Trust Issues: Uncertainty Estimation Does Not Enable Reliable OOD Detection On Medical Tabular Data

Dennis Ulmer DENNIS.ULMER@MAILBOX.ORG
Lotta Meijerink LOTTA.MEIJERINK@PACMED.NL
Giovanni Cinà GIOVANNI.CINA@PACMED.NL
Pacmed BV - Amsterdam, Netherlands

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[‡], Stephanie L. Hyland[‡]

Abstract

When deploying machine learning models in high-stakes real-world environments such as health care, it is crucial to accurately assess the uncertainty concerning a model's prediction on abnormal inputs. However, there is a scarcity of literature analyzing this problem on medical data, especially on mixed-type tabular data such as Electronic Health Records. We close this gap by presenting a series of tests including a large variety of contemporary uncertainty estimation techniques, in order to determine whether they are able to identify out-of-distribution (OOD) patients. In contrast to previous work, we design tests on realistic and clinically relevant OOD groups, and run experiments on real-world medical data. We find that almost all techniques fail to achieve convincing results, partly disagreeing with earlier findings.

Keywords: Uncertainty Estimation, OOD Detection, Electronic Health Records

1. Introduction

In healthcare, tremendous potential has been identified for applications deep learning methods for e.g. prediction, screening, finding or designing better treatments for patients (Sheikhalishahi et al., 2019; Esteva et al., 2019). However, safe deployment in practice demands certain properties from models: in the context of healthcare, models have to prove to be interpretable and deliver trustworthy predictions (He et al., 2019). In particular, a trustworthy application should be able to identify new samples outside the training population. Although we always hope for models to generalize to novel data points, a significant shift in the data can render the current predictor unreliable. The COVID-19 pandemic offers an example of this phenomenon on a large scale: the training data used for models deployed before the outbreak might not have contained patients with these new combinations of symptoms. Thus, the predictions of said models on COVID-patients should have a lower degree of certainty to avoid misleading end-users like doctors and nurses. The same could happen due to other, more common factors such as changes in the patient demographics, evolving clinical protocols or scaling of a device (Curth et al., 2019).

Deep neural networks are notorious for delivering predictions in an overconfident manner, be it by confidently predicting the wrong class after minimal perturbation of an input (Goodfellow et al., 2015), assigning high probability to OOD samples (Nalisnick et al.,

2018) or simply being badly-calibrated in general (Guo et al., 2017). Such observations have motivated the development of a series of techniques to more accurately estimate the uncertainty about an outcome (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Blundell et al., 2015). Alas, there is still a limited volume of experiments on tabular data and classification tasks, a context of particular importance for health care applications. Notably, Ovadia et al. (2019) conducted an extensive study of the quality of uncertainty estimates under covariate shift, by artificially constructing corrupted OOD samples. We argue that this procedure is insufficient to evaluate these techniques in a real-world healthcare application: the generalization of performance to OOD examples in an artificial scenario is no guarantee of robustness, especially if the model in question is unable to detect OOD reliably. Inability to recognize OOD could result in confident predictions on abnormal data points for which the model actually under-performs.

We study the efficacy of a large array of uncertainty estimation methods for OOD detection in the context of a binary classification task with clinical significance, i.e. prediction of mortality for ICU patients, by running experiments on two large mixed-type, highly unbalanced tabular data sets. We provide three main contributions.[1] First, we design a number of clinically relevant OOD experiments, including scenarios for changing patient population, variation of clinical protocol and data corruption. Second, we show that all tested uncertainty estimation techniques for neural discriminators fail to reliably detect OOD samples across experiments. Lastly, we provide novel benchmarks for OOD detection on two large tabular medical data sets.

1. All code is publicly available under https://github.com/Pacmed/ehr_ood_detection.

2. Related Work

Uncertainty from a modelling perspective is often separated into aleatoric and epistemic uncertainty. While the former usually refers to *non-reducible* uncertainty, e.g. uncertainty inherent in the data generating process, the latter denotes *reducible* uncertainty. This encompasses both the lack of knowledge about the ideal model for a problem, as well as about its best parameters (Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2019). In our case, we are concerned with methods that try to estimate either only the epistemic or both types of uncertainty in a post-hoc manner.

One method to appraise uncertainty estimation methods is the study of their behaviour in presence of OOD samples (Ovadia et al., 2019). The topic of OOD detection has been studied extensively; we present here a selection of articles. A simple baseline for discriminators for the task was introduced by Hendrycks and Gimpel (2017). Lee et al. (2018) fit a generative model on the pre-softmax outputs of neural discriminators and employ the Mahalanobis distance of data samples as a confidence metric. Jiang et al. (2018) employ the distance between a sample and the next sample of another class as a non-probabilistic trust score. Other articles use density estimation models instead (Xiao et al., 2020; Choi et al., 2018), while some identify their shortcomings when it comes to detecting OOD samples in the image domain (Nalisnick et al., 2018; Kirichenko et al., 2020). Eduardo et al. (2020) introduce a variant of the variational autoencoder for tabular data, which is able to assign OOD scores to both single cells (i.e. features) and row entries (i.e. whole data points) alike. A related line of work investigates the phenomenon of covariate shift: Ovadia et al. (2019) conduct a large study to analyze model calibration under increasing shift. Park et al. (2020) propose a

procedure to explicitly adapt models to shifting data distributions. There is also some work in the medical context about this topic, with e.g. Curth et al. (2019) deriving formal guarantees for domain adaptation procedures and testing them on clinical data sets.

Given the importance of the reliability of models in a health care setting, the application of such methods to tabular medical data sets constitutes important related work. However, the literature on the topic is scarce. Myers et al. (2020) introduce a model-independent reliability metric based on the difference between the model's prediction and predictions for patients of the same class. Ruhe et al. (2019) and Meijerink et al. (2020) test different models for mortality prediction on electronic health records, albeit with limited experiments, while Dusenberry et al. (2020) also test models and their uncertainty in predicting diagnoses. The latter work however restricts its scope to only two types of models and does not evaluate uncertainty estimates for novel patient groups.

3. Background

First of all, we lay out some relevant definitions and notations. Let $p : \mathbb{R}^D \to [0, 1]$ denote a probability density function from which data samples $\boldsymbol{x}_i \in \mathbb{R}^D$ stem from and let $p(\boldsymbol{x}, y) : \mathbb{R}^D \times \{0, 1\} \to [0, 1]$ denote the joint distribution of points and labels with $y \in \{0, 1\}$ in a binary classification setting. We indicate with $\mathcal{D}$ the set of tuples contained in a data set. We consider a data point out-of-distribution if its underlying distribution q underwent a *covariate shift* relative to our model's training distribution p, namely if $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are different.[2] Covariate shift is especially prevalent in non-stationary environments (Moreno-Torres et al., 2012), which abound in healthcare (Curth et al., 2019).

We consider the following metrics for OOD detection: the maximum softmax probability baseline by Hendrycks and Gimpel (2017) and the standard deviation of probabilities of class 1 aggregated from multiple different sources, e.g. different members of an ensembles or predictions produced under distinct sets of weights samples from a posterior distribution. In these cases, we also compute the predictive entropy (Gal, 2016) or mutual information (Smith and Gal, 2018) across different predictions. Due to space limitations, we refer the reader to Appendix A for a more in-depth explanation of these methods.

4. Methodology

We describe the experimental setup, the models employed and the evaluation metrics.

4.1. Experimental Design

Mortality Prediction Task As we are interested in testing the OOD-detection capabilities of models trained on a clinically relevant classification task, we focused only on prediction of in-hospital mortality based on the first 48 hours of data from intensive care admissions. This task has already been investigated extensively as a relevant task for clinical prediction models (Sheikhalishahi et al., 2019; Ruhe et al., 2019; Curth et al., 2019; Harutyunyan et al., 2019; Meijerink et al., 2020) and provides a clear example of binary classification with unbalanced outcome.

Clinical Data Sets We performed experiments on the MIMIC-III data set (Johnson et al., 2016), which comprises health data from ICU admissions from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. From the 46476 patients and 61532 ICU stays in the MIMIC-III data set, data points are selected based on patient age

2. Other authors such as Shimodaira (2000) and Moreno-Torres et al. (2012) give more restrictive definitions of covariate shift (e.g. assuming $p(y|\boldsymbol{x}) = q(y|\boldsymbol{x})$) that do not apply in our setting.

($\geq$ 18 years) and length of stay ($\geq$ 48 hours). Furthermore, stays were excluded when there was no data in the first 48 hours or when there were multiple IC transfers within one hospital admission. After selection, the cohort consisted of 21139 IC stays corresponding to 18094 patients, with a mortality rate of 13.23%. We also employed the eICU data set (Pollard et al., 2018), which contains ICU data from patients admitted to different hospitals in the United States in 2014 and 2015. We pre-processed the data using the general pipeline presented in the work of Sheikhalishahi et al. (2019) and applied the same filtering criteria used for MIMIC-III, resulting in 38072 ICU stays, with one unique patient per stay and a mortality rate of 12.53%.

Feature Engineering To simplify our experimental pipeline and allow more direct comparisons between the two data sets, we only keep clinical variables that were present in both data sets (see appendix Table 2). We then engineered features calculating six statistics on seven sub-sequences of a time series, in line with the feature engineering for the logistic regression model of Harutyunyan et al. (2019). The sub-sequences consist of the full time series, the first 10% and last 10%, the first and last 25%, and the first and last 50%. The statistics are minimum, maximum, mean, standard deviation, skew and the number of measurements, which were all standard-scaled and mean-imputed if necessary, resulting in 588 different features. In both cases, the data sets were split into 70 % training, 15 % validation and 15 % test set.

Clinically Relevant OOD Groups Using the help of medical professionals, we select clinically relevant patient groups on both datasets. We first extract groups of patients based on age (newborns for MIMIC-III[3]), on ethnicity (white, black) and on gender, to test the model's behavior under changing patient demographics. We also separate patients by admission type (elective vs. emergency admissions) to simulate change in protocol and according to diagnosis (acute and unspecified renal failure, epilepsy, hypertension with complications and secondary hypertension and thyroid disorders) as we also envision novel sets of symptoms to emerge in patients in practice. To ensure that the resulting patients groups are sufficiently different, we perform a feature-wise Welch's t-test and report the percentage of features that differ in a statistically significant manner from the rest of the training set; the percentages are reported in the plots of the results in Section 5.

3. We were unable to identify enough newborns matching our filter criteria in eICU.

4.2. Models

In the following, we briefly discuss the models and corresponding OOD detection metrics used for our experiments. Due to space constraints we refer to the cited papers for a more detailed description of the models.

Single Models We consider different variations of a standard feed-forward neural network with a final sigmoid activation function. Besides a normal network with ReLU activation functions and intermediate dropout layers, we also consider a model which was additionally calibrated with temperature scaling (Guo et al., 2017) on the validation set. As uncertainty metrics, the model's entropy over the class distribution as well as its maximum softmax probability are considered. We also add a logistic regression baseline.

Bayesian Models We include models using MC Dropout (Gal and Ghahramani, 2016) and Bayes-by-Backprop (BBB, Blundell et al., 2015). As uncertainty metrics, we consider the standard deviation, their predictive entropy as well as their mutual information. We also attempted to train Gaussian Processes as a Bayesian baseline, but it could not scale to

the size of the data set and did not converge to useful solutions on subsets.

Ensemble Models We furthermore investigate the performance of different ensemble models, namely a group of standard feed-forward neural networks (Lakshminarayanan et al., 2017), a bootstrapped ensemble, meaning that every member is trained on a different (but not disjoint) subset of the data, as well as the anchored bayesian ensembles introduced by Pearce et al. (2020). The same metrics are employed as for the bayesian models.

Density Estimation Baselines Lastly, we train two simple density estimation models: Probabilistic principal component analysis (PPCA; Bishop, 1999), where we simply use the log-probability of a sample as an indication of its novelty, as well as an auto-encoder (AE). For the latter, its reconstruction error serves as uncertainty metric, since we expect the model to reconstruct samples closer to the training distribution in a more reliable way. Although AEs are not explicitly estimating the data density, there is a known connection to PPCA as they project samples into a latent space, albeit in a non-linear fashion (cp. e.g. Valpola, 2015).

4.3. Model Training & Evaluation

Random hyperparameter search (Bergstra and Bengio, 2012) was performed for all model types on MIMIC-III and eICU separately. The options considered, number of trials and best found hyperparameters are reported in Appendix C. All models were generally trained for at most 10 epochs using the adam optimizer (Kingma and Ba, 2014) and dropout (Srivastava et al., 2014) regularization, with early stopping if the validation loss did not decrease after 3 epochs. Logistic regression was trained using l_2 regularization.

To evaluate all the models enumerated in the last section, we consider the area under the receiver operating characteristic curve (AUC-ROC). In the case of mortality prediction, AUC-ROC is used as a standard classification metric. In the case of OOD detection, AUC-ROC is calculated for a classification task in which the classes are in-distribution and out-of-distribution. Here, the model is not evaluated based on its predicted class probabilities (about a patient's mortality), but rather on the uncertainty about its prediction. In short, AUC-ROC is employed to measure whether uncertainty helps in discriminating OOD samples.

5. Results

In this section we describe the main experimental findings on mortality prediction and OOD detection.

5.1. Mortality Prediction

Before delving deeper into OOD detection, we compare the models' performances on the mortality prediction task, the results of which are given in Table 1. Almost all models are able to solve the task well, achieving very similar results across data sets. Neural models slightly outperform the logistic regression baseline. One exception is the anchored ensemble, which seems to perform slightly worse. We attribute this effect to the regularization procedure using sampled anchored points, which might prevent ensemble members to converge to the same local minimum. This seems to constitute a trade-off between the bayesian guarantees and performance, which will become relevant again in the following section. Furthermore, BBB performs considerably worse than the other models while also displaying a comparatively high standard deviation in results. We observed this even after an extensive hyperparameter search (for details see Appendix C). We note that previous work applied BBB to similar data sets, but either using less features (Ruhe et al.,

2019) or more homogeneous ones (MNIST in the case of Blundell et al., 2015); we hypothesize that a mixture prior of only two gaussian distributions for the weights might not be expressive enough to find a suitable weight distribution for the high number and complexity of features we employ.[4]

Model	MIMIC	eICU
AnchoredNNEnsemble	0.837 ± 0.006	0.832 ± 0.004
BBB	0.628 ± 0.081	0.611 ± 0.066
BootstrappedNNEnsemble	0.847 ± 0.000	0.847 ± 0.000
LogReg	0.835 ± 0.000	0.823 ± 0.000
MCDropout	0.849 ± 0.002	0.844 ± 0.001
NNEnsemble	0.847 ± 0.000	0.847 ± 0.000
NN	0.847 ± 0.002	0.842 ± 0.002
PlattScalingNN	0.847 ± 0.002	0.844 ± 0.002

Table 1: Mortality prediction AUC-ROC on MIMIC-III and eICU. Results are averaged over $n = 5$ runs.

5.2. OOD Experiments

We conduct a series of experiments to determine whether the presented uncertainty estimation techniques are able to successfully detect OOD data. In a clinical setting, a change in the patient population is often more complex than the shift in a single feature, but it rather appears as a new configuration of covariates. A model may encounter OOD examples for a variety of reasons: a change in protocol that modifies the provenance of the patients, an intrinsic shift in the patient population, or corruption in the data due to machine or human error, to name a few. To simulate these scenarios we confront our models with novel data obtained via three methods. First, data corruption: we shift the data artificially, scaling one feature at a time by an increasingly large degree. Second, changing patient population: some clinically relevant patient groups (based on demographics, provenance or pathology), are excluded from the training set and reintroduced later for testing (such as in Meijerink et al., 2020). Third, changing data source: we repeat the same procedure, but this time using patient stays from eICU as OOD data for models trained on MIMIC-III, and vice versa.

5.2.1. Artificial Data Corruption

We investigate the ability of models to identify novel data points with abnormal feature values. To simulate extreme or possibly corrupted feature values, we scale features by a factor of 10, 100, 1000 and 10000. This is achieved by selecting a single feature at random and scaling it by the chosen factor across the whole test set. The corrupted test set and the original one are then compared in terms of uncertainty using AUC-ROC, to check whether uncertainty helps in identifying OODs. For each factor, we repeat this procedure one hundred times per model, sampling without replacement the feature to scale. We then report the average OOD detection AUC-ROC score on these samples, which is depicted for eICU in Figure 1; due to their similarity, the results for MIMIC-III were moved to Figure 8 in the appendix. In this case, we would like to see the AUC-ROC increase as the scaling of a random feature increases, i.e. the covariate of a sample becomes more and more abnormal compared to the training distribution. However, we see quite the opposite in most models regardless of the applied uncertainty estimation technique, as AUC-ROC actually *decreases*, resulting in scores much worse than random guessing. In practice, this would result in a model being overly confident in a prediction for a corrupt data point. The only notable exceptions to this trend are AE and PPCA, which can be sorted into an entirely different family of methods

4. The sensitivity of variational methods - including Bayes-by-Backprop - towards the choice of prior is part of an ongoing discussion, refer e.g. to the blog post by Gelada and Buckman (2020).

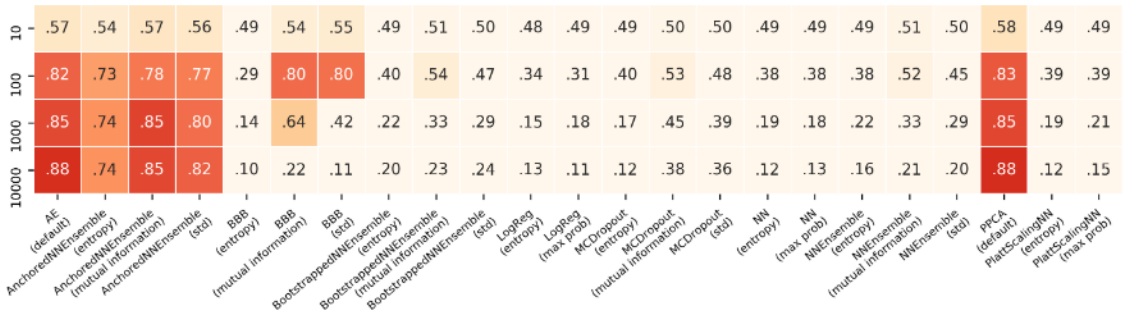

Figure 1: Perturbation experiment results for eICU measured via the OOD detection AUC-ROC. Scales are given on the y-axis, tested models and metrics on the x-axis. Results are averaged over $n = 100$ different, randomly selected perturbed features.

and, surprisingly, the anchored ensemble and BBB.

5.2.2. Clinical OOD Groups

Several articles testing OOD detection capabilities employ images from other data sets as OOD samples (e.g. Pearce et al., 2020 or Kirichenko et al., 2020). In these cases, the OOD property is constituted by a disjunct set of class labels, e.g. when models are trained on MNIST digits but tested on clothing items. We deem this to be an inapplicable methodology for our predictive task, as it is very unlikely that new classes would be introduced in a clinical binary classification problem. We thus use the clinically relevant groups defined earlier, separate them from the training set and treat them as OOD during inference, staying within the same classification framework.

The results in Figure 2 paint a discouraging picture:[5] except for newborns in the MIMIC-III data sets, which have the highest percentage of significantly different features among all OOD groups, all models fail to discern in-distribution from out-of-distribution samples. This even holds for AE and PPCA. The difference between the model families is that density estimation models do seem to struggle for many groups, but do not fail as spectacularly as discriminators (see e.g. the "Elective admissions" group in Figure 2).

5.2.3. Clinical data sets as OOD

Lastly, we treat the whole of eICU as OOD examples for models trained on MIMIC-III and vice versa. This is different from a domain adaptation problem for mortality prediction, where we test the generalization of model performance on data from a different source.[6] We are instead interested to see whether models can identify the data as having a different origin. In the medical domain, this is especially relevant, as e.g. changing hospital protocols could have impact on patients' covariates and therefore influence a model's prediction.

The results in Figure 3 illustrate the intricacies of the problem: again, the density estimation baselines perform best in both cases. By plotting both data sets in the same feature space, it becomes apparent that they are almost completely disjoint, exemplifying a clear case where $p(\boldsymbol{x}) \neq q(\boldsymbol{x})$ (see Figure 6 in the appendix). Furthermore, this fact

5. Results for eICU are displayed in Figure 7 in the appendix, while mortality prediction AUC-ROC scores for the OOD groups are given in Figure 10.

6. These results are given in tab. 5 in the appendix.

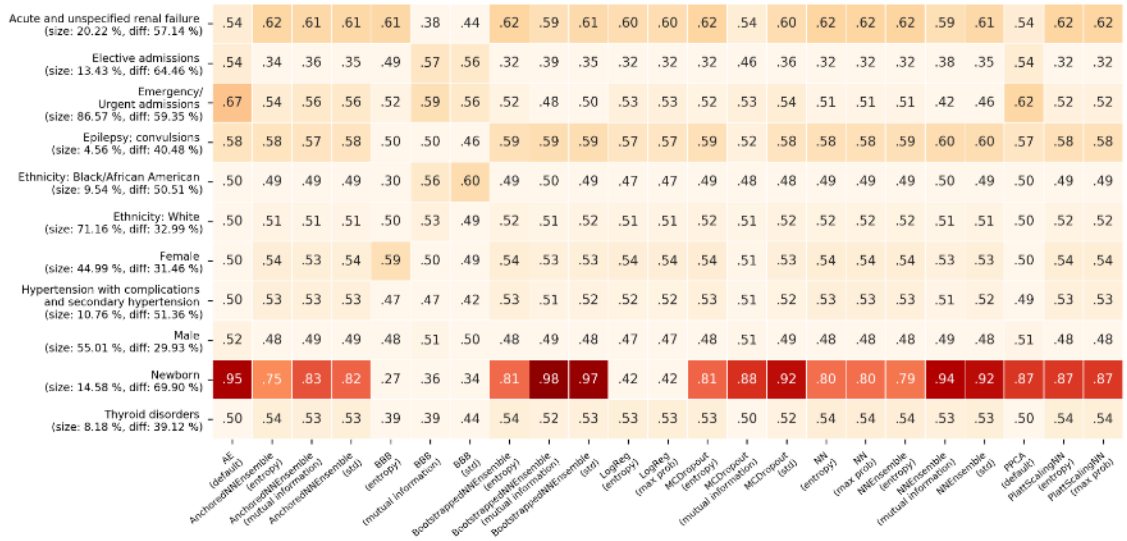

Figure 2: OOD AUC-ROC score for different models and uncertainty metrics for groups in the MIMIC-III data set. *size* denoted the relative size of the OOD group compared to the full training set, *diff* the percentage of OOD features that were different in a statistically significant manner compared to the remaining training set, using a Welch's t-test and $p < 0.01$. Results are averaged over $n = 5$ runs.

also underlines the importance of this issue for medical applications: models trained on either data set might produce unreliable results for patients originating from a different source, whence the need to carefully assess the scaling of medical devices. In our case, classification performance generalizes well across both data sets (see Table 10 in the appendix), however, there are documented cases in the literature where this is not the case (Curth et al., 2019). There are some notable exceptions in the MIMIC-III to eICU case, namely BBB and anchored ensemble , but also some slightly improved scores in the case of the bootstrapped ensemble and the MC Dropout model. Nonetheless, the fact that these scores cannot be reproduced in the reverse case - eICU to MIMIC-III - shows that these results cannot be seen as reliable but rather due to the idiosyncrasies of the training data and contingent model weights.

6. Discussion

For several deep neural discriminators, we have demonstrated that the tested uncertainty estimation methods do not detect OOD samples reliably on medical tabular data. The intuition behind this shortcoming can be explained by the toy example in Figure 4: By separating the feature space into potentially open-ended decision regions, neural discriminators will create large areas with high confidence where no training data were observed, making them prone to failure when presented with OOD samples. Every method aggregating multiple predictions will only enlarge the areas with low confidence due to the overlap of slightly different decision boundaries, but is unlikely to remove this effect entirely and reliably. We did see a few methods stand out in some experiments, but they were not able to do so consistently. In particular, the BBB

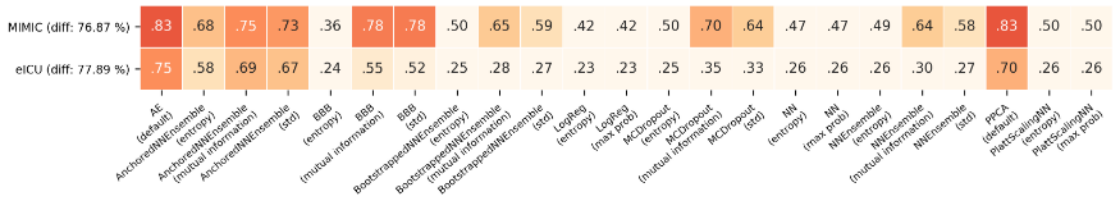

Figure 3: OOD detection AUC-ROC for models trained on eICU and tested on MIMIC-III (first row) and vice versa (second row). Results are averaged over $n = 5$ runs.

and anchored ensemble performed comparatively well in the perturbation experiments.

For the latter model, we conjecture that this is due to the special regularization, namely the l_2 distance to anchor points, which are sampled in the beginning of the training. This seems to incentivize the ensemble members to diversify,[7] an effect that we could also observe when comparing the predictive entropy scores for normal and anchored ensembles on toy data (cp. Figure 9 in the appendix). When different models arrive at diverse solutions, their decision regions have less overlap, creating areas where combined discordant predictions produce higher uncertainty. A similar effect might be at play for BBB, although the diversity that seems to enable the model to identify OODs is more likely due to underfitting.[8]

For simple density-estimation models, many clinically-relevant patient groups also seem to be hard to identify as OOD. This may be because said groups appear to be interspersed with the remaining training population, even if a lot of their feature values differ significantly (cp. Figure 5 in the appendix for the "elective admissions" group in the eICU data set). An explanation could be that the ability to distinguish samples probably does not only depend on the number of differently distributed features, but also on the features themselves, as models are likely learning to prioritize some features over others in terms of importance. More powerful density-estimation models might be able to distinguish these patients; it is an exciting direction for future work to assess whether models such as the one presented in Eduardo et al. (2020), which are tailored towards tabular data, or hybrid models, e.g. in the work

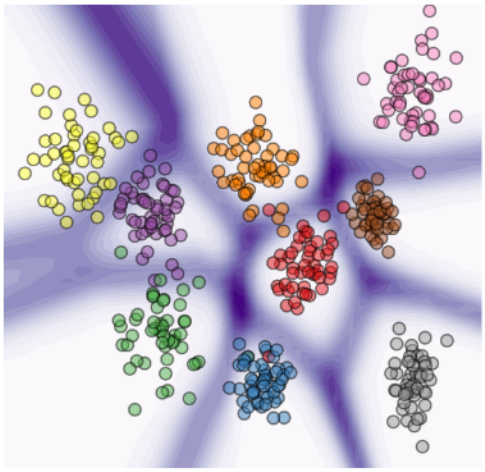

Figure 4: Predictive entropy of a single neural classifier for a toy multi-class classification problem.

7. Such a trade-off between diversity and regularization has also been noted by the original authors in a follow-up work (Pearce et al., 2018).
8. BBB generally struggled to converge during training, even under various hyperparameter configurations (see Section 5 for further discussion).

of Grathwohl et al. (2020), can achieve better OOD detection on such patient groups.

Although the generalization capabilities of deep neural networks have been hailed tremendously in the past - and often rightfully so - they might produce a false sense of security in users for applications with critical outcomes. This is why we advocate to put into context results like those of Ovadia et al. (2019) or Smith and Gal (2018), which suggest that some of these models are relatively robust under covariate shift. Such robustness was only shown in a limited circumstances and does not imply the ability to *detect* covariate shift, as proven in this work. Our results should serve as a cautionary tale, since a strong belief in robustness to OOD coupled with a failure in OOD detection can cause serious problems in practice. We illustrate this with an example. Consider the emergency/urgent admission group in MIMIC-III. The models are barely as good as random guessing when trying to identify the group as OOD. Moreover, the performance on mortality prediction for this group is significantly worse than performance on general population or elective admissions (mortality prediction AUC-ROC of $.71 - .74$, see Figure 10 in the appendix). Thus, in a scenario in which this group is OOD, the models in question would confidently give poor predictions. Such a discrepancy would take a considerable time to detect in practice – several OOD patients would have to be treated before realizing that the performance of the model is declining – time in which the model's mispredictions could already produce harm.

As advocated by Thagaard et al. (2020), a more mindful application of uncertainty estimation techniques is required, especially in safety-critical domains such as health care. This work provides a suite of tests for OOD detection to benchmark uncertainty estimation techniques on two large public medical data sets; more investigation will be necessary to enable safe applications.

References

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

Christopher M Bishop. Bayesian pca. In *Advances in neural information processing systems*, pages 382–388, 1999.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Alicia Curth, Patrick Thoral, Wilco van den Wildenberg, Peter Bijlstra, Daan de Bruin, Paul Elbers, and Mattia Fornasa. Transferring clinical prediction models across hospitals and electronic health record systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 605–621. Springer, 2019.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 204–213, 2020.

Simao Eduardo, Alfredo Nazábal, Christopher KI Williams, and Charles Sutton. Robust variational autoencoders for outlier detection and repair of mixed-type data. In *International Conference on Artificial Intelligence and Statistics*, pages 4056–4066, 2020.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1(3), 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1050–1059, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Carles Gelada and Jacob Buckman. Bayesian neural networks need not concentrate. https://jacobbuckman.com/2020-01-22-bayesian-neural-networks-need-not-concentrate/, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. URL https://openreview.net/forum?id=HkxzxONtDB.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1321–1330, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *arXiv preprint arXiv:1910.09457*, 2019.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. To trust or not to trust A classifier. In *Advances in Neural Information Processing Systems 31:*

Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, pages 5546–5557, 2018. URL http://papers.nips.cc/paper/7798-to-trust-or-not-to-trust-a-classifier.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

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

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. *arXiv preprint arXiv:2006.08545*, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6402–6413, 2017. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

Lotta Meijerink, Giovanni Cinà, and Michele Tonutti. Uncertainty estimation for classification and risk prediction in medical settings. *arXiv preprint arXiv:2004.05824*, 2020.

Jose G Moreno-Torres, Troy Raeder, RocíO Alaiz-RodríGuez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Paul D Myers, Kenney Ng, Kristen Severson, Uri Kartoun, Wangzhi Dai, Wei Huang, Frederick A Anderson, and Collin M Stultz. Identifying unreliable predictions in clinical risk models. *NPJ digital medicine*, 3(1):1–8, 2020.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *CoRR*, abs/1810.09136, 2018. URL http://arxiv.org/abs/1810.09136.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002, 2019.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, pages 3219–3229, 2020. URL http://proceedings.mlr.press/v108/park20b.html.

Tim Pearce, Nicolas Anastassacos, Mohamed Zaki, and Andy Neely. Bayesian inference with anchored ensembles of neural networks, and application to exploration

in reinforcement learning. *arXiv preprint arXiv:1805.11324*, 2018.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately bayesian ensembling. In *International Conference on Artificial Intelligence and Statistics*, pages 234–244, 2020.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5:180178, 2018.

David Ruhe, Giovanni Cina, Michele Tonutti, Daan de Bruin, and Paul Elbers. Bayesian modelling in practice: Using uncertainty to improve trustworthiness in medical applications. *International Conference on Machine Learning, 9-15 June 2019, Long Beach, California, USA, AI for Social Good Workshop*, 2019.

Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset, 2019.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244, 2000.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569, 2018. URL http://auai.org/uai2018/proceedings/papers/207.pdf.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D Hansen, and Anders B Dahl. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *23rd International Conference on Medical Image Computing and Computer Assisted Intervention,*, 2020.

Harri Valpola. From neural pca to deep unsupervised learning. In *Advances in independent component analysis and learning machines*, pages 143–171. Elsevier, 2015.

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *arXiv preprint arXiv:2003.02977*, 2020.

Appendix A. Uncertainty metrics

In the following, we denote a neural network as a function $p_{\boldsymbol{\theta}} : \mathbb{R}^D \to [0, 1]$ parameterized by a parameter vector $\boldsymbol{\theta}$. The baseline for OOD detection introduced by Hendrycks and Gimpel (2017) involves reporting the highest softmax probability across all classes:

$$y_{\max} = \max_{c \in C} p_{\boldsymbol{\theta}}(y = c|\boldsymbol{x}_i)$$

The underlying intuition is that the model would predict more uniform distributions for samples with higher uncertainty, therefore producing a lower $y_{\max}$ score. The uncertainty estimation techniques we consider in this work try to approximate the uncertainty by retrieving a set of K predictions $p^{(k)}_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)$ for a sample $\boldsymbol{x}_i$ from either independently trained models in an ensemble, multiple forward passes with different dropout masks or models with distinct weights sampled from a (variational) posterior. These can then be

aggregated in different manners. One of the simplest ways in a binary classification setting is to compute the standard deviation of the positive class:

$$\sigma_{\mathcal{P}} = \mathbb{E}_{\mathcal{P}}\Big[p_{\boldsymbol{\theta}}(y=1|\boldsymbol{x}_i) - \mathbb{E}_{\mathcal{P}}\big[p_{\boldsymbol{\theta}}(y=1|\boldsymbol{x}_i)\big]\Big]$$

where $\mathbb{E}_{\mathcal{P}}\big[p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)\big]$ is commonly evaluated as a Monte-Carlo estimate using a set of model weights $\mathcal{P} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ s.t. $\frac{1}{K}\sum_{k=1}^{K} p_{\boldsymbol{\theta}}^{(k)}(y|\boldsymbol{x}_i) \approx \mathbb{E}_{\mathcal{P}}\big[p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)\big]$. Thus, a high standard deviation can be interpreted as a large degree of disagreement or uncertainty between classifiers. Another common metric is the predictive entropy (Gal, 2016):

$$\tilde{\mathbb{H}}[p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)] = \mathbb{H}\Big[\mathbb{E}_{\mathcal{P}}\big[p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)\big]\Big]$$

where $\mathbb{H}$ denotes Shannon entropy. As the entropy is computed on the averaged aggregated predictions, the entropy is low when probability mass is distributed uniformly across all classes, implying that all predictions where not able to single out a clear candidate class. Conversely, entropy is high when all classifiers accumulate mass on a single class, showing a low degree of uncertainty about the current sample. Lastly, we also consider mutual information approximated as in Smith and Gal (2018):

$$\mathbb{I}(y, \boldsymbol{\theta}|\boldsymbol{x}_i) \approx \mathbb{H}\Big[\overline{p}_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)\Big] - \mathbb{E}_{\mathcal{P}}\Big[\mathbb{H}\big[p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)\big]\Big]$$

where $\overline{p}_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i) = \mathbb{E}_{\mathcal{P}}\big[p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)\big]$. Intuitively, this is supposed to measure the information gain about the ideal model parameters by receiving a label y. If the possible gain is low, that means that current parameters are close to the optimal ones, demonstrating a low uncertainty. While other metrics capture both aleatoric and epistemic uncertainty of the model, approximate mutual information is exclusively concerned with the latter.

Appendix B. MIMIC-III and eICU Clinical Variables

An overview of the used clinical features and their corresponding names in MIMIC-III and eICU is given in Table 2. In all experiments, only features engineered based on those variables were used.

Appendix C. Hyperparameter Search

For the hyperparameter search, a random search procedure (Bergstra and Bengio, 2012) is employed, sampling hyperparameters for every run from scratch from the options / distributions listed in Table 4. Per data set, we test 40 runs for the autoencoder, vanilla neural network and MC Dropout model, while BBB receives a total of 60 runs each. We also test all configurations for `C` for logistic regression on both data sets. The same hyperparameters found for the neural network model is then also used for its platt-scaling variant and all the ensemble models. The prior scale parameter λ for the anchored ensemble (Pearce et al., 2020) is set to $\sqrt{2/k}$, where k refers to the number of rows of a weight matrix / bias. The number of components for PPCA was set to 15 for both data sets to be comparable with the autoencoder. All models were trained exclusively on CPU.

The best hyperparameters found (excluding the cases above where the same hyperparameters were reused for similar models) are listed in Table 3.

Appendix D. Additional Results

This section contains additional results and plots that were omitted due to spatial constraints.

Description	MIMIC	eICU
Diastolic Blood Pressure	Diastolic blood pressure	Invasive BP Diastolic
Systolic Blood Pressure	Systolic blood pressure	Invasive BP Systolic
Fraction of inspired oxygen	Fraction inspired oxygen	FiO2
Glasgow coma scale (verbal)	Glasgow coma scale verbal response	Verbal
Glasgow coma scale (eyes)	Glasgow coma scale eyes opening	Eyes
Glasgow coma scale (motor functions)	Glasgow coma scale motor response	Motor
Glasgow coma scale (total)	Glasgow coma scale total	GSC Total
Blood glucose level	Glucose	glucose
Heart rate	Heart Rate	Heart Rate
Mean arterial pressure	Mean blood pressure	MAP (mmHg)
Blood oxygen saturation	Oxygen saturation	O2 Saturation
Respiratory rate	Respiratory rate	Respiratory Rate
Body temperature	Temperature	Temperature (C)
Blood pH value	pH	pH

Table 2: Overview of all clinical variables used for the mortality prediction task, including their names in the MIMIC-III and the eICU data set.

Model	Hyperparameter	Value MIMIC-III	Value eICU
AE	hidden_sizes	75	100
AE	latent_dim	15	15
AE	lr	0.006897	0.005216
LogReg	C	10	1000
NN	hidden_sizes	30	75
NN	dropout_rate	0.157483	0.381918
NN	lr	0.000538	0.000904
MCDropout	hidden_sizes	50	50
MCDropout	dropout_rate	0.333312	0.333312
MCDropout	lr	0.000526	0.000526
BBB	hidden_sizes	[25, 25, 25]	[30, 30]
BBB	dropout_rate	0.177533	0.038759
BBB	lr	0.002418	0.002287
BBB	posterior_mu_init	0.22187	0.518821
BBB	posterior_rho_init	−5.982621	−4.475038
BBB	prior_pi	0.233419	0.858602
BBB	prior_sigma_1	0.740818	0.904837
BBB	prior_sigma_2	0.606531	0.67032

Table 3: Best hyperparameters found on the MIMIC-III and eICU data set.

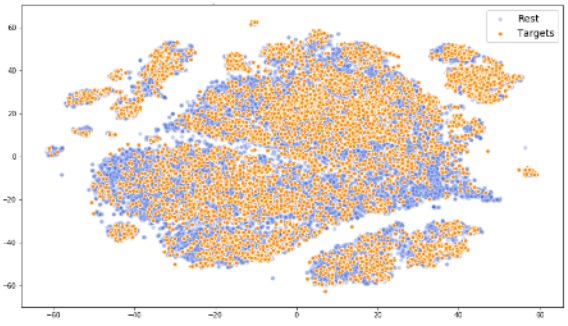

Figure 5: Plot of patients by elective admissions in eICU, projected into a two-dimensional space using t-SNE.

Hyperparameter	Description	Used in	Chosen from
`C`	Inverse l_2 regularization weight	`LogReg`	$\{10, 100, 1000, 10000\}$
`hidden_sizes`	Number / size of hidden layers	All `NN` models, `AE`	1-4 layers of either 25, 30, 50, 75, 100 units
`latent_dim`	Dimensionality of latent space	`AE`	$\{5, 10, 15, 20\}$
`lr`	Learning rate	All models except `PPCA`	$\mathcal{U}(\log(10^{-4}), \log(0.1))$
`dropout_rate`	Dropout rate	All `NN` models	$\mathcal{U}(0, 0.5)$
`posterior_rho_init`	Variance parameter of weight posterior$^{(*)}$	`BBB`	$\mathcal{U}(-8, -2)$
`posterior_mu_init`	Mean parameter of weight posterior$^{(*)}$	`BBB`	$\mathcal{U}(-0.6, 0.6)$
`prior_pi`	Mixture component of prior$^{(*)}$	`BBB`	$\mathcal{U}(\exp(0.1), \exp(0.9))$
`prior_sigma_1`	Variance of prior mixture component 1$^{(*)}$	`BBB`	$\mathcal{U}(\exp(-0.8), \exp(0.1))$
`prior_sigma_2`	Variance of prior mixture component 2$^{(*)}$	`BBB`	$\mathcal{U}(\exp(-0.8), \exp(0.1))$

Table 4: Distributions or options that hyperparameters were sampled from during the random hyperparameter search. $^{(*)}$For more information about these hyperparameters the reader is referred to the work of Blundell et al. (2015).

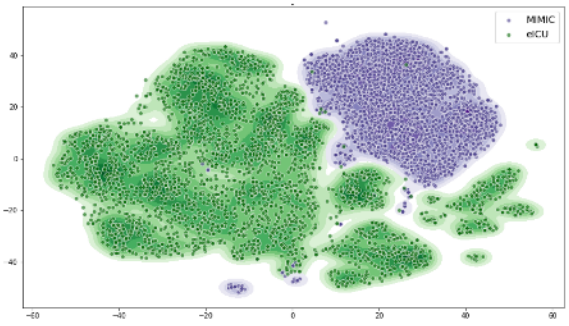

Figure 6: Joint plot of the eICU and MIMIC-III manifolds, projected into a two-dimensional space using t-SNE.

Model	eICU → MIMIC	MIMIC → eICU
`AnchoredNNEnsemble`	0.91 ± 0.00	0.88 ± 0.00
`BBB`	0.66 ± 0.02	0.67 ± 0.03
`BootstrappedNNEnsemble`	0.87 ± 0.00	0.87 ± 0.00
`LogReg`	0.84 ± 0.00	0.86 ± 0.00
`MCDropout`	0.87 ± 0.00	0.87 ± 0.00
`NNEnsemble`	0.87 ± 0.00	0.87 ± 0.00
`NN`	0.87 ± 0.00	0.87 ± 0.00
`PlattScalingNN`	0.87 ± 0.00	0.87 ± 0.00

Table 5: Mortality prediction AUC-ROC on the domain adaptation task. Results are taken over $n = 5$ runs.

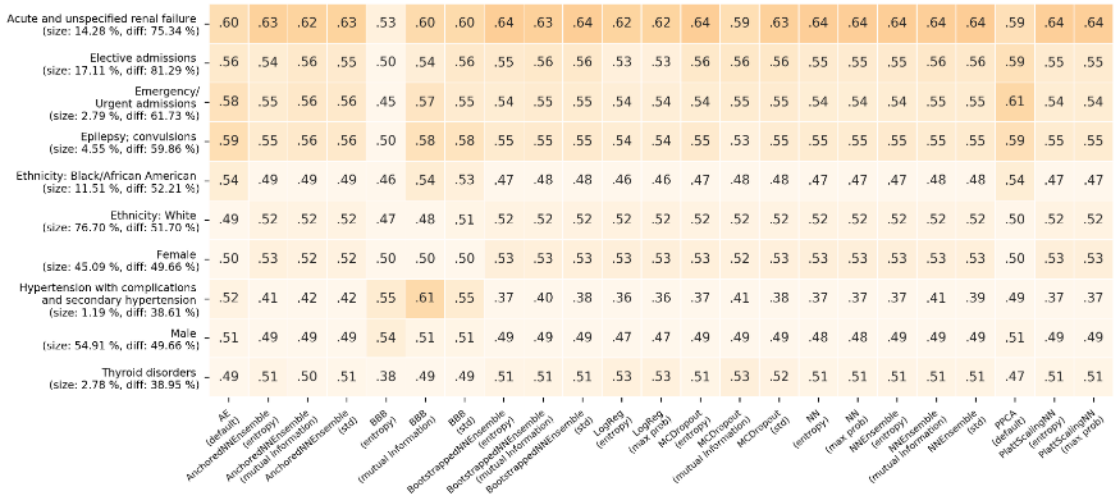

Figure 7: OOD detection AUC-ROC for different models and metrics given pre-defined groups in the eICU data set. *size* denoted the relative size of the OOD group compared to the full training set, *diff* the percentage of OOD features that were statistically significant compared to the remaining training set, using a Welch's t-test and $p < 0.01$. Results are averaged over $n = 5$ runs.

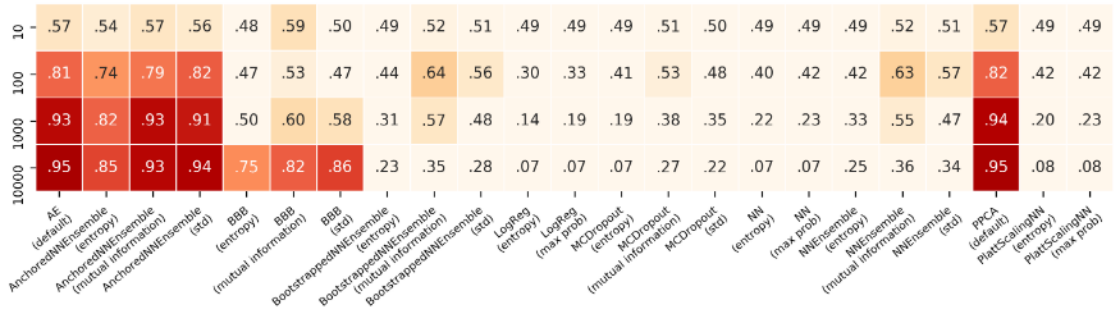

Figure 8: Perturbation experiment results for MIMIC-III measured via the OOD detection AUC-ROC. Scales are given on the y-axis, tested models and metrics on the x-axis. Results are averaged over $n = 100$ different, randomly selected perturbed features.

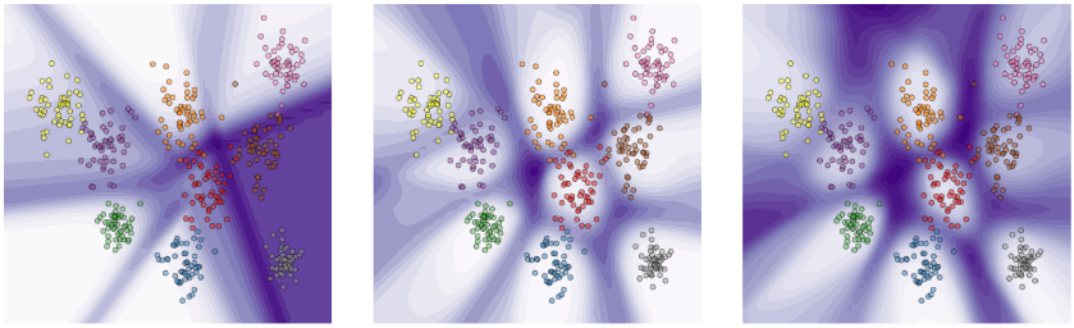

Figure 9: Comparing the predictive entropy scores across a feature space for a multi-class toy example for a single neural discriminator (left), an ensemble of discriminators (center) and an anchored ensemble (right; Pearce et al. (2020)). While the single discriminator produces open regions with low uncertainty even far away from the training data, this effect is somewhat mitigated by ensembling, where overlapping slightly different decision boundaries create regions of higher uncertainty. In the case of the anchored ensemble, this effect is even more pronounced due to the diversity of ensemble members.

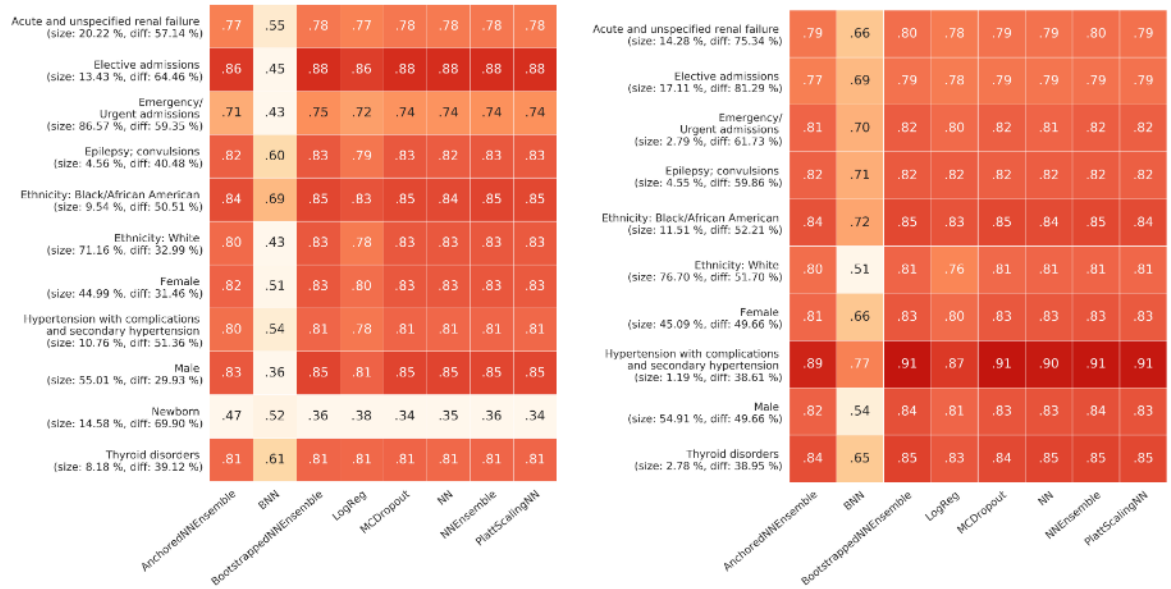

Figure 10: Mortality prediction AUC-ROC scores for different OOD groups of the MIMIC-III (left) and eICU (right) data set. Results are averaged over $n = 5$ runs.