Multi-Model Evaluation with Labeled and Unlabeled Data

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

Comparing machine learning classification models requires access to labeled data, but high-quality labeled data is often prohibitively expensive to obtain. Despite the practical importance of model comparison in label-constrained settings, there has been relatively little work focused on this question. We propose a framework for model comparison that uses labeled *and* unlabeled data. Our key insight is to directly estimate the joint class-conditional distribution of classifier scores using a mixture model, where the latent mixture variable is the true class (for an extended discussion of related work, please see Sec. B). This framework has several advantages in that it can characterize any number of classifiers, accepts continuous predictions from each classifier, and exploits both labeled and unlabeled data. The resulting mixture model can estimate both individual (e.g. calibration error) and relative performance metrics (e.g. relative accuracy).

Method Details We consider a setting in which a consumer (e.g., a hospital system) must choose between several classification models. Formally, we have a set of M classification models $\{f_1, f_2, \ldots, f_M\}$ designed for the same prediction task, so $f_i : X \to \Delta^D$, where X is the domain of the input and Δ^D is the D-simplex representing a model's predicted probability for each of the D classes. Models in the set may differ by function class, training data, or training hyperparameters. For each point x, we let $\hat{\mathbf{f}}(x) = [f_1(x), \ldots, f_M(x)] \in (\Delta^D)^M$ be the model set's predictions; that is, we observe M simplex draws.

During evaluation, we have access to two datasets: (1) a small labeled dataset, $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{\ell}$ and (2) a larger unlabeled dataset $\mathcal{D}_U = \{(x_i)\}_{i=1}^{u}$, which are both drawn i.i.d. from the same distribution. We maximize the likelihood of our datasets, which we write as: $P(\mathcal{D}_U, \mathcal{D}_L) \propto \sum P(\mathbf{f}(x_i)|y_i)P(y_i)$. Splitting this sum into labeled and unlabeled components, we get $\lambda_L \sum_{x_i \in \mathcal{D}_L} P(\mathbf{f}(x_i)|y_i = y_{\text{true}}) + \sum_{x_i \in \mathcal{D}_U} \sum_y P(\mathbf{f}(x_i)|y_i = y)P(y_i = y)$, where λ_L modulates the weight of the labeled data in the likelihood.

We instantiate $P(\mathbf{f}(x)|y)$ as a multivariate logit-normal distribution. Through the use of *compositional data transforms* to map Δ^D to $\mathbb{R}^{D-1}$ (Aitchison, 1982), we are able to model the transformed data with a multivariate normal distribution (see Fig. 1 for distribution of transformed probabilities). This parameterization allows us to use Gaussian Mixture Models, which can be easily and quickly fit with Expectation-Maximization (EM). However, the approach is flexible to any parameterization amenable to semi-supervised estimation.

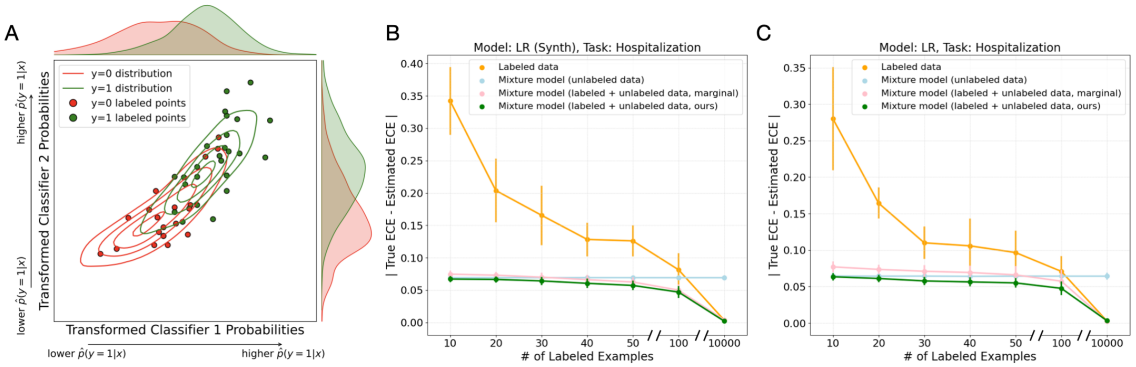

Figure 1: (A) Joint distribution of model predictions. (B) Estimation of ECE for synthetic model set. (C) Estimation of ECE for real model set. We report results across 10 validation/test splits, and include results for remaining models in XXX.

2 Preliminary experiments

We report results using two classification model sets: one containing three synthetic classifiers, in which we simulate classifier scores to follow a multivariate normal distribution, and one containing three classifiers trained on real health data (Xie et al., 2022). We begin with synthetic classifier scores to illustrate performance when the multivariate normal mixture model is well-specified. On the real data, each classification model is evaluated on its ability to predict a patient's risk of hospitalization, based on features available during triage in the emergency department. We elaborate on the dataset and classifiers in Sec. A, and include results for additional classifiers, metrics, tasks in Sections C, D, and E.

Figure 1B and C compare the performance of the proposed framework to using labeled data alone (orange). We also compare to mixture models fit to the unlabeled data alone (blue) and to the scores of a single classification model (i.e. the marginal distribution of classifier scores, pink). We observe greatest benefits relative to labeled data alone when measuring expected calibration error, an important metric in risk-sensitive applications such as clinical medicine. 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 — which measure discriminative power — do not bin predictions and thus are more competitive with labeled data alone.

Notably, fitting the mixture model on unlabeled data alone provides better ECE estimates than labeled data alone when the amount of labeled data is small (< 100 points). Furthermore, using the full *joint* distribution of scores yields better ECE estimates than only using the *marginal* distribution of scores from a single classifier, as multiple distinct predictions for a given example offer more information about underlying ground truth than any one alone (Dawid and Skene, 1979; Ratner et al., 2017; Platanios et al., 2017).

Future directions First, we intend to extend the framework to non-Gaussian distributions, since these frequently occur in real-world clinical tasks. The multivariate normal performs less well in these settings (Sec. 3), and it represents a simple extension of the framework. Second, classifier choice is ubiquitous across applications, and we hope to apply our method to tasks beyond healthcare.

References

Hervé Abdi and Paul Molin. Lilliefors/Van Soest's test of normality.

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].

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, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 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].

Harvey Goldstein, James Carpenter, Michael G Kenward, and Kate A Levin. Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197, October 2009. ISSN 1471-082X, 1477-0342. doi: 10.1177/1471082X0800900301. URL http://journals.sagepub.com/doi/10.1177/1471082X0800900301.

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].

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 Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. URL https://physionet.org/content/mimiciv/2.1/.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised Learning with Deep Generative Models, October 2014. URL http://arxiv.org/abs/1406.5298. arXiv:1406.5298 [cs, stat].

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.

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*, pages 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].

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.

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*, pages 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/.

Appendix A. Experimental Details

Dataset We validate the proposed framework on MIMIC-IV (Johnson et al.), a large dataset of electronic health records describing 418K patient visits to a Boston-area emer-

gency 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 directly sample from our mixture model by first sampling a *true label* y according to the estimated overall class prevalences and then simulating the classifier scores by sampling from the corresponding class-conditional score distribution. For each classifier, we transform these scores back into (predicted) probabilities and then estimate the performance metrics accordingly.

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).

Appendix 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 and 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 and 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 (Goldstein et al., 2009; 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 and Štrumbelj, 2019). Their focus is optimal aggregation, in contrast to our own, which is evaluation.

Appendix C. Results across classifiers

Figure 2 reports ECE estimation results on the hospitalization task across all three synthetic candidate classifiers. As expected, the use of labeled data alone (orange) is poor with very few labeled data points. Applying the mixture model to the data *without* incorporating any labels (blue) outperforms the use of labeled data alone given small amounts of labeled data (less than 100 examples). The mixture model fit to the labeled and unlabeled scores for a single classifier (pink) provides a slight improvement. The proposed approach (green) outperforms each of these baselines, across all amounts of labeled data. We also report results for the fully supervised mixture model, for which all labels in the estimation split are revealed. While this is not a realistic baseline, it serves as useful sanity check for whether the proposed model converges to ground truth given abundant labeled data.

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

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.

Figure 3 reports ECE estimation results for the same task, on real candidate classifiers, which have been trained to predict a patient's risk of hospitalization based on data in the train split. While the proposed approach outperforms the use of labeled data alone, as previously seen, our results provide evidence of model misspecification. Indeed, when applying a test for normality, the Lilliefores test (Abdi and Molin), to the class-conditional score distributions for each classifier-task combination, we find that the distribution is not normal ($p < 0.001$ for every task and classifier).

Consider the ECE estimation performance for the MLP (Figure 3, right). The mixture model fit to all (x, y) (gray) in the estimation split performs worse than the mixture model fit to all $(x,)$ in the estimation split (pink). We see that the performance of the joint mixture model (green) *worsens* with additional labeled data. Ultimately, these preliminary results provide motivation to extend the framework to more flexible parametrizations.

Appendix D. 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 5 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.

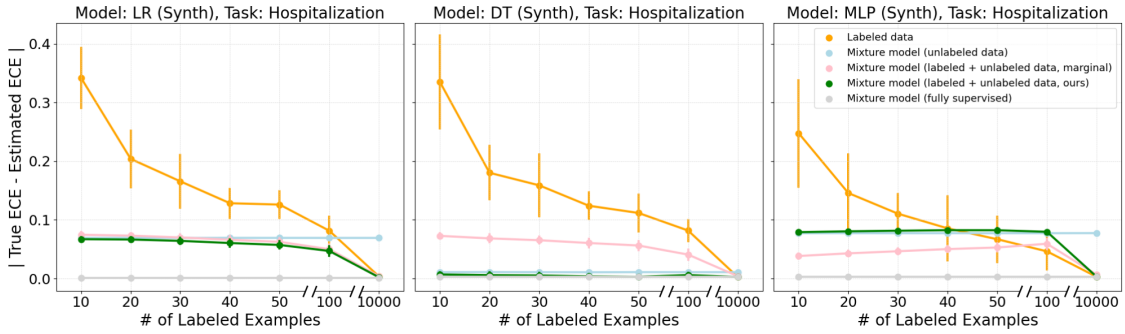

Figure 2: **Mixture model performance on simulated classifiers (ECE).** We measure the absolute error when estimating expected calibration across different amounts of labeled data using different approaches. Here, the mixture model is well-specified; that is, when the joint distribution of transformed classifier predictions follows a multivariate normal distribution. The mixture model fit to estimate the joint distribution outperforms all other considered baselines. Encouragingly, the mixture model fit to all (x, y) in the estimation split (gray) achives 0 estimation error, and serves to upper bound the performance of our approach.

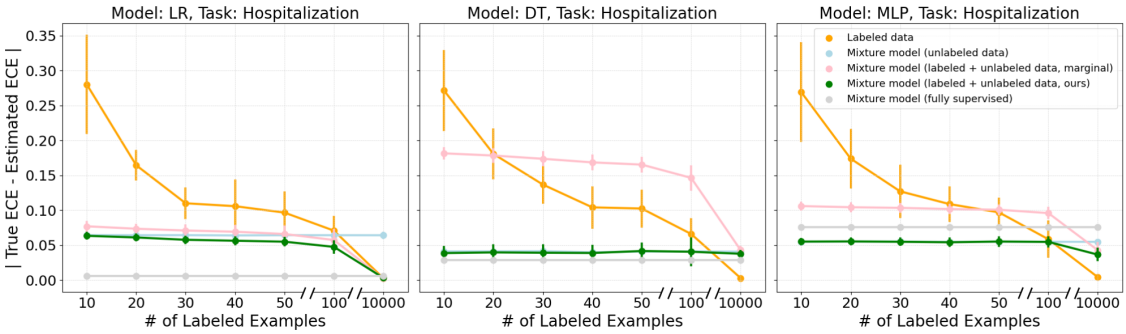

Figure 3: **Mixture model performance on real classifiers (ECE).** Fitting a mixture model to estimate the joint distribution of classifier predictions (green) outperforms the use of labeled data alone, under limited labeled data (< 100 points). Note, however, that the mixture model with access to all labels in the estimation split (gray) fails to recover the true ECE; this suggests that the mixture model is misspecified. For the MLP (right), for example, fitting a mixture model to all (x, y) in the estimation split (gray) produces worse ECE estimates than fitting to only $(x,)$ (pink). This behavior forms our motivation to extend the framework to more flexible parameterizations.

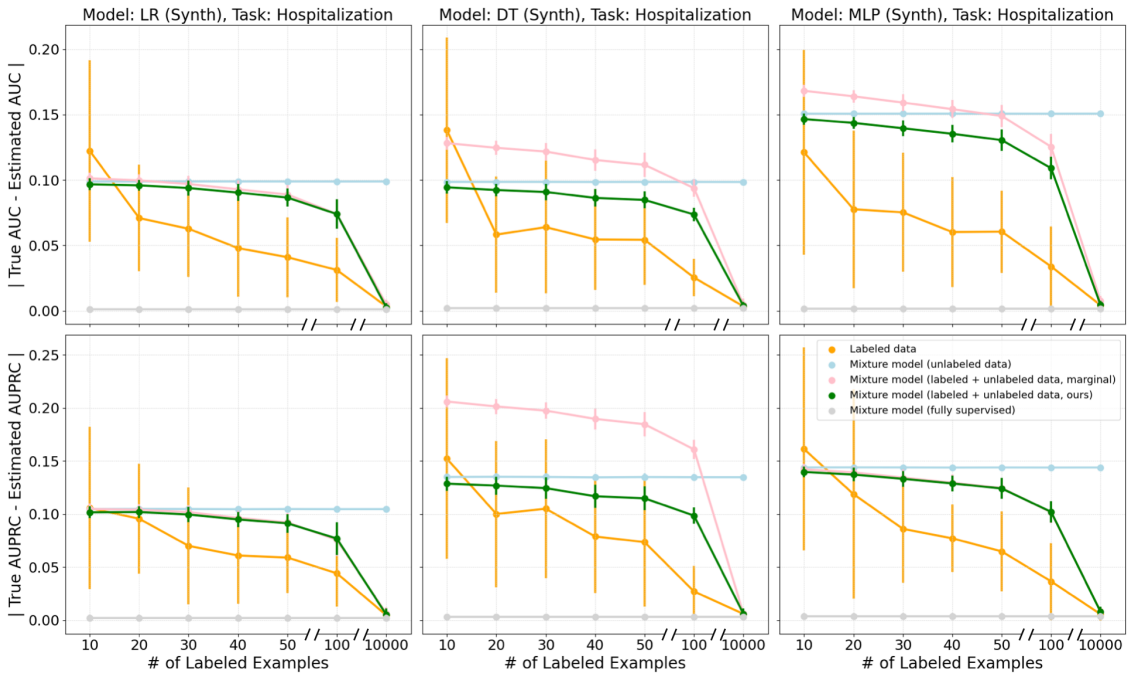

Figure 4: **Mixture model performance on synthetic classifiers (AUC, AUPRC).** While the mixture model can offer more accurate estimates of AUC and AUPRC given extremely few labeled examples (e.g. 10), the difference is not meaningful. Further, there are cases where fitting the mixture model to a single classifier's scores (pink) outperforms fitting the mixture model to the joint distribution of classifier scores. We suspect we would see larger variability in estimates of AUC using 10 labeled points under higher class imbalance.

Appendix E. Results on additional tasks

Thus far, we have discussed results in the context of predicting patient hospitalization. Here we consider how our results generalize to two other clinical tasks: predicting a critical outcome for a patient (defined as death or admission to the intensive care unit) and predicting whether a patient with revisit the emergency department within three days. Both tasks demonstrate much lower prevalence of positive visits (.06 and .03 respectively) compared to hospitalization, for which 45% of visits are positive. We restrict our discussion of these results to the synthetic classifier case, where we simulate classification model scores based on the empirical mean and variance of each classifier's class-conditional distribution of scores, since we have established the role of model misspecification with respect to the real classifiers.

Figure E plots results for the synthetic classifiers, equating to a setting in which the model is well-specified. The Gaussian mixture model fit to the joint distribution of transformed model scores (green) and the Gaussian mixture model fit to the marginal distribu-

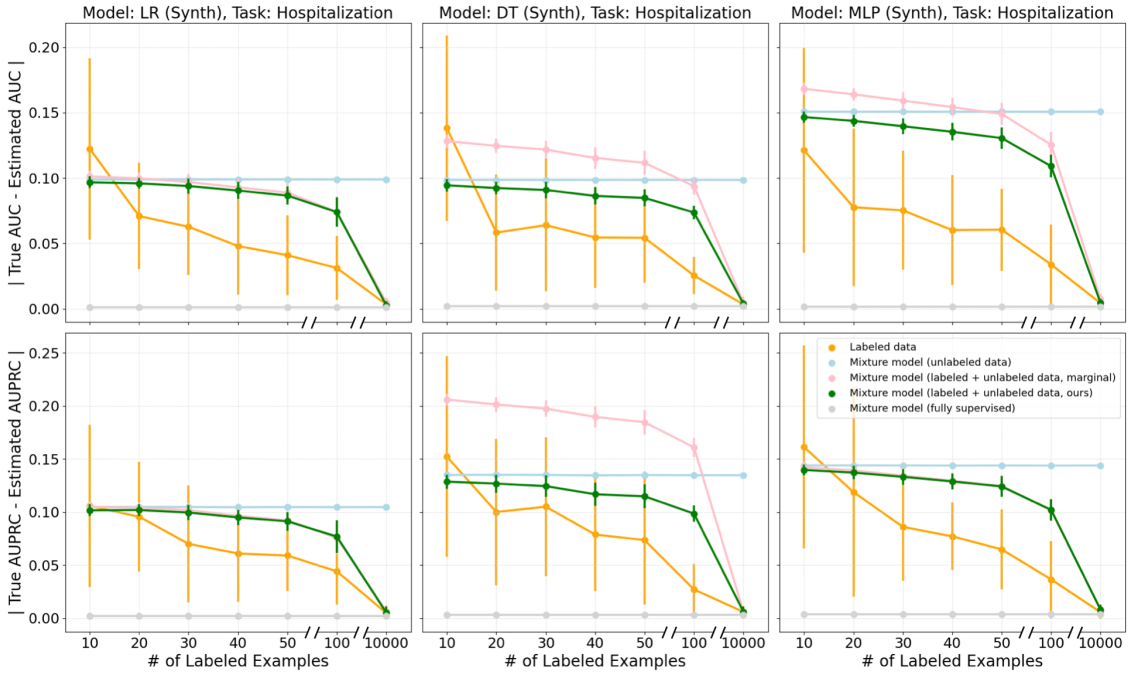

Figure 5: **Mixture model performance on synthetic classifiers (AUC, AUPRC).** While the mixture model can offer more accurate estimates of AUC and AUPRC given extremely few labeled examples (e.g. 10), the difference is not meaningful. Further, there are cases where fitting the mixture model to a single classifier's scores (pink) outperforms fitting the mixture model to the joint distribution of classifier scores. We suspect we would see larger variability in estimates of AUC using 10 labeled points under higher class imbalance.

tion of transformed model scores (pink) both fail to match the performance of labeled data alone. The stark difference in performance can be attributed to class imbalance; at very small sample sizes and very low positive prevalences, it is difficult to observe both classes. These results suggest the importance of support for both classes in the labeled dataset, and further suggest that there are certain tasks for which mixture modeling may be too difficult.

Appendix F. 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:

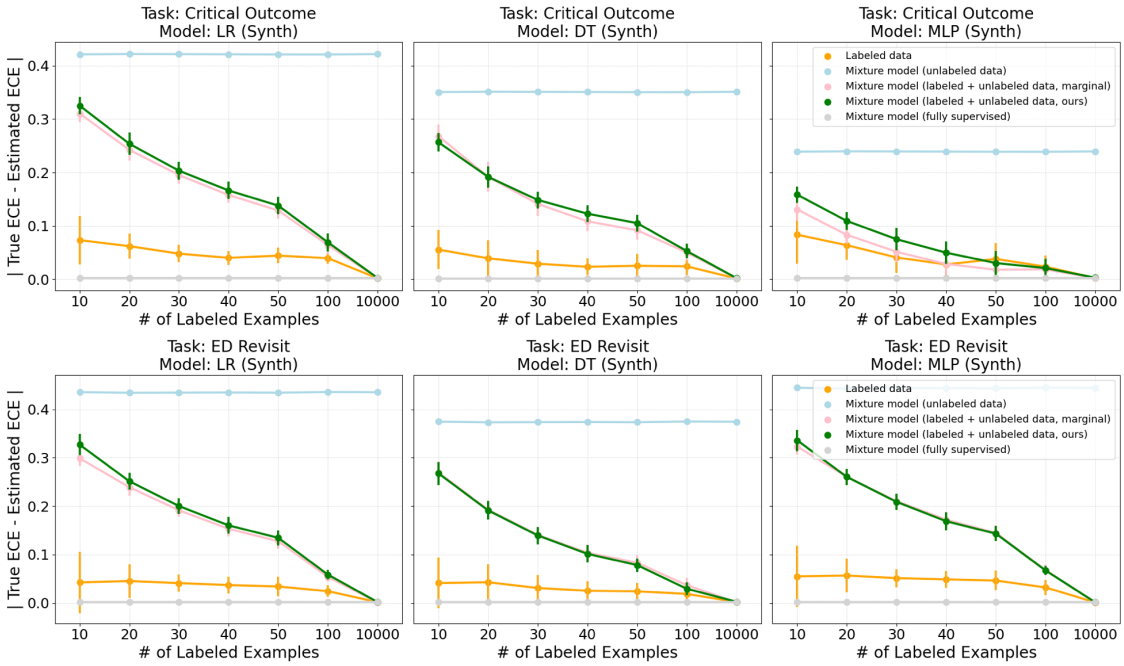

Figure 6: **Mixture model performance (ECE) on synthetic classifiers across two additional tasks.** On two additional tasks (top: prediction of a critical outcome, bottom: prediction of an ED revisit), the mixture model parametrization as a multivariate normal performs poorly. Fitting the GMM to a single model's scores (pink) outperforms estimating the joint distribution by a small margin, but neither compete with the use of a small amount of labeled data. Both tasks suffer from class imbalance, and thus small samples (less than 100 labeled data points) may contain no positive examples. A potentially fruitful assumption to make is that the labeled dataset contains support for all classes.

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

For each point x, we can access $\mathbf{f}(x) \in (\Delta^D)^M$. Compositional data transforms provide one-to-one mappings $g : \Delta^D \to \mathbb{R}^{D-1}$. Thus, we can transform each classifier's scores $f_j(x)$ to $g(f_j(x)) \in \mathbf{R}^{D-1}$ without losing any information, which we concatenate across all classifiers to get $g(\mathbf{f}(x)) \in \mathbf{R}^{(D-1)\times M}$. 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.

Once the mixture model parameters are fit, we can estimate any metrics we'd like, such as ECE, by (1) sampling a true label y, (2) sampling the classifier scores $\mathbf{f}(x)$, sampled from our fitted distribution for $g(\mathbf{f}(x))|y$ and applying g^{-1}, and (3) measuring the discrepancies between the sampled scores for each classifier and true labels.

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 and 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 $g(\mathbf{f}(x))|y = c \sim \mathcal{N}(\mu_c, \Sigma_c)$, 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. As noted in Appendix C, each of the class-conditional score distributions on our real-world health dataset were not normal ($p < 0.001$) across all tasks and classifiers we examined, so alternative parameterizations are necessary to explore.

Directions to explore include semi-supervised class-conditional variational autoencoders (Kingma et al., 2014) and Dirichlet mixture models.