An Empirical Analysis of Uncertainty Estimation in Genomics Applications

Sepideh Saran^{1,4}
sepideh.saran@mdc-berlin.de

Mahsa Ghanbari^{1,2} *
mahsa.ghanbari@mdc-berlin.de

Uwe Ohler^{1,2,3*}
uwe.ohler@mdc-berlin.de

[1]The Berlin Institute for Medical Systems Biology
Max Delbrück Center for Molecular Medicine
[2]Department of Biology, Humboldt Universität zu Berlin
[3]Department of Computer Science, Humboldt Universität zu Berlin
[4]Department of Electrical Engineering and Computer Science, Technical University of Berlin

Abstract

The usability of machine learning solutions in critical real-world applications relies on the availability of an uncertainty measure that reflects the confidence in the model predictions. In this work, we present an empirical analysis of uncertainty estimation approaches in Deep Learning models. We contrast Bayesian Neural Networks (BNN) against Monte Carlo-dropout (MC-dropout) methods to evaluate their performance and uncertainty scores in two classification tasks with different dataset characteristics.

1 Introduction

Neural Networks have become the state-of-the art methods for identifying functional elements in the genome (Eraslan et al., 2019). Superior performance, scalability, and the rise of interpretation approaches (Carvalho et al., 2019; Samek et al., 2020; Bach et al., 2015; Sundararajan et al., 2017) have made Deep Learning models a popular choice in many genomics applications. For example, Deep Learning models have been used to accurately map the genomic sequences to the associated functional molecular readouts, such as protein binding information (Ghanbari and Ohler, 2020). In doing so, they identify short sequence elements called "motifs" that serve as target binding sites for proteins like transcription factors (TFs) and RNA-binding proteins (RBPs). These motifs can be discovered through interpretation of the network (Sundararajan et al., 2017), leading to new insights and a better understanding of the genomic associations.

For the ultimate use of these models in many critical downstream tasks (e.g., predicting genetic variant effects), it is essential to provide a measure for the confidence in the model's outputs. Providing uncertainty measurements enhances the credibility of the proposed machine learning solution and helps clinicians in the subsequent decision-making process.

Bayesian approaches such as Bayesian neural networks and Gaussian processes provide uncertainty scores besides the predictive probabilities (Hüllermeier and Waegeman, 2021). BNNs bring stochasticity into the network by learning a distribution for each weight instead of a single point estimate (Blundell et al., 2015). A prior distribution is selected for each parameter and is represented by its

[*] These authors contributed equally to this work.

Bayesian Deep Learning workshop, NeurIPS 2021.

mean and standard deviation (SD), increasing the number of trainable parameters in the Bayesian neural network to twice the size of the equivalent deterministic architecture. Since exact inference is computationally intractable for these networks, Variational Inference (Jordan et al., 1999; Wainwright and Jordan, 2008; Graves, 2011; Blei et al., 2017) is used to approximate the posterior on weights and thus train the network (Filos et al., 2019). A predictive distribution for the output can be generated by drawing values for weights from the approximated posterior distribution (i.e., Monte Carlo sampling).

Monte-Carlo dropout has been introduced as a probabilistic approach that approximates Bayesian inference in deep Gaussian processes (Gal and Ghahramani, 2016). MC-dropout extends the now well-established practice of using Dropouts (Srivastava et al., 2014) during the training phase to the test phase, bringing stochasticity to the network at inference time. MC-dropout has the same number of parameters as the deterministic equivalent network and uses the same training procedure, making it computationally more efficient than Bayesian neural networks. MC-dropout has revealed promising results on image datasets (Filos et al., 2019), but it has not been tested on other data types like genomic data.

In this work, we aim to investigate predictive uncertainty estimation methods in a genomics application. Experimental noise, incorrect dataset labels, out-of-distribution samples, class imbalance, and presence of multiple motifs (i.e., multi-label setting) are the major reasons for uncertainty in computational models in biology. Our goal is to find the most suitable strategy that well-reflects the model's uncertainty in its prediction without sacrificing predictive performance. Furthermore, these uncertainty scores should be reliable for further investigating the source of uncertainty, as well as selecting high-quality predictions, which in turn can be used in downstream analysis (Monti et al., 2021).

To this end, we empirically compare the BNN approach and the MC-dropout in the task of predicting RNA binding protein binding sites. This presents a novel application of uncertainty estimation methods on multi-label imbalanced genomic data that goes beyond the standard image processing applications. However, we additionally use MNIST dataset as a benchmark to compare the performance of both approaches in different scenarios common in a genomics application, namely imbalance data, and scarce data. Specifically, the increased number of parameters and training through approximation in BNNs results in the need for more computational resources and larger datasets. In applications where data is scarce, BNNs might not perform well in their prediction and the quality of uncertainty estimates. In our comparisons, we, therefore, assess the usability, performance, and agreement of the uncertainty estimates provided by these two approaches in each application.

The remainder of this paper is organized as follows: In section 2, we describe the two classification tasks and the two uncertainty estimation methods we examine in our experiments. Section 3, presents our evaluations comparing the two methods and discussion on our observations. Finally, section 4 concludes this paper and presents insights into the future work.

2 Analysis of Uncertainty Estimation Approaches

To study the predictive uncertainty estimation methods in deep learning, we have compared a Bayesian neural network with variational inference against a Monte Carlo-dropout network on two classification tasks. We conduct experiments under different dataset settings to cover scenarios that are particularly of interest in biological applications, namely imbalance data, scarce data, multi-label and single-label classification. In this section, we describe our selected applications, datasets, and modeling choices.

2.1 MNIST Classification

The MNIST classification task (LeCun et al., 1998) is a well-established benchmark in machine learning. The dataset contains $70,000$ images of handwritten digits in grayscale. The dataset is normalized and balanced, with each digit centered in a 28x28 pixel box. We run several experiments on this multi-class classification problem as a benchmark for comparing MC-dropout and the BNN in terms of the prediction performance and the estimated uncertainty. To further evaluate these approaches, we repeat the experiments on an imbalanced version of the MNIST dataset with around $42,000$ samples. The imbalanced version of the dataset is generated by sub-sampling from the original dataset using to the following fractions: class 0: 0.5, class 1: 0.7, class 2: 0.8, class 3: 0.3, class 4: 0.6, class 5: 0.8, class 6: 0.4, class 7: 0.2, class 8: 0.9 and class 9: 0.8 (see Figure1).

2.2 RNA Binding Protein Site Classification

Proteins regulate different stages of gene expression through binding to specific locations in the genome or the RNA. Proteins that bind to RNA can affect post-transcriptional processes such as splicing, RNA localization, translation, and degradation (Gerstberger et al., 2014). These RBPs are selective of their target binding sites in terms of patterns in the sequence (i.e. sequence motifs) or the RNA structure (i.e. structural motifs). Some RBPs share preferred sequence patterns for binding, and some have distinct motifs. Identifying RBP binding preferences opens the door to understanding their function in the downstream biological processes. Following Ghanbari and Ohler (2020), we formulate the problem of RBP binding prediction as a multi-label classification task (i.e., to predict the RBP class, given sections of the RNA sequence) and extend our experiments for comparing the uncertainty estimation approaches for this problem. PAR-CLIP experiments provide high-resolution binding site sequences for the RBP of interest (Hafner et al., 2010). These bindings sites are small RNA fragments of a typical size of a few dozen nucleotides, which contain sequence patterns of typically 4-8 nucleotides in length. We select PAR-CLIP data (Mukherjee et al., 2019) for three RBPs, namely MBNL1, PUM2, and QKI, for this task. Our dataset is a combination of different published PAR-CLIP datasets from the HEK293 cell-line that are processed uniformly through a single pipeline (Corcoran et al., 2011). The dataset contains approximately $16,500$ samples in total. It is imbalanced and includes approximately 475 multi-label samples, meaning one sequence can have the binding motifs for multiple RBPs. For further analysis, we sub-sample from the larger classes to build a balanced multi-label dataset containing about $4,500$ (including 460 multi-label) samples (see Figure1) and a balanced single-label version of the same dataset. We design our experiments based on the work of Ghanbari and Ohler (2020) and take their proposed deep learning model called DeepRipe as our deterministic architecture for this task.

2.3 Experimental design

We design experiments to compare the prediction and uncertainty measurements of the two selected approaches for estimating uncertainty, namely, the Bayesian neural network with variational inference and the MC-dropout method. Our main interest is to study the applicability of these methods on imbalanced or small datasets, which is often the case in biological applications. To this end, we define two scenarios for MNIST classification task for balanced and imbalanced datasets, and three scenarios for the RBP binding classification task: multi-label setting with balanced dataset, multi-label setting with imbalanced dataset and finally single-label setting with balanced dataset. We have implemented variations of the two probabilistic models, as well as deterministic models for each scenario. Here, we describe the implementation details of these models.

The base model architecture for our experiments in both classification tasks is a simple neural network with two convolutional layers, each followed by a max-pooling layer and a final fully-connected layer. For MNIST, we use the softmax activation function in the last layer, whereas, in the RBP classification, we use Sigmoid to address the multi-label setting.

In the deterministic baseline models, we use dropout for regularization. The BNN is built on a similar architecture, but with Gaussian prior and posterior on the trainable parameters and mean-field variational inference (Peterson and Anderson, 1987; Blundell et al., 2015) as the training approach. To speed up the training process, we use the Flipout Monte Carlo estimator that de-correlates the gradients within each mini-batch (Wen et al., 2018). Our MC-dropout networks have the same architecture as the deterministic models, with the difference of activating the dropout in the final layers at testing time.

The deterministic and MC-dropout networks are trained with a batch size of 128 for 30 epochs with early stopping based on the validation loss with the patience set to 8. However, for the results of the networks to be comparable, we needed to train the BNNs longer to reach similar accuracy to MC-dropout's. For all the experiments, we have used the Adam optimizer.

After training the probabilistic networks, we infer predictions on the test set for 1000 iterations, building an output distribution for each test example through Monte Carlo sampling. The mean of the distribution is regarded as the model prediction, and the standard deviation represents the model's uncertainty for the given test sample. We fix the test set across all models for each task and compare the performance and uncertainties of the models.

Experiments are repeated with different random seeds, and the models are implemented in python using the Tensorflow library (Abadi et al., 2015).

3 Empirical Evaluation

This section presents our observations in our experiments with the Bayesian neural network trained with variational inference and the MC-dropout network. As explained in the previous section, we train different variations of each model and evaluate their predictive distributions for the test set.

3.1 Usability in different scenarios

We consider the accuracy of the deterministic model as our baseline for the performance of the models in each scenario. The probabilistic models are then trained to reach similar performance, to ensure comparability, and to avoid sacrificing accuracy when estimating uncertainty. As expected, we observe that BNNs require a much longer training time and more computational resources in all scenarios. Moreover, training BNNs for the multi-label setting in the RBP binding classification task appeared unstable when repeated with different random seeds, with instances not converging or getting stuck in local minima. However, the training process was improved when we reduced the learning rate from 0.01 to 0.005 and allowed for longer patience for early stopping. We found BNNs to be sensitive to the choice of optimizer and hyper-parameters such as learning rate, whereas training MC-dropout networks seemed robust against these choices. However, selecting the dropout rate in MCD, and choosing which layers to have MC-dropout activated, can influence the performance.

3.2 Evaluating predictive performance

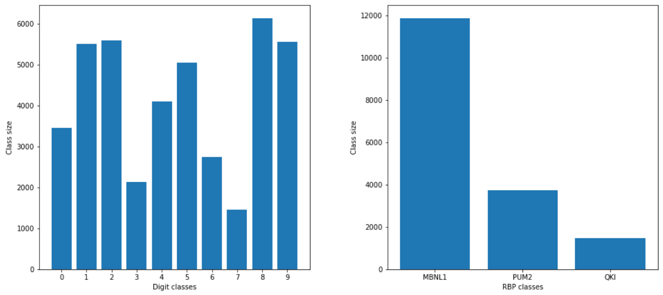
Following our experimental design, we compare methods with similar performance to the deterministic model. All models show high predictive performance on both tasks. The models' performance on the RBP prediction task are shown in Figure 2. With similar performance, BNNs tend towards prediction values closer to zero or one compared to deterministic and MDC models for both tasks (See Figure 3 for the RBP binding task). Furthermore, we calculated Pearson correlation and Spearman correlation between the means of the predictive distributions to illustrate agreements of the methods in their predictions. Figure 4 shows these correlations between MCD and BNN on the three scenarios for the RBP binding task.

3.3 Comparing uncertainty estimations

To assess the agreement and quality of the uncertainty measures provided by the two probabilistic methods, we compare the standard deviations of the corresponding predictive distributions. Similar to the means, we calculated Pearson correlation and Spearman correlation between the standard deviations of the predictive distributions. Compared to the means, correlations of SDs are lower across the models (see Figure 4 for RBP binidng task). This suggests that models with different uncertainty estimation approaches produce different ranges of uncertainty scores. In particular, BNNs seem to have a narrower range of uncertainty scores compared to MCD models. This observation may be specific for our implementation choices for BNN models and needs to be verified in other settings.

Additionally, we calculate Kendall's tau test (Kendall, 1938) between SDs of all models. Kendall rank coefficient correlation (Kendall's tau) measures the correspondence between ordinal data. The test returns the tau value and the p-value. Kendall's tau value ranges between -1 and 1, with values close to 1 indicating strong agreement. Tests on both tasks yielded significant p-values, rejecting the null hypothesis of no association between the uncertainties. In all of our scenarios, the tau values are positive, however the agreement between uncertainty estimates of different MCD methods is stronger than the agreement between MCD and BNN or different BNNs with each other (see Figure 5).

To further investigate the quality of uncertainty estimates provided by the two approaches, we selected the most uncertain samples (i.e highest standard deviation) and analyzed commonalities and differences between the BNN and MCD. Figure 6 and Figure 7 illustrate the overlap between the most uncertain samples identified by BNN and MCD in the two classification tasks. Moreover, we visualized the samples that are considered uncertain in one approach but not the other for the MNIST task (Figure8). From looking into the most uncertain samples per class, we find the uncertainty estimations of the MCD to be of a higher quality and more similar to what we would consider an ambiguous case. At this point, this is done only by visual explorations and not objectively quantified.



(a) imbalanced MNIST dataset

(b) imbalanced RBP dataset

Figure 1: Distribution of class sizes in the imbalanced version of the MNIST dataset and the imbalanced version of the RBP dataset

4 Conclusion and Future Work

This paper presented our findings in experimenting with BNNs and MC-dropout for uncertainty estimation in two classification tasks. We observed strong agreement between models' predictions and a high correlation between their uncertainty scores in different scenarios. This suggests that MC-dropout can potentially be used as a computationally cheaper alternative to BNNs for uncertainty estimation. Especially in applications where a trained deterministic model is already in place, transition to a MC-dropout model requires less effort compared to BNN.

We want to extend experiments in this work to evaluate the quality of uncertainty scores with out-of-distribution samples (e.g., random genome sequence) and to assess the use of Concrete Dropout (Gal et al., 2017).

In our future work, we would like to study the use of uncertainty estimates in several downstream tasks, especially for model interpretation. Uncertainty estimation, coupled with interpretability methods, can identify the source of predictive uncertainty from the model's point of view. We want to use interpretability methods to visualize the most uncertain samples for the RBP binding task to get more insights on sequence motifs and RBP binding preferences. Comparing uncertain sequence motifs to the known motifs for a given RBP can help us select a more suitable approach for uncertainty estimation tailored to our application. Moreover, the uncertainty scores of individual samples can be used in building population-level descriptive patterns (e.g., consensus motif for a family of RBPs), which is an active area of research in the field of explainable Artificial Intelligence.

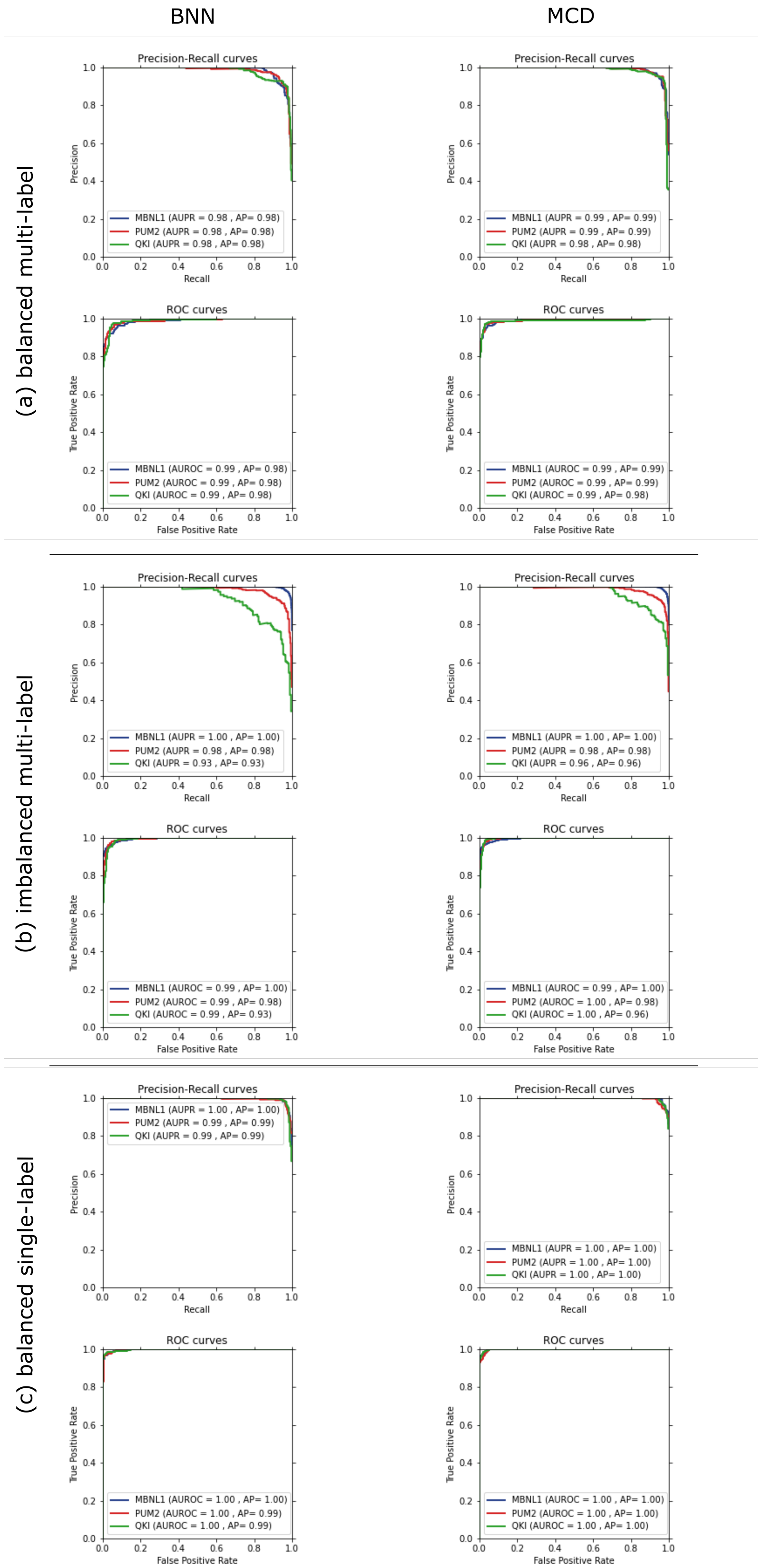


Figure 2: Predictive performance of an instance of BNN and instance of MC-dropout (shown as MCD) in the RBP binding task in three different scenarios: (a) balanced dataset in multi-label classification setting, (b) imbalanced dataset in multi-label classification setting and (c) balanced dataset in single-label classification setting.The metrics presented here for evaluating the performance in each class are as follows: average precision (AP), area under the receiver-operating characteristic curve (AUROC) and finally the area under the precision-recall curve (AUPR). AUPR is preferred over the AUROC for evaluation in the imbalanced dataset settings.

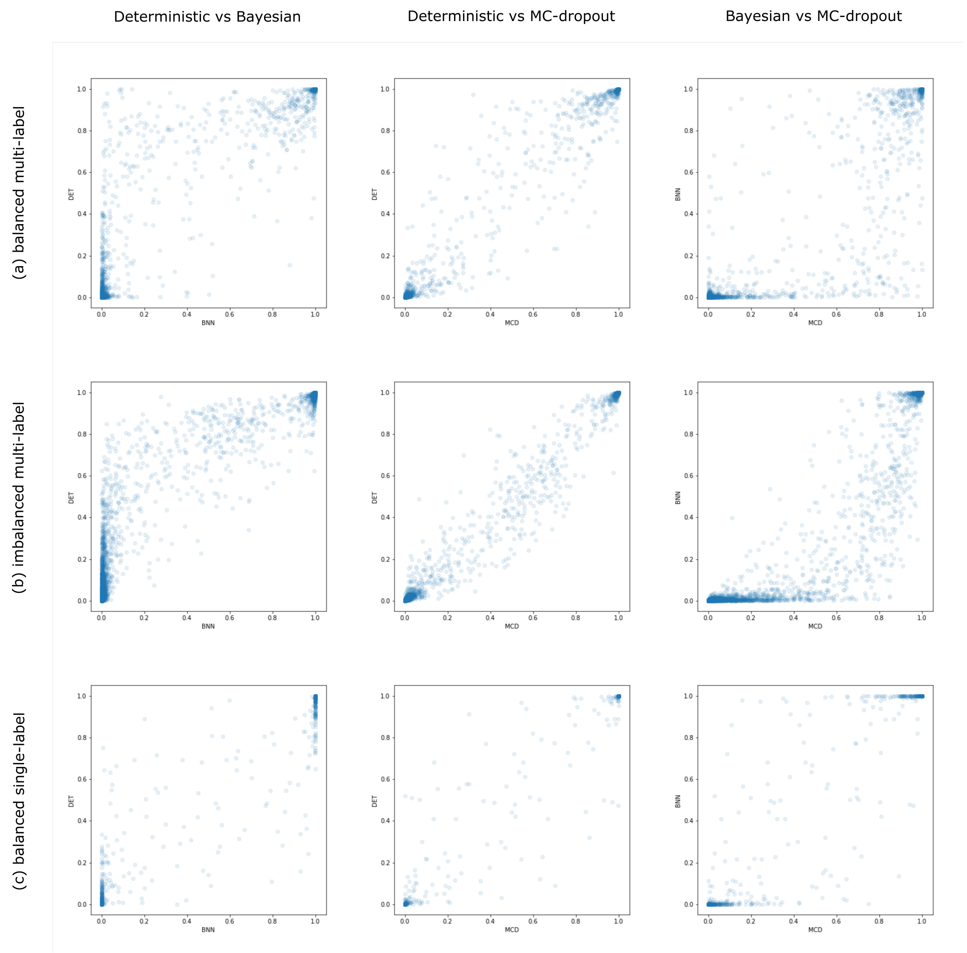
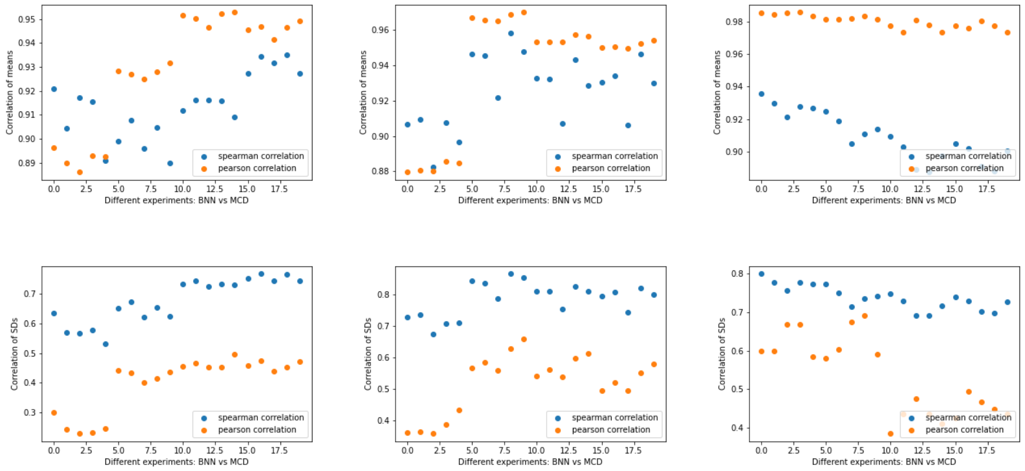
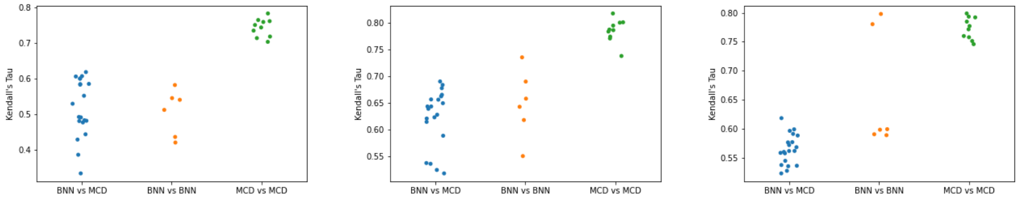


Figure 3: Comparison of the predicted values of the probabilistic and deterministic models in the RBP binding tasks in three different scenarios: (a) balanced dataset in multi-label classification setting, (b) imbalanced dataset in multi-label classification setting and (c) balanced dataset in single-label classification setting. The left column plots the predictions of the deterministic model against BNN for each scenario. Similarly, the middle column plots predictions of the deterministic model against MCD, and the column on the right shows the same for BNN against MCD.



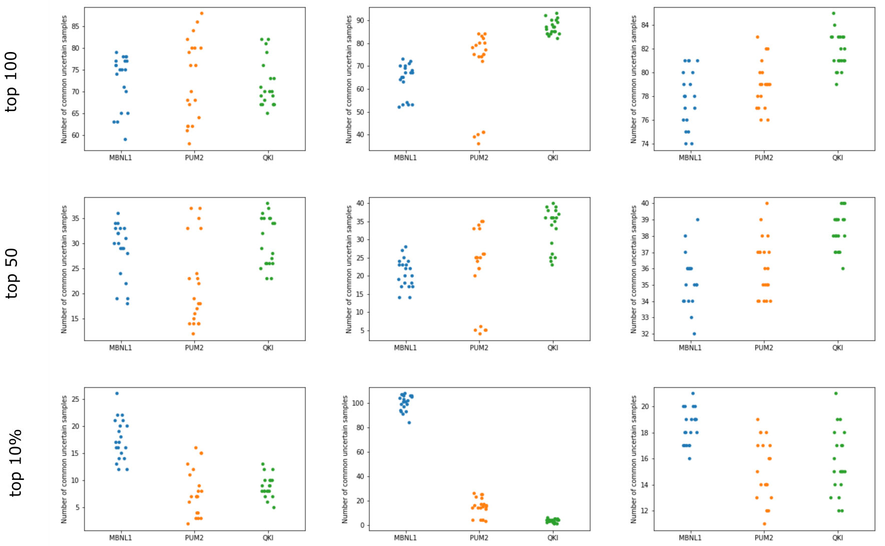
(a) balanced multi-label (b) imbalanced multi-label (c) balanced single-label

Figure 4: Comparing BNN models' predictions and uncertainty scores against MCD models' in the RBP binding tasks in three different scenarios: (a) balanced dataset in multi-label classification setting, (b) imbalanced dataset in multi-label classification setting and (c) balanced dataset in single-label classification setting. This is illustrated by showing the Pearson and Spearman correlations between the means in the first row, and between the SDs in the second row, for each experiment (i.e. comparing pairs of BNN vs MCD).



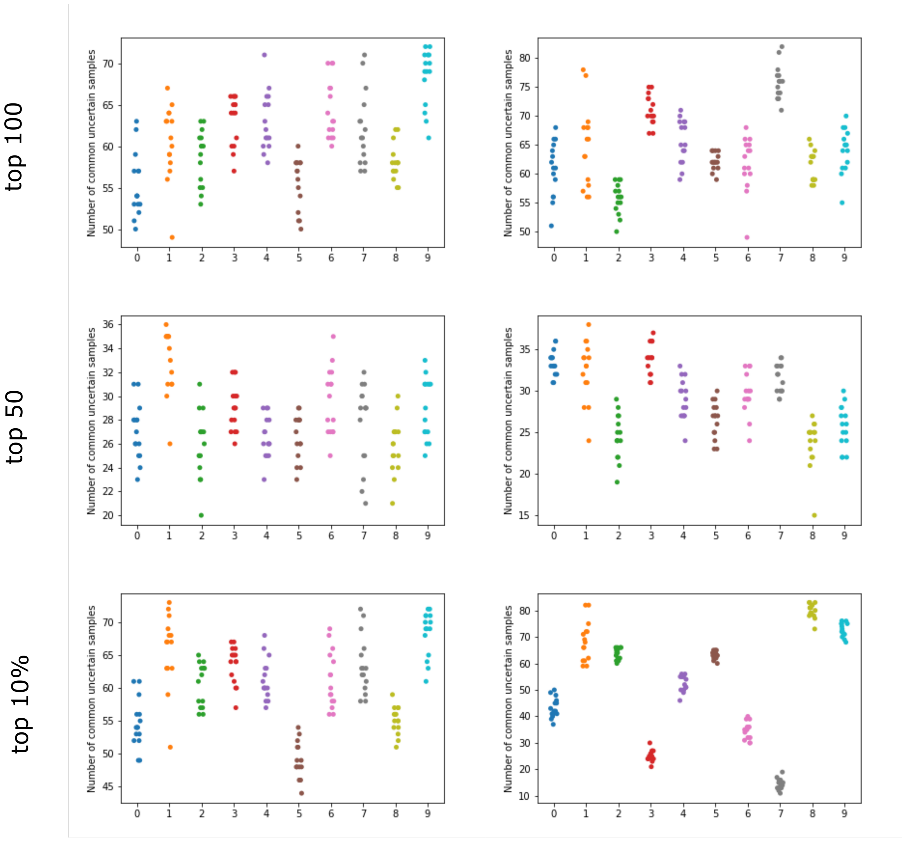
(a) balanced multi-label (b) imbalanced multi-label (c) balanced single-label

Figure 5: Kendall's tau value for measuring correlation between the uncertainty estimates of BNN and MCD in the RBP binding tasks in three different scenarios: (a) balanced dataset in multi-label classification setting, (b) imbalanced dataset in multi-label classification setting and (c) balanced dataset in single-label classification setting. Kendall's tau values range from -1 to 1, with 1 showing strong agreement and -1 strong disagreement of the two ordinal datasets.



(a) balanced multi-label (b) imbalanced multi-label (c) balanced single-label

Figure 6: Comparing the number of samples that are considered uncertain by both of the BNN and MCD models in the RBP binding tasks in three different scenarios: (a) balanced dataset in multi-label classification setting, (b) imbalanced dataset in multi-label classification setting and (c) balanced dataset in single-label classification setting. This is illustrated in the first row by the count of common uncertain samples between top 100 most uncertain samples for each class as indicated by the BNN compared to of those indicated by MCD. Similarly, the count of common uncertain samples in top 50 most uncertain samples are shown in the second row, and the top 10% most uncertain samples in the third row.



(a) balanced

(b) imbalanced

Figure 7: Comparing the number of samples that are considered uncertain by both of the BNN and MCD models in the MNIST tasks in two different scenarios: (a) balanced dataset and (b) imbalanced dataset. This is illustrated in the first row by the count of common uncertain samples between top 100 most uncertain samples for each class as indicated by the BNN compared to of those indicated by MCD. Similarly, the count of common uncertain samples in top 50 most uncertain samples are shown in the second row, and the top 10% most uncertain samples in the third row.

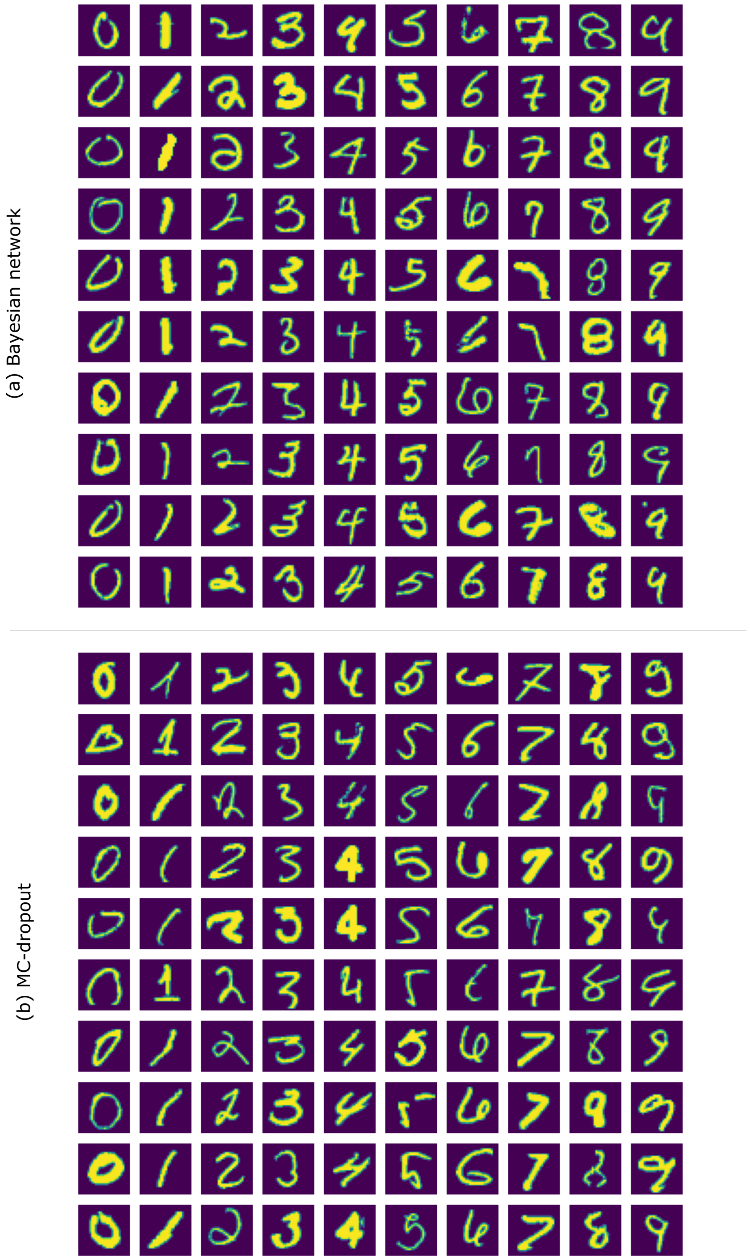


Figure 8: Comparing samples that are considered uncertain by only one of the BNN or MCD models in the MNIST task with balanced dataset. This is illustrated as follow: (a) shows 10 samples from each class, that are among the top 100 most uncertain samples based on uncertainty scores provided by BNN model, but are not among the 100 most uncertain of the corresponding MCD model. (b) shows 10 samples from each class, that are among the top 100 most uncertain samples based on uncertainty scores provided by MCD model, but are not among the 100 most uncertain of the corresponding BNN model.

Acknowledgments and Disclosure of Funding

S. Saran and U. Ohler are partially supported by BMBF grant 01DQ19008 and acknowledge support by the Helmholtz-Einstein International Berlin School in Data Science (HEIBRiDS). M. Ghanbari and U. Ohler are supported by BMBF funding for the Berlin Institute for the Foundations of Learning and Data (BIFOLD).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* 10, 7 (2015). https://doi.org/10.1371/journal.pone.0130140

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.* 112, 518 (Apr 2017), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight Uncertainty in Neural Networks. arXiv:1505.05424 [stat.ML]

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019). https://doi.org/10.3390/electronics8080832

David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, and Uwe Ohler. 2011. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology* 12 (2011). https://doi.org/10.1186/gb-2011-12-8-r79

Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. 2019. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* 20 (2019), 389–403. https://doi.org/10.1038/s41576-019-0122-6

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. 2019. A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks. arXiv:1912.10481 [stat.ML]

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Appendix. arXiv:1506.02157 [stat.ML]

Yarin Gal, Jiri Hron, and Alex Kendall. 2017. Concrete Dropout. arXiv:1705.07832 [stat.ML]

Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. 2014. A census of human RNA-binding proteins. *Nature Reviews Genetics* 15 (2014), 829–845. https://doi.org/10.1038/nrg3813

Mahsa Ghanbari and Uwe Ohler. 2020. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Research* 30 (2020), 214–226. https://doi.org/10.1101/gr.247494.118

Alex Graves. 2011. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf

Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. 2010. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* 141, 1 (2010), 129–141. https://doi.org/10.1016/j.cell.2010.03.009

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* 110 (2021), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Michael Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning* 37 (01 1999), 183–233. https://doi.org/10.1023/A:1007665907178

M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93. http://www.jstor.org/stable/2332226

Yann LeCun, León Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, Vol. 86. 2278–2324. http://yann.lecun.com/exdb/publis/pdf/lecun-98.pdf

Remo Monti, Pia Rautenstrauch, Mahsa Ghanbari, Alva Rani James, Uwe Ohler, Stefan Konigorski, and Christoph Lippert. 2021. Identifying interpretable gene-biomarker associations with functionally informed kernel-based tests in 190,000 exomes. *bioRxiv* (2021). https://doi.org/10.1101/2021.05.27.444972 arXiv:https://www.biorxiv.org/content/early/2021/05/27/2021.05.27.444972.full.pdf

Neelanjan Mukherjee, Hans-Hermann Wessels, Svetlana Lebedeva, Marcin Sajek, Mahsa Ghanbari, Aitor Garzia, Alina Munteanu, Dilmurat Yusuf, Thalia Farazi, Jessica I Hoell, Kemal M Akat, Altuna Akalin, Thomas Tuschl5, and Uwe Ohler. 2019. Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Research* 47 (2019), 570–581. https://doi.org/10.1093/nar/gky1185

Carsten Peterson and James R. Anderson. 1987. A Mean Field Theory Learning Algorithm for Neural Networks. *Complex Systems* 1 (1987), 995–1019.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. 2020. Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *CoRR* abs/2003.07631 (2020). arXiv:2003.07631 https://arxiv.org/abs/2003.07631

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs.LG]

Martin Wainwright and Michael Jordan. 2008. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* 1 (01 2008), 1–305. https://doi.org/10.1561/2200000001

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. 2018. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. arXiv:1803.04386 [cs.LG]