Proceedings of the
9th Graz Brain-Computer Interface Conference 2024

10.3217/978-3-99161-014-4-016

UNCERTAINTY QUANTIFICATION FOR CROSS-SUBJECT MOTOR IMAGERY CLASSIFICATION

Prithviraj Manivannan*, Ivo Pascal de Jong*, Matias Valdenegro-Toro, Andreea Ioana Sburlea

Department of Artificial Intelligence, Bernoulli Institute, University of Groningen, The Netherlands

E-mail: ivo.de.jong@rug.nl

ABSTRACT: Uncertainty Quantification aims to determine when the prediction from a Machine Learning model is likely to be wrong. Computer Vision research has explored methods for determining epistemic uncertainty (also known as model uncertainty), which should correspond with generalisation error. These methods theoretically allow to predict misclassifications due to inter-subject variability. We applied a variety of Uncertainty Quantification methods to predict misclassifications for a Motor Imagery Brain Computer Interface. Deep Ensembles performed best, both in terms of classification performance and cross-subject Uncertainty Quantification performance. However, we found that standard CNNs with Softmax output performed better than some of the more advanced methods.

INTRODUCTION

Machine Learning systems for Brain Computer Interfaces (BCI) are normally optimised to their predictive accuracy. The availability of public datasets and benchmarking systems allow for faster progress in this direction. However, for successful BCI systems there are more aspects that need to be explored.

This study explores the options of Uncertainty Quantification (UQ) for Machine Learning models [1] as applied to non-invasive Motor Imagery BCIs. Uncertainty Quantification aims to estimate how likely a prediction from a Machine Learning model is to be correct. For this two types of uncertainty are commonly considered.

Two types of Uncertainty: Aleatoric uncertainty (also referred to as data uncertainty) is the uncertainty inherent in the data. This cannot be reduced by better models, only by better EEG recordings or better paradigms. Noisy EEG recordings or extracted features that are poorly correlated to the to-be-predicted classes introduce aleatoric uncertainty.

Epistemic uncertainty (also referred to as model uncertainty) is the uncertainty in the model. This kind of uncertainty can be reduced by collecting more training samples that are similar to what the model is being evaluated on. In BCI contexts this uncertainty can come from limited amounts of training data [2], but also from between-subject variability [3].

While there is some Motor Imagery BCI research dedicated to UQ [2–5], it is worth noting that simple methods of estimating aleatoric uncertainty are often readily available. For example, Neural Networks used for classification generally use Softmax or Sigmoid activation functions for the output, which also gives a crude estimate of aleatoric uncertainty.

This study, like most research on modelling epistemic uncertainty is mostly done in the domain of Deep Learning.

Using Uncertainty for Rejection: UQ is often considered as a method for improving interpretability of predictions from a Machine Learning model [6]. There, the goal is to have a precise and well calibrated prediction of the class probability. This means that a prediction with 90% certainty should be correct 90% of the time. This results in methods aimed at addressing overconfidence of Neural Networks [7].

However, for BCIs there is often no time for human interpretation of the classification. Instead, the system should automatically deal with certain and uncertain predictions. Typically this means "rejecting" the uncertain predictions and abstaining from sending a control command to the device. We focus on this rejection case, as it aligns with how BCIs are implemented in practice, and highlight that it comes with different methods and metrics.

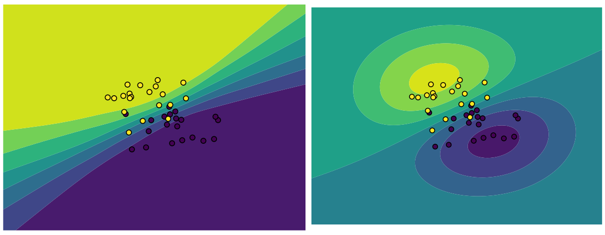
Research Aim: This paper investigates whether UQ methods that account for epistemic uncertainty can identify wrong predictions in cross-subject classification. This expands on previous work [3, 5] by exploring a larger variety of UQ methods and by applying a leave-one-subject-out cross validation paradigm to get a more realistic estimate of model performance.

We investigate whether available UQ methods for CNNs that account for epistemic uncertainty are actually able to reject the uncertain predictions when applied cross-subject better than the crude methods readily available.

Previous work has shown success with rejection methods [5], but a comparison with simple baseline methods such as Softmax is missing. Moreover, by using different measures of uncertainty we can see how much aleatoric and epistemic uncertainty contribute to the total uncertainty. This disentangling of uncertainties has not been applied to BCIs before [8]. Lastly, we cover a wider range of UQ methods and explain how they have different underlying assumptions.

[*] These authors contributed equally to this work

Published by
Verlag der Technischen Universität Graz



(a) Discriminative model (b) Generative model

Figure 1: An illustration of a discriminative and a generative model. The yellow and purple dots indicate the training samples of two different classes. The background indicates the prediction. The green color indicates uncertainty.

Background on uncertain models: Following [9] we consider two assumptions for how epistemic uncertainty may be modelled. [9] calls these two assumptions *discriminative* and *generative* models.

Discriminative models learn a boundary that optimally separates the classes. Samples that are far away from this boundary are considered "certain", whereas samples that are close to this boundary are considered uncertain. When the model uncertainty is considered, these methods consider multiple decision boundaries that are all valid with the training data. When samples fall between different decision boundaries this is considered epistemic uncertainty. Figure 1a shows what this looks like in a 2D feature space. This could be the band power following 2 CSP filters, but a similar concept can also be applied at a higher dimensional space for Neural Networks. In contrast, generative models learn the distribution of each class. A sample that matches the distribution of the training data is considered "certain", whereas a sample that is far away from the training data is considered "uncertain". Figure 1b visualises this concept.

Both approaches have similar behaviour under aleatoric uncertainty. This is seen in the parts where the two classes overlap. However, they exhibit very different behaviour under epistemic uncertainty. Since it is not known which of these underlying assumptions is most suitable it is important to consider models from either family.

Bayesian Neural Networks: Bayesian Neural Networks (BNNs) fall under the category of discriminative models. Standard Neural Networks learn a single optimal vector θ of the parameters learned on the training data D. They then do classification according to the Softmax function to capture aleatoric uncertainty.

BNNs instead consider a weight distribution $p(\theta|D)$. This captures all possible weights for the Neural Networks, based on how well they fit the data. Inference is then made according to the predictive posterior distribution:

$$p(y=c|x) = \int \underbrace{p(y=c|x,\theta)}_{aleatoric}\,\underbrace{p(\theta|D)}_{epistemic}\,d\theta. \qquad (1)$$

Truly Bayesian Neural Networks are computationally infeasible, so instead various methods to approximate it have been proposed [1, 8]. We will be considering MC-Dropout [10], MC-DropConnect [11], Deep Ensembles [12] and Flipout [13].

While they have differences in approximation quality, implementation complexity, and computational cost, they all rely on BNN fundamentals.

Deterministic Uncertainty Quantification (DUQ): DUQ [14] uses a different approach to Uncertainty Quantification in Neural Networks. DUQ uses a standard Neural Network as a feature extractor, and then learns a centroid for each class. Samples that are far away from the centroids are deemed uncertain, whereas samples that are close to a centroid are deemed certain.

This different underlying assumption of how uncertainty should arise is inspired by generative models, though DUQ is not actually a generative model. A true generative model models the distribution of the training samples directly, whereas DUQ only models class centroids. Still, this makes it fundamentally different from the discriminative BNNs, and may therefore give different results than the BNN approach. It also means that aleatoric and epistemic uncertainty cannot be clearly distinguished, but they are both included in the predicted uncertainty.

METHODS

Dataset: We used the public Motor Imagery dataset: BCI Competition IV, dataset 2a [15]. This dataset contains 22 channel EEG and 3 monopolar EOG channel recordings of 9 subjects performing one of 4 different motor imagery tasks— left hand (class 1), right hand (class 2), both feet (class 3) and tongue (class 4).

The sampling rate was 250Hz and the dataset comes preapplied with a 50Hz notch filter and a bandpass filter of 0.5Hz to 100Hz.

The Braindecode [16] and MNE [17] Python libraries were used to load and pre-process the data.

The training setup (shown for a single subject in figure 2) was designed to allow the observation of aleatoric uncertainty and the combination of aleatoric and epistemic uncertainty. This allows the impact of epistemic uncertainty to be observed in isolation.

We used leave-one-subject-out cross-validation with a slight variation. Normally leave-one-subject-out involves splitting $N-1$ subjects into a training set and leaving the last subject as the out-of-population (cross-subject) set. Our variation to this procedure is as follows: 10% of the data from each training set subject is used as a within-population test set[*]. This within-population dataset allows for an observation with minimal epistemic uncertainty, and comparing it to the cross-subject set allows us to isolate the impact of cross-subject generalisation.

Preprocessing: Some EEG pipelines employ extensive signal processing and feature extraction in order to operate with ML algorithms. However, it is often unclear what value each processing step introduces, and various

[*] The remaining training data was in turn split into 90% train and 10% validation for hyperparameter optimisation

10.3217/978-3-99161-014-4-016

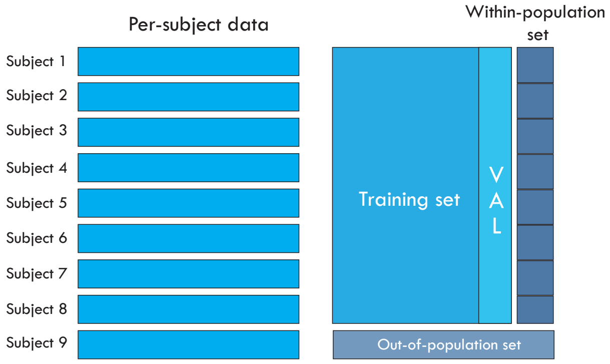


Figure 2: Training setup for a single subject. One subject is excluded and used as an out-of-population set while the other 10% of the data from each subject is separated into a within-population set. The data of the remaining subjects are concatenated and split 90-10 into a training and validation set. This procedure is repeated for every subject.

researchers and labs use different pipelines. The use of CNNs (and DL methods in general) in EEG is promising because of their ability to automatically extract features from raw data and perform classifications, with minimal preprocessing required [18, 19].

Hence the following preprocessing steps are very minimal. It consists of: dropping the EOG channels, converting the EEG signals from volts to microvolts (μV), applying an exponential moving standardisation with parameters described by [16] and epoching from 0.5 seconds before the trial cue at $t = 2s$ to end of the trial at $t = 6s$ (for a total trial window of 4.5 seconds). Creating epochs as such leads to a single trial being a matrix (C, T) with $C = 22$ being the number of channels and $T = 1125$ being the number of timestamps.

Model Architecture: We used Keras [20] to implement the Shallow ConvNet CNN [16], and the Keras Uncertainty library [21] to implement the UQ adaptations. [*]

Although all UQ methods followed the same Shallow ConvNet architecture, minor differences existed in the implementation of the UQ layers. Two standard models regularised with Dropout and DropConnect were used as baselines.

MC-Dropout and MC-DropConnect and their standard counterparts both had only a single UQ layer. In MC-Dropout this layer was positioned before the dense classification layer with a drop rate of 0.2. In MC-DropConnect it was positioned after the second convolutional layer with a drop rate of 0.1. A grid search was done on a single subject (due to computational complexity) to decide this configuration. Normal Dropout and DropConnect sets the value of a node or weight to 0 during training. The equivalent UQ versions retain this during testing, resulting in slightly different predictions each forward pass, thereby representing epistemic uncertainty.

The Ensemble model simply consisted of 10 standard Shallow ConvNet CNNs, regularised with dropout identical to the dropout baseline model. Disagreement between these 10 models represents epistemic uncertainty.

Flipout changes the final dense classification layer to a standard dense layer using ReLU activation with 10 units, following which are two flipout layers. Both flipout layers use a prior $P(\theta) = \mathcal{N}(0, 1.0^2) + \pi\, \mathcal{N}(0, 2.5^2)$ with $\pi = 0.1$. Additionally, the first flipout layer had 10 units. Both sets of parameters were determined using a grid search.

MC-Dropout, MC-Dropconnect and Flipout are stochastic during inference. Therefore, a number of forward passes T needs to be selected. T was chosen to be 50 as it has been found to be point where the improvement in accuracy stabilises [3].

DUQ changes the final layer of the Shallow ConvNet CNN to a dense layer with 100 units using a ReLU activation, following which is an RBF classification layer with a length scale of 0.4 with trainable centroids of dimension 100. These parameters were found using a grid search. Additionally, compared to the categorical cross-entropy loss used by the other methods, DUQ utilizes binary cross-entropy.

Other hyperparameters follow common practice in Deep Learning literature. Specifically we set the learning rate (1×10^{-4}), loss function (categorical cross entropy), and optimiser (Adam).

Uncertainty Measures: The BNN-based methods rely on T forward passes from a stochastic model. Each forward pass predicts class probabilities p_c, resulting in a distribution over probabilities. To this we can apply various Uncertainty Measures to measure either aleatoric uncertainty, epistemic uncertainty or the total uncertainty [8].

The total uncertainty is based on the mean of the predicted probability for each class and is measured by the Predictive Entropy:

$$\mathbb{H}_{\text{pred}}(p) = -\sum_c \bar{p}_c \log \bar{p}_c. \tag{2}$$

The Expected Entropy first determines the uncertainty of each forward pass, and then takes the average over those uncertainties.

$$\mathbb{H}_{\mathbb{E}}(p) = -T^{-1} \sum_t \sum_c p_{ct} \log p_{ct} \tag{3}$$

In this approach, Expected Entropy takes the "average uncertainty" of each individual model. As such, it only corresponds to aleatoric uncertainty [22].

Lastly, subtracting the aleatoric uncertainty from the total uncertainty results in the remaining epistemic uncertainty. This measure is referred to as Mutual Information [23]:

$$\mathbb{I}(p) \approx \mathbb{H}_{\text{pred}}(p) - \mathbb{H}_E(p) \tag{4}$$

Predictive Entropy and Expected entropy may be applied to a standard Neural Network, but they will result in the same prediction. This approximation for Mutual Information cannot be applied to standard Neural Networks.

[*] All code is available at `https://github.com/p-manivannan/UQ-Motor-Imagery`

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-016

Because DUQ does not follow the same discriminative assumptions for uncertainty, these measures of uncertainty do not apply. Instead, it gives a single uncertainty measure that responds to both aleatoric and epistemic uncertainty.

RESULTS

Classification accuracy for each method is given in table 1. It can be seen that performance is higher within-population than out-of-population, with Ensembles outperforming all other methods for both groups. The performance of the Ensemble is in-line with benchmarks for out-of-population and within-population accuracies [19] while the other methods are slightly underperforming.

To find out whether UQ can improve performance, uncertainty estimation was treated as a binary classification task, where the aim was to classify wrong predictions as uncertain. Therefore, the Area Under the ROC curve (AUROC) is considered as a performance metric [24]. Note that this can never approach 100 as the uncertain samples are "guessed", which will be correct 25% of the time. These would be labelled as false positives in this framework.

This metric is chosen in place of common metrics like Expected Calibration Error [7], because our goal is to detect misclassifications, whereas ECE aims to detect overconfidence or underconfidence.

The uncertainty AUROC scores for each method and each uncertainty measure on the within-population set is given in table 2a. This table shows that Mutual Information (which corresponds only to epistemic uncertainty) performs the worst. Predictive Entropy and Expected Entropy perform similarly, suggesting that the modelling of epistemic uncertainty is not beneficial to the uncertainty estimation. It also shows that DUQ has the worst uncertainty estimation, and that most discriminative models show similar performance.

Table 2b shows the performance of uncertainty estimation on the out-of-population dataset. The performance of uncertainty estimation is consistently lower here than on the within-population set. Mutual Information, which represents epistemic uncertainty, still does not offer better uncertainty estimation. This suggests that none of the available models are able to fully account for the epistemic uncertainty introduced by cross-subject classification. We again see that DUQ has noticeably worse UQ performance.

It can be seen that the quality of uncertainty estimation is worse cross-population than within-population. This behaviour is inevitable for measures of aleatoric uncertainty, but measures of epistemic uncertainty should be more robust to this [25].

When predictive entropy is disentangled into aleatoric and epistemic uncertainty, it can be seen that epistemic uncertainty based thresholding is consistently slightly worse than aleatoric uncertainty based thresholding. This suggests either that aleatoric uncertainty is more prevalent than epistemic uncertainty, or that epistemic uncertainty is not captured well by the models. Since the accuracy does go down when moving to cross-population, it is clear that there must be an increase in epistemic uncertainty which the models are not accounting for.

Table 1: Mean accuracy per subject for each method. Within-population accuracy is higher overall than cross population accuracy, with ensembles outperforming other methods in both categories. Standard DropConnect performs noticeably worse, but most methods perform similar to Standard Dropout.

Method	Within pop. Acc%	Cross pop. Acc %
Dropout	68.98 ± 2.73	55.54 ± 7.95
MC-Dropout	69.00 ± 2.73	55.56 ± 7.94
DropConnect	66.67 ± 2.23	53.51 ± 11.67
MC-DropConnect	69.27 ± 1.34	54.96 ± 9.76
Flipout	69.90 ± 2.55	54.99 ± 8.67
Ensembles	**73.05 ± 2.22**	**59.05 ± 8.11**
DUQ	70.47 ± 2.93	55.42 ± 9.16

It can be seen that no BNN method is substantially better than another at uncertainty quantification. Only DUQ performs substantially worse than other methods, performing even lower than standard neural networks.

DISCUSSION

Surprisingly, we find that the specific UQ methods designed to observe epistemic uncertainty are not able give better uncertainty estimations than a similar Neural Network with Softmax activation. It is still possible for all methods to reject some of the uncertain samples to increase accuracy, but this is trivial.

A possible reason for this is that since aleatoric uncertainty seems more prevalent, the ability of these UQ methods to take into account epistemic uncertainty does not help, hence explaining how standard models are able to achieve comparable performance. However, it is clear that the decrease in accuracy should be attributable to an increase in epistemic uncertainty. This could be caused by how these methods model uncertainty, but the results show that the discriminative models and DUQ suffer the same problems.

Relation to background: Our findings contradict the expectation that cross-subject classification should introduce epistemic uncertainty, and that therefore BNNs should perform better.

Epistemic uncertainty should arise when a model is tested on data that is different from the data it was trained on. In this case, the cross-subject testing samples are different from the data that the model is trained on, but the models capturing epistemic uncertainty were not able to offer better uncertainty estimates.

It is difficult to attribute this to problems with a specific approximation of BNNs, as a variety of approximations show this effect consistently. We also cannot attribute this to flaws in the discriminative model as shown in Figure 1a, because this problem is consistent even when using DUQ which has a fundamentally different assumption of uncertainty.

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-016

Table 2: Uncertainty AUROC scores for each method both within-population and out-of-population. Predictive Entropy and Expected Entropy both perform equally well for all BNN models. At the same time Mutual Information performs noticeably worse, and shows more difference for the different models. The uncertainty of all models and all uncertainty measures is consistently worse when moving out-of-population.

(a) Within-population

Method	Predictive Entropy (Ale+Epi)	Expected Entropy (Ale)	Mutual Information (Epi)
Standard Dropout	76.07 ± 2.918	76.07 ± 2.918	-
MC-Dropout	76.07 ± 2.927	76.06 ± 2.927	74.24 ± 3.398
Standard DropConnect	75.44 ± 2.691	75.44 ± 2.691	-
MC-DropConnect	75.3 ± 3.303	75.29 ± 3.297	73.33 ± 2.835
Flipout	75.56 ± 2.461	76.70 ± 2.460	70.56 ± 2.488
Ensembles	76.92 ± 2.868	76.66 ± 3.046	70.02 ± 2.064
DUQ	73.19 ± 2.379	-	-

(b) Out-of-population

Method	Predictive Entropy (Ale+Epi)	Expected Entropy (Ale)	Mutual Information (Epi)
Standard Dropout	67.46 ± 4.646	67.46 ± 4.646	-
MC-Dropout	67.43 ± 4.611	67.43 ± 4.611	66.6 ± 4.164
Standard DropConnect	68.23 ± 4.532	68.23 ± 4.532	-
MC-DropConnect	68.48 ± 4.625	68.48 ± 4.626	66.82 ± 5.311
Flipout	67.79 ± 5.156	67.79 ± 5.152	63.95 ± 4.024
Ensembles	67.39 ± 5.446	67.29 ± 5.564	63.86 ± 4.354
DUQ	65.30 ± 4.01	-	-

The previous studies in this direction [3, 5] show more positive findings for approximations of BNNs, but by considering an equivalent CNN and using Softmax as a baseline we were able to that those results can also be achieved with simpler methods.

Limitations: Our study also only focuses on the use of uncertainty for rejecting difficult samples, and does not actively look at the absolute epistemic uncertainty. It may be that the epistemic uncertainty did increase for cross-subject samples, but if this happens uniformly for a given subject we are not able to capture it. This does not affect the validity of the findings, but does make it harder to know why these Bayesian Neural Networks are not performing well.

There may also be limitations underlying how Predictive Uncertainty is disentangled into aleatoric and epistemic uncertainty. The proposed approach follows a line of existing work [22, 23], but there is also a line of work that assumes an entirely different formulation for disentangling uncertainty [26, 27]. There, the BNNs have two outputs. One for predicting the prediction, and one for the variance. The mean of the variances is then the aleatoric uncertainty, and the variance of the predictions is then epistemic uncertainty. This approach explicitly models aleatoric and epistemic as part of the model, which may give more favourable results.

Directions for future research: We showed that UQ did not work to reject the cross-subject samples with the most epistemic uncertainty. However, it may still be usable for deciding whether or not to make a prediction under noisy EEG, or for identifying a model well suited for a certain subject, or even for detecting off-task thoughts.

CONCLUSION

Available Deep Learning methods that capture and disentangle epistemic uncertainty are not able to improve the robustness of within-subject nor cross-subject Motor Imagery BCIs in the context of a benchmark dataset. However, there are other contexts in BCIs where epistemic uncertainty may be expected. Off-task thoughts, rare artifacts, or insufficient training data can all introduce epistemic uncertainty, and the methods demonstrated here may be able to improve robustness in those cases. This has not yet been investigated.

We want to emphasise the need to study the behaviour and uses of uncertainty estimates from non-Deep Learning models. Classical Machine Learning models for classification often come with an adaptation to return class probabilities, but the behaviour of these may vary substantially. Assessing their ability to reject segments of EEG that are likely to be false positives may allow for more robust BCI systems. The robustness promised by good UQ may be a step towards making BCIs more usable outside of the lab.

REFERENCES

[1] M. Abdar *et al.*, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Information Fusion*, vol. 76, pp. 243–297, 2021.

[2] T. Duan, Z. Wang, S. Liu, Y. Yin, and S. N. Srihari, “Uncer: A framework for uncertainty estimation and reduction in neural decoding of eeg signals,” *Neurocomputing*, vol. 538, p. 126 210, 2023.

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-016

[3] D. Milanés-Hermosilla *et al.*, "Monte carlo dropout for uncertainty estimation and motor imagery classification," *Sensors*, vol. 21, no. 21, 2021. [Online]. Available: `https://www.mdpi.com/1424-8220/21/21/7241`.

[4] E. I. Chetkin, S. L. Shishkin, and B. L. Kozyrskiy, "Bayesian opportunities for brain–computer interfaces: Enhancement of the existing classification algorithms and out-of-domain detection," *Algorithms*, vol. 16, no. 9, p. 429, 2023.

[5] D. Milanés-Hermosilla *et al.*, "Robust motor imagery tasks classification approach using bayesian neural network," *Sensors*, vol. 23, no. 2, p. 703, 2023.

[6] A. Campbell, L. Qendro, P. Liò, and C. Mascolo, "Robust and efficient uncertainty aware biosignal classification via early exit ensembles," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3998–4002.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1321–1330.

[8] I. P. de Jong, A. I. Sburlea, and M. Valdenegro-Toro, "Uncertainty quantification in machine learning for biosignal applications–a review," *arXiv preprint arXiv:2312.09454*, 2023.

[9] S. J. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[10] Y. Gal and Z. Ghahramani, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*, 2016. arXiv: `1506.02142 [stat.ML]`.

[11] A. Mobiny, H. V. Nguyen, S. Moulik, N. Garg, and C. C. Wu, *Dropconnect is effective in modeling uncertainty of bayesian deep networks*, 2019. arXiv: `1906.04569 [cs.LG]`.

[12] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, "Flipout: Efficient pseudo-independent weight perturbations on mini-batches," *arXiv preprint arXiv:1803.04386*, 2018.

[14] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International conference on machine learning*, PMLR, 2020, pp. 9690–9700.

[15] M. Tangermann *et al.*, "Review of the bci competition iv," *Frontiers in Neuroscience*, vol. 6, 2012.

[16] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, 2017. [Online]. Available: `http://dx.doi.org/10.1002/hbm.23730`.

[17] A. Gramfort *et al.*, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, p. 267, 2013.

[18] N. Tibrewal, N. Leeuwis, and M. Alimardani, "The promise of deep learning for bcis: Classification of motor imagery eeg using convolutional neural network," *bioRxiv*, 2021. eprint: `https://www.biorxiv.org/content/early/2021/06/18/2021.06.18.448960.full.pdf`.

[19] A. Zancanaro, G. Cisotto, J. R. Paulo, G. Pires, and U. J. Nunes, "Cnn-based approaches for cross-subject classification in motor imagery: From the state-of-the-art to dynamicnet," in *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2021, pp. 1–7.

[20] F. Chollet *et al.*, *Keras*, `https://keras.io`, 2015.

[21] M. Valdenegro, *Keras-uncertainty*, `https://github.com/mvaldenegro/keras-uncertainty`, 2023.

[22] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty," *arXiv preprint arXiv:2102.11582*, vol. 2, 2021.

[23] L. Smith and Y. Gal, *Understanding measures of uncertainty for adversarial example detection*, 2018. arXiv: `1803.08533 [stat.ML]`.

[24] X. Huang, J. Yang, L. Li, H. Deng, B. Ni, and Y. Xu, "Evaluating and boosting uncertainty quantification in classification," *arXiv preprint arXiv:1909.06030*, 2019.

[25] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty: A simple baseline," *arXiv preprint arXiv:2102.11582*, 2021.

[26] M. Valdenegro-Toro and D. S. Mori, "A deeper look into aleatoric and epistemic uncertainty disentanglement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2022, pp. 1508–1516.

[27] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

Published by
Verlag der Technischen Universität Graz