Uncertainty Quantification for Motor Imagery BCI Machine Learning vs. Deep Learning

Joris Suurmeijer   Ivo Pascal de Jong   Matias Valdenegro-Toro   Andreea Ioana Sburlea

Department of Artificial Intelligence, Bernoulli Institute – University of Groningen

university of groningen

Introduction

We investigate class probabilities, i.e. the uncertainty/confidence of a prediction.

- Confidence should be well-calibrated, classification 75% confidence should be correct 75% of the time.
- We want to avoid being overconfident or underconfident.
- Confidence should be able to separate between knowing and guessing.
- Good uncertainty can reject bad samples, preventing unwanted commands being sent to a device.

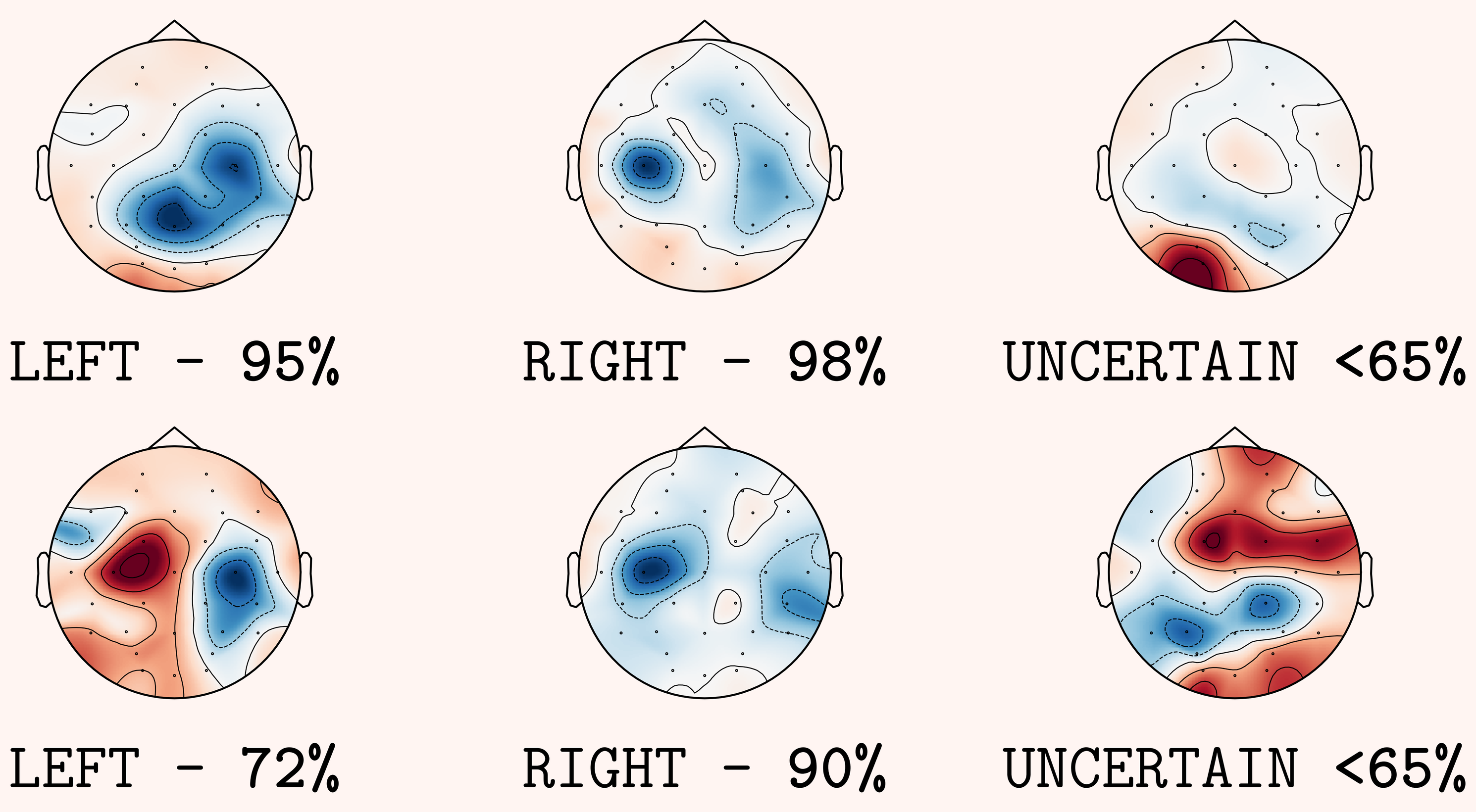

Figure 1. Event Related Desynchronisation (blue) and Synchronisation (red) trials in left vs. right hand motor imagery classification, with predictions and probabilities. Models should be uncertain when EEG is too noisy.

Models

- MDRM calculates Riemannian distance of covariance matrices to class means. Softmax turns this into probabilities.
- We use a ShallowConvNet CNN as Deep Learning method.
- Deep Ensembles (DE) and Direct Uncertainty Quantification (DUQ) are specialised Uncertainty Quantification methods for Deep Learning.

Temperature Scaling – MDRM-T

MDRM is underconfident, as shown in Figure 2a.

We solve this by adding Temperature Scaling. By scaling up the distances, the probabilities get pushed towards extremes (0, 1).

$$\hat{y} = \frac{\exp(-d_i^2/T)}{\sum_j \exp(-d_j^2/T)} \quad (1)$$

Here d_i represents the distance to a class mean, $\hat{y}$ is the class probability, and T is the Temperature parameter that is optimized after training the model.

Optimal T minimizes the Expected Calibration Error.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \quad (2)$$

where B_m are predictions binned by confidence.

MDRM-T makes uncertainties well-calibrated

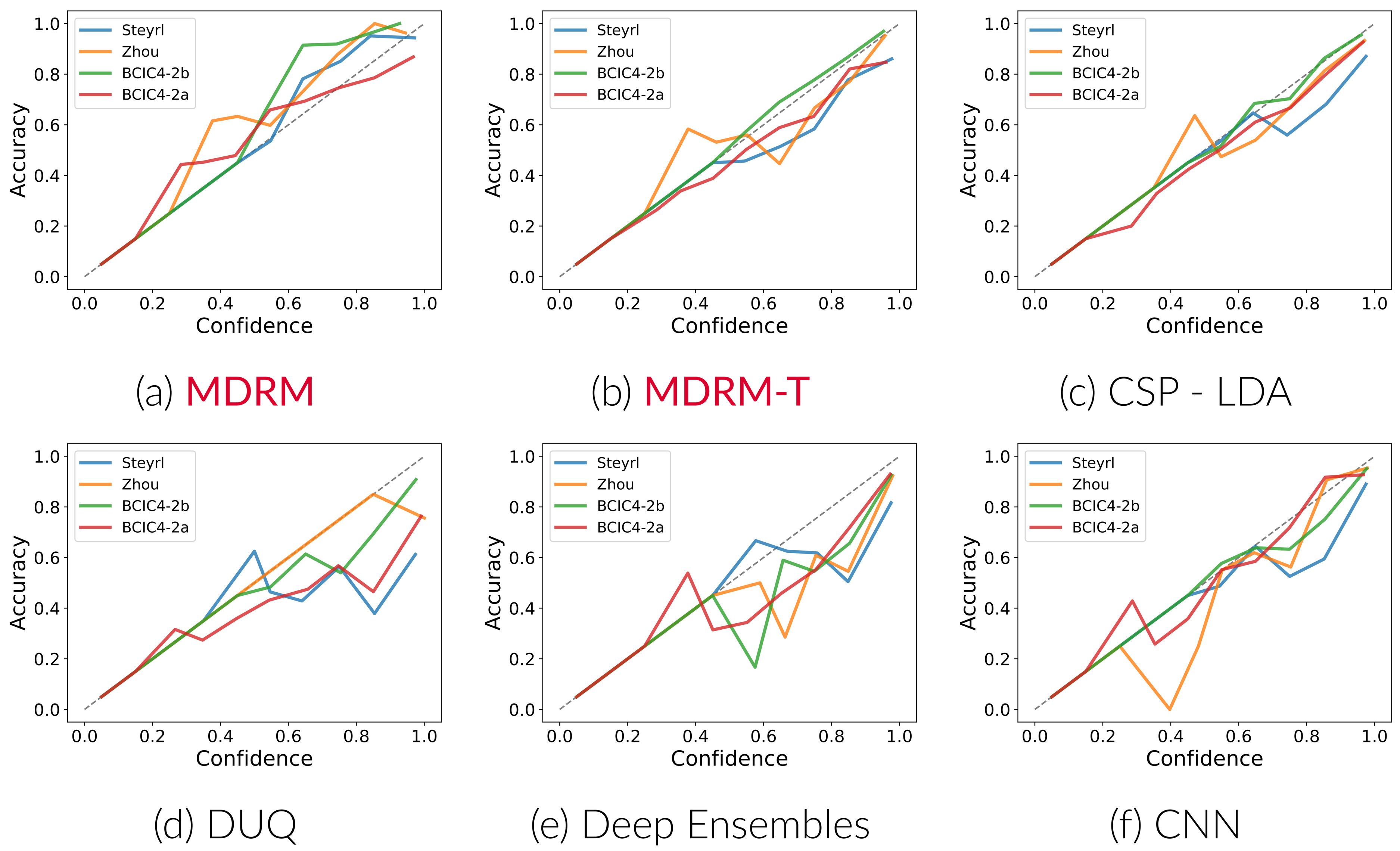

Figure 2. Uncertainty calibration plots. Perfect calibration has confidence (x) equal accuracy (y), following the diagonal line. Deep Learning (d-f) is overconfident, MDRM (a) is underconfident. MDRM-T (b) and CSP-LDA (c) are well-calibrated.

Further Results

Metric	Dataset	MDRM	MDRM-T	CSP-LDA	DUQ	DE	CNN
Acc. %↑	Steyrl	70.3%	70.3%	75.9%	51.3%	71.0%	70.8%
	Zhou	72.3%	72.3%	77.6%	76.7%	83.0%	84.3%
	BCIC4-2b	71.4%	71.4%	72.7%	77.1%	79.1%	79.0%
	BCIC4-2a	58.2%	58.2%	66.5%	56.8%	72.3%	71.8%
ECE ↓	Steyrl	0.163	0.155	0.231	0.257	0.276	0.214
	Zhou	0.163	0.148	0.122	0.233	0.264	0.202
	BCIC4-2b	0.186	0.066	0.074	0.164	0.209	0.121
	BCIC4-2a	0.156	0.146	0.136	0.265	0.198	0.164

Table 1. Within-subject performances. Highlighted scores indicates the best performing model for a metric and dataset. Deep Learning gives better classifications (Acc.), but MDRM-T and CSP-LDA give better uncertainties (ECE).

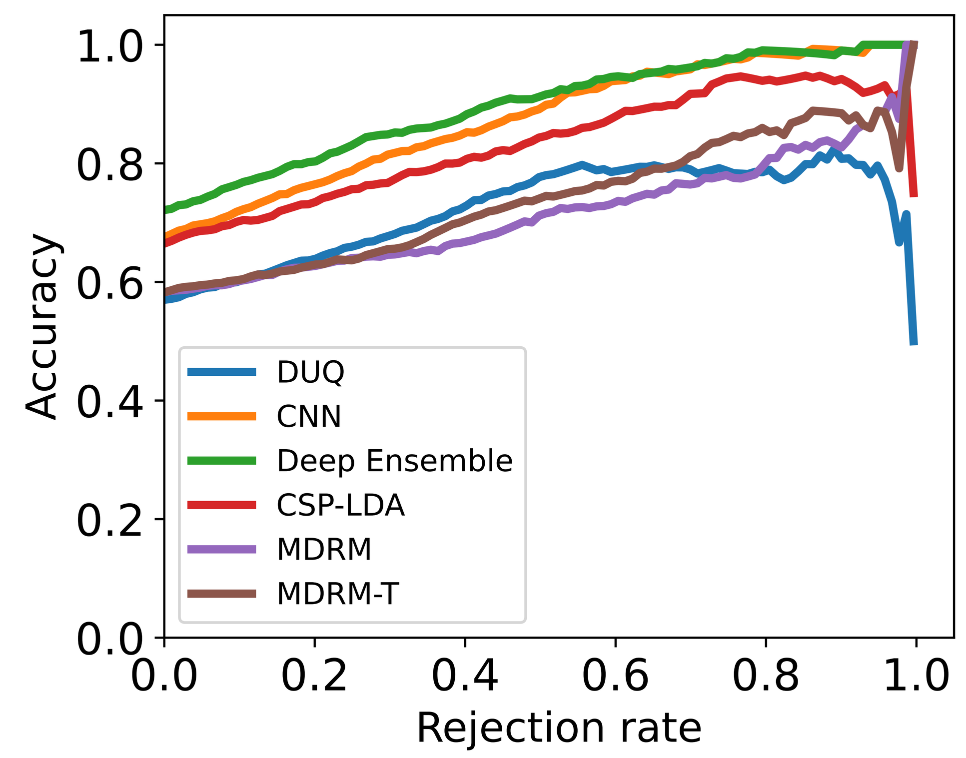

Figure 3. Rejection rate (x) and accuracy (y), calculated on remaining samples. Accuracy is improved by rejecting uncertain samples.

Conclusions

- MDRM is underconfident.
- MDRM-T improves calibration with Temperature Scaling, and does not affect accuracy.
- Well designed BCI models have better uncertainty calibration than Deep Learning models.
- Models are able to improve accuracy by rejecting difficult samples using uncertainty.