Evaluating Out-of-Domain Performance Prediction Methods Under Domain Shifts

P. Weerasinghe[1]*, R. Tennakoon[2], J. Thangarajah[2], D. Herath[1]

[1]*University of Peradeniya, Peradeniya 20400, Sri Lanka*
[2]*RMIT University, Melbourne, VIC 3000, Australia*
**pramukaw@sci.pdn.ac.lk*

Deep neural networks have achieved remarkable success in various applications where the test data follows the same distribution as the training data. However, in many real-world scenarios, this assumption often fails due to domain shifts, where test data referred to as out-of-domain (OoD) data differs from the training distribution. Such shifts can significantly degrade model performance and are particularly pronounced in data-constrained domains such as medical imaging, arising from variations in imaging devices, acquisition protocols, and patient demographics. While numerous techniques have been proposed to address domain shift, their effectiveness remains limited, especially when labeled target data is unavailable. This study investigates algorithms designed to predict OoD performance on unlabeled datasets and evaluates their reliability under different types of domain shifts, specifically diversity shifts (emergence of new patterns) and correlation shifts (changes in feature–label relationships). A systematic benchmarking framework is adopted to evaluate a diverse set of state-of-the-art OoD performance prediction algorithms across datasets exhibiting both shift types. The evaluation includes benchmark datasets such as PACS, CIFAR-100-C, and ColoredMNIST. OoD performance prediction methods Average Threshold Confidence (ATC), Difference of Confidence (DoC), Mandoline, and AugMax are assessed under consistent experimental settings using metrics such as robustness accuracy and mean absolute error. The results reveal significant variability in algorithm performance across shift types, indicating a lack of consistent generalization. Under correlation shift (MNIST to ColoredMNIST), ATC achieves substantially higher robustness accuracy (0.9466) compared to DoC (0.2096), demonstrating its effectiveness in confidence-based estimation. In contrast, under diversity shift (CIFAR-100 to CIFAR-100-C), Mandoline outperforms ATC (0.5164 vs. 0.2655), highlighting its sensitivity to distributional variation. Overall, no single method performs reliably across both shift types, emphasizing limitations of current approaches in complex real-world scenarios. These findings highlight the need for shift-aware or hybrid OoD performance prediction frameworks capable of handling diverse and interacting domain shifts. This work provides a benchmark for future research and offers critical insights for advancing reliable OoD performance estimation.

Keywords: out-of-domain performance, domain shift, robustness evaluation