Bayesian Modelling in Practice

Using Uncertainty to Improve Trustworthiness in Medical Applications

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

[1]University of Amsterdam, [2]Pacmed, [3]University Medical Center, Amsterdam

1 Introduction

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 patients. 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 (Blundell et al., 2015) 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.

2 The Loss Bounds

Through the Bhatia-Davis inequality we can derive an upper and lower bound on the obtainable binary cross-entropy loss (first term of Eq. 2.3) with respect to the predictive uncertainty (variance in the predictions, Eq. 2.1 and 2.2). The bounds (Eq. 1) are depicted in Figure 1. In Figure 2 we observe how the data empirically follow these bounds. Areas of low loss cannot be achieved with low predictive uncertainty. In other words, uncertainty will drive the loss from the extremes. This means that out of domain predictions will never result in extreme prediction risk and loss, a flaw that is commonly observed in neural networks.

$$-\log\left(\frac{1}{2}+\frac{1}{2}\sqrt{1-4\cdot Var(y_i|X_i)}\right)\leq \mathcal{L}(y_i|X_i)\leq -\log\left(\frac{1}{2}-\frac{1}{2}\sqrt{1-4\cdot Var(y_i|X_i)}\right)$$

Equation 1: The bounds of the loss with respect to predictive uncertainty.

$$P(y_i|\boldsymbol{x}_i)=\frac{1}{T}\sum_{t=1}^{T}P(y_{it}|f^{\hat{\omega}_t}(\boldsymbol{x}_i)) \quad (1)$$

$$Var(y_i|\boldsymbol{x}_i)=\frac{1}{T}\sum_{t=1}^{T}(P(y_i|\boldsymbol{x}_i)-P(y_i|f^{\hat{\omega}_t}(\boldsymbol{x}_i))^2 \quad (2)$$

$$\mathcal{L}_{VI}:=\log P(\boldsymbol{y}|\boldsymbol{X})-\mathrm{KL}\left(q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||P(\boldsymbol{\omega}|\boldsymbol{X},\boldsymbol{y})\right) \quad (3)$$

Equation 2: 1) Posterior predictive was approximated by Monte Carlo integration. 2) Uncertainty was measured by the variance in the predictions. 3) Minimized ELBO criterion.

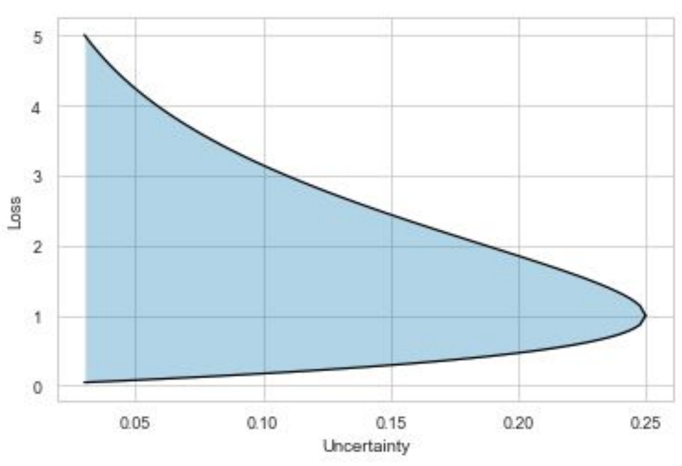

Figure 1: The bounds of the loss (Equation 1) visualized.

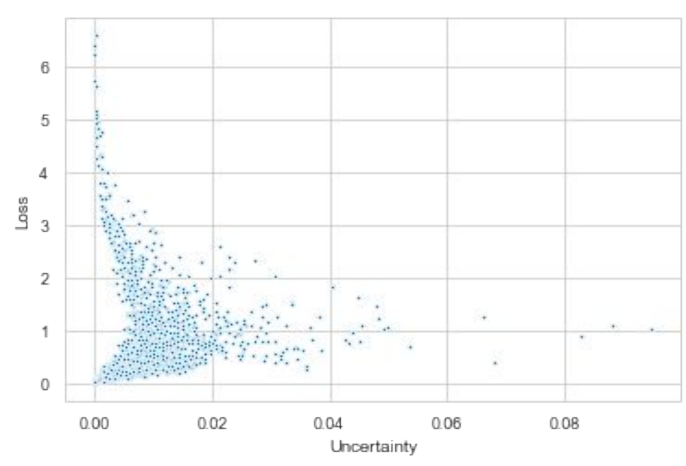

Figure 2: Empirical illustration of the bounded data.

3 Loss vs. Uncertainty

We empirically evaluate whether uncertain examples yield a higher loss. In figure 3 we ordered the predictive loss values according to the predictive uncertainty. We observe that the loss obtained on the 20% most uncertain data is 20 times as high as the 20% most certain data. Figure 4 shows a peculiar finding: this result even holds for the predictive loss of an unrelated gradient boosting model.

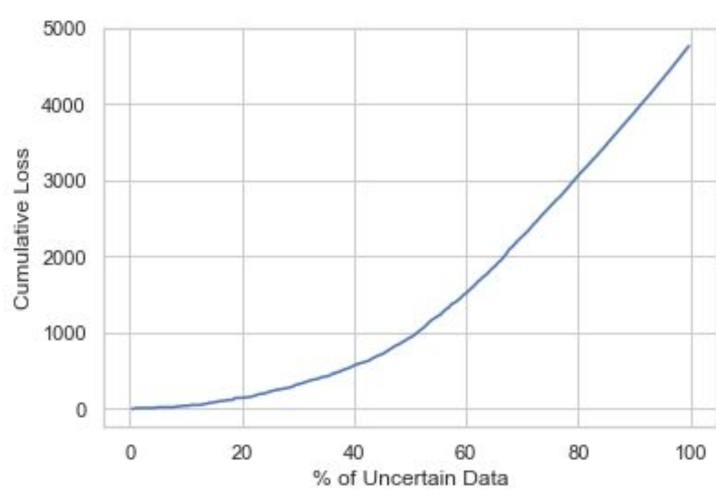

Figure 3:
Cumulative Loss of BNN with respect to Uncertainty

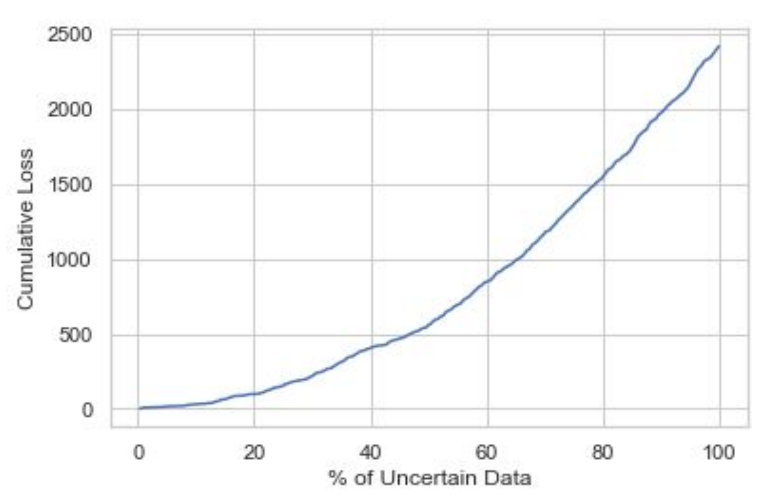

Figure 4:
Cumulative Loss of Gradient Boosting Model with respect to Uncertainty

4 Detecting Out-of-Domain Patients

Ethnic minorities tend to be underrepresented in medical research. Regrettably, study results are still applied to the entire population in practice, causing poorly substantiated treatments for these people. In Figure 6a we observe the predictive uncertainty (y-axis) as a function of prediction risk (x-axis). In the central region of the graph we notice that patients can both have low and high predictive uncertainty. The latter case is caused by patients that lie further from the domain. In figure 5 we see that the BNN can effectively identify out-of-domain patients (newborns, in this case). In figure 6b we depict the same plot as Figure 6a, though this time for the out-of-domain newborns. Additional experiments showed that the BNN was significant in identifying ethnic minorities (given the ICU signal data), meaning that it can indicate its uncertainty about treatment of these patients.

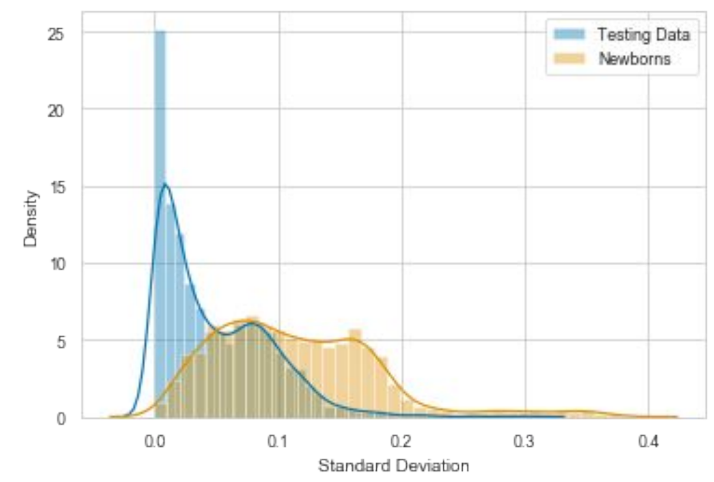

Figure 5:
Uncertainty density on the testing data and the Out-of-Domain data.

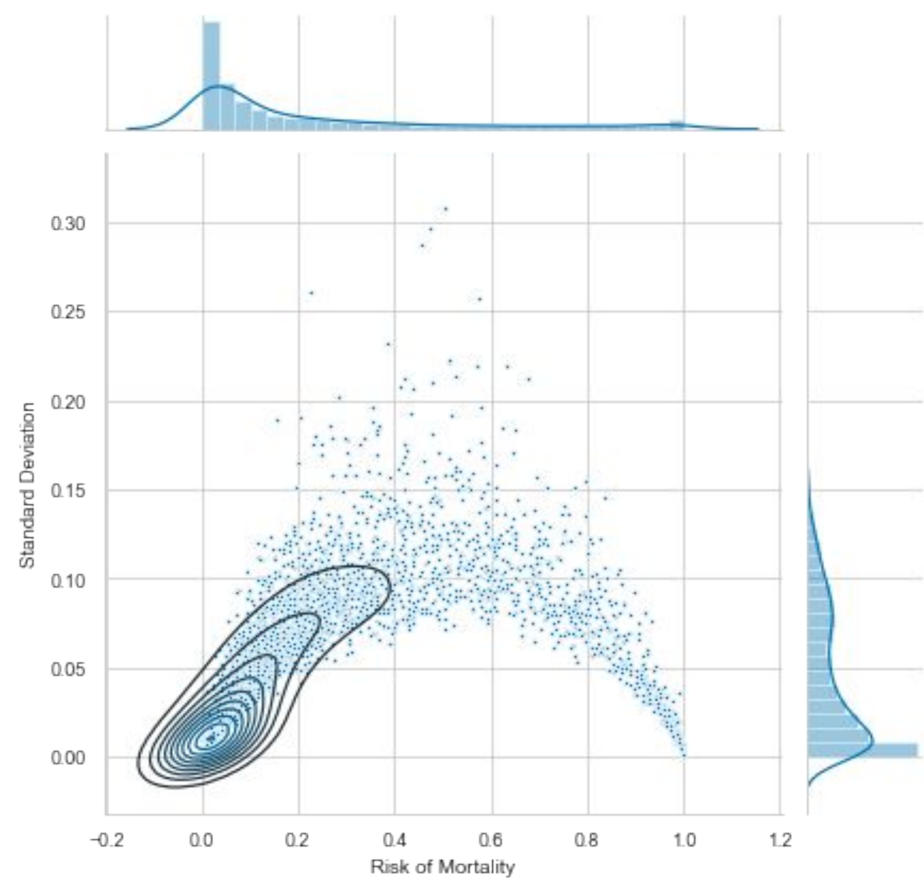

Figure 6a:
Uncertainty as a function of Predictive Risk

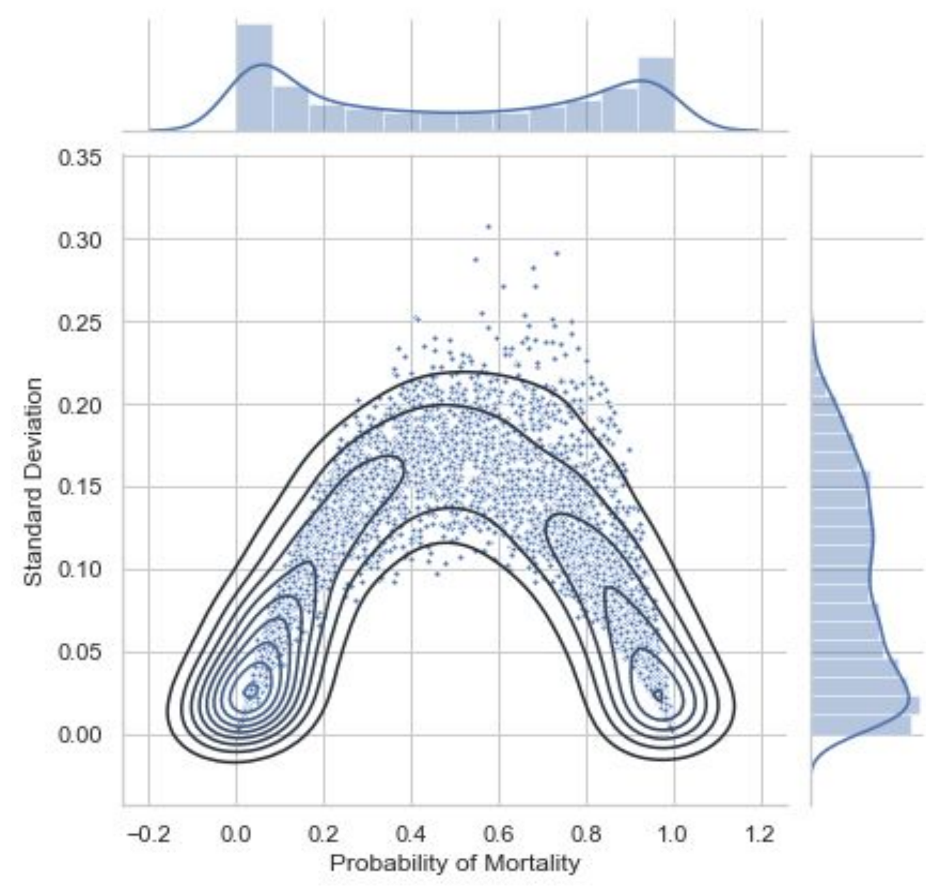

Figure 6b:
Uncertainty as a function of Predictive Risk

Furthermore, we follow the approach of Hendrycks & Gimpel (2016). The authors suggest that deterministic neural networks can effectively be used to identify out-of-domain data points. In this experiment, we compare this to our BNN. We compare the area under the ROC curve (AUROC) and area under the precision-recall (AUPR) curves for detecting out of domain patients between a neural network with sigmoid activation and a BNN. We see that the BNN is much more effective in identifying out-of-domain (newborn) cases.

Model	AUROC	AUPR In	AUPR Out
BNN STD	75.9	80.5	69.7
NN Sigmoid	49.8	61.7	42.5

Table 1:
Effectivity of detecting out-of-domain patients of a BNN compared to a deterministic BNN.

5 Conclusion

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. We trained a BNN on the MIMIC critical care dataset. Using derived bounds on cross-entropy loss with respect to predictive uncertainty, we found that uncertainty is able to mitigate performance risk and loss. We applied this intuition to show empirically that the loss of the prediction on a test set 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 a BNN can effectively identify patients outside of its previously observed domain. Overall, this work demonstrates that uncertainty is effective in enhancing model trustworthiness in a critical setting like the ICU. A mathematical intuition as to why predictive uncertainty generalizes to other models remains unclear. This can be an interesting direction for future research.

ICML

AI for Social Good