Bayesian Modelling in Practice: Using Uncertainty to Improve Trustworthiness in Medical Applications

David Ruhe [1][2] **Giovanni Cinà** [2] **Michele Tonutti** [2] **Daan de Bruin** [2] **Paul Elbers** [3]

Abstract

The Intensive Care Unit (ICU) is a hospital department where machine learning has the potential to provide valuable assistance in clinical decision making. However, classical machine learning models usually only provide point-estimates and no uncertainty of predictions. In practice, uncertain predictions should be presented to doctors with extra care in order to prevent potentially catastrophic treatment decisions. In this work we show how Bayesian modelling and the predictive uncertainty that it provides can be used to mitigate risk of misguided prediction and to detect out-of-domain examples in a medical setting. We derive analytically a bound on the prediction loss with respect to predictive uncertainty. The bound shows that uncertainty can mitigate loss. Furthermore, we apply a Bayesian Neural Network to the MIMIC-III dataset, predicting risk of mortality of ICU patients. Our empirical results show that uncertainty can indeed prevent potential errors and reliably identifies out-of-domain patients. These results suggest that Bayesian predictive uncertainty can greatly improve trustworthiness of machine learning models in high-risk settings such as the ICU.

1. Introduction

The Intensive Care Unit (ICU) is a resource-intensive environment where patients receive care that is not readily available elsewhere in the hospital. ICU patients usually are in life-threatening conditions. Therefore, adequate assessment of illness severity and expected effectiveness of interventions is of utmost importance for clinical decision making. Recent advancements in machine learning have cleared the path to deployment of machine learning based decision support software systems in critical areas like the ICU. Examples are predicting risk of readmission after discharge (Thoral et al., 2018), optimizing sepsis treatment (Raghu et al., 2017), or predicting mortality risk given the current state of the patient (Pirracchio et al., 2015). While these applications have proven to be effective, model trustworthiness remains elusive. Most machine learning models only provide point estimates of their parameters and corresponding predictions. A recent study has shown that these models can output high-risk predictions on datapoints that lie far from their observed dataset (Nguyen et al., 2015). Especially in a critical area like the ICU these flaws could result in catastrophes. For effective machine learning based decision support, we think it is crucial that machine learning methods output uncertainty estimates beside predictions. Bayesian modelling has the desirable property of expressing predictive uncertainty through stochasticity in its parameters. In this work, we show how *Bayesian Neural Networks*, through predictive uncertainty (hereafter referred to as uncertainty), can correctly identify predictions that are likely to be misguided. When the model encounters a datapoint that lies far from its observed set, a practitioner can be notified through the model's uncertainty and observe that the patient has a combination of symptoms that hitherto has not been seen. Therefore, the practitioner should proceed with care.

In previous work, Leibig et al. (2017) are among the first to apply model uncertainty to healthcare by diagnosing diabetic retinopathy from fundus images. Similarly, Nair et al. (2018); Wang et al. (2019); Orlando et al. (2019) also estimate uncertainty to image analysis using MC Dropout (Gal, 2016). We extend on previous work by making the following constributions.

1. We provide mathematical bounds on the obtainable loss with respect to uncertainty.

2. Through these bounds, model performance is directly related to uncertainty. Uncertainty prevents high-loss prediction errors.

3. We show that BNNs can identify out-of-domain patients competently in a real use-case.

[1]University of Amsterdam [2]Pacmed BV, Amsterdam [3]Univeristy Medical Center, Amsterdam. Correspondence to: David Ruhe <david.ruhe@gmail.com>.

Additionally, we are (to our best knowledge) novel in the approach of using Bayes By Backprop (Blundell et al., 2015) instead of MC Dropout to estimate model distribution parameters directly. This choice has two motivations. First, MC Dropout rates have to be carefully adjusted to obtain well-calibrated uncertainties. This requires tuning all dropout probabilities, which is unfeasible for deep neural networks. Second, the MC dropout approximate posterior does not contract with more data, and therefore the approach has been questioned (Osband, 2016). We also extend on earlier publications by applying Bayesian uncertainty to (ICU) signal processing.

The rest of the paper is structured as follows. In section 2 we discourse required background knowledge. In section 3 we elaborate on methodological decisions for signal processing and modelling. In section 4 we provide an overview of the results, showing that model uncertainty can mitigate the prediction loss and how uncertainty relates to out-of-domain observations. In section 5 we conclude and provide suggestions for future research directions.

2. Background

In this section we summarize important background theory about Bayesian modelling.

2.1. Bayesian Neural Networks

In deterministic modelling, given a dataset of observations $X = \{x_1, \ldots, x_N\}$ and labels $y = \{y_1, \ldots, y_N\}$, we restrict model parameters ω to a point estimate and optimize the likelihood function $p(y|X, \omega)$ directly. That is, we try to find

$$w^* = \arg\max_{\omega} P(y|X, \omega). \tag{1}$$

To capture model uncertainty, we strive to find the posterior distribution over the model parameters:

$$P(\omega|X, y) = \frac{P(y|X, \omega)P(\omega)}{P(y|X)}, \tag{2}$$

where $P(\omega)$ is a prior. The marginal likelihood $P(y|X)$ is computed by marginalisation: $\int_{\omega} P(y|X, \omega)P(\omega)$. Given many continuous model parameters this term cannot be evaluated analytically, which is the case when ω are the parameters of a (deep) neural network. Consequently, the posterior is intractable as well. Therefore, we approximate it with a tractable variational distribution $q_{\theta}(\omega)$ through a procedure called *variational inference* (Hinton & van Camp, 1993). The goal is to minimize the KL divergence between the true and approximate posterior:

$$\theta^* = \arg\min_{\theta} \mathrm{KL}\left(q_{\theta}(\omega)||P(\omega|X, y)\right). \tag{3}$$

This KL divergence, too, is analytically intractable. Instead, we can minimize it by maximizing the *evidence lower bound*:

$$\mathcal{L}_{VI}(\theta) := \int_{\omega} q_{\theta}(\omega) \log P(y|X, \omega) - \mathrm{KL}(q_{\theta}(\omega)||P(\omega)) \tag{4}$$

$$= \log P(y|X) - \mathrm{KL}\left(q_{\theta}(\omega)||P(\omega|X, y)\right)$$

Since the KL-divergence is non-negative, the evidence lower bound floors the marginal likelihood. Thus, we are able to minimize the gap between the true and approximate posterior.

In this paper, we used the approach of Blundell et al. (2015), coined Bayes By Backprop (BBB), to maximize the evidence lower bound. Predictions for a data point x_i were made by sampling model parameters from the variational posterior distribution $q_{\theta}(\omega)$ and taking the mean of the predictions of all sampled models. That is:

$$P(y_i|x_i) = \frac{1}{T}\sum_{t=1}^{T} P(y_{it}|f^{\hat{\omega}_t}(x_i)) \tag{5}$$

with $\hat{\omega}_t \sim q_{\theta}^*(\omega)$. Predictive uncertainty was computed as the variance in the predictions:

$$Var(y_i|x_i) = \frac{1}{T}\sum_{t=1}^{T}(P(y_i|x_i) - P(y_i|f^{\hat{\omega}_t}(x_i))^2 \tag{6}$$

3. Methods

3.1. Data and Preprocessing

Data was obtained from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database (Johnson et al., 2016). This dataset contains signal data for 46,520 patients and 58,976 ICU admissions. $7,821$ of the patients in the MIMIC-III dataset are newborns; these were excluded from the dataset as newborns come with different characteristics and treatment requirements. For each patient, features were constructed by aggregating relevant clinical information, lab values and vital signs. The objective of the model was to find patients that have a high risk of mortality during the admission. More information on the task and data processing is included in the supplementary material.

3.2. Model Architecture

Following Blundell et al. (2015), we used a Gaussian scale mixture as our prior. We used two 128-neuron hidden layers, ReLU intermediate activation and Sigmoid final activation to obtain the probability distribution $P(y_i|f^{\omega}(x_i))$. All weight distributions were initialized with $\mu \sim \mathcal{U}(-.2, .2)$ and $\rho \sim \mathcal{U}(-5, -4)$. Adam was used as the optimizer with the configurations as suggested by the authors (Kingma & Ba, 2014).[1]

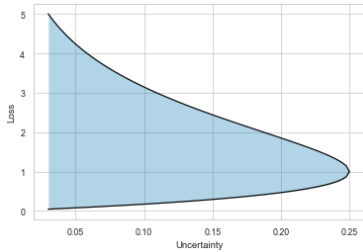[1]Code is available at https://github.com/Pacmed/aisg_2019

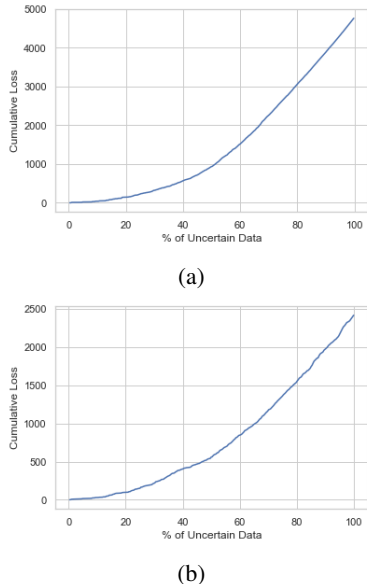Figure 1. The reachable loss region in a Bayesian binary classification problem.

(a)

(b)

Figure 2. Cumulative prediction loss (y-axis) related to the amount of uncertain testing data included in the analysis (x-axis). The uncertainty effectively mitigates prediction loss for both the BNN (a) and the Gradient Boosting model (b).

4. Results

We discuss our results in the following two subsections.

4.1. Uncertainty Mitigates Prediction Loss

In the supplementary material we derive the upper and lower bounds of the binary cross-entropy loss $\mathcal{L}(y_i|\boldsymbol{x_i})$ with respect to uncertainty:

$$-\log\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\cdot Var(y_i|\boldsymbol{x_i})}\right)$$
$$\leq \mathcal{L}(y_i|\boldsymbol{x_i}) \leq -\log\left(\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\cdot Var(y_i|\boldsymbol{x_i})}\right). \tag{7}$$

These are depicted in figure 1. In the supplementary material we show how the data follows these bounds. Areas of both low and high loss cannot be reached with high un-

certainty. In other words, when the variational posterior distribution is a good approximation of the true posterior, far-from-domain examples that cause high uncertainty will mathematically drive the loss away from the extreme values. From this line of thought, we hypothesize that uncertainty is able to mitigate prediction loss. Deterministic models can output very confident predictions on inputs that lie far from their training domain, yielding high performance loss. From equation 7 we see that this is not possible for a BNN. Additionally, the cause of a highly certain wrong prediction may be due to errors in the data, for example due to mislabelled observations. If the variational posteriors is *not* a good approximation of the true posterior, low uncertainty can still be achieved on wrong predictions, leading to bad consequences. Therefore, we evaluate empirically.

Figure 2 illustrates how uncertainty relates to predictive performance. When the observations $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$ are sorted according to their predictive uncertainty, it can be seen that the loss increases superlinearly for each additional data point included. That is, the most uncertain 20% of the data contribute about 20 times the loss the most certain 20% do. In the supplementary material, we show that for the 20% most certain data the area under the receiver operating characteristic curve (AUROC) approaches unity. Thus, low-uncertainty patients are more often classified correctly. Therefore, by restricting classification to certain data points, we can effectively employ predictive uncertainty to calibrate the performance of the model.

A peculiar finding is depicted in figure 2b. We observe that uncertainty generalizes to a gradient boosting decision tree model, trained on the same data. In the case of tabular data like MIMIC, this result could prove to be particularly interesting, as neural networks are often outperformed by other non-linear models in similar tasks (Fernández-Delgado et al., 2014). This observation could mean that a Bayesian model can be deployed for the sole purpose of its uncertainty output, leaving the classification task to a second higher-performing model. The total obtained cumulative loss is lower for the tree-based model, meaning that it had more confident correct predictions. This is likely due to the fact that the prediction risk for the decision tree is not mathematically bound to uncertainty. Note that we have not investigated when this does or does not hold. This remains worthy of further investigation.

4.2. Detecting Out-of-Domain Patients

Plotting the relationship between predictions and uncertainty results in the moon-shaped scatter plot depicted in figure 3a. This corresponds to the intuition given earlier, showing that extreme predictions are only made on observations with low uncertainty. In the central region of the graph, we see that the uncertainty has a wider spread, meaning that there are
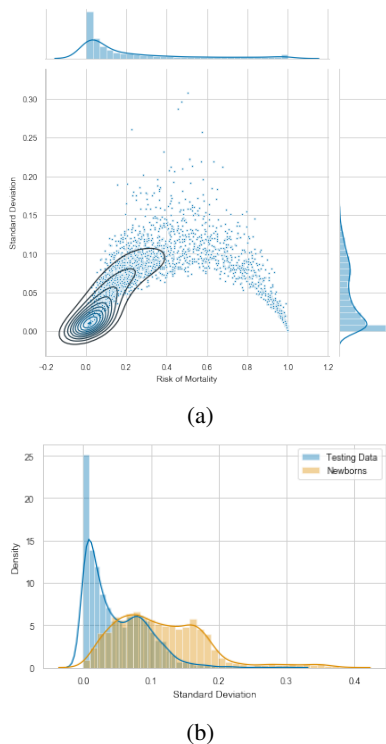
(a)

(b)

Figure 3. In plot 3a we depict the predictive uncertainty (x-axis) related to predictive uncertainty (y-axis). In plot 3b we see how the BNN effectively identifies out-of-domain examples.

both observations with feature values occurring frequently (low uncertainty) and patients with sets of features that lie further from the previously observed domain. Most of the mass expectedly concentrates around the surviving group of patients.

In the medical field, research is often conducted on a biased sample of the population of patients and therefore minorities are underrepresented (Baird, 1999; Swanson & Ward, 1995; Giuliano et al., 2000; Bonevski et al., 2014). Regrettably, models and treatments are still applied to the true population, including these minorities, which can lead to adverse results. In the MIMIC dataset, there is a group of $7,821$ newborns. These were set aside as out-of-domain patients, and excluded from training. We replaced obvious differences like weight and age with training data averages. We observe in figure 3b that when the model is presented newborns, it becomes about 230% more uncertain on average. This means that a high-uncertainty output can be used to warn a practitioner that the (combination of) symptoms and vital signs have rarely been observed before, and he or she should proceed with care. A plot for the newborns similar to figure 3a is given in the supplementary material. Since the characteristics of newborns greatly differ from the rest of the population, we also experimented with uncertainty on ethnic minorities. Among others, Carson et al. (1999) inves-

tigated the differences in responses to heart failure therapies between ethnicities. African American people did show different responses as a result of different features compared to the baseline. This motivated us to investigate model uncertainty on such an ethnic minority. After setting it apart, we observed that the BNN became 130% more uncertain ($p < .0001$) on this group. Comparison to a deterministic baseline on these tasks is given in the supplementary material.

5. Conclusions

In this paper, we explored the application of Bayesian Neural Networks to improve the safety of machine-learning-based clinical decision support tools in critical areas such as the ICU. Following the findings of Blundell et al. (2015), we trained a BNN on the MIMIC critical care dataset. The objective of the model was to predict patient mortality given clinical observations and lab values. We derived bounds on cross-entropy loss with respect to predictive uncertainty. Through these bounds, uncertainty is able to mitigate performance risk and loss. Empirically, we showed that the loss of test set predictions increased superlinearly when the patients were sorted according to their corresponding uncertainty. Secondly, the results reveal that the uncertainty of the predictions increases significantly on out-of-domain patients. This suggests that in an applied setting a BNN can effectively identify patients outside of its previously observed domain. Overall, this work demonstrates that uncertainty is effective in enhancing model trustworthiness and mitigating prediction risk and loss in an critical setting like the ICU. A mathematical intuition as to why uncertainty generalizes to other models remains unclear. This can be an interesting direction for future research.

References

Baird, K. L. The new nih and fda medical research policies: targeting gender, promoting justice. *Journal of Health Politics, Policy and Law*, 24(3):531–565, 1999.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., Brozek, I., and Hughes, C. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC medical research methodology*, 14(1):42, 2014.

Carson, P., Ziesche, S., Johnson, G., Cohn, J. N., Group, V.-H. F. T. S., et al. Racial differences in response to therapy for heart failure: analysis of the vasodilator-heart

failure trials. *Journal of cardiac failure*, 5(3):178–187, 1999.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Gal, Y. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

Giuliano, A. R., Mokuau, N., Hughes, C., Tortolero-Luna, G., Risendal, B., Ho, R. C., Prewitt, T. E., and Mccaskill-Stevens, W. J. Participation of minorities in cancer research: the influence of structural, cultural, and linguistic factors. *Annals of epidemiology*, 10(8):S22–S34, 2000.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pp. 5–13, New York, NY, USA, 1993. ACM. ISBN 0-89791-611-5. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

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

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1): 17816, 2017.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–663. Springer, 2018.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Orlando, J. I., Seebӧck, P., Bogunović, H., Klimscha, S., Grechenig, C., Waldstein, S., Gerendas, B. S., and Schmidt-Erfurth, U. U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. *arXiv preprint arXiv:1901.07929*, 2019.

Osband, I. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., and van der Laan, M. J. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52, 2015.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Swanson, G. M. and Ward, A. J. Recruiting minorities into clinical trials toward a participant-friendly system. *JNCI Journal of the National Cancer Institute*, 87(23): 1747–1759, 1995.

Thoral, P. et al. Right data, right now: developing a big data machine-learning based prediction model to prevent icu readmission. *Intensive Care Medicine Experimental 2018, 6*, 2018. URL https://icm-experimental.springeropen.com/track/pdf/10.1186/s40635-018-0201-6.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 2019.
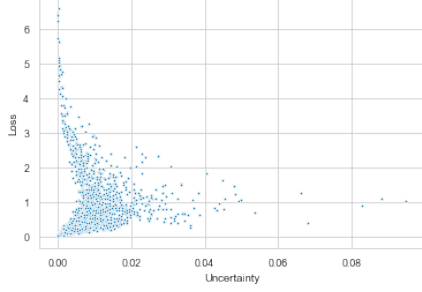
Figure 4. Illustration of how the data follows the bounds.

A.

In this section, we show the derivation of the bounds on the binary cross-entropy criterion relating to predictive uncertainty. Uncertainty is computed as the variance in the predictions for each datapoint x_i given T samples from a Bayesian posterior $P(\boldsymbol{\omega}|\boldsymbol{X}, \boldsymbol{Y})$. We start off with the Bhatia-Davis inequality on variance given a distribution maximum M and distribution minimum m:

$$\sigma^2 \leq (M - \mu)(\mu - m). \tag{8}$$

Solving the equality case for μ,

$$\mu^2 - (M + m)\mu + mM + \sigma^2 = 0 \tag{9}$$

$$\mu = \frac{M + m}{2} \pm \frac{1}{2}\sqrt{(M - m)^2 - 4\sigma^2}, \tag{10}$$

gives us bounds for the mean with respect to the variance:

$$\frac{M + m}{2} - \frac{1}{2}\sqrt{(M - m)^2 - 4\sigma^2} \leq \mu$$
$$\leq \frac{M + m}{2} + \frac{1}{2}\sqrt{(M - m)^2 - 4\sigma^2}. \tag{11}$$

In our case $\mu = P(y_i|x_i) = \frac{1}{T}\sum_{t=1}^{T} P(y_{it}|f_t^{\hat{\omega}}(x_i))$, $\sigma^2 = Var(y_i|x_i) = \frac{1}{T}\sum_{t=1}^{T}(P(y_i|x_i) - P(y_i|f_t^{\hat{\omega}}(x_i))^2$. $P(y_{it}|f_t^{\boldsymbol{\omega}}(x_i))$ is obtained from Sigmoid activation, therefore $m = \inf_{\boldsymbol{\omega}} P(y_i|x_i) = 0$ and $M = \sup_{\boldsymbol{\omega}} P(y_i|x_i) = 1$. Plugging these values and bounds in 11 into the binary cross-entropy criterion,

$$\mathcal{L}(y_i|x_i) = \begin{cases} \log(P(y_i|x_i)), & \text{for } y = 1 \\ \log(1 - P(y_i|x_i)), & \text{for } y = 0 \end{cases}, \tag{12}$$

gives us a single bound (for both $y = 1$ and $y = 0$) on the loss w.r.t. the uncertainty:

$$-\log\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4 \cdot Var(y_i|x_i)}\right)$$
$$\leq \mathcal{L}(y_i|x_i) \leq -\log\left(\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4 \cdot Var(y_i|x_i)}\right). \tag{13}$$
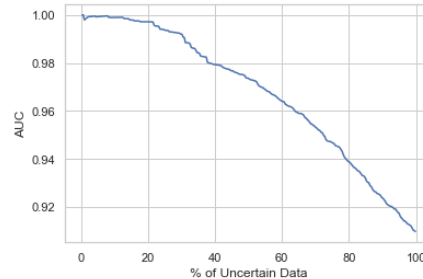
In figure 4 we depict how the data follows these bounds.

B.

In figure 5 we show that the uncertainty can calibrate performance in AUC for both a BNN (a) and a Gradient Boosting model, motivating researching the usage of uncertainty in combination with readily deployed models.
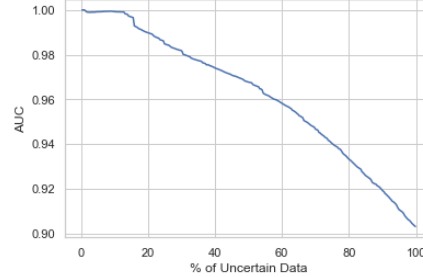
(a)

(b)

Figure 5. Predictive performance measured in AUROC (y-axis) related to the amount of uncertain data included in the analysis (x-axis).

C.

Patient data from the MIMIC-III database was used (Johnson et al., 2016). Vital signs, lab values and patient characteristics that were most abundantly available were gathered. Examples are blood pressure, potassium and age, respectively. To keep interpretability, we restricted ourselves to 25 clinically relevant features. Arterial and non-invasive blood pressures were combined, using the arterial blood pressure where possible. Features that lied further than 8 interquartile ranges were regarded as outliers and removed. Labels were obtained directly from the MIMIC tables, regarding expirement during the last hospital admission as a positive label. 9,237 ICU stays were set apart for testing purposes, leaving 36,944 patients for training.

D.

In figure 6 we observe how the predictions on the newborns follow the same moon shape as the trained dataset. However, average uncertainty is much higher.
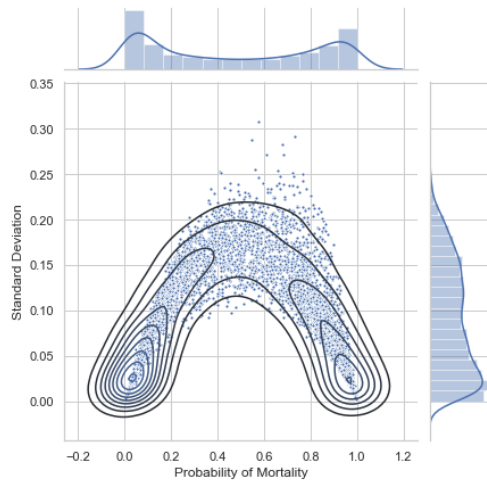
Figure 6.

E.

Following the approach of Hendrycks & Gimpel (2016), we compare how well sigmoid probabilities are able to detect correct classifications and out of domain patients compared to Bayesian uncertainty.

Table 1. Comparison of error and success detection between deterministic baseline and BNN.

MODEL	AUROC	AUPR SUCC	AUPR ERR
BNN STD	83.7	97.7	32.6
NN SIGMOID	79.7	95.8	37.7

Table 2. Comparison of in and out of domain detection between deterministic baseline and BNN.

MODEL	AUROC	AUPR IN	AUPR OUT
BNN STD	75.9	80.5	69.7
NN SIGMOID	49.8	61.7	42.5