Deep EM with Hierarchical Latent Label Modelling for Multi-Site Prostate Lesion Segmentation

Wen Yan[1], Yipei Wang[1], Shiqi Huang[1], Natasha Thorley[2], Mark Emberton[3], Vasilis Stavrinides[4,5], Shonit Punwani[2], Yipeng Hu[1], Dean Barratt[1]

[1] UCL Hawkes Institute; Department of Medical Physics and Biomedical Engineering, University College London, Gower St, London WC1E 6BT, U.K.
wen-yan@ucl.ac.uk

[2] Centre for Medical Imaging, Division of Medicine, University College London, Gower St, London WC1E 6BT, U.K.

[3] Division of Surgery and Interventional Science, University College London, 10 Pond St, London NW3 2PS, U.K.

[4] Cancer Institute, Urology Department, UCLH, University College London, 16-18 Westmoreland St, London W1G 8PH, U.K.

[5] Radiology Department, Imperial College Healthcare, The Bays, S Wharf Rd, London W2 1NY, U.K.

Abstract. Label variability is one of the major challenges in prostate lesion segmentation. In multi-site datasets, annotations often reflect site-specific contouring protocols. This can bias segmentation networks towards local annotation styles and reduce generalisation to unseen sites during inference. In this study, we treat each observed annotation as a noisy observation of an underlying latent "clean" lesion mask, and propose a hierarchical expectation-maximisation (HierEM) framework that alternates between: (i) inferring a voxel-wise posterior distribution over the latent mask, (ii-1) training a UNet that uses this posterior as a soft target, and (ii-2) estimating site-specific sensitivity and specificity with a logistic-normal hierarchical prior. The hierarchical prior decomposes label quality into global, site- and case-level components, which regularises site-level deviations, encouraging the model to avoid overfitting to centre-specific annotation styles, thereby promoting site-invariant segmentation. Experiments on three-site datasets demonstrate that the proposed HierEM framework enhances cross-site generalisation compared to state-of-the-art methods. In the pooled datasets evaluation, per-site mean DSC ranges from 29.50% to 39.69% across datasets; for leave-one-site-out generalisation, it ranges from 27.91% to 32.67%, yielding statistically significant improvements over comparison methods ($p < 0.039$). The method also produces interpretable per-site latent label-quality estimates (e.g., sensitivity α ranging from 31.5% to 47.3% at specificity $\beta \approx 0.99$). These results indicate that explicitly modelling site-dependent annotation can improve cross-site generalisation. Code is available via https://github.com/yanwenCi/HierEM.

Keywords: Lesion segmentation · Hierarchical latent label · EM.

1 Introduction

Multiparametric MRI (mpMRI), including T2-weighted (T2W), diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) maps and contrast-enhanced (DCE) modalities, is widely used to assess suspected prostate cancer (PCa) [2]. Clinicians often contour lesions to guide diagnosis, treatment planning and follow-up. Automated deep-learning methods aim to reduce this workload, yet they suffer from substantial ground-truth variability. This label variability is rooted in site-specific annotation protocols, institutional contouring styles shaped by local expert training, and imaging protocols that influence how lesion boundaries are perceived. Together, these factors create site-specific label bias. Inter-reader agreement for prostate lesion contours on mpMRI is only moderate, with reported Dice often around 40% [5,8]. The current paradigm creates a critical bottleneck: segmentation networks easily overfit to the local contouring style of the training sites. When deployed to a new institution, the model generalises poorly, as single-site training achieves only $4\% - 28\%$ Dice on a held-out site [10]. By contrast, within-site testing achieves around $20 - 60\%$ Dice [16,3,13,1] with different datasets. This gap can be alleviated via test-site finetuning or calibration. However, such methods often lead to test-site-biased accuracy assessments. They force the model to match the local "observed" labels, which are themselves imperfect and biased. A truly multi-site training set (more than 10 sites) is expensive to curate for individual tasks that require coordination in labelling. PI-CAI is an example study with such a diverse collection of reference standards, but restricted to a primarily patient-level classification [12]. In many real-world applications, further finetuning or calibration at deployment may be infeasible or impractical. While cross-site performance gaps may also stem from both imaging differences and label variation, pre-processing can partially mitigate the former. In this study, we focus on the impact of label-level variation across sites.

We propose a latent-label modelling approach to address cross-site annotation variability using site-level sensitivity and specificity parameters. Our use of sensitivity and specificity is related to the classical formulation used in STAPLE [15], but the problem setting is fundamentally different. STAPLE addresses multi-annotator label fusion for a fixed image, where multiple annotations of the same case are available and the goal is to infer a clean label while estimating annotator reliability. In contrast, our setting typically contains only one annotation per case, and the annotation source is determined by the acquisition site rather than by multiple independent raters. We therefore do not perform label fusion. Instead, we model each observed site-specific annotation as a noisy label of an unobserved latent lesion mask during training. The segmentation network estimates this latent mask, while site-dependent sensitivity and specificity parameters account for systematic differences in annotation behaviour across sites. Thus, sensitivity and specificity are used not for multi-rater fusion, but to characterise site-dependent label noise and support learning of a site-agnostic lesion representation from single, site-dependent noisy labels. We impose logistic-normal hierarchical priors on the sensitivity and specificity parameters, which are decomposed into three components: (1) global population factors capturing

lesion characteristics shared across sites; (2) site-specific effects modelling systematic offsets due to local contouring and imaging protocols; and (3) case-level variability capturing intrinsic ambiguity, such as small or low-contrast lesions, that may affect annotation independently of site protocol. This decomposition enables partial pooling across multi-site data while accounting for both systematic site-level variation and individual-case ambiguity.

Learning is performed with an EM procedure [6]. In the E-step, we infer a voxel-wise posterior over the latent clean mask by combining the network's image-based prediction with the likelihood of the observed site-specific annotation. In the M-step, this posterior is used as a soft target to update the UNet, while the hierarchical sensitivity and specificity parameters are re-estimated to capture site- and case-dependent label variability. By explicitly modelling site-specific contouring behaviour as label noise, HierEM encourages the network to learn more site-agnostic lesion representations and provides interpretable estimates of annotation tendencies across sites and cases, thereby improving robustness when deploying to unseen sites without requiring additional site-specific fine-tuning.

2 Method

Let Ω denote the voxel space of an mpMRI volume. For each case $k \in \{1, \ldots, K\}$ we observe an image X_k and a single binary lesion annotation $Y_k \in \{0,1\}^{\Omega}$ provided by one site $s_k \in \{1, \ldots, S\}$. The (unobserved) latent "clean" lesion mask is $G_k \in \{0,1\}^{\Omega}$. Our goal is to learn a segmentation model that is robust to site-specific annotation protocol, while also estimating site-/case-level sensitivity and specificity for observed labels relative to the latent true labels.

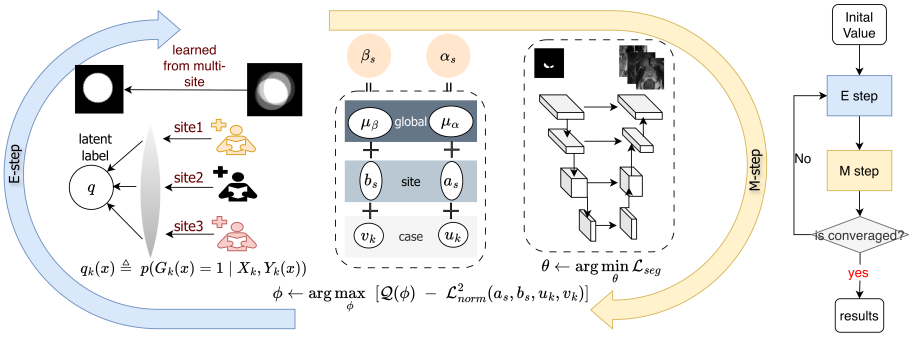

Fig. 1. Overview of proposed deep mixed EM framework with hierarchical latent label-quality parameters. The E-step infers a latent clean mask, and the M-step updates the segmentation network and site-specific sensitivity and specificity.

Segmentation network We model the latent clean mask with an image-conditioned voxel-wise probability map predicted by the network: $\pi_k(x) \triangleq p_\theta(G_k(x) = 1 \mid X_k) = \sigma(f_\theta(X_k)_x)$, where $x \in \Omega$ is a voxel in mpMRI, $f_\theta(\cdot)$ is a neural network producing logits per voxel and $\sigma(\cdot)$ is the sigmoid.

2.1 Hierarchical site-case label variability model

We treat each site as having a distinct annotation protocol and model the observed label $Y_k(x)$ as a noisy observation of $G_k(x)$. Conditioned on (G_k, s_k), voxels are assumed independent with site- and case-level sensitivity and specificity:

$$p(Y_k(x) = 1 \mid G_k(x) = 1) = \alpha_{s_k,k}, \quad p(Y_k(x) = 0 \mid G_k(x) = 0) = \beta_{s_k,k}, \ \forall x \in \Omega_k. \tag{1}$$

To stabilise site-/case-level label-quality estimates, we use a logistic-normal hierarchical prior that partially pools sensitivity and specificity towards global means μ_α, μ_β:

$$\text{logit}(\alpha_{s,k}) = \mu_\alpha + a_s + u_k, \qquad \text{logit}(\beta_{s,k}) = \mu_\beta + b_s + v_k, \tag{2}$$

where (a_s, b_s) capture persistent site-specific contouring behaviour and (u_k, v_k) capture case difficulty. Eq. (2) decomposes annotation reliability into three hierarchical components: global annotation tendency, site-level labelling differences, and case-specific uncertainty. This decomposition allows HierEM to distinguish common annotation tendencies from site-dependent bias and individual-case difficulty. We place zero-mean Gaussian priors (equivalently, ℓ_2 penalties under MAP): $a_s \sim \mathcal{N}(0, \sigma_a^2)$, $b_s \sim \mathcal{N}(0, \sigma_b^2)$, $u_k \sim \mathcal{N}(0, \sigma_u^2)$, $v_k \sim \mathcal{N}(0, \sigma_v^2)$. For identifiability we constrain $\sum_s a_s = 0$, $\sum_s b_s = 0$. We further calculate site-level sen/spec as ($\sigma(\cdot)$ is sigmoid function): $\alpha_s = \sigma(\mu_\alpha + a_s)$, $\beta_s = \sigma(\mu_\beta + b_s)$.

EM learning with hierarchical sensitivity/specificity modelling We estimate segmentation network parameters θ and the latent label-quality parameters $\phi = \{\mu_\alpha, \mu_\beta, a_{1:S}, b_{1:S}, u_{1:K}, v_{1:K}\}$ by maximizing the (regularized) marginal likelihood of $\{Y_k\}$ with latent $\{G_k\}$. We optimise a MAP objective via an EM procedure.

E-step: posterior of the latent mask. Given current (θ, ϕ), the posterior factorizes over voxels:

$$q_k(x) \triangleq p(G_k(x) = 1 \mid X_k, Y_k(x)) \tag{3}$$

$$= \frac{\pi_k(x)\, p(Y_k(x) \mid G_k(x) = 1)}{\pi_k(x)\, p(Y_k(x) \mid G_k(x) = 1) + (1 - \pi_k(x))\, p(Y_k(x) \mid G_k(x) = 0)}. \tag{4}$$

With the sensitivity and specificity model, the likelihood terms are

$$p(Y{=}1 \mid G{=}1) = \alpha_{s,k}, \quad p(Y{=}1 \mid G{=}0) = 1 - \beta_{s,k}, \tag{5}$$

$$p(Y{=}0 \mid G{=}1) = 1 - \alpha_{s,k}, \quad p(Y{=}0 \mid G{=}0) = \beta_{s,k}. \tag{6}$$

$$p(Y|G;\alpha,\beta) = \begin{cases} \alpha^Y(1-\alpha)^{1-Y}, & G=1, \\ (1-\beta)^Y\beta^{1-Y}, & G=0. \end{cases} \tag{7}$$

Thus $q_k(x)$ provides a *soft clean* mask that fuses the image prior and the site/case-level sensitivity and specificity.

M-step (A): update segmentation network. We update the segmentation network with θ by minimising a soft-label cross-entropy plus Dice loss:

$$\mathcal{L}_{\text{seg}}(\theta) \;=\; \sum_{k=1}^{K}\sum_{x\in\Omega} \text{CE}\big(q_k(x), \pi_k(x)\big) - \sum_{k=1}^{K} \text{Dice}(q_k(x), \pi_k(x)), \tag{8}$$

where $\text{CE}(q,\pi) = -q\log\pi - (1-q)\log(1-\pi)$ and $\text{Dice}(q,\pi) = \frac{\sum_{x\in\Omega} 2q_k(x)\cdot\pi_k(x)}{\sum_{x\in\Omega} q_k(x) + \sum_{x\in\Omega}\pi_k(x)}$

M-step (B): update hierarchical latent label-quality by aggregated expected counts. Define expected per-case sufficient statistics:

$$TP_k = \sum_{x\in\Omega} q_k(x)\, Y_k(x), \qquad P_k = \sum_{x\in\Omega} q_k(x), \tag{9}$$

$$TN_k = \sum_{x\in\Omega} (1-q_k(x))\,(1-Y_k(x)), \qquad N_k = \sum_{x\in\Omega}(1-q_k(x)). \tag{10}$$

Using posterior q to approximate latent Ground Truth G, the expected complete-data log-likelihood for latent label-quality parameters is:

$$\begin{aligned} \mathcal{Q}(\phi) &= \mathbb{E}_{G\sim q}\big[\log p(Y|G;\phi)\big] && (11)\\ &= \sum_{x\in\Omega}\Big[\mathbb{1}\{G_k(x)=1\}\Big(Y_k(x)\log\alpha_{s_k,k} + \big(1-Y_k(x)\big)\log\big(1-\alpha_{s_k,k}\big)\Big) \\ &\quad + \mathbb{1}\{G_k(x)=0\}\Big(\big(1-Y_k(x)\big)\log\beta_{s_k,k} + Y_k(x)\log\big(1-\beta_{s_k,k}\big)\Big)\Big] && (12)\\ &= \sum_{k=1}^{K}\Big[TP_k\log\alpha_{s_k,k} + (P_k - TP_k)\log(1-\alpha_{s_k,k}) \\ &\quad + TN_k\log\beta_{s_k,k} + (N_k - TN_k)\log(1-\beta_{s_k,k})\Big], && (13) \end{aligned}$$

with $\alpha_{s,k}, \beta_{s,k}$ defined by (2). We obtain a MAP estimate by maximising MLE and L2 penalisation on prior variables in ϕ:

$$\phi \leftarrow \arg\max_{\phi}\; [\mathcal{Q}(\phi) \;-\; \mathcal{L}^2_{norm}], \tag{14}$$

where $\mathcal{L}^2_{norm} = \frac{1}{2\sigma_a^2}\sum_s a_s^2 + \frac{1}{2\sigma_b^2}\sum_s b_s^2 + \frac{1}{2\sigma_u^2}\sum_k u_k^2 + \frac{1}{2\sigma_v^2}\sum_k v_k^2$ subject to $\sum_s a_s = 0$ and $\sum_s b_s = 0$. Because (14) depends on the data only via (TP_k, P_k, TN_k, N_k), it can be optimised efficiently with a few iterations of a second-order L-BFGS [7] over low-dimensional parameters. L_2 penalty on the hierarchical latent variables in ϕ corresponds to a Gaussian prior and provides shrinkage that stabilises site/case quality estimates, prevents degenerate sensitivity and specificity values, and mitigates overfitting when annotations are sparse or noisy.

Algorithm 1 Hierarchical EM for multi-site lesion segmentation

Require: Images $\{X_k\}$, masks $\{Y_k\}$, site labels $\{s_k\}$
1: Initialize network θ; initialize ϕ (μ_α, μ_β near logit(0.9), $a_s = b_s = u_k = v_k = 0$)
2: **for** EM iterations $t = 1, \ldots, T$ **do**
3: **E-step:** compute $\pi_k(x) = \sigma(f_\theta(X_k)_x)$ and $q_k(x)$ via (4)
4: Compute (TP_k, P_k, TN_k, N_k) via (9)–(10)
5: **M-step (A):** update θ by minimizing $\mathcal{L}_{\text{seg}}(\theta)$ in (8)
6: **M-step (B):** update ϕ by MAP optimization of (14), subject to $\sum_s a_s = 0$ and $\sum_s b_s = 0$
7: **end for**
8: **Output:** network θ and learned site-/case-level parameters ϕ

Training procedure We alternate the E-step (4) with M-steps (8) and (14) until convergence. We used the Adam optimiser with a learning rate of 10^{-4}. In practice, we perform 5-epoch gradient steps for θ per EM iteration, and 5 optimiser steps for ϕ using the aggregated sufficient statistics in (13). Algorithm 1 summarises the procedure. To stabilise early optimisation, we add an auxiliary supervision in (8) using the observed label. The corresponding loss weight was gradually decayed to zero over the first five epochs using a sigmoid schedule.

2.2 Uncertainty and interpretability

We quantify voxel-wise uncertainty for case k using the predictive entropy of the segmentation probability $\pi_{k,i}$: $H_{k,i} = -\pi_{k,i} \log \pi_{k,i} - (1 - \pi_{k,i}) \log(1 - \pi_{k,i})$. Entropy is low for confident predictions and maximal at $\pi_{k,i} = 0.5$. Risk-coverage curves are obtained by retaining the lowest-entropy fraction $c \in [0.5, 0.9]$ of voxels, $S_{k,c} = \{i : H_{k,i} \leq t_{k,c}\}$, $|S_{k,c}|/|\Omega_k| = c$, and averaging the risk over test cases: $\text{Risk}(c) = \frac{1}{K} \sum_{k=1}^{K} (1 - \text{Dice}(S_{k,c}))$.

3 Experiment

3.1 Datasets

We utilised datasets from $S = 3$ sites. Each site provides T2W, $\text{DWI}_{\text{high } b}$, and ADC images, together with a single expert lesion contour annotated on T2W. Site 1 was obtained from UCLH with ethical approval, described in [16]. The dataset comprises 850 patients of 1201 mpMR combinations with multiple high b-values, each with PI-RADS ≥ 3 lesions. Site 2 is a public cohort [14], which includes 391 patients who have Likert $\geq$3 lesions. Site 3 was collected under ethically approved clinical protocols across multiple institutions in Miami, FL, containing 1265 mpMRI scans with PIRADS ≥ 3 lesion contours.

Split A (Pooled patient-level held-out evaluation). We split each dataset by 4:1:1, resulting in pooled mixed train, validation and per-patient level held-out sets. The held-out sets were only used for model evaluation.

Split B (Leave-one-site-out, LOSO, cross-site generalisation). We perform leave-one-site-out evaluation across the 3 sites. For each fold, we hold out all cases from one site as test, and train on the remaining two sites. Within the training pool, we reserve 10% of cases (stratified by site) as validation for early stopping and hyperparameter tuning.

3.2 Baselines and evaluation metrics

Baselines In this study, to ensure a fair and transparent comparison across methods, we use UNet [11] with nnUNet-derived preprocessing and hyperparameters [4] as the segmentation backbone for the proposed method and all compared methods. We compare against: (i) Supervised UNet; (ii) label bootstrapping (soft self-training) [9], where the standard supervised loss is replaced by a bootstrapped self-training loss that combines the provided labels with the model's own predictions as soft auxiliary targets: $\mathcal{L}_{bootstrap}^{t+1} = (1-\lambda)\mathcal{L}_{seg}(\omega^{t+1}, y) + \lambda\mathcal{L}_{seg}(\omega^{t+1}, \omega^{t})$, where ω^t denotes the network prediction at iteration t, and λ is a ramp-up hyperparameter with a maximum value of 0.5. This encourages smoother decision boundaries and reduces sensitivity to annotation noise. (iii) Site-as-reader EM without hierarchy, where independent α_s, β_s are estimated for each site rather than using the hierarchical formulation in Eq. (2). This progressive comparison serves both as a baseline comparison and as an ablation study, isolating the contribution of the key components: soft-label self-training, explicit EM-based noise modelling, and hierarchical site/case reliability modelling.

Evaluation metrics for segmentation and uncertainty On each test fold and site, we report Dice and HD95 for segmentation results. For calibration and uncertainty metrics, we report selective segmentation (risk-coverage) curves based on predicted uncertainty (entropy of $\pi_k(x)$). We compare methods using a paired t-test on the same test cases within each held-out site. Multiple comparisons are controlled within each held-out site using Benjamini-Hochberg FDR at 5%.

4 Results

Table 1 reports cross-site performance using Dice and HD95. Under the pooled patient-level held-out split, all methods achieve moderate Dice scores, approximately from 27% to 40%, while HierEM gives the best mean Dice on all three sites (Site 1: 39.69%, Site 2: 29.50%, Site 3: 35.60%) and comparable or slightly improved HD95 (e.g., Site 2: 20.46 vs 21.25 mm for Supervised). Bootstrap performs similarly to Supervised on the pooled split. In the pooled held-out split, significance is not observed for every comparison, likely due to limited test-set size. In addition, pooled multi-site training improves baseline performance, so our method shows a smaller margin in this setting. Under the more challenging LOSO evaluation, performance drops substantially for methods trained without explicit cross-site label variability modelling. Supervised Dice decreases to 25.50/24.66/31.20% across Sites 1-3, with a concurrent increase in boundary

Table 1. Sites 1-3 contained 252, 87, and 265 cases in the pooled patient-level held-out set, and 1201, 391, and 1265 cases in the LOSO experiments, respectively. Asterisks denote baseline results for which HierEM is significantly better at $\alpha = 0.05$.

Methods	Site 1		Site 2		Site 3	
	Dice(%)	HD95(mm)	Dice(%)	HD95(mm)	Dice(%)	HD95(mm)
	Split A: Pooled patient-level held-out split					
UNet	38.63±10.41	17.04±2.96	27.47±9.62*	21.25±4.50	33.49±6.98*	22.06±3.92
Bootstrap	38.49±9.94	16.97±3.14	26.69±10.59*	22.29±5.12*	33.14±7.05*	**21.41±5.84**
Site-EM	36.46±11.72*	18.41±7.07*	23.74±9.73*	22.99±5.52*	34.62±6.15	21.86±4.91
HierEM	**39.69±11.88**	**16.92±3.29**	**29.50±9.13**	**20.46±5.58**	**35.60±7.85**	21.66±4.11
	Split B: Leave-one-site-out (LOSO)					
UNet	25.50±7.33*	22.10±4.65*	24.66±5.29*	22.46±3.15*	31.20±9.80*	**21.05±5.64**
Bootstrap	26.34±12.92*	21.02±7.02*	21.66±5.20*	22.54±3.56*	31.64±8.92*	22.85±5.17
Site-EM	24.00±18.09*	23.74±10.79*	24.92±13.29*	21.83±7.13	31.90±13.43*	22.50±7.28
HierEM	**28.11±10.21**	**20.51±4.68**	**27.91±10.16**	**21.12±5.83**	**32.67±8.08**	22.82±2.50

error compared to pooled held-out testing. In contrast, HierEM consistently improves generalisation, yielding higher Dice on Sites 1 and 2 (28.11% and 27.91%) with lower HD95 (20.51 and 21.12 mm), while remaining competitive on Site 3 (Dice 32.67%; HD95 22.82 mm). Overall, HierEM provides the most consistent cross-site performance, especially in the LOSO setting. Fig 2-left shows risk-coverage curves that assess uncertainty for selective segmentation. HierEM has lower risks than supervised UNet at a given coverage range for all three LOSO test sites, which indicates that the method's uncertainty concentrates errors in the rejected region, enabling reliable abstention. In Figure 2-right, HierEM achieved the highest sensitivity under the matched specificity ($\simeq 0.99$) in all LOSO experiments. The gap between pooled and LOSO highlights a strong

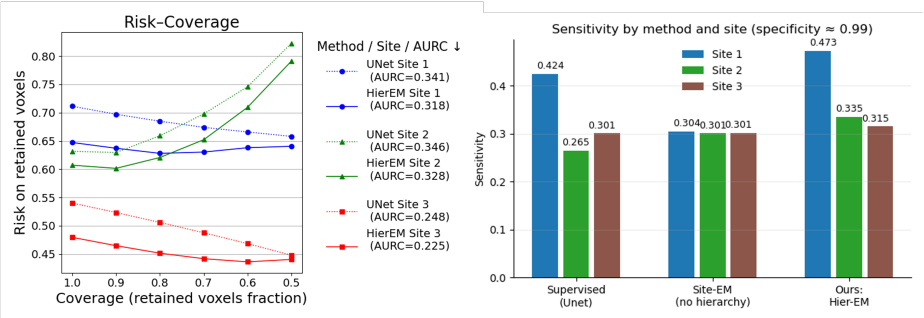

Fig. 2. Left: Risk-coverage curves on three-site datasets under LOSO experiments, where smaller AURC represents lower risk; Right: Estimated sensitivity of baselines and HierEM with *spec* $\simeq 0.99$. Such high specificity is expected because lesions occupy only a small fraction of the image.

domain shift: models that perform reasonably well on pooled held-out testing degrade when the test site is unseen, suggesting that performance in pooled splits largely comes from learning site-specific contouring biases rather than a fully transferable latent "clean" lesion mask. HierEM narrows this gap, supporting the hypothesis that explicitly modelling site/case-dependent label noise helps separate site-specific annotation from image evidence. By inferring a soft latent clean mask (q) and updating sensitivity and specificity, HierEM reduces overfitting to training-site annotation styles when applied to a new site, which is reflected in both higher Dice and generally competitive HD95 on unseen sites.

5 Conclusion

We propose a Deep EM framework with a hierarchical prior that treats each site's lesion contour as a noisy observation of a latent clean mask. The E-step infers a voxel-wise posterior over this latent mask, while the M-step updates the segmentation network and estimates site-specific sensitivity and specificity using hierarchical priors. Rather than representing individual-reader performance, the estimated sensitivity and specificity represent pooled site-level labelling behaviour, capturing the effective annotation tendency of each dataset/site. By separating latent labels from site- and case-dependent annotation variability, HierEM improves robustness and cross-site generalisation while providing interpretable diagnostics of site-level annotation behaviour. The framework is backbone-agnostic, and future work will extend it to multi-site, multi-annotator datasets and richer models of clinical annotation variability.

Acknowledgments. This work is supported by the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK [C28070/A30912; C73666/A31378; EDDAMC-2021/100011], Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fouladi, S., Darvizeh, F., Gianini, G., Di Meo, R., Di Palma, L., Damiani, E., Maiocchi, A., Fazzini, D., Alì, M.: Exploring UNet-based models for prostate lesion segmentation from multi-sequence MRI (T2W, ADC, DWI). World Wide Web **29**(1), 4 (2026)
2. Haider, M.A., Van Der Kwast, T.H., Tanguay, J., Evans, A.J., Hashmi, A.T., Lockwood, G., Trachtenberg, J.: Combined t2-weighted and diffusion-weighted MRI for localization of prostate cancer. American journal of roentgenology **189**(2), 323–328 (2007)
3. Hambarde, P., Talbar, S., Mahajan, A., Chavan, S., Thakur, M., Sable, N.: Prostate lesion segmentation in MR images using radiomics based deeply supervised UNet. Biocybernetics and Biomedical Engineering **40**(4), 1421–1435 (2020)

4. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
5. Liechti, M.R., Muehlematter, U.J., Schneider, A.F., Eberli, D., Rupp, N.J., Hötker, A.M., Donati, O.F., Becker, A.S.: Manual prostate cancer segmentation in MRI: interreader agreement and volumetric correlation with transperineal template core needle biopsy. European radiology **30**(9), 4806–4815 (2020)
6. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal processing magazine **13**(6), 47–60 (1996)
7. Nocedal, J.: Updating quasi-newton matrices with limited storage. Mathematics of computation **35**(151), 773–782 (1980)
8. Piert, M., Shankar, P.R., Montgomery, J., Kunju, L.P., Rogers, V., Siddiqui, J., Rajendiran, T., Hearn, J., et al.: Accuracy of tumor segmentation from multiparametric prostate MRI and 18F-choline PET/CT for focal prostate cancer therapy applications. EJNMMI research **8**(1), 23 (2018)
9. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014)
10. Rodrigues, N.M., de Almeida, J.G., Castro Verde, A.S., Mascarenhas Gaivão, A., Bilreiro, C., Santiago, I., Ip, J., Belião, S., Moreno, R., et al.: Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data. Computers in Biology and Medicine p. 108216 (2024). https://doi.org/10.1016/j.compbiomed.2024.108216
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
12. Saha, A., Bosma, J., Twilt, J., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., et al.: Artificial intelligence and radiologists at prostate cancer detection in MRI-the PI-CAI challenge. In: Medical Imaging with Deep Learning, short paper track (2023)
13. Wang, W., Pan, B., Ai, Y., Li, G., Fu, Y., Liu, Y.: Paracm-pnet: A cnn-tokenized mlp combined parallel dual pyramid network for prostate and prostate cancer segmentation in MRI. Computers in Biology and Medicine **170**, 107999 (2024)
14. Wang, Y., Gayo, I., Thorley, N., Ng, A., Yan, W., Barratt, D., Stavrinides, V., Emberton, M., Kasivisvanathan, V., Punwani, S., Hu, Y.: A derived promis data set (Jun 2025). https://doi.org/10.5281/zenodo.15683922, https://doi.org/10.5281/zenodo.15683922
15. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)
16. Yan, W., Chiu, B., Shen, Z., Yang, Q., Syer, T., Min, Z., Punwani, S., Emberton, M., et al.: Combiner and hypercombiner networks: Rules to combine multimodality MR images for prostate cancer localisation. Medical Image Analysis **91**, 103030 (2024)