Failures of Uncertainty Estimation on Out-Of-Distribution Samples: Experimental Results from Medical Applications Lead to Theoretical Insights

Karina Zadorozhny [*1] **Dennis Ulmer** [*2] **Giovanni Cinà** [1]

Abstract

Detection of Out-Of-Distribution (OOD) samples in real time is a crucial feature for safety-critical applications of ML models. We extend previous work showing that, on a number of experiments, uncertainty estimation techniques fail at detecting OOD samples on real-world tabular medical data. The experiments described here can serve as benchmark tasks for evaluating OOD detection on medical tabular data. These results suggest that neural discriminators are especially overconfident about their ability to detect OOD samples. Following this intuition we show that, for a class of widely-used network architectures and a list of common uncertainty metrics, neural discriminators generalize their confidence level to previously unseen areas of the feature space, effectively defusing the possibility to detect OOD reliably.

1. Introduction

Neural networks and machine learning models have achieved remarkable performance on a variety of medical tasks ranging from medical imaging to clinical risk assessment (Tang, 2019; Imai et al., 2020). However, deployed models assume that samples are coming from the same distribution as training (that is, in-distribution) data and their performance degrades rapidly when this assumption is violated (Beede, 2020). In this paper we focus on the phenomenon of covariate shift, namely the changes in the feature distributions (Shimodaira, 2000; Moreno-Torres et al., 2012). In medical settings, this could amount to receiving samples from different hospitals, changes in a patient population, observing patients with a previously unseen disease, or receiving corrupted data (Curth et al., 2020; Mårtensson et al., 2020). Therefore, reliable detection of OOD samples in real-time is crucial for real-world applications in high-stakes areas such as healthcare.

The problem is exacerbated by the fact that neural networks (NNs) often make wrong predictions with very high confidence. In particular, neural networks are susceptible to small perturbations of inputs (Goodfellow et al., 2015), the uncertainty of predictions of ReLU networks is often miscalibrated (Guo et al., 2017; Lee et al., 2018), and they produce over-confident predictions on OOD samples (Lakshminarayanan et al., 2017; Hendrycks & Gimpel, 2017; Liang et al., 2018). Despite these challenges, the importance of being able to assess reliability of deployed models has led to the development of new methods for uncertainty quantification and OOD detection.

While there are proposed benchmarks on other data types, (Ovadia et al., 2019; Ren et al., 2019), many of the uncertainty estimation techniques focus mainly on the imaging domain, an area where the OOD detection problem has received a lot of attention in recent years (Lee et al., 2018; Liang et al., 2018; Cao et al., 2020). However, the problem of medical tabular and mixed-type datasets has not been addressed sufficiently.

The key contributions of our work are the following:

- We show that the problem of OOD detection is far from solved in the medical domain: we extend previous work showing that uncertainty estimation techniques fail at detecting OOD samples on real-world tabular medical data.
- We take a cue from the results of perturbation experiments to prove that a class of NN over-generalizes confidence level from training to OOD data, crippling their ability to detect OOD reliably.
- We provide an open-source implementation of the OOD detection models and all the experiments. This can serve as a benchmark for OOD detection on publicly available medical tabular data.

2. Related Work

In order to ensure safe integration into practical applications (Bhatt et al., 2020; D'Amour et al., 2020; Kompa et al.,

[*]Equal contribution [1]Pacmed BV [2]ITU Copenhagen. Correspondence to: Karina Zadorozhny <Karina.Zadorozhny@pacmed.nl>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

2021), recent research efforts yielded a plethora of methods to quantify a neural network's certainty in its prediction, of which we present a selection here. In a classification setting, these include ensembling techniques (Lakshminarayanan et al., 2017; Pearce et al., 2020; Wilson & Izmailov, 2020), approaches based on variational inference (Blundell et al., 2015; Gal & Ghahramani, 2016) or neural parameterizations of Dirichlet distributions (Malinin & Gales, 2018; Joo et al., 2020; Charpentier et al., 2020), hybrid modelling using density estimation (Grathwohl et al., 2020; Mukhoti et al., 2021), modernized RBF networks (Van Amersfoort et al., 2020), and more tractable neural approximations of Gaussian Processes (Liu et al., 2020; Adlam et al., 2020; van Amersfoort et al., 2021).

At the same time, other works have analyzed limitations and failure modes of such techniques. These span from concerns about model calibration (Guo et al., 2017) to missing robustness to distributional shift (Ovadia et al., 2019; Koh et al., 2020) and inadequate uncertainty estimates (Kompa et al., 2020; Kopetzki et al., 2020). Furthermore, it has been shown empirically that the popular method of bootstrapping is not beneficial for neural networks (Nixon et al., 2020) and that even exact Bayesian neural network inference is not robust to distributional shift (Izmailov et al., 2021) and often underperforms compared to point estimates when the posterior is not tempered (Wenzel et al., 2020). Theoretical analyses have further uncovered that neural discriminators yield arbitrarily high-confidence predictions in regions of low-data density (Hein et al., 2019) and that they - in a similarly haphazard manner - generalize uncertainty levels in regions of the feature space without any training data (Ulmer & Cinà, 2021), as we discuss in section 4.2. Mukhoti et al. (2021) show that epistemic and aleatoric uncertainty cannot be disentangled successfully purely based on the output distribution of a single network. Even for models that exclusively model the data density, analyses by Nalisnick et al. (2019) and Lan & Dinh (2020) have cast doubt on their ability to identify OOD inputs.

Unfortunately, most of the aforementioned works evaluate their approaches on the image or text domain. For many practical applications, data is supplied in tabular form containing variables of many different types. This creates unique challenges for which above uncertainty estimation and OOD detection techniques are pathologically understudied. In line with our experiments in section 4.1, we present a number of studies based on such tabular data in a medical context, namely electronic health records (EHRs): Ruhe et al. (2019) present a Bayesian modelling approach in this regard, while Curth et al. (2020) formalize and implement a domain adaptation procedure for EHRs. Dusenberry et al. (2020) test models and their predictive uncertainty using time series data from EHRs, while Ulmer et al. (2020) analyze the behavior of models and their confidence when faced with corrupted feature values and unseen patient groups. Myers et al. (2020) identify groups for which models might underperform. Chan et al. (2020) show that using unlabeled data can improve calibration under distributional shift.

3. Methods

For the OOD detection tasks, we consider several families of models with appropriate uncertainty quantification metrics:

- Logistic Regression (LogReg) baseline with maximum probability and entropy metrics.
- Standard Neural Network (NN) and temperature-scaled NN (PlattScallingNN; Guo et al., 2017) with maximum probability and entropy.
- Ensemble networks. Standard Ensembles (NNEnsemble; Lakshminarayanan et al., 2017), Bootstrapped Ensembles (BootstrappedNNEnsemble), and Anchored Ensembles (AnchoredNNEnsemble; Pearce et al., 2020) with entropy, standard deviation, and mutual information.
- Bayesian Neural Networks. Bayes-by-Backprop (BBB; Blundell et al., 2015) and Monte Carlo Dropout (MC-Dropout; Gal & Ghahramani, 2016) with entropy, standard deviation, and mutual information.
- Neural Gaussian Process. Deep Kernel Learning via Deterministic Uncertainty Estimator (DUE; van Amersfoort et al., 2021) with standard deviation and entropy.
- Density estimators. Autoencoder (AE), beta-Variational Autoencoder (VAE; Higgins et al., 2017) with the reconstruction error metric, and Probabilistic PCA (PPCA) with the log-likelihood metric.
- Clustering-based Local Outlier Factor (LOF; Breunig et al., 2000) with density scores.

Implementation and training details, selected hyperparameters, and performance of the models on a mortality classification task can be found in the Appendix A and in Ulmer et al. (2020).

We stress that the experiments are constructed on the publicly available datasets eICU and MIMIC-III (MIMIC henceforth) and that the implementation of the models, and the code for running the experiments is open source.[1] Therefore, we propose to use these experiments as a new benchmark for OOD detection on medical tabular data.

[1]The code for all the models and the experiments is available at https://github.com/Pacmed/ehr_ood_detection.

4. Results

4.1. Experiments

All the predictor models were trained and optimized for a binary classification task of in-hospital mortality prediction (Appendix B). Given the low rates of in-hospital fatalities, this is a highly unbalanced classification task.

4.1.1. Scaled Features

To analyze whether the current OOD detection approaches can detect scaled samples, we compared popular techniques on a perturbation experiment. Random features in the test set were scaled by a factor of 10, 100, 1000 or 10,000. We report the AUC-ROC score of the detection of the corrupted test samples compared to the original test set (Figure 1 and Appendix C). Many commonly used neural discriminators (with the exception of Anchored Ensembles) coupled with metrics such as maximum probability, entropy, and standard deviation failed to detect the corrupted samples (Figure 1). Moreover, the AUC-ROC score for these models actually decreases below 0.5 as the scale factor increases indicating that the models are assigning very low uncertainty scores to the most corrupt inputs. The density estimators performed better at detecting samples with scaled features. This suggests that the neural discriminators have high certainty in the areas far away from the training distribution. Similar results were obtained on MIMIC (Appendix C).

4.1.2. OOD Groups

To investigate the ability of the models to flag groups of samples that they have not seen before, we separated clinically relevant groups of patients from the training data and presented them to the models during testing. The AUC-ROC scores on the MIMIC dataset are shown in Appendix D. Apart from the most distinct clinical group of newborns, the models consistently failed to detect the OOD groups. Similar under-performance was observed for the eICU dataset (Appendix D). We also investigated whether models trained on one dataset assign higher uncertainty scores to samples coming from a different source. To evaluate this, we trained the models on eICU and tested on the MIMIC data, and vice versa (Appendix E). Again, we observed that neural discriminators, on average, tend to perform worse than density estimators (with the exception of Anchored Ensembles).

4.2. Theoretical results

After illustrating how many uncertainty estimation techniques fail to discern in- from out-of-distribution data in section 4.1, we now turn to finding a theoretical explanation for this behavior, given in detail in Ulmer & Cinà (2021). Earlier work showed how neural discriminators with ReLU activations behave like linear classifiers on polytopal regions of the feature space (Arora et al., 2018; Hein et al., 2019). We then derive how scaling single feature values in the limit lets the softmax probabilities, as well as uncertainty scores produced by a variety of metrics, converge to a fixed point, regardless of the density of the training data. These insights are illustrated in Figure 2 and lead to the following theorem, stated informally here due to spatial constraints:[2]

Theorem 1 (Convergence of uncertainty in the limit) *Given a set of ReLU networks, suppose that their Jacobian matrices with respect to the input do not contain any zero entries. Then, whenever uncertainty is measured via either of the following metrics*

1. *Max. softmax probability (Hendrycks & Gimpel, 2017)*
2. *Class variance (Smith & Gal, 2018)*
3. *Predictive entropy (Gal & Ghahramani, 2016)*
4. *Mutual information (Smith & Gal, 2018)*

the network(s) will converge to fixed uncertainty scores when scaling a feature of an input in the limit.

The mentioned set of ReLU networks can stem from approaches such as ensembling (Lakshminarayanan et al., 2017; as diff. members), MC Dropout (Gal & Ghahramani, 2016; diff. forward passes) or Bayes-by-backprop (Blundell et al., 2015; diff. sampled network parameters), all of which we group under the umbrella of Bayesian model averaging, as argued for by Wilson & Izmailov (2020).

5. Discussion

We refer the reader to Ulmer et al. (2020) and Ulmer & Cinà (2021) for in-depth discussion of the experimental and theoretical findings, and we highlight here what we take to be the most important insights.

First, the task of OOD detection on tabular data still requires scrutiny. Perturbation of a single feature is a rather specific way of constructing OOD, which in reality can come in a variety of novel feature combinations. While some models seem to perform reasonably well in the experiment displayed here, e.g. the density estimation models, results are more mixed in further experiments such as those presented in Ulmer et al. (2020) and Appendix D. For this reason, we stress the importance of a benchmark on public data such as the one presented here. Second, simple neural discriminators seem to be poorly equipped for OOD detection, given their inherent tendency to generalize from seen to unseen data. This points to the fact that different approaches should

[2] A precise and unabridged version and proof is given in Ulmer & Cinà (2021).

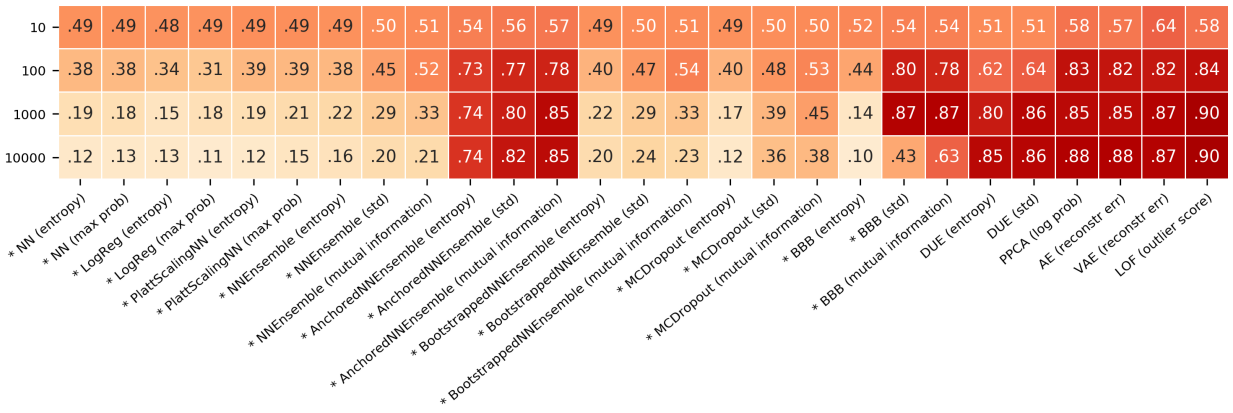
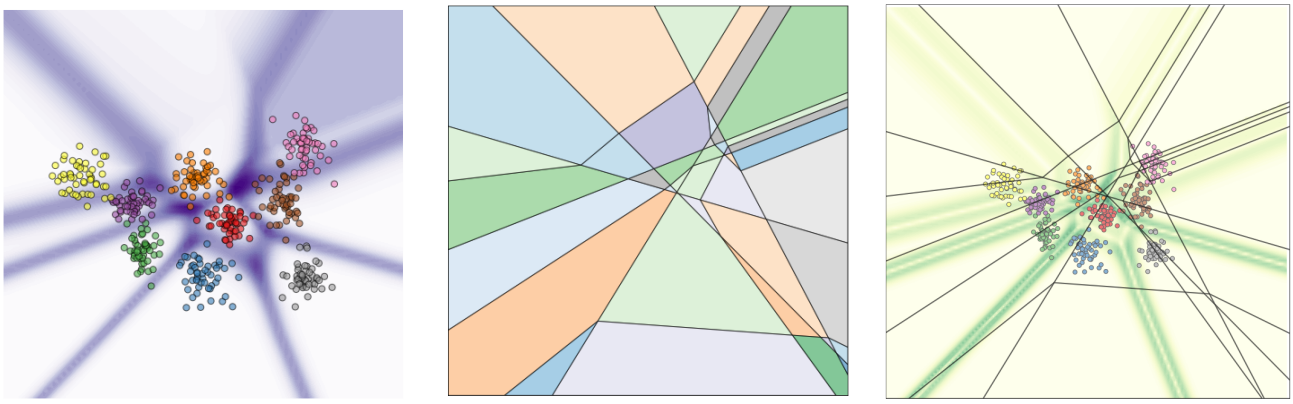


Figure 1. Perturbation experiments on the eICU dataset. AUC-ROC of OOD detection shown for different scaling factors. Results are averaged over n = 100 different, randomly selected perturbed features. The models that fall under the conclusions of the Theorem 1 are marked with an asterisk.



(a) Predictive entropy of ReLU classifier.

(b) Polytopal linear regions induced by same classifier (Arora et al., 2018).

(c) Magnitude of gradient of predictive entropy w.r.t. input.

Figure 2. (a) Uncertainty of a neural classifier with ReLU activations measured by predictive entropy on synthetic data, illustrated by increasing shades of purple with white denoting absolute certainty. (b) Polytopal, linear regions in the feature space induced by the same classifier (as introduced by Arora et al. (2018)). (c) Norm of the gradient of the predictive entropy plotted by increasing shades of green, showing how small perturbations in the input have a decreasing influence on the uncertainty of the network as we stray away from the training data, creating large areas in which uncertainty levels are overgeneralized. Figure taken from Taken from Ulmer & Cinà (2021).

probably get priority. We should add that density estimation techniques also come with some shortcomings – scaling is a well-known problem of some models in this family – and their efficacy has been debated (Nalisnick et al., 2019). The combination of neural discriminators with other techniques (as in van Amersfoort et al., 2021) could be a promising way forward.

Third, it is worth noting that ensembling may work locally: we see in the experiments that anchored ensembles (in which ensemble members retain a certain level of diversification among them) seem to hold their ground. However, Theorem 1 and other experiments lead us to believe that in general the more the ensemble members coincide the more they misbehave with respect to OOD, so there is a trade-off between OOD detection (where increased disagreement is beneficial) and performance (for which less disagreement is beneficial). Fourth, the small print of Theorem 1, namely the conditions under which the theorem is applicable, could serve as a point of reflection on what sort of generalization behaviour must be prevented to enable OOD detection. It is also still an open question whether the theoretical results can be extended to a wider class of networks, a wider class of uncertainty metrics or categorical features.

Finally, we remark that this line of work, albeit targeted to medical application, does concern all kind of applications revolving around tabular data and (unbalanced) classification tasks, such as fraud detection from financial data or prediction of mortgage default.

References

Adlam, B., Lee, J., Xiao, L., Pennington, J., and Snoek, J. Exploring the Uncertainty Properties of Neural Networks' Implicit Priors in the Infinite-Width Limit. *arXiv preprint arXiv:2010.07355*, 2020.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding Deep Neural Networks with Rectified Linear Units. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Beede, E. Healthcare AI Systems that put People at the Center, 2020.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., et al. Uncertainty as a Form of Transparency: Measuring, Communicating, and using Uncertainty. *arXiv preprint arXiv:2011.07586*, 2020.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight Uncertainty in Neural Networks. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622, 2015.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. LOF: Identifying Density-based Local Outliers. *ACM SIGMOD Record*, 29:93–104, 2000.

Cao, T., Huang, C.-W., Hui, D. Y.-T., and Cohen, J. P. A Benchmark of Medical Out-of-Distribution Detection. *arXiv:2007.04250 [cs, stat]*, 2020.

Chan, A., Alaa, A., Qian, Z., and Van Der Schaar, M. Unlabelled Data Improves Bayesian Uncertainty Calibration under Covariate Shift. In *International Conference on Machine Learning*, pp. 1392–1402. PMLR, 2020.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty Estimation without OOD samples via Density-Based Pseudo-Counts. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Curth, A., Thoral, P., van den Wildenberg, W., Bijlstra, P., de Bruin, D., Elbers, P., and Fornasa, M. Transferring Clinical Prediction Models Across Hospitals and Electronic Health Record Systems. Communications in Computer and Information Science, pp. 605–621, Cham, 2020. Springer International Publishing.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. Underspecification presents challenges for credibility in modern machine learning, 2020.

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., and Dai, A. M. Analyzing the Role of Model Uncertainty for Electronic Health Records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 204–213, 2020.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing Model Uncertainty in Deep Learning. In *International conference on Machine Learning*, pp. 1050–1059, 2016.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

Grathwohl, W., Wang, K., Jacobsen, J., Duvenaud, D., Norouzi, M., and Swersky, K. 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.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On Calibration of Modern Neural Networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why ReLU Networks yield High-confidence Predictions far away from the Training Data and how to Mitigate the Problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Hendrycks, D. and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in neural networks. 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France,*

April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. Improving Explorability in Variational Inference with Annealed Variational Objectives. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Imai, S., Takekuma, Y., Kashiwagi, H., Miyai, T., Kobayashi, M., Iseki, K., and Sugawara, M. Validation of the Usefulness of Artificial Neural Networks for Risk Prediction of Adverse Drug Reactions used for Individual Patients in Clinical Practice. *PloS One*, 15, 2020.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. What are Bayesian Neural Network Posteriors really like? *arXiv preprint arXiv:2104.14421*, 2021.

Joo, T., Chung, U., and Seo, M. Being Bayesian about Categorical Probability. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4950–4961. PMLR, 2020.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A Benchmark of in-the-wild Distribution Shifts. *arXiv preprint arXiv:2012.07421*, 2020.

Kompa, B., Snoek, J., and Beam, A. Empirical Frequentist Coverage of Deep Learning Uncertainty Quantification Procedures. *arXiv preprint arXiv:2010.03039*, 2020.

Kompa, B., Snoek, J., and Beam, A. L. Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning. *NPJ Digital Medicine*, 4(1):1–6, 2021.

Kopetzki, A.-K., Charpentier, B., Zügner, D., Giri, S., and Günnemann, S. Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable? *arXiv preprint arXiv:2010.14986*, 2020.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

Lan, C. L. and Dinh, L. Perfect Density Models cannot Guarantee Anomaly Detection. *arXiv preprint arXiv:2012.03808*, 2020.

Lee, K., Lee, K., Lee, H., and Shin, J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*, 2018.

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Malinin, A. and Gales, M. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A Unifying View on Dataset Shift in Classification. *Pattern recognition*, 45 (1):521–530, 2012.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. Deterministic Neural Networks with Appropriate Inductive Biases Capture Epistemic and Aleatoric Uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

Myers, P. D., Ng, K., Severson, K., Kartoun, U., Dai, W., Huang, W., Anderson, F. A., and Stultz, C. M. Identifying Unreliable Predictions in Clinical Risk Models. *NPJ digital medicine*, 3(1):1–8, 2020.

Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M. G., Taylor, J.-P., Hort, J., Snædal, J., Kulisevsky, J., Blanc, F., Antonini, A., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A., Aarsland, D., and Westman, E. The Reliability of a Deep Learning Model in Clinical Out-Of-Distribution MRI Data: A Mlticohort Study. *Medical Image Analysis*, 66:101714, 2020.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. Do Deep Generative Models Know What They Don't Know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Nixon, J., Lakshminarayanan, B., and Tran, D. Why are Bootstrapped Deep Ensembles not Better? In *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*, 2020.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you Trust your Model's Uncertainty?

Evaluating Predictive Uncertainty under Dataset Shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Pearce, T., Leibfried, F., and Brintrup, A. Uncertainty in Neural Networks: Approximately Bayesian Ensembling. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 234–244, 2020.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood Ratios for Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ruhe, D., Cina, G., Tonutti, M., de Bruin, D., and Elbers, P. 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.

Shimodaira, H. Improving Predictive Inference under Covariate Shift by Weighting the Log-likelihood Function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Smith, L. and Gal, Y. 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*, pp. 560–569, 2018.

Tang, X. The Role of Artificial Intelligence in Medical Imaging Research. *BJR Open*, 2(1), 2019. ISSN 2513-9878.

Ulmer, D. and Cinà, G. Know Your Limits: Uncertainty Estimation with ReLU Classifiers Fails at Reliable OOD Detection. *arXiv:2012.05329*, 2021.

Ulmer, D., Meijerink, L., and Cinà, G. Trust Issues: Uncertainty Estimation does not Enable Reliable OOD Detection on Medical Tabular Data. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136, pp. 341–354, 2020.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In *International Conference on Machine Learning*, pp. 9690–9700. PMLR, 2020.

van Amersfoort, J., Smith, L., Jesson, A., Key, O., and Gal, Y. Improving Deterministic Uncertainty Estimation in Deep Learning for Classification and Regression, 2021.

Wenzel, F., Roth, K., Veeling, B. S., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How Good is the Bayes Posterior in Deep Neural Networks Really? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10248–10259. PMLR, 2020.

Wilson, A. G. and Izmailov, P. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Appendix

A. Extended Methods

The hyperparameters for all models were found using an extensive hyperparameter search on both the eICU and the MIMIC datasets. Hyperparameters used for NN, temperature-scaled NN, Ensembles, MCDropout, BBB, PPCA, and AE can be found in Ulmer et al. (2020).

For the DUE model, we followed the implementation described in van Amersfoort et al. (2021). The final hyperparameters used for eICU were the following: the Matern 1/2 kernel function, 50 inducing points, the distance-preserving neural network contained 4 layers, each with 256 units, the Lipschitz coefficient was set to 0.5, and the learning rate was 0.002. For the MIMIC dataset, the number of inducing points was set to 20 and the learning rate to 0.004. Other hyperparameters were the same as for eICU.

The beta-VAE models was implemented according to Higgins et al. (2017). An option to perform beta-annealing (a deterministic warm-up; Huang et al., 2018) was added as a hyperparameter. The hyperparameters for the eICU data were the following: both the encoder and the decoder had one layer with 100 units, 20 latent dimensions, the learning rate was 0.003, and the value of beta was set to 1.8 with annealing. For the MIMIC dataset, the encoder and the decoder contained three layers with 50 units each an 5 latent dimensions. The beta-VAE was trained with a learning rate of 0.005 and a beta-value of 1.6 with annealing.

B. Mortality Prediction

We trained the predictor models on a binary mortality classification task on both the eICU and the MIMIC dataset. The final AUC-ROC scores for the models are shown in Table 1.

Both datasets have unbalanced classes. This is due to the intrinsic nature of the problem — the in-hospital mortality rates are low. The percentage of positive labels in the training data for the eICU and the MIMIC datasets are 12.5% and 13.5% respectively.

C. Perturbation Experiments on MIMIC Dataset

The ability of models to detect scaled inputs was tested on the eICU dataset (Section 4.1.1) and the MIMIC dataset (Figure 3). Similar to the results in the main section, neural discriminators generally perform worse at flagging the scaled inputs and their performance tends to decrease with greater scaling.

Table 1. AUC-ROC score for the mortality prediction task on the eICU and the MIMIC datases. Results were averaged over 5 runs and displayed with standard deviation.

	eICU	MIMIC
AnchoredNNEnsemble	0.832 ± 0.004	0.839 ± 0.006
BootstrappedNNEnsemble	0.847 ± 0.000	0.848 ± 0.001
DUE	0.828 ± 0.003	0.838 ± 0.002
LogReg	0.823 ± 0.000	0.834 ± 0.000
MCDropout	0.844 ± 0.001	0.847 ± 0.002
NNEnsemble	0.847 ± 0.000	0.848 ± 0.001
NN	0.842 ± 0.002	0.847 ± 0.003
PlattScalingNN	0.844 ± 0.002	0.845 ± 0.002

D. Clinical OOD Groups

As outlined in the section 4.1.2, we selected clinically relevant groups of patients and withheld these groups during training. The models then scored the groups at test-time.

To compare a feature-wise difference of the groups and the training data, we performed a feature-wise Welch's t-test (with $p < 0.01$). The percentages shown in the plots (Figures 4 and 5) indicate how many features were significantly different from the features of the training set.

Note that the problem is different from the predictive performance of the models on new data given that generalization of models is desirable if the input is sufficiently similar. What is being tested is the ability of models to detect new samples.

E. Clinical Datasets as OOD

We investigated whether the models are able to recognize samples with a different origin as described in the section 4.1.2. To this end, the models were trained on either the eICU dataset or the MIMIC dataset and during inference, were presented with samples from the other dataset. The ability to flag these samples as novel is show in Figure 6.

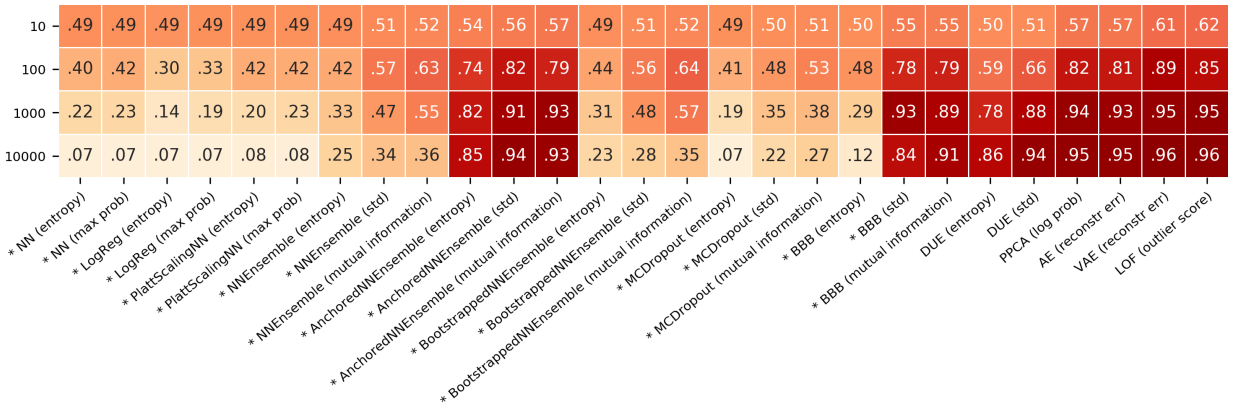


Figure 3. Perturbation experiments on the MIMIC dataset. AUC-ROC of OOD detection shown for different scaling factors.Results are averaged over n = 100 different, randomly selected perturbed features. The models that fall under the conclusion of the Theorem 1 are marked with an asterisk.

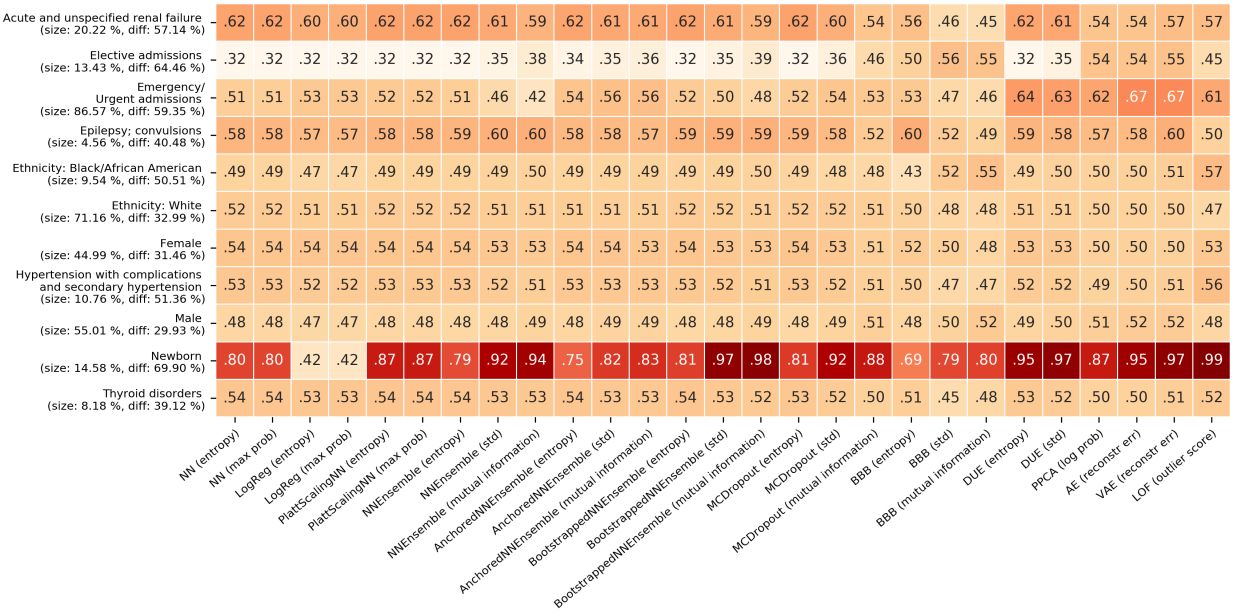


Figure 4. OOD groups for the MIMIC dataset. The AUC-ROC scores of detecting the indicated OOD groups at test time. For each group, its size compared to the training data is shown (*size*) along with a percentage of features that are significantly different from the training set (*diff*). Average results over 5 runs are shown.

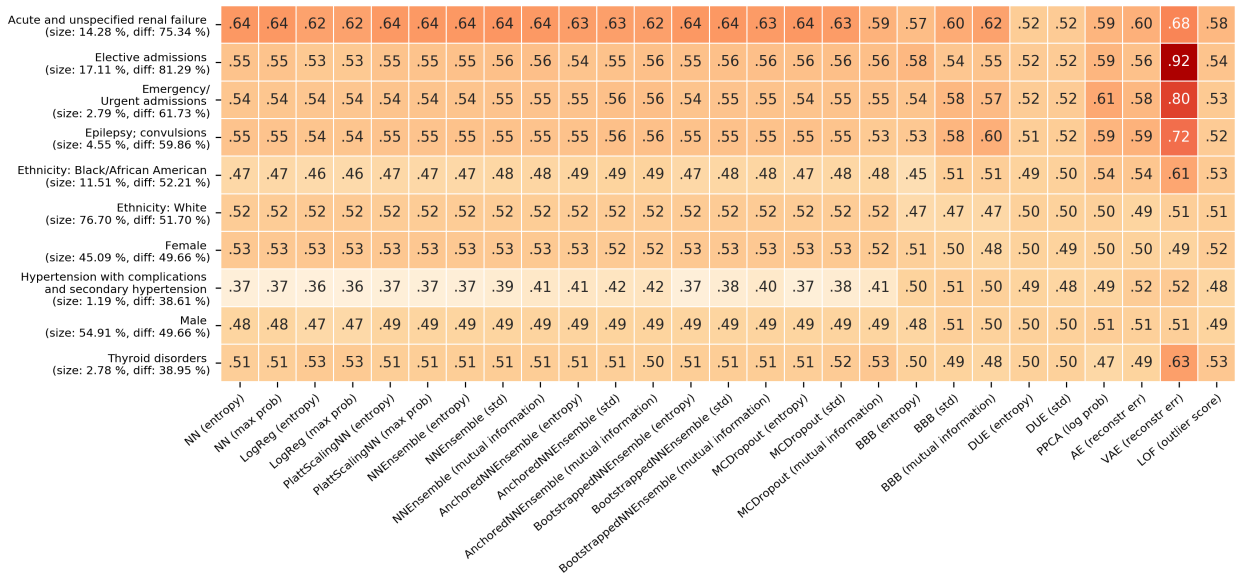


Figure 5. OOD groups for the eICU dataset. The AUC-ROC scores of detecting the indicated OOD groups at test time. For each group, its size compared to the training data is shown (*size*) along with a percentage of features that are significantly different from the training set (*diff*). Average results over 5 runs are shown.

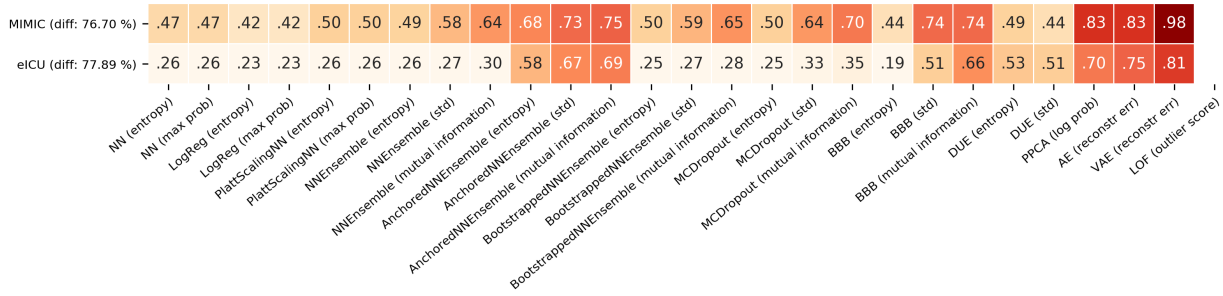


Figure 6. The AUC-ROC scores of detecting new clinical datasets as OOD. The models were trained on the eICU dataset and presented with MIMIC samples (top row), or visa versa (bottom row). The percentage of features that are significantly different from the training set is added (*diff*). Average results over 5 runs are shown.