Semi-supervised Learning from Coarse Histopathology Labels

Fahimeh Fooladgar[a], Minh Nguyen Nhat To[a], Golara Javadi[a], Samira Sojoudi[a], Walid Eshumani[b], Antonio Hurtado[b], Silvia Chang[b], Peter Black[b], Parvin Mousavi[c], and Purang Abolmaesumi[a]

[a]Department of Electrical and Computer Engineering, University of British Columbia, BC, Canada;
[b] Vancouver General Hospital, Vancouver, BC, Canada;
[c] School of Computing, Queen's University, Kingston, ON, Canada

ARTICLE HISTORY
Compiled September 15, 2022

ABSTRACT
Ultrasound imaging is commonly used to guide sampling the prostate tissue in transrectal biopsies, followed by detection of cancer through histopathological analysis and coarse labeling (cancer or benign) of sampled tissue. The biopsy procedure is currently blind to the underlying tissue microstructure. Ideally, the procedure should be improved by developing machine learning solutions that can identify the presence of cancer in ultrasound images to guide the biopsy procedure. Training a fully supervised learning model for ultrasound using coarse histopathology labels suffers from weakly annotated data. Weak labels of biopsy samples also introduce label noise for each image pixel, making the training process even more challenging, as deep learning models are prone to overfit to noisy samples which negatively affects their generalization. To address this challenge, we propose a semi-supervised framework for learning with noisy labels. To prepare the labeled and unlabeled sets from weakly annotated data, we leverage a two-component mixture model to cluster the training data into clean and noisy label samples based on their loss values. To prevent the model from overfitting and memorizing noisy samples, we propose two models trained in a co-teaching pipeline. Then, during the semi-supervised training phase, we utilize the well-known MixMatch algorithm which incorporates consistency regularization, entropy minimization, and the Mixup regularization as well as the cross-entropy loss function for unlabeled and labeled sets, respectively. We evaluate the proposed framework with prostate ultrasound data obtained from 71 subjects, while sampling 264 biopsy cores. We achieve balanced accuracy, sensitivity, and specificity of 78.6%, 80.0%, and 77.1%, respectively. In a detailed comparison study, we demonstrate that our proposed framework outperforms the fully supervised method with state-of-the-art robust loss functions to deal with label noise.

KEYWORDS
Noisy labels; Semi-supervised learning; Prostate cancer detection

1. Introduction

Prostate cancer (PCa) is the second most commonly diagnosed malignancy in men (Rawla (2019)). Regardless, the risk of mortality imposed by PCa can be avoided by proper treatment and management that follow the timely and accurate screening of

CONTACT Fahimeh Fooladgar. Email: fahimeh.fooladgar@ubc.ca

cancer. It is, hence, critical that a diagnosis procedure is capable of detecting clinically significant PCa. Systematic transrectal ultrasound (TRUS)-guided prostate biopsy is currently regarded as the gold standard for PCa diagnosis, where prostate tissues are systematically sampled from a standard 8 to 12-core template. This conventional approach is limited because clinically significant PCa far from the sampled locations can be easily missed or under-diagnosed (Serefoglu et al. (2013)). Multi-parametric magnetic resonance imaging (mp-MRI) has recently evolved as a promising technique for PCa assessment. Despite the growing progress of mp-MRI, its wide adoption for image-guided biopsy is still challenged by the added cost, screening time, and demanding domain-expertise (Johnson et al. (2019)). In this regard, TRUS-guided biopsy is still a more rapid, accessible, and cost-effective method, especially in hospitals with limited resources (Moe and Hayne (2020)).

The advantages of TRUS-guided biopsy can be leveraged by incorporating an automatic diagnosis system capable of identifying high-risk cancer lesions. Among various methods proposed for identifying PCa in ultrasound, supervised deep learning using a sequence of ultrasound images has shown promising results (Sedghi et al. (2020); Javadi et al. (2020)). In these methods, training labels are obtained from histopathology reports of biopsy samples, which include the percentage of cancer in cores, known as "involvement". These labels capture statistical representations of cancer distribution rather than the exact distribution of cancer, leading to weak supervision at a pixel level. As the consequence, a large number of labels are wrongly assigned to individual ultrasound pixels (i.e. noisy labels), which negatively affects the PCa classification performance.

In machine learning literature, the problem of reducing the influence of label noise is known to be challenging as deep learning models can easily overfit the noisy labels (Liu et al. (2020)). Existing techniques for learning with noisy labels can be divided into two categories. The first category focuses on developing cost functions that have a high tolerance for noisy labels. Recent examples of these methods include asymmetric loss functions (Zhou et al. (2021)), which take advantage of clean-labels-domination-assumption to eliminate the effect of label noise, and early-learning regularization (Liu et al. (2020)), which establishes a regularization term to prevent the memorization of false labels in the late training phase. The second group designs sophisticated training strategies to identify and correct or clean label noise. For instance, co-teaching (Han et al. (2018)) proposes a robust training approach that considers two models with different initializations and uses the results of each model to update the other. The Mixup data augmentation (Zhang et al. (2018)) is another way of providing robustness against label noise through the combination of clean and noisy samples (both inputs and labels) to compute a more representative loss to guide the training process. Most recent SSL methods use pseudo-labeling (Sohn et al. (2020); Lee et al. (2013)) and consistency regularization (Berthelot et al. (2019); Tarvainen and Valpola (2017); Sohn et al. (2020)) to hypothesize labels for unlabeled samples. In order to create this label hypothesis, they use the model's predicted output after randomly perturbing the input sample. The FixMatch algorithm (Sohn et al. (2020)) generates weak and strong augmentations for each unlabeled input; the prediction of weak augmented input is then used as a target when predicting strongly augmented inputs along with a cross-entropy loss function. The authors of Mean Teacher algorithm (Tarvainen and Valpola (2017)), take an exponential moving average (EMA) of model parameters to guess labels for two weakly augmented versions of each unlabeled sample. Mean square error (MSE) is used for consistency regularization as a loss function. The method DivideMix (Li et al. (2020)) proposes semi-supervised learning to leverage the discarded

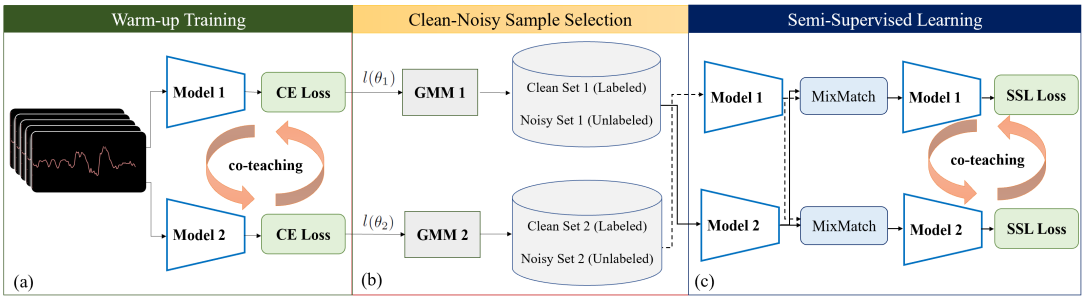


Figure 1. An overview of the proposed method, (a) Step 1: warm-up two models by training on all data using the standard cross-entropy loss function for a few epoch. (b) Step 2: in each epoch after warm-up training, apply GMM on the loss values to divide the whole training set into clean and noisy sets. (c) Step 3: each model is trained in semi-supervised manner using MixMatch algorithm (Berthelot et al. (2019)) in every mini-batch on labeled-unlabeled sets generated by another model.

noisy labels as unlabeled data.

These methods have achieved outstanding results under simulated or real-world noise situations that are totally different from scenarios involving noisy labels in the medical domain. In particular, label noise is tougher to deal with in the field of medical image analysis where datasets are often small and labeling requires domain expertise. More recently, Javadi et al. (2022b) and To et al. (2022) propose label refinement algorithms to deal with label noise during training. (Javadi et al. (2021)) proposed a supervised method for PCa detection to address the noisy labeling problem by formulating the loss function based on cancer distribution statistics. In another study, the authors of (Javadi et al. (2022a)) argued that the statistical distribution of cancer within the core could be considered as aleatoric uncertainty. Therefore, they proposed an uncertainty-aware method to estimate the uncertainty associated with the labels as well as the model. Nonetheless, both(Javadi et al. (2021) and Javadi et al. (2022a)) focused primarily on easy/clean (high cancer involvement) cores, and reduce the negative effects of data with hard/noisy (low cancer involvement) by an explicit weighting strategy. In this study, we empirically demonstrate that it is more effective to learn from the hard/noisy cores in an unsupervised manner as we completely avoid using excessively noisy labels for supervised training.

In this paper, we present a semi-supervised framework for learning with noisy labels (Semi-supervised-LNL) to classify pixels in prostate ultrasound images as benign or cancer using coarse, core-level labels produced from histopathology. With this training paradigm, we can leverage both clean and noisy labels by splitting the dataset into labeled and unlabeled sets, respectively. To create noisy and clean sets, initially, we design two models based on the co-teaching framework (Han et al. (2018)) and warm up these models for a small number of epochs on all training data in a supervised manner using Cross-Entropy (CE) loss function (Figure 1.a). Next, clean and noisy clusters are constructed by fitting two Gaussian Mixture Models (GMMs) to the per-sample loss distribution to form the labeled and unlabeled sets of our Semi-supervised-LNL method (Figure 1.b). Once we formed these steps, we leverage the well-known semi-supervised learning method, MixMatch algorithm (Berthelot et al. (2019)), to fine-tune the models on clean and noisy samples as labeled and unlabeled sets (Figure 1.c). The last two steps are applied iteratively throughout training until the models converge. By considering the noisy samples as unlabeled samples and training in a semi-supervised manner, our model learns the features from the whole training set while reducing the effect of mislabeled ground truth samples in the training process. As a result, it yields

great improvements in accuracy and provides clinicians with a more generalized model for computer-assisted cancer detection from ultrasound. To the best of our knowledge, this paper is the first study that incorporates semi-supervised learning to deal with noisy labels in pixel-wise PCa detection using ultrasound images. The workflow of our proposed semi-supervised learning method is illustrated in Figure 1. The main contributions are as follows:

- We propose SSL for pixel-wise detection of cancer in ultrasound images with only coarse (core-level) histopathology labels.
- We leverage two models and incorporate co-teaching to prevent overfitting to noisy label samples and prepare reasonable labeled/unlabeled sets for SSL step.
- We experimentally show that integrating noisy samples as unlabeled samples significantly improves the results of prostate cancer detection.
- We extensively evaluate the impact of the different components of our workflow. Specifically we use three different SSL algorithms as substitutes for our SSL building block and compare their performance in the proposed pipeline.

2. Materials and Method

2.1. Data

This study includes data from 71 patients who underwent systematic TRUS-guided prostate biopsy. We utilized Temporal enhanced UltraSound (TeUS) (Moradi et al. (2010)) as our input data, which has been shown to outperform the analysis of B-mode data in characterization of tissue in a number of studies (Imani et al. (2015); Bayat et al. (2017)). To extract the data from the biopsied locations, we first use the B-mode images to identify a rectangular region of size 2×18 mm (constrained by the biopsy needle geometry) corresponding to the location of the biopsy sample. Subsequently, we extract 200 consecutive samples as a time series, for each image pixel in the corresponding region from Radio Frequency (RF) ultrasound data. The number of time series contained within each core varies from approximately 4000 to 12000, depending on the depth and location of the region in the prostate. We visually inspect the data acquisition quality on B-mode and exclude biopsy cores if excessive hand motion before firing the needle is detected, or if the percentage of the cancer is smaller than 40% of the biopsy area. The remaining 264 biopsy cores are randomly split into training (33 patients: 90 benign, 62 cancer), validation (25 patients: 37 benign, 20 cancer), and test (13 patients: 35 benign, 20 cancer) sets, where sets have mutually exclusive patients. As a result, the total number of time-series signals in our dataset are 1133478, 451850, and 439579 for training, validation, and test sets, respectively. Biopsies are labeled as cancer or benign based on the histopathology results. Individual time-series within each core are given the same label as its coarse biopsy core label (Benign or Cancer), inherently introducing label noise in the pixel-wise annotation. Therefore, each time-series data and its label are considered as the input signal and output of our model, respectively. Figure 2 depicts the general pipeline of our preprocessing procedure.

2.2. Method

We aim to propose an architecture using semi-supervised learning to deal with noisy labels caused by the coarse histopathology annotation. Hence, we first need to divide

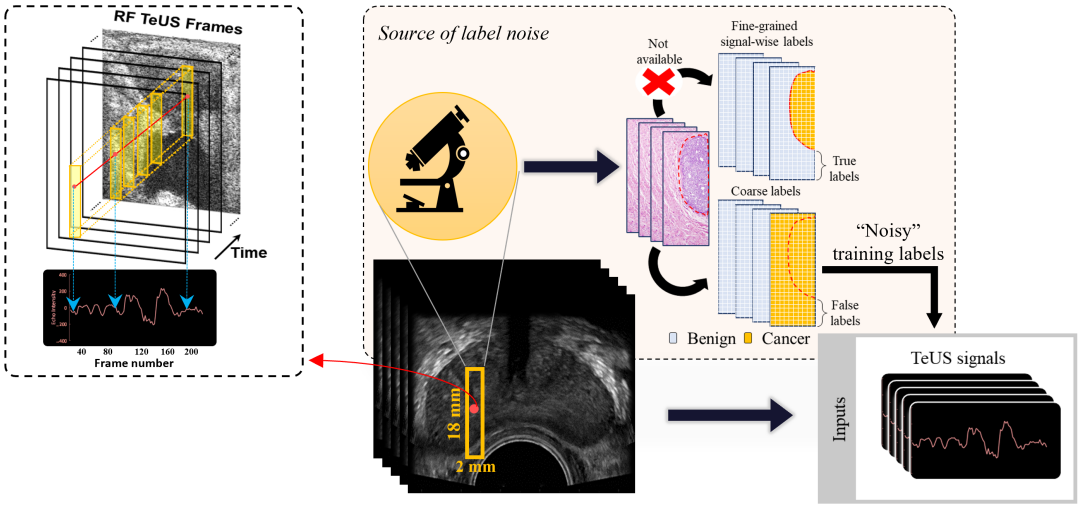


Figure 2. We use the B-mode images to identify a rectangular region of size 2×18 mm (constrained by the biopsy needle geometry) corresponding to the location of biopsy sample. Then, 200 consecutive time series per image pixel in the region from the corresponding RF data are extracted. The annotation is made based on the coarse biopsy core label (Benign or Cancer), which causes label noise in pixel-wise annotation.

the coarse annotated training data into labeled and unlabeled sets. To do so, we train two networks on the entire dataset for a few epochs. Then, we model the per-sample loss distributions of these networks by using two GMMs to form the two desired sets. Next, each network performs semi-supervised learning guided by the other network using both labeled and unlabeled samples. Summarily, our method consists of three main steps: 1) warm-up training, 2) clean and noisy sample selection, and 3) semi-supervised learning, which are explained in detail in the following subsections.

Problem Formulation: Let the whole training set be defined as $D = (\mathcal{X}; \mathcal{Y}) = \{(x_i, y_i)\}_{i=1,}^{N}$, where x_i is i-th sample (a single time-series within the biopsy core) and $y_i \in \{0, 1\}^C$ is its one-hot noisy label over two classes (Benign and Cancer) and N denotes the total number of time-series signals in our training set. We define the Softmax probabilities produced by the CNN model as $p(y_i|x_i; \theta_t)$, where θ_t is the parameters of classifier t and $t \in \{1, 2\}$ as we have two classifiers.

Step 1. Warm-up Training: In general, CNNs tend to learn simple patterns first before fitting label noise. Hence, to divide the training data into clean and noisy samples, we could leverage the value of the loss function as a criterion to find clean and noisy samples at the early steps of the training phase. In the presence of the high rate of label noise, the model quickly overfits to noisy samples leading to overconfident prediction on training data and generating low loss values for all training data. To overcome this challenge, we design two separate models that are randomly initialized and trained simultaneously in a co-teaching framework for some epochs to prevent the memorization of noisy samples. In this framework, each model updates itself by propagating the error selected by its peer model. This error is computed in every mini-batch, using the useful examples of its peer model. Such examples are chosen from training samples with lower loss values and it is controlled by the epoch number, where a higher number of samples are selected in earlier epochs. Therefore, the output of these two models prepares early predictions for the subsequent steps (Figure 1(a)).

Step 2. Clean and Noisy Sample Selection: After warm-up training of two models, we divide the samples into clean and suspected noisy labeled samples according to their loss values. We further involve the suspected noisy samples in unsupervised learning as unlabeled data points instead of totally discarding them from training process. The authors of (Arazo et al. (2019)) argue that by applying the two-component mixture model (e.g., Beta Mixture Model (BMM)) to the loss value of each sample during training, they can estimate the probability that a sample is mislabelled. Then, Li et al. (2020) shows that using GMM instead of BMM generates two sharper clusters to distinguish correctly versus incorrectly classified samples. Therefore, in the presence of label noise and before overfitting, we can hypothesize that the correctly classified samples are the clean label samples while the other ones are noisy labels or complex samples. As a result, we construct two GMMs and fit them to the loss values of classifier θ_1 and θ_2, where the per sample loss value for all N training samples is computed as follows:

$$l(\theta_1) = \Big\{ l_i \Big\}_{i=1}^{N} = \Big\{ -\sum_{c=1}^{C} y_i^c \log \; p^c(y_i|x_i;\theta_1) \Big\}_{i=1}^{N}, \tag{1}$$

where $p^c(y_i|x_i;\theta_1)$ is the Softmax probability output of the first model for class c. Henceforth, at each epoch, we divide the whole training data into clean and noisy samples which are then considered as $\mathcal{X}$ and unlabeled $\mathcal{U}$ sets, respectively. If each model is trained on $\mathcal{X}$ and $\mathcal{U}$ sets generated by itself, it possibly can be biased to its mislabeled outputs, hence the GMM generated from the prediction of the model θ_1 is used to split the training data for the second model θ_2 and vice versa. Finally, we provide two GMMs with two sets of $\mathcal{X}$ and $\mathcal{U}$ in each training epoch that builds the labeled-unlabeled training set of our semi-supervised learning step below (Figure 1(b)).

Step 3. Semi-supervised Learning: This step leverages the noisy label samples as unlabeled data by enforcing the model to predict low entropy values on the unlabeled data and consistent values on the perturbed input by using the holistic MixMatch algorithm (Berthelot et al. (2019)). It combines labeled and unlabeled data and guesses low-entropy labels for unlabeled augmented examples. In this method, the CE loss function is computed for the labeled set and the entropy minimization loss function is considered for the unlabeled ones. In every mini-batch of the learning process, we use one batch of labeled ($\mathcal{X}_b$) and unlabeled ($\mathcal{U}_b$) samples as the input to the MixMatch algorithm, which generates one augmented labeled batch ($\mathcal{X}'_b$) and one augmented unlabeled batch ($\mathcal{U}'_b$) with its guessed label for each of unlabeled sample. This algorithm has three stages as follows:

A. Data Augmentation: For each sample x_b in $\mathcal{X}_b$, we generate a transformed version $\hat{x}_b = Augment(x_b)$ by applying simple data augmentation. Besides, for each unlabeled sample u_b in $\mathcal{U}_b$, we generate K augmented versions $\hat{u}_{b,k} = Augment(u_b)$.

B. Label Guessing: For each unlabeled sample(u_b), we need to estimate a label q, hence we use the prediction of the corresponding model for K augmented versions to guess its label. Thus, the average of K predictions of the model has been utilized as a label for unlabeled sample (u_b), which can be computed by $\bar{q} = \frac{1}{K}\sum_{k=1}^{K} p(y|\hat{u}_{b,k},\theta)$.

C. Mixup: In this stage, for each augmented labeled sample pair $(\hat{x}_b, y_b)$, one random pair of samples in the mini-batch is selected, then the randomly Mixup method proposed in (Berthelot et al. (2019)) is applied to generate $(\hat{x}'_b, p'_b)$. This operation is repeated for all of the labeled and unlabeled augmented samples in every mini-batch and generated an augmented mixed up version determined by $\mathcal{X}'$ and $\mathcal{U}'$.

Consequently, the semi-supervised learning loss function is defined by combining simple CE loss on generated labeled set $\mathcal{X}'$, and L2 loss applied on unlabeled $\mathcal{U}'$ set as follows:

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda \mathcal{L}_{\mathcal{U}}, \tag{2}$$

$$\mathcal{L}_{\mathcal{X}} = -\frac{1}{|\mathcal{X}'|} \sum_{(x,p)\in\mathcal{X}'} \sum_{c} p_c \; log \; p^c(y|x;\theta), \tag{3}$$

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{|\mathcal{U}'|} \sum_{(u,q)\in\mathcal{U}'} ||q - p(y|x;\theta)||_2^2, \tag{4}$$

where λ is used to control the effect of the unsupervised loss function in the training process. In the semi-supervised learning phase of our training, we again leverage the effect of the co-teaching technique, where the two models teach each other based on the above loss function. Figure 1(c) illustrates the building blocks of the proposed semi-supervised learning step.

3. Experimental setup

We design a one-dimensional CNN model constructed as seven sequentially cascaded blocks, each consisting of one-dimensional depth-wise separable convolution followed by ReLU, batch normalization, and drop-out layers. The number of filters in convolution layers is 32, 32, 64, 64, 128, 128, and 256, respectively. Thereafter, three fully-connected layers are incorporated. The two first layers are followed by ReLU and drop-out, while a third layer using the Softmax activation function is considered the final classification layer. We use the same network architecture with random parameter initialization for our two models. The input and output of our models are individual signals within the core and their predicted label (Cancer or Benign), respectively. The predicted probability of labels for each biopsy core is computed by averaging the prediction of all signals within the core. All evaluations (and metrics) are then performed and reported for biopsy cores which are associated with pathology labels we assume to be free of label noise.

The proposed architecture is trained using Stochastic Gradient Descent (SGD) optimizer with a learning rate of $1e-1$ and a batch size of 2048. We train our two models in a supervised way for 10 epochs as a warm-up training, then both models are trained for 90 more epochs with an early stopping strategy. The hyper-parameter of the semi-supervised learning loss function, λ, is set to 25. The augmentation methods utilized in the SSL step of the MixMatch algorithm include adding noise, scaling, magnitude, time, and window warping, which are randomly applied for each time-series input.

Table 1. Comparison of the proposed Semi-supervised-LNL method with state-of-the-art robust loss functions in terms of balanced accuracy, sensitivity and specificity (%).

Method	ACC_B	$SEN.$	$SPE.$
Co-teaching (Han et al. (2018))	60.4	75.0	45.7
SCE (Wang et al. (2019))	60.7	70.0	51.4
GCE (Zhang and Sabuncu (2018))	61.8	75.0	48.5
NCE+RCE (Ma et al. (2020))	63.5	**90.0**	37.1
AGCE+NCE (Zhou et al. (2021))	65.0	**90.0**	40.0
Max-Inv. constraints (Javadi et al. (2021))	67.0	78.0	57.0
Semi-supervised-LNL (Ours)	**78.6**	80.0	**77.1**

For weak and strong augmentation in the FixMatch and Mean Teacher algorithms, we used the same augmentation sets but with a higher rate of change for strong ones.

To account for the imbalanced ratio between cancer and benign samples in our dataset, we choose the balanced accuracy (ACC_B), specificity ($SPE.$), and sensitivity ($SEN.$) as our evaluation metrics. We compare the proposed method with Max-Inv. constraints (Javadi et al. (2021)), a previous approach for prostate cancer detection that formulates the coarse labeling of histopathology as an optimization problem. They leveraged the cancer involvement of each core as auxiliary information, and designed a multi-constraint loss function to deal with label noise in a co-teaching framework. Besides, we compare our approach with state-of-the-art robust loss function methods presented in the literature to handle learning with noisy labels. We consider the Symmetric Cross Entropy (SCE) (Wang et al. (2019)), Generalized cross entropy (GCE) (Zhang and Sabuncu (2018)), Normalized loss functions (NCE+RCE) (Ma et al. (2020)), and Asymmetric loss functions (AGCE+NCE) (Zhou et al. (2021)), which were specially designed for training in the presence of extremely high label noise. To evaluate the effect of different SSL methods in the proposed pipeline, we implement two other SSL methods, Mean Teacher (Tarvainen and Valpola (2017)) and FixMatch (Sohn et al. (2020)), as a substitute for MixMatch algorithm and compare their performance in the proposed pipeline.

4. Results and Discussions

Table 1 shows the cancer detection performance of different methods for learning with noisy labels on our dataset. Overall, the model trained with the Semi-supervised-LNL ($ACC_B = 78.6$) outperforms other models by a large margin ($ACC_B \leq 67.0$). Interestingly, the results on sensitivity and specificity show that the improved classification accuracy of our method was largely contributed by a substantially higher specificity ($SPE. = 77.1$). In comparison, models trained with other methods suffer from a high rate of false positives ($SPE.$ $\leq$57.0). We speculate that the label noise in the cancer cores is much higher than in the benign ones since the coarse annotation blindly flips the true-benign labels of cores with lower involvements (e.g. 0.4 or 0.5) into cancer. Therefore, the less robust models may struggle to restrain from overfitting to those flipped labels, resulting in a tendency to predict cancer class for true-benign signals. This also indicates that the sensitivity of the less robust models could be exaggerated and may not truly reflect the classification performance. On the other hand, our model demonstrates higher tolerance to noisy labels and achieves better generalizability by effectively identifying samples with noisy labels and learning from them as unlabelled

Table 2. Comparison of state-of-the-art SSL methods in our proposed pipeline in terms of balanced accuracy, sensitivity and specificity (%).

SSL Method	ACC_B	$SEN.$	$SPE.$
Mean Teacher (Tarvainen and Valpola (2017))	65.7	**80.0**	51.4
FixMatch (Sohn et al. (2020))	72.5	65.0	**80.0**
MixMatch (Berthelot et al. (2019))	**78.6**	**80.0**	77.1

Table 3. Effect of each component of the proposed method on classification performance in terms of test accuracy (%).

Method	Mixup	ACC_B	$SEN.$	$SPE.$
Supervised-LNL	-	60.4	75.0	45.7
Semi-supervised-LNL	-	71.1	**85.0**	57.1
Semi-supervised-LNL	✓	**78.6**	80.0	**77.1**

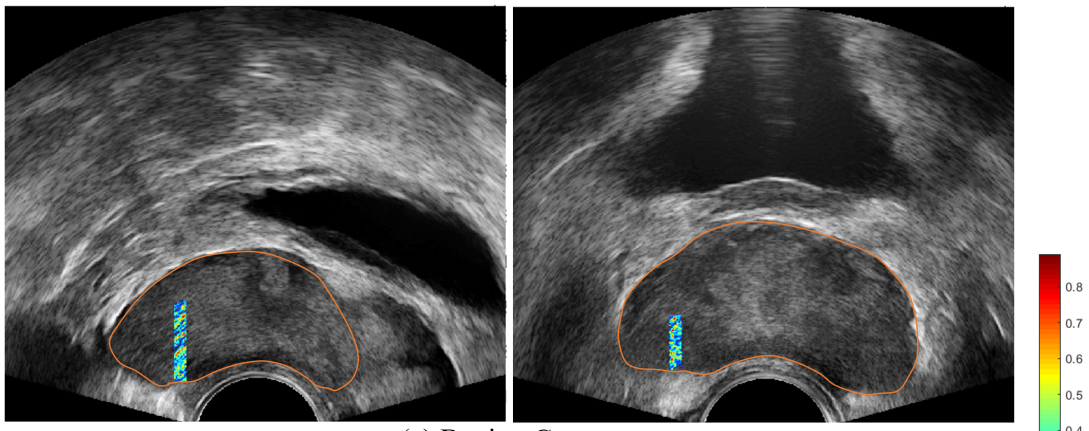
data.

Table 2 shows the performance of these three SSL methods. Overall, the model trained with the MixMatch algorithm ($ACC_B = 78.6$) outperforms the two other methods. As part of the unsupervised loss function, FixMatch algorithm needs weak and strong augmentations applied to the input data. Strong augmentations may cause challenges for our dataset, since it may change the time-series labels. In the computer vision, the rates of strong augmentation can be evaluated visually and their such augmentations do not result in changes in their labels (For example strongly augmented cat image is still a cat). In our case, strong augmentation may result in changes in the label of the time-series data. As a results, we are limited to the extent of strong augmentation we can applied. Hence, the performance of the FixMatch is not as good as the MixMatch. Furthermore, our experiment shows that EMA in Mean Teacher algorithm does not significantly affect the performance of the SSL pipeline, since the results suffer from a high rate of false positives ($SPE. = 51.4$).

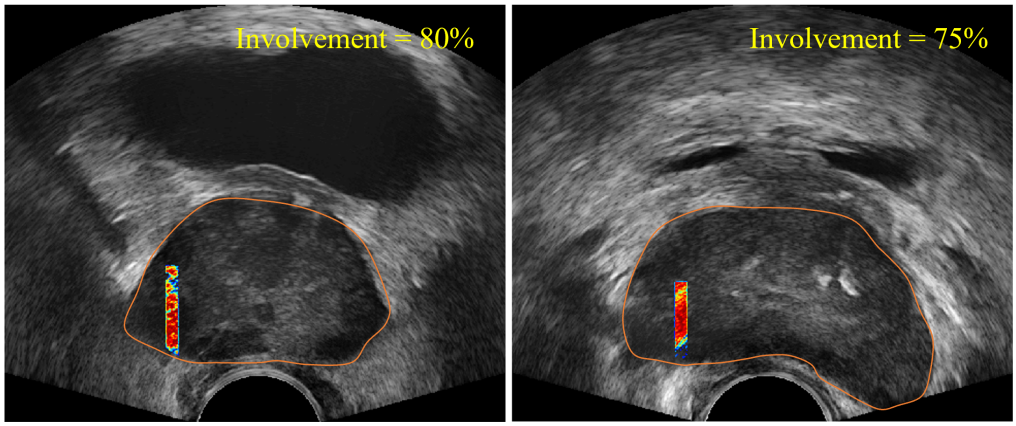
Figure 3 illustrates the cancer likelihood maps overlaid on B-mode ultrasound images for four biopsy cores with benign and cancer pathology, as well as cancer involvement in the cancerous cores. The blue colors depict a high probability of benign tissue, while red colors present a high probability of cancer.

Ablation Study: In Table 3, we demonstrate the benefits of different modules in our proposed method. First, we establish a baseline with co-teaching training ($ACC_B = 60.4$) without any further steps. Next, we show that by using GMM for the clean-noisy sample selection strategy, we can efficiently construct an unlabelled dataset from noisy samples that can be learned in an unsupervised manner with MixMatch. Since Mixup augmentations in MixMatch were originally proposed for image data, we evaluate the performance of semi-supervised learning with and without Mixup separately (78.6 and 71.1 ACC_B, respectively). The results show 10% improvement in balanced accuracy in the presence of Mixup, suggesting that this augmentation can also effectively leverage time series data. These results justify our design choices in building a model that can focus more on clean samples and learn general and specific patterns of the data instead of memorizing the features and overfitting to noisy examples.

Our study has some limitations. Firstly, the clinical evaluation of our approach is not present. Regardless, the application of the proposed method during the inference phase is straightforward, and the prediction can be easily made for the whole prostate rather than per core. Future studies will examine how the cancer likelihood maps generated



(a) Benign Core



(b) Cancer Core

Figure 3. Cancer likelihood maps overlaid on the prostate B-Mode ultrasound images. The red and blue colors indicate cancer and benign predictions, respectively. Involvement shows the gold-standard percentage of the cancerous area in the needle region based on histopathology.

by our model would affect the current systematic prostate biopsy. Secondly, our study does not compare the performance of MixMatch with other state-of-the-art methods in semi-supervised learning. Nevertheless, our work shows promising results in computer-aided diagnosis of prostate cancer in ultrasound with MixMatch. We also notice that our approach is versatile and can incorporate virtually any semi-supervised strategy. The upcoming studies will focus on how data from outside the biopsy core region can be effectively utilized as unlabelled data to improve cancer detection performance.

5. Conclusion

Training deep learning models on real clinical data can be challenging due to extremely high label noise rates. We demonstrate that it is possible to leverage semi-supervised learning along with a robust training technique, such as co-teaching, to handle the samples with noisy labels, while taking advantage of all data in the training process without discarding any noisy samples. We utilized an unsupervised GMM to construct the labeled and unlabeled sets of semi-supervised learning. We applied this method for prostate cancer detection in ultrasound imaging, where only coarse histopathology

labels are available. We compared the proposed framework against state-of-the-art methods designed to address learning with noisy labels. Our experiments demonstrated that the accuracy of the proposed training framework outperforms those robust loss functions in terms of balanced accuracy and specificity.

References

Arazo E, Ortego D, Albert P, O'Connor N, McGuinness K. 2019. Unsupervised label noise modeling and loss correction. In: International Conference on Machine Learning. PMLR. p. 312–321.

Bayat S, Azizi S, Daoud MI, Nir G, Imani F, Gerardo CD, Yan P, Tahmasebi A, Vignon F, Sojoudi S, et al. 2017. Investigation of physical phenomena underlying temporal-enhanced ultrasound as a new diagnostic imaging technique: theory and simulations. IEEE transactions on ultrasonics, ferroelectrics, and frequency control. 65(3):400–410.

Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA. 2019. Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems. 32:5049—-5059.

Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, Tsang I, Sugiyama M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in Neural Information Processing Systems. 31:8536—-8546.

Imani F, Abolmaesumi P, Gibson E, Khojaste A, Gaed M, Moussa M, Gomez JA, Romagnoli C, Leveridge M, Chang S, et al. 2015. Computer-aided prostate cancer detection using ultrasound rf time series: in vivo feasibility study. IEEE Transactions on Medical Imaging. 34(11):2248–2257.

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

Javadi G, Samadi S, Bayat S, Pesteie M, Jafari MH, Sojoudi S, Kesch C, Hurtado A, Chang S, Mousavi P, et al. 2020. Multiple instance learning combined with label invariant synthetic data for guiding systematic prostate biopsy: a feasibility study. International Journal of Computer Assisted Radiology and Surgery. 15:1023–1031.

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

Javadi G, Samadi S, Bayat S, Sojoudi S, Hurtado A, Eshumani W, Chang S, Black P, Mousavi P, Abolmaesumi P. 2022b. Training deep neural networks with noisy clinical labels: toward accurate detection of prostate cancer in us data. International Journal of Computer Assisted Radiology and Surgery:1–9.

Johnson DC, Raman SS, Mirak SA, Kwan L, Bajgiran AM, Hsu W, Maehara CK, Ahuja P, Faiena I, Pooli A, et al. 2019. Detection of individual prostate cancer foci via multiparametric magnetic resonance imaging. European Urology. 75(5):712–720.

Lee DH, et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML; vol. 3. p. 896.

Li J, Socher R, Hoi SC. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations.

Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C. 2020. Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems. 33:20331–20342.

Ma X, Huang H, Wang Y, Romano S, Erfani S, Bailey J. 2020. Normalized loss functions for

deep learning with noisy labels. In: International Conference on Machine Learning. PMLR. p. 6543–6553.
Moe A, Hayne D. 2020. Transrectal ultrasound biopsy of the prostate: does it still have a role in prostate cancer diagnosis? Translational Andrology and Urology. 9(6):3018.
Moradi M, Abolmaesumi P, Mousavi P. 2010. Tissue typing using ultrasound rf time series: experiments with animal tissue samples. Medical physics. 37(8):4401–4413.
Rawla P. 2019. Epidemiology of prostate cancer. World Journal of Oncology. 10(2):63.
Sedghi A, Mehrtash A, Jamzad A, Amalou A, III WMW, Kapur T, Kwak JT, Turkbey B, Choyke P, Pinto P, et al. 2020. Improving detection of prostate cancer foci via information fusion of MRI and temporal enhanced ultrasound. International Journal of Computer Assisted Radiology and Surgery. 15:1215–1223.
Serefoglu EC, Altinova S, Ugras NS, Akincioglu E, Asil E, Balbay MD. 2013. How reliable is 12-core prostate biopsy procedure in the detection of prostate cancer? Canadian Urological Association Journal. 7(5-6):E293.
Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li CL. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems. 33:596–608.
Tarvainen A, Valpola H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems. 30:1195—-1204.
To MNN, Fooladgar F, Javadi G, Bayat S, Sojoudi S, Hurtado A, Chang S, Black P, Mousavi P, Abolmaesumi P. 2022. Coarse label refinement for improving prostate cancer detection in ultrasound imaging. International Journal of Computer Assisted Radiology and Surgery. 17(5):841–847.
Wang Y, Ma X, Chen Z, Luo Y, Yi J, Bailey J. 2019. Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 322–330.
Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. 2018. mixup: Beyond empirical risk minimization. International Conference on Learning Representations.
Zhang Z, Sabuncu MR. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In: 32nd Conference on Neural Information Processing Systems (NeurIPS).
Zhou X, Liu X, Jiang J, Gao X, Ji X. 2021. Asymmetric loss functions for learning with noisy labels. In: International Conference on Machine Learning. PMLR. p. 12846–12856.