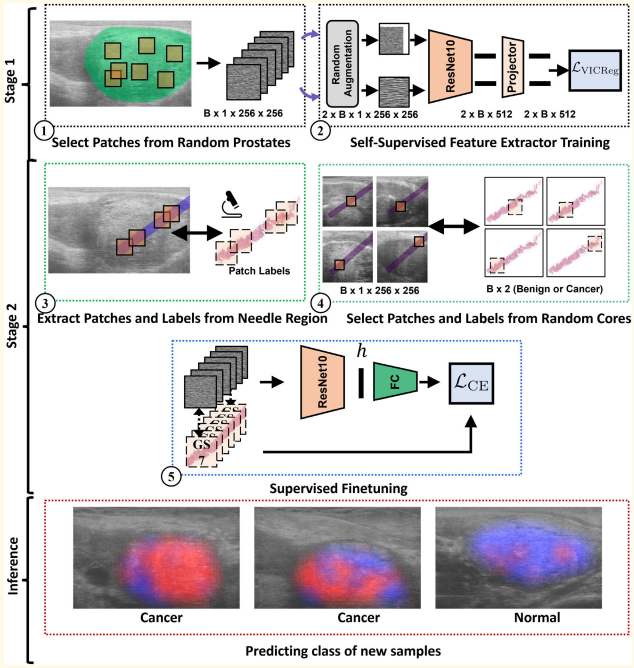
Self-Supervised Learning With Limited Labeled Data for Prostate Cancer Detection in High-Frequency Ultrasound

Paul F. R. Wilson, Mahdi Gilany, Amoon Jamzad, Fahimeh Fooladgar, Minh Nguyen Nhat To, Brian Wodlinger, Purang Abolmaesumi, *Senior Member, IEEE*, and Parvin Mousavi, *Senior Member, IEEE*

***Abstract*—Deep learning-based analysis of high-frequency, high-resolution micro-ultrasound data shows great promise for prostate cancer (PCa) detection. Previous approaches to analysis of ultrasound data largely follow a supervised learning (SL) paradigm. Ground truth labels for ultrasound images used for training deep networks often include coarse annotations generated from the histopathological analysis of tissue samples obtained via biopsy. This creates inherent limitations on the availability and quality of labeled data, posing major challenges to the success of SL methods. However, unlabeled prostate ultrasound data are more abundant. In this work, we successfully apply self-supervised representation learning to micro-ultrasound data. Using ultrasound data from 1028 biopsy cores of 391 subjects obtained in two clinical centers, we demonstrate that feature representations learned with this method can be used to classify cancer from noncancer tissue, obtaining an AUROC score of 91% on an independent test set. To the best of our knowledge, this is the first successful end-to-end self-SL (SSL) approach for PCa detection using ultrasound data. Our method outperforms baseline SL approaches, generalizes well between different data centers, and scales well in performance as more unlabeled data are added, making it a promising approach for future research using large volumes of unlabeled data. Our code is publicly available at www.github.com/MahdiGilany/SSL_micro_ultrasound.**



***Index Terms*—Micro-ultrasound, prostate cancer (PCa), prostate imaging, self-SL (SSL), ultrasound imaging.**

Manuscript received 29 May 2023; accepted 17 July 2023. Date of publication 21 July 2023; date of current version 29 August 2023. This work was supported in part by Canadian Institutes for Health Research (CIHR) and in part by Natural Sciences and Engineering Research Council of Canada (NSERC). The work of Parvin Mousavi was supported in part by a Canada CIFAR AI Chair and in part by the Vector Institute. *(Purang Abolmaesumi and Parvin Mousavi are co-senior authors.) (Paul F. R. Wilson and Mahdi Gilany contributed equally to this work.) (Corresponding author: Mahdi Gilany.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Boards (IRBs) of each center (ethical board), and followed the rules of patient recruiting and informed consent of the trial (clinicaltrials.gov NCT02079025).

Paul F. R. Wilson, Mahdi Gilany, Amoon Jamzad, and Parvin Mousavi are with the School of Computing, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: 1pfrw@queensu.ca; abbasgilani1996@gmail.com; mousavi@queensu.ca).

Fahimeh Fooladgar, Minh Nguyen Nhat To, and Purang Abolmaesumi are with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

Brian Wodlinger is with Exact Imaging, Markham, ON L3R 2N2, Canada.

Digital Object Identifier 10.1109/TUFFC.2023.3297840

I. Introduction

A. Image-Based Cancer Detection

PROSTATE cancer (PCa) is the second most common cancer diagnosed in men worldwide [1]. Early and accurate detection and staging of PCa is critical to guide treatment decisions. The standard of care for diagnosing PCa is histopathological analysis of tissue samples using the Gleason grading system where microscopic patterns of the tissue are used to determine cancer grades (1 to 5, higher grade is more severe cancer), and the grades of the two dominant tissue patterns are added and reported as the Gleason score (GS). Tissue samples are obtained via needle biopsy, typically under the guidance of transrectal ultrasound (TRUS). TRUS is used for navigation but not for targeting the biopsy, as historically it has lacked sufficient accuracy in identifying cancerous lesions [2]. Instead, freehand prostate biopsy is primarily *systematic* where a number of biopsy cores are collected, in a specific pattern, from the prostate with the aim of obtaining

Highlights

- **Novelty: Self-supervised learning on abundant unlabeled data is used to classify prostate cancer in high-frequency ultrasound, achieving an impressive AUROC score of 91%.**
- **Results: Self-supervised learning outperforms supervised learning methods by a significant margin in classifying cancer from non-cancer tissue in micro-ultrasound data.**
- **Implications: Self-supervised learning has the potential to greatly improve prostate cancer detection and could be applied to other medical imaging modalities.**

enough samples to identify cancer, should it be present. Still, cancer is frequently missed and many patients with a negative biopsy will eventually require re-biopsy and be diagnosed with cancer [3]. In other cases, men undergo unnecessary biopsies for benign pathologies or indolent cancers where a watch-and-wait approach may be preferable. Biopsies carry the risk of adverse events [3]. Any improvements in the ability to detect or rule out cancer by direct analysis of ultrasound images would have a major impact on patient outcomes.

A substantial body of literature has established the limitations of B-mode ultrasound for PCa detection [2]. However, analysis of raw radio frequency (RF) echo data is more promising, as the frequency and phase information contained therein has been shown to correlate with tissue properties. This reasoning is the foundation for quantitative ultrasound (QUS) methods which compute envelope statistics and backscatter coefficients from RF data and associate them with tissue microstructure [4], [5]. QUS has shown to improve PCa detection compared to analysis of B-mode images [6], [7]. In particular, manual features selected from QUS combined with machine learning methods has seen considerable success [6], [7]. Although handcrafted feature selection allows for better explainability and suffers less from overfitting, it is restricted to a relatively small number of features and may miss unknown properties in the raw RF data that correlate with PCa. Deep learning approaches, however, allow feature extraction from raw data and are increasingly outperforming classical models in other areas of medical imaging [8].

Other ultrasound-based methods, other than QUS, have also been proposed to improve PCa detection. Doppler ultrasound has been used for cancer detection by measuring angiogenesis associated with tumor formation [9], [10]. Elastography-based methods have shown promise for detecting PCa by measuring tissue stiffness [11], [12], while temporal enhanced ultrasound has been employed for PCa detection by enhancing the resolution of imaging of tissue microsctructures [13], [14], [15]. Another technique that employs intravenously administered microbubble contrast agents to enhance the visibility of blood vessels and other structures during ultrasound imaging is CEUS [16], [17]. CEUS has been shown to improve the sensitivity and specificity of ultrasound in detecting PCa, especially in cases where the cancer is small and difficult to detect with conventional ultrasound techniques. Furthermore, combining different ultrasound modalities, such as B-mode, shear-wave elastography, CEUS, etc., in multiparametric Ultrasound (mpUS) has been proposed to take advantage of all methods and further enhance PCa detection [18]. However, these methods all use conventional clinical imaging frequencies (6–14 MHz) that only afford limited spatial resolution, and may hinder the ability to robustly identify PCa.

The emerging state-of-the-art is to combine TRUS with multiparametric MRI (mp-MRI), which has a higher sensitivity (88%–96% compared to 42%–55% for conventional TRUS) in detection of PCa [3], [19]. Fusion of mp-MRI with TRUS enables biopsy targeting by identifying suspicious lesions in the prostate [20], [21]. However, the availability of MRI is limited, and fusion biopsy requires image registration that can be prone to errors due to patient movement. The improvement of biopsy targeting using TRUS directly is, therefore, highly desirable.

Recently developed high-frequency "micro-ultrasound" technology allows imaging the prostate at a much finer spatial resolution [22]. Clinical studies show that micro-ultrasound has a sensitivity comparable to mp-MRI using the qualitative ultrasound-based "PRI-MUS" scoring system [23], [24], and a recent meta-analysis of 13 published studies with 1125 participants concludes that micro-ultrasound-guided biopsy has similar PCa detection rates as mp-MRI fusion biopsy [25]. However, analysis of RF data remain relatively unexplored for micro-ultrasound: these are limited to a single study using machine learning with QUS [26] and two studies [27], [28] using deep learning.

B. Machine Learning-Based Cancer Detection

While promising, there are key challenges to the development of machine learning-based cancer detection models that are clinically useful. We argue that not all of these challenges have been adequately addressed for micro-ultrasound, and present a self-SL (SSL) approach as a solution. The challenges are as follows.

1) Weak Labeling: Machine learning methods for PCa detection rely on pathology annotations as ground truth labels for corresponding ultrasound data. These annotations are only coarse approximations of the distribution of cancer in the tissue. When using the label of "malignant," for instance, it is not known precisely which areas within the needle region were cancer and which were benign. This weak labeling can severely impact the robustness of deep learning models which tend to memorize incorrect labels [29].

2) Heterogeneity: Prostate tissue includes normal tissue, benign conditions, precancerous changes, and cancers ranging

from indolent to highly aggressive. Within each of these categories there is significant variability in tissue characteristics as well. Ultrasound is subject to noise and imaging artifacts that further increase heterogeneity in tissue appearance. It is, therefore, challenging to train models that generalize to unseen data, specifically to tissue variations that may appear as out-of-distribution (OOD).

3) Distribution Shift: Ultrasound data are prone to major distribution shifts due to differences in equipment, clinical settings, and patient populations [30]. A typical example is the distribution shift between clinical centers. Standard deep learning approaches are not robust to distribution shifts [31], limiting the clinical translation of these models.

4) Data Scarcity: Obtaining labeled ultrasound data requires a biopsy and additional time-intensive annotation from a human expert. This strongly limits the availability of large datasets to train and evaluate models. Deep learning methods have historically relied on copious amounts of data to learn and generalize well; labeled data scarcity is a major challenge on its own that also exacerbates other previously mentioned challenges.

Previously various solutions to these challenges have been proposed. For weak labeling, Zou et al. [32] propose a noisy annotation-tolerant network for breast ultrasound segmentation. Javadi et al. [33] and [34] propose to use multiinstance learning and coteaching for PCa detection. Uncertainty estimation allows a model to express uncertainty when seeing OOD data rather than making false predictions; such methods have been applied to prostate [28], [35] and breast [36], [37] ultrasound to address heterogeneity. Shao et al. [27] propose to handle intercenter distribution shift on micro-ultrasound data by training an adversarial auxiliary model to predict data center origin from the hidden features of a cancer detection model, thereby encouraging the detection model to be invariant under distribution shifts. For data scarcity, transfer learning from larger natural image datasets [38] or learning from synthetic data [39] has been applied to B-mode (but not RF) ultrasound. While these strategies have addressed individual angles of the problem, none are a unified solution, and none adequately address data scarcity.

Additionally, these approaches to PCa detection all follow a supervised learning (SL) approach, while *self-supervised* methods remain unexplored. However, there are intuitive, theoretical, and empirical reasons that make SSL a potentially unified solution. SSL allows learning without labels, sidestepping issues of weak label memorization. SSL has been empirically found to improve robustness to label corruptions [40], and improve model uncertainty. SSL has also been shown to be more robust to dataset imbalance [41] which may improve performance on under-represented natural tissue variations. Self-supervised models can be more robust to dataset-level distribution shift [42] and have better transfer learning performance [43] than their supervised counterparts. The benefits of transfer learning using SSL on domain-specific data have been shown for a variety of X-ray and histology slide image tasks [44]. Finally, and possibly the most compelling, is that SSL enables learning with much more abundant unlabeled data, addressing the data scarcity challenge directly.

C. Contributions

We conduct a study on SSL methods for RF micro-ultrasound data. To the best of our knowledge, this is the first application of SSL for automatic PCa detection using RF ultrasound (either at micro-ultrasound *or* conventional frequencies). Through extensive experiments on data from two clinical centers involving 391 total subjects (1028 total biopsy cores), we demonstrate that the following.

1) SSL significantly improves PCa detection compared to SL alone. By using unlabeled data, SSL allows the model to learn features from a greater volume of data and wider range of tissue types, addressing the issues of data scarcity and heterogeneity.
2) Even when using matched amounts of data, SSL still significantly outperforms SL by avoiding label memorization and reducing the impact of weak labels.
3) SSL models outperform SL models when used for transfer learning between datasets. By capturing useful general features of RF ultrasound data that are independent of specific distributions, SSL alleviates distribution shift.

II. Materials

A. Data Acquisition

We use data from a multicenter clinical trial (Multicenter Trial of High-resolution TRUS Versus Standard Low-resolution TRUS for the Identification of Clinically Significant PCa, clinicaltrials.gov, NCT02079025). Data collected from 391 patients at two sites were included in our study. Urology of Virginia (UVA), Virginia Beach, VA, USA and Centre de Recherche sur le Cancer (CRCEO), Quebec City, QC, Canada. Subjects underwent systematic TRUS-guided prostate biopsy using the ExactVu micro-ultrasound system (ExactVu, Markham, ON, Canada). The system consists of a side-mounted linear array with 512 evenly spaced transducers covering an area of 46.06 mm. The system operates with a pulse frequency up to 29 MHz with center frequency of 21 MHz (compared to the standard 9–14 MHz range of conventional ultrasound), capturing a high-resolution ultrasound image of the prostate down to 70 μm, producing 300% improvement over the traditional systems standards [22]. For each biopsy location, raw RF ultrasound images of the tissue were saved immediately prior to the biopsy gun being fired. The biopsy needle enters the tissue at a fixed angle relative to the imaging plane. The approximate needle trace region was determined using this known angle and the penetration depth. RF scans consist of 512 RF lines (lateral dimension) with 10 016 samples (axial dimension). Therefore, each RF scan is an image of shape $(10\,016, 512)$ consisting of echo intensity values. These images correspond to a physical tissue extent of 28 mm axially and 46.06 mm laterally. Moreover, to avoid any potential dependencies to system settings, ultrasound image parameters affecting RF signals were fixed during all acquisitions for systems in both centers; these parameters include: transducer frequency, focal depth, image depth, gain, and power.

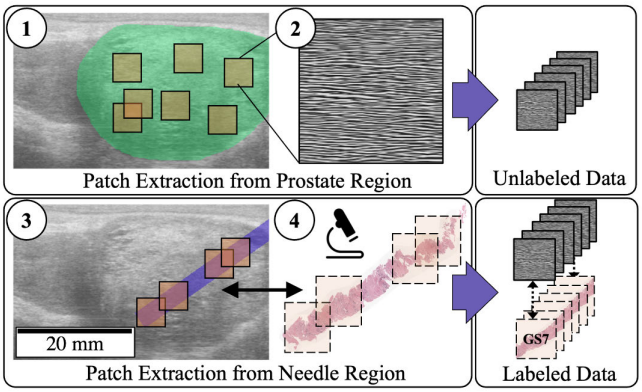


Fig. 1. Summary of data preparation from raw RF ultrasound to model input. (1) For unlabeled data, patches are selected from the prostate region, (2) close up of an RF ultrasound patch, (3) for labeled data, patches are selected from the needle trace region and paired with, and (4) tissue labels from histopathology of the corresponding tissue sample.

Patients underwent a standard 12-core biopsy procedure. Cores were analyzed histopathologically to determine the GS, primary and secondary Gleason grades, and an approximate percentage of cancer (termed the "involvement"). The GS was converted to binary labels of 0 (benign; GS $\leq$ 6) and 1 (cancer; GS $\geq$ 7) which were assigned to patches in the needle trace region of the corresponding biopsy core. Additional data available for each patient are prostate-specific antigen (PSA), mean value 6.7 UVA, 6.3 CRCEO; age, mean 63.5 UVA, 63.9 CRCEO; and family history of cancer, true for 22% of UVA and 28% of CRCEO patients. Among 3804 total cores obtained, the majority (86.4% UVA, 85.4% CRCEO) are benign. Cancerous cores with GS 7, 8, 9, and 10, respectively, makeup 9% UVA, 8% CRCEO; 2% UVA, 2% CRCEO; 1% UVA, 1% CRCEO; and 0.5% UVA, 0% CRCEO of the remaining cores. This selection is done by randomly choosing patients for the test set until the number of cancer cores in the test set compared to the remaining cores reaches the desired ratio of approximately 25% to 75%. To balance the cancerous and benign classes, we under-sample the benign cores to match the number of cancerous cores.

B. Data Preprocessing

Following the previous literature [26], [27], patches corresponding to a tissue area of 5 $\times$ 5 mm are extracted from the RF images as input to our models. A prostate segmentation mask was manually drawn for each core. The prostate occupied an average 45.46% of the area of each image, while the needle region occupied 7.2%. For the unlabeled datasets, patches were extracted from within the prostate region with the shift of 3 mm in axial or lateral directions for extracting each patch (similar to convolutional neural networks strides). For the labeled datasets, patches were selected from within the intersection of needle trace region and prostate region with the shift of 1 mm, and labeled 0 (benign) or 1 (malignant) based on the pathology findings of the core. We considered a patch to be within the needle or prostate region if it overlaps by at least 66% with the needle trace or prostate mask, respectively (Fig. 1). The labeled and unlabeled datasets corresponding to the UVA and CRCEO centers are denoted by $\mathcal{D}^{\text{UVA}}_{\text{needle}}$, $\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$, $\mathcal{D}^{\text{UVA}}_{\text{prostate}}$, and $\mathcal{D}^{\text{CRCEO}}_{\text{prostate}}$. The quantity of data used for pretraining, training, and testing in these different datasets is summarized in Table I, with the quantity of data reported based on the number of patches, cores, and patients analyzed.

TABLE I
NUMBER OF DATA IN EACH DATASET FOR SEED 0 AFTER PREPROCESSING

	Pretrain/Train			Test		
	Patches	Cores	Patients	Patches	Cores	Patients
$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	11,230	286	119	3,447	81	27
$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	7,363	198	81	2,194	106	25
$\mathcal{D}^{\text{UVA}}_{\text{prostate}}$	123,039	286	119	-	-	-
$\mathcal{D}^{\text{CRCEO}}_{\text{prostate}}$	72,104	198	81	-	-	-

Each 5 $\times$ 5 mm patch, originally consisting of 1794 $\times$ 55 pixels, was downsampled axially and linearly interpolated laterally to a size of 256 $\times$ 256 pixels; this is to adapt to the ResNet input size and ensure that the patches have a suitable shape for the designed convolutional filters. Note the RF lines' data sampling frequency is 280 MHz, so no loss of information due to aliasing occurs during this resizing. Each patch was instance-normalized by computing its mean and standard deviation, truncating pixels which fall above or below four standard deviations from the mean, and rescaling to the range (0, 1).

C. Data Augmentations

Selection of data augmentations is an important consideration when using SSL methods for computer vision as they drive the learning objective: Networks are trained to extract features which are invariant under different augmentations. In doing so the network learns to extract high-level semantic features that do not depend on the specific pixel-level features of an image. Augmentations should significantly distort the input image such that the task is difficult, but not so much as to destroy high-level features relevant to downstream tasks.

We reasoned that the standard natural image augmentation pipeline for SSL (resized cropping, random application of Gaussian filters, and random color jitter [45], [46]) is unlikely to be the correct choice for RF ultrasound, as these transformations alter frequency content which contains important tissue information. Instead, we use a combination of rigid transformations and masking.

1) `random_translation`: Translating the image by a factor of up to 0.2 in the horizontal and vertical directions independently, and filling the empty pixel values with 0.5.
2) `random_erasing`: Selecting and filling with the value 0.5 a single rectangular patch with a height and width factor between 0.02 and 0.1 the image size. Note 0.5 is selected as the mean of pixel values to avoid introducing bias.
3) `random_vertical_flip`.
4) `random_horizontal_flip`.

We also introduce several augmentations handcrafted specifically to the physics of RF ultrasound. Ultrasound physics-inspired augmentations have been proposed previously for B-mode [47] but not RF. Our approach considers the decomposition of an RF line into envelope and instantaneous frequency. Let $s(t)$ for $t \in \mathbb{R}$ denote an RF echo (the following analysis can be easily converted to discrete time samples, but for clarity we use the continuous time variable). Then the analytic representation of s, denoted by $s_a(t)$, is given by

$$s_a(t) = s(t) + j\mathcal{H}(s)(t) \tag{1}$$

where $\mathcal{H}(\cdot)$ is the Hilbert transform operator

$$\mathcal{H}(f)(t) = \frac{1}{\pi}\text{p.v.}\int_{-\infty}^{\infty}\frac{f(t)}{\tau - t}d\tau \tag{2}$$

(where $p.v$ denotes the Cauchy principle value definition of the integral), and j is the imaginary unit. When s is real, so is $\mathcal{H}(s)$; thus the original signal can be recovered via $s(t) = \Re\mathfrak{e}(s_a(t))$. By writing the analytic representation as $s_a(t) = A(t)e^{j\phi(t)}$, we can call $\phi(t)$ the *instantaneous phase*, $\phi'(t)$ the *instantaneous frequency*, and $A(t)$ the *instantaneous amplitude* (or envelope) of the signal. Our augmentations are:

1) `random_phase_shift`—Shifts the phase of the signal without changing the instantaneous envelope or frequency.

$$T(s)(t) = \Re\mathfrak{e}\left(s_a(t)e^{i\theta}\right);\ \theta \sim \text{Uniform}(0, 2\pi).$$

2) `random_envelope_distort`—Alters the envelope but not the phase or frequency ("Noise" is low-pass filtered white noise)

$$T(s)(t) = \Re\mathfrak{e}(s_a(t)(1 + n(t)));\ n \sim \text{Noise}.$$

Note that when applied to patches, the same randomly sampled transformation is applied to each RF line in the patch.

All of these augmentations are used together during self-supervised and (optionally) supervised training. Each type of augmentation is chosen randomly to either transform a patch or skip it with a probability of 0.5. It should be noted that each type of augmentation is independently applied to each patch, so it is possible for different combinations of augmentations to be applied to the same patch during training.

III. Methods

We use SSL to learn high-quality feature representations for RF ultrasound using unlabeled data. This allows us to sidestep any problems associated with weak labels while drastically increasing the pool of data available for training. Following standard practice in SSL, we use a two-stage pipeline consisting of: 1) self-supervised training (also referred to as "pretraining") and 2) supervised finetuning and evaluation. This pipeline is illustrated in Fig. 2.

A. Self-Supervised Representation Learning

We use SSL to train a backbone network (in our case, a modified ResNet [48] architecture with one convolutional layer instead of two in each block, with a total of ~6M parameters) to extract low-dimensional abstract feature representations from high-dimensional raw RF ultrasound data. We study a number of methods following the successful "Siamese neural networks" concept (see for instance [45], [46], [49], [50]), where two different views (augmentations) of an instance image are mapped to two low-dimensional representation vectors using the shared backbone neural network [Fig. 2(1)].

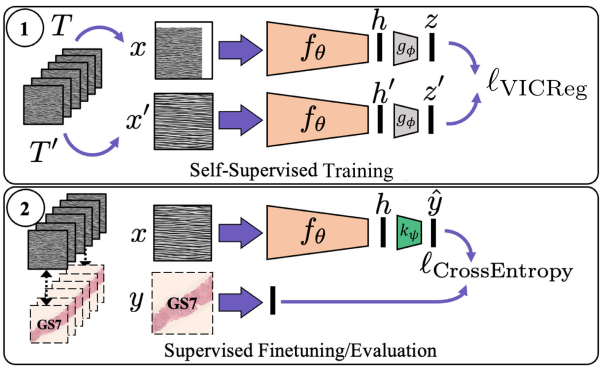


Fig. 2. Summary of our learning approach. (1) SSL: a feature extractor is trained to reduce the VICReg loss [46] between representations from pairs of augmented views of RF ultrasound data. (2) Supervised finetuning/evaluation: the model is trained to minimize the cross-entropy loss between true histopathology labels and predicted labels, using the (possibly frozen) pretrained feature extractor, f_θ, as a backbone network.

Formally, given an extracted raw RF patch x_i^{raw}, two transformations t and t' are sampled from a distribution τ, and two augmented views are produced, $x_i = t(x_i^{\text{raw}})$ and $x_i' = t'(x_i^{\text{raw}})$. The augmented data are then mapped to representation vectors h_i and h_i' using the backbone network

$$h_i = f_\theta(x_i). \tag{3}$$

Next, the representations are projected to z_i and z_i' using the projection network $z_i = g_\phi(h_i)$. During self-supervised training, θ and ϕ are tuned to minimize a cost function with respect to the pairs (z_i, z_i').

We propose to use the recent Variance-Invariance-Covariance Regularization (VICReg) [46] method which we found to have the best performance among the methods tested. VICReg maximizes the agreement between representations z_i and z_i' using the mean-squared error loss. A trivial solution would be for the network to return constant output regardless of input, a phenomenon called representation collapse. VICReg avoids collapse by encouraging the variance across features in a batch to be above a certain threshold and the covariance between features to be as low as possible; this acts to maximize the information content of representations.

The VICReg loss $\ell(Z, Z')$ is the weighted sum of three regularization terms, called the invariance $i(Z, Z')$, variance $v(Z)$, and covariance $c(Z)$ losses defined as follows:

$$\begin{aligned}\ell(Z, Z') = {} & \lambda i(Z, Z') \\ & + \mu\left[v(Z) + v(Z')\right] \\ & + \nu\left[c(Z) + c(Z')\right]\end{aligned} \tag{4}$$

where the weights λ, μ, and ν are hyperparameters and Z and Z' denote the batches of projection vectors z_i and z_i'

(i.e., Z is the matrix whose rows are $[z_1, z_2, \ldots, z_n]$ for batch size n). z^j denotes the jth column of Z, that is the vector composed of the jth feature of each projection vector in the batch.

The invariance term $i(Z, Z')$ is the mean-squared error loss between each pair of z_i and z_i'

$$i(Z, Z') = \frac{1}{n}\sum_{i=1}^{n}\|z_i - z_i'\|^2. \tag{5}$$

Minimizing this term forces the network to learn features which are invariant under augmentations of the same input data.

The variance regularization term $v(Z)$ is defined as the hinge function of standard deviation of the projections along with batch dimension, namely

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d}\max\left(0, \gamma - \sigma\left(z^j\right)\right) \tag{6}$$

where d is the feature dimension and

$$\sigma\left(z^j\right) = \sqrt{\epsilon + \mathrm{Var}\left(\left\{z_i^j | i = 1, \ldots, n\right\}\right)} \tag{7}$$

and γ is a threshold value set to 1 in our experiments, and ϵ is a small number added to reduce numerical instability in the standard deviation. Minimizing this loss maintains variance in each feature of the representations, thereby avoiding mode collapse.

The covariance regularization term $c(Z)$ is

$$c(Z) = \frac{1}{d}\sum_{i\neq j} C(Z)_{ij}^2 \tag{8}$$

where $C(Z)$ is the covariance matrix of Z given by

$$C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)\left(z_i - \bar{z}\right)^T, \quad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i. \tag{9}$$

Notice $c(Z)$ is the sum of the off-diagonal coefficients of covariance matrix C. By minimizing this, the covariance between features is forced to be close to 0 which minimizes redundancies due to intercorrelations between features.

During pretraining, the parameters θ and ϕ of the feature extractor network $f_\theta(.)$ and projection network $g_\phi(.)$ are optimized to minimize the VICReg loss across many mini-batches of augmented data pairs.

B. Supervised Finetuning

Following SSL, the feature extractor $f_\theta(.)$ is paired with a linear classification head $k_\psi(.)$ which projects feature representations of input patches to a 2-D vector corresponding to the probability for the benign (0) and cancer (1) class. This network is trained to minimize the cross-entropy loss between the predicted probability and the ground truth class labels within a labeled dataset. Following common practice in SSL literature (e.g., see VICReg [46]), we consider two finetuning strategies; either *linear finetuning*, when only the linear layer weights ψ are optimized; or *semi-supervised finetuning*, where both the linear layer and feature extractor weights ψ and θ are optimized. For all experiments involving finetuning, the results of both finetuning strategies are included in the tables in the results section.

C. Quantitative Evaluation

We measure our model's performance in terms of cancer classification in two ways: First, we measure its performance for patch classification. As previously noted, patch-wise labels are weak labels and classification performance on these weak labels is not a perfect measurement of true performance. (To illustrate this, consider, a model *correctly* predicting the label "benign" on a benign patch from a cancerous core would be registered as a *false* prediction, because the patch would be labeled according to the overall core label of "cancer"). Still, we assume that patch classification performance on weak labels strongly correlates with true performance. Second, we measure the performance of the model for core classification by defining the predicted class probabilities for a core to be the mean of predicted classes (0 or 1) for patches within that core. For both patch-wise and core-wise performance, we measure balanced accuracy (ACC-B, the average of sensitivity and specificity), average precision (Avg-Prec), and area under the receiver operating characteristic curve (AUROC).

When computing the performance metrics for our cancer detection model, we follow a convention used by previous literature [26], [27], [28] and exclude cores with cancer involvement of less than 40% from both our training and testing sets. This choice is justified as: 1) cores with low involvement are less likely to contain features representative of cancer and 2) weak labeling means that using a low-involvement core will result in more incorrectly labeled patches than correctly labeled ones. Training with these cores is, therefore, problematic and the performance metrics using them are invalid.

D. Qualitative Evaluation

To demonstrate the output of our models, we allow the models to predict the tissue type (benign or malignant) of each patch within the intersection of prostate region and needle. These are compiled into a heatmap, where regions of blue correspond to benign predictions and regions of red correspond to malignant predictions. The heatmaps are overlaid over the corresponding B-mode image to show the model's prediction of the spread of cancer. We can roughly compare the output of model to the involvement of cancer estimated in the pathology reports. The number of "cancer" predictions compared to total predictions can be considered the "predicted involvement" of the model for that needle region, and can be approximated visually or calculated numerically. If the predicted involvement is close to true involvement, this reflects good model performance. Note that this comparison is not entirely precise as: 1) the extracted patches have large overlaps and this induces inaccuracies in predicted involvement and 2) the pathology-reported involvement is a rough measurement and may not accurately show the percentage of cancer length in each biopsy core.

IV. Experiments

We designed our experiments to answer four key questions relevant to the clinical application of our model.

1) Does using SSL improve the performance of the models compared to SL alone?
2) Does SSL outperform SL on transfer learning for downstream tasks (e.g., PCa detection on a different dataset)?
3) Are SSL models more robust to intercenter distribution shifts in PCa data than their SL counterparts?
4) Are there statistically significant performance differences between different SSL methods for PCa detection?

To answer these questions, we designed four experiments.

A. Comparing SSL to SL in Performance

We conduct a comparison between self-supervised pretraining followed by finetuning and fully SL. To ensure a comprehensive evaluation, we investigate both scenarios of using UVA dataset and CRCEO. Additionally, we compare our approach with Gilany et al.'s model [28], which employs supervision and coteaching to address label noise and uncertainty. By utilizing the same testing set, we are able to make a meaningful comparison between the two models.

B. Comparing SSL to SL in Transfer Learning

To evaluate the transfer learning capabilities of our approach, we conduct two experiments. We use the UVA dataset as a pretraining dataset and transfer the learned feature extractor model to the CRCEO dataset, allowing the model to re-tune its weights. We further assess the consistency of our findings by conducting the same experiment in reverse data order, CRCEO as pretraining, and UVA as a training and testing dataset.

C. Comparing SSL to SL in Distribution Shift Robustness

To test the robustness to distribution shift, we train the model on the UVA dataset and test the model on the CRCEO dataset. The difference between this and the transfer learning experiment is that in this case, the model does *not* re-tune its weights on the new dataset. We compare the distribution shift performance of supervised and self-supervised models to the "home" distribution performance of a supervised model trained from scratch on CRCEO. To further assess the consistency and robustness of self-supervised models to distribution shift, we conduct the same experiment in reverse data order by training the model on the CRCEO dataset and evaluating it on the UVA dataset.

D. Comparing Contemporary SSL Methods

We compared the quality of representations learned using three different SSL methods, which may be considered as representative of three broad classes: SimCLR [45] representing the contrastive learning family, BYOL [49] representing the momentum-teacher family, and VICReg [46] representing the more recent noncontrastive family. We measure linear finetuning performance which is a direct measurement of the quality of SSL representations (only a linear layer is trained; the feature extractor weights learned using SSL are unchanged).

Implementation: For SSL pretraining and linear evaluation, we use the Adam [51] optimizer. For finetuning and SL, we use the NovoGrad [52] optimizer which we found resulted in improved training stability. For all training protocols, we use learning rate schedule consisting of linear warmup to 1e-4 over ten epochs followed by a cosine annealing schedule over the remaining epochs. This schedule was chosen based on precedent in SSL literature (e.g., Barlow Twins [50], VICReg [46]), and the base rate of 1e-4 was selected empirically using a hyperparameter search. Self-supervised pretraining was carried out for 200 epochs, while finetuning or fully SL converged quickly and was only done for 50 epochs. For VICReg pretraining, we use variance, invariance, and covariance loss weights of 25, 25, 1, respectively, following the original article [46]. In all experiments, a random subset of the training set was used for cross-validation. Following training, we restore the weights from the epoch during which the best AUROC for the validation set was recorded, and measure the performance on the test set. Each experiment is run 16 times with different random model initializations and random train/validation splits; mean and standard deviation of performance across runs is reported.

V. Results and Discussion

A. Comparing SSL to SL in Performance

Table II summarizes the results of first experiment, i.e., Experiment a), the comparison of SSL to SL using the quantitative metrics. Row 1 is the Evidential deep learning (EDL) + coteaching model of [28] trained and tested on UVA dataset. The remaining part of table is divided into two parts; first the experiments trained and tested on UVA, and second the experiments trained and tested on CRCEO. In each of these two parts, we compared the performance of a fully supervised experiment in one row to VICReg methods in four rows with different training datasets, and finetuning protocols, i.e., datasets with different sizes $\mathcal{D}_{\text{needle}}$ or $\mathcal{D}_{\text{prostate}}$, and protocols for freezing feature extractor weights, Linear or Semi-sup.

Comparing the experiments with UVA as the train and test (TableII first part including EDL + coteaching) dataset shows that the only significant difference between the EDL + coteaching model and SL model is that the former had superior patch-wise performance. Overall, comparing the best SSL versus SL models, we see an improvement in core-wise AUROC (+3%), core-wise Avg-Prec (+3.9%, not in the table), ACC-B (+3.9%), patch-wise AUROC (+5.5%), and patch-wise Avg-Prec (+7.3%, not in the table). The only metric that does not significantly improve is Patch-ACC-B, where the EDL + coteaching and SSL models are comparable. Improvements moving from baseline SL to SSL are more pronounced when considering patch-wise versus core-wise metrics, so the SSL model is likely better at localizing cancer precisely. Similarly, comparing the best SSL versus SL models using CRCEO as the train and test dataset suggests significant

TABLE II
COMPARISON OF SUPERVISED VERSUS SELF-SUPERVISED MODELS ON PCA DETECTION

Method	Pretrain	Train&Test	Finetuning	AUROC	ACC-B	Patch-AUROC	Patch-ACC-B
EDL+Co-teaching [28]	None	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	N/A	87.76±1.8	77.79±4.2	-	71.25±1.2
Supervised	None	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	N/A	87.87±2.1	76.17±6.1	74.39±1.6	66.76±1.4
VICReg	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	Linear	89.83±1.3	80.73±2.8	79.43±0.8	70.71±1.2
			Semi-sup	89.30±3.0	77.94±9.7	78.23±2.5	68.31±3.7
	$\mathcal{D}^{\text{UVA}}_{\text{prostate}}$		Linear	**90.99±3.2**	**81.66±3.6**	**79.90±2.9**	**71.44±2.4**
			Semi-sup	90.60±2.4	80.91±4.9	79.49±1.6	70.79±2.0
Supervised	None	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	N/A	67.73 ±3.6	53.96±6.2	59.21±1.8	53.33 ±2.0
VICReg	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	Linear	77.36±5.7	**69.84±4.0**	66.26±3.1	60.30±2.3
			Semi-sup	**80.72±4.7**	66.90±7.0	**67.52±2.6**	**61.01±2.2**
	$\mathcal{D}^{\text{CRCEO}}_{\text{prostate}}$		Linear	77.42±3.4	74.38±4.2	66.56±1.8	60.65±1.2
			Semi-sup	79.62±4.7	67.95±6.4	67.52±2.6	60.86±2.8

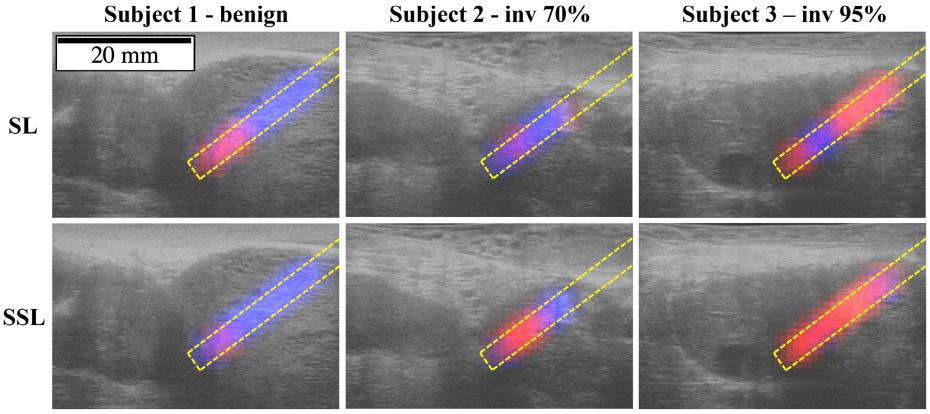


Fig. 3. Demonstration of cancer detection models. For each ultrasound image, the model predicts whether patches in the intersection of prostate and needle trace region (yellow dashed line) are benign (blue) or malignant (red). Three subjects are shown demonstrating model performance on benign (Subject 1, left), malignant with 70% involvement (Subject 2, middle) and malignant with 95% involvement (Subject 3, right) tissue. We demonstrate two models, namely the fully SL model and SSL model when all layers are finetuned. Note that for all subjects the SSL model makes fewer incorrect prediction, and the ratio of malignant (red) to benign (blue) using the SSL model better correlates with the pathology reported involvement of cancer in each biopsy core.

improvements of +13% in core-wise and +8% in patch-wise AUROC, showing the consistent improvement of SSL over SL models.

Including more unlabeled data in SSL pretraining ($\mathcal{D}_{\text{prostate}}$ versus $\mathcal{D}_{\text{needle}}$) resulted in consistent performance improvements in both parts except for CRCEO as the train and test dataset when using semi-supervised finetuning. This suggests that performance may improve further if even more unlabeled data were added, strengthening the case for SSL as a method to address data scarcity. Further experiments with larger volumes of data would be needed to confirm this. It is important to note that pretraining on the same amount of data is also a factor in improving performance, as seen when comparing SL to SSL pretrained on $\mathcal{D}_{\text{needle}}$.

To qualitatively study the performance of these models, we selected three biopsy cores and computed the heatmaps (Fig. 3) as explained in Section III-D. These correspond to a benign, malignant (70% involvement), and malignant (95%) moving from left to right. We compare predictions of SL and SSL models. Across all three patients, we see that the area of "cancer" predictions compared to the total needle region is closer to the true involvement for the SSL model, agreeing with the quantitative measurements of improved performance.

B. Comparing SSL to SL in Transfer Learning

Table III presents a summary of the quantitative evaluation results obtained from the transfer learning experiment, i.e., Experiment b). The table is divided into two parts, each of which includes experiments with the same train and test dataset, but different pretraining. The first row of both parts shows the results of the baseline supervised model, which was trained from scratch on CRCEO and UVA without any pretraining, for comparison purposes. The second and third rows of the supervised methods show the models that were pretrained using SL on UVA/CRCEO and finetuned on CRCEO/UVA, respectively. It should be noted that SL pretraining is equivalent to supervised training on a dataset before re-tuning the weights on the second dataset.

When comparing the performance of the supervised models with and without pretraining, we observe inconsistencies in their performance improvement. Specifically, for CRCEO as the train and test dataset, the performance improves with pretraining, while it deteriorates when UVA is used as the train and test dataset. This indicates that the transfer learning capability of the SL method is inconsistent, as some features are transferable and can improve the performance, while others learned using a different dataset do not transfer.

TABLE III
COMPARISON OF TRANSFER LEARNING PERFORMANCE FOLLOWING EITHER SUPERVISED PRETRAINING, SELF-SUPERVISED PRETRAINING (VICREG), OR NO PRETRAINING (BASELINE)

Method	Pretrain	Train&Test	Finetuning	AUROC	Avg-Prec	Patch-AUROC	Patch-Avg-Prec
Supervised	None	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	N/A	87.87±2.1	86.83±1.6	74.39±1.5	71.03±1.8
	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	Linear	81.5 ±2.0	79.9 ±7.6	69.9 ±0.9	67.3 ±4.1
			Semi-sup	77.7 ±9.1	75.6 ±5.9	69.4 ±4.0	67.2 ±0.1
VICReg	$\mathcal{D}^{\text{CRCEO}}_{\text{prostate}}$	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	Linear	88.05±3.0	-	74.99±1.8	-
			Semi-sup	**90.30±1.4**	-	**77.24±1.9**	-
Supervised	None	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	N/A	67.73±3.6	62.67±3.5	59.41±1.3	57.18±1.8
	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	Linear	70.44±5.1	64.95±3.6	62.33±1.9	59.92±1.6
			Semi-sup	71.84±4.6	66.43±4.9	62.67±2.0	59.23±2.1
VICReg	$\mathcal{D}^{\text{UVA}}_{\text{prostate}}$	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	Linear	70.35±3.0	65.81±2.6	65.64±2.1	62.30±2.3
			Semi-sup	**75.33±5.5**	**69.40±4.3**	**67.12±3.2**	**64.65±3.3**

However, we see a consistent and considerable performance increase when adopting SSL comparing to supervised methods with pretraining or without it. This suggests that self-supervised pretraining is an effective method to learn transferable features. Of the two VICReg methods in Table III, the ones using semi-supervised finetuning are significantly better. In summary, self-supervised training improves performance compared to no pretraining (+7.6% core-wise AUROC, +7.7% patch-wise AUROC) and compared to supervised pretraining (+3.5% core-wise AUROC, +4.5% patch-wise AUROC) when CRCEO is used as the train and test dataset. When UVA is used as the train and test dataset, self-supervised training is consistent and still improves the performance compared to no pretraining (+2.4% core-wise AUROC, +2.9% patch-wise AUROC) and compared to supervised pretraining (+8.8% core-wise AUROC, +7.3% patch-wise AUROC).

As self-supervised pretraining strongly benefited performance on both datasets compared to training from scratch, this confirms that transfer learning using other micro-ultrasound datasets is very beneficial for PCa detection. SSL is better than SL for transfer learning, but this is considerable when the feature extractor backbone is finetuned on the new data (semi-supervised finetuning). We reason that SSL improves transfer performance by giving good *initializations* of the feature extractor weights rather than producing a feature extractor which transfers directly; at least some re-tuning of these weights is needed to optimize performance.

C. Comparing SSL to SL in Distribution Shift Robustness

Table IV summarizes quantitative results for the distribution shift experiment, i.e., Experiment c. We analyzed the robustness to distribution shift by training and testing on two different datasets, and divided the experiments into two parts; in the first part, the CRCEO dataset is used for pretraining and/or training and the UVA dataset for testing, and in the second, the UVA dataset is used for pretraining and/or training and the CRCEO dataset for testing.

In both parts, an extra supervised experiment using the same train and test dataset is placed in the table to form the baseline (referred as in-distribution test set). The baselines are expected to show that there exists distribution shift between UVA and CRCEO datasets, and this shift deteriorates the performance when OOD test set is used. However, in the first part (CRCEO as test), we observe the opposite; the supervised experiment with in-distribution test set underperforms the supervised with OOD test.

Nonetheless, the best performance is achieved by SSL experiments (VICReg experiments) in both parts. Comparing the best SSL to supervised training with OOD CRCEO test dataset, we see an improvement in core-wise AUROC (+8.7%) and patch-wise AUROC (+2.5%). When UVA is used as test dataset, the core-wise AUROC improves by +8.9% and patch-wise AUROC by +6.9%.

We note that, while the distribution of cancer grades, involvement, PSA values and patient age, were similar between the two centers, distribution shifts in the imaging features are present, as also previously reported by [27]. We performed an experiment where a classifier was trained to identify the center a given ultrasound patch was from, with an accuracy of 67% on a center-balanced test set.

To explain the consistently better performance of models on the UVA compared to CRCEO test dataset across all experiments, is it likely that the UVA training dataset has more representative features, and that the UVA test dataset is "easier." We observed variability in the performance outcomes for different training and validation splits, but that the variance and mean of the validation performance was similar between the two centers. This variability should be addressed in future work by using larger datasets for testing.

D. Comparing Contemporary SSL Methods

Table V summarizes the comparison between the various SSL methods tested. We found that methods rank as follows: VICReg, SimCLR, BYOL. VICReg led by a significant margin in terms of performance, and required relatively little tuning of hyperparameters to achieve stable training and good results. However, SimCLR still performs fairly well, but required a large batch size. BYOL does not appear to work at all for RF data, although this finding is limited in that we did not perform an exhaustive hyperparameter search.

We speculate that the improved performance of VICReg may be due to the noncontrastive nature of the algorithm.

TABLE IV
COMPARISON OF SL AND SSL ON DISTRIBUTION SHIFT

Method	In-dist-test	Pretrain	Train	Test	Finetuning	AUROC	Patch-AUROC
Supervised	Yes	None	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	N/A	87.87±2.1	74.39±1.6
	No		$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$			81.51±5.7	67.41±1.8
VICReg	No	$\mathcal{D}^{\text{CRCEO}}_{\text{prostate}}$	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	Linear	**90.37±2.7**	74.31±1.9
					Semi-sup	88.57±2.8	**75.48±2.9**
Supervised	Yes	None	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	N/A	67.73±3.6	59.41±1.3
	No		$\mathcal{D}^{\text{UVA}}_{\text{needle}}$			72.54±4.91	66.94±1.39
VICReg	No	$\mathcal{D}^{\text{UVA}}_{\text{prostate}}$	$\mathcal{D}^{\text{UVA}}_{\text{needle}}$	$\mathcal{D}^{\text{CRCEO}}_{\text{needle}}$	Linear	73.71±3.0	64.61±1.6
					Semi-sup	**81.23±4.3**	**69.45±1.4**

TABLE V
COMPARISON OF LINEAR EVALUATION PERFORMANCE FOR DIFFERENT SSL METHODS ON ULTRASOUND RF DATA. ALL METHODS ARE PRETRAINED ON $\mathcal{D}^{\text{UVA}}_{\text{PROSTATE}}$ AND FURTHER EVALUATED BY TRAINING ON $\mathcal{D}^{\text{UVA}}_{\text{NEEDLE}}$

Method	AUROC	Avg-Prec	Patch-AUROC	Patch-Avg-Prec
VICReg	**90.99±3.2**	**90.74±3.6**	**79.90±2.9**	**78.78±3.0**
SimCLR	82.90±3.0	81.23±3.0	73.52±3.1	70.82±2.8
BYOL	51.25±5.8	52.31±3.8	50.75±2.8	51.49±2.4

In SSL, contrastive methods (e.g., SimCLR [45]) explicitly push feature representations of different input patches apart, which may rely on the unrealistic assumption that no instances in the same batch represent the same tissue type (and so *should* have similar representations). Noncontrastive methods may be better because they do not use this assumption, although the benefits observed may be due to another property of the VICReg algorithm.

E. Limitation of Micro-Ultrasound Depth

While our study used micro-ultrasound imaging with a maximum depth of 28 mm, it is worth noting that micro-ultrasound imaging can provide imaging depths of up to 5–6 cm in some cases. This is still less than the typical 8–12 cm depth achieved by conventional ultrasound. Nonetheless, the limited depth of micro-ultrasound imaging in our study may have excluded certain areas of the prostate, particularly in the transition zone, where tissue types can be more variable. However, the higher resolution of micro-ultrasound imaging allows for better visualization and characterization of smaller structures and finer tissue details, which can aid in the detection of clinically significant PCa. Overall, the use of micro-ultrasound imaging with its improved resolution and raw data source has its own unique advantages for PCa detection, but it should be interpreted in light of its limitations, including its shallower depth compared to conventional ultrasound.

VI. CONCLUSION

We carried out a multicenter study on PCa detection using self-supervised representation learning for RF micro-ultrasound data. We argued that SSL is a promising approach to address several characteristic challenges associated with this dataset, including weak labeling, high data heterogeneity, distribution shift, and data scarcity. We showed strong empirical evidence that SSL is beneficial for training models that can detect PCa and whose knowledge can be effectively transferred between data centers. We showed that recently proposed SSL methods outperform older methods on our data, giving hope for continued improvements in the future. Future work should focus on two directions: First, to pretrain larger models with data from large, multicenter datasets which may further improve PCa detection, and second to apply other techniques for handling weak labeling and uncertainty in combination with SSL for a unified learning approach.

REFERENCES

[1] L. Smith, S. Bryan, and P. De, "Canadian cancer statistics advisory committee. Canadian cancer statistics 2018," Can. Cancer Statist., Canada, Special Rep., 2018.

[2] M. Smeenge, J. J. M. C. H. de la Rosette, and H. Wijkstra, "Current status of transrectal ultrasound techniques in prostate cancer," *Current Opinion Urol.*, vol. 22, no. 4, pp. 297–302, 2012.

[3] H. U. Ahmed et al., "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study," *Lancet*, vol. 389, no. 10071, pp. 815–822, Feb. 2017.

[4] M. L. Oelze, "Quantitative ultrasound techniques and improvements to diagnostic ultrasonic imaging," in *Proc. IEEE Int. Ultrason. Symp.*, Oct. 2012, pp. 232–239.

[5] M. L. Oelze and J. Mamou, "Review of quantitative ultrasound: Envelope statistics and backscatter coefficient imaging and contributions to diagnostic ultrasound," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 63, no. 2, pp. 336–351, Feb. 2016.

[6] E. J. Feleppa, M. J. Rondeau, P. Lee, and C. R. Porter, "Prostate-cancer imaging using machine-learning classifiers: Potential value for guiding biopsies, targeting therapy, and monitoring treatment," in *Proc. IEEE Int. Ultrason. Symp.*, Sep. 2009, pp. 527–529.

[7] E. J. Feleppa, "Ultrasonic tissue-type imaging of the prostate: Implications for biopsy and treatment guidance," *Cancer Biomarkers*, vol. 4, nos. 4–5, pp. 201–212, Nov. 2008.

[8] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[9] E. D. Nelson, C. B. Slotoroff, L. G. Gomella, and E. J. Halpern, "Targeted biopsy of the prostate: The impact of color Doppler imaging and elastography on prostate cancer detection and Gleason score," *Urology*, vol. 70, no. 6, pp. 1136–1140, Dec. 2007.

[10] S. Khanduri et al., "Evaluation of prostatic lesions by transrectal ultrasound, color Doppler, and the histopathological correlation," *Cureus*, vol. 9, no. 7, pp. 1–8, Jul. 2017.

[11] G. Salomon et al., "Evaluation of prostate cancer detection with ultrasound real-time elastography: A comparison with step section pathological analysis after radical prostatectomy," *Eur. Urol.*, vol. 54, no. 6, pp. 1354–1362, Dec. 2008.

[12] T. Abrar Aleef et al., "Quasi-real time multi-frequency 3D shear wave absolute vibro-elastography (S-WAVE) system for prostate," 2022, *arXiv:2205.04038*.

[13] M. Moradi, P. Abolmaesumi, D. R. Siemens, E. E. Sauerbrei, A. H. Boag, and P. Mousavi, "Augmenting detection of prostate cancer in transrectal ultrasound images using SVM and RF time series," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 9, pp. 2214–2224, Sep. 2009.

[14] S. Azizi et al., "Detection and grading of prostate cancer using temporal enhanced ultrasound: Combining deep neural networks and tissue mimicking simulations," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 8, pp. 1293–1305, Aug. 2017.

[15] S. Azizi et al., "Deep recurrent neural networks for prostate cancer detection: Analysis of temporal enhanced ultrasound," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2695–2703, Dec. 2018.

[16] R. R. Wildeboer, R. J. G. van Sloun, P. Huang, H. Wijkstra, and M. Mischi, "3-D multi-parametric contrast-enhanced ultrasound for the prediction of prostate cancer," *Ultrasound Med. Biol.*, vol. 45, no. 10, pp. 2713–2724, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030156291930225X

[17] C. K. Mannaerts et al., "Detection of clinically significant prostate cancer in biopsy-naive men: Direct comparison of systematic biopsy, multiparametric MRI-and contrast-ultrasound-dispersion imaging-targeted biopsy," *BJU Int.*, vol. 126, no. 4, pp. 481–493, 2020.

[18] R. R. Wildeboer et al., "Automated multiparametric localization of prostate cancer based on B-mode, shear-wave elastography, and contrast-enhanced ultrasound radiomics," *Eur. Radiol.*, vol. 30, no. 2, pp. 806–815, Feb. 2020.

[19] M. M. Siddiqui et al., "Comparison of MR/ultrasound fusion–guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer," *J. Amer. Med. Assoc.*, vol. 313, no. 4, pp. 390–397, 2015.

[20] B. P. Rai, C. Mayerhofer, B. K. Somani, P. Kallidonis, U. Nagele, and T. Tokas, "Magnetic resonance imaging/ultrasound fusion-guided transperineal versus magnetic resonance imaging/ultrasound fusion-guided transrectal prostate biopsy—A systematic review," *Eur. Urol. Oncol.*, vol. 4, no. 6, pp. 904–913, Dec. 2021.

[21] M. M. Siddiqui et al., "Magnetic resonance imaging/ultrasound–fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy," *Eur. Urol.*, vol. 64, no. 5, pp. 713–719, 2013.

[22] C. M. L. Klotz, "Can high resolution micro-ultrasound replace MRI in the diagnosis of prostate cancer?" *Eur. Urol. Focus*, vol. 6, no. 2, pp. 419–423, Mar. 2020.

[23] R. Abouassaly, E. A. Klein, A. El-Shefai, and A. Stephenson, "Impact of using 29 MHz high-resolution micro-ultrasound in real-time targeting of transrectal prostate biopsies: Initial experience," *World J. Urol.*, vol. 38, no. 5, pp. 1201–1206, May 2020.

[24] S. Ghai et al., "Assessing cancer risk on novel 29 MHz micro-ultrasound images of the prostate: Creation of the micro-ultrasound protocol for prostate risk identification," *J. Urol.*, vol. 196, no. 2, pp. 562–569, Aug. 2016.

[25] P. Sountoulides et al., "Micro-ultrasound–guided vs multiparametric magnetic resonance imaging-targeted biopsy in the detection of prostate cancer: A systematic review and meta-analysis," *J. Urol.*, vol. 205, no. 5, pp. 1254–1262, May 2021.

[26] D. Rohrbach, B. Wodlinger, J. Wen, J. Mamou, and E. Feleppa, "High-frequency quantitative ultrasound for imaging prostate cancer using a novel micro-ultrasound scanner," *Ultrasound Med. Biol.*, vol. 44, no. 7, pp. 1341–1354, Jul. 2018.

[27] Y. Shao, J. Wang, B. Wodlinger, and S. E. Salcudean, "Improving prostate cancer (PCa) classification performance by using three-player minimax game to reduce data source heterogeneity," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3148–3158, Oct. 2020.

[28] M. Gilany et al., "Towards confident detection of prostate cancer using high resolution micro-ultrasound," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2022, pp. 411–420.

[29] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[30] M. Blaivas, L. N. Blaivas, and J. W. Tsung, "Deep learning pitfall: Impact of novel ultrasound equipment introduction on algorithm performance and the realities of domain adaptation," *J. Ultrasound Med.*, vol. 41, no. 4, pp. 855–863, Apr. 2022.

[31] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do CIFAR-10 classifiers generalize to CIFAR-10?" 2018, *arXiv:1806.00451*.

[32] H. Zou, X. Gong, J. Luo, and T. Li, "A robust breast ultrasound segmentation method under noisy annotations," *Comput. Methods Programs Biomed.*, vol. 209, Sep. 2021, Art. no. 106327.

[33] G. Javadi et al., "Multiple instance learning combined with label invariant synthetic data for guiding systematic prostate biopsy: A feasibility study," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 6, pp. 1023–1031, Jun. 2020.

[34] G. Javadi et al., "Training deep networks for prostate cancer diagnosis using coarse histopathological labels," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2021, pp. 680–689.

[35] F. Fooladgar et al., "Uncertainty-aware deep ensemble model for targeted ultrasound-guided prostate biopsy," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[36] P. Mojabi, V. Khoshdel, and J. Lovetri, "Tissue-type classification with uncertainty quantification of microwave and ultrasound breast imaging: A deep learning approach," *IEEE Access*, vol. 8, pp. 182092–182104, 2020.

[37] Z. Zhang, Y. Li, W. Wu, H. Chen, L. Cheng, and S. Wang, "Tumor detection using deep learning method in automated breast ultrasound," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102677.

[38] G. Ayana, K. Dese, and S.-W. Choe, "Transfer learning in breast cancer diagnoses via ultrasound imaging," *Cancers*, vol. 13, no. 4, p. 738, Feb. 2021.

[39] B. Behboodi and H. Rivaz, "Ultrasound segmentation using U-Net: Learning from simulated data and testing on real data," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6628–6631.

[40] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[41] H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma, "Self-supervised learning is more robust to dataset imbalance," 2021, *arXiv:2110.05025*.

[42] Y. Zhong, H. Tang, J. Chen, J. Peng, and Y.-X. Wang, "Is self-supervised learning more robust than supervised learning?" 2022, *arXiv:2206.05259*.

[43] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5410–5419.

[44] S. Azizi et al., "Robust and efficient medical imaging with self-supervision," 2022, *arXiv:2205.09723*.

[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[46] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," 2021, *arXiv:2105.04906*.

[47] M. Tirindelli, C. Eilers, W. Simson, M. Paschali, M. F. Azampour, and N. Navab, "Rethinking ultrasound augmentation: A physics-inspired approach," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2021, pp. 690–700.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[49] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[50] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[52] B. Ginsburg et al., "Stochastic gradient methods with layer-wise adaptive moments for training of deep networks," 2019, *arXiv:1905.11286*.