Domain Transfer Through Image-to-Image Translation for Uncertainty-Aware Prostate Cancer Classification

Meng Zhou simon.zhou@queensu.ca
Medical Informatics Laboratory, Queen's University, Kingston, ON, Canada
Amoon Jamzad a.jamzad@queensu.ca
Medical Informatics Laboratory, Queen's University, Kingston, ON, Canada
Jason Izard jason.izard@kingstonhsc.ca
Kingston Health Sciences Research Centre, Kingston, ON, Canada
Alexandre Menard alexandre.menard@kingstonhsc.ca
Kingston Health Sciences Research Centre, Kingston, ON, Canada
Robert Siemens robert.siemens@kingstonhsc.ca
Kingston Health Sciences Research Centre, Kingston, ON, Canada
Parvin Mousavi mousavi@queensu.ca
Medical Informatics Laboratory, Queen's University, Kingston, ON, Canada

Abstract

Prostate Cancer (PCa) is often diagnosed using High-resolution 3.0 Tesla(T) MRI, which has been widely established in clinics. However, there are still many medical centers that use 1.5T MRI units in the actual diagnostic process of PCa. In the past few years, deep learning-based models have been proven to be efficient on the PCa classification task and can be successfully used to support radiologists during the diagnostic process. However, training such models often requires a vast amount of data, and sometimes it is unobtainable in practice. Additionally, multi-source MRIs can pose challenges due to cross-domain distribution differences. In this paper, we have presented a novel approach for unpaired image-to-image translation of prostate mp-MRI for classifying clinically significant PCa, to be applied in data-constrained settings. First, we introduce domain transfer, a novel pipeline to translate unpaired 3.0T multi-parametric prostate MRIs to 1.5T, to increase the number of training data. Second, we estimate the uncertainty of our models through an evidential deep learning approach; and leverage the dataset filtering technique during the training process. Furthermore, we introduce a simple, yet efficient *Evidential Focal Loss* that incorporates the focal loss with evidential uncertainty to train our model. Our experiments demonstrate that the proposed method significantly improves the Area Under ROC Curve (AUC) by over 20% compared to the previous work (98.4% vs. 76.2%). We envision that providing prediction uncertainty to radiologists may help them focus more on uncertain cases and thus expedite the diagnostic process effectively. Our code is available at `https://github.com/med-i-lab/DT_UE_PCa`.

Keywords: Deep Learning, Image Translation, Uncertainty Estimation, Prostate Cancer

1. Introduction

Prostate Cancer (PCa) is a prevalent form of cancer among men (Reda et al., 2018), and the clinically significant PCa is defined by the Gleason score > 6 or the histopathology ISUP grade ≥ 2 (Smith et al., 2004; Arif et al., 2020). The current PCa diagnosis procedure involves a combination of the prostate-specific antigen test and the histopathology analysis of the Transrectal Ultrasound-guided biopsy (TRUS) taken from 10-12 regions on the prostate gland (Fletcher, 2019; Grebenisan et al., 2021). However, the histopathology analysis on TRUS can miss up to 20% of clinically significant PCa due to the limited number of biopsy samples (Grebenisan et al., 2021; Reda et al., 2018). Multi-parametric Magnetic Resonance Imaging (mp-MRI) has emerged as an effective alternative to TRUS for the early detection of PCa. mp-MRI uses a combination of anatomical and functional sequences of MRI that can further highlight the differences between normal and abnormal (cancer) cells. 3.0T MRI generally has higher image quality and spatial resolution (Ladd et al., 2018) than 1.5T MRI, but the latter is widely used in local, small clinical centers due to its lower price (Mushlin et al., 1997). The evaluation and reporting guideline of prostate mp-MRI was first introduced in the Prostate Imaging Reporting and Data System (PI-RADS) (Barentsz et al., 2012; Weinreb et al., 2016; Barentsz et al., 2016). The guideline provides a comprehensive scoring schema for suspicious prostate lesions and mp-MRI sequences. An extensive Prostate MRI imaging study (PROMIS) (Bosaily et al., 2015) reported that targeted biopsy using mp-MRI has higher sensitivity and negative-predictive value (NPV) but lower specificity compared to TRUS biopsy (Bosaily et al., 2015; Stabile et al., 2020; Ahmed et al., 2017). The study also showed that 27% of the patients did not need to undergo biopsy, had mp-MRI been used for screening. Although PROMIS provides strong practical implications for mp-MRI in PCa diagnosis, the low specificity indicates that mp-MRI can be plausibly improved by advanced analyses.

In recent years, deep learning methods have emerged as a powerful tool for image classification tasks, and have provided promising performance in detecting and segmenting PCa on multi-parametric Prostate MRIs (Saha et al., 2021; Le et al., 2017; Yoo et al., 2019; Iqbal et al., 2021; Pellicer-Valero et al., 2022). A more recent grand challenge, ProstateX (Armato et al., 2018), has further shown the ability of deep learning approaches in detecting clinically significant PCa on 3.0T mp-MRI data. Several groups have developed Convolutional Neural Network (CNN)-based models that achieve high performance for PCa classification (Litjens et al., 2014; Liu et al., 2017; Mehrtash et al., 2017; Armato et al., 2018; Grebenisan et al., 2020, 2021). These methods have a great potential for clinical translations by highlighting abnormal lesions for radiologists during the PCa diagnostic process.

While deep learning has shown promising results in detecting PCa on mp-MRI, there are several challenges in deploying deep models in local clinics with limited data and patient throughput. Training deep models typically requires a large amount of data, making them difficult to deploy in small clinics. Moreover, MRI data may be acquired under different magnetic strengths, vendors, and protocols, which can affect the performance of deep models. For example, prostate MRI typically with high magnetic strength of 3.0T (Ullrich et al., 2017) is preferred because it produces high-resolution images and provides detailed information. In contrast, the low magnetic strength MRI (1.5T) may result in fuzzy boundaries (Ladd et al., 2018) and not be able to offer detailed information. The performance of deep models will be significantly affected if there is a difference between training and test distribution. There are efforts in the literature on solving related problems in federated learning (Li et al., 2020; Adnan et al., 2022) which is not a focus on our study. Furthermore, classical deep models are designed to predict a label when inferring data from a test set, regardless of whether or not the test image is in or out of the training set distribution. These models are not able to identify data samples that belong to unrelated distributions (Sensoy et al., 2018), or indicate how confident they are in their prediction. These limitations make models hard to interpret and hence, there are concerns about the reliability of such models. Hence, reusing and deploying models for local PCa detection is challenging. It is essential to address the above limitations and drawbacks when deploying deep learning models to real clinical routines. Thus, two main questions arise in this context:

1. For small local clinical centers, can they take advantage of the large high-resolution 3.0T public MRI data and enhance the classification performance on their limited low-resolution local 1.5T MRI data?

2. When deploying models in clinical centers, could we offer additional information regarding the confidence of the model's predictions, in addition to the final result, to enhance the reliability of the models?

In this work, we aim to answer the two questions above. We propose a novel 2-stage learning framework for clinically significant PCa classification using multi-parametric, multi-center MRI data that can simultaneously provide an estimate of the predictive confidence and the corresponding predicted label to improve classification performance. In the first stage, we introduce a data preprocessing pipeline that translates prostate mp-MRI data from 3.0T to 1.5T via a Generative Adversarial Network (GAN) approach in order to increase the number of training samples. This step addresses the challenge of limited data in local clinics with low patient throughput (see Section 4.1). In the second stage, we propose an uncertainty-aware PCa classification approach. Specifically, we design three different model architectures and leverage the *co-teaching* framework (Han et al., 2018) to address the noisy label problem (see Section 4.2.1). During the training phase, we incorporate dataset filtering using *evidential uncertainty estimation* (Sensoy et al., 2018) to eliminate the training data samples with high prediction uncertainty to improve the robustness of our models. Finally, we extend the work of Sensoy et al. (2018) to design a novel *Evidential Focal Loss* to optimize our classification models during training (see Section 4.2.2). Experiments demonstrate the effectiveness of the proposed framework in significantly improving the classification performance compared to previous work.

Contributions: In summary, our work makes three main contributions:

1. We develop a GAN-based framework to translate unpaired prostate mp-MRIs from 3.0T to 1.5T, which we termed as domain transfer. This framework would align different data distributions and increase the number of training data for deep classification models.

2. We incorporate the Theory of Evidence (Yager and Liu, 2008) into our model, enabling it to identify and filter out highly uncertain training data and making the model more robust. We propose a novel loss function termed *Evidential Focal Loss* that combines the original Focal Loss (Lin et al., 2017) and the evidential uncertainty (Sensoy et al., 2018) for the binary PCa classification task.

3. Using the uncertainty and filtering on the training set, our results outperform the state-of-the-art and improve the interpretability of model predictions. By providing confidence estimates for the predictions, radiologists can make informed decisions during the PCa diagnostic process and effectively expedite the process.

2. Related Work

2.1 Domain Adaptation

Machine learning algorithms usually perform well when training and test data share the same distribution and feature space. However, in real-world applications, the distribution of test data often shifts, leading to biased or inaccurate predictions. In addition, it is time-consuming or infeasible to acquire new training data and fully repeat training steps. Domain Adaptation (DA) is an approach that addresses this issue by mitigating the dataset bias or domain shift problem caused by different distributions. There has been a lot of work on this topic in the past few years, which can be grouped into the following three general tasks (Cui et al., 2020): (1) unsupervised DA tasks (Ganin and Lempitsky, 2015; Ganin et al., 2016; Long et al., 2016; Saito et al., 2018; Long et al., 2014) focus on addressing the domain shift problem without requiring labeled target domain data; (2) semi-supervised DA tasks(Yao et al., 2015; Saito et al., 2019; Li et al., 2021) aim to explore the partially labeled target domain data to further enhance the performance of domain adaptation algorithms; and (3) multi-source DA tasks (Hoffman et al., 2012; Xu et al., 2018; Peng et al., 2019) deal with scenarios where multiple source domains are available for adaptation. DA methods are often used to extract domain-invariant features for transferring knowledge between source and target domains. These methods incorporate various learning objectives with deep neural networks Wang and Deng (2018) for distribution matching: (1). **Discrepancy Measurement-based** methods aim to align feature distributions between two domains by fine-tuning deep models, e.g., using statistic criterion like Maximum Mean Discrepancy (Long et al., 2015; Yan et al., 2017; Kumagai and Iwata, 2019), and class criterion (Tzeng et al., 2015; Hinton et al., 2015; Motiian et al., 2017). Some of these methods often require large labeled target domain data to diminish the domain shift problem, which is sometimes infeasible to get such medical data in the real-life scenario. (2). **Adversarial-based** methods aim to confuse domain discriminators from Generative Adversarial Networks (GANs) to enhance the invariant feature extraction (Ganin et al., 2016; Bousmalis et al., 2017; Hong et al., 2018). One common scenario involves utilizing noise vectors, either with or without source images, to generate realistic target images while preserving the source features. However, training GANs are hard and sometimes results in generator degradation, e.g., mode collapse (Karras et al., 2020). (3). **Reconstruction-based** methods, in addition to the general GANs approach from the above category, aim to reconstruct source-like images as an auxiliary task to preserve domain invariant features through an adversarial reconstruction paradigm (Hoffman et al., 2018; Zhu et al., 2017). These methods usually have superior performance over the conventional GANs approach because they have an explicit reconstruction task to supervise the entire pipeline and make the training process more stable.

CycleGAN (Zhu et al., 2017) is one of the state-of-the-art unsupervised adversarial reconstruction-based methods that is widely used for unpaired image-to-image translation. Its cycle consistency loss ensures the pixel-level similarity between two images through a reconstruction task, i.e., the source image s is translated to the target domain $\hat{s}$ and then translated back $\tilde{s}$, where it should be identical to the original image ($s = \tilde{s}$). However, one drawback of cycle consistency loss is the harsh constraint on the pixel-level, which will degrade the performance of GANs in some tasks (Zhao et al., 2020). To address this limitation, Zhao et al. (2020) purpose the adversarial consistency loss GAN (ACL-GAN) that replaces the pixel-level similarity with the distance between distributions, i.e., instead of forcing $s = \tilde{s}$, we let the distribution of $\tilde{s}$ to be similar to the distribution of s. The ACL-GAN can retain important features from the source images and overcomes the disadvantage of the cycle-consistency constraint. Therefore, we adapt the ACL-GAN model and build our framework based on it.

In medical imaging, domain shift problems usually fall into two variations: subject-related variation (age, gender, etc.), and acquisition-related variation (MRI vendor, field strength, imaging protocol, etc.) (Kouw et al., 2017). To solve the problem, one intuitive approach is to fine-tune a model that is pre-trained on the source domain with the new data from the target domain. Khan et al. (2019) propose to use the pre-trained VGG model on the ImageNet dataset (Deng et al., 2009) to learn robust high-level features of natural images, and then fine-tune it on the labeled MR images for the Alzheimer's Disease (AD) classification task to achieve state-of-the-art performance. Similarly, Ghafoorian et al. (2017) study the impact of the fine-tuning techniques on the brain lesion segmentation task, demonstrating that fine-tuning with only a small number of target domain training samples can outperform models trained from scratch. Another approach is to use domain adaptation as an intermediate step to reduce variance in image acquisition parameters from both domains and then use it for downstream tasks. Researchers have attempted to address the problem of acquisition variation in MRI data for several years. Kouw et al. (2017) propose a feature-level representation learning method to either extract acquisition-invariant features or remove acquisition-variant features from paired 1.5T and 3.0T brain MRIs. The learned features are then used for a downstream classification task. However, obtaining paired 1.5T and 3.0T MRI data in real-life scenarios is impractical. Another way to align acquisition-invariant features is to synthesize images from different types of acquisition parameters using GAN-based adversarial reconstruction methods. GANs have been applied to perform cross-modality image translation between different medical images or generate synthetic images from random noise. The objective of such translation tasks is to retain the underlying structure while changing the appearance of the image (Armanious et al., 2020). Researchers have attempted to estimate images in the target modality from the source modality, such as MRI-CT translation (Hiasa et al., 2018; Nie et al., 2017; Oulbacha and Kadoury, 2020; Armanious et al., 2019), and X-ray to CT translation (Ying et al., 2019; Ge et al., 2022). Other areas that have been explored include intra-modality translation, such as 3.0T-7.0T MRI translation (Nie et al., 2018), T1/T2-FLAIR translation (Hu, 2021; Uzunova et al., 2020) and pure data augmentation by generating synthetic images from random noise vectors (Radford et al., 2015; Frid-Adar et al., 2018; Huang and Jafari, 2021; Kwon et al., 2019). However, most of the works do not consider the real clinical practicality, for example, for 3.0T-7.0T MRI translation in Nie et al. (2018), the training data is paired, which is not feasible in real clinical settings. Generating synthetic images from noise does not take advantage of the publicly available data and ignores *a-priori* information. The current limitations provide great potential for unpaired image translation for medical images, which we employ in this work.

2.2 Deep Learning for PCa Classification

The use of 3D-CNN models has gained widespread popularity for classifying PCa based on volumetric image data due to their excellent performance. Mehrtash et al. (2017) propose a feature fusion 3D-CNN to classify clinically significant PCa using mp-MRI data. They use ADC maps, DWI, and K^{trans} 3.0T MR data to enable the model to learn multi-modal information. Inspired by the VGG architecture (Simonyan and Zisserman, 2014), the model has three VGG-like feature extractors for each image modality, followed by the concatenation between outputs of each extractor and a vector represents the zonal information of the suspicious region. On the test set, the proposed model achieves the area under the receiver operating characteristic (AUC) curve of 0.80 on 140 unseen patients. Liu et al. (2017) propose a similar VGG-like 3D-CNN architecture for the same PCa classification task. Different from Mehrtash et al. (2017), they only have one model for feature extraction. To obtain the multi-modal information, they stack three images from each of the ADC maps, DWI, K^{trans} into one 3-channel image as the input. The model achieves the AUC of 0.84 on the test set.

In Yoo et al. (2019), a probabilistic approach using mp-MRI data is employed for PCa classification. The authors develop an automated pipeline for the classification of clinically significant PCa using 3.0T DWI images from 427 patients. The pipeline consists of three parts: classification of each DWI slice using the pre-activated ResNet model (He et al., 2016), extraction and selection of first-order statistics from the CNN outputs, and final class label prediction using a random forest classifier. On the test set, the model achieves an AUC of 0.87. While the aforementioned studies may yield favorable AUCs, the reproducibility of the model might be challenging clinics with limited patient (data) throughput. Recently, Grebenisan et al. (2021, 2020) address the data-hungry problem by introducing a disentangled representation learning approach (SDNet) to synthesize public 3.0T MRI images into 1.5T MRI images to increase the training data size for centres with limited 1.5T data. Their approach aims to separate the anatomy- and modality-specific features present in images, subsequently merging the 1.5T modality features with the 3.0T anatomical features to generate MRI images resembling those acquired at 1.5T. Finally, a simple 3D-CNN classifier is used for the binary classification of clinically significant PCa. The model outperforms the state-of-the-art performance in PCa classification through domain alignment between different data sources.

Although current methods for PCa classification can achieve good performance, they do not provide a confidence score for their prediction, making them less interpretable in clinical practice.

2.3 Uncertainty Estimation

Recent studies in medical imaging have highlighted the detrimental impact of label noise on the performance of modern deep learning models (Karimi et al., 2020). Conventional regularization techniques such as dropout, batch normalization, weight decay, etc. can not properly address the problem (Arpit et al., 2017; Zhang et al., 2021). Methods proposed to mitigate such problem can be summarized as those that use (Song et al., 2022): (1) Robust loss functions and loss adjustments (Van Rooyen et al., 2015; Charoenphakdee et al., 2019; Zhang and Sabuncu, 2018) aiming to stabilize the model performance when optimizing its parameters; (2) Sample selection (Jiang et al., 2018; Malach and Shalev-Shwartz, 2017; Wang et al., 2020) aiming to select a subset of "clean" data from a batch of samples to compute the loss; and (3) Robust architectures (Han et al., 2018; Yu et al., 2019) aiming to learn the same data by training multiple models with different initialization assess output stability. While these methods inherently handle the noisy label problem, they can not provide explicit uncertainty estimation in terms of confidence in their output. Moreover, the capability of deep learning models to effectively identify irrelevant samples is still limited. For instance, when a model trained on prostate MRIs is presented with a CT scan of the prostate at the time of inference, it is unclear whether the model can provide meaningful predictions or simply indicate a lack of in-domain knowledge and perform a human-in-the-loop analysis instead. In recent years, research has been conducted on uncertainty estimation for deep learning models. Gal and Ghahramani (2016, 2015) develop the *dropout neural networks* framework to represent the prediction uncertainty of deep learning models, where the dropout layers in the model are formed by Bernoulli distributed random variables. During the test phase, the predictive uncertainty can be determined by enabling dropout layers and averaging the results of multiple runs. An alternative approach to modeling uncertainty involves the use of *evidential neural networks* (Sensoy et al., 2018), which formulate uncertainty by fitting a Dirichlet distribution - acting as the conjugate prior of the categorical distribution — to the class probabilities acquired from neural networks (Sensoy et al., 2018). This method considers model predictions as multinomial subjective opinions (Jøsang, 2016) or beliefs (Dempster, 1968), which can be further modeled explicitly using subjective logic. The "evidential" approach emphasizes the ability of the model to deliver certain predictions and exhibits superiority compared to the dropout approach (Gal and Ghahramani, 2016).

In clinical practice, uncertainty estimation is crucial. By integrating uncertainty information into prediction outcomes, misclassification rates can be significantly reduced. For instance, in radiograph classification task (Ghesu et al., 2019), the authors employ the Dempster-Shafer Theory of Evidence (Dempster, 1968) and the principles of subjective logic (Jøsang, 2016) to develop a framework that jointly estimates per-class probabilities and provides predictive uncertainty. Later, this approach has been extended to abdominal ultrasound and brain MR images (Ghesu et al., 2021). In the context of breast cancer classification, Tardy et al. (2019) apply the evidential neural networks approach (Sensoy et al., 2018) to effectively diagnose breast cancer. A similar approach is used for the same task by Yuan et al. (2020) through the evidence adjustment technique, which focuses on the difference in the risks of uncertain samples from different classes. Consequently, we build upon the work from Sensoy et al. (2018) by adding uncertainty estimation to improve the robustness of the model and the interpretability of predictions.

3. Materials

3.1 Data

In this work, we use both large publicly available ProstateX data and small private local clinical data. A visualization of sample images from both datasets is presented in Figure 1.

ProstateX Grand Challenge Data (3.0T). The 3.0T data is provided by the International Society of Optics and Photonics in the "ProstateX" challenge (Litjens et al., 2017). The dataset contains T2-weighted (T2), maximum b-value diffusion, diffusion-weighted imaging (DWI) with apparent diffusion coefficient (ADC) maps, and K^{trans} images of 346 patients undergoing prostate biopsies. T2 images show the anatomical structure of the prostate, and both the ADC maps and K^{trans} could further highlight the differences between normal and abnormal (cancer) cells in the MRI scans (Kasivisvanathan et al., 2018; Kasson et al., 2018). We only use 204 of the total 346 patients in this work since these are reserved as training data, and hence they are provided with the spatial location of the suspicious finding, and a binary label indicating whether or not there is cancer. The remaining 142 patients are reserved as the test set and no labels are provided, hence, we exclude those from our work. **Kingston Health Science Center Data (1.5T)**. The local 1.5T data is obtained from the Kingston Health Science Center (KHSC), which contains 104 patients with the corresponding biopsy-confirmed cancer and the Gleason Score. For the local data, only T2, ADC, and b-value images are available. All patients MRI have the spatial location of the suspicious finding(s), the Gleason Score, and the binary label indicating whether it is a cancer lesion or not.

Since all patients in both datasets have complete T2 and ADC data, our focus in this work is solely on these two types of images. Each MRI data in our study is associated with a single patient. Both datasets are processed similarly unless stated.

3.2 Pre-processing

T2 and ADC sequences from both datasets are $160 \times 160 \times C$, where C is the total number of slices in the MRI. We resample all 3D data to have the same voxel spacing. To reduce aliasing artifacts, the most common voxel spacing ($0.5 \times 0.5 \times 3$ mm^3) is used across all data, and the consine-windowed interpolation is utilized during sampling. We normalize pixel intensities to $[-1, 1]$ for all data. For the translation purpose from 3.0T to 1.5T, we further resample all 3D data to $256 \times 256 \times C$ and split into C 2D gray-scale slices.

Augmentation: For each patient, the MRI volume undergoes rotation ranging from 0 to 100 degrees in 5-degree increments, hence expanding the data size 20-fold.

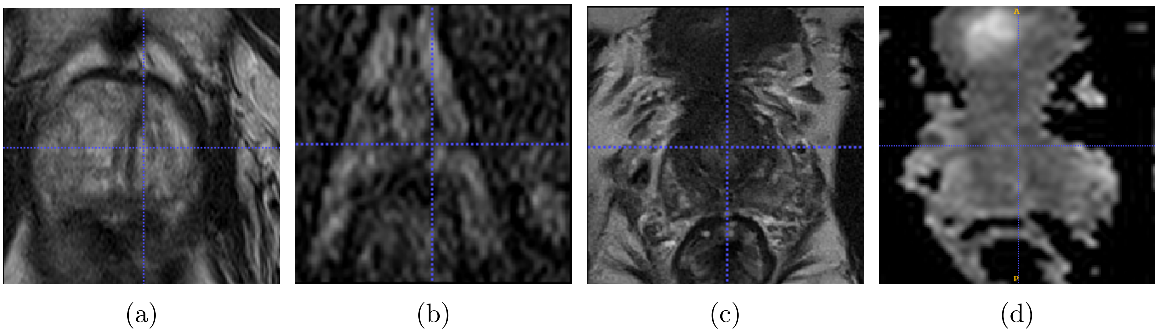

(a) (b) (c) (d)

Figure 1: Visualization of sample data. 1a and 1b are the 1.5T T2 and ADC images from KHSC, respectively. Similarly, 1c and 1d are the 3.0T T2 and ADC images from the "ProstateX" Challenge, respectively.

Cropped Patches: To reduce the computational cost, cropped patches of the MRI volume were employed. The process involves identifying the suspicious slice (i_s) based on the provided spatial location. Recognizing that PCa lesions can span multiple slices, two neighboring slices (i_{s-1} and i_{s+1}) are selected as well and cropped around the biopsy location to generate a patch of size $64 \times 64 \times 3$.

4. Methods

Figure 2 summarizes an overview of our proposed approach. The domain transfer framework aims to reduce the distribution-level discrepancy between two prostate MRI datasets. The framework matches the acquisition parameters of publicly available, large 3.0T prostate mp-MRI data with local, small 1.5T prostate mp-MRI data. Once all the data from 3.0T are translated to 1.5T, a subsequent classifier is trained to classify clinically significant PCa. Furthermore, during the training process, the uncertainty is calculated along with the class output. We also introduce a novel evidential focal loss for the PCa classification task. Lastly, we utilize dataset filtering to improve robustness and accuracy by eliminating uncertain data samples from the training set.

4.1 The Domain Transfer Framework

We adapt the ACL-GAN model (Zhao et al., 2020) discussed in Section 2.1 to perform unpaired MR image translation from 3.0T to 1.5T. There are two generators in this model namely $G_{T->S}$ and $G_{S->T}$, $G_{T->S}$ translates the images from the target domain to the source domain given the input x and a noise vector z sampled from $\mathcal{N}(0, 1)$. $G_{S->T}$ is the reverse process of $G_{T->S}$ which translates the image from the source domain to the target domain. There are three discriminators, D_S, D_T, and $\hat{D}$ in this model. The first two ensure that translated images are in the correct domain by optimizing adversarial losses, and $\hat{D}$ ensures that translated images retain anatomical features in 3.0T by distinguishing the pair (Source, Trans. Source1) and (Source, Trans. Source2), as shown in the bottom of Figure 2. The loss function of ACL-GAN (Zhao et al., 2020) is defined as in equation (1):

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{acl}\mathcal{L}_{acl} + \lambda_{idt}\mathcal{L}_{idt} + \lambda_{mask}\mathcal{L}_{mask} \tag{1}$$

Where $\mathcal{L}_{adv}$ is the traditional adversarial loss for both source domain S and target domain T, i.e., $\mathcal{L}_{adv} = \mathcal{L}^S_{adv} + \mathcal{L}^T_{adv}$, to ensure the translated image is in the correct domain. $\mathcal{L}_{acl}$ is the adversarial consistency loss that is used to preserve important features of the source image in the translated image, $\mathcal{L}_{idt}$ is the identity loss, which encourages the generators to perform approximately identity mapping when images in the corresponding domain are provided, and $\mathcal{L}_{mask}$ is used to force both generators to only modify certain regions of the source image and keep the rest of the areas unchanged. Readers are encouraged to refer to the original paper (Zhao et al., 2020) for more details.

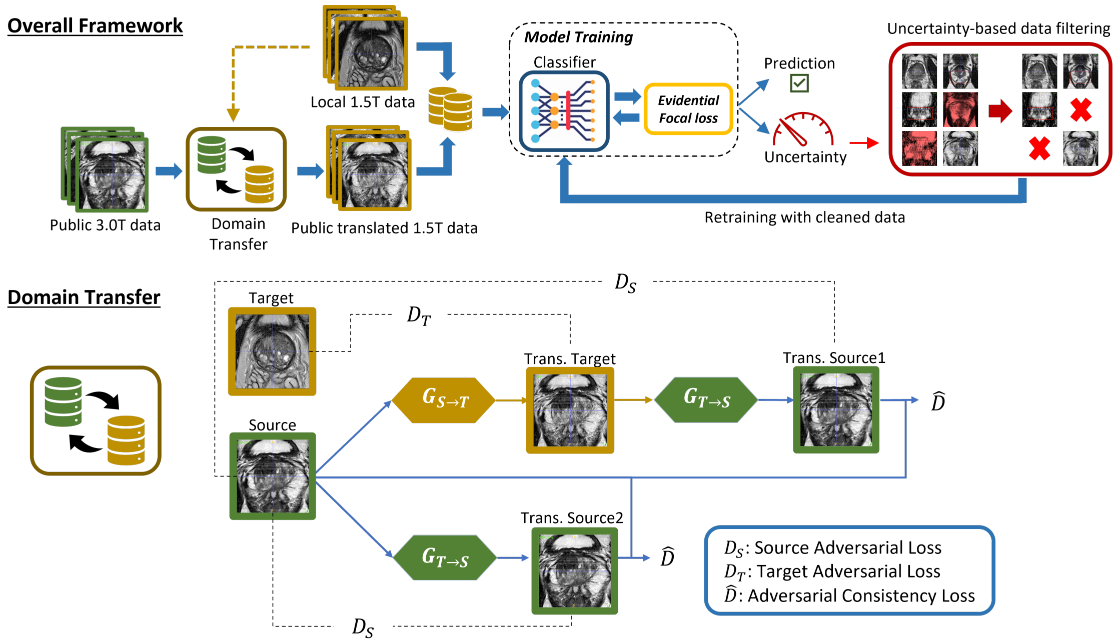


Figure 2: Detailed schematic of the proposed method. The overall framework of our proposed method contains two stages: 1), domain translation to map public 3.0T MRI with local 1.5T MRI; 2), uncertainty-aware clinically significant PCa classification. The bottom figure is the training schema for domain transfer. The upper right portion of the figure illustrates the PCa classification training process, which involves training the classifier using the Evidential Focal loss, filtering the training set based on uncertainty, and retraining the classifier on the filtered data to obtain the final classifier.

4.2 Uncertainty-aware PCa Classification

4.2.1 Classifier architectures

The traditional CNN approach is used for the clinically significant PCa binary classification task. Specifically, we explore three different model architectures for combinations of T2 and ADC patches. The first architecture, called the multi-stream CNN ("M.S. MpMRI"), treats T2 and ADC patches as separate inputs, as shown in Figure 3. The model takes 3D patches of T2 and ADC as parallel inputs, which are then processed by the same feature extractor to extract deep semantic representations. The output representations of T2 and ADC are then concatenated channel-wise and fed into another convolutional layer followed by a fully-connected layer to produce the class probabilities.

In the second architecture, we combine ADC and T2 patches as a single input to the network. We stack cropped 3D patches of T2 and ADC along the channel axis and obtain the input data size of $64 \times 64 \times 6$. Another way to combine is we only consider the located suspicious slice i_s for both T2 and ADC, and stack them along the channel axis to obtain the input data size of $64 \times 64 \times 2$. The model architecture for both combinations is similar to Figure 3, where there is only one branch and no concatenation afterward. We name the model with input size of 64 × 64 × 6 (resp. input size 64 × 64 × 2) as "Vol. MpMRI"(resp. "MpMRI").

Lastly, we use only 3D T2 patches as input to match with the previous work (Grebenisan et al., 2021). The model architecture is as same as the one for MpMRI, and we call this model "T2-only".

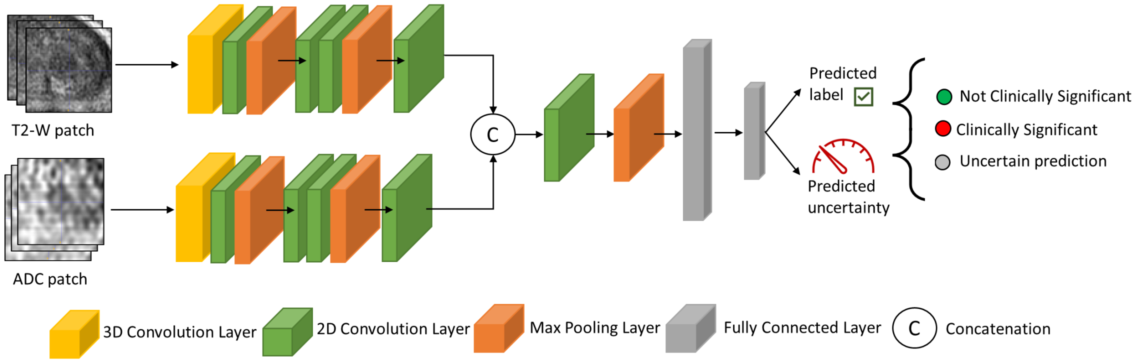


Figure 3: Detailed architecture of "M.S. MpMRI" model. The first sequence of CNN layers contains 1 × 3D convolution layer and 4 × 2D convolution layers, 2 × Max Pooling layers with window size 2 × 2. Both extracted feature maps of T2 and ADC are concatenated channel-wise. After that, another set of convolution-max pooling layers is utilized. Finally, the extracted 2D features are reshaped to 1D and fed into a Fully connected layer follow by a softmax layer with 2 outputs representing the probabilities of which class the input data belongs to.

To combat the potential noisy label problem, the *co-teaching* framework (Han et al., 2018) is also utilized in this work. In co-teaching, two models with the same architecture and configurations are trained simultaneously, as shown in Figure 4. In every mini-batch, two models are trained in parallel. Each model first feeds forward all data in the current batch and selects some of the data that has the clean labels with high probability; then, two models decide on what data in the current batch should be used for training; finally, each model uses the data selected by the peer model to update itself. Denote f be the first model and g be the second model. The number of instances selected by both models is controlled by a non-increasing function $R(T)$ on T defined in equation (2). At epoch T, each model only calculate loss based on the $R(T)$ portion of the batch instances.

$$R_T = (1 - \tau \cdot \min(T/T_k, 1)) \times 100\% \tag{2}$$

where T is the total training epochs, τ is the same as noise rate, and T_k is the epochs for linear drop rate. We use the "MpMRI" as the backbone model (model A and B in Figure 4) in the co-teaching framework.

4.2.2 Evidential Focal Loss

Dataset filtering during the training phase could reduce the effect of the noisy label on the deep model. Followed by Ghesu et al. (2019), the process of uncertainty-based filtering is

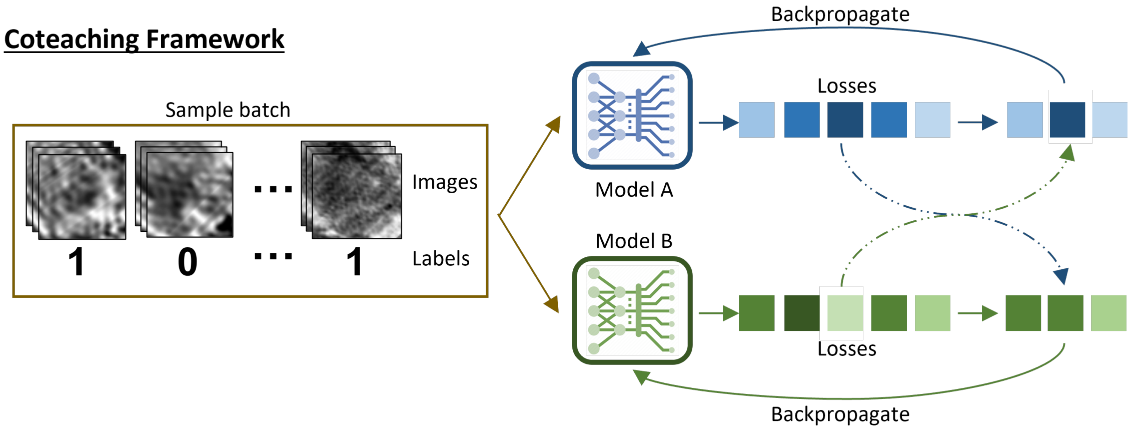


Figure 4: The co-teaching framework, where A and B are two models with the same architectures that are trained in parallel and simultaneously.

shown at the top of Figure 2: Firstly, we calculate the uncertainty value for each sample in the training set. We then remove a portion of the training samples that exhibit high predictive uncertainty. Finally, we retrain the model using the remaining "clean" training data.

Following the work from Sensoy et al. (2018), we extend and combine the idea of subjective logic (Jøsang, 2016) with the focal loss (Lin et al., 2017) for the clinically significant PCa binary classification task. In the context of the Theory of Evidence, a belief mass is assigned to individual attributes, e.g., the possible class label of a specific data sample. The belief mass is generally calculated from the evidence collected from the observed data (Dempster, 1968). Let K be the number of classes, and $b_k \geq 0, k \in [1, K]$ is the belief mass for class k and $u \geq 0$ is the overall uncertainty measure. Let $e_k \geq 0$ be the evidence computed for k^{th} class, then followed by Sensoy et al. (2018), the belief b_k and the uncertainty u are computed as the following:

$$b_k = \frac{e_k}{S} \quad and \quad u = \frac{K}{S} \tag{3}$$

where $S = \sum_{i=1}^{K}(e_i + 1)$. For our binary task ($K = 2$), we can further simplify Equation (3) to $b_0 = \frac{e_0}{e_0+e_1+2}, b_1 = \frac{e_1}{e_0+e_1+2}$, and $u = \frac{2}{e_0+e_1+2}$. The belief mass assignment, e.g., subjective opinion, is corresponding to the Dirichlet distribution with parameters $\alpha_k = e_k+1$, and we could rewrite $S = \sum_{i=1}^{K} \alpha_k$ as the Dirichlet strength. The formal definition of Dirichlet distribution can be found in Sensoy et al. (2018). The expected probability for the k^{th} class is calculated by the mean with the associate Dirichlet distribution $\hat{p}_k = \frac{\alpha_k}{S}, k \in [1, ..., K]$ (Sensoy et al., 2018).

Given the training set contains N data samples, $D := \{x_i, y_i\}_{i=1}^{N}$, where x_i is the i^{th} data sample and $y_i \in [0, 1]$ is the corresponding label, 0 is the negative sample and 1 is the positive sample. We further denote $\mathbf{y_i}$ as the one-hot encoding label for sample i, e.g., $\mathbf{y_i} = [1, 0]$ for class 0 and $\mathbf{y_i} = [0, 1]$ for class 1. The original focal loss is designed by Lin et al. (2017), aiming to address the imbalanced class problem and further penalized well-classified samples. The focal loss for binary classification is defined by $FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$ where $p_t = p$ if $y_i = 1$ for i^{th} sample, otherwise $p_t = 1 - p$ with probability output p from the model. Let $\mathbf{P}_i$ be a vector that contains the probability of i^{th} sample for both classes from our model output; $p_{i,j}$ is the probability of i^{th} sample belonging to j^{th} class; K is the number of classes, and β_j is the class weight of j^{th} class. γ is the focusing parameter to reduce the loss for well-classified samples, and we fix $\gamma = 2$ in this task. We could define Evidential Focal Loss as the following:

$$\mathcal{L}_i^{cls}(\theta) = \int \sum_{j=1}^{K} -\beta_j(1-p_t)^\gamma \log(p_{ij}) \frac{1}{\mathcal{B}(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\mathbf{P}_i \tag{4}$$

Rewriting class probabilities in vector form, the equation (4) can be simplified to (5) by the definition of expectations:

$$\mathcal{L}_i^{cls}(\theta) = -\sum_{j=1}^{K} \beta_j \mathbf{E}[(1-\mathbf{P}_i)^2 \log(p_{ij})] \tag{5}$$

Following the idea of focal loss, we replace the constant term 1 in the original focal loss function with $\mathbf{y_i}$, with the goal to tackle the hard-to-classified samples and reduce the loss of well-classified samples for both classes. Recall that expected probability $\hat{p}_k$ for the k^{th} class is α_k/S, then by the linearity of expectations and the definition of expectations of Dirichlet distribution, we could simply to:

$$\mathcal{L}_i^{cls}(\theta) = \sum_{j=1}^{K} \boldsymbol{\beta}(y_{ij} - (\alpha_j/S))^2(\psi(S_i) - \psi(\alpha_{ij})) \tag{6}$$

where $\psi(\cdot)$ is the digamma function, y_{ij} is the j^{th} class label in the one hot encoding representation $\mathbf{y_i}$ and $\boldsymbol{\beta}$ is the class weight vector of length K.

To ensure that highly uncertain data samples, referred to as "I do not know" decisions, do not impact the overall data fit and to minimize their associated evidence, we adopt the approach presented in Sensoy et al. (2018). This involves utilizing the Kullback-Leibler (KL) divergence as a regularization term to penalize the unknown predictive distributions, effectively shrinking their influence towards zero. The KL divergence is as same as it is defined in Sensoy et al. (2018).

Finally, our total loss is defined as:

$$\mathcal{L}^{total}(\theta) = \sum_{i=1}^{N} \mathcal{L}_i^{cls}(\theta) + \lambda_t \sum_{i=1}^{N} KL[D(\mathbf{P}_i|\alpha_i)||D(\mathbf{P}_i|\mathbf{1})] \tag{7}$$

where $\mathbf{1}$ is an one-vector, λ_t is the balancing factor between $\mathcal{L}^{cls}$ and the KL divergence loss, and is defined as $\lambda_t = min(1.0, t/10) \in [0, 1]$, where t is the current number of epochs of training.

Finally, we introduce two proposed methods for filtering training samples based on the calculated uncertainty.

Patch-driven filtering: Given the uncertainty for each training patch, we simply eliminate $x\%, x \in [10, 20]$ of the *patches* with highest uncertainty and retrain the model on the rest of the samples on the training set.

Patient-driven filtering: Similar to Patch-driven filtering, we first calculate the uncertainty of each training patch. To determine the uncertainty of each patient, we calculate the average uncertainty value across their corresponding patches (20 patches per patient as mentioned in Section 3.2). We then eliminate $x\%, x \in [10, 20]$ of the training *patients* with high uncertainty value and retrain the model on the rest on the training set.

5. Experiments & Setup

Data Split: For the domain transfer task, as mentioned in Section 3.2 and 4, the resampled T2 and ADC "images" with size 256×256 from both ProstateX and local datasets is used for training the ACL-GAN model. Particularly, we allocate 90% of images in both datasets as training and keep 10% as validation set to avoid overfitting the ACL-GAN model. Importantly, we ensure that the images corresponding to each patient are exclusively present in either the training or validation set, but not both. To improve the robustness and enhance the ability of the ACL-GAN to capture feature-level representations of 1.5T images, we use all data from our local hospital. However, it is important to note that this approach does not yield any additional impacts on the subsequent classification task. The model only modifies image regions that have visual differences caused by acquisition parameters of various MRI machines, but it does not alter the context of the prostate itself. For the PCa classification task, we use cropped and augmented T2 and ADC "patches" from both datasets. As mentioned before, this includes 204 ProstateX patients translated to 1.5T, as well as 104 patients from our local hospital captured in 1.5T. Regarding the data split for the classification, we keep patches of 34 patients from our local center as test set. From the remaining patches (70 local patients and all ProstateX patients, we allocate 80% for training and 20% for validation, assuring patches from the same patient not included in both of these sets.

Domain Transfer: The first experiment is translating the ProstateX MRI data from 3.0T to 1.5T using our proposed ACL-GAN model through the conversion process mentioned previously. We then evaluate the effectiveness of this approach by using the translated MRI data in a downstream binary classification task for clinically significant PCa.

Classification: We divide our classification experiments into two categories. In the first category, we use the conventional training paradigm without any filtering or uncertainty estimation. We use different model architectures for these experiments as discussed in Section 4.2.1. In the second category, we use the dataset filtering method and evidential focal loss proposed in Section 4.2.2 for training our models. We select two methods (MpMRI and M.S. MpMRI) that achieve the best performance from the first category to conduct experiments in the second category. Additionally, we conduct several ablation studies and report the results. Finally, we focus on dataset filtering during deployment and examine how this technique affects the classification performance of the test data.

5.1 Experimental Details

We train two ACL-GAN models separately for T2 and ADC images as part of our domain transfer framework. The optimizer used for both models is Stochastic Gradient Descent with Adam update rule (Kingma and Ba, 2014), with an initial learning rate of 0.0001 and weight decay of 0.0001 to prevent overfitting. The batch size is 3 and are trained for 30,000 epochs. Moreover, when training the model for T2 images, we set the $\lambda_{mask} = 0.0025, \lambda_{idt} = 1, \lambda_{acl} = 0.2$ in Equation (1) and lower and upper mask threshold to be 0.005 and 0.1, respectively. When training the model for ADC images, the value of $\lambda_{mask}, \lambda_{idt}, \lambda_{acl}$ are the same as in the T2 model with lower and upper mask threshold is set to 0.001 and 0.005, respectively. The Least-Square (LS) loss (Mao et al., 2017) is utilized to calculate $\mathcal{L}_{adv}$ and $\mathcal{L}_{acl}$ in Equation (1).

Converting 3.0T to 1.5T: Once we have obtained two ACL-GAN models, we need to standardize the acquisition parameters of 3.0T prostate MRIs to match those of the 1.5T data in our local dataset. To achieve this, we divide the original MRI into multiple 2D grayscale slices. For each 2D slice, we use the generator G_T and a noise vector z randomly sampled from $\mathcal{N}(0, 1)$ to translate the slice to 1.5T, e.g., $I_{1.5\mathrm{T}} = G_T(I_{3.0\mathrm{T}}, z)$ in Section 4.1. We repeat this process for all 2D slices and then stack them back together to reconstruct the 3D MRI for each patient. The voxel spacing remains unchanged before and after the translation process. The above process is applied to both T2 and ADC data.

All classification models are trained with Stochastic Gradient Descent with Adam, and batch normalization is used to speed up convergence. In the first category of classification experiments, the traditional focal loss (Lin et al., 2017) with $\gamma = 2$ is used to train the model. Specifically, all models except co-teaching are trained for 300 epochs with a learning rate of 0.0001, weight decay of 0.01, and a batch size of 10. To train the co-teaching model, we set the noise rate to 0.1; the forget rate $\tau = 0.1$; the number of epochs for linear drop rate $T_k = 10$ in Equation (2). The model is trained for 300 epochs; the batch size is set to 10, and the learning rate is 0.00001.

For experiments in the second category, to train the "MpMRI" model for patch-driven filtering, we set the learning rate to 0.0001; weight decay to 0.01; total training epochs to 300, and batch size to 10. The class weights $\boldsymbol{\beta}$ in Equation (6) are set to [0.25, 0.75] for filtering 10%, and [0.25, 1.25] for filtering 20% of the training data. For patient-driven filtering, all parameters are the same except the class weights $\boldsymbol{\beta}$ are set to [0.25, 1] for filtering 10% of the training data. Last but not least, we set the initial learning rate to 0.0001; total training epochs to 300; batch size to 10; the learning rate decayed by a factor of 0.1 for every 200 epochs, and the class weights $\boldsymbol{\beta}$ are set to [0.25, 1] for filtering 20% of the training data.

To train the "M.S. MpMRI" model for patch-driven filtering, we set the initial learning rate to 0.0001; weight decay to 0.01; total training epochs to 300. The class weights $\boldsymbol{\beta}$ in Equation (6) are set to [0.25, 1] for both filtering 10% and 20% of the training data. For patient-driven filtering, all parameters were the same except the class weights $\boldsymbol{\beta}$ are set to [0.25, 1] for filtering 10% and [0.25, 1.25] for filtering 20% of the training data.

5.2 Evaluation

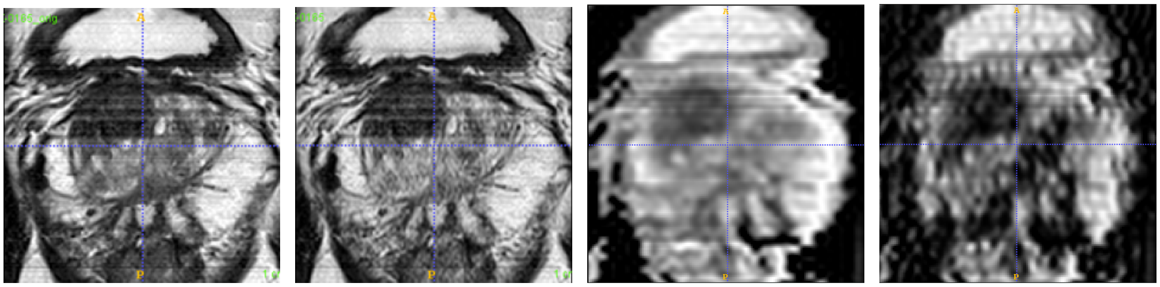
The traditional classification metrics e.g., accuracy, sensitivity, specificity, and AUC for this task. Reporting the patient-level performance are more relevant to the real clinical setting. However, since we used patches as input to the model, to be able to calculate the performance metrics at patient-level, we need to aggregate individual results obtained from the individual patches. To achieve this, we first use the classifier to predict test patches, which are then grouped sequentially into x groups, for each patient (20 probabilities due to the augmentation mentioned in Section 3.2) and then compute the *median* probability $\tilde{p}_i$ as the aggregated probability of that patient. Finally, a threshold of 0.5 is used to determine whether the patient has PCa or not, i.e., assigning label 1 if $\tilde{p}_i > 0.5$, and 0 otherwise.

6. Results and Discussion

In this section, we report **patient-based** classification results; the performance of our methods on patches are reported in Appendix A.

6.1 Translated samples from Domain Transfer

Figure 5 shows the visualization of the difference between original 3.0T T2 and ADC images in the ProstateX Challenge and the corresponding translated 1.5T T2 and ADC images using our proposed domain transfer framework. As shown, domain transfer reduces the image contrast and results in loss of minor details in the original 3.0T images. However, to evaluate the effectiveness of the domain transfer framework, a user study involving radiologists' interpretation is necessary. Our proposed domain transfer framework shows potential improvements over the baseline model SDNet for the following reason. In SDNet (Grebenisan et al., 2021), modality features from randomly selected 1.5T images are merged with anatomical features from 3.0T rather than from the whole distribution of all 1.5T images. In contrast, our method learns the overall data distribution in an adversarial manner, capturing the entire distribution of 1.5T images and performing the translation. Moreover, our method ensures the translated image contains the crucial features of the original image, and the generator only modifies certain parts of the image. Our image translation method is thus more suitable for this task and can be further validated by the classification performance presented in the following sections.



(a) Original T2 image in 3.0T

(b) Translated T2 image in 1.5T

(c) Original ADC image in 3.0T

(d) Translated ADC image in 1.5T

Figure 5: Translation from 3.0T T2 and ADC images to 1.5T-like images of a random patient in the "ProstateX" Challenge dataset. *left two figures:* the original 3.0T T2 image, and the same image that is translated to 1.5T. *right two figures:* the original 3.0T ADC image, and the same image that is translated to 1.5T.

6.2 PCa classification without filtering

Table 1 summarizes the main results of this study, which contains the PCa classification performance of all experiments we conducted. The table is divided into three sections. The first section corresponds to experiments conducted without using either the dataset filtering or uncertainty estimation, as described in the first category of Section 5. The second and third sections represent experiments with dataset filtering and uncertainty estimation described in the second category of Section 5. In the first section of Table 1, we observe that the AUC of using the co-teaching framework with "MpMRI" architecture as the base model achieves the best AUC and outperforms the baseline.

We also noticed that the sensitivity increases by approximately 50% while the specificity only decreases 10% for our co-teaching model compared to the baseline model, indicating our model has better learning abilities for classifying both positive and negative data samples. In the training process, we adopt a greedy approach of assuming 10% of the samples to be noisy. Consequently, both models need to designate a portion of the data in each batch as "clean" to update the parameters. This strategy allows our model to prioritize learning from the clean data, leading to enhanced robustness.

Ablation Study: We embed the results of the ablation study in the first section of Table 1. Our ablation here is two-fold: alteration of the number of input modalities, and alteration of the architecture of the model. To examine the effect of data modalities on the classification performance, we compare the T2-only model with (Vol.)MpMRI and M.S. MpMRI models, both use T2 and ADC patches as input. We observe a significant improvement in classification performance with the addition of the ADC modality, suggesting that multi-modal information is useful in guiding the model to classify clinically significant PCa. To examine the effect of model architecture on classification performance, we compare the MpMRI and M.S. MpMRI models, which have different architectures, and the co-teaching model. We find that the model with simpler inputs, MpMRI, performs better, and the results can be further improved by using the co-teaching framework.

6.3 PCa classification with filtration

In this section, we conduct experiments using two different architectures (MpMRI and M.S. MpMRI) and with training set filtering at various rates. The evidential focal loss described in Section 4.2.2 is used to optimize the models. The co-teaching framework is excluded from this section for the following reason: while co-teaching *implicitly* handles noisy labels or samples in the training set, the training set filtering in Section 4.2.2 is an explicit alternative way of dealing with them. The co-teaching framework will first update its model parameters with simpler and cleaner samples during training. However, through the filtering process, data samples with high uncertainty values are considered potentially noisy and do not involve in the training process. We argue that it would be a duplicate procedure if we use co-teaching and filtering the training data simultaneously. Our hypothesis for training set filtering is by explicitly eliminating those highly uncertain data samples from the set and optimizing only on the rest of the "confident" samples using the evidential focal loss (Section 4.2.2), we could produce a more robust model. Therefore, to coalesce our proposed loss function with training set filtering, we do not use co-teaching and instead, we select MpMRI and M.S. MpMRI for experiments in this section. We use the MpMRI (resp. M.S. MpMRI) model

	Data	F.R.	F.M.	Acc.	Sen.	Spec.	AUC
SDNet(baseline)	T2	0%	N/A	79.4	28.6	**92.6**	76.2±17.5
T2-only	T2	0%	N/A	64.7	71.4	63.0	77.8±18.0
MpMRI	T2+ADC	0%	N/A	79.4	**85.7**	77.8	84.7±15.5
Vol. MpMRI	T2+ADC	0%	N/A	67.6	71.4	66.7	68.9±26.1
M.S. MpMRI	T2+ADC	0%	N/A	73.5	71.4	74.1	82.5±14.3
MpMRI+co-teaching	T2+ADC	0%	N/A	**82.3**	**85.7**	81.2	**88.4±10.6**
MpMRI	T2+ADC	10%	patch	82.4	85.7	81.5	85.7±9.9
M.S.MpMRI	T2+ADC	10%	patch	82.4	71.4	85.2	83.6±13.5
MpMRI	T2+ADC	20%	patch	**85.3**	**100**	81.5	**98.4±1.6**
M.S. MpMRI	T2+ADC	20%	patch	**85.3**	71.4	**88.9**	92.6±7.4
MpMRI	T2+ADC	10%	patient	**88.2**	**100**	**85.2**	**92.6±7.4**
M.S. MpMRI	T2+ADC	10%	patient	85.3	**100**	81.5	86.2±9.0
MpMRI	T2+ADC	20%	patient	73.5	85.7	70.4	86.8±12.6
M.S. MpMRI	T2+ADC	20%	patient	73.5	71.4	74.1	84.6±14.2

Table 1: **Patient-based results** of all performed experiments. Acc., Sen., Spec., and AUC are the shorts for Accuracy, Sensitivity, Specificity, and AUC, respectively. Standard deviations are computed from the 95% bootstrap confidence interval with $n = 3000$ samples. The "F.R." and "F.M." represents the filtering rate and the filtering method, respectively. The best results for each section are in black **bold**; blue **bold**, and red **bold** separately. All units of the numeric values are in %.

to compute the uncertainty for all training data first, and then the filtering process can be either done in patch-driven or patient-driven on the training set, as we discussed in Section 4.2.2.

In the second part of Table 1, we present the **patient-based results** on filtering 10% and 20% of training *patches* in the training set for the two selected models, while the third part of the table reports the results of filtering 10% and 20% of training *patients*. The results from both sections demonstrate that the binary classification performance improves when filtering more uncertain data for both models. Comparing these results with those from the first section of Table 1, we can conclude that the dataset filtering method applied to the training set, together with the evidential focal loss we proposed, can effectively improve the classification performance.

Moreover, an interesting observation is that the performance gradually deteriorates in *patient-driven* filtering. The reason behind this may be that in patch-driven filtering, we simply exclude some training patches with a high uncertainty value in the training process, no matter which patient the patches belong to. However, in the case of patient-driven filtering, we have to consider the average uncertainty of the 20 patches for each patient. If the average uncertainty of a patient is below the threshold, all the patches would be used for training, regardless of whether a specific patch has a very high uncertainty value. Therefore, there is a risk that we may falsely retain high uncertain patches because the corresponding patient has relatively low uncertainty on average, which can affect the model's performance. This is also the reason why patch-driven filtration results are better.

Ablation Study: As previously mentioned, the first section of Table 1 corresponds to experiments conducted without using evidential focal loss or filtering. On the other hand, the second and third sections encompass experiments that incorporate both these elements. In order to solely examine the influence of our proposed loss, we conducted an experiment where the evidential focal loss was employed without any filtering (0%). This results are summarized in Table 2 in comparison with 20% patch-based filtering approach. As can be seen, even without any data filtering during the training, we could correctly classify all patients with clinically significant PCa (sensitivity = 100%), which demonstrates a significant improvement compared to the baseline result in Grebenisan et al. (2021). As expected, the addition of data filtering further improves the results. The original results based on image patches of Table 2 can be found in Appendix A.

	filter 0%				filter 20%			
	Acc.	Sen.	Spec.	AUC	Acc.	Sen.	Spec.	AUC
MpMRI	82.4	**100**	77.8	89.4±9.1	**85.3**	**100**	**81.5**	**98.4±1.6**
M.S. MpMRI	76.5	**100**	70.4	82.0±12.7	**85.3**	71.4	**88.9**	**92.6±7.4**

Table 2: Ablation on employing proposed evidential focal loss with and without data filtering for the two selected architectures. The **Patient-based** results are reported. Acc., Sen., Spec., and AUC are the shorts for Accuracy, Sensitivity, Specificity, and Area Under (ROC) Curve, respectively. Standard deviations are computed from the 95% bootstrap confidence interval with $n = 3000$ samples. Best results are in **bold**. All units of the numeric values are in %.

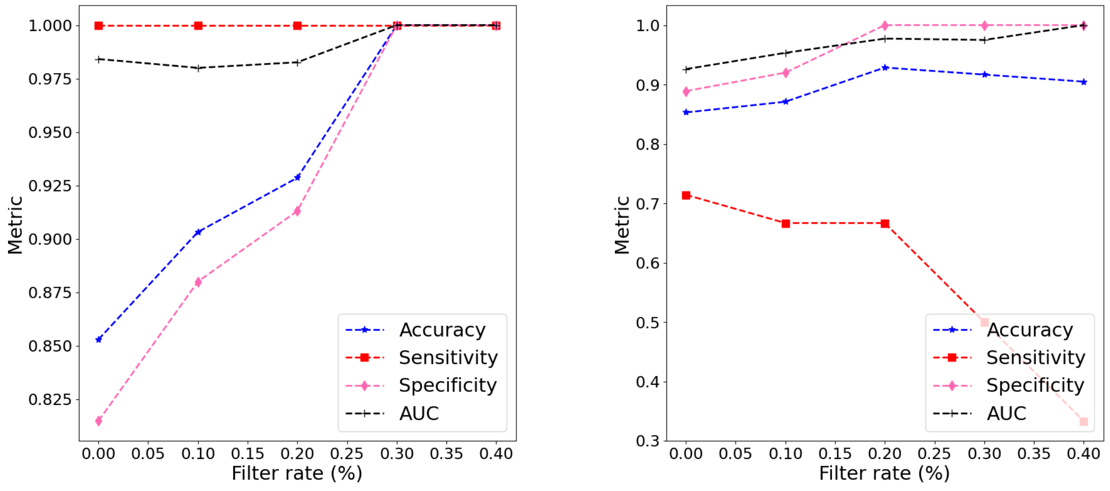
6.4 Filtering during deployment

So far, we explored the effect of the data filtering strategy during training to improve the model robustness. It is also possible to apply filtering on the test set, i.e. when deploying the model to real clinical routines. This is equivalent to refraining from making decisions on the test samples that are identified as highly uncertain. We use the pre-trained MpMRI and M.S. MpMRI models, each with 0% and 20% training filtering rate as final models and evaluate their performance on the test set. Figure 6 shows the experiment on filtering 0% to 40% of the test data when deploying the pre-trained models with 20% filtering during training. The performance of the other two models (0% filtering) can be found in Appendix B. We observe that the model improves its performance when filtering out highly uncertain patients from the test set, and eventually classified all patients correctly, as shown in Figure 6a.

This approach has practical applications under real clinical settings, as it can help radiologists save much time by focusing on patients that have been filtered out (with high uncertainty value) during the diagnostic process, rather than well-classified patients.

7. Conclusion

In this work, we presented a novel approach for unpaired image-to-image translation of prostate mp-MRI and developed a robust deep-learning model for classifying clinically sig-



(a)

(b)

Figure 6: Performance of filtering 0% to 40% of test data on the selected models. 6a represents the test performance of "MpMRI" model with 20% filtration on the training set, and 6b represents the test performance of "M.S. MpMRI" model with 20% filtration on the training set.

nificant PCa using evidential focal loss. We demonstrated the effectiveness of our method on our local dataset, reinforced by a publicly available one, and showed that uncertainty-aware filtering during both training and deployment can significantly improve the PCa classification performance. Out method has the potential to assist with and expedite the diagnostic process by suggesting highly uncertain patients where clinicians can focus on precise diagnosis and fast track those with high prediction certainty.

While our approach has shown promising results, there are still opportunities for improvement. One potential area for future work is to consider the spatial dependency between slices in volumetric MRI. Currently, our domain transfer framework only accepts 2D images as input and output, and we reshape the volumetric MRI into several 2D slices. However, explicitly splitting 3D images into 2D slices may eliminate the spatial dependency within each MRI data and affect the classification results. Therefore, one possible solution is to translate the 3D MRI as a whole from 3.0T to 1.5T instead of translating a single slice at a time.

Lastly, there is great potential for further improving the classification performance by combining more images from different MRI functional sequences, such as b-value and K^{trans}. We have already demonstrated that incorporating additional ADC images significantly enhances classification performance. We believe that if we successfully translate other images from b-value or K^{trans} acquired at 3.0T to 1.5T and incorporate them into the classification, the results could be further improved. However, the additional MRI sequences may not be available in the local 1.5T dataset. The conversion process may become feasible if we acquire those sequences from local hospitals.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Council of Canada, Canadian Institutes for Health Research, and Queen's University. Parvin Mousavi is supported by Canada CIFAR AI Chair and the Vector AI Institute.

Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022.

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

Muhammad Arif, Ivo G Schoots, Jose Castillo Tovar, Chris H Bangma, Gabriel P Krestin, Monique J Roobol, Wiro Niessen, and Jifke F Veenland. Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric mri. *European radiology*, 30(12):6582–6592, 2020.

Karim Armanious, Chenming Jiang, Sherif Abdulatif, Thomas Küstner, Sergios Gatidis, and Bin Yang. Unsupervised medical image translation using cycle-medgan. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

Samuel G Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S Kirby, Nicholas Petrick, George Redmond, Maryellen L Giger, Kenny Cha, Artem Mamonov, et al. Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501, 2018.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

Jelle O Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J Fütterer. Esur prostate mr guidelines 2012. *European radiology*, 22(4):746–757, 2012.

Jelle O Barentsz, Jeffrey C Weinreb, Sadhna Verma, Harriet C Thoeny, Clare M Tempany, Faina Shtern, Anwar R Padhani, Daniel Margolis, Katarzyna J Macura, Masoom A Haider, et al. Synopsis of the pi-rads v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use. *European urology*, 69(1):41, 2016.

A El-Shater Bosaily, C Parker, LC Brown, R Gabe, RG Hindley, R Kaplan, M Emberton, HU Ahmed, PROMIS Group, et al. Promis—prostate mr imaging study: a paired validating cohort study evaluating the role of multi-parametric mri in men with clinical suspicion of prostate cancer. *Contemporary clinical trials*, 42:26–40, 2015.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pages 961–970. PMLR, 2019.

Shuhao Cui, Xuan Jin, Shuhui Wang, Yuan He, and Qingming Huang. Heuristic domain adaptation. *Advances in Neural Information Processing Systems*, 33:7571–7583, 2020.

Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

Robert H. Fletcher. Guideline: Experts recommend against prostate cancer screening with prostate-specific antigen test. *Annals of Internal Medicine*, 170(2):JC2, 2019. doi: 10.7326/ACPJC-2019-170-2-002. URL `https://www.acpjournals.org/doi/abs/10.7326/ACPJC-2019-170-2-002`. PMID: 30641553.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Rongjun Ge, Yuting He, Cong Xia, Chenchu Xu, Weiya Sun, Guanyu Yang, Junru Li, Zhihua Wang, Hailing Yu, Daoqiang Zhang, et al. X-ctrsnet: 3d cervical vertebra ct reconstruction and segmentation directly from 2d x-ray images. *Knowledge-Based Systems*, 236:107680, 2022.

Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017.

Florin C Ghesu, Bogdan Georgescu, Eli Gibson, Sebastian Guendel, Mannudeep K Kalra, Ramandeep Singh, Subba R Digumarthy, Sasa Grbic, and Dorin Comaniciu. Quantifying and leveraging classification uncertainty for chest radiograph assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 676–684. Springer, 2019.

Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Eli Gibson, RS Vishwanath, Abishek Balachandran, James M Balter, Yue Cao, Ramandeep Singh, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855, 2021.

Andrew Grebenisan, Alireza Sedghi, Alexandre Menard, Jason Izard, Robert Siemens, and Parvin Mousavi. Towards democratizing ai in mr-based prostate cancer diagnosis: 3.0 to 1.5 tesla. In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11315, pages 184–189. SPIE, 2020.

Andrew Grebenisan, Alireza Sedghi, Jason Izard, Robert Siemens, Alexandre Menard, and Parvin Mousavi. Spatial decomposition for robust domain adaptation in prostate cancer detection. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1218–1222. IEEE, 2021.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using cyclegan. In *International workshop on simulation and synthesis in medical imaging*, pages 31–41. Springer, 2018.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *European Conference on Computer Vision*, pages 702–715. Springer, 2012.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2018.

Xiaobin Hu. Multi-texture gan: Exploring the multi-scale texture translation for brain mr images. *arXiv preprint arXiv:2102.07225*, 2021.

Gaofeng Huang and Amir Hossein Jafari. Enhanced balancing gan: Minority-class image generation. *Neural Computing and Applications*, pages 1–10, 2021.

Saqib Iqbal, Ghazanfar Farooq Siddiqui, Amjad Rehman, Lal Hussain, Tanzila Saba, Usman Tariq, and Adeel Ahmed Abbasi. Prostate cancer detection using deep learning and traditional techniques. *IEEE Access*, 9:27085–27100, 2021.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018.

Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

Veeru Kasivisvanathan, Antti S Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A Mynderse, Markku H Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G Hindley, et al. Mri-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, 378(19):1767–1777, 2018.

Matthew Kasson, Michael Ortman, Krishnanath Gaitonde, Sadhna Verma, and Abhinav Sidana. Imaging prostate cancer using multiparametric magnetic resonance imaging: past, present, and future. In *Seminars in Roentgenology*, volume 53, pages 200–205. Elsevier, 2018.

Naimul Mefraz Khan, Nabila Abraham, and Marcia Hon. Transfer learning with intelligent training data selection for prediction of alzheimer's disease. *IEEE Access*, 7:72726–72735, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Wouter M Kouw, Marco Loog, Lambertus W Bartels, and Adriënne M Mendrik. Mr acquisition-invariant representation learning. *arXiv preprint arXiv:1709.07944*, 2017.

Atsutoshi Kumagai and Tomoharu Iwata. Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4106–4113, 2019.

Gihyun Kwon, Chihye Han, and Dae-shik Kim. Generation of 3d brain mri using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–126. Springer, 2019.

Mark E Ladd, Peter Bachert, Martin Meyerspeer, Ewald Moser, Armin M Nagel, David G Norris, Sebastian Schmitter, Oliver Speck, Sina Straub, and Moritz Zaiss. Pros and cons of ultra-high-field mri/mrs for human application. *Progress in nuclear magnetic resonance spectroscopy*, 109:1–50, 2018.

Minh Hung Le, Jingyu Chen, Liang Wang, Zhiwei Wang, Wenyu Liu, Kwang-Ting Tim Cheng, and Xin Yang. Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks. *Physics in Medicine & Biology*, 62 (16):6497, 2017.

Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1104–1113, 2021.

Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE transactions on medical imaging*, 33(5):1083–1092, 2014.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Prostatex challenge data. *Cancer Imaging Arch*, 10:K9TCIA, 2017.

Saifeng Liu, Huaixiu Zheng, Yesu Feng, and Wei Li. Prostate cancer diagnosis using deep learning with 3d multiparametric mri. In *Medical imaging 2017: computer-aided diagnosis*, volume 10134, pages 581–584. SPIE, 2017.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1417, 2014.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.

Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *Advances in neural information processing systems*, 30, 2017.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

Alireza Mehrtash, Alireza Sedghi, Mohsen Ghafoorian, Mehdi Taghipour, Clare M Tempany, William M Wells III, Tina Kapur, Parvin Mousavi, Purang Abolmaesumi, and Andriy Fedorov. Classification of clinical significance of mri prostate findings using 3d convolutional neural networks. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101342A. International Society for Optics and Photonics, 2017.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

Alvin I. Mushlin, Cathleen Mooney, Robert G. Holloway, Allan S. Detsky, David H. Mattson, and Charles E. Phelps. The cost-effectiveness of magnetic resonance imaging for patients with equivocal neurological symptoms. *International Journal of Technology Assessment in Health Care*, 13(1):21–34, 1997. doi: 10.1017/S0266462300010205.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention*, pages 417–425. Springer, 2017.

Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12):2720–2730, 2018.

Reda Oulbacha and Samuel Kadoury. Mri to ct synthesis of the lumbar spine from a pseudo-3d cycle gan. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 1784–1787. IEEE, 2020.

Oscar J Pellicer-Valero, José L Marenco Jiménez, Victor Gonzalez-Perez, Juan Luis Casanova Ramón-Borja, Isabel Martín García, María Barrios Benito, Paula Pelechano Gómez, José Rubio-Briones, María José Rupérez, and José D Martín-Guerrero. Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Scientific reports*, 12 (1):1–13, 2022.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Islam Reda, Ashraf Khalil, Mohammed Elmogy, Ahmed Abou El-Fetouh, Ahmed Shalaby, Mohamed Abou El-Ghar, Adel Elmaghraby, Mohammed Ghazal, and Ayman El-Baz. Deep learning role in early diagnosis of prostate cancer. *Technology in cancer research & treatment*, 17:1533034618775530, 2018.

Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. End-to-end prostate cancer detection in bpmri via 3d cnns: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis*, 73:102155, 2021.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ryan P Smith, S Bruce Malkowicz, Richard Whittington, Keith VanArsdalen, Zelig Tochner, and Alan J Wein. Identification of clinically significant prostate cancer by prostate-specific antigen screening. *Archives of internal medicine*, 164(11):1227–1230, 2004.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Armando Stabile, Francesco Giganti, Andrew B Rosenkrantz, Samir S Taneja, Geert Villeirs, Inderbir S Gill, Clare Allen, Mark Emberton, Caroline M Moore, and Veeru Kasivisvanathan. Multiparametric mri for prostate cancer diagnosis: current status and future directions. *Nature reviews urology*, 17(1):41–61, 2020.

Mickael Tardy, Bruno Scheffer, and Diana Mateus. Uncertainty measurements for the reliable classification of mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 495–503. Springer, 2019.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076, 2015.

T Ullrich, M Quentin, C Oelers, F Dietzel, LM Sawicki, C Arsov, R Rabenalt, P Albers, G Antoch, D Blondin, et al. Magnetic resonance imaging of the prostate at 1.5 versus 3.0 t: A prospective comparison study of image quality. *European journal of radiology*, 90: 192–197, 2017.

Hristina Uzunova, Jan Ehrhardt, and Heinz Handels. Memory-efficient gan-based domain translation of high resolution 3d medical images. *Computerized Medical Imaging and Graphics*, 86:101801, 2020.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

Zhen Wang, Guosheng Hu, and Qinghua Hu. Training noise-robust deep neural networks via meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4524–4533, 2020.

Jeffrey C Weinreb, Jelle O Barentsz, Peter L Choyke, Francois Cornud, Masoom A Haider, Katarzyna J Macura, Daniel Margolis, Mitchell D Schnall, Faina Shtern, Clare M Tempany, et al. Pi-rads prostate imaging–reporting and data system: 2015, version 2. *European urology*, 69(1):16–40, 2016.

Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3964–3973, 2018.

Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer, 2008.

Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2272–2281, 2017.

Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2142–2150, 2015.

Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10628, 2019.

Sunghwan Yoo, Isha Gujrathi, Masoom A Haider, and Farzad Khalvati. Prostate cancer detection using deep convolutional neural networks. *Scientific reports*, 9(1):1–10, 2019.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.

Bin Yuan, Xiaodong Yue, Ying Lv, and Thierry Denoeux. Evidential deep neural networks for uncertain data classification. In *International Conference on Knowledge Science, Engineering and Management*, pages 427–437. Springer, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

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

Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *European Conference on Computer Vision*, pages 800–815. Springer, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Appendix A. Original Results

In this section, we provide all original results based on image patches (Patch-based results) of all experiments we performed in the paper. See Table 3 for the main results.

	Data	Rate	Method	Acc.	Sen.	Spec.	AUC
SDNet(baseline)	T2	0%	N/A	77.8	27.9	**90.7**	74.8±4.1
T2-only	T2	0%	N/A	65.6	**75.7**	63.0	77.9±3.9
MpMRI	T2+ADC	0%	N/A	75.1	69.3	76.7	79.9±3.8
Vol. MpMRI	T2+ADC	0%	N/A	64.9	62.9	65.4	63.4±5.5
M.S. MpMRI	T2+ADC	0%	N/A	70.4	74.3	69.4	80.1±3.6
MpMRI+co-teaching	T2+ADC	0%	N/A	**78.4**	72.1	80.0	**80.7±3.9**
MpMRI	T2+ADC	10%	patch	77.5	**80.7**	76.7	82.4±3.4
M.S.MpMRI	T2+ADC	10%	patch	79.7	70.7	82.0	81.1±3.8
MpMRI	T2+ADC	20%	patch	**83.8**	80.0	**84.8**	**89.7±2.6**
M.S. MpMRI	T2+ADC	20%	patch	82.4	73.6	84.6	87.6±2.8
MpMRI	T2+ADC	10%	patient	75.3	**80.7**	73.9	**86.1±2.9**
M.S. MpMRI	T2+ADC	10%	patient	75.6	75.6	**76.9**	82.4±3.6
MpMRI	T2+ADC	20%	patient	72.5	74.3	72.0	81.5±3.6
M.S. MpMRI	T2+ADC	20%	patient	**76.3**	74.3	**76.9**	80.1±4.2

Table 3: **Patch-based results** of all performed experiments. Acc., Sen., Spec. and AUC are the shorts for Accuracy, Sensitivity, Specificity and Area Under (ROC)Curve, respectively. Standard deviations are computed from the 95% bootstrap confidence interval with $n = 3000$ samples. The "Rate" column shows the filtering rate, and the "Method" column represents the filtering method. Best results for each section are in black **bold**; blue **bold**, and red **bold** separately. All units of the numeric values are in %.

Table 4 shows the original results based on image patches for the ablation study conducted in Section 6.3.

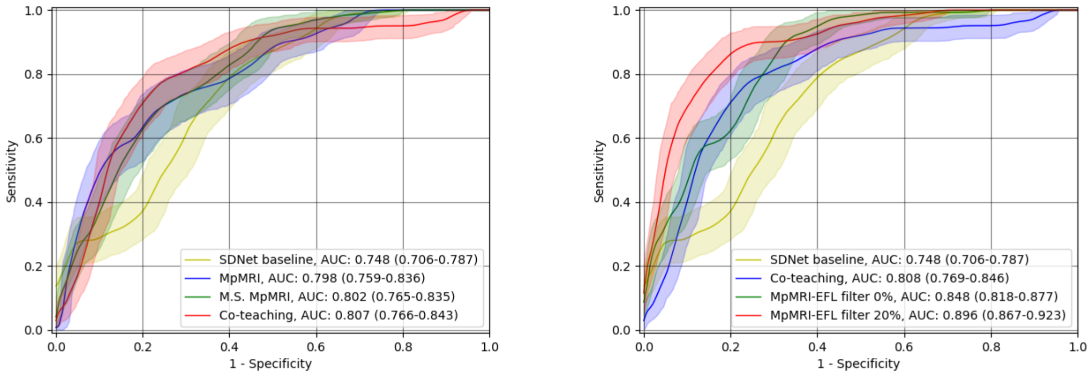
	filter 0%				filter 20%			
	Acc.	Sen.	Spec.	AUC	Acc.	Sen.	Spec.	AUC
MpMRI	74.1	79.3	72.8	**84.8±2.9**	83.8	80.0	84.8	**89.7±2.6**
M.S. MpMRI	73.7	86.4	70.4	80.4±3.1	82.4	73.6	84.6	87.6±2.8

Table 4: Patch-based results for disable filtration on the training set for two models.

we also provide the visualization of patch-based AUC curves for the selected experiments in Section 6.2 and 6.3, along with the 95% confidence interval against the baseline model in Figure 7.

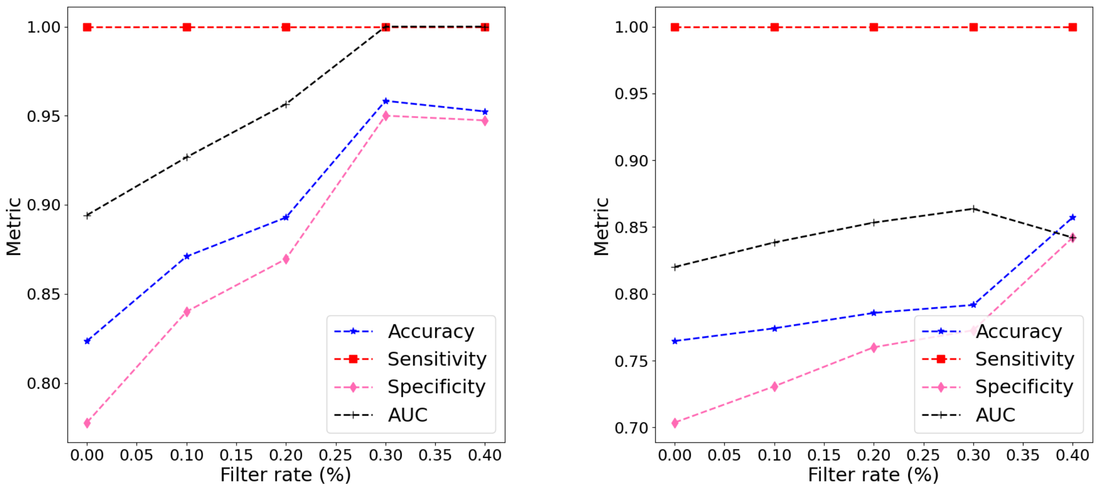
Appendix B. Filtration while deploying

Next, we provide the performance for test set filtering using pre-trained MpMRI and M.S. MpMRI with 0% filtering rate on the training set in Figure 8.



(a) (b)

Figure 7: Both figures demonstrate the original AUC results. 7a shows the comparison of AUC curves between the baseline and the models without filtration (experiments in the first category); 7b shows the comparison of AUC curves between the baseline, the best model without filtration, and the best model with filtration on the training set. “EFL” is short for Evidential Focal Loss. The shaded areas in both figures represent the 95% confidence intervals (CI) of each model. CIs are obtained by using Bootstrap with $n = 3000$.



(a) (b)

Figure 8: Performance of filtering 0% to 40% of test data on the selected models. 8a represents the test performance of “MpMRI” model with 0% filtration on the training set, and 8b represents the test performance of “M.S. MpMRI” model with 0% filtration on the training set.