Post-hoc Bayesian uncertainty for a deployed tumour classifier: calibration and out-of-distribution detection under platform shift

C. Stacy[1] and S. Das[1]

[1]School of Pharmacy and Biomolecular Sciences, RCSI University of Medicine and Health Sciences, Dublin, Ireland.

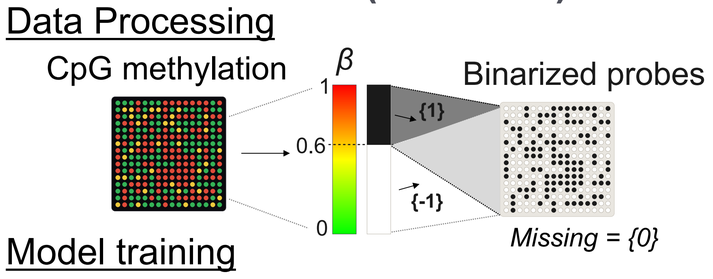
RCSI UNIVERSITY OF MEDICINE AND HEALTH SCIENCES

Introduction

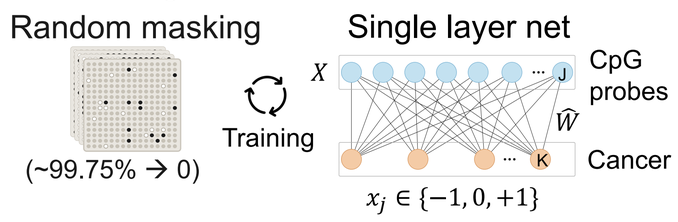
- DNA methylation signatures can classify cancers.
- Most classifiers are platform (data type) specific.
- Newly developed classifier works across all existing data types.

THE CLASSIFIER (crossNN)[1]

Data Processing



Model training

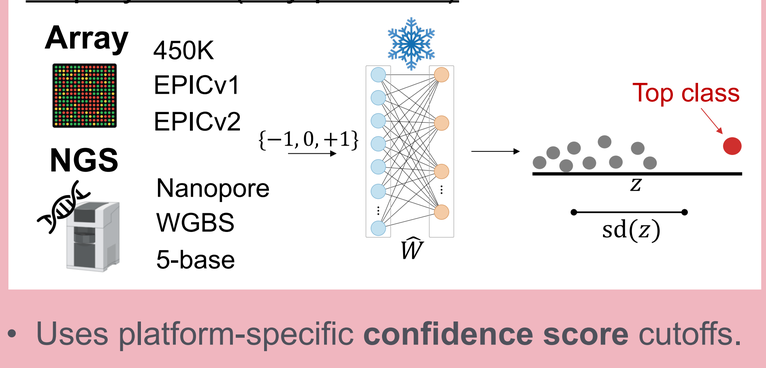


$$z_k = \sum_j \widehat{W}_{\{k,j\}} x_j$$

Confidence score:

$$\mathrm{softmax}_k\left(\frac{z}{\mathrm{sd}(z)}\right)$$

Deployment (any platform)



- Uses platform-specific **confidence score** cutoffs.
  - Current values: sequencing (**0.2**) vs array (**0.4**)
- Excellent classification performance: AUROC>0.95

Rare sinonasal tumours can be classified too

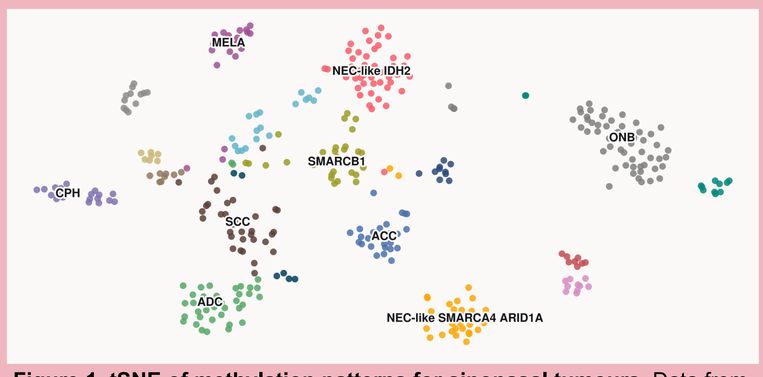


Figure 1. tSNE of methylation patterns for sinonasal tumours. Data from Jurmeister et al.[2] 395 reference tumours in 18 classes.

- **7 of the 18** classes hold fewer than 15 samples

Methods & Data

- **Model.** crossNN[1]: single linear layer on ternary features.
  – No model was refit to calibrate, and no diagnosis changed.
- **Data.** 2,801 CNS[3] and 7,001 TCGA primary tumours using released weights; 447 head & neck; 415 nanopore runs, 59 of them in known CNS classes.
- **Calibration.** Measured as *top-label* ECE with 15 equal-width bins and cross-checked against the class-wise KL calibration error of a leave-one-out KDE.
- **Fit.** a solves E[σ(e(a))] = accuracy was set from CNS.

Accurate but underconfident

- Accurate: **96.2%** CNS, **98.4%** head & neck.
- Confidence score not a probability.

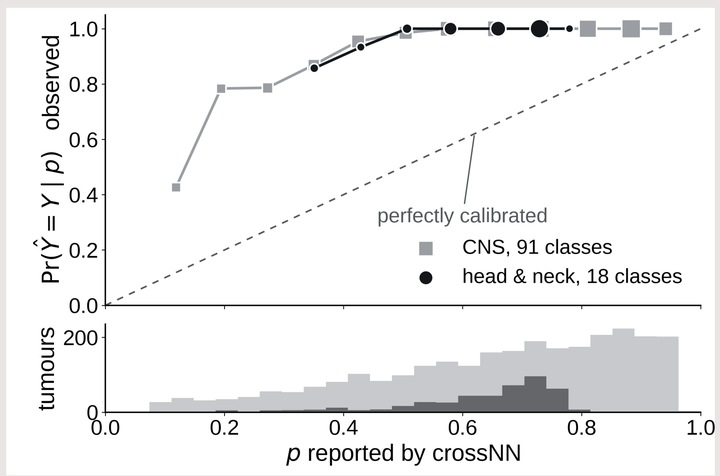


Figure 2. The score sits far below the accuracy it should report. Reliability (top) and where the scores fall (bottom). The head & neck score never exceeds 0.796, so no call reaches 90% confidence.

The missing coefficient

- Published rule: **softmax(z / sd(z))**.
- Expanded: **softmax(z / (1 · sd(z)))**, so the coefficient is 1.
- Laplace posterior on the weights do not support a=1.

$z = Wx$ $\quad s = \mathrm{sd}(z)$ $\quad \sigma(t) = (1 + e^{-t})^{-1}$
$p^{(a)} = \mathrm{softmax}(z/(as))$

(1) $$\mathrm{softmax}\left(\frac{z}{s}\right) = \mathrm{softmax}\left(\frac{z}{\mathbf{1} \cdot s}\right)$$

(2) $$u_k = \frac{z_{\max} - z_k}{s} \qquad u_2 = \min_{k \neq \mathrm{top}} u_k$$

$\mathrm{sd}(u) = 1$ for every sample

(3) $$\max_k p_k^{(a)} = \frac{1}{\sum_k e^{-(z_{\max} - z_k)/(as)}}$$

(4) $$= \frac{1}{\sum_k e^{-u_k/a}}$$

(5) $$= \frac{1}{1 + e^{-u_2/a} N(a)} = \sigma(e(a))$$

(6) $$N(a) = \sum_{k \neq \mathrm{top}} e^{-(u_k - u_2)/a} \qquad e(a) = \frac{u_2}{a} - \ln N(a)$$

(7) $$\mathbb{E}_n[\sigma(e(a))] = A \qquad A = \text{accuracy}$$

$a = 0.34$ $\qquad 1/a = 2.9$ nats

Confirmed by a proper score

- Mean reported confidence = observed accuracy gives **a = 0.3404**.
- The log-likelihood puts its minimum at **0.3405**, on the same data.

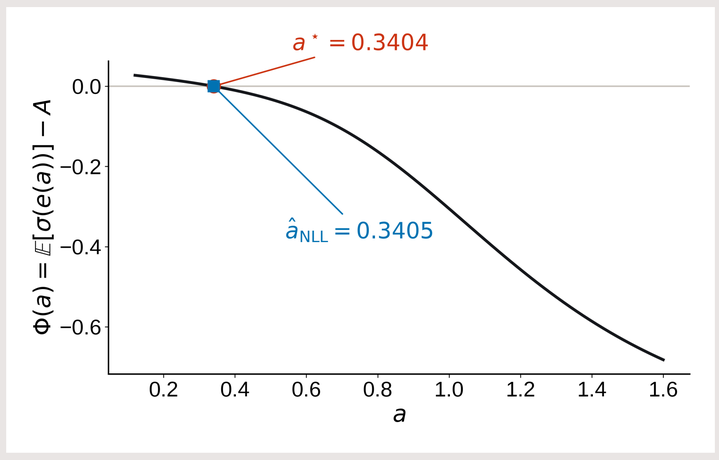


Figure 3. Two estimators agree to four decimals. Φ(a) = mean reported confidence − observed accuracy, 2,801 CNS tumours. Circle: its root. Square: the log-likelihood minimiser, on the same data.

Bayesian motivation

- Laplace posterior on the *weights*; MacKay evidence[4] prior.
- Its probit predictive divides the logits by $\sqrt{(1 + \pi v/8)}$, which is near-constant across samples, i.e. a **temperature**.
- Set equal to a·sd(z) on CNS: **a = 0.44**, no calibration labels. Matching accuracy gives **0.34**.

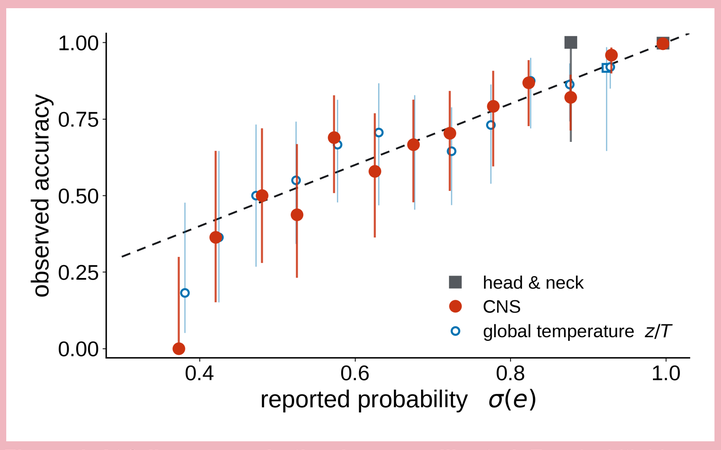


Figure 4. At full coverage both rules are calibrated. Equal-width bins, Wilson 95% intervals. ECE for z/(a·s) is 0.016 and 0.007, against 0.016 and 0.006 for the global temperature.

Limits of the posterior

- The posterior gives the right **form**, the wrong **scale**.
- Logit is a coherent sum (∝|S|); its posterior spread is incoherent (∝|S|½).
- No prior identified closes a **0.48** exponent gap.
- Curvature unusable: **71%** of samples past float64's limit.

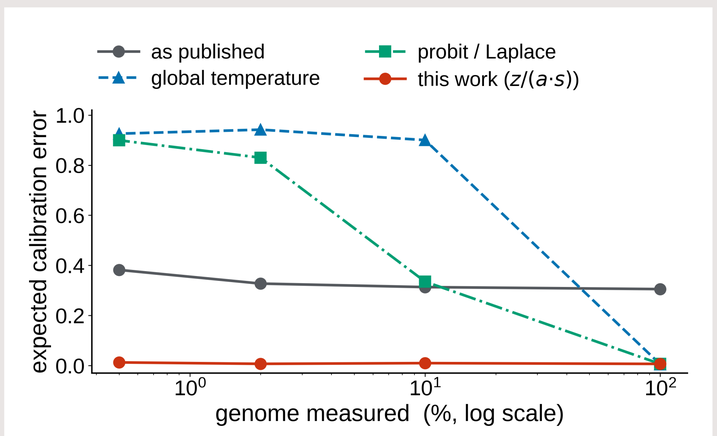


Figure 5. Only a·s survives the loss of coverage. All three alternatives are set at full coverage, then coverage is reduced.

Arrays to nanopore

- Both scalars were fitted on arrays and read unchanged on nanopore at **~1.7%** coverage.
- A temperature is calibrated only at the coverage it was fitted at; **a·s** holds.

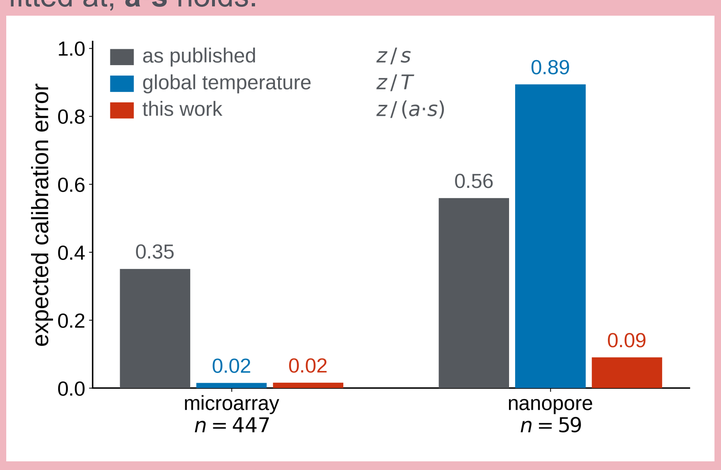


Figure 6. A temperature fitted on arrays fails on nanopore. All three rules are softmax(z / divisor); only the divisor differs. Accuracy 0.949 on nanopore, unchanged by any of them.

References. 1. Yuan D et al. Nat Cancer 6:1283–1294 (2025). 2. Jurmeister P et al. Nat Commun 13:7148 (2022). 3. Capper D et al. Nature 555:469–474 (2018). 4. MacKay DJC. Neural Comput 4:720–736 (1992). 5. Guo C et al. ICML PMLR 70:1321–1330 (2017).

Result

a = 0.34 was set once on CNS and frozen; the Laplace posterior gives 0.44 with no labels

model	K	n	ECE shipped	ECE at a
CNS (their weights)	91	2,801	0.305	**0.007**
head & neck (ours)	18	447	0.351	**0.016**
pan-cancer (their weights)	178	7,001	0.403	**0.032**

ECE is top-label, 15 equal-width bins, and a biased low estimate. Class-wise KL error by Dirichlet-KDE, the consistent estimator, agrees: 0.046 to 0.006 on head & neck and 0.009 to 0.002 on CNS. Accuracy unchanged.

What the cutoffs mean

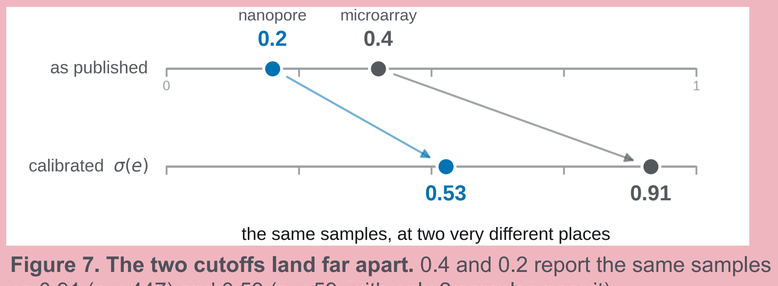


Figure 7. The two cutoffs land far apart. 0.4 and 0.2 report the same samples as 0.91 (n = 447) and 0.53 (n = 59, with only 2 samples near it).

Conclusion

- A deployed classifier's confidence was off probability by a coefficient.
- Post-hoc Bayes sets the scale; tracking sd(z) carries it across platforms.
- One number fixes **18, 91 and 178** classes, with nothing refit to calibrate and no diagnosis changed.
- **1/a = 2.9**. The shipped rule fixes this coefficient at 1.
- **Limits.** OOD ranking barely moves (AUROC 0.908 to 0.912); no scalar sees a tumour outside the K classes; pan-cancer's slope is 0.66 until a is refit.

Acknowledgements Health Research Board TRANSCAN

Supported by the Health Research Board under SPELCASTER (TRANSCAN-3 ERA-NET Cofund, EU). Das Lab, RCSI.

WWW.RCSI.COM