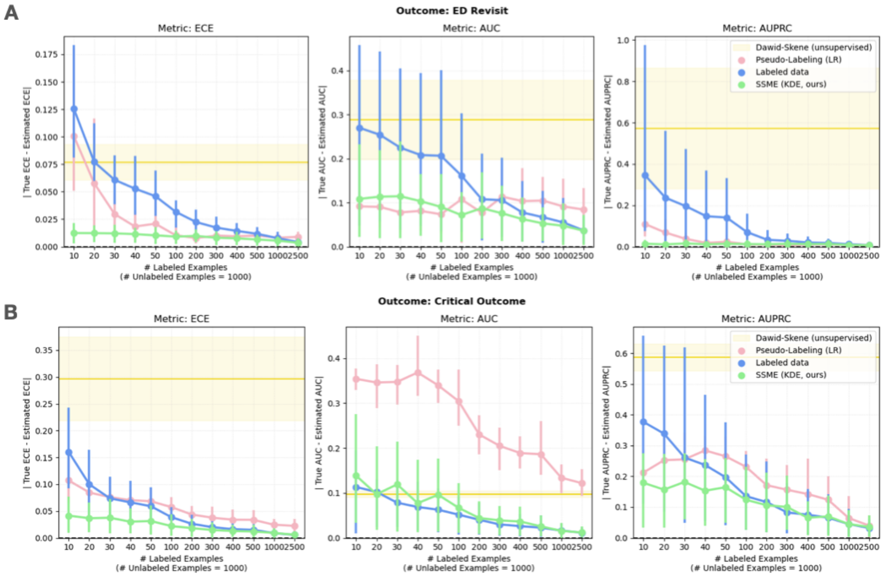

Figure 2: **Comparison to baselines.** We compare SSME to three baselines: labeled data alone, pseudo-labeling with a logistic regression trained on the available labeled data, and Dawid-Skene, an unsupervised method to aggregate multiple potentially noisy binary annotations. For estimating expected calibration error (ECE), SSME outperforms labeled data alone across each of the four outcomes. For AUC, SSME outperforms or matches labeled data for three outcomes, and worsens estimates of AUC for hospitalization prediction. For AUPRC, SSME again outperforms or matches labeled data on three of the four considered outcomes. Plots for the remaining two outcomes are in the supplement.

equivalent to choosing at random. The addition of 1000 unlabeled examples significantly improves this percentage, to approximately 80%. With 50 labeled examples, the rate of correctly ranking the best model rises to 55%, while an additional 1000 unlabeled examples increases this value to 83%. It is easier to identify the best model according to AUC and AUPRC given labeled data alone; with 10 labeled examples, model selection is better than random, and with 50 labeled examples, the best model selection rate rises to over 70%. Unlabeled data produces significant improvements in our ability to rank models by AUC (reaching near perfect accuracy with 1000 unlabeled examples), and matches the performance of labeled data alone for AUPRC. Just as the heatmaps indicate, knowledge of the prior helps substantially; Figure 4B plots the improvements in best model selection when the prior is available, and exhibits steeper improvements with the addition of unlabeled data.

5 Discussion

Our empirical results are promising and indicate that there are advantages to augmenting model evaluation with unlabeled data, particularly for the estimation of calibration error with few labeled examples. Perhaps most importantly, the use of unlabeled data can be useful for the practical question of identifying the best model for a particular task, across each of the three metrics we consider.

Several open questions remain. We consider exclusively binary tasks and two datasets; it will be important to verify the generalizability of these results to additional tasks and domains. We also employ a particular mixture modeling technique based on kernel density estimators. While the model is well-motivated by the practical considerations of multi-model evaluation, kernel density estimators are known to generalize poorly to higher dimensions. For cases with large model sets, or multiple classes, alternative techniques that scale to high dimensions (for example, normalizing flows) could be preferable.

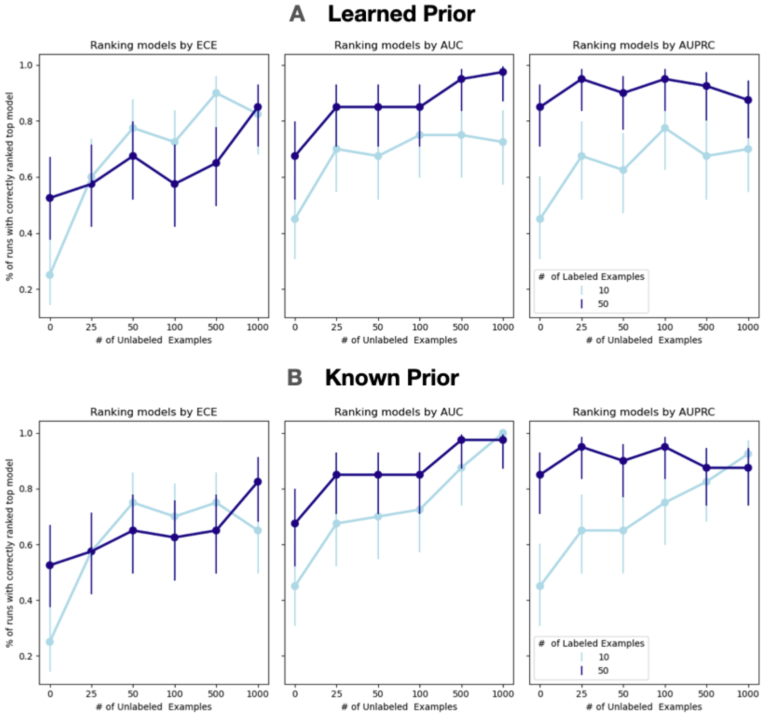

Figure 3: **Improvements in model ranking.** We plot the probability that we correctly rank the best model, according to different metrics, across 50 runs. When unlabeled data is zero, we use the labeled data alone. When we learn the prior (top), we see significant improvements in our ability to identify the b eset model with the addition of unlabeled data. We see greatest benefits for identifying the best calibrated model. When we introduce knowledge of the prior (bottom), there are steeper improvements in model rankings with more unlabeled data.

References

J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 0035-9246. URL https://www.jstor.org/stable/2345821. Publisher: [Royal Statistical Society, Wiley].

Sebastian Pineda Arango, Fabio Ferreira, Arlind Kadra, Frank Hutter, and Josif Grabocka. Quick-Tune: Quickly Learning Which Pretrained Model to Finetune and How, July 2023. URL http://arxiv.org/abs/2306.03828. arXiv:2306.03828 [cs].

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the Structure of Generative Models without Labeled Data, September 2017. URL http://arxiv.org/abs/1703.00854. arXiv:1703.00854 [cs, stat].

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art, May 2017. URL http://arxiv.org/abs/1703.09207. arXiv:1703.09207 [stat].

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? arXiv, November 2022. doi: 10.48550/arXiv.2211.13972. URL http://arxiv.org/abs/2211.13972. arXiv:2211.13972 [cs].

Lingjiao Chen, Matei Zaharia, and James Zou. Estimating and Explaining Model Performance When Both Covariates and Labels Shift, September 2022. URL http://arxiv.org/abs/2209.08436. arXiv:2209.08436 [cs, stat].

Alexandra Chouldechova, Siqi Deng, Yongxin Wang, Wei Xia, and Pietro Perona. Unsupervised and Semi-supervised Bias Benchmarking in Face Recognition. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pp. 289–306, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19778-9. doi: 10.1007/978-3-031-19778-9_17.

Cynthia S. Crowson, Elizabeth J. Atkinson, and Terry M. Therneau. Assessing Calibration of Prognostic Risk Scores. *Statistical methods in medical research*, 25(4):1692–1706, August 2016. ISSN 0962-2802. doi: 10.1177/0962280213497434. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3933449/.

A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 20–28, 1979. ISSN 0035-9254. doi: 10.2307/2346806. URL https://www.jstor.org/stable/2346806. Publisher: [Wiley, Royal Statistical Society].

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance, October 2022. URL http://arxiv.org/abs/2201.04234. arXiv:2201.04234 [cs, stat].

Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2): 2053951720943234, 2020.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with Confidence on Unseen Distributions, August 2021. URL http://arxiv.org/abs/2107.03315. arXiv:2107.03315 [cs, stat].

Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *International conference on machine learning*, pp. 4615–4630. PMLR, 2020.

Disi Ji, Padhraic Smyth, and Mark Steyvers. Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference, October 2020. URL http://arxiv.org/abs/2010.09851. arXiv:2010.09851 [cs, stat].

Disi Ji, Robert L. Logan, Padhraic Smyth, and Mark Steyvers. Active Bayesian Assessment of Black-Box Classifiers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 7935–7944, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i9.16968. URL https://ojs.aaai.org/index.php/AAAI/article/view/16968. Number: 9.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. Assessing Generalization of SGD via Disagreement, May 2022. URL http://arxiv.org/abs/2106.13799. arXiv:2106.13799 [cs, stat].

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, June 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2018340118. URL https://pnas.org/doi/full/10.1073/pnas.2018340118.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Yuzhe Lu, Yilong Qin, Runtian Zhai, Andrew Shen, Ketong Chen, Zhenlin Wang, Soheil Kolouri, Simon Stepputtis, Joseph Campbell, and Katia Sycara. Characterizing Out-of-Distribution Error via Optimal Transport, May 2023. URL http://arxiv.org/abs/2305.15640. arXiv:2305.15640 [cs].

Benjamin A. Miller, Jeremy Vila, Malina Kirn, and Joseph R. Zipkin. Classifier Performance Estimation with Unbalanced, Partially Labeled Data. In *Proceedings of The International Workshop on Cost-Sensitive Learning*, pp. 4–16. PMLR, August 2018. URL https://proceedings.mlr.press/v88/miller18a.html. ISSN: 2640-3498.

Rajiv Movva, Divya Shanmugam, Kaihua Hou, Priya Pathak, John Guttag, Nikhil Garg, and Emma Pierson. Coarse race data conceals disparities in clinical risk score performance, August 2023. URL http://arxiv.org/abs/2304.09270. arXiv:2304.09270 [cs, stat].

Alfredo Nazabal, Pablo Garcia-Moreno, Antonio Artes-Rodriguez, and Zoubin Ghahramani. Human Activity Recognition by Combining a Small Number of Classifiers. *IEEE journal of biomedical and health informatics*, 20(5):1342–1351, September 2016. ISSN 2168-2208. doi: 10.1109/JBHI.2015.2458274.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, January 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1219097111. URL http://arxiv.org/abs/1303.3257. arXiv:1303.3257 [cs, stat].

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, December 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00040. URL https://direct.mit.edu/tacl/article/43448.

Gregor Pirš and Erik Štrumbelj. Bayesian Combination of Probabilistic Classifiers using Multivariate Normal Mixtures. *Journal of Machine Learning Research*, 20(51):1–18, 2019. ISSN 1533-7928. URL http://jmlr.org/papers/v20/18-241.html.

Emmanouil Platanios, Hoifung Poon, Tom M Mitchell, and Eric J Horvitz. Estimating Accuracy from Unlabeled Data: A Probabilistic Logic Approach. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/95f8d9901ca8878e291552f001f67692-Abstract.html.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, November 2017. ISSN 2150-8097. doi: 10.14778/3157794.3157797. URL http://arxiv.org/abs/1711.10160. arXiv:1711.10160 [cs, stat].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Jacob Steinhardt and Percy S Liang. Unsupervised Risk Estimation Using Only Conditional Independence Structure. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/f2d887e01a80e813d9080038decbbabb-Abstract.html.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Peter Welinder, Max Welling, and Pietro Perona. A Lazy Man's Approach to Benchmarking: Semisupervised Classifier Evaluation and Recalibration. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3262–3269, June 2013. doi: 10.1109/CVPR.2013.419. ISSN: 1063-6919.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9:658, October 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01782-9. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9610299/.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning, June 2021. URL `http://arxiv.org/abs/2102.11005`. arXiv:2102.11005 [cs].

A Experimental Details

Dataset We validate the proposed framework on MIMIC-IV (Johnson et al., 2023), a large dataset of electronic health records describing 418K patient visits to a Boston-area emergency department. We focus on three clinically relevant tasks: **hospitalization** (predicting hospital admission based on features available during triage, $P(y = 1) = .45$), **critical outcomes** (predicting inpatient mortality or a transfer to the ICU within 12 hours, $P(y = 1) = .06$), and **emergency department revisits** (predicting a patient's return to the emergency department within 3 days, $P(y = 1) = .03$). We split and preprocess data according to prior work (Xie et al., 2022; Movva et al., 2023); for a full list of features, please refer to Table S1 in Movva et al. (2023). We divide the available data into three splits, where no patient appears in more than one split. We reserve 30% of the data for classifier training. We reserve an additional 35% for estimating performance, which we refer to as the estimation split. Our mixture models are fit to this data. The test split contains the remaining 35% of available data. No method sees data from the test split, which is used to estimate ground truth performance. For all experiments where relevant, we set λ_l to be 2000.

Classification Models We use two sets of candidate classifiers, one with real data and one with synthetic. The real set of candidate classifiers contains three clinical risk scores. Each risk score is generated by different machine learning classifier: a logistic regression (LR), a decision tree (DT), and a multi-layer perceptron (MLP) fit to data in the train split. To illustrate performance of the proposed mixture model in a well-specified setting, we also generate a synthetic set of candidate classifiers, in which we simulate each classifier's scores based on the empirical mean and variance of its class-conditional score distributions. The resulting joint distribution of classifier scores is a multivariate Gaussian (in which classifier scores are still correlated) and allows us to understand the extent to which mixture model misspecification may play a role in our results.

Metrics and Evaluation We estimate three continuous performance metrics for each classifier, which are of broad interest to machine learning practitioners: area under the receive-operating curve (AUROC), area under the precision-recall curve (AUPRC), and the expected calibration error (ECE). To do so, we sample a true label for each unlabeled point using our fitted posterior distribution for $P(y_i|\mathbf{s}_i)$. We then compare the sampled true label to the scores $\mathbf{s}_i[j] = f_j(x_i)$ for each model j.

These metrics capture both each classifier's ability to differentiate classes (AUROC, AUPRC) and measure how semantically meaningful the predicted probabilities are (calibration), which is considered a pre-requisite to the effective, non-discriminatory clinical decision-making (Crowson et al., 2016; Berk et al., 2017).

B Related work

Our work builds on two areas of literature: methods which use a combination of labeled and unlabeled data to 1) evaluate a single classifier or 2) evaluate the accuracy of multiple proxies. We elaborate on each below, and provide a taxonomy of related work in Table 1.

Semi-supervised classifier evaluation concerns the evaluation of a single classifier, using both labeled and unlabeled data. There are two types of assumptions works rely on to produce a semi-supervised estimate of performance. The first type of assumption places parametric constraints on the distribution of classifier scores. Several works attempt to fit a mixture model to the distribution of classifier scores (Welinder et al., 2013; Chouldechova et al., 2022; Miller et al., 2018), as we do, while others apply techniques from Bayesian calibration (Ji et al., 2020; 2021). Our work differs in that the proposed framework naturally accommodates and exploits multiple classifiers, and as our results show, doing so results in improved estimates of ground truth. As Garg et al. (2022) establish, estimating accuracy on the unlabeled data is impossible absent assumptions about the nature of the distribution shift. Examples of these assumptions include covariate shift (Chen et al., 2022; Lu et al., 2023), conditional independence of features (Steinhardt & Liang, 2016), and calibration on the unlabeled data (Guillory et al., 2021; Jiang et al., 2022). Here too, a majority of existing work focuses on evaluating individual classifiers and often rely on larger amounts of labeled data than we assume (on the order of hundreds of labeled data points). In contrast, our focus is on the evaluation of *multiple* classifiers, when the number of labeled data points is too small to reliably learn any model of distribution shift between the labeled and unlabeled data.

Semi-supervised evaluation of multiple proxies was first introduced by Dawid & Skene (1979), who proposed a method to estimate ground truth in the presence of multiple potentially noisy proxies. Many follow-on works inherited Dawid-Skene's strong assumption of class-conditional independence of proxies (Parisi et al., 2014; Platanios et al., 2017). Such an assumption is plausible in the context of medical diagnostics that use different biological features, but does not naturally translate to sets of candidate classifiers, whose predictions are likely to be correlated. Subsequent work has made an effort to relax the assumption of class-conditional independence, replacing it with independence conditional on a latent notion of example difficulty **?**Paun et al. (2018) or a or annotator quality (Ratner et al., 2017; Bach et al., 2017). However, these methods are designed to estimate the accuracy of *binary* proxies; they do not exploit the continuous probabilities available in multi-classifier evaluation. Recent work has made progress towards accommodating continuous proxies (Nazabal et al., 2016; Pirš & Štrumbelj, 2019). Their focus is optimal aggregation, in contrast to our own, which is evaluation.

	Multiple classifiers	**Continuous predictions**	**Unlabeled data**	**Labeled data**
Dawid-Skene and others	51	55	51	51
Unsupervised OOD evaluation	55	51	51	55
Semi-supervised evaluation of single classifiers	55	51	51	51
Our method	51	51	51	51

Table 1: A comparison of prior work and our proposed method. Whereas previous works only use at most three sources of information, our method is able to estimate the true labels from multiple classifiers' continuous predictions with both labeled and unlabeled examples.

C Additional results

Relative value of labeled and unlabeled data Here, we restrict our focus to CivilComments and ECE because 1) the model set exhibits larger differences in performance between models and 2) the KDE produces the greatest improvements with respect to ECE. In the previous section, we considered a single setting for the number of unlabeled data points provided to each method. We next study performance estimation using the proposed framework as a function of both the number of unlabeled examples and the number of labeled examples (Figure 4, where darker cells correspond to better estimates of ECE). As expected, cells darken quickly as we advance down the heatmap and increase the number of labeled examples available.

A significant challenge to estimating performance in the absence of abundant labeled data is identifying the underlying prevalence of the outcome. The accuracy of the estimated prior has a direct impact on the accuracy of the estimated metric, for each of the metrics we consider. To disentangle the difficulty of estimating prevalence from estimating performance, we replicate the contour plot with a revealed prior; that is, the KDE estimates $P(\mathbf{f}(x_i)|y)$, but $P(y)$ is set to be the ground truth prior. We see that cells corresponding to few labeled examples and many unlabeled examples darken, suggesting that knowledge of the underlying prevalence can be useful towards semi-supervised model evaluation in label-constrained settings.

Results on additional metrics As discussed, the mixture model can be used to estimate any metric that measures discrepancies between $\hat{p}(y = 1|x)$ and y, including AUC and AUPRC. Figure **??**

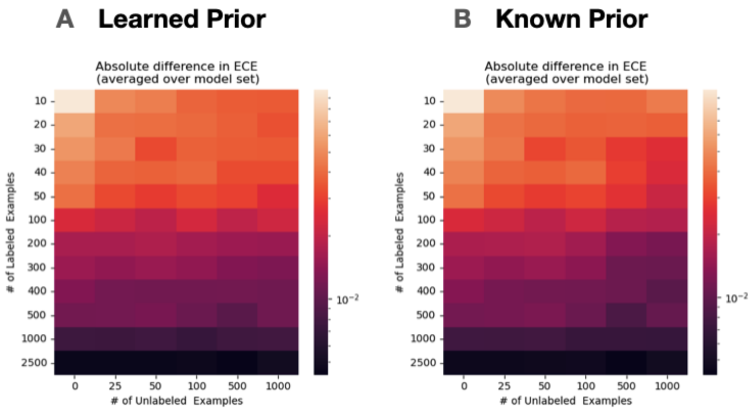

Figure 4: **Comparing the effectiveness of additional labeled and unlabeled data.** We plot improvements in ECE as we increase unlabeled data (x-axis) and as we increase labeled data (y-axis); darker cells correspond to better estimates of ECE, averaged across models. When the kernel density estimator learns the prior (left), we see that increasing labeled data and increasing unlabeled data both improve estimates of ECE, albeit at different rates. Adding knowledge of the prior (right) improves the utility of additional unlabeled data (for example, the estimation performance of 30 labeled examples and 1000 unlabeled examples improves with the addition of the known prior).

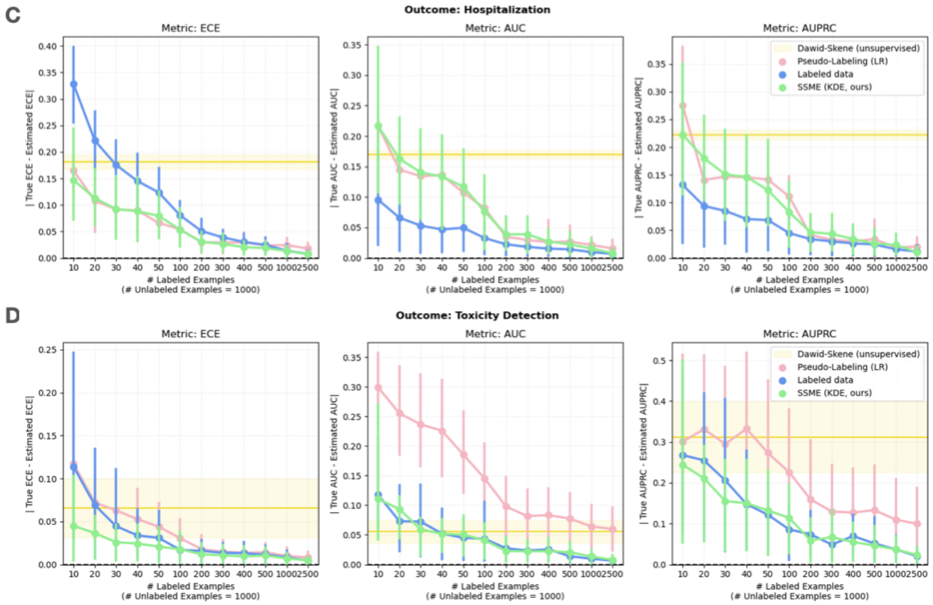

Figure 5: **Comparison to baselines on other datasets.** Our method is able to recover the ground truth performance metrics with less labeled data on both the hospitalization and toxicity classification datasets.

describes the mixture model's ability to recover AUC and AUPRC when well-specified (i.e. on the the set of synthetic classifiers). At very small amounts of labeled data (10 labeled examples), the mixture model offers improvements over using labeled data alone. The near-perfect performance of the fully-supervised mixture model suggests that it is possible for the mixture model to estimate AUC and AUPRC accurately. However, the gain relative to labeled data alone may be smaller in class-balanced, binary classification settings.

Results on additional datasets Figure 5 provides results on CivilComments and MIMIC hospitalization.

D Extensions

Multi-class settings Here we explored our method's performance in *binary* outcome settings. However, our method can be easily extended to *multi-class* outcomes as well. For a C class mixture, our likelihood can be written as follows:

$$P(\mathcal{D}_U, \mathcal{D}_L) \propto \underbrace{\lambda_L \sum_{x_i \in \mathcal{D}_L} P(\mathbf{s}_i | y_i = y_{true})}_{\text{labeled likelihood}} + \underbrace{\sum_{x_i \in \mathcal{D}_U} \sum_{y=1}^{C} P(\mathbf{s}_i | y_i = y) P(y_i = y)}_{\text{unlabeled likelihood}}$$

For each point x, we can access $\mathbf{s} \in [0,1])^M$. Compositional data transforms provide one-to-one mappings $g : [0,1] \to \mathbb{R}$. Thus, we can transform each classifier's scores $f_j(x)$ to $g(f_j(x)) \in \mathbf{R}$ without losing any information, which we concatenate across all classifiers to get $g(\mathbf{f}(x)) \in \mathbf{R}^M$. We can then fit any mixture distribution to these scores; for instance, a multivariate normal distribution enables us to model the covariance in class-conditional classifier scores.

Multi-model performance metrics While we utilized the *joint* distribution of classifier scores to fit our mixture model, each of the metrics we examined were *single* classification model scores. In this sense, we used the joint distribution to improve our estimates of the ground truth labels, but not in the classifier evaluation stage itself.

A growing body of literature on *multi-classifier metrics* provides some motivation to measure properties of the classifiers as a set. For instance, some recent work has demonstrated *systemic failures* (Bommasani et al., 2022; Kleinberg & Raghavan, 2021) across classifiers, where, for instance, a set of separate classifiers produces errors on the *same* instances. Given that we model the full joint distribution of predictions, our method can be extended to incorporate systemic failure and multi-classifier metrics as well.

Alternative mixture parameterizations Until now, we've let $P_\theta(\mathbf{f}(x)|y)$ be parameterized by a KDE, but alternative parameterizations are also possible, provided (1) they can accommodate both labeled and unlabeled data and (2) can be fit to the mixture model framework described above.

In particular, we've implemented normalizing flow based mixture models Izmailov et al. (2020). Normalizing flows apply a learnable and invertible transform f_θ to a distribution $z \sim D_1$ to obtain $f_\theta(z) \sim D_2$. By modeling z in a latent space (e.g. as a Gaussian mixture distribution), one can move back and forth between the two distributions, as $f_\theta^{-1}(f_\theta(z)) = z$.