LensePro: **L**ab**e**l **N**oi**s**e-Tol**e**rant **Pro**totype-Based Network for Improving Cancer Detection in Prostate Ultrasound with Limited Annotations

Minh Nguyen Nhat To[1*], Fahimeh Fooladgar[1†], Paul Wilson[2†], Mohamed Harmanani[2†], Mahdi Gilany[2], Samira Sojoudi[1], Amoon Jamzad[2], Silvia Chang[1], Peter Black[1], Parvin Mousavi[2*], Purang Abolmaesumi[1*]

[1*]Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada.
[2]School of Computing, Queen's University, Kingston, Canada.

*Corresponding author(s). E-mail(s): tnnhatminh@gmail.com;
†These authors contributed equally to this work.

Abstract

Purpose: The standard-of-care for prostate cancer (PCa) diagnosis is the histopathological analysis of tissue samples obtained via transrectal ultrasound (TRUS) guided biopsy. Models built with deep neural networks (DNNs) hold the potential for direct PCa detection from TRUS, which allows targeted biopsy and subsequently enhances outcomes. Yet, there are ongoing challenges with training robust models, stemming from issues such as noisy labels, out-of-distribution (OOD) data, and limited labeled data.

Methods: This study presents LensePro, a unified method that not only excels in label efficiency but also demonstrates robustness against label noise and OOD data. LensePro comprises two key stages: first, self-supervised learning to extract high-quality feature representations from abundant unlabeled TRUS data, and second, label noise-tolerant prototype-based learning to classify the extracted features.

Results: Using data from 124 patients who underwent systematic prostate biopsy, LensePro achieves an AUROC, sensitivity, and specificity of 77.9%, 85.9%, and 57.5%, respectively, for detecting PCa in ultrasound. Our model shows it is effective for detecting OOD data in test time, critical for clinical deployment. Ablation studies demonstrate that each component of our method improves PCa

detection by addressing one of the three challenges, reinforcing the benefits of a unified approach.
Conclusion: Through comprehensive experiments, LensePro demonstrates its state-of-the-art performance for TRUS-based PCa detection. Although further research is necessary to confirm its clinical applicability, LensePro marks a notable advancement in enhancing automated computer-aided systems for detecting prostate cancer in ultrasound.

Keywords: Ultrasound Imaging, Image-Guided Interventions, Prostate Biopsy, Noisy Labels, Out of Distribution Data.

1 Introduction

Prostate cancer (PCa) is the second most common cancer diagnosed in men worldwide. The standard of care for PCa diagnosis is the histopathological analysis of tissue samples obtained by transrectal ultrasound-guided (TRUS) biopsy. However, the challenges associated with precisely detecting cancerous lesions in TRUS [1] lead to biopsies that are *systematic* than *targeted*. TRUS biopsy has relatively low sensitivity, a high chance of missing significant cancer, and a risk of biopsy-related adverse events [2, 3]. Targeted biopsy can alleviate these limitations, but currently requires pre-procedure multi-parametric MRI imaging followed by time-consuming registration with TRUS, posing cost and time constraints. As a result, there is a growing demand for targeted biopsy that is guided directly by TRUS imaging.

Deep neural networks (DNNs) offer considerable promise for ultrasound-based PCa detection. Despite the successes achieved in applying DNNs to several ultrasound modalities, such as contrast-enhanced ultrasound [4], micro-ultrasound [5], and temporal-enhanced ultrasound [6], major challenges remain in training robust models for PCa detection. Specifically, DNNs typically require extensive high-quality data and corresponding labels. Obtaining such data for ultrasound-based PCa detection is prohibitive, as it requires a biopsy procedure followed by expert labeling of the tissue. Two further issues inherent to ultrasound-based PCa detection that exacerbate the challenge are noisy labels and out-of-distribution (OOD) data. Noisy labels occur as histopathology analysis of biopsy samples, used as ground truth labels, only provide coarse descriptions of tissue properties. Assigning coarse labels to the corresponding ultrasound images introduces potentially mislabeled data [7]. OOD data arise as a result of imaging artifacts and inclusion of different tissue types in the imaging plane (e.g., non-prostatic and fibromuscular tissue, and benign prostatic hyperplasia) than those being classified (i.e. benign vs. cancer). Without addressing these challenges, DNNs may memorize noisy labels and OOD data, leading to poor predictions on unseen data and failure during clinical deployment.

While methods such as self-supervised and semi-supervised learning, learning with label noise, and OOD detection have been extensively applied in the computer vision literature, they are typically studied independently and evaluated on benchmark datasets with simulated noise and OOD. However, medical imaging data present a

unique challenge, as they simultaneously suffer from noisy labels, OOD data, and a scarcity of labeled data. Several strategies have been employed in medical imaging to tackle specific challenges. For addressing label noise, [8] used a "teacher" model trained on a small, clean training set to weight the loss function for the noisy labels, [9] used a loss modeling method to identify clean and noisy labels in histopathology images, and [10] used multi-instance learning to alleviate the effects of noisy labels in ultrasound data. For OOD detection, the predominant approach has been to use uncertainty [5, 11] although more recently [12] proposed to detect outliers through spectral analysis of DNN embeddings. To address labeled-data scarcity, self-supervised [13, 14] and semi-supervised learning [15] methods have been applied. Notably, these techniques have been applied in isolation and have not collectively addressed the challenges of noisy labels, OOD, and labeled-data scarcity in a unified manner.

We introduce **L**ab**e**l **n**oi**s**e-tol**e**rant **Pro**totype-based network (**LensePro**), a novel model for TRUS-based PCa detection that is data-efficient, and robust to label noise and OOD. It consists of two stages: (1) self-supervised learning to extract high-quality feature representations from unlabeled TRUS data; and (2), a prototype-based DNN [16] to classify the learned features in the presence of label noise and detect OOD samples when effectively calibrated. The resulting model achieves an AUROC of 77.9%, while rejecting OOD data during testing further improves the results. We conduct extensive experiments to demonstrate the effectiveness of our approach compared to baseline methods. To our knowledge, this is the first unified method to address labeled-data scarcity, OOD, and label noise for PCa detection in TRUS. Code is available at: *https://github.com/minhto2802/LensePro.*

2 Materials

2.1 Data Acquisition

The dataset used in this study includes 789 biopsy cores obtained from 124 patients who underwent systematic prostate biopsy at Vancouver General Hospital between 2018 and 2021. The study was approved by the institutional research ethics board and patients consented to be included. TRUS data at each biopsy location were obtained by using a *BK3500* ultrasound machine to acquire 200 consecutive radio frequency (RF) TRUS frames over 5 seconds while holding the transducer steady before firing the biopsy needle to collect tissue samples. To ensure data quality and maintain consistency with prior studies [5, 17], biopsy cores with excessive hand motion[1] or less than 40% cancer involvement were excluded from the dataset. The patient cohort was divided into three exclusive sets: the first batch for training (65 patients - 352 benign and 53 cancer cores), the second for validation (18 patients - 78 benign and 24 cancer cores), and the last for testing (41 patients - 250 benign and 32 cancer cores). Each core was labeled as benign or cancer based on the histopathology report, and this label was considered the ground truth for all of the patches extracted in that core.

Previous studies [7, 17] suggested a link between cancer involvement in a core and the rate of label noise. In this study, we hypothesize that using only a subset of cancer

[1] Excessive hand motion is detected by checking the B-mode videos recorded during the biopsy procedure.

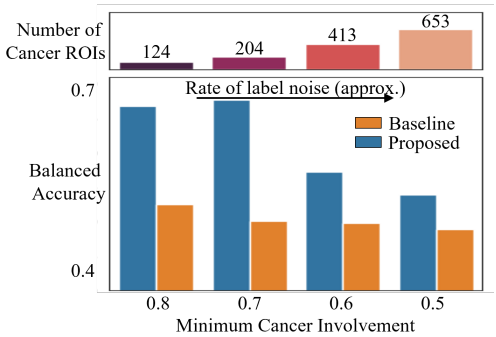

Fig. 1 Model's performance at varying rates of label noise approximated by the cancer involvement.

cores with high involvement can create a cleaner training set, potentially enhancing model performance. Testing this on the validation set, we varied the minimum cancer involvement for training without altering the test and validation sets. The subset with a minimum cancer involvement of 70% yielded the highest accuracy with our proposed method (Fig 1) and is consequently used for all subsequent evaluations.

2.2 Data Preprocessing

The input data for our model consist of regions of interest (ROI) extracted from the needle area of each core as follows: (i) all RF frames for the core were averaged to improve the signal-to-noise ratio; (ii) the needle tip was identified on the B-mode frames collected during the firing of the needle gun, and a 2×18 mm needle mask (based on the needle geometry) was created to match the tip position; (iii) the mask was converted to the RF coordinate system, and 32×32 pixel ROIs were selected, centered on the mapped needle outline, and moved in 32-pixel strides axially. On average, 20 ROIs were extracted per biopsy core; and (iv) each ROI was quantile-normalized based on its pixels' intensity.

3 Method

Our approach, as outlined in Figure 2, integrates IsoMax loss [18], Gaussian Mixture Model (GMM), and VICReg [19] to address the unique challenges present in the prostate ultrasound dataset. While IsoMax loss is effective for training prototypes in classification and OOD detection, the high label noise rate compromises the robustness of learned prototypes. To mitigate this, we leverage clean training samples by employing GMM to identify and filter data with corrupted labels in the training set. However, our results show that both IsoMax and GMM are insufficient to prevent the model from quickly overfitting to the limited and noisy labeled data when trained from scratch. Therefore, we complement our framework by adopting VICReg, a self-supervised learning approach known for its resilience across various batch sizes, to pretrain the backbone network before training the prototypes in a supervised manner. Through careful design, our method unifies the advantages of individual modules and significantly outperforms their performance when used in isolation.

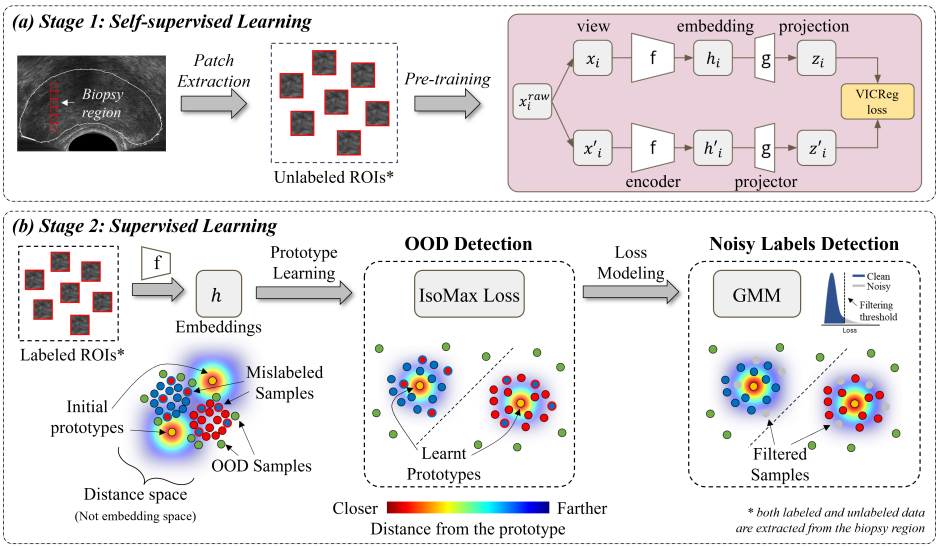

Fig. 2 Overview of the training pipeline: (a) self-supervised learning; (b) OOD/noisy label detection.

3.1 Self-supervised Learning

We employ a ResNet model as the backbone network to extract concise representations of unlabeled input data. This backbone network excels at capturing relevant features while filtering out noise and less relevant low-level details (see Figure 2.a). Achieving this is made possible through self-supervised learning, a more data-efficient approach that does not rely on labels, thus mitigating the risk of overfitting due to limited or noisy labels. Our self-supervised learning method of choice is the Variance-Invariance-Covariance Regularization (VICReg) [19], which is highly suitable for medical datasets as it performs well regardless of training batch size.

As a required step for self-supervised learning, we augment each ROI, x_i^{raw}, by randomly sampling two augmentations t and t' from a distribution[2] and applying them to the ROI. The resulting augmented versions $x_i = t(x_i^{\text{raw}})$ and $x'_i = t'(x_i^{\text{raw}})$ are passed through the backbone network $f_\theta(.)$ to generate two sets of representation vectors h_i and h'_i, respectively. These representations are then projected onto a new feature space using an additional projection network g_ϕ, resulting in z_i and z'_i.

During self-supervised pretraining, we generate batches Z and Z' of these projections and optimize $g_\phi(.)$ and $f_\theta(.)$ to minimize the VICReg loss. This loss is the weighted sum of three regularization terms, i.e. invariance $s(Z, Z')$, variance $v(Z)$, and covariance $c(Z)$ losses defined as follows:

$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')] \tag{1}$$

where λ, μ, and ν are hyperparameters controlling the importance of each term (fixed to 25, 25, and 1, following [19]). The invariance loss, $s(Z, Z') = \frac{1}{n}\sum_{i=1}^{n} \|z_i - z'_i\|^2$, is the

[2] We used a composition of random rotation (± 20 deg), contrast fluctuation, and crops.

mean-squared error loss between columns of Z and Z', encouraging the representations to be invariant to different augmentations of the input data; the variance loss, $v(Z) = \frac{1}{d}\sum_{j=1}^{d}\max(0, \gamma - \sigma(z^j))$, ensures the diversity across each batch by maintaining a minimum standard deviation for each feature (j^{th} row in Z), therefore maximizing the information content. Lastly, the covariance loss, $c(Z) = \frac{1}{d}\sum_{i\neq j} Cov(Z)_{ij}^2$, minimizes the redundancy of features in the representation space, where $Cov(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T, \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, is the features covariance matrix.

3.2 Prototype Learning

3.2.1 Prototype-based Network:

Following self-supervised learning, the high-level features associated with ultrasound patches are used to train learnable prototypes associated with benign and cancer classes (Figure 2.b). Since SoftMax loss does not enforce the high density of embeddings in the feature space, we replace it with IsoMax loss [18], a distance-based loss, to train a prototype-based classifier that assigns the embeddings to either benign or cancer using on their proximity to the nearest prototype.

For each of the K classes, $p_\psi^{(i)} : i = 1, ..., K \subset \mathbb{R}^d$ is a random 2048-dimensional vector (same number of dimensions with embeddings extracted by ResNet50) drawn from a normal distribution with μ mean and σ standard deviation. Using a feature extractor network $f_\theta(.) : \mathbb{R}^n \to \mathbb{R}^d$, K prototypes are learned by minimizing the loss function across data-label pairs (X, y), defined as:

$$\mathcal{L}_{\text{IsoMax}}(X, y) = -\log\left(\frac{\exp\big(-E_s|d_s|D(f_\theta(X), p_\psi^{(y)})\big)}{\sum_{i=1}^{K}\exp\big(-E_s|d_s|D(f_\theta(X), p_\psi^{(i)})\big)}\right), \tag{2}$$

where $D(x, y) = \|\frac{x}{\|x\|} - \frac{y}{\|y\|}\|$ is the Euclidean distance between normalized vectors, d_s is a learnable *distance scale* that is fixed for all classes, and E_s is a fixed scalar hyperparameter called the *entropic scale*. The entropic scale is introduced to avoid overconfident low-entropy predictions often seen on models trained with SoftMax loss.

3.2.2 Prototype-based Out-of-distribution Detection:

During inference, prototypes trained with IsoMax loss can also be used to detect OOD samples. The minimum distance score (MDS), which presents the distance to the nearest prototype, is proposed as an OOD score [18]. MDS between the features of the input sample, $f_\theta(X)$, and all class prototypes p_ψ^j, is given by

$$\text{MDS} = \min_j D\big(f_\theta(X), p_\psi^j\big). \tag{3}$$

The smaller the MDS, the more likely the sample is in-distribution; conversely, the larger the MDS, the more likely it is OOD. However, while achieving promising performance on computer vision datasets, our empirical results show that MDS, when used

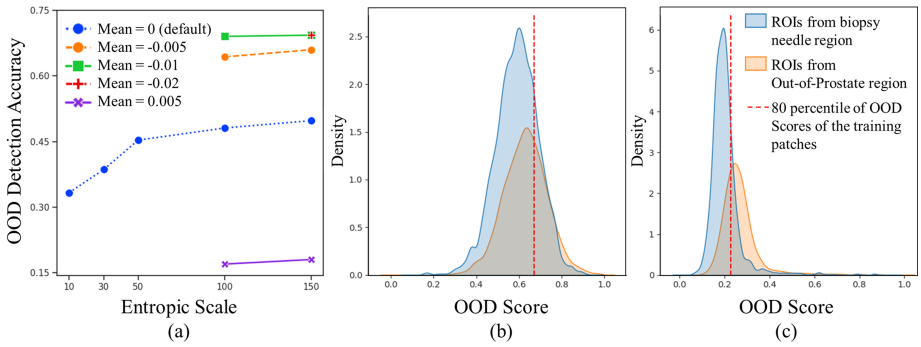

Fig. 3 Results of the OOD detection calibration. (a) OOD detection accuracy at different entropic scales (E_s) and means of prototype initialization (μ). OOD score distribution of the training and OOD sets (b) before calibration ($E_s = 10$, $\mu = 0$), and (c) after calibration ($E_s = 100$, $\mu = -0.01$).

with default entropic scale ($E_s < 10$) and prototype initialization (vectors with $\mu = 0$ and $\sigma = 1$), has limited success in detecting OOD US data, e.g., non-prostate tissues.

To this end, we create an extra dataset, called the OOD set, consisting of ROIs from *outside the prostate*. At different combinations of entropic scale and prototype initialization, we train the prototypes for a small number of epochs and evaluate how these factors affect the separation between the OOD and the training sets. The combination of entropic scale and prototype initialization that best separates the OOD set and the training set is subsequently used for prototype learning and OOD removal during inference. The highest OOD detection accuracy is obtained when initializing the prototypes at $Mean = -0.01$ and using $E_s = 100$ (Figure 3). These hyperparameters are subsequently used for the following experiments of our method.

3.3 Label Noise-tolerant Learning

The label noise in our dataset stems from aligning histopathology findings with corresponding patches at the biopsy location. For example, when designating a core with 60% cancer involvement as cancer and subsequently labeling individual ROIs extracted from that core as cancer, mislabeled ROIs are inevitably introduced to the training set, as some ROIs likely originate from the non-cancerous area of that core. DNNs tend to learn clean labels faster than noisy ones, resulting in lower loss values associated with clean samples [20]. Herein, we adopt the Gaussian Mixture Model (GMM) to automatically detect and exclude noisy ROIs based on their loss values before updating the prototypes. In particular, a two-component GMM is used to learn the distribution of loss values across all training samples, then classify a sample as "noisy" if its probability of belonging to the component with the larger mean is at least 0.5. We subsequently ignore the noisy labels and train only on the clean ones, resulting in learning better class prototypes. Additionally, we employ a peer-networks approach: because a single network choosing clean examples based on its own loss values is prone to confirmation bias, we train two different "peer" networks simultaneously, with clean examples for one network chosen based on the loss values of the other network. This approach

follows [21, 22]. In our study, the different networks are initialized based on the same feature extractor $f_\theta(.)$ but with different randomly initialized sets of prototypes.

3.4 Experiments

To determine the efficacy of our method, we designed three main experiments: 1) we compared our performance to several PCa detection methods, which separately address label noise [17, 10] and labeled-data scarcity [14]; 2) we studied the impact of OOD detection by testing the improvement in our performance when excluding data flagged as OOD, at test time; and 3) we studied the contribution of individual components of our method. Self-supervised VICReg training is performed for 200 epochs with ResNet architectures. ResNet50, which was modified in the first layer to accept 1-channel input, is used as the backbone network for feature extraction from 32×32 patches. We then train the two prototypes for ROI classification using an SGD optimizer with a one-cycle learning rate scheduler ($maximum = 0.001$) and a batch size of 128. After 5 warmup epochs, the loss modeling and noise exclusion are performed every epoch for the remainder of the training. Most hyperparameters are tuned based on validation accuracy, except the entropic scale and prototype initialization, which were manually calibrated using an additional OOD set. We report balanced accuracy (ACC_B), AUROC, sensitivity (SEN), and specificity (SPE), averaged across the three best epochs and three last epochs. Each experiment is repeated with 10 random seeds, and the average and standard deviation of performance are reported. All metrics are computed for core-wise performance, where a core is predicted as cancer if 40% or more of its ROIs are classified as cancer (corresponding to the lowest cancer involvement in the dataset). One-tailed paired Wilcoxon signed rank is used for statistical tests.

4 Results and Discussion

4.1 Quantitative Results

Table 1 shows a comparison of the performance of our method (bottom row) to other methods for PCa detection. In a broader comparison in learning with label noise (LNL), we included NCE-AGCE [23], GCE [24], and DivideMix [21]. Both NCE-AGCE and GCE are tailored for weakly supervised learning through noise-robust loss functions, while DivideMix performs semi-supervision by leveraging samples with detected noisy labels as unlabeled data. In general, our method shows the best performance, exceeding the second-best method in balanced accuracy (+3%), AUROC (+3%), and sensitivity (+5%). On the contrary, methods for LNL struggled in handling labeled data scarcity leading to rapid overfitting on our training set, evidenced by the large performance difference between the best and last epochs. In addition, our method is competitive with the sensitivity and specificity of mp-MRI, as reported by the PROMIS trial (88% and 45% for mp-MRI, respectively, compared to 86% and 57%, respectively, for our approach) [2], indicating the potential for clinically useful biopsy targeting capabilities similar to mp-MRI.

A benefit of our model is its ability to detect OOD data at test time. We hypothesize that abstaining from making predictions on OOD data would improve test

Table 1 Prostate cancer detection performance against competing methods.

Methods		ACC_B	AUROC	SEN	SPE
Co-Teaching [17]	Best	57.3±0.8	59.8±2.0	75.6±8.6	39±8.8
	Last	54.5±1.5	60.8±0.7	72.4±5.2	36.6±3.4
NCE-AGCE [23]	Best	64.7±1.0	67.5±1.0	68.3±7.3	61.1±8.0
	Last	59.0±1.2	67.0±0.7	43.4±3.5	74.6±3.0
GCE [24]	Best	66.3±1.0	69.9±1.1	70.9±7.5	53.7±6.6
	Last	60.2±1.4	67.6±0.6	60.1±4.3	60.3±3.1
VICReg [14]	Best	68.6±1.1	75.0±1.3	80.9±9.5	56.3±9.3
	Last	64.3±1.1	73.1±0.4	59.4±4.2	69.1±3.1
DivideMix [21]	Best	62.7±1.4	67.0±2.5	71.4±10.7	54.1±10.7
	Last	55.9±3.2	64.3±2.1	32.6±16.1	**79.2**±11.8
Multi-Instance Learning [10]		66	68	77	55
LensePro (ours)	Best	**71.7**±0.5	**77.9**±1.0	**85.9**±3.3	57.5±3.2
	Last	68.5±0.7	77.3±0.4	77.6±1.4	59.4±1.3

Table 2 Prostate cancer detection performance with OOD removal at test time. The removal thresholds are calculated using the percentiles of MDS of the OOD set constructed in Section 3.2.2.

Removal percentile	ACC_B	AUROC	SEN	SPE
No removal	71.7	77.9	**85.9**	57.5
20	73.7	79.0	82.6	64.8
30	**73.9**	**79.1**	81.7	**66.2**
50	73.8	77.7	82.1	65.6

accuracy. This hypothesis is tested in Table 2, where we recalculate the metrics while ignoring predictions made on the top N percent of ROIs with the highest MDS OOD score. The significant increase in ACC_B (p-values< 0.001) of OOD removal at all levels compared to without removal can be attributed to the substantial improvement in the model's specificity (SPE). The higher SPE following the removal of out-of-distribution (OOD) samples suggests that OOD patches might be prone to erroneous identification as cancer. Simultaneously, instances accurately predicted as cancer are also classified as OOD and eliminated, leading to a reduction in sensitivity (SEN). We also note that ODD detection is not as effective when too many samples (small MDS threshold is used) are removed, likely as in-distribution samples are being removed by mistake.

4.2 Ablation Study

The results of the ablation study on the different components of our method are shown in Table 3. The balanced accuracy and AUROC of the full model are significantly better (p-values< 0.005) than other configurations. We find a slight reduction of performance when removing label noise-tolerant learning (third row), and a further, more significant reduction of performance when replacing IsoMax loss with cross-entropy loss for supervised learning after pretraining the model with VICReg (second row). The removal of self-supervised learning results in a severe decline in the performance

Table 3 Ablation results on the effectiveness of different modules in our approach using the models at the last training epoch.

Prototype learning	Self-supervised learning	Label noise learning	ACC_B	AUROC
✓			50.0 ±0.0	58.8 ±0.9
	✓		64.3 ±1.1	73.1 ±0.4
✓	✓		67.8 ±1.1	76.8 ±0.5
✓	✓	✓	**68.5** ±0.7	**77.3** ±0.4

of the model, even when keeping the prototype-based approach (first row). We hypothesize that this is due to the rapid overfitting when training the prototypes and feature extractor together from scratch. This observation partially explains the poor performance of co-teaching in Table 1, which also trains the feature extractor from scratch. This ablation study underlines the necessity of our proposed unified approach since the individual components of our method do not work as well in isolation.

4.3 Qualitative Results

Qualitative results for our method are shown in Figure 4, where we generate heatmaps showing the predictions made by our model. Patient 1 had all biopsy cores positive for cancer, and correspondingly high cancer scores throughout the prostate, while Patients 2 and 3 had no positive biopsy, and the heatmaps show much lower probabilities of cancer. In the right-most column, we demonstrate the positive impact of test-time OOD removal, which reduces false positives caused by artifacts.

5 Conclusion

We propose LensePro, a unified framework for TRUS focusing on efficient label utilization, robustness against label noise, and out-of-distribution data. Our model excels in PCa detection in TRUS, surpassing other leading methods. A limitation of our study is the exclusion of biopsy cores with less than 40% cancer involvement during both training and evaluation. Dealing with the high label noise in these cases necessitates stronger methods for learning with noisy labels and a more sophisticated approach for aggregating patch-wise predictions into core-wise predictions, aspects we plan to address in future work. Additionally, our approach lacks a mechanism to differentiate hard-to-learn samples from noisy ones, potentially resulting in under-learning from rare cases and limiting clinical applicability. To retain hard samples, future work will explore implementing a dedicated module to filter samples in which the peer models highly diverge from the set of noisy labels [25]. We will also entail larger multi-center datasets to evaluate the generalizability of our approach. While more research is needed to validate its clinical potential, LensePro represents a significant step forward in improving automated computer-aided diagnosis systems for PCa in ultrasound.

Acknowledgments. This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (CIHR). Parvin Mousavi is supported by Canada CIFAR AI Chair and the

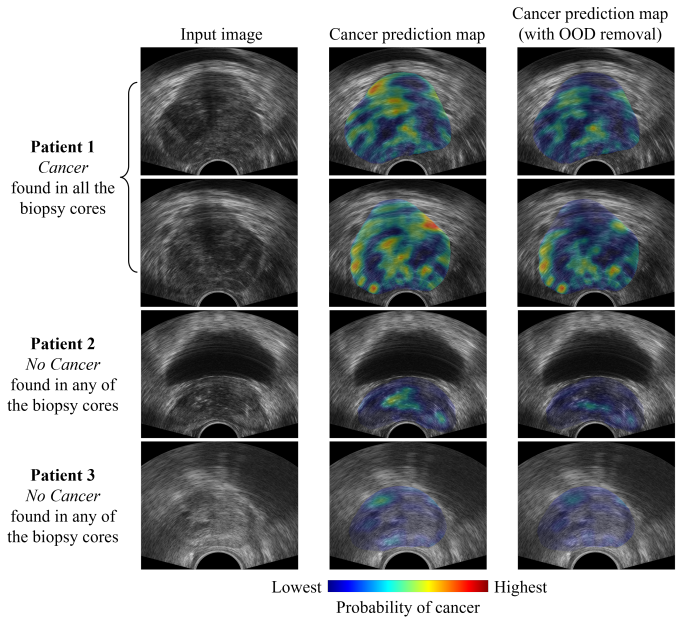

Fig. 4 Qualitative Results. Cancer maps were generated for the whole biopsy imaging plane of a patient with large cancer lesions, and two patients with all benign biopsies.

Vector Institute. We acknowledge the staff at Vancouver General Hospital who assisted with data acquisition for our study.

6 Ethics declarations

- Conflict of interest: All authors confirm that there are no known conflicts of interest with this publication.
- Ethical approval: All studies involving human participants were in accordance with the ethical standards of the institutional research board and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.
- Informed consent: Informed consent was obtained from all individual participants included in the study.

References

[1] Martijn Smeenge, Jean JMCH de la Rosette, and Hessel Wijkstra. Current status of transrectal ultrasound techniques in prostate cancer. *Current Opinion in Urology*, 22(4):297–302, 2012.

[2] Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, et al. Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822, 2017.

[3] Adam Madej, Jacek Wilkosz, Waldemar Różański, and Marek Lipiński. Complication rates after prostate biopsy according to the number of sampled cores. *Central European Journal of Urology*, 65(3):116, 2012.

[4] Yujie Feng, Fan Yang, Xichuan Zhou, Yanli Guo, Fang Tang, Fengbo Ren, Jishun Guo, and Shuiwang Ji. A deep learning approach for targeted contrast-enhanced ultrasound based prostate cancer detection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(6):1794–1801, 2018.

[5] Mahdi Gilany, Paul Wilson, Amoon Jamzad, Fahimeh Fooladgar, Minh Nguyen Nhat To, Brian Wodlinger, Purang Abolmaesumi, and Parvin Mousavi. Towards confident detection of PCa using high resolution micro-ultrasound. In *Medical Image Computing and Computer Assisted Interventions*, pages 411–420, 2022.

[6] Fahimeh Fooladgar, Minh Nguyen Nhat To, Golara Javadi, Samareh Samadi, Sharareh Bayat, Samira Sojoudi, Walid Eshumani, Antonio Hurtado, Silvia Chang, Peter Black, et al. Uncertainty-aware deep ensemble model for targeted ultrasound-guided prostate biopsy. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[7] Golara Javadi, Samareh Samadi, Sharareh Bayat, Samira Sojoudi, Antonio Hurtado, Silvia Chang, Peter Black, Parvin Mousavi, and Purang Abolmaesumi. Training deep networks for prostate cancer diagnosis using coarse histopathological labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 680–689, 2021.

[8] Han Le, Dimitris Samaras, Tahsin Kurc, Rajarsi Gupta, Kenneth Shroyer, and Joel Saltz. Pancreatic cancer detection in whole slide images using noisy label annotations. In *Medical Image Computing and Computer Assisted Intervention*, pages 541–549. Springer, 2019.

[9] Murtaza Ashraf, Willmer Rafell Quiñones Robles, Mujin Kim, Young Sin Ko, and Mun Yong Yi. A loss-based patch label denoising method for improving whole-slide image analysis using a convolutional neural network. *Scientific Reports*, 12(1):1392, 2022.

[10] Golara Javadi, Samareh Samadi, Sharareh Bayat, Mehran Pesteie, Mohammad H Jafari, Samira Sojoudi, Claudia Kesch, Antonio Hurtado, Silvia Chang, Parvin Mousavi, et al. Multiple instance learning combined with label invariant synthetic data for guiding systematic prostate biopsy: a feasibility study. *International Journal on Computer Assisted Radiology and Surgery*, 15(6):1023–1031, 2020.

[11] Jasper Linmans, Stefan Elfwing, Jeroen van der Laak, and Geert Litjens. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83:102655, 2023.

[12] Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE Transactions on Artificial Intelligence*, 2022.

[13] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,

pages 3478–3488, 2021.

[14] Paul FR Wilson, Mahdi Gilany, Amoon Jamzad, Fahimeh Fooladgar, Minh Nguyen Nhat To, Brian Wodlinger, Purang Abolmaesumi, and Parvin Mousavi. Self-supervised learning with limited labeled data for prostate cancer detection in high frequency ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2023.

[15] Yanyun Jiang, Xiaodan Sui, Yanhui Ding, Wei Xiao, Yuanjie Zheng, and Yongxin Zhang. A semi-supervised learning approach with consistency regularization for tumor histopathological images analysis. *Frontiers in Oncology*, 12:7200, 2022.

[16] David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano LI Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection. In *2021 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2021.

[17] Minh Nguyen Nhat To, Fahimeh Fooladgar, Golara Javadi, Sharareh Bayat, Samira Sojoudi, Antonio Hurtado, Silvia Chang, Peter Black, Parvin Mousavi, and Purang Abolmaesumi. Coarse label refinement for improving prostate cancer detection in ultrasound imaging. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–7, 2022.

[18] David Macêdo and Teresa Ludermir. Enhanced isotropy maximization loss: Seamless and high-performance out-of-distribution detection simply replacing the softmax loss. *arXiv preprint arXiv:2105.14399*, 2021.

[19] Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022.

[20] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9686, 2022.

[21] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019.

[22] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018.

[23] Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *International conference on machine learning*, pages 12846–12856. PMLR, 2021.

[24] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[25] Xiaobo Xia, Bo Han, Yibing Zhan, Jun Yu, Mingming Gong, Chen Gong, and Tongliang Liu. Combating noisy labels with sample selection by mining high-discrepancy examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1843, 2023.