Beyond temperature scaling:
Obtaining well-calibrated multiclass probabilities with Dirichlet calibration

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, Peter Flach

University of Tartu · University of Bristol · sphere · The Alan Turing Institute · Universidade Federal da Paraíba ESTATÍSTICA

Class probabilities predicted by most multiclass classifiers are uncalibrated, often tending towards over-confidence. With neural networks, calibration can be improved by temperature scaling, a method to learn a single corrective multiplicative factor for inputs to the last softmax layer. On non-neural models the existing methods apply binary calibration in a pairwise or one-vs-rest fashion. We propose a natively multiclass calibration method applicable to classifiers from any model class, derived from Dirichlet distributions and generalising the beta calibration method from binary classification. It is easily implemented with neural nets since it is equivalent to log-transforming the uncalibrated probabilities, followed by one linear layer and softmax.

Contributions

- **Dirichlet calibration**:
  - Parametric multi-class calibration method
  - General-purpose (acts in class probability space)
  - Easy to implement as a neural layer or as multinomial logistic regression on log-transformed class probabilities

	Logit space	Class probability space
Binary classification	Derived from Gaussian distribution Platt scaling[1]	Derived from Beta distribution Beta calibration[2] (+ constrained variants)
Multi-class classification	Matrix scaling[3] (+ vector scaling, + temperature scaling)	Derived from Dirichlet distribution Dirichlet calibration (+ constrained variants)

- **ODIR** (Off-Diagonal and Intercept Regularisation):
  - A new regularization method for Dirichlet calibration and matrix scaling
- Clarifications in calibration evaluation of multi-class classifiers.

Is my multiclass classifier calibrated?

Multiclass classifer: $\hat{\mathbf{p}}(X) = (\hat{p}_1(X), \dots, \hat{p}_k(X)) \in \Delta_k \subset [0,1]^k$

Actual class: $Y \in \{1, \dots, k\}$

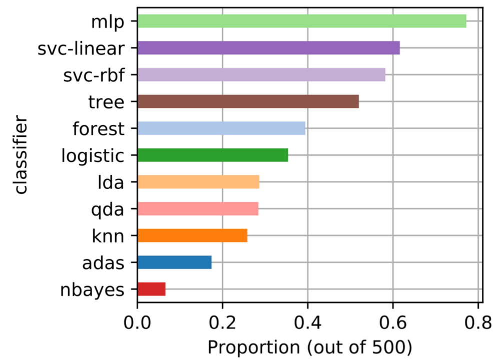
Multiclass-calibrated: $P(Y = i \mid \hat{\mathbf{p}}(X) = \mathbf{q}) = q_i$ for $\mathbf{q} \in \Delta_k;\ i = 1, \dots, k$

Classwise-calibrated: $P(Y = i \mid \hat{p}_i(X) = q_i) = q_i$ for $q_i \in [0,1];\ i = 1, \dots, k$

Confidence-calibrated: $P(Y = \arg\max \hat{\mathbf{p}}(X) \mid \max \hat{\mathbf{p}}(X) = c) = c$ for $c \in [0,1]$

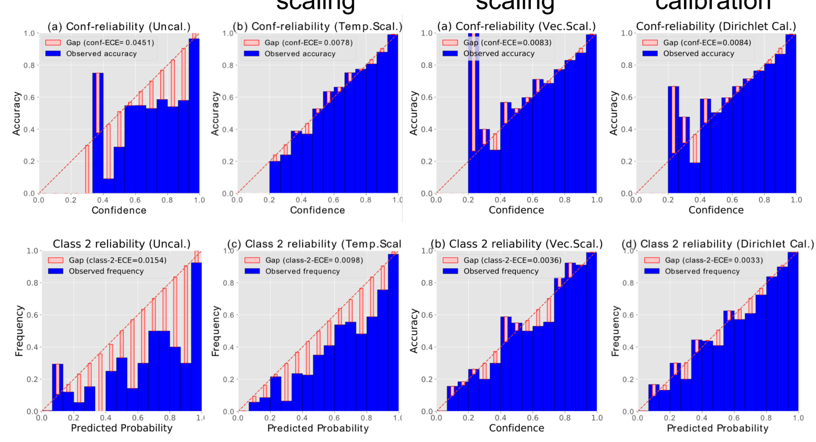
How often are classifiers classwise-calibrated?

The proportion of cases where the null hypothesis that the model is classwise-calibrated was accepted, across 20 datasets (5x5-fold cross-validation on each):



Example on a neural network

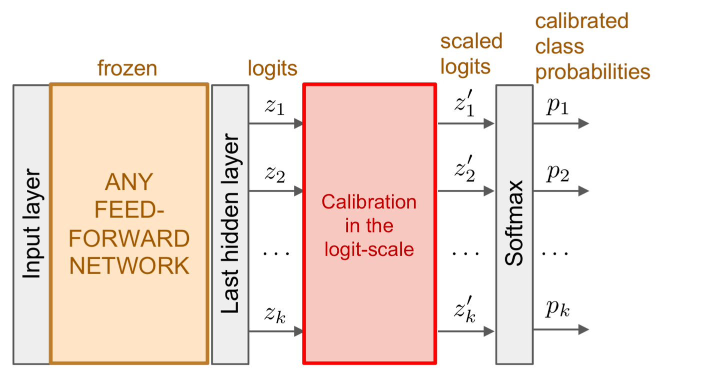
ResNet Wide 32 network trained on CIFAR 10:

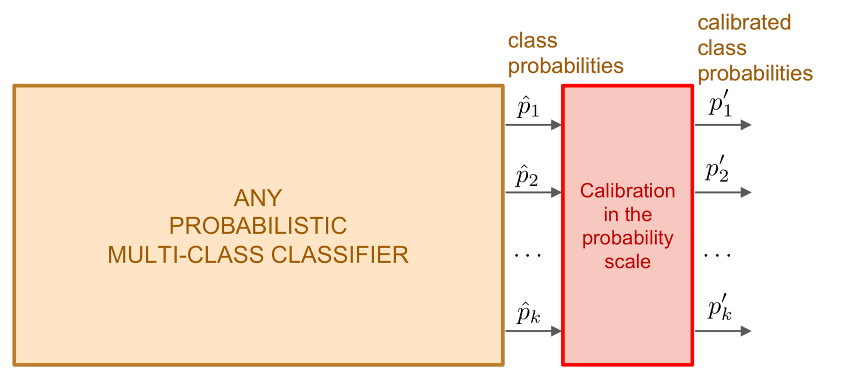


[1] J. Platt. Probabilities for SV machines. In *Advances in Large Margin Classifiers, pages 61–74, MIT Press*, 2000.
[2] M. Kull, T. Silva Filho, P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. AISTATS 2017
[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. ICML 2017

How to calibrate a multiclass classifier:

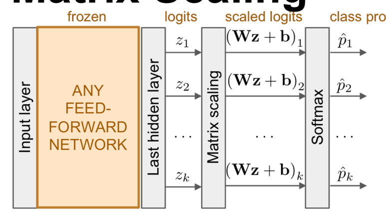
1. Choose logit-space or class probability space





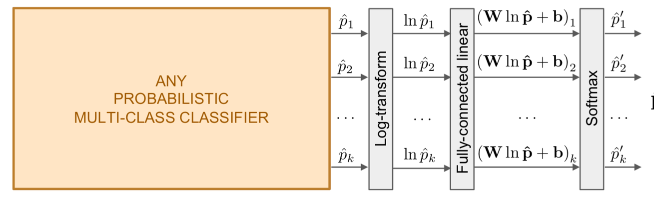
2. Choose a calibration map family

Matrix Scaling



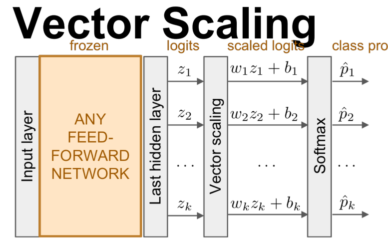
$$\hat{\mathbf{p}} = \sigma\left(\begin{pmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k1} & w_{k2} & \dots & w_{kk} \end{pmatrix} \mathbf{z} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}\right)$$

Dirichlet Calibration = matrix scaling on pseudo-logits



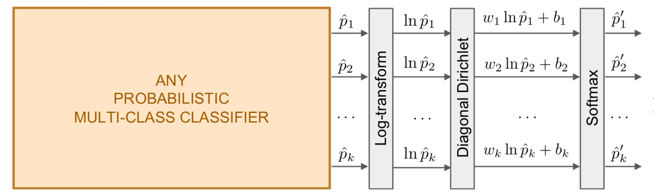
$$\hat{\mathbf{p}}' = \sigma\left(\begin{pmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k1} & w_{k2} & \dots & w_{kk} \end{pmatrix} \ln \hat{\mathbf{p}} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}\right)$$

Vector Scaling



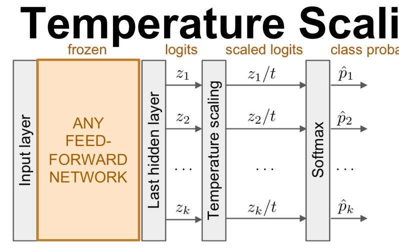
$$\hat{\mathbf{p}} = \sigma\left(\begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_k \end{pmatrix} \mathbf{z} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}\right)$$

Diagonal Dirichlet Cal. = vector scaling on pseudo-logits



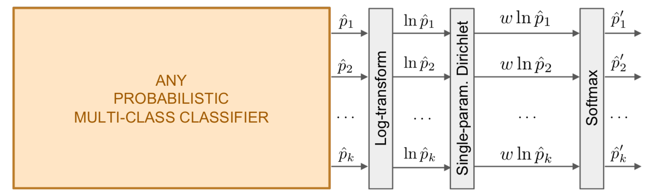
$$\hat{\mathbf{p}}' = \sigma\left(\begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_k \end{pmatrix} \ln \hat{\mathbf{p}} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}\right)$$

Temperature Scaling



$$\hat{\mathbf{p}} = \sigma\left(\begin{pmatrix} 1/t & 0 & \dots & 0 \\ 0 & 1/t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/t \end{pmatrix} \mathbf{z} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right)$$

Single-param. Dirichlet Cal. = temp. scaling on pseudo-logits



$$\hat{\mathbf{p}}' = \sigma\left(\begin{pmatrix} w & 0 & \dots & 0 \\ 0 & w & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w \end{pmatrix} \ln \hat{\mathbf{p}} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right)$$

3. Fit the calibration map

by minimising cross-entropy on the validation data and optionally regularise (L2 or ODIR)

ODIR = Off-Diagonal and Intercept regularisation

$$L = \frac{1}{n}\sum_{i=1}^{n} logloss\left(\hat{\mu}_{DirLin}(\hat{\mathbf{p}}(\mathbf{x}_i); \mathbf{W}, \mathbf{b}), y_i\right) + \lambda \cdot \left(\frac{1}{k(k-1)}\sum_{i \neq j} w_{ij}^2\right) + \mu \cdot \left(\frac{1}{k}\sum_j b_j^2\right)$$

Derivations of calibration maps

$$\mathbf{z}(X) \mid Y{=}j \sim \text{Gaussian}(\mu^{(j)}, \sigma^2) \longrightarrow \text{Platt scaling}$$
$$\hat{\mathbf{p}}(X) \mid Y{=}j \sim \text{Beta}(\alpha^{(j)}, \beta^{(j)}) \longrightarrow \text{Beta calibration}$$
$$\hat{\mathbf{p}}(X) \mid Y{=}j \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(j)}) \longrightarrow \text{Dirichlet calibration}$$

Parametrisations of Dirichlet calibration maps

generative parametrisation: $\mu_{DirGen}(\mathbf{q}; \boldsymbol{\alpha}, \boldsymbol{\pi}) = (\pi_1 f_{Dir(\boldsymbol{\alpha}^{(1)})}(\mathbf{q}), \dots, \pi_k f_{Dir(\boldsymbol{\alpha}^{(k)})}(\mathbf{q}))/z$

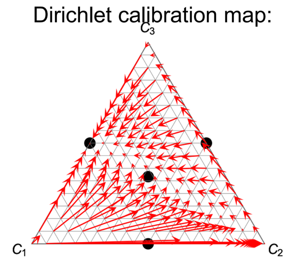
linear parametrisation: $\mu_{DirLin}(\mathbf{q}; \mathbf{W}, \mathbf{b}) = \boldsymbol{\sigma}(\mathbf{W} \ln \mathbf{q} + \mathbf{b})$

canonical parametrisation: $\mu_{Dir}(\mathbf{q}; \mathbf{A}, \mathbf{c}) = \boldsymbol{\sigma}(\mathbf{A} \ln \frac{\mathbf{q}}{1/k} + \ln \mathbf{c})$

Interpretation of Dirichlet calibration maps

MLP on abalone dataset

Dirichlet calibration map:



Canonical parametrization:

Class	1	2	3	c
1	0.54	0.00	0.50	0.32
2	0.04	1.10	0.00	0.32
3	0.00	0.01	0.94	0.36

ConvNet on SVHN dataset

canonical parametrization:

	0	1	2	3	4	5	6	7	8	9	c
0	2.16	0.00	0.21	0.25	0.13	0.37	0.00	0.23	0.00	0.16	0.05
1	0.15	2.13	0.39	0.01	0.12	0.18	0.17	0.00	0.04	0.21	0.02
2	0.12	0.09	1.82	0.07	0.00	0.00	0.09	0.06	0.63	0.09	0.10
3	0.17	0.14	0.24	2.12	0.08	0.13	0.22	0.01	0.21	0.08	0.02
4	0.00	0.09	0.10	0.00	1.35	0.22	0.12	0.15	0.23	0.27	0.23
5	0.10	0.10	0.12	0.22	0.05	1.94	0.14	0.17	0.03	0.15	0.14
6	0.27	0.08	0.03	0.23	0.09	0.26	2.02	0.05	0.05	0.13	0.05
7	0.12	0.11	0.08	0.14	0.17	0.27	0.15	1.89	0.23	0.00	0.08
8	0.07	0.19	0.00	0.10	0.09	0.26	0.05	0.10	1.88	0.33	0.13
9	0.16	0.08	0.06	0.08	0.05	0.29	0.24	0.09	0.05	1.64	0.17

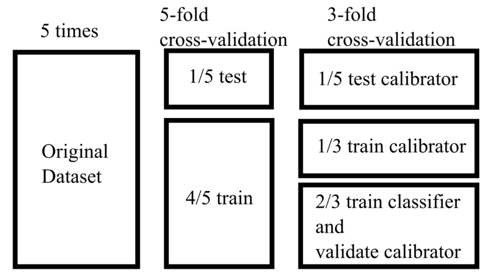
changes to the confusion matrix after applying Dirichlet calibration map:

	0	1	2	3	4	5	6	7	8	9
0	18.00	-1.00	2.00	-8.00	4.00	0.00	18.00	1.00	0.00	2.00
1	-6.00	-16.00	11.00	1.00	1.00	0.00	3.00	3.00	1.00	2.00
2	-12.00	-12.00	50.00	-3.00	0.00	1.00	0.00	1.00	-26.00	1.00
3	9.00	-4.00	3.00	-14.00	3.00	0.00	0.00	0.00	2.00	1.00
4	-1.00	-11.00	-6.00	-1.00	16.00	7.00	0.00	-4.00	-4.00	4.00
5	-3.00	-5.00	-3.00	-7.00	1.00	19.00	0.00	-7.00	1.00	4.00
6	-24.00	-9.00	2.00	-1.00	1.00	0.00	28.00	1.00	2.00	0.00
7	4.00	-10.00	0.00	-6.00	3.00	4.00	0.00	-3.00	6.00	2.00
8	1.00	-34.00	17.00	-5.00	3.00	1.00	2.00	-2.00	19.00	-2.00
9	-12.00	-10.00	2.00	-2.00	1.00	1.00	1.00	-7.00	6.00	20.00

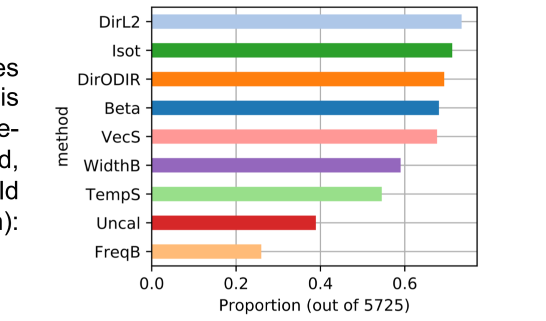
Non-neural experiment

21 UCI datasets and 11 sklearn classifiers = 231 settings

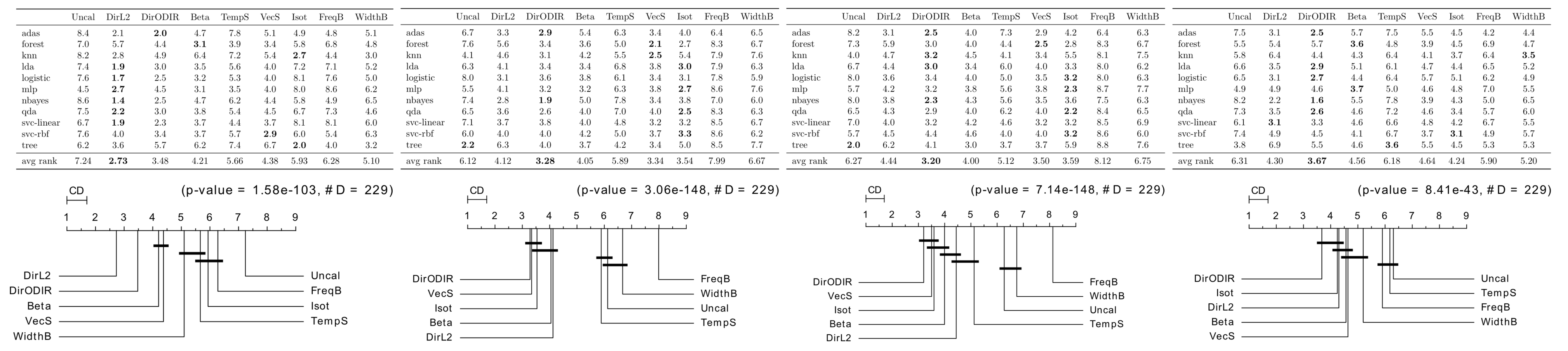
(actually 229 as Naive Bayes and QDA ran too long on *shuttle*)



The proportion of cases where the null hypothesis that the model is classwise-calibrated was accepted, across 229 settings (5x5-fold cross-validation on each):



Log-loss (average ranks) · **Classwise-ECE** (average ranks) · **Confidence-ECE** (average ranks) · **Error rate** (average ranks)

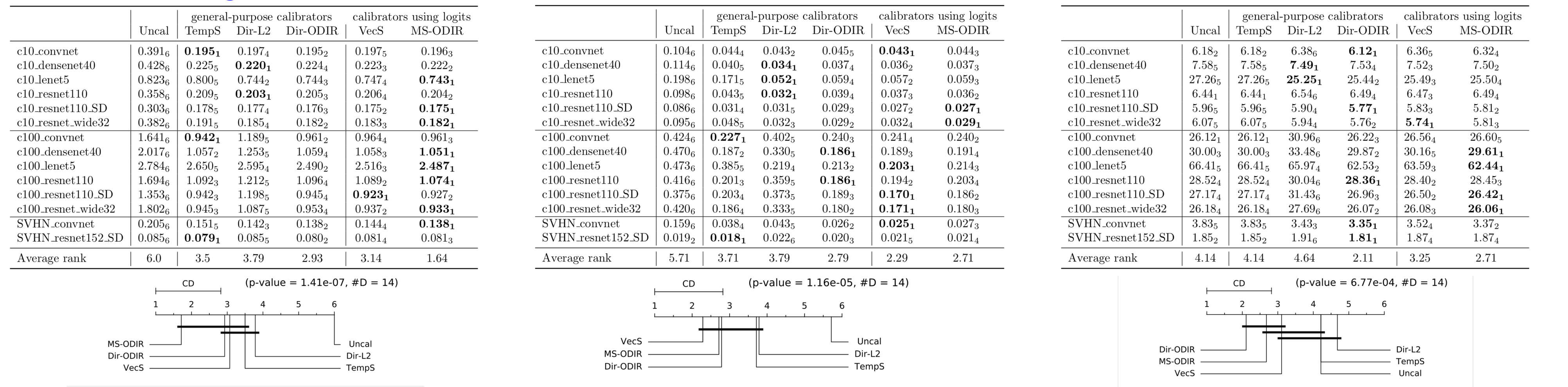


Deep neural networks experiment

14 CNNs for CIFAR-10, CIFAR-100, SVHN 14

Calibration maps trained on 5000 validation instances with 5-fold-crossvalidation

Log-loss · **Classwise-ECE** · **Error rate**



Conclusion:

- Dirichlet calibration:
  - New parametric general-purpose multiclass calibration method
  - Natural extension of two-class Beta calibration
  - Easy to implement as a neural layer or as multinomial logistic regression on log-transformed class probabilities
  - Best or tied best average rank across 21 datasets x 11 classifiers
- ODIR regularisation:
  - Matrix scaling with ODIR is tied best in log-loss
  - Dirichlet with ODIR is tied best in error rate

Contact: meelis.kull@ut.ee