Calibrated Diverse Ensemble Entropy Minimization for Robust Test-Time Adaptation in Prostate Cancer Detection

Mahdi Gilany[1](✉), Mohamed Harmanani[1], Paul Wilson[1],
Minh Nguyen Nhat To[2], Amoon Jamzad[1], Fahimeh Fooladgar[2],
Brian Wodlinger[3], Purang Abolmaesumi[2], and Parvin Mousavi[1]

[1] School of Computing, Queen's University, Kingston, Canada
mahdi.gilany@queensu.ca
[2] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
[3] Exact Imaging, Markham, Canada

Abstract. High resolution micro-ultrasound has demonstrated promise in real-time prostate cancer detection, with deep learning becoming a prominent tool for learning complex tissue properties reflected on ultrasound. However, a significant roadblock to real-world deployment remains, which prior works often overlook: model performance suffers when applied to data from different clinical centers due to variations in data distribution. This distribution shift significantly impacts the model's robustness, posing major challenge to clinical deployment. Domain adaptation and specifically its test-time adaption (TTA) variant offer a promising solution to address this challenge. In a setting designed to reflect real-world conditions, we compare existing methods to state-of-the-art TTA approaches adopted for cancer detection, demonstrating the lack of robustness to distribution shifts in the former. We then propose Diverse Ensemble Entropy Minimization (DEnEM), questioning the effectiveness of current TTA methods on ultrasound data. We show that these methods, although outperforming baselines, are suboptimal due to relying on neural networks output probabilities, which could be uncalibrated, or relying on data augmentation, which is not straightforward to define on ultrasound data. Our results show a significant improvement of 5% to 7% in AUROC over the existing methods and 3% to 5% over TTA methods, demonstrating the advantage of DEnEM in addressing distribution shift.

Keywords: Ultrasound Imaging · Prostate Cancer · Computer-aided Diagnosis · Distribution Shift Robustness · Test-time Adaptation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73284-3_36.

X. Xu et al. (Eds.): MLMI 2024, LNCS 15241, pp. 361–371, 2025.
https://doi.org/10.1007/978-3-031-73284-3_36

1 Introduction

Prostate cancer (PCa) diagnosis remains a pivotal challenge, where early and accurate detection can significantly influence treatment outcomes. The standard diagnostic approach is transrectal ultrasound-guided biopsy (TRUS), which *systematically* samples biopsy cores from uniform locations across the prostate. However, this method cannot specifically target suspicious areas due to similar acoustic tissue properties and low resolution, often resulting in missed diagnoses or unnecessary interventions. Alternatively, advanced ultrasound modalities aim to improve tissue characterization and offer targeted diagnostic strategies. In particular, high-frequency micro-ultrasound (micro-US) has recently emerged for visualization of prostate tissue at significantly higher resolutions than conventional imaging [4], facilitating the identification of potential cancer lesions. Using a qualitative approach, micro-US has demonstrated comparable PCa detection to multi-parametric MRI targeted biopsy [1,3], with the distinct advantage of operating in real-time. However, quantitative approaches interpreting this modality is in early-stage research with few studies [5,22] proposed for user-independent and objective PCa detection. Recent studies have utilized deep learning to analyze micro-US images for identifying PCa [6,27]. While deep learning can effectively extract complex tissue features from noisy ultrasound images, its adaptability to any data distribution can undermine model *robustness* during deployment when data distributions shift. Such shifts significantly impact model performance, as the model may learn biases in the training distribution and form spurious correlations between input images and tissue types [14,23]. This study argues that, despite recent successes, the performance and effectiveness of existing micro-US PCa detection studies are overestimated. These studies typically assume similar data distributions between training and deployment scenarios, which is unrealistic in clinical settings where shifts in data distribution are likely to occur due to differences in patient populations, operator techniques, and imaging devices.

To improve PCa detection robustness and mitigate distribution shifts, we explore the promising field of domain adaptation, where techniques are developed to adjust trained models to perform well on a target domain with a shift in distribution [15]. Specifically, we investigate a more challenging variant, *test-time adaptation* (TTA), which employs unsupervised objectives to fine-tune trained models on a single test sample during inference [18]. This approach eliminates the need to access the entire test dataset and accounts for variability in data distributions of test patients or biopsy cores. In practice, several challenges hinder the direct application of state-of-the-art (SOTA) TTA methods in PCa detection. Recent studies indicate that prior methods may fail in real-world scenarios with multiple types of distribution shifts [20], and their effectiveness varies depending on the shift type [31]. Moreover, the most common and promising TTA methods utilize either self-supervised learning (SSL) or entropy loss in their core [25,26]. However, choice of SSL data augmentation for ultrasound is not straightforward, and current entropy-based methods rely on neural networks generated probabilities (thus, entropy), known to be biased and uncalibrated [9]. This issue is exacerbated by distribution shifts and ultrasound artifacts and noisy labels [21].

An ideal TTA method should be tailored for micro-US shifts, produce calibrated entropy, and not rely on augmentations. To this end, we propose DEnEM, a novel TTA method for addressing clinical center distribution shifts in micro-US PCa detection. DEnEM utilizes an ensemble of neural networks diversified through mutual information minimization to produce calibrated marginal entropy without the need for data augmentations. Our key contributions are:

1. We investigate the existing SOTA PCa detection methods in a real-world setting, revealing the impact of distribution shift on their performance.
2. We propose leveraging SOTA TTA methods, for the first time, to address clinical center distribution shifts in PCa detection. We demonstrate their effectiveness in significantly improving over existing methods while also examining their limitations on micro-US data.
3. Inspired by TTA methods, we propose DEnEM, a novel approach tailored for micro-US that addresses entropy calibration and the reliance on data augmentations. We demonstrate that DEnEM significantly outperforms both baselines and SOTA TTA methods across several evaluation metrics.

2 Related Works

Micro-US PCa Detection: Recent studies have employed quantitative approaches for micro-US PCa detection [5,22,24,28]. Rohrbach et al. [22] introduced the first quantitative method using SVMs on manually extracted ultrasound spectral features, which is less effective than deep learning and requires a reference phantom, limiting its usability. Gilany et al. [5] demonstrated the success of CNNs in automatically extracting features from RF images to detect PCa. More recent methods [6,27] addressed labeled data scarcity and weak labeling, improving detection performance. However, these studies often neglect clinical integration and robust deployment across various clinical centers, leading to overestimated and unrealistic performance due to shifting data distributions.

Test-Time Adaptation: TTA is a challenging form of unsupervised domain adaptation, requiring the adaptation of a model trained on a source domain to a shifted target domain during inference, without accessing the *source domain* to preserve data privacy, and without the *entire target domain*, enabling immediate application post-deployment [18]. The core of TTA lies in defining a proxy objective to adapt the model to test samples unsupervised. These objectives include entropy minimization [26,30], self-supervised learning (SSL) [2,25], and feature alignment [31]. "TENT" [26] fine-tunes the model, at inference, by minimizing the entropy on test prediction probabilities p: $H(p) = -\sum_c p^c \log p^c$. "MEMO" [30] extends TENT by minimizing the entropy of *marginal* probability distribution calculated across a set of augmentations. While entropy minimization is a theoretically well established TTA objective [7], neural networks often produce biased and uncalibrated entropy [9]. Moreover, "TTT" [25] and "MT3" [2] use SSL proxy objectives. These methods modify the neural architecture by attaching a self-supervision head to the feature extractor network.

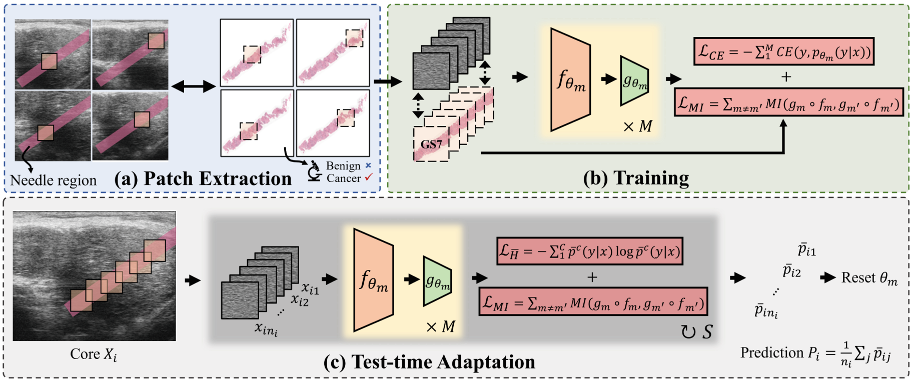
Needle region
(a) Patch Extraction
Benign ×
Cancer ✓
GS7
f_{θ_m}
g_{θ_m}
× M
$\mathcal{L}_{CE} = -\sum_1^M CE(y, p_{\theta_m}(y|x))$
+
$\mathcal{L}_{MI} = \sum_{m \neq m'} MI(g_m \circ f_m, g_{m'} \circ f_{m'})$
(b) Training
Core X_i
x_{i1}
x_{i2}
x_{in_i}
f_{θ_m}
g_{θ_m}
× M
$\mathcal{L}_{\overline{H}} = -\sum_1^C \bar{p}^c(y|x) \log \bar{p}^c(y|x)$
+
$\mathcal{L}_{MI} = \sum_{m \neq m'} MI(g_m \circ f_m, g_{m'} \circ f_{m'})$
↻ S
$\bar{p}_{i1}$
$\bar{p}_{i2}$
$\bar{p}_{in_i}$
Reset θ_m
Prediction $P_i = \frac{1}{n_i}\sum_j \bar{p}_{ij}$
(c) Test-time Adaptation

Fig. 1. Overview of DEnEM method. (a) RF patches extraction from needle region. (b) Deep ensemble training with cross entropy and mutual information losses. (c) Model adaptation at inference to each core with marginal entropy loss before the prediction.

During inference, this head and SSL objective adapt the feature extractor to the test distribution. However, these approaches have two drawbacks: they alter the neural architecture, potentially degrading performance in some tasks, and they only fine-tune the feature extractor, leaving the classification head unadapted. The latter is particularly impactful on distribution shift [13].

3 Materials and Method

3.1 Data

Our study employs a dataset collected from five clinical centers, with consented data from 693 patients who underwent systematic TRUS-guided prostate biopsy as part of a clinical trial (NCT02079025 clinicaltrials.gov). The procedure leverages high-frequency micro-US technology, typically capturing 10-12 biopsy cores per patient. Prior to firing the biopsy gun, a single RF image of the prostate with 28 mm depth and 46 mm width is captured. These images, paired with corresponding histopathology reports of the cores constitute our dataset. The reports include a binary label indicating the presence of cancer, an estimate of the cancer length or "involvement of cancer", and the aggressiveness of cancer as determined by the Gleason Score [19]. Overall, our dataset includes 6607 biopsy cores with 5727 of those being benign. More details of the data distribution among centres are provided in the supplementary material.

The needle trace region in the RF image, where histopathology labels are assigned, is first identified [5,6,24]. From this region, we perform patch extraction by capturing overlapping regions of interest (ROIs) that correspond to a tissue area of 5 mm by 5 mm, using 1 mm by 1 mm strides (see Fig. 1 (a)). An RF patch is considered to be within the needle region if there is at least a 60% overlap, a threshold set to compensate for potential inaccuracies in identifying the needle region [27]. RF patches are resized from their original dimensions of 1780 × 55 to 256 × 256 to align with our deep network architectures.

3.2 Method

Figure 1 shows an overview of our proposed approach. This approach trains a deep ensemble [16] of neural networks to minimize both cross-entropy loss and mutual information loss. During inference, marginal entropy is utilized to adapt the deep ensemble to the distribution of patches from a biopsy core in test set.

Deep Ensemble: Our model borrows ResNet10 [10] image encoder from prior PCa detection literature [5,27]. Deep ensemble is a collection of M encoders and classifiers with different random weight initialization. Unlike single network models, deep ensemble not only provides a highly calibrated entropy, but also this calibration remains less affected by distribution shift [21].

Mutual Information Loss: Randomly initialized neural networks in deep ensemble may still learn similar encoders and classifiers, providing biased and uncalibrated entropy. To overcome this, we propose ensuring statistical independence of predictions across the networks. Inspired by "DivDis" [17], we minimize the mutual information between the predictions of each network pair. The mutual information loss ($\mathcal{L}_{MI}$) and the total training loss ($\mathcal{L}$) are defined as follows:

$$\mathcal{L}_{MI}(g_m \circ f_m, g_{m'} \circ f_{m'}) = \mathcal{D}_{\mathcal{KL}}(p_{\theta_m,\theta_{m'}}(y_m, y_{m'}) || p_{\theta_m}(y_m) \otimes p_{\theta_{m'}}(y_{m'})) \quad (1)$$

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \sum_{m \neq m'} \mathcal{L}_{MI}(g_m \circ f_m, g_{m'} \circ f_{m'}) \quad (2)$$

where $p_{\theta_m}(y|x)$ is the output probability distribution of a data sample from the m'th model with parameters θ_m. $\mathcal{D}_{\mathcal{KL}}$ refers to $\mathcal{KL}$ divergence between joint probability distribution of models m and m', and is calculated empirically [17]. Note that unlike "DivDis", our mutual info. loss is calculated on the same training/adaptation data. $\mathcal{L}_{CE}$ represents the sum of cross-entropy loss for all networks, and $\lambda = 10$ controls the importance of mutual info. loss.

Marginal Entropy Loss: Entropy minimization [26] and marginal entropy minimization [30] are promising proxy objectives for test-time robustness and adaptation to distribution shift with theoretically well established literature [7]. However, neural networks may not provide a well calibrated entropy, significantly undermining the effectiveness of these methods. Additionally, data augmentations may not be feasible or properly studied for a specific data type like ultrasound RF images, proposing new challenges to the problem. Therefore, we propose to find marginal entropy across deep ensemble networks instead of across various augmentations [30]. We first find marginal distribution by averaging M output probabilities: $\overline{p}(y|x) = \mathbb{E}_M[p_{\theta_m}(y|x)] = \frac{1}{M}\sum_{m=1}^{M} p_{\theta_m}(y|x)$. Then, the marginal entropy loss is calculated as: $\mathcal{L}_{\overline{H}} = -\sum_{c=1}^{C} \overline{p}^c(y|x) \times \log \overline{p}^c(y|x)$.

Test-Time Adaptation: To adapt to test distribution, we fine-tune the ensemble networks by minimizing both unsupervised objectives of marginal entropy

($\mathcal{L}_{\overline{H}}$) and mutual info. ($\mathcal{L}_{MI}$) losses on RF patches of each test biopsy core $X_i = \{x_{i1}, x_{i2}, ..., x_{in_i}\}$, separately. After S iteration of adaptation, with learning rate lr_{adapt}, the parameters of ensemble networks, i.e. θ_m are recovered to the values after initial training. This is equivalent to episodic adaptation on each biopsy core, and we leave online adaptation for future work due to the challenging hyperparameter tuning [31]. Lastly, we follow SOTA TTA methods [29] and substitute all batch norm layers [11] with batch-agnostic group norm layers [29]. More discussions on this effective approach are presented in the next sections.

4 Experiments

We organize our experiments in following parts: (i) *SOTA baseline comparison.* We compare our model's performance with prior SOTA methods for PCa detection in micro-US. This includes patch classification with single or with an ensemble of ResNet10 [5], patch classification with SSL pre-training on unlabeled data [27], and core classification with TRUSFormer [6] which collects patch extracted features with a transformer aggregator. (ii) *SOTA TTA comparison.* Additionally, we compare our model against SOTA TTA methods, including TENT [26] with ResNet10 backbone, MEMO [30] with marginal entropy calculated across various augmentations proposed in [27], TTT [25] with BYOL [8] as the SSL objective, and its extended meta learning version MT3 [2]. We also include "SAR" which filters unreliable samples based on entropy and minimizes a smoothed entropy loss [20]. (iii) *Ablation studies.* We perform two ablation studies and a qualitative analysis. First, we evaluate the contribution of each of our model's component to the performance. Second, we follow TTA methods [20] and study the effect of batch-agnostic norm layers. In particular, we replace ResNet10 batch norm [11] layers with group norm [29] and compare.

Evaluation Strategy: We evaluate all methods using a leave-one-center-out (LOCO) strategy, holding all data from one clinical center for testing in each experiment. This approach reflects real-world deployment. The remaining four centers are divided into training and validation with patient-wise stratification. To ensure balanced representation, the dataset from each center is first divided into training and validation sets and then combined. The performance is measured at core-level by averaging output probabilities of RF patches inside a core. Following previous works [5,22], cores with cancer involvement $\leq$ 40% are removed from test for more stable evaluation and AUROC and balanced-accuracy (sensitivity and specificity average) are reported. We additionally include AUROC-All where all test set cores, regardless of cancer involvement, are included.

Implementation Details: We re-implement all baseline models to evaluate them in a realistic LOCO setting with data distribution shift. We use PyTorch 2.1, with training and validation of our method taking approximately 4.6 h on a single NVIDIA A40 GPU. Inference on a biopsy core, including adaptation, takes around 250 ms. We manually tune hyperparameters based on validation. Adam optimizer [12] with learning rate of $1e-4$ and a scheduler with cosine annealing and a linear warm-up were used. For adaptation during inference, we employed SGD with the best learning rate selected from $lr_{adapt} = \{1e-1, 1e-2, 1e-3\}$ and number of iterations from $S = \{1, 5\}$. Additional details are provided in the supplementary document.

5 Results and Discussion

Table 1. PCa detection performance for baselines, TTAs, and our proposed method.

Method	AUROC	AUROC-All	Balanced-Acc.
Baselines			
ResNet10	75.2 ± 7.0	68.3 ± 6.0	68.0 ± 4.6
Ensem. [16]	75.8 ± 6.4	68.1 ± 6.3	68.0 ± 6.2
SSL + ResNet [27]	75.1 ± 4.0	67.7 ± 5.3	68.0 ± 4.3
TRUSFormer [6]	75.3 ± 4.9	70.4 ± 2.6	68.8 ± 4.0
Test-time adaptations			
TENT [26]	77.3 ± 4.2	71.0 ± 3.6	69.7 ± 5.4
MEMO [30]	77.4 ± 4.2	70.9 ± 3.6	69.4 ± 6.4
TTT [25]	77.8 ± 5.1	71.1 ± 3.4	66.5 ± 8.8
MT3 [2]	77.7 ± 5.7	70.3 ± 4.4	63.3 ± 4.8
SAR [20]	77.6 ± 4.4	71.2 ± 3.7	70.7 ± 5.0
DEnEM (ours)	$\mathbf{80.9 \pm 4.5}$	$\mathbf{75.0 \pm 2.9}$	$\mathbf{75.6 \pm 3.4}$

Table 2. Ablation study on different components of the proposed method.

Method	Group-Norm	Ensem.	Mutual-Info.	Test-Adapt	AUROC-All
ResNet10	✓	–	–	–	71.7 ± 3.7
Ensem.	✓	✓	–	–	72.9 ± 4.1
Ensem.+$\mathcal{L}_{MI}$	✓	✓	✓	–	73.8 ± 3.5
Ensem.+$\mathcal{L}_{\overline{H}}$	✓	✓	-	✓	73.1 ± 4.0
Ensem.+$\mathcal{L}_{MI}$+$\mathcal{L}_{\overline{H}}$ (ours)	✓	✓	✓	✓	75.0 ± 2.9

SOTA Baseline Comparison: Table 1 summarizes the results of SOTA comparison. Overall, our proposed method (DEnEM) outperforms all baselines by a significant margin in core AUROC, AUROC-All, and Balanced-ACC by around 5% to 7%. In particular, our patch classification approach outperforms SOTA core classification TRUSFormer by ∼5% indicating the effectiveness of DEnEM. Future works may leverage the orthogonal benefits of these two methods. Moreover, comparing the baselines shows that the leading method in PCa detection may fall behind relatively simpler approaches in LOCO setting, highlighting the importance of this realistic evaluation. Lastly, note that the high standard deviations are due to averaging the results of *different* test datasets with high variability in performance. In ablation study, we will show the standard deviation for each individual test datasets.

SOTA TTA Comparison: Similar to baseline comparisons, our proposed method substantially improves over SOTA TTA methods in all metrics by 3% to 5%. Comparing TTA methods show that TENT and MEMO produce similar results, indicating that marginal entropy across various augmentations is not effective. Considering this in conjunction with SSL-pretraining results, we hypothesize that the proposed augmentations for RF images [27] might be ineffective. In conclusion, the performance of DEnEM compared with marginal differences between TTA results justifies the proposed marginal entropy in DEnEM.

Ablation Studies: Table 2 shows the effect of different components on the performance of our method. Overall, each component increases the performance individually, with $\mathcal{L}_{MI}$ improving by 0.9% and $\mathcal{L}_{\overline{H}}$ improving by 0.2%. However, when combined together, the gain is compounded with improvement of 2.1%. Unlike the baselines, deep ensemble improves AUROC-All over ResNet10 when group norm is adopted. This expected outcome demonstrates the harmful effect of batch norm on baselines with LOCO setting. Recent TTA methods [20] have also highlighted the potential drawbacks of batch norm under distribution shift as the mean and variance estimation in batch norm layers will be biased. In Fig. 2 (b), we compare the AUROC-All of ResNet10 using batch norm versus group norm (used in our experiment) with separate bar plots for each center left out for test. This figure further confirms the substantial improvement of replacing batch norm with group norm. Additionally, it reveals a high variability in performance, ranging from 68% to 77% across different centers. We hypothesize that the quantity and quality of data in both training and test sets contribute to this variability, though further research is needed in future works.

Qualitative analysis: We qualitatively analyze our model's prediction by using heatmaps, comparing the predictions when no TTA is adopted (e.g. ResNet10 baseline) with DEnEM, as detailed in Fig. 2 (a). These heatmaps are created by running the model across the entire RF image using sliding patches. They showcase two examples of benign (top row) and cancerous (bottom row) cores with red indicating cancer predictions and blue for benign predictions. These maps, overlaid on corresponding B-mode images, not only demonstrate the failure of baseline compared to DEnEM, but also qualitatively show the capability of our model in localizing cancer and offering reliable guidance for targeted biopsy.

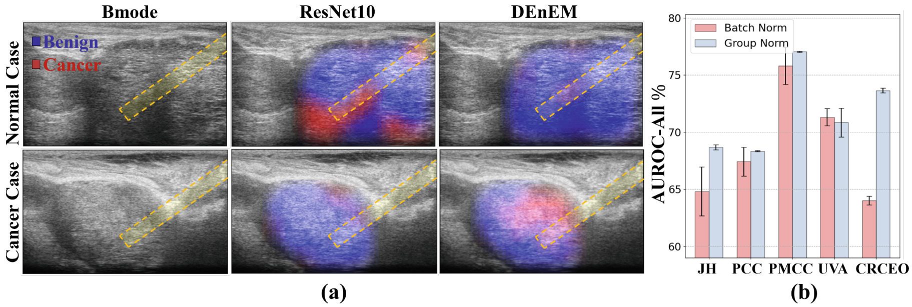

Fig. 2. (a) Heatmap comparison of ResNet10 and DEnEM with cancer (red) vs. benign (blue) areas. Top row: benign core; bottom row: cancerous core (Gleason score 3+4). (b) Baseline ResNet10 Batch norm vs. group norm comparison for different test center. (Color figure online)

6 Conclusion

This study examined the robustness of deep learning models to clinical center distribution shifts in micro-US PCa detection. We showed that existing methods are vulnerable to these shifts, with leading approaches sometimes outperformed by simpler ones. To address this, we adopted TTA, which improved detection but faced limitations due to ultrasound augmentations and biased entropy of neural networks. We proposed DEnEM, which calculates calibrated marginal entropy across a diverse ensemble of networks without requiring data augmentations. DEnEM significantly enhanced performance, outperforming previous TTA methods and demonstrating its potential to improve PCa detection robustness.

References

1. Abouassaly, R., Klein, E.A., El-Shefai, A., Stephenson, A.: Impact of using 29 mhz high-resolution micro-ultrasound in real-time targeting of transrectal prostate biopsies: initial experience. World J. Urol. **38**(5), 1201–1206 (2020)
2. Bartler, A., Bühler, A., Wiewel, F., Döbler, M., Yang, B.: Mt3: meta test-time training for self-supervised test-time adaption. In: International Conference on Artificial Intelligence and Statistics, pp. 3080–3090. PMLR (2022)
3. Cotter, F., Perera, S., Sathianathen, N., Lawrentschuk, N., Murphy, D., Bolton, D.: Comparing the diagnostic performance of micro-ultrasound-guided biopsy versus multiparametric magnetic resonance imaging-targeted biopsy in the detection of clinically significant prostate cancer: A systematic review and meta-analysis. Société Internationale d'Urologie Journal **4**(6), 465–479 (2023)
4. Ghai, S., et al.: Assessing cancer risk on novel 29 mhz micro-ultrasound images of the prostate: creation of the micro-ultrasound protocol for prostate risk identification. J. Urol. **196**(2), 562–569 (2016)
5. Gilany, M., Wilson, P., Jamzad, A., Fooladgar, F., To, M.N.N., Wodlinger, B., Abolmaesumi, P., Mousavi, P.: Towards confident detection of prostate cancer using high resolution micro-ultrasound. In: International Conference on Medical

Image Computing and Computer Assisted Intervention, pp. 411–420. Springer (2022). https://doi.org/10.1007/978-3-031-16440-8_40

6. Gilany, Met al.: Trusformer: improving prostate cancer detection from micro-ultrasound using attention and self-supervision. Inter. J. Comput. Assisted Radiol. Surgery, 1–8 (2023)
7. Goyal, S., Sun, M., Raghunathan, A., Kolter, J.Z.: Test time adaptation via conjugate pseudo-labels. Adv. Neural. Inf. Process. Syst. **35**, 6204–6218 (2022)
8. Grill, J.B., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural. Inf. Process. Syst. **33**, 21271–21284 (2020)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. pmlr (2015)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. arXiv preprint arXiv:2204.02937 (2022)
14. Koh, P.W., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning, pp. 5637–5664. PMLR (2021)
15. Kouw, W.M., Loog, M.: A review of domain adaptation without target labels. IEEE Trans. Pattern Anal. Mach. Intell. **43**(3), 766–785 (2019)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inform. Process. Syst. **30** (2017)
17. Lee, Y., Yao, H., Finn, C.: Diversify and disambiguate: out-of-distribution robustness via disagreement. In: The Eleventh International Conference on Learning Representations (2022)
18. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. arXiv preprint arXiv:2303.15361 (2023)
19. Michalski, J.M., Pisansky, T.M., Lawton, C.A., Potters, L.: Chapter 53 - prostate cancer. In: Gunderson, L.L., Tepper, J.E. (eds.) Clinical Radiation Oncology (Fourth Edition), pp. 1038–1095.e18. Elsevier, Philadelphia, fourth edition edn. (2016)
20. Niu, S., et al.: Towards stable test-time adaptation in dynamic wild world. In: The Eleventh International Conference on Learning Representations (2023)
21. Ovadia, Y., et al.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Adv. Neural Inform. Process. Syst. **32** (2019)
22. Rohrbach, D., Wodlinger, B., Wen, J., Mamou, J., Feleppa, E.: High-frequency quantitative ultrasound for imaging prostate cancer using a novel micro-ultrasound scanner. Ultrasound Med. Biol. **44**(7), 1341–1354 (2018)
23. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
24. Shao, Y., Wang, J., Wodlinger, B., Salcudean, S.E.: Improving prostate cancer (pca) classification performance by using three-player minimax game to reduce data source heterogeneity. IEEE Trans. Med. Imaging **39**(10), 3148–3158 (2020)

25. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning, pp. 9229–9248. PMLR (2020)
26. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
27. Wilson, P.F., et al.: Self-supervised learning with limited labeled data for prostate cancer detection in high frequency ultrasound. IEEE Trans. Ultrasonics Ferroelectrics Frequency Control (2023)
28. Wilson, P., et al.: Toward confident prostate cancer detection using ultrasound: a multi-center study. Inter. J. Comput. Assisted Radiol. Surgery (2024)
29. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19 (2018)
30. Zhang, M., Levine, S., Finn, C.: Memo: test time robustness via adaptation and augmentation. Adv. Neural. Inf. Process. Syst. **35**, 38629–38642 (2022)
31. Zhao, H., Liu, Y., Alahi, A., Lin, T.: On pitfalls of test-time adaptation. arXiv preprint arXiv:2306.03536 (2023)