

# Towards Confident Detection of Prostate Cancer using High Resolution Micro-ultrasound

Mahdi Gilany<sup>1</sup>(✉), Paul Wilson<sup>1</sup>, Amoon Jamzad<sup>1</sup>, Fahimeh Fooladgar<sup>2</sup>, Minh Nguyen Nhat To<sup>2</sup>, Brian Wodlinger<sup>3</sup>, Purang Abolmaesumi<sup>2</sup>, Parvin Mousavi<sup>1</sup>

<sup>1</sup>School of Computing, Queen’s University, Kingston, Canada

**mahdi.gilany@queensu.ca**

<sup>2</sup>Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

<sup>3</sup>Exact Imaging, Markham, Canada

**Abstract.** MOTIVATION: Detection of prostate cancer during transrectal ultrasound-guided biopsy is challenging. The highly heterogeneous appearance of cancer, presence of ultrasound artefacts, and noise all contribute to these difficulties. Recent advancements in high-frequency ultrasound imaging - micro-ultrasound - have drastically increased the capability of tissue imaging at high resolution. Our aim is to investigate the development of a robust deep learning model specifically for micro-ultrasound-guided prostate cancer biopsy. For the model to be clinically adopted, a key challenge is to design a solution that can confidently identify the cancer, while learning from coarse histopathology measurements of biopsy samples that introduce weak labels. METHODS: We use a dataset of micro-ultrasound images acquired from 194 patients, who underwent prostate biopsy. We train a deep model using a co-teaching paradigm to handle noise in labels, together with an evidential deep learning method for uncertainty estimation. We evaluate the performance of our model using the clinically relevant metric of accuracy vs. confidence. RESULTS: Our model achieves a well-calibrated estimation of predictive uncertainty with area under the curve of 88%. The use of co-teaching and evidential deep learning in combination yields significantly better uncertainty estimation than either alone. We also provide a detailed comparison against state-of-the-art in uncertainty estimation.

**Keywords:** prostate cancer · micro-ultrasound · uncertainty · weak labels

## 1 Introduction

Prostate cancer (PCa) is the second most common cancer in men worldwide [18]. The standard of care for diagnosing PCa is histopathological analysis of tissue samples obtained via systematic prostate biopsy under trans-rectal ultrasound (TRUS) guidance. TRUS is used for anatomical navigation rather than cancer targeting. The appearance of cancer on ultrasound is highly heterogeneous and

is further affected by imaging artifacts and noise, resulting in low sensitivity and specificity in PCa detection based on ultrasound alone.

Substantial previous literature and large multi-center trials report low sensitivity of systematic TRUS biopsy. In [3], authors compare diagnostic accuracy of TRUS biopsy and multi-parametric MRI (mp-MRI). They report sensitivity of systematic TRUS biopsy as low as 42-55% compared to 88-96% for mp-MRI. However, they report low specificity of 36-46% for mp-MRI compared to 94-98% for TRUS.

Fusion of mp-MRI imaging with ultrasound can enable targeted biopsy by identifying cancerous lesions in the prostate [13,17]. Fusion biopsy involves either manual or semi-automated registration of lesions identified in mp-MRI with real-time TRUS. This process can be time-consuming and inaccurate due to registration errors and patient motion. It is therefore highly desirable to improve the capability of biopsy targeting using ultrasound imaging alone at the point of care.

The recent development of high frequency “micro-ultrasound” technology allows for the visualization of tissue at higher resolution than conventional ultrasound. A qualitative scoring system based on visual analysis of micro-ultrasound images called the PRI-MUS (prostate risk identification using micro-ultrasound) protocol [6] has been proposed to estimate PCa likelihood. Several studies have shown that micro-ultrasound can detect PCa with sensitivity comparable to that of mp-MRI using this grading system [2,4]. A recent systematic review and meta analysis analyzing 13 published studies with 1,125 total participants found that micro-ultrasound guided prostate biopsy and mp-MRI imaging targeted prostate biopsy resulted in comparable detection rates for PCa [19]. Research on this technology is in early stages and relatively few quantitative methods are reported. Rohrbach et al. [14] use a combination of manual feature selection with machine learning as the first quantitative approach to this problem. Shao et al. [16] use a deep learning strategy with a three-player minimax game to tackle data source heterogeneity. While these studies show significant potential of micro-ultrasound as a diagnostic tool for PCa, methods to-date primarily focus on improving accuracy for cancer prediction. We argue that in addition, confidence in detection of cancer can play a significant role for adoption of this technology to ensure that predictions can be clinically trusted. Towards this end, we propose to address several key challenges.

Machine learning models built from ultrasound data rely on ground truth labels from histopathology that are coarse and only approximately describe the spatial distribution of cancer in a biopsy core [11,14,16]. The lack of finer labels cause two challenges: first, labels assigned to patches of ultrasound images in a biopsy core may not match the ground truth tissue, resulting in weak labels; second, biopsies include other types of tissue such as fibromuscular cells, benign prostatic hyperplasia and precancerous changes. Many of these tissues are unlabeled in a histopathology report, which will result in out-of-distribution (OOD) data. Therefore, effective learning models for micro-ultrasound data should be robust to label noise and OOD samples.

Several solutions have been presented to address the above issues, mainly by quantifying the uncertainty of predictions [1,12,5]. Predictive uncertainty can be used as a tool to discard unreliable and OOD samples. Evidential deep learning (EDL) [15] and ensemble methods [12] are amongst such approaches. In particular, evidential learning is computationally light, run-time efficient and theoretically grounded, hence it fits our clinical purpose here. Learning from noise in labels (i.e. weak labels) has also been addressed before using methods that 1) estimate noise; 2) modify the learning objective function, or 3) use alternative optimization [8]. Among these, co-teaching [9] has been shown to be a successful baseline that can be easily integrated with any uncertainty quantification method.

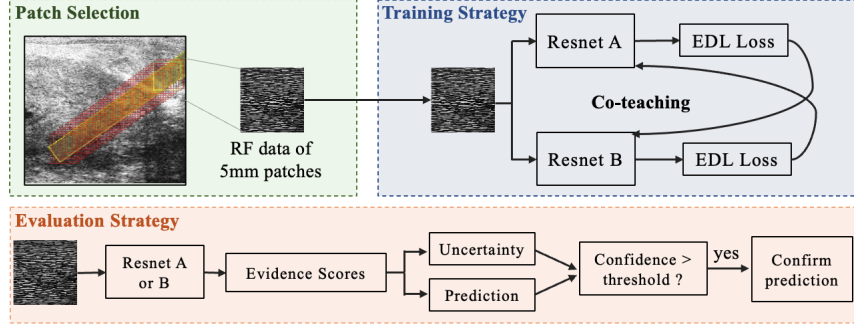
In this paper, for the first time, we propose a learning model for PCa detection using micro-ultrasound that can provide an estimate of its predictive confidence and is robust to weak labels and OOD data. We address label noise using co-teaching and utilize evidential learning to estimate uncertainty for OOD rejection, resulting in confident detection of PCa. We assess our approach by examining the classification accuracy and uncertainty calibration (i.e. the tendency of the model to have high levels of certainty on correct predictions). We compare our methodology to a variety of uncertainty methods with and without co-teaching and demonstrate significant improvements over baseline. We show that applying an adjustable threshold to discard uncertain predictions yields great improvements in accuracy. By allowing correct and confident predictions, our approach could provide clinicians with a powerful tool for computer-assisted cancer detection from ultrasound.

## 2 Materials and Methods

### 2.1 Data

Data is obtained from 2,335 biopsy cores of 198 patients who underwent transrectal ultrasound-guided prostate biopsy through a clinical trial and after institutional ethics approval is provided. A 29 MHz micro-ultrasound system and transducer (ExactVu, Markham) was used for data acquisition. A single sagittal ultrasound image composed of 512 lateral radio frequency (RF) lines was obtained prior to the firing of the biopsy gun for each core. Primary and secondary Gleason grades, together with an estimate of the fraction of cancer relative to the total core area (the so-called “involvement of cancer”) are also provided for each patient. We under-sampled benign cores in order to obtain an equal proportion of cancerous and benign cores during training and evaluation, resulting in 300 benign and 300 cancerous cores, respectively. As in [14], we exclude cores with involvement less than 40% to learn from data that better represents PCa. We hold out the data from 27 patients as a test set, with the remaining 161 used for training and cross-validation.

**Pre-processing:** For each RF ultrasound image, a rectangular region of interest (ROI) corresponding to the approximate needle trace area is determined by using



**Fig. 1.** Top left: Patches are extracted from the needle region. Top Right: During training, “clean” examples are selected by the peer model for training updates. Bottom: The model predicts evidence scores which are used to calculate predictions and uncertainty. Predictions with high uncertainty are rejected.

the angle and location of the probe-mounted needle relative to the imaging plane (Fig. 1, yellow region). This ROI is intersected with a manually drawn prostate segmentation mask to exclude non-prostatic tissue. Overlapping patches are extracted corresponding to  $5\text{ mm} \times 5\text{ mm}$  tissue regions with an overlap of 90% covering the ROI. These patches are up-sampled in the lateral direction and down-sampled in the axial direction by factors of 5 to obtain a uniform physical spacing of pixels in both directions. This results in a patch of 256 by 256 pixels. Ultrasound data in each patch are normalized to a mean of 0 and standard deviation of 1. Patches are assigned a binary label of 0 (benign) or 1 (cancerous) depending on the pathology of the core. The patches and their associated labels are inputs to our learning algorithms.

## 2.2 Methodology

We propose a micro-ultrasound PCa detection learning model that is robust to challenges associated with weak labels and OOD samples. In this section, we first define the problem followed by descriptions of co-teaching as a strategy for dealing with weak labels. Next, we incorporate evidential deep for quantifying prediction uncertainty and excluding suspected OOD data. Finally, we present evaluation metrics to assess our methods.

**Weak Labels and OOD:** Let  $X_i = \{x_1, x_2, \dots, x_{n_i}\}$  refer to a biopsy core where  $n_i$  number of patches extracted from needle region (Figure 1). For each biopsy core  $X_i$ , pathology reports a label  $Y_i$  and the length of cancer  $L_i$  in core, which is a rough estimate between zero and the biopsy sample length. Following previous work in PCa detection [14,11], we assign coarse pathology labels  $Y_i$  to all extracted patches  $\{x_1, x_2, \dots, x_{n_i}\}$  due to the lack of finer patch-level labels. Therefore, many assigned labels to patches may not necessarily

match with the ground truth and they are inherently weak. Additionally, other tissue than cancer, present in the core, does not have any gold standard labels. Therefore, there is also OOD data.

**Co-teaching:** We propose to use a state-of-the-art method, co-teaching, to address label noise for micro-ultrasound data [9]. For weak label methods, we rely on the findings of [11] showing the success of co-teaching method, and [20], which found that co-teaching significantly out-performed other methods such as robust loss functions. This approach simultaneously trains two similar neural networks with different weight initializations. According to the theory of co-teaching, neural networks initially learn simpler and cleaner samples then overfit to noisy input. Therefore, during each iteration, each network picks a subset of samples with lower loss values as potentially clean data and trains the other network with those samples. In a batch of data with size  $N$ , only  $R(e) * N$  number of samples are selected by each network as clean samples, where  $R(e)$  is the ratio of selection starting from 1 and gradually decreasing to a fixed value  $1 - \gamma$ . Formally we have  $R(e) = 1 - \min(\frac{e}{e_{\max}}, \gamma)$ , where  $\gamma \in [0, 1]$  is a hyper-parameter, and  $e$  and  $e_{\max}$  are the current and maximum number of epochs, respectively. Using two networks prevents confirmation bias from arising.

**Evidential Deep Learning:** Evidential deep learning (EDL) [15] uses the concepts of *belief* and *evidence* to formalize the notion of uncertainty in deep learning. A neural network is used to learn the parameters of a prior distribution for the class likelihoods instead of point estimates of these likelihoods. Given a binary classification problem where  $P(y = 1|x_i) = p_i$ , instead of estimating  $p_i$ , the network estimates parameters  $e_0, e_1$  such that  $p_i \sim \text{Beta}(e_0 + 1, e_1 + 1)$ . These parameters are then referred to as evidence scores for the classes, and used to generate a belief mass and uncertainty assignment, via  $b_0 = \frac{e_0}{S}, b_1 = \frac{e_1}{S}, U = \frac{2}{S}$ , where  $S = \sum_{i=0}^1 e_i + 1$ . Note that  $b_0 + b_1 + U = 1$ .  $U$  ranges between 0 and 1 and is inversely proportional to our overall level of belief or evidence for each class. It is worth mentioning that term confidence is also used often instead of uncertainty with confidence being  $1 - U$ .

The network is trained to minimize an objective function based on its Bayes Risk as an estimator of the likelihoods  $p_i$ . If the network produces evidences  $e_0, e_1$  for sample  $i$ , the loss and predicted uncertainty for this sample are

$$\mathcal{L}_i = \sum_{i=1}^n E_{p_i \sim \text{Beta}(e_0+e_1)}(|p_i - y_i|^2), U_i = \frac{2}{e_0 + e_1 + 2}, \quad (1)$$

where  $e_0$  and  $e_1$  are the network outputs. The loss also incorporates a KL divergence term, which encourages higher uncertainty on predictions that do not contribute to data fit. The method offers a combination of speed (requiring only a single forward pass for inference) and well-calibrated uncertainty estimation with a solid theoretical foundation.

**Table 1.** Effect of co-teaching on accuracy and calibration error.

| Method                      | AUC                        | Sensitivity         | Specificity         | Patch B-accuracy           | ECE                           |
|-----------------------------|----------------------------|---------------------|---------------------|----------------------------|-------------------------------|
| EDL                         | <b>88.27</b><br>$\pm 2.66$ | 71.32<br>$\pm 1.23$ | 84.80<br>$\pm 7.01$ | 67.47<br>$\pm 2.47$        | 0.1989<br>$\pm 0.0142$        |
| EDL + Co-teaching<br>(ours) | <b>87.76</b><br>$\pm 1.82$ | 67.38<br>$\pm 4.91$ | 88.20<br>$\pm 6.85$ | <b>71.25</b><br>$\pm 1.16$ | <b>0.1379</b><br>$\pm 0.0258$ |

**Clinical Evaluation Metrics:** The goal of our model is to provide the operator with clinically relevant information, such as real-time identification of potential biopsy targets. It should also state the degree of confidence in its predictions such that the operator can decide when to accept the model’s suggestions or defer to their own experience. To measure these success criteria, we propose several evaluation metrics.

Accuracy reported at the level of patches (the basic input to the model) can be misleading due to weak labeling (some correct predictions are recorded as incorrect because of incorrect labels). Therefore, we propose accuracy reported at the level of biopsy cores as a more relevant alternative. We determine core-based accuracy using core-wise predictions aggregated from patch-wise predictions for the core. Specifically, the average of patch predicted labels is used as a probability score that cancer exists in the core [20,21]. To model uncertainty at the core level, patch-wise predictions that do not meet a specified confidence threshold are ignored when calculating this score, and if more than 40% of the patch predictions for a core fall below this threshold, the entire core prediction is considered “uncertain”.

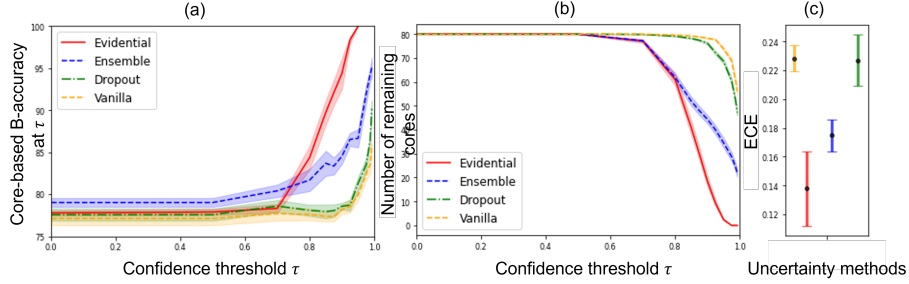
We also use “uncertainty calibration”, a metric that assesses how accurate and representative the predicted uncertainty or confidence is (in terms of true likelihood). To compute calibration, we compute Expected Calibration Error (ECE) [7], which measures the correspondence between predictive confidence and empirical accuracy. ECE is calculated by grouping the predictions so that each prediction falls into one of the  $S$  equal bins produced between zero and one based on its confidence score:

$$\text{ECE} = \sum_{s=1}^S \frac{n_s}{N} |\text{acc}(s) - \text{conf}(s)|, \quad (2)$$

where  $S$  denotes the number of bins (10 used in this paper),  $n_s$  the number of predictions in bin  $s$ ,  $N$  the total number of predictions, and  $\text{acc}(s)$  and  $\text{conf}(s)$  the relative accuracy and average confidence of bin  $s$ , respectively.

### 3 Experiments and Results

From all data, 161 patients (392 cores, 12664 patches) are used for training and a further 40 patients are used as a validation set for model selection and tuning.

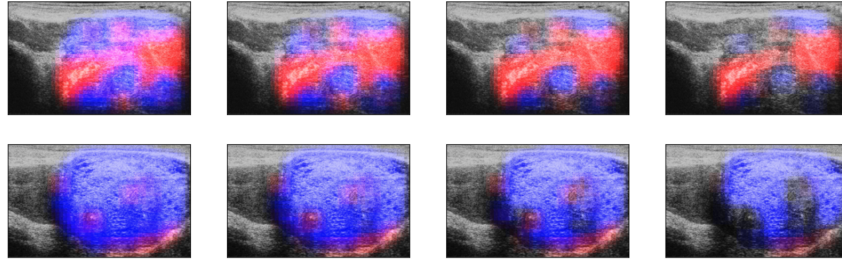


**Fig. 2.** Left: accuracy vs. confidence plot. As we increase the confidence threshold  $\tau$  and retain only confident predictions, the balanced accuracy increases accordingly. Middle: The number of remaining cores following exclusion based on the confidence threshold. Right: the Expected Calibration Error (ECE) error bar plot for all presented uncertainty quantification methods (lower is better).

We hold out a set of randomly selected, mutually exclusive, patients as test set (27 patients, 80 cores, 2808 patches). All experiments, except for the ensemble method, are repeated nine times with three different validation sets, each with three different initializations; the average of all runs is reported. For the ensemble method, as suggested in [12], five different models with different initialization are used for estimating true prediction probabilities,  $p(y_i|x_i)$ . This process is done with five different validation sets, resulting in a total of 25 runs. As a backbone network, we modify ResNet18 [10] by using only half of the layers in each residual block. We found this reduction in layers to improve model performance, likely by reducing overfitting. Two copies of modified ResNet with different initializations are used for the co-teaching framework. For our choice of  $\gamma$ , we empirically found 0.4 to be the best. We employ the NovoGrad optimizer with learning rate of  $1e-4$ .

### 3.1 Effect of Co-teaching

To determine the effects of weak labels and the added value of co-teaching, we design an experiment comparing EDL with co-teaching to EDL alone. Table 1 shows a promising improvement in both ECE score and patch-based balanced accuracy (Patch B-accuracy) when the co-teaching is employed. We report sensitivity, specificity and area under the curve (AUC) metrics for cores. Counter-intuitively, we observe that gains in patch-wise accuracy with co-teaching are not reflected in these metrics. We hypothesize that the averaging from patch-wise to core-wise predictions may sufficiently smooth the effects of noisy labels at this level. We emphasize that the AUC for *both* methods is at least 10% higher than AUC achieved using conventional ultrasound machines [11], underlining the strong capabilities of high-frequency ultrasound.



**Fig. 3.** Heatmaps representing predictions of cancer (red) or benign (blue). The confidence threshold is increased from left to right as  $[0.7, 0.8, 0.85, 0.9]$ , progressively excluding more uncertain predictions. The top row is from cancerous core with Gleason score 4+3; the bottom row is from a benign core.

### 3.2 Comparison of Uncertainty Methods

Quantification of predictive uncertainty could help clinical decision making during the biopsy procedure by only relying on highly confident predictions and discarding OOD and suspect samples. We examine EDL predictive uncertainty using *accuracy vs. confidence plots* in this section, and illustrate how it may be utilised to eliminate uncertain predictions while achieving high accuracy on the confident ones. Then, we compare EDL predictive uncertainty with MC Dropout [5] and deep ensemble [12] methods.

In our *accuracy vs. confidence plot*, Figure 2 (a), we plot core-based balanced accuracy as a function of the confidence threshold  $\tau \in [0, 1]$  used to filter out underconfident patch-level predictions. Patches with predicted confidence less than  $\tau$ , i.e. predictive uncertainty more than  $1 - \tau$ , are discarded. If at least 60% of extracted patches for a biopsy core remain, the average of the remaining patch predictions is used as core-based prediction. We observe the increase in core-based accuracy as the threshold increases, showing that confident predictions tend to be correct. As shown in Figure 2 (b), there is a natural trade-off, with increased threshold values also resulting in increased numbers of rejected cores, yet with well-calibrated uncertainty methods it is not necessary to discard a high fraction of cores in order for uncertainty thresholding to result in meaningful accuracy gains. In Figure 2 (c), we compare the quality of predictive uncertainty of all methods via ECE score. Our experiments show that EDL achieves the best calibration error while providing the best balance between high accuracy and core retention at different threshold levels.

### 3.3 Model Demonstration

As a proof-of-concept for the clinical utility of our method, we applied our model as a sliding window over entire RF images and generated a heatmap, where red corresponds to a prediction of cancer and blue to a prediction of benign. Uncertainty thresholds at various levels were applied to discard uncertain predictions



- discarded predictions had their opacity decreased to 0. These maps were overlaid over the corresponding B-mode images to visualize the spread of cancer. An example of heatmaps for a cancerous and benign core are shown in Figure 3. The cancerous image shows a large amount of red which focuses on two main regions as the confidence threshold increases. By the results of Figure 1, we can say that these loci are very likely to be cancerous lesions and good biopsy targets. The benign image, on the other hand, shows a dominance of blue, with two small red areas that disappear as the threshold increases. These are most likely areas of OOD features on which the model correctly reported high levels of uncertainty. These images show the subjective quality of our model’s performance and the utility of an adjustable uncertainty threshold.

## 4 Conclusion

We proposed a model for confident PCa detection using micro-ultrasound. We employed co-teaching to improve robustness to label noise, and used evidential deep learning to model the predictive uncertainty of the model. We find these strategies to yield a significant improvement over baseline in the clinically relevant metrics of accuracy vs. confidence. Our model provides crucial confidence information to interventionists weighing the recommendations of the model against their own expertise, which can be critical for the adoption of precision biopsy targeting using TRUS.

**Acknowledgement.** This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (CIHR).

## References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* **76**, 243–297 (2021)
2. Abouassaly, R., Klein, E.A., El-Shefai, A., Stephenson, A.: Impact of using 29 mhz high-resolution micro-ultrasound in real-time targeting of transrectal prostate biopsies: initial experience. *World journal of urology* **38**(5), 1201–1206 (2020)
3. 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., 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)
4. Eure, G., Fanney, D., Lin, J., Wodlinger, B., Ghai, S.: Comparison of conventional transrectal ultrasound, magnetic resonance imaging, and micro-ultrasound for visualizing prostate cancer in an active surveillance population: a feasibility study. *Canadian Urological Association Journal* **13**(3), E70 (2019)

5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
6. Ghai, S., Eure, G., Fradet, V., Hyndman, M.E., McGrath, T., Wodlinger, B., Pavlovich, C.P.: Assessing cancer risk on novel 29 mhz micro-ultrasound images of the prostate: creation of the micro-ultrasound protocol for prostate risk identification. *The Journal of urology* **196**(2), 562–569 (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
8. Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I.W., Kwok, J.T., Sugiyama, M.: A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406* (2020)
9. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems* **31** (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. 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. Springer (2021)
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* **30** (2017)
13. Rai, B.P., Mayerhofer, C., Somani, B.K., Kallidonis, P., Nagele, U., Tokas, T.: Magnetic resonance imaging/ultrasound fusion-guided transperineal versus magnetic resonance imaging/ultrasound fusion-guided transrectal prostate biopsy—a systematic review. *European Urology Oncology* **4**(6), 904–913 (2021)
14. Rohrbach, D., Wodlinger, B., Wen, J., Mamou, J., Feleppa, E.: High-frequency quantitative ultrasound for imaging prostate cancer using a novel micro-ultrasound scanner. *Ultrasound in medicine & biology* **44**(7), 1341–1354 (2018)
15. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems* **31** (2018)
16. 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)
17. Siddiqui, M.M., Rais-Bahrami, S., Truong, H., Stamatakis, L., Vourganti, S., Nix, J., Hoang, A.N., Walton-Diaz, A., Shuch, B., Weintraub, M., et al.: Magnetic resonance imaging/ultrasound–fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. *European urology* **64**(5), 713–719 (2013)
18. Smith, L., Bryan, S., De, P., et al.: Canadian cancer statistics advisory committee. *canadian cancer statistics 2018* (2018)
19. 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)
20. To, M.N.N., Fooladgar, F., Javadi, G., Bayat, S., Sojoudi, S., Hurtado, A., Chang, S., Black, P., Mousavi, P., Abolmaesumi, P.: Increasing diagnostic yield of prostate cancer during ultrasound guided biopsy in the presence of label noise
  21. To, M.N.N., Fooladgar, F., Javadi, G., Bayat, S., Sojoudi, S., Hurtado, A., Chang, S., Black, P., Mousavi, P., Abolmaesumi, P.: Coarse label refinement for improving prostate cancer detection in ultrasound imaging. *International Journal of Computer Assisted Radiology and Surgery* **17**(5), 841–847 (2022)