Towards Confident Prostate Cancer Detection using Ultrasound: A Multi-Center Study

Paul F. R. Wilson[1*], Mohamed Harmanani[1], Minh Nguyen Nhat To[2], Mahdi Gilany[1], Amoon Jamzad[1], Fahimeh Fooladgar[2], Brian Wodlinger[3], Purang Abolmaesumi[2], Parvin Mousavi[1]

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

*Corresponding author(s). E-mail(s): 1pfrw@queensu.ca;

Abstract

Purpose: Deep learning-based analysis of micro-ultrasound images to detect cancerous lesions is a promising tool for improving prostate cancer (PCa) diagnosis. An ideal model should confidently identify cancer while responding with appropriate uncertainty when presented with out-of-distribution inputs that arise during deployment due to imaging artifacts and the biological heterogeneity of patients and prostatic tissue. **Methods:** Using micro-ultrasound data from 693 patients across 5 clinical centers who underwent micro-ultrasound guided prostate biopsy, we train and evaluate convolutional neural network models for PCa detection. To improve robustness to out-of-distribution inputs, we employ and comprehensively benchmark several state-of-the-art uncertainty estimation methods. **Results:** PCa detection models achieve performance scores up to **76%** average AUROC with a 10-fold cross validation setup. Models with uncertainty estimation obtain expected calibration error scores as low as **2%**, indicating that confident predictions are very likely to be correct. Visualizations of the model output demonstrate that the model correctly identifies healthy vs. benign tissue. **Conclusion:** Deep learning models have been developed to confidently detect PCa lesions from micro-ultrasound. The performance of these models, determined from a large and diverse dataset, is competitive with visual analysis of magnetic resonance imaging, the clinical benchmark to identify PCa lesions for targeted biopsy. Deep learning with micro-ultrasound should be further studied as an avenue for targeted prostate biopsy.

Keywords: uncertainty calibration, deep learning, prostate cancer, ultrasound

1 Introduction

Prostate cancer (PCa) is the second most common cancer in men worldwide. Early and accurate diagnosis and staging are crucial for guiding treatment decisions. Currently, in clinical practice, the detection of PCa involves a systematic biopsy (SB) of the prostate under trans-rectal ultrasound (TRUS) guidance. Targeted biopsy, using pre-procedural multi-parametric Magnetic Resonance Imaging (mp-MRI) with the Prostate Imaging Reporting and Data System (PI-RADS) [1] to identify suspicious lesions, has been shown to significantly improve the sensitivity of the biopsy procedure compared to systematic biopsy [2]. However, the need for MRI is prohibitive in many clinical settings and there is a pressing need to develop biopsy targeting techniques that rely solely on ultrasound.

While conventional ultrasound has low sensitivity and specificity for cancer detection [3], micro-ultrasound is a recently developed technology that utilizes high transducer frequencies to achieve significantly higher spatial resolution than conventional ultrasound, improving the visualization of tissue microstructure and enabling the identification of cancerous lesions. The Prostate Risk Assessment using Micro-Ultrasound (PRI-MUS) [4] protocol is a qualitative method relying on visual analysis of B-mode micro-ultrasound data for assessing PCa risk having comparable performance to mp-MRI [5]. However, this protocol is subjective, susceptible to inter-observer variability and may overlook quantitative information embedded in raw radio-frequency (RF) micro-ultrasound data, related to tissue micro-structure [6]. Thus, there is a need for a complementary objective, user-independent, and quantitative PCa detection tools with micro-ultrasound data.

Deep learning has the potential to contribute to such a tool. In PCa diagnosis, deep learning has achieved human expert performance levels in PCa histopathology [7] and has been used in MRI-based PCa detection [8]. It has been applied for ultrasound-based PCa detection as well, for example from temporal-enhanced ultrasound [9, 10], contrast-enhanced ultrasound [11], and micro-ultrasound [12–14]. However, PCa detection from ultrasound is challenging: prostate ultrasound data is highly heterogeneous due to the large variability in types of cancerous and benign tissue, variations in patient populations and clinical practices, and the frequent presence of ultrasound artifacts, making it difficult to train models that generalize well to new patients and clinical settings. When seeing unfamiliar data, deep learning models have the tendency to make overconfident but incorrect predictions [15]. This seriously undermines the user's confidence in the model as a reliable tool for clinical decision making.

To address this problem, several solutions have been explored. Deep ensembles [16] have been used for PCa detection from temporal-enhanced ultrasound [9], and for automatic segmentation of cancerous regions in histopathology slides [17]. Test-time augmentations have been applied for uncertainty estimation for prostate segmentation [18]. In a previous work [13] we explored the evidential method [19] for

ultrasound-based PCa detection. Although their findings were promising, these methods were often studied in the context of small or single-center datasets, which likely fail to capture the variability of data in real-world clinical settings across different populations. Additionally, common validation techniques such as random splitting of data risk making overoptimistic performance estimates which may not generalize. Studies validating models on many patients from multiple clinical centers are needed to adequately estimate real-world performance of these methods.

Recognizing this need, we perform a study to benchmark the performance of deep-learning PCa detection models using a multi-center clinical dataset, aiming to obtain the best possible understanding of how models are likely to perform in a clinical setting. Using a dataset including 693 patients from five clinical centers, we benchmark numerous models for PCa detection along with state of the art uncertainty calibration methods. We study the performance of the models for cancer detection and uncertainty calibration; in addition, we qualitatively evaluate the models via the generation of cancer "heatmaps". To maximize the statistical validity of our findings, we employ a nested 10-fold cross validation scheme. To our knowledge this is the first multi-center study on uncertainty calibration for high-resolution ultrasound based PCa detection. Our models were able to detect PCa with AUROC up to 76%. Utilizing uncertainty calibration techniques, our models can achieve a very good level of uncertainty calibration, meaning that confident predictions are very likely to be correct. Performance is competitive with both PI-RADS and PRI-MUS systems. These contributions collectively pave the way for accurate and reliable diagnosis and improved patient outcomes.

2 Materials

2.1 Data

We use data collected from 693 patients who underwent prostate biopsy in five centers as part of a clinical trial. The trial was approved by institutional ethics boards, and informed consent was given for the use of data. Systematic sextant biopsy was carried out under trans-rectal ultrasound (TRUS) guidance using the ExactVu micro-ultrasound system [20]. The system uses an ultra-high resolution (22.5 MHz center frequency; upper band range of 29 MHz), 512-element side-firing linear array, capturing images in the sagittal plane. The transducer includes a fixed biopsy needle guide with an angle of 30 degrees relative to the linear array. Immediately prior to firing the biopsy gun, a raw RF ultrasound image was saved. Histopathology analysis was performed to determine the Gleason score (a summary of the presence and aggressiveness of cancer in the entire tissue sample) for each core. The scores were converted to binary indicators of 0 (benign) and 1 (cancer, Gleason score ≥ 7) and assigned as the core label. In total, the dataset consists of 6607 total cores, of which 5727 (87%) are benign and 880 (13%) contain cancer. The precise breakdown of grades per center is in Table 1.

Table 1 Summary of the data, with the number of patients and cores originating from each center, and the proportion of benign and cancerous cores of different grades.

Center	Location	Patients	Cores	Benign	GS7	GS8	GS9	GS10
JH	Baltimore, USA	60	616	568	32	10	6	0
UVA	Charlottesville, USA	236	2335	2018	221	57	28	11
PCC	Calgary, Canada	171	1599	1400	162	23	14	0
PMCC	Toronto, Canada	71	588	486	90	12	0	0
CRCEO	Quebec City, Canada	155	1469	1255	170	32	12	0
Total	-	693	6607	5727	675	134	60	11

2.2 Data Preprocessing

Following the paradigm established in previous studies [13], we perform cancer detection by extracting patches from the image in a sliding window fashion and predicting a cancer likelihood score for each patch. The depth and width of the image are 28 mm and 46.06 mm, respectively. We use a patch size of 5×5 mm, which corresponds to 447×55 pixels (axially and laterally) of raw RF data, and we use a stride of 1×1 mm. Patches are then resampled to a shape of 256×256 pixels to match a typical convolution network (CNN) input[1]. For each core, patches overlapping the needle region are selected and labeled with the core label. Cancer cores may in general include a combination of benign and cancerous tissues. Therefore, following previous work [10, 13, 14, 21], only cancer cores with high involvement (defined as at least 40% cancer by area) are included, ensuring patches extracted from these cores will intersect cancer, and making this labeling strategy valid. At inference time, the model may be deployed as a "sliding window" to generate a heatmap of cancer likelihood for any larger region of interest in the image.

2.3 Cohort Selection

In order to provide the best estimates of how the model may generalize, we use a nested 10-fold cross-validation scheme. The patients are divided into 10 folds with an equal number of patients from each center in each partition. To ensure generalization across centers, an additional leave-one-center-out cross-validation scheme was also tested for a subset of the experiments. Naive K-fold cross-validation yields over-optimistic estimates if the test performance is used to tune hyperparameters for training. Hence, we use a nested cross-validation scheme where for each fold, the training set is further subdivided into training (80% of patients) and validation (20% of patients) sets. This validation set is used to select the best-performing model from each training run, and for temperature calibration (further explained in Section 3.4). Following patient selection, cores are selected as follows: similar to previous work [10, 13, 14, 21], only cancer cores with involvement of more than 40% are included. Benign cores in the training (but not testing and validation) set are randomly undersampled to match the number of cancer cores, to address the class imbalance in the data.

[1] Because the RF signal is band-limited, no loss of information occurs in resampling.

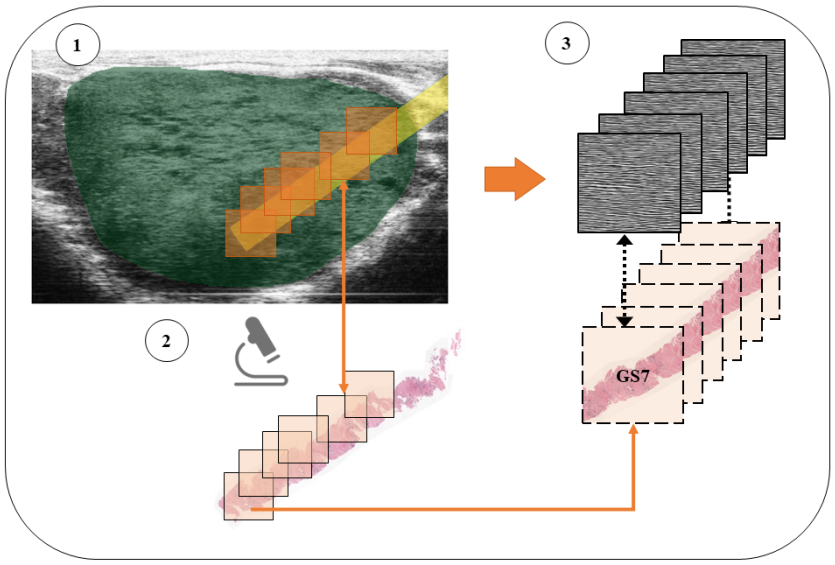

Fig. 1 Preprocessing a micro-ultrasound image into patches. (1) Delineation of the needle region and selection of patches within that region. (2) Labelling of cores via histopathological analysis of tissue retrieved from the core during biopsy. (3) Patch extraction and labeling according to the core label.

3 Methods

Here, we describe our basic learning paradigm (baseline), followed by state-of-the-art uncertainty quantification methods we evaluate, and how we improve the calibration of uncertainty estimates.

3.1 Baseline

Our deep learning model consists of a feature extractor $f_\theta(\cdot)$ which maps an input patch X to a 512-dimensional feature representation, followed by a classification head $g_\phi(\cdot)$ which maps the feature representation to a scalar value representing the log-probability that the patch contains cancer. For the feature extractor $f_\theta(\cdot)$, we use a modified ResNet18 [22], where each residual block has only one sequence of convolution, activation, and batch normalization (compared to the original paper, where 2 such sequences are used per block). This modification empirically improves performance, possibly by mitigating overfitting [13]. This feature extractor architecture is used for all methods studied. The classification head $g_\phi(\cdot)$ is a single linear layer. Both components of the model are trained using supervised learning with cross-entropy loss $\sum_{(X,\hat{y})\in\text{data}} -\hat{y}\log(y) - (1-\hat{y})\log(1-y))$, where $\hat{y}$ is the binary label assigned to the patch based on pathology findings.

3.2 Uncertainty Estimation

Deep ensembles [16] involve several models independently trained with different parameter-initializations, where their predictions are aggregated. In other words, with

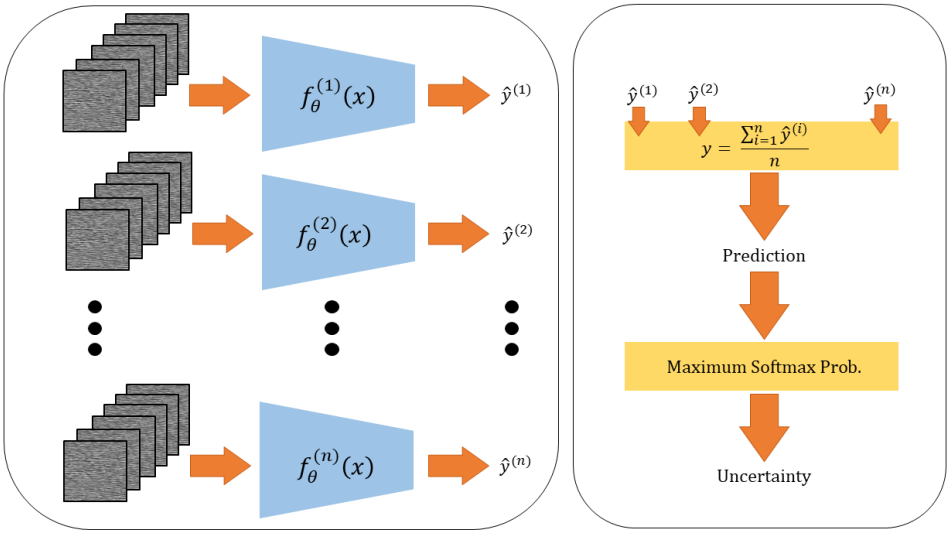

Fig. 2 Visualization of the deep ensemble architecture. **Left**: n models are trained to classify patches as benign or cancer. The output vector $y^{(i)}$ of model i is a 2×1 vector of logits. **Right**: The predictions made by each model are averaged to produce a final prediction y for the batch. The maximum softmax probability of y is then used as an uncertainty score.

the set of parameterizations $\{(\theta_i, \phi_i) | i = 1, ..., N\}$ arising from N independent training runs, we have the ensemble prediction:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} g_{\phi_i}(f_{\theta_i}(X)). \tag{1}$$

We train each model in the ensemble using the baseline method described in Subsection 3.1, and $N = 10$ models are used. Due to multiple inferences from separate models being required, deep ensembles greatly increase computational costs. Hence, we also study two leading methods in single-model uncertainty as a potential alternative.

Spectral normalized Gaussian processes (SNGP) [23] make two modifications to the baseline method. First, spectral normalization [23] is applied to the weight matrices of the model, which enforces Lipschitz continuity of the feature extractor f_θ. Informally, this means approximate distance preservation between the data manifold and the representation space. Second, the linear classification head is replaced by a Gaussian process classifier with a finite-dimensional approximation of the radial basis function kernel, allowing end-to-end training of the Gaussian process. Gaussian processes are guaranteed to have higher uncertainty for inputs that are very different from the training data, yielding better uncertainty estimation than a typical linear layer [23].

The IsoMaxPlus [24] method learns prototypes for each of the K classes, $\{p_i | i = 1, ..., K\}$, in the feature space. The probability of class i given input X is calculated

as follows:

$$P(y_i|X) = \frac{\exp\big(-K \cdot D(f_\theta(X), p_i)\big)}{\sum_{i=1}^{K} \exp\big(-K \cdot D\big(f_\theta(X), p_j\big)\big)}, \tag{2}$$

where $D(x, y) = \|\frac{x}{\|x\|} - \frac{y}{\|y\|}\|$ is the Euclidean distance between normalized vectors, and K is a learnable scaling factor. This improves uncertainty estimates for OOD data since the features for these data are likely to be far from all prototypes.

3.3 Temperature Calibration

Temperature Calibration (TC) [15] is a simple method for dealing with poor uncertainty calibration of neural networks. Given trained feature extractor f_θ and classifier g_ϕ, two additional scalar parameters τ and β, respectively called the temperature and bias, are added to slightly adjust the model's outputs via:

$$\hat{y} = \sigma\big(g_\phi(f_\theta(X))/\tau + \beta\big). \tag{3}$$

These parameters optimized to minimize the total cross-entropy loss on the *validation set*, while the model is in *evaluation mode*, i.e. with batch normalization and dropout disabled. For ensembles, we may apply temperature calibration to each member in the ensemble (early TC), or calibrate the aggregated predictions of the ensemble (late TC).

3.4 Evaluation Methods

The proposed methods are aimed for use as detection models which would be applied as a "sliding window" across an ultrasound image. The scores from multiple overlapping small patches are taken in the aggregate to identify overall suspicious regions and biopsy them. To simulate this aggregation, we apply our models to each patch within the needle trace region for a core and average their predictions into an overall core-level prediction. We measure the AUROC, sensitivity, and specificity when comparing this prediction to the biopsy ground truth.

To measure uncertainty calibration, we used the Expected Calibration Error (ECE) [15], and Brier score [23]. We use the probability assigned to the predicted class (also called the maximum softmax probability [23]), as a measure of model confidence. These metrics are computed at the patch-level. In addition, we perform performance-rejection analysis: for various rejection rates r between 0 and 100%, "reject" the r lowest-confidence predictions and compute the AUROC score for the remaining predictions. The AUROC is expected to increase at the cost of making fewer predictions, and the best model would have the best trade-off between the highest AUROC relative to the number of rejections needed to achieve that AUROC.

Finally, we qualitatively analyze the models by displaying their probability scores as heatmaps overlaid on the B-mode rendering of the corresponding ultrasound image. We also simulate the "reject" option by displaying versions of the heatmaps where low confidence predictions are made transparent.

4 Results

4.1 Quantitative Evaluation

Table 2 Evaluation of different learning frameworks based on classification performance and uncertainty calibration. Each model is also tested with and without temperature calibration (TC). Classification metrics are computed from core-wise predictions, and uncertainty metrics from patch-wise predictions. The best results for each column are shown in bold.

Method	**AUROC** ↑	**Sens.** ↑	**Spec.**↑	**ECE**↓	**Brier**↓
Baseline	74.1 ± 4.6	65.9 ± 7.3	69.4 ± 3.8	11.4 ± 9.3	48.6 ± 3.7
Baseline + TC	75.0 ± 4.5	65.5 ± 7.1	$\mathbf{70.8} \pm 3.5$	2.7 ± 2.8	44.9 ± 5.6
IsoMaxPlus [24] + TC	72.7 ± 7.6	$\mathbf{70.5} \pm 9.1$	64.3 ± 6.1	10.1 ± 5.2	43.4 ± 1.7
SNGP [23]	75.5 ± 5.6	67.9 ± 7.9	70.2 ± 3.4	6.0 ± 4.0	45.0 ± 2.3
SNGP [23] + TC	74.7 ± 6.0	67.6 ± 7.2	67.0 ± 4.7	$\mathbf{2.1} \pm 1.3$	43.6 ± 1.2
Ensemble [16]	75.8 ± 5.6	63.3 ± 19.3	70.2 ± 21.5	8.6 ± 6.8	43.7 ± 2.1
Ensemble [16] + early TC	$\mathbf{76.0} \pm 5.7$	67.6 ± 7.4	70.2 ± 3.8	3.2 ± 1.9	$\mathbf{42.7} \pm 1.2$
Ensemble [16] + late TC	75.8 ± 5.6	68.0 ± 8.3	70.3 ± 3.9	2.3 ± 1.2	42.8 ± 1.5

The quantitative performance benchmarks for 10-fold cross validation are shown in Table 2. As seen, most methods have an overall similar classification performance of 75% AUROC, with a small absolute difference (3% between the best and worst performance). The ensemble method exhibits a small improvement in AUROC and sensitivity compared to the baseline. The IsoMaxPlus method obtains the lowest performance of all models, with an AUROC score of 72.7%, 1% lower than that of the baseline. Leave-one-center-out cross-validation results are similar in aggregate and per-center performance, indicating that generalization to new centers does not entirely depend on the presence of data from those centers in the training set. There are differences in the performance of the models across centers regardless of the cross-validation approach, with PCC and JH typically performing about 10% worse than the average.

In terms of uncertainty estimation, there is a clear advantage to using temperature calibration, as all models exhibit an improvement of 4-9% in ECE and 2-5% in Brier score. Overall, the SNGP model obtains the lowest ECE score at 2.1% (9% improvement over the baseline), while the deep ensemble method obtains the lowest Brier score at 42.7% (5% improvement over the baseline). The IsoMaxPlus method exhibits the second highest ECE score (10.1), only 1% improvement better than the baseline ECE score, and 7% worse than the baseline + TC method's ECE score.

The performance-rejection analysis is shown in Figure 3. Most methods show a monotonically increasing AUROC when rejecting low-confidence patch predictions, indicating good uncertainty calibration. The exception is IsoMax, which shows a drop in performance for the highest confidence predictions. For core-wise predictions, all models exhibit a drop in performance at the highest rejection rate, with the exception of the baseline. The IsoMaxPlus method shows a severe drop in performance at the highest threshold, and the Ensemble and SNGP methods show a less significant one. As can be seen from both plots, deep ensemble has the highest AUROC at every

rejection percentage for patch predictions, and has the highest AUROC at the majority of thresholds for core predictions.

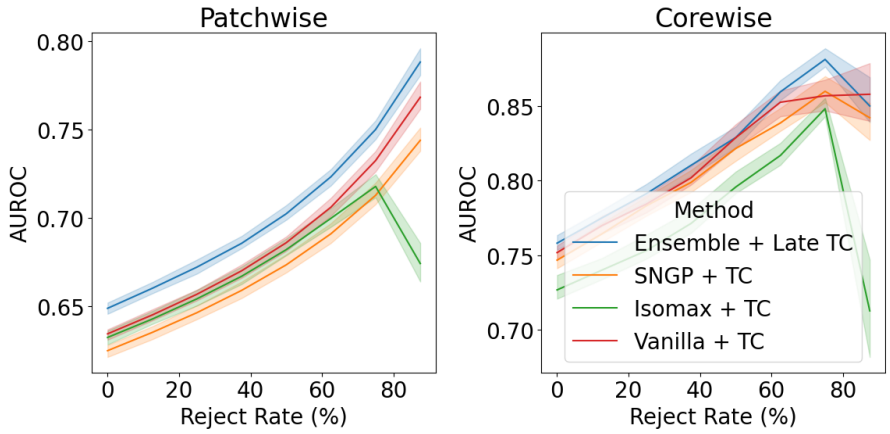

Fig. 3 AUROC for patch-wise predictions (left) and core-wise predictions (right) as the lowest confidence predictions are rejected. Lines represent means and bands represent the 50% confidence intervals across the 10 folds.

4.2 Comparison with clinical benchmarks

Table 3 Comparing the performance of our best model with clinically established biopsy targeting methods as reported in clinical studies.

Method	No. Patients	Sens.	Spec.	B. Acc.
PI-RADS + mp-MRI [1]	576	88	45	66.5
PI-RADS + mp-MRI [25]	1024	90	22	56
PRI-MUS + micro-ultrasound [25]	1024	94	22	58
PRI-MUS + micro-ultrasound [26]	139	92	44	68
Deep Ensembles + micro-ultrasound (ours)	693	68	70	69

Table 3 shows the comparison of our highest performing method with the performance of biopsy targeting protocols in clinical use as reported in clinical studies. Visual analysis of mp-MRI using PI-RADS is historically considered the clinical standard and has high sensitivity but relatively low specificity. Visual analysis of micro-US using the PRI-MUS protocol has a slightly higher range of reported specificity than mp-MRI, with similar sensitivity. Our proposed method has a sensitivity that is around 20% lower than PI-RADS or PRIMUS (68% versus a range of 88%-94%), but has significantly higher specificity (70% versus a range of 22% to 45%. Balanced accuracy scores are similar between our method and the range observed for PI-RADS and PRI-MUS. Although comparison is limited by the fact that these studies used different populations, in general our method performs competitively, underlining its potential as a

useful clinical tool. The higher specificity potentially provided by our method is valuable, as the low specificity of mp-MRI can be associated with superfluous biopsies or inappropriate treatment [27].

4.3 Qualitative evaluation

Figure 4 shows the prediction heatmaps overlaid on the ultrasound image for two cancer and two benign cases, using the Ensemble + Late TC method, and increasing rejection of low-confidence predictions moving from left to right. The model's output generally matches the corresponding biopsy results. Cancer predictions which remain even at high rejection rates (i.e. are confident predictions) are likely to correspond to true cancer, and such regions could be targeted for biopsy.

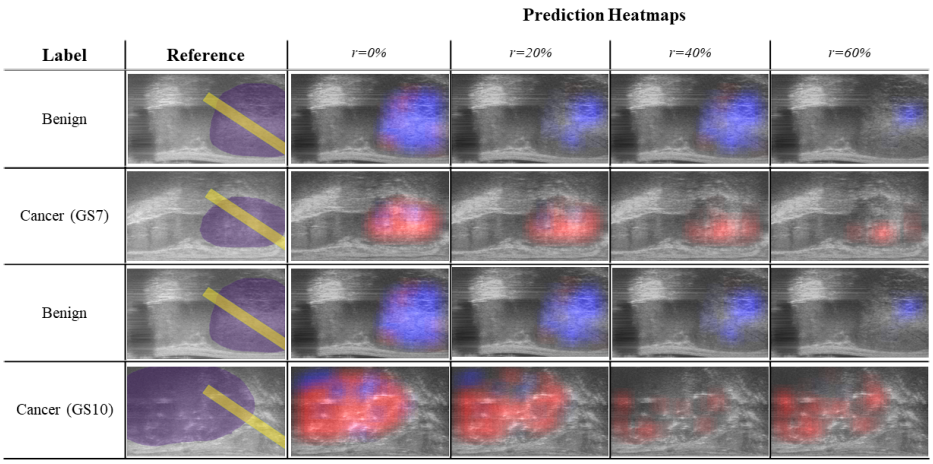

Fig. 4 Visualizing deep ensemble's predictions at four uncertainty thresholds representing rejection rates (r) of 0%, 20%, 40% and 60%. Cancer and benign predictions are highlighted in red and blue, respectively. The prostate region is outlined in purple and the needle trace region is outlined in yellow in the second column.

5 Discussion and Conclusion

As shown in Table 2 and Figure 3, deep ensembles with TC achieve the highest classification performance and the best overall Brier score. SNGP with TC also performs well, but lags behind in terms of Brier score and overall reliability with regards to performance-rejection analysis. Despite its simplicity, TC yields the best gains in uncertainty calibration, even compared to SNGP and IsoMax which currently state-of-the-art methods in computer vision. Key differences in our dataset compared to benchmark computer vision datasets may explain the mediocre performance of the latter two methods: first, the methods make the assumption that the training dataset is "clean", i.e. that the types of data contained therein all represent the types of data that one would want to make confident predictions for – second, they assume low

amounts of irreducible uncertainty (for instance, most training examples in computer vision are clearly distinguishable by human annotators). However, our dataset is not "clean" in this sense - it contains many instances of imaging artifacts, where one would not want to make confident predictions, but which cannot be excluded from training, as well as irreducible uncertainty due to limitations of the imaging modality.

A noteworthy finding of our study is that the accuracy of the models' predictions depends on the zonal anatomy of the prostate. Our observations suggest better performance of the models in the peripheral zone than in the central and transitional zones. Tissue micro-structure varies by zone, and the appearance of cancer is known to be different between zones [28]. It is perhaps not surprising that our model learned the features of cancer for the peripheral zone, since 80% of cancers are in the peripheral zone, and the needle trace predominantly intersects with the peripheral zone tissue due to its proximity to the rectal wall. To improve cancer detection in the anterior, transition, and central zones, future studies should incorporate transperineal biopsy data, which better sample these zones. Our results also suggest worse performance at certain centers independent of whether or not training data from that center were used. More research is needed to determine factors influencing these differences, which could be due to operator technique or patient populations.

A limitation of our study is that we did not evaluate the models' performances on cores with very low involvement of cancer. Due to this limitation, we cannot conclude whether or not our method would be able to identify very small cancerous lesions. This is in part due to the method of aggregating predictions by simple averaging. In addition, all imaging acquisitions used in our study had the same imaging settings, such as depth, focal depth, and transducer frequency. Although these are realistic operating conditions for the system, the models are not guaranteed to generalize if these settings were changed. The comparisons of Section 4.2 on clinical benchmarks are based on values reported from clinical studies on different patient populations. Further studies are needed to directly compare deep learning methodologies with clinical benchmarks on common data and better establish clinical conclusions. Finally, racial/ethnic information for study participants were not collected in our dataset, preventing us from analyzing the connection between these characteristics and model performance, which should be a priority in future work to ensure fair generalization of the models.

Despite these limitations, the potential impact of this work is clear: deep learning methods with uncertainty estimation are a potential solution for direct ultrasound-based biopsy targeting using micro-ultrasound. They have performance that is competitive with PI-RADS and PRI-MUS. Additionally, their strong uncertainty calibration gives a useful "rejection" option by which a clinician could adjust the uncertainty threshold to achieve a desired level of performance and fall back to standard protocols when confident predictions are not available. A possible implementation would be to run the model in real time during biopsy procedure, visualize its predictions as heatmaps, and acquire biopsy samples from confident cancer predictions in addition to the systematic biopsy cores. Based on the findings of this study, this method has the potential to improve the cancer yield of the biopsy procedure compared to PRI-MUS-based targeting alone or no targeting. Future prospective validation and clinical studies should investigate the efficacy of this strategy.

Declarations

This work was supported by the 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 Vector Institute.

Brian Wodlinger is Vice President of Clinical and Engineering at Exact Imaging, and provided access to this dataset. No other author has any potential conflict of interest to disclose.

All patient data were used with informed consent and approval of institutional ethics boards.

References

[1] Ahmed, H.U., Bosaily, A.E.-S., Brown, L.C., Gabe, R., Kaplan, R., Parmar, M.K., Collaco-Moraes, Y., Ward, K., Hindley, R.G., Freeman, A., Kirkham, A., Oldroyd, R., Parker, C., Emberton, M.: 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)

[2] Drost, F.-J.H., Osses, D., Nieboer, D., Bangma, C.H., Steyerberg, E.W., Roobol, M.J., Schoots, I.G.: Prostate magnetic resonance imaging, with or without magnetic resonance imaging-targeted biopsy, and systematic biopsy for detecting prostate cancer: a cochrane systematic review and meta-analysis. European urology **77**(1), 78–94 (2020)

[3] Correas, J.-M., Halpern, E.J., Barr, R.G., Ghai, S., Walz, J., Bodard, S., Dariane, C., Rosette, J.: Advanced ultrasound in the diagnosis of prostate cancer. World journal of urology **39**, 661–676 (2021)

[4] Ghai, S., Eure, G., Fradet, V., Hyndman, M.E., McGrath, T., Wodlinger, B., Pavlovich, C.P.: Assessing cancer risk on novel 29 mhz micro-us images of the prostate: creation of the micro-us protocol for prostate risk identification. The Journal of urology **196**(2), 562–569 (2016)

[5] Sountoulides, P., Pyrgidis, N., Polyzos, S.A., Mykoniatis, I., Asouhidou, E., Papatsoris, A., Dellis, A., Anastasiadis, A., Lusuardi, L., Hatzichristou, D.: Micro-ultrasound–guided vs multiparametric magnetic resonance imaging-targeted biopsy in the detection of prostate cancer: a systematic review and meta-analysis. The Journal of urology **205**(5), 1254–1262 (2021)

[6] Cloutier, G., Destrempes, F., Yu, F., Tang, A.: Quantitative ultrasound imaging of soft biological tissues: a primer for radiologists and medical physicists. Insights into Imaging **12**, 1–20 (2021)

[7] Linmans, J., Elfwing, S., Laak, J., Litjens, G.: Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Medical Image Analysis **83**, 102655 (2023)

[8] Turkbey, B., Haider, M.A.: Deep learning-based artificial intelligence applications in prostate mri: brief summary. The British Journal of Radiology **95**(1131), 20210563 (2022)

[9] Fooladgar, F., To, M.N.N., Javadi, G., Samadi, S., Bayat, S., Sojoudi, S., Eshumani, W., Hurtado, A., Chang, S., Black, P., Mousavi, P., Abolmaesumi, P.: Uncertainty-aware deep ensemble model for targeted ultrasound-guided prostate biopsy. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022)

[10] Javadi, G., Samadi, S., Bayat, S., Sojoudi, S., Hurtado, A., Chang, S., Black, P., Mousavi, P., Abolmaesumi, P.: Training deep networks for prostate cancer diagnosis using coarse histopathological labels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 680–689 (2021)

[11] Feng, Y., Yang, F., Zhou, X., Guo, Y., Tang, F., Ren, F., Guo, J., Ji, S.: 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)

[12] 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 Transactions on Medical Imaging **39**(10), 3148–3158 (2020)

[13] 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 (2022). Springer

[14] Wilson, P.F., Gilany, M., Jamzad, A., Fooladgar, F., To, M.N.N., Wodlinger, B., Abolmaesumi, P., Mousavi, P.: Self-supervised learning with limited labeled data for prostate cancer detection in high frequency ultrasound. arXiv preprint arXiv:2211.00527 (2022)

[15] 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). PMLR

[16] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)

[17] Ikromjanov, K., Bhattacharjee, S., Sumon, R.I., Hwang, Y.-B., Rahman, H., Lee, M.-J., Kim, H.-C., Park, E., Cho, N.-H., Choi, H.-K.: Region segmentation of

whole-slide images for analyzing histological differentiation of prostate adenocarcinoma using ensemble efficientnetb2 u-net with transfer learning mechanism. Cancers **15**(3), 762 (2023)

[18] Xu, X., Sanford, T., Turkbey, B., Xu, S., Wood, B.J., Yan, P.: Polar transform network for prostate ultrasound segmentation with uncertainty estimation. Medical Image Analysis **78**, 102418 (2022)

[19] Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems **31** (2018)

[20] Rohrbach, D., Wodlinger, B., Wen, J., Mamou, J., Feleppa, E.: High-frequency quantitative ultrasound for imaging pca using a novel micro-us scanner. Ultrasound in medicine & biology **44**(7), 1341–1354 (2018)

[21] Javadi, G., Bayat, S., Kazemi Esfeh, M.M., Samadi, S., Sedghi, A., Sojoudi, S., Hurtado, A., Chang, S., Black, P., Mousavi, P., Abolmaesumi, P.: Towards targeted ultrasound-guided prostate biopsy by incorporating model and label uncertainty in cancer detection. International journal of computer assisted radiology and surgery **17**(1), 121–128 (2022)

[22] 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)

[23] Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B.: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Advances in Neural Information Processing Systems **33**, 7498–7512 (2020)

[24] Macêdo, D., Ludermir, T.: Enhanced isotropy maximization loss: Seamless and high-performance out-of-distribution detection simply replacing the softmax loss. arXiv preprint arXiv:2105.14399 (2021)

[25] Klotz, L., Lughezzani, G., Maffei, D., Sánchez, A., Pereira, J.G., Staerman, F., Cash, H., Luger, F., Lopez, L., Sanchez-Salas, R., *et al.*: Comparison of micro-ultrasound and multiparametric magnetic resonance imaging for prostate cancer: A multicenter, prospective analysis. Canadian Urological Association Journal **15**(1), 11 (2021)

[26] Dias, N., Colandrea, G., Botelho, F., Rodriguez-Sanchez, L., Lanz, C., Macek, P., Cathelineau, X.: Diagnostic accuracy and clinical utility of micro-ultrasound guided biopsies in patients with suspected prostate cancer. Central European Journal of Urology **76**(1), 25 (2023)

[27] Arafa, M.A., Rabah, D.M., Khan, K., Farhat, K.H., Ibrahim, N.K., Albekairi,

A.A.: False-positive magnetic resonance imaging prostate cancer correlates and clinical implications. Urology Annals (2022)

[28] Steiger, P., Thoeny, H.C.: Prostate mri based on pi-rads version 2: how we review and report. Cancer Imaging **16**(1), 9 (2016)