

SCOPE-Intent: Segment-Conditioned Probability Calibration for E-Commerce Purchase and Repurchase Prediction

Xiang Li*

Qingdao No.58 High School of Shandong Province, Qingdao, China

**Corresponding author: Xiang Li.*

Abstract

Purchase-intent models can rank prospective buyers well while returning probabilities whose reliability varies across behavioral contexts. We evaluate SCOPE-Intent, a validation-only segment-conditioned calibration layer, on session purchase prediction (UCI Online Shoppers) and customer-anchor 30-day repurchase prediction (UCI Online Retail). The pipeline fits prediction and soft routing on training data, fits all calibration maps on validation data, and evaluates only once on held-out records. Five-seed paired bootstrap analysis finds that SCOPE-Intent improves Brier score versus global temperature scaling on both tasks (Online Shoppers: 0.00068, 95% CI [0.00055, 0.00080]; Online Retail: 0.00040, [0.00032, 0.00049]). The result is not universal: Platt and beta calibration are substantially better on Online Shoppers, and hard or unshrunk segment controls are not stably worse on Online Retail. The evidence supports task- and comparator-conditioned probability calibration only, not marketing uplift, revenue, causal demand effects, or deployment readiness on these two public benchmarks.

Keywords

purchase-intent prediction, probability calibration, e-commerce, consumer-demand estimation, behavioral segmentation

1. Introduction

Personalized marketing systems often rank customers or sessions by an estimated purchase probability. Ranking alone, however, does not ensure that a reported probability is reliable for every behavioral context. A poorly calibrated value can be too confident or too conservative even when it orders cases well. This is a probability-estimation problem, not evidence that contacting a consumer causes a purchase.

We study this problem on two deliberately distinct public e-commerce tasks: Online Shoppers session purchase prediction and a leakage-safe Online Retail customer-anchor, next-30-day repurchase task. The distinction matters. The first predicts a supplied session outcome, whereas the latter is constructed from pre-anchor transaction history and permits repeated observations of a customer across months. Neither task observes marketing treatment, campaign response, revenue, or causal consumer demand.

This paper evaluates SCOPE-Intent (Segment-Conditioned Optimization for Probability Estimation), a post-hoc probability layer. It fits a predictor and soft behavioral routing from training data, learns all calibration maps from validation data only, and evaluates the frozen pipeline on test data. Rather than assume that this conditioning is broadly superior, we compare it fairly with global temperature, Platt, beta, isotonic, hard-segment, and unshrunk soft-segment calibration on the same predictor outputs.

Our contributions are empirical and reproducible. First, we specify an unweighted partial-pooling temperature penalty and report soft support separately, avoiding an unsupported sparse-group interpretation. Second, we persist row-level predictions and routing evidence for paired session- or customer-cluster bootstrap inference. Third, we report counterexamples alongside favorable comparisons: SCOPE-Intent improves on global temperature scaling on both tasks, but it loses to Platt and beta calibration on Online Shoppers and has no stable advantage over hard or unshrunk routing on Online Retail. The result is therefore a task- and comparator-conditioned calibration finding, not a universal method claim.

2. Related Work

E-commerce purchase-intent modelling is commonly evaluated on sessions or transaction histories, but these are distinct prediction targets. The Online Shoppers benchmark concerns a session purchase label [1, 2]; the Online Retail source records transactions and has also supported customer-segmentation analysis [3, 4]. Our derived retail target is instead a customer's next 30-day repurchase after an anchor. Neither task is a causal personalized-marketing or demand-intervention study.

Post-hoc calibration distinguishes score ranking from the reliability of the reported probability. Classical logistic/Platt maps [5], isotonic maps [6, 7], Bayesian binning [8], beta calibration [9], temperature scaling [10], and more flexible Dirichlet maps [11] offer different bias–variance trade-offs. Recent work further examines flexible nonparametric and ranking-aware calibration [12–14], while surveys and evaluation studies warn that a single calibration score is not a complete reliability assessment [15–17]. We therefore compare these maps on the same fixed outputs and show reliability curves, Brier score, ECE, and PR-AUC.

Group-conditional reliability is motivated by multicalibration [18], but SCOPE-Intent does not assert a multicalibration guarantee. It uses train-fitted soft routing and a validation-only temperature layer, then reports each route's support, outcome rate, temperature, Brier score, ECE, calibration intercept/slope, and the soft–hard routing discrepancy. Its contribution is consequently procedural and empirical: a transparent group-conditioned calibration comparison, not a new theoretical guarantee or a new representation architecture.

The method is also separate from training-time calibration objectives [19, 20] and from the standard repeated-measurement models that motivate partial pooling [21, 22]. Its quadratic penalty is deliberately simple; whether it helps remains an empirical question answered only by the held-out, paired comparison.

3. Method

3.1 Problem Setting

For each e-commerce record, a fixed predictor maps information observed before the prediction anchor to an intent logit. The objective is to post-process that logit into a probability without changing the outcome definition and without using test labels during fitting. We call the procedure *SCOPE-Intent* (*Segment-Conditioned Optimization for Probability Estimation*). The principal predictor is a tabular MLP; XGBoost is a separately reported robustness backbone. Thus, the calibration comparison does not conflate a new predictor with a new probability map.

3.2 Segment-Conditioned Calibration

Fig. 1 shows three leakage-separated stages. Training features fit the fixed predictor and a behavioral segmenter; validation logits and labels fit every calibrator; and the test split is used only once to score the completed maps. Let x denote a preprocessed record, z its fixed base logit, and K the number of segments. Train-fitted k -means centers [23] define normalized soft routing weights.

$$w_k(x) = \frac{\exp(-\|x - c_k\|_2^2/\tau)}{\sum_{j=1}^K \exp(-\|x - c_j\|_2^2/\tau)}.$$

Here c_k is the k th train-fitted center, j is the summation index, and $\tau > 0$ is the routing-temperature hyperparameter. The segmenter is fit with 20 initializations on the train-only preprocessed feature matrix; validation and test memberships reuse these centers unchanged.

The global temperature T_g is fit on validation logits and labels. The segment temperatures T_k are then fit on the same validation partition, and the calibrated probability is

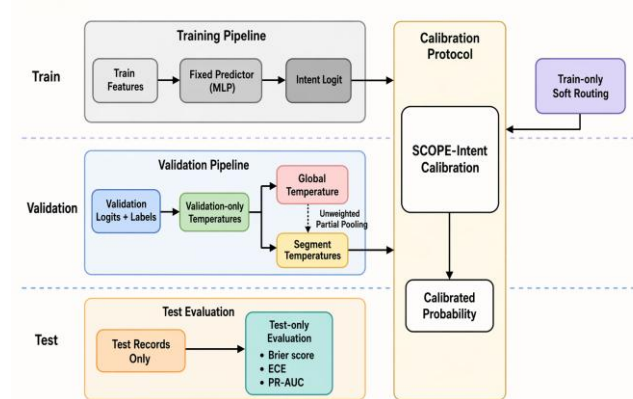
$$\hat{p}(z, x) = \sum_{k=1}^K w_k(x) \sigma(z/T_k),$$

where σ is the logistic function. Let $\mathcal{V}$ denote the validation-record index set and let $n_k = \sum_{i \in \mathcal{V}} w_k(x_i)$ be the validation soft support of segment k . We report n_k as a diagnostic; it is not an extra penalty multiplier. With $y \in \{0,1\}$ denoting the validation outcome, the validation objective is the unweighted partial-pooling penalty

$$\mathcal{L} = \text{BCE} \left(y, \sum_{k=1}^K w_k(x) \sigma(z/T_k) \right) + \lambda \sum_{k=1}^K (\log T_k - \log T_g)^2.$$

This is a deterministic regularizer that pulls each segment temperature toward the global estimate, analogous in intent to partial pooling but not a fitted hierarchical Bayesian model [21, 22]. It therefore does *not* justify a claim that sparse groups receive stronger shrinkage. Temperature parameters are represented in log space and clamped to $[0.05, 20]$ at evaluation. The global log-temperature is optimized by L-BFGS; the segment log-temperatures start at $\log T_g$ and are optimized with Adam [24] for the declared validation steps.

Figure 1: SCOPE-Intent uses a fixed predictor (MLP in the primary experiment), train-fitted soft behavioral routing, and validation-fitted temperatures. The test branch only evaluates Brier score, ECE, and PR-AUC; it never updates the predictor, router, or calibrator



3.3 Controls and Ablations

On identical MLP outputs, the primary comparison includes Constant-prevalence, global temperature scaling [10], Platt scaling [5], beta calibration [9], isotonic regression [6], hard-segment temperature scaling, hard-segment Platt scaling, soft routing without shrinkage, and SCOPE-Intent. Logistic regression, the uncalibrated MLP, and XGBoost [25] provide predictor baselines; XGBoost repeats the calibrator set as a robustness analysis. The ablations are global-only calibration (no segment conditioning), no shrinkage, the raw MLP probability, and random routing. Lower Brier score [26] and ECE are reliability diagnostics; PR-AUC is reported to detect material ranking changes. None of these comparisons estimates a marketing treatment effect.

4. Experiments

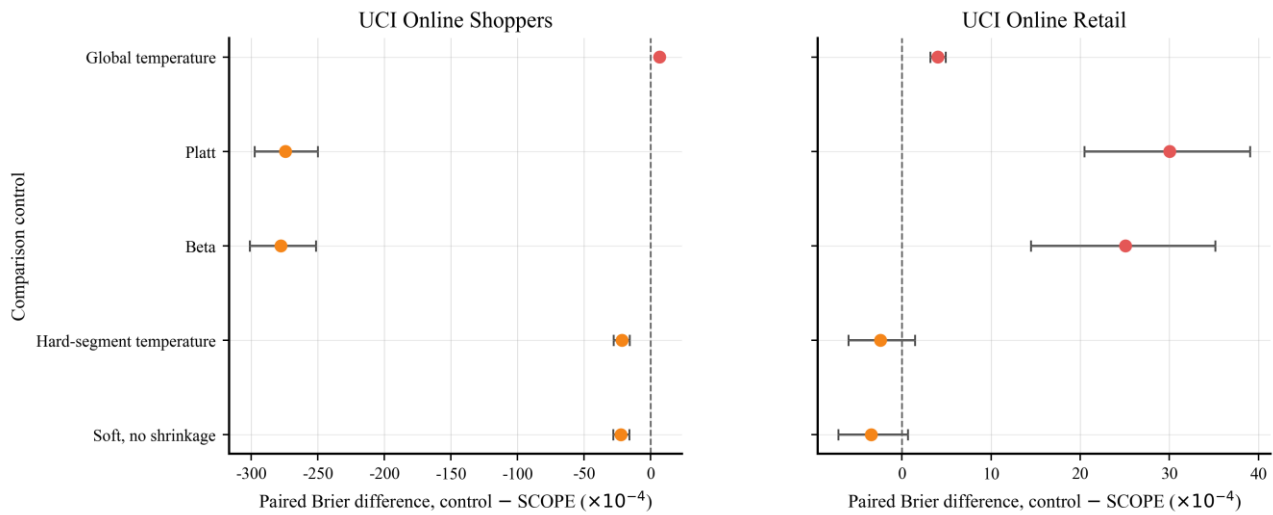
4.1 Data and Leakage-Safe Protocol

Online Shoppers is a session-purchase task: its supplied Revenue label is used once per session [2]. The data contain 12,330 sessions and 1,908 positive labels (15.47%). For each of the five prespecified seeds (13, 23, 37, 53, and 71), stratified splitting produces 7,398 training, 2,466 validation, and 2,466 test sessions. The same predictor output is passed to every calibration control; training fits the predictor and preprocessing, validation fits every calibration map, and test records are never used for fitting.

Online Retail is a different task: customer-anchor 30-day repurchase prediction, not session purchase. Starting from the official transaction data [3, 4], we remove missing identifiers, cancellation invoices, and nonpositive quantity or price; aggregate invoice lines; and retain a customer at an anchor only after at least two earlier invoices. Its label is a positive non-cancellation invoice in [anchor, anchor + 30 days). The 12,440 anchors have 4,877 positive labels (39.20%). Training uses May–July 2011 (4,095 records), validation August–September (3,737), and test October–November (4,608); respective positive rates are 40.27%, 37.09%, and 39.97%. All eight history summaries use invoices strictly before the anchor. A customer may legitimately appear at multiple monthly anchors, so retail inference re-samples anonymized customer clusters rather than treating repeated records as independent.

The primary backbone is the MLP with ReLU widths 64 and 32, dropout 0.1 [27], AdamW learning rate 10^{-3} , and 32 epochs. XGBoost is a robustness backbone, not a pooled primary result. The primary score is Brier score [26, 28]; ECE (15 equal-width bins) and PR-AUC are secondary diagnostics; the full PR-AUC results are reported in anonymous Supplementary Table S1. We preserve a row-level test artifact for every method and seed: anonymized case and customer keys, anchor, label, probability, method label, soft memberships, hard routing, and segment temperatures. The primary interval is the 2,000-replicate paired bootstrap of control Brier minus SCOPE Brier: Online Shoppers re-samples sessions and Online Retail re-samples customer clusters. Positive intervals favor SCOPE-Intent. Seed standard deviations describe training variability only and do not replace this interval. Fig. 2 makes this direction and uncertainty visible: each point is a paired Brier difference and each horizontal interval is its 95% bootstrap interval.

Figure 2: Paired baseline comparisons using Brier-score differences with 95% percentile bootstrap intervals. Positive values favor SCOPE-Intent. Online Shoppers uses session-level resampling and Online Retail uses anonymized customer-cluster resampling. Intervals crossing zero are inconclusive



4.2 Main Results

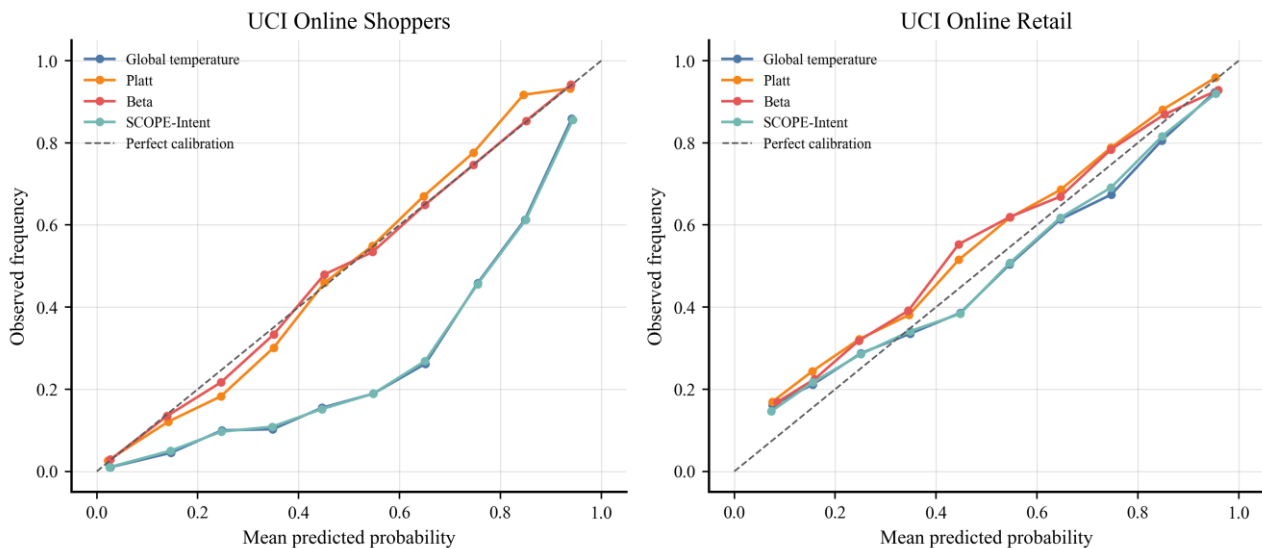
Table I reports mean $\pm$ seed standard deviation. On Online Shoppers, SCOPE-Intent improves on global temperature scaling by 0.00068 Brier points; its 95% interval is [0.00055, 0.00080]. It remains substantially worse than Platt and beta calibration. On Online Retail, the improvements over global temperature, Platt, and beta are 0.00040, 0.00301, and 0.00251. Their intervals are [0.00032, 0.00049], [0.00205, 0.00391], and

[0.00145, 0.00351]. Comparisons with hard-segment temperature scaling and soft routing without shrinkage are inconclusive: the differences are -0.00024 and -0.00034, with intervals [-0.00060, 0.00014] and [-0.00071, 0.00007]. Thus the data do not support an overall superiority claim. With XGBoost, the corresponding improvements over global temperature are 0.00069 [0.00058, 0.00081] on Online Shoppers and 0.00039 [0.00032, 0.00045] on Online Retail; the broader pattern remains task- and comparator-dependent.

Table 1: Held-out MLP-backbone calibration results (mean $\pm$ seed standard deviation, $n = 5$). Lower Brier/ECE is better. PR-AUC is reported in anonymous Supplementary Table S1; constant-prevalence is a non-discriminative calibration reference

Method	Shoppers Brier	Shoppers ECE	Retail Brier	Retail ECE
Constant-prevalence	.13091 $\pm$.00000	.00022 $\pm$.00007	.23996 $\pm$.00000	.00295 $\pm$.00000
MLP	.10497 $\pm$.00622	.10721 $\pm$.00687	.20863 $\pm$.00117	.05709 $\pm$.00569
Global temperature	.10275 $\pm$.00587	.11819 $\pm$.00416	.20708 $\pm$.00085	.04362 $\pm$.00401
Platt	.07465 $\pm$.00332	.02010 $\pm$.00257	.20969 $\pm$.00158	.06703 $\pm$.00431
Beta	.07431 $\pm$.00346	.01959 $\pm$.00393	.20919 $\pm$.00160	.06519 $\pm$.00396
Isotonic	.07504 $\pm$.00366	.01863 $\pm$.00612	.20979 $\pm$.00167	.06662 $\pm$.00469
Hard-segment temperature	.09993 $\pm$.00563	.11012 $\pm$.00351	.20644 $\pm$.00080	.04140 $\pm$.00532
Hard-segment Platt	.07457 $\pm$.00348	.01944 $\pm$.00409	.20697 $\pm$.00128	.05695 $\pm$.00360
Soft, no shrinkage	.09985 $\pm$.00565	.10964 $\pm$.00334	.20634 $\pm$.00079	.04246 $\pm$.00407
SCOPE-Intent	.10208 $\pm$.00595	.11565 $\pm$.00361	.20668 $\pm$.00087	.04211 $\pm$.00293

Figure 3: Reliability curves for global temperature, Platt, beta, and SCOPE-Intent, pooled across the five prespecified held-out seed splits. The perfect-calibration line is a reference. Curves are descriptive; the paired bootstrap effects in Fig. 2 provide the primary uncertainty analysis



4.3 Segment Diagnostics and Supplementary Analysis

The segment artifact supports diagnosis rather than a sparse-group claim. Mean validation soft support ranges from 254 to 857 in Online Shoppers and from 604 to 1,175 in Online Retail; mean fitted temperatures range from 1.28 to 1.38 and 1.09 to 1.17, respectively. For every seed we report each segment's support, positive rate, temperature, Brier score, ECE, calibration intercept/slope, and test hard count. The nearest-center hard route agrees with the maximum soft-membership route by construction; the meaningful soft-versus-hard comparison is therefore the held-out hard-segment comparison in Fig. 2. Fig. 3 plots pooled held-out reliability curves for global temperature, Platt, beta, and SCOPE-Intent, together with the perfect-calibration reference; it is descriptive rather than the source of the primary confidence intervals. The prespecified sensitivity grids are reported in the anonymous supplementary material and were not used to select the headline configuration.

5. Limitations

The study uses two public datasets and one derived temporal retail target; it does not establish generalization to another merchant, time period, or intervention policy. Online Retail deliberately retains repeat customer observations across monthly anchors, which is why its uncertainty analysis clusters by anonymous customer, but residual temporal and source-specific dependence may remain. Calibration improvements do not estimate campaign uplift, revenue, consumer welfare, or deployment utility. Finally, the soft segments are data-adaptive routing devices rather than validated consumer categories, and the quadratic regularizer is not a hierarchical Bayesian model or a multicalibration guarantee.

6. Conclusion

SCOPE-Intent is a reproducible, validation-only segment-conditioned probability calibration procedure. Paired bootstrap evidence shows a stable advantage over global temperature scaling on the two evaluated e-commerce tasks, but not over every reasonable calibrator. In particular, Platt and beta calibration are stronger on Online Shoppers, while hard and unshrunk segment controls remain competitive on Online Retail. The appropriate conclusion is therefore conditional: segmentation can alter probability calibration, but its value depends on the target task, backbone, and comparator. Future work should test new time-separated datasets and externally meaningful segments before considering any operational use.

References

- [1] Sakar, C. O. and Kastro, Y., “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks,” *Neural Computing and Applications*, vol. 31 pp. 6893–6908 2019, doi: 10.1007/s00521-018-3523-0.
- [2] Sakar, C. O. and Kastro, Y., “Online shoppers purchasing intention dataset,” 2018, doi: 10.24432/C5F88Q.
- [3] Chen, D., Sain, S. L., and Guo, K., “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” *Journal of Database Marketing and Customer Strategy Management*, vol. 19 no. 3 pp. 197–208 2012, doi: 10.1057/dbm.2012.17.
- [4] Chen, D., “Online retail dataset,” 2015, doi: 10.24432/C5BW33.
- [5] Platt, J. C., “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, pp. 61–74 1999.
- [6] Zadrozny, B. and Elkan, C., “Transforming classifier scores into accurate multiclass probability estimates,” *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 694–699 2002, doi: 10.1145/775047.775151.
- [7] Niculescu-Mizil, A. and Caruana, R., “Predicting good probabilities with supervised learning,” *Proceedings of the 22nd international conference on machine learning*, pp. 625–632 2005, doi: 10.1145/1102351.1102430.
- [8] Naeini, M. P., Cooper, G., and Hauskrecht, M., “Obtaining well calibrated probabilities using bayesian binning,” *Proceedings of the AAAI conference on artificial intelligence*, vol. 29 pp. 2901–2907 2015.
- [9] Kull, M., Menezes e Silva Filho, T. de, and Flach, P., “Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers,” *Proceedings of the 20th international conference on artificial intelligence and statistics*, vol. 54 pp. 623–631 2017.
- [10] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., “On calibration of modern neural networks,” *Proceedings of the 34th international conference on machine learning*, vol. 70 pp. 1321–1330 2017.
- [11] Kull, M., Nieto, M. P., Kängsepp, M., Filho, T. S., Song, H., and Flach, P., “Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration,” *Advances in neural information processing systems*, vol. 32 2019.
- [12] Wenger, J., Kjellström, H., and Triebel, R., “Non-parametric calibration for classification,” *Proceedings of the 23rd international conference on artificial intelligence and statistics*, vol. 108 pp. 178–190 2020.

- [13] Ma, X. and Blaschko, M., “Meta-cal: Well-controlled post-hoc calibration by ranking,” *Proceedings of the 38th international conference on machine learning*, vol. 139 pp. 7235–7245 2021.
- [14] Berta, E., Bach, F., and Jordan, M. I., “Classifier calibration with ROC-regularized isotonic regression,” *Proceedings of the 27th international conference on artificial intelligence and statistics*, vol. 238 pp. 1972–1980 2024.
- [15] Filho, T. S., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., and Flach, P., “Classifier calibration: A survey on how to assess and improve predicted class probabilities,” *Machine Learning*, vol. 112 pp. 3211–3260 2023, doi: 10.1007/s10994-023-06336-7.
- [16] Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. B., “Evaluating model calibration in classification,” *Proceedings of the 22nd international conference on artificial intelligence and statistics*, vol. 89 pp. 3459–3467 2019.
- [17] Kumar, A., Sarawagi, S., and Jain, U., “Trainable calibration measures for neural networks from kernel mean embeddings,” *Proceedings of the 35th international conference on machine learning*, vol. 80 pp. 2805–2814 2018.
- [18] Hébert-Johnson, Úrsula, Kim, M. P., Reingold, O., and Rothblum, G. N., “Multicalibration: Calibration for the (computationally-identifiable) masses,” *Proceedings of the 35th international conference on machine learning*, vol. 80 pp. 1939–1948 2018.
- [19] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P., “Calibrating deep neural networks using focal loss,” *Advances in neural information processing systems*, vol. 33 2020.
- [20] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M., “Revisiting the calibration of modern neural networks,” *Advances in neural information processing systems*, vol. 34 pp. 15682–15694 2021.
- [21] Laird, N. M. and Ware, J. H., “Random-effects models for longitudinal data,” *Biometrics*, vol. 38 no. 4 pp. 963–974 1982, doi: 10.2307/2529876.
- [22] Greenland, S., “Principles of multilevel modelling,” *International Journal of Epidemiology*, vol. 29 no. 1 pp. 158–167 2000, doi: 10.1093/ije/29.1.158.
- [23] MacQueen, J. B., “Some methods for classification and analysis of multivariate observations,” *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*, vol. 1 pp. 281–297 1967.
- [24] Kingma, D. P. and Ba, J., “Adam: A method for stochastic optimization,” *International conference on learning representations*, 2015.
- [25] Chen, T. and Guestrin, C., “XGBoost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794 2016, doi: 10.1145/2939672.2939785.
- [26] Brier, G. W., “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78 no. 1 pp. 1–3 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- [27] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15 no. 56 pp. 1929–1958 2014.
- [28] Murphy, A. H., “A new vector partition of the probability score,” *Journal of Applied Meteorology*, vol. 12 no. 4 pp. 595–600 1973, doi: 10.1175/1520-0450(1973)012<0595:ANVOTP>2.0.CO;2.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

