ABSTRACT
Multi-zonal segmentation is a critical component of computer-aided diagnostic systems for detecting and staging prostate cancer. Previously, convolutional neural networks such as the U-Net have been used to produce fully automatic multi-zonal prostate segmentation on magnetic resonance images (MRIs) with performance comparable to human experts, but these often require large amounts of manually segmented training data to produce acceptable results. For institutions that have limited amounts of labeled MRI exams, it is not clear how much data is needed to train a segmentation model, and which training strategy should be used to maximize the value of the available data. This work compares how the strategies of transfer learning and aggregated training using publicly available external data can improve segmentation performance on internal, site-specific prostate MR images, and evaluates how the performance varies with the amount of internal data used for training. Cross training experiments were performed to show that differences between internal and external data were impactful. Using a standard U-Net architecture, optimizations were performed to select between 2D and 3D variants, and to determine the depth of fine-tuning required for optimal transfer learning. With the optimized architecture, the performance of transfer learning and aggregated training were compared for a range of 5-40 internal datasets. The results show that both strategies consistently improve performance and produced segmentation results that are comparable to that of human experts with approximately 20 site-specific MRI datasets. These findings can help guide the development of site-specific prostate segmentation models for both clinical and research applications.

INDEX TERMS
Automatic prostate segmentation, magnetic resonance imaging, convolutional neural networks, 3D U-Net.

I. INTRODUCTION
Prostate cancer is the second most prevalent cancer diagnosed in men, accounting for approximately 1.3 million cases per year worldwide [1]. The majority of cases progress slowly and require only active surveillance, but others require aggressive treatment that poses the risk of serious side effects [2]. Segmenting the prostate into the distinct regions of peripheral zone (PZ) and central gland (CG), here defined as the entire non-PZ region of the prostate, would provide important information for planning biopsies and focal therapies, while improving cancer detection performance. Cancer identified within the PZ typically differs in prevalence and prognosis from cancer diagnosed within the CG [3], [4]. By incorporating this region-wise information in computer-aided diagnostic systems employing predictive models for cancer detection based on multiparametric MRI, performance can be improved [5], [6]. However, to effectively use these models prospectively, accurate segmentation of an individual’s prostate is required.
Manual segmentation of the prostate requires time and expertise: radiologists typically require 5-10 minutes to segment the whole prostate and ~15 minutes to perform two-zone (PZ & CG) segmentation [7], [8] from typical magnetic resonance images (MRI). Thus the development of computational methods to automate the segmentation process has long been a topic of research. In recent years, automated methods employing convolutional neural networks (CNNs) have shown the best performance, with segmentation accuracy comparable to that of human experts. Training these networks requires large, manually-segmented training datasets containing dozens or even hundreds of cases [9], [10]. Furthermore, a trained network will generally not perform well on images acquired with acquisition parameters different from those used in the training set, and it is common for institutions to customize their acquisitions. The need for a large, labeled corpus of training data with matching acquisition parameters limits the generalizability of CNN-based segmentation methods in research and clinical practice.

In this work we consider the problem of applying a CNN-based method for automatically segmenting prostate MR images that were generated with site-specific acquisition parameters. We refer to these site-specific acquisitions as internal data. If the site has a large number of internal image sets available, they could manually label a subset of the images that could then be used to train an CNN to be applied to the remaining images. If only a small number of internal image sets are available, it is possible to take advantage of publicly available datasets, measured with different acquisition parameters, that have already been manually segmented. There are two well-established strategies for utilizing these external image sets: transfer learning, in which a model is trained first on the external data and fine-tuned with internal data, and aggregated training, which combines internal and external data to train a single model. While these methods have both been used extensively in the literature it is not clear which approach performs better, and if the choice depends on the amount of internal data available. This work seeks to determine which strategy for incorporating external data sets is most appropriate, and to determine how many labelled internal data sets are needed to give acceptable performance.

This manuscript is structured as follows: Section II discusses relevant research on prostate segmentation, including several recent deep learning techniques as well as traditional methods. Section III describes the multi-institutional, heterogeneous datasets used in this work, which includes two publicly available datasets as well as one internal dataset. Section IV details the neural network architecture and general training procedures. Section V presents a series of experiments evaluating architectural and training choices and evaluating the consistency of three datasets. The experiments culminate with a performance comparison of three training strategies (independent training, transfer learning, aggregated training) evaluated on a variable number of internal cases (5-40) to assess the impact of the internal data size on the relative performance of the different strategies. The paper concludes with discussion and conclusions in Sections VI and VII.

II. RELATED WORK
A. STATE OF THE ART SEGMENTATION METHODS
In recent years, segmentation methods using CNNs have surpassed the performance of traditional methods in automatic prostate segmentation, as demonstrated in the PROMISE12 challenge in which participants developed algorithms to segment the whole prostate in multi-institutional MRI data [11]. The top 5 results of the original 2012 online challenge were achieved by an automatic active appearance model (Sorenson-Dice score DSC = 0.88), automatic hierarchical discriminative learning method (0.87), an interactive Smart Paint method (0.87), a semi-automated contour optimization method (0.84), and semi-automated active appearance models (0.82) [11]–[16]. However, the leaderboard is constantly updated, and today the current top 7 PROMISE12 leaders implement CNNs that outperform these traditional methods [17]. These state-of-the-art results essentially match the performance of human experts, as determined by inter-expert variability: a DSC of approximate 0.90 for whole prostate segmentation, 0.75-0.85 for transition zone (TZ), and 0.71-0.80 for PZ [11], [18], [10]. We use this level of performance as a benchmark for evaluating our results herein.

A variety of convolutional architectures have been used for prostate segmentation. Nearly all of these feature encoding and decoding networks connected with skip connections, exemplified by the U-Net architecture [19], [20]. Liu et al. trained a Fully Convolutional Network with Feature Pyramid Attention on 250 MRIs, then tested on both internal and external datasets, achieving PZ DSC of 0.74 and TZ DSC of 0.86 on the internal dataset but only PZ DSC of 0.74 and TZ DSC of 0.79 on the external dataset, indicating room for improvement in the model’s generalizability [18].

Zabihollahy et al. trained U-Nets on 225 T2w and apparent diffusion coefficient (ADC) MRIs to segment the whole prostate and CG. Segmentations on T2w and ADC images achieved a mean DSC for whole prostate of 0.93 and 0.90, CG 0.91 and 0.86, and PZ 0.8622 and 0.83 respectively, suggesting T2w MRIs enable more accurate segmentations than ADC images with this method [8].

Cuocolo et al. used over two hundred T2w MRIs from the PROSTATEX public dataset [21], [22] to train an efficient neural network, Enet, to segment the CG and PZ which slightly out-performed the standard U-Net in spite of containing roughly 10% as many parameters, achieving a DSC of 0.91 on whole prostate, 0.87 on TZ, and 0.71 PZ. In comparison, their implementation of a U-Net achieved DSC of 0.88 on whole prostate, 0.86 on TZ, and 0.70 on PZ [23], [24].

Aldoj et al. implemented a Dense-2U-Net to explore segmentation accuracy of models trained on finely versus coarsely annotated datasets using 188 T2w MRIs [25].
Segmentations predicted on the finely annotated dataset achieved DSC of 0.921 whole prostate, 0.895 CG, and 0.781 PZ, only slightly higher than the segmentations predicted on the coarsely annotated dataset, which lagged behind by a DSC of approximately 0.01 [25].

Nie et al. used an Attention based Semi-supervised Deep Networks (ASDNet) which utilizes unlabeled data in conjunction with confidence maps to predict segmentations with a DSC of 0.911 on an internal validation dataset and 0.90 on the external PROMISE12 dataset [26].

Each of the above-cited architectures use variations of the encoder-decoder strategy, but even a basic U-Net architecture can achieve excellent results. Isensee et al. proposed that the variations in architecture are not as important to overall performance as appropriate hyperparameter tuning [27]. Their no-new U-Net approach uses only a basic U-Net architecture but is currently a top-5 performer on the PROMISE12 leaderboard. Based on their rationale and success, we chose to use a simple U-Net architecture in this work and focus on its performance using limited data.

B. TRAINING WITH LIMITED DATA
The performance of conventionally-trained CNNs decreases with smaller datasets. Meyer et al. explored the effect of corpus size by training a 3D U-Net with a range of 2-78 prostate cases [10]. They found that the greatest improvement in performance occurred when increasing from 8 to 16 training cases, and only slight improvement when greater than 32 training cases were used.

Transfer learning is a well-established technique that can improve performance when only small training datasets are available [28]. This has been previously applied in prostate segmentation. Motamed et al. [29] trained a modified 3D U-Net using only 8 internal cases of diffusion-weighted (DW) MRIs and reported whole prostate segmentations with a DSC of only 0.58. However, when initially trained on over 500 external DW MRIs and then fine-tuned on their 8 internal cases, performance improved, resulting in a DSC of 0.79 for whole prostate segmentation [29]. They also showed that the benefit of transfer learning decreased if more internal data were available. Another example is that of Sanford et al., who used transfer learning to perform two-zone prostate segmentation with CNNs trained and validated on over six hundred single-site T2w MRIs and tested on hundreds of external multi-institutional cases [30]. Their average model achieved a whole prostate DSC of 0.92 and TZ DSC of 0.90 on external data [30].

While transfer learning has been widely used, the specific procedures for fine-tuning the model for the target dataset are not standardized and often not published. Generally, some of the model weights from the initial training are fixed (or frozen), while other model weights are refined with training on the target dataset. In experiment V.C below we explore different levels of tuning to optimize the transfer learning process.

An alternative to transfer learning is aggregated training, which can provide performance comparable to transfer learning and may improve generalizability of the results. Zavala-Romero et al. developed a neural network to perform two-zone prostate segmentation that utilized axial, coronal, and sagittal slices from 550 T2w MRIs acquired from multiple vendors aggregated together (Siemens and GE) [9]. When trained on all 550 cases, their model achieved a mean DSC of 0.89 for whole prostate and 0.81 for PZ on the Siemens dataset and a DSC of 0.83 for whole prostate and 0.79 for peripheral zone on the GE dataset, demonstrating variability in model performance with multi-site data [9]. Rundo et al. trained a single 3D U-Net on a combination of two datasets acquired at different sites, each containing approximately 20 cases of T2w MRIs, and found that aggregating data improved generalizability of their model without sacrificing accuracy when compared to the same model trained on data from one site and applied directly to another [31], without any fine tuning. Liu et al. extended the aggregation strategy by using a novel network architecture to separate site-specific and more general features from a heterogeneous dataset [32]. Their model predicted whole-prostate segmentations with a mean DSC of 0.92, superior to other recently published methods but did not explore multi-zone segmentation [32].

C. CONTRIBUTION OF THIS WORK
The use of CNNs for prostate segmentation is a well-studied topic, in part due to the availability of public datasets. In this work we intentionally do not propose new architectures or training strategies; rather, we seek to determine how to best apply existing techniques for optimal training of two-zone CNN models when a limited amount of training data is available, a practical scenario in both research and clinical practice. This work extends the variable training data exploration of Meyer et al. [10] by adding transfer learning and aggregated training strategies, and comparing their relative performance.

III. DATA
For this study, external data provided as part of the NCI-ISBI2013 Challenge was used in conjunction with an internal dataset [5], [33]–[35]. The first external dataset, A, contains 30 T2w cases acquired at Radboud University Nijmegen Medical Center with a Siemens Trio 3T scanner with an external surface coil [33]. The second external dataset, B, contains 30 T2w cases acquired at Boston Medical Center with a Phillips Achieva 1.5T MRI scanner with both an endorectal coil (ERC) and a surface coil [34]. The third dataset, C, consists of 50 T2w cases acquired internally at the University of Minnesota Center for Magnetic Resonance Research (CMRR) with a Siemens 3T scanner with an ERC as part of an IRB approved research study. All of these 50 internal cases were acquired from participants diagnosed with prostate cancer. Significant inter-individual variation in prostate geometry exists due to the possible presence of cancer and benign proliferative processes. Figure 1, featuring
three cases from dataset C, provides an example showing how the presence of cancer can distort the prostate and complicate regional segmentation. The PZ and CG of all internal cases were manually segmented using ITK Snap [36] by Ph.D. students under the guidance of an expert radiologist with > 50 years of experience.

IV. MODEL ARCHITECTURE AND TRAINING

A. MODEL ARCHITECTURE

This work focuses on the U-Net, a CNN widely used in medical image segmentation due to its high performance on small datasets [19] and its dominant success across a wide range of semantic segmentation challenges in medical imaging [27]. It consists of an encoding branch and a decoding branch, each containing several layers. The initial layer of the encoding branch identifies information from the full resolution image such as texture and edges. At each subsequent layer, the image is downsampled, and the deepest layers provide information about the overall shape and location. The decoding branch utilizes this information to predict segmentations from the MRIs. Both 2D and 3D U-Nets are commonly used in medical image segmentation today. It is not clear apriori which architecture will perform better for a specific problem; therefore in this work we implemented both 2D and 3D U-Nets and compared the accuracy of each to select the preferred architecture for this two-zone prostate segmentation using limited data.

The 3D U-Net in this work contained an encoding arm of four layers, each layer containing the following steps: 3D convolution (with filters $3 \times 3 \times 3$), batch normalization, 3D convolution, batch normalization, and max pooling (with filters $3 \times 3 \times 3$). The fifth layer consisted of only three steps, including a 3D convolution, batch normalization, and a second 3D convolution. Each of the four layers of the decoding arm contained a 3D convolutional transpose (with filters $3 \times 3 \times 3$), concatenation, 3D convolution, batch normalization, and 3D convolution. Soft max activation was the final layer. A diagram of both the 2D and 3D U-Nets is seen in Figure 2. The 2D U-Net mimics the 3D U-Net described above, with corresponding 2D components replacing 3D components, as depicted in Figure 2. These models were implemented with Python 3.7, Keras 2.3.0-tf, and TensorFlow 2.2.0 backend on a 24 GB Titan RTX GPU with CUDA version 11.0.

B. PRE-PROCESSING & AUGMENTATION

All data were resampled to $256 \times 256 \times 19$ and cropped to $192 \times 192 \times 19$. Bias field correction was applied to dataset C with N4ITK [37], [38]. Each dataset was independently normalized prior to use in the model, and data augmentation (rotation, translation, scaling, and left-right flip) was performed [39]. As described in the original U-Net paper, augmentation is used to improve performance of models trained on small datasets [19]. Augmentation has been shown specifically to improve accuracy of prostate segmentation by an average DSC of 0.022 for whole prostate and 0.03 for the CG [30]. Differences in patient positioning within a MR scanner result in slight differences of the prostate’s location within an image, which may be captured by translation and rotation commonly used in augmentation. Scaling reflects variation in prostate size, and left-right flip reflects the symmetry of the prostate. Augmenting the dataset trains the model to adapt to these variations, improving performance.

C. CROSS-VALIDATION

Five-fold cross validation was implemented for all training strategies, with 80% of the cases used as training data and the remaining 20% used for validation. When fewer than 40 internal cases were used in training, the model was trained on a randomly chosen subset of the training data, leaving validation data untouched. Regardless of the number of cases the model was trained on, 20% of the entire dataset was used as validation data for each of the five folds, and segmentations predicted for each validation case were saved. Performance on datasets A, B, and C was evaluated on all cases from each

FIGURE 1. Examples of central slices (halfway between apex and base) of prostate from cases containing the minimum, average, and maximum total volumes of prostate cancer found in the internal dataset C. Left: 0.06 cm$^3$ in PZ, Center: 2.00 cm$^3$ in PZ, Right: 14.38 cm$^3$ of cancer in both the PZ & CG.
S. L. Saunders et al.: Training Convolutional Networks for Prostate Segmentation With Limited Data

FIGURE 2. Diagram of the 2D and 3D U-Net architectures, with the legend for each component on the bottom left. Left: 2D U-Net. MR images and ground truth segmentations of size 192 × 192 are input into the U-Net and down-sampled in each subsequent layer, as indicated by the dimensions in the figure. Convolutions and transpose convolutions each consisted of a kernel size (3,3), with the number of filters used in each indicated in the above diagram. Max pooling consisted of a pool size of (3,3) with strides (2,2). Right: 3D U-Net. MR images and ground truth segmentations of size 19 × 192 × 192 are input into the U-Net and down-sampled in each subsequent layer, as indicated by the shape. Convolutions and transpose convolutions each consisted of a kernel size (3,3,3), with the number of filters used in each indicated in the above diagram. Max pooling consisted of a pool size of (3,3,3) with strides (1,2,2) to accommodate the relatively large slice thickness compared to axial resolution.

TABLE 1. Segmentation accuracy of 2D v. 3D U-Nets trained on 5 and 40 cases of single-site data. Mean (standard deviation) is shown for all cases.

| Dataset C (CMRR): 5 training cases | Dataset C (CMRR): 40 training cases |
|------------------------------------|-------------------------------------|
| 2D U-Net                          | 3D U-Net                            |
|                                    |                                     |
| DSC                                | DSC                                 |
| PZ 0.41 (0.14)                     | PZ 0.73 (0.11)                      |
| CG 0.64 (0.14)                     | CG 0.82 (0.06)                      |
| Whole 0.66 (0.13)                  | Whole 0.87 (0.05)                   |
|                                    |                                     |
| 95% HD (mm)                        | 95% HD (mm)                         |
| 6.73 (2.37)                        | 2.31 (3.77)                         |
| 4.06 (2.09)                        | 1.56 (0.66)                         |
| 5.82 (2.40)                        | 1.69 (3.22)                         |

D. LOSS METRICS

In training, the loss was measured by average DSC of the PZ, CG, and background. The Adam optimizer (learning rate of 1e-04) was used, and the model trained for a maximum of 500 epochs, with early stopping set to 100 [40]. Performance was evaluated on the validation data for all training strategies using the 3D DSC and 95% 2D Hausdorff Distance (95% HD) using the seg-metrics package [41], [42].

V. EXPERIMENTS

Four sets of experiments are described below. The first three provide guidance on the choice of model architecture and training strategies, whereas the fourth experiment provides the primary comparison between training strategies for varying data set sizes.

A. 2D VS. 3D U-Net

The first experiment sought to determine whether a 2D or 3D U-Net architecture was more optimal for this segmentation task. The comparison between the performance of these two architectures is shown in Table 1 for both the minimum and maximum number of cases. With the maximum number of cases (n = 40), the 3D U-Net slightly outperforms the 2D U-Net for each zone and for the whole prostate by an average DSC of 0.03, but with fewer cases the 3D model outperforms the 2D architecture by an average DSC of 0.18. Based on these findings the 3D architecture was used for subsequent experiments because it is more robust to small sample size.

B. DIRECT TRANSFER LEARNING

The next set of experiments evaluated the performance of training a model from one data set and directly applying that model to a different data set, omitting the fine-tuning step commonly used in transfer learning. Such experiments are
FIGURE 3. A 3D U-Net trained on A, B, and C (all 40 cases) individually as well as the combination of A and B was tested on validation data of each of A, B, and C without further training, as described by the independent training and direct transfer learning methods. Mean DSC for the PZ, CG, and Whole prostate are shown.

C. FINE-TUNING FOR TRANSFER LEARNING

This experiment sought to optimize the transfer learning process by experimentally determining the degree of fine-tuning needed to give the best performance. For this assessment, the initial model is the aggregate A+B model trained in the previous experiment, evaluated on 50 cases from dataset C. This represents a limiting case scenario for transfer learning, in which there are multiple public datasets available but minimal internal data.

As is common practice for transfer learning with U-Nets, the encoding layers were frozen and not retrained on the new dataset, but various portions of the decoding layers of the U-Net (the last 50%, 25%, and 12.5% of the entire model) were retrained on the new data. Results of these experiments are shown in Table 2. Due to the superiority of its 95% HD and similarity of its DSC to other models, the model in which only the last 25% was fine-tuned was used for the remainder of the transfer learning experiments in this work.

D. COMPARISON OF THREE TRAINING STRATEGIES: INDEPENDENT TRAINING, TRANSFER LEARNING, AND AGGREGATED TRAINING

The final experiment compares the three training strategies – independent training using only internal data with five-fold cross validation, transfer learning from external data with fine tuning, and aggregated training using both internal and external data. Each of these three strategies was evaluated on a range of 5-40 cases of internal data (site C) to measure how the data size impacts final performance and whether the optimal strategy depends on the size of the internal data set.

For transfer learning, a model was trained using all 60 cases of external data (A and B) and fine-tuned on the selected samples from internal data (C).

For aggregated training, all 60 cases of external data were combined with the selected cases from the internal data for training. Performance between the three strategies was assessed using the DSC and 95% HD metrics, and analyzed using the Wilcoxon signed-rank test at a significance level of \( p < 0.05 \).

Table 3 shows the results comparing all three strategies on the minimum and maximum numbers of internal cases. When all 40 cases are used, all three strategies give similar performance, with transfer learning and aggregated training only slightly outperforming independent training.

Figure 4 plots the effect of dataset size on the relative performance of these three strategies. With few cases of internal data, both transfer learning and aggregated training greatly

useful for assessing the similarity between datasets, determining the need for fine-tuning for transfer learning, and establishing baseline performance metrics for comparing with the more optimized models described below.

Using the 3D U-Net architecture, three models were trained on data from each of sites A, B, and C independently. A fourth model was trained using an aggregation of sites A and B (the external data sets). All four models were subsequently evaluated on the validation subsets for all three sites.

The performance of these evaluations is shown in matrix form in Figure 3. This matrix shows the performance of both independent training (training and testing on data acquired at a single institution, e.g., train on A, assess on A) and direct transfer learning without fine tuning (e.g., train on A, assess on B). The independent training results were generally good; for dataset C the results are comparable to human expert performance. The direct transfer learning gave poorer results, attributable to the differences in data acquisition between sites. Training with combined external datasets A and B and directly assessing on C was better than using either A or B alone for training, but all of these options underperformed independent training for dataset C. These results showed that the site-specific differences were substantial, and fine-tuning of transfer learning models would be required to improve performance over simple independent training.
TABLE 2. Results of three models trained first on public datasets A & B, then fine-tuned with transfer learning using only 5 cases of internal data from C. Mean (standard deviation) is shown for models in which fine-tuning occurred solely on the last 50%, 25%, and 12.5% of the U-Net.

|          | 50% DSC | 95%HD (mm) | 25% DSC | 95%HD (mm) | 12.5% DSC | 95%HD (mm) |
|----------|---------|------------|---------|------------|-----------|------------|
| PZ       | 0.72 (0.11) | 2.46 (3.57) | 0.73 (0.11) | 2.09 (1.17) | 0.74 (0.08) | 1.95 (1.29) |
| CG       | 0.83 (0.06) | 2.32 (3.96) | 0.83 (0.05) | 1.73 (0.93) | 0.83 (0.06) | 1.98 (2.35) |
| Whole    | 0.88 (0.04) | 1.90 (3.73) | 0.88 (0.04) | 1.36 (0.68) | 0.88 (0.03) | 1.55 (1.74) |

TABLE 3. Comparison of models trained using 5 and 40 cases from (C) with only internal data, transfer learning, and with aggregated data. Mean (standard deviation) is shown for all cases. Bold numbers indicate the strategy with the lowest 95%HD.

|          | Single Internal Site | Transfer Learning | Aggregated Public and Internal Data |
|----------|----------------------|-------------------|-------------------------------------|
| C: 5 cases | DSC | 95%HD (mm) | DSC | 95%HD (mm) | DSC | 95%HD (mm) |
| PZ       | 0.65 (0.12) | 2.67 (1.50) | 0.73 (0.11) | 2.09 (1.17) | 0.73 (0.08) | 1.94 (1.02) |
| CG       | 0.77 (0.08) | 2.54 (1.27) | 0.83 (0.05) | 1.73 (0.93) | 0.83 (0.05) | 1.70 (0.79) |
| Whole    | 0.83 (0.06) | 2.14 (1.23) | 0.88 (0.04) | 1.36 (0.68) | 0.88 (0.03) | 1.27 (0.47) |
| C: 40 cases | DSC | 95%HD (mm) | DSC | 95%HD (mm) | DSC | 95%HD (mm) |
| PZ       | 0.76 (0.07) | 1.51 (0.84) | 0.77 (0.07) | 1.56 (1.21) | 0.77 (0.08) | 1.41 (0.81) |
| CG       | 0.85 (0.04) | 1.42 (0.78) | 0.85 (0.04) | 1.29 (0.45) | 0.85 (0.04) | 1.39 (0.73) |
| Whole    | 0.89 (0.03) | 1.15 (0.65) | 0.90 (0.02) | 1.08 (0.42) | 0.90 (0.03) | 1.07 (0.46) |

FIGURE 4. Comparison PZ, CG, and whole prostate segmentation accuracy as measured by mean DSC and 95% HD across all validation data for independent, transfer learning, and aggregated methods with respect to number of internal cases used in training. Higher DSC values correspond to more accurate segmentation, in contrast to the 95% HD, in which lower values indicate more accurate segmentation. Statistically significant differences in independent v. transfer learning, independent v. aggregated method, and transfer learning v. aggregated method models as determined by the Wilcoxon signed-rank test (p < 0.05) are indicated with α, Δ, and +, respectively, placed below the legend.

improved the model performance compared to training only on internal data. The differences between independent training and either of the other two methods was significant with both DSC and 95% HD metrics for most of the comparisons. The performance difference between transfer learning and aggregation was small: no significant differences found with
DSC, whereas with 95% HD several tests showed a modest but significant performance advantage for aggregated training in the PZ.

These plots quantify the benefit of transfer learning and aggregated training strategies. Looking across all 6 metrics (DSC and 95% HD, in PZ, CG, and whole prostate), the performance of 5 cases with aggregated training is comparable to ~15-20 cases without; 20 cases with aggregated training is comparable to ~40 without. Transfer learning provides similar improvements. Using either strategy, the model trained with only 20 internal cases has performance comparable to that of human experts.

Though the DSC metric is commonly used to judge segmentation accuracy, a comparison of DSC and 95% HD reveals that the 95% HD is more sensitive to small changes in segmentation accuracy, as seen in Figure 4. Additionally, since the 95% HD is a distance metric and invariant of surface area to volume ratios, unlike the DSC, it more uniformly judges segmentation accuracy across regions of different volumes, such as the PZ and CG.

Examples of segmentations produced by independent training, transfer learning, and aggregated training with both 5 and 40 internal cases are shown in Figure 5. For each method, a central prostate slice from a single subject that most accurately represented the entire dataset as measured by DSC and 95% HD is shown. Additionally, a video, suppl1_prostate.mov, containing the entire prostate and segmentationspredicted by each of these methods can be found in the supplemental material.

### VI. DISCUSSION

Neural networks such as the U-Net provide a valuable resource for two-zone prostate segmentation but typically require a large amount of data to achieve optimal performance when trained on internal data alone. Due to the heterogeneity of acquisition methods, such as the use of an ERC versus surface coil, models trained on external data did not directly transfer well to internal data. However, it is possible to take advantage of external datasets to achieve state-of-the-art segmentation performance with limited internal data using methods such as transfer learning and training with aggregated data. Our transfer learning and aggregate models both achieved a mean DSC of 0.90 for whole prostate, 0.77 for PZ, and 0.85 for CG when trained on 40 internal cases, comparable to results shown by Litjens et al., Liu et al., and Meyer et al. [11], [18], and [10] and similar to the performance of human experts. When trained on only 10 cases, both the transfer learning and aggregate models performed similarly, sacrificing a DSC of only 0.01-0.02 when compared to the respective models trained on 40 cases and far exceeding the performance of a model trained on 10 internal cases alone.

The independent training model contained 6,475,731 parameters and required approximately 15 hours to train on 40 cases. In contrast, the model fine-tuned with transfer learning required roughly 5 hours of training to fine-tune the model with 40 internal cases after 19 hours of training on 60 cases of public data. Similarly, the aggregated method required approximately 24 hours to train on 40 cases. Therefore, if training time is a factor, transfer learning is the most efficient method if a model trained on external data is already available and only fine-tuning is necessary. However, once the models are trained, both can predict segmentation masks on a new 3D MRI in seconds. Our work demonstrates that the use of external data in training a model using either transfer learning or aggregated data can compensate for the lack of large internal data sets, making automatic segmentation more accessible to groups without the significant resources necessary to acquire and segment hundreds of MR scans often used to train neural networks.

Future work includes use of this automated two-zone prostate segmentation in an algorithm to detect and diagnose prostate cancer. Since prevalence and presentation of prostate cancer varies within different regions of the prostate [3], [4], these segmentations will identify the location of the cancer, potentially improving diagnosis. Additionally, we plan to analyze segmentation accuracy with relation to prostate cancer size, location, and staging.
In this study we optimized a U-Net CNN for two-zone prostate segmentation and compared the performance improvement with transfer learning and aggregated training using external, publicly available data. We found that transfer learning and aggregated training performed similarly, reducing the number of internal training cases required, and can produce performance comparable to human experts with only 20 training cases.

COMPLIANCE WITH ETHICAL STANDARDS

Internal data used in this research was acquired as part of an IRB-approved research study at the University of Minnesota, Center for Magnetic Resonance Research.

ACKNOWLEDGMENT

The authors have no conflicts of interest.

REFERENCES

[1] P. Rawla, “Epidemiology of prostate cancer,” World J. Oncol., vol. 10, no. 2, pp. 63–89, 2019, doi: 10.14740/wjon191.
[2] M. S. Litwin and H. Tan, “The diagnosis and treatment of prostate cancer: A review,” J. Amer. Med. Assoc., vol. 317, no. 24, pp. 2532–2542, Jun. 2017, doi: 10.1001/jama.2017.7248.
[3] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, S. S. Raman, and K. Sung, “Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention,” IEEE Access, vol. 7, pp. 163626–163632, 2019, doi: 10.1109/ACCESS.2019.2952534.
[4] O. Ronneberger, P. Fischer, and T. Brox, “UNet: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI) ( Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Munich, Germany: Springer, 2016, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[5] G. Litjens, O. Bentsen, N. Karssemeijer, and H. Huisman, “ProstateX challenge data,” 2017, doi: 10.7937/K9TYC.2017.MRUS5CL.
[6] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, “Computer-aided detection of prostate cancer in MRI,” IEEE Trans. Med. Imag., vol. 33, no. 5, pp. 1083–1092, May 2014, doi: 10.1109/TMI.2014.2303821.
[7] A. Ushinsky, M. Bardis, J. Glavvis-Bloom, E. Uchio, C. Chantaduly, M. Nguyent, D. Chow, P. D. Chang, and R. Houshyar, “A 3D-2D hybrid U-Net convolutional neural network approach to prostate organ segmentation of multiparametric MRI,” Amer. J. Roentgenol., vol. 216, no. 1, pp. 111–116, Jan. 2021, doi: 10.2214/AJR.19.22168.
[8] F. Zbibhollah, N. Schieda, S. K. Jeyaraj, and E. Ukwatta, “Automated segmentation of prostate zonal anatomy on T2-weighted (T2W) and apparent diffusion coefficient (ADC) map MR images using U-Nets,” Med. Phys., vol. 46, pp. 3078–3090, Jul. 2019, doi: 10.1002/mp.13550.
[9] O. Zavala-Romero, A. L. Breto, I. R. Zu, Y.-C.-C. Chang, N. Gauthey, A. D. Pra, M. C. Abramowicz, A. Pollack, and R. Stoyanova, “Segmentation of prostate and prostate zones using deep learning,” Strahlentherapie Onkologie, vol. 196, no. 10, pp. 932–942, Mar. 2020, doi: 10.1007/s00066-020-01607-x.
[10] A. Meyer, M. Rakr, D. Schindele, S. Blaschke, M. Schostak, A. Fedorov, and C. Hanses, “Towards patient-individual PI-rad v2 sector map,” CNN for automatic segmentation of prostastic zones from T2-weighted MRI,” in Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI), Apr. 2019, pp. 696–700, doi: 10.1109/ISBI.2019.8759572.
[11] A. Ha. Vargas, O. Akin, T. Franier, D. A. Goldman, K. Udo, K. A. Touijer, J. Jin, L. Zhang, E. Leng, G. J. Metzger, and J. S. Koopmeiners, “Detection of prostate cancer: Quantitative multiparametric MR imaging models developed using registered correlative histopathology,” Radiology, vol. 279, no. 3, pp. 805–816, Jun. 2016, doi: 10.1148/radiol.2015151089.
[12] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, and R. Strand, “Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge,” Med. Image Anal., vol. 18, no. 2, pp. 359–373, Feb. 2014, doi: 10.1016/j.media.2013.12.002.
[13] G. Vincent, G. Guillard, and M. Bowes, “Fully automatic segmentation of the prostate using active appearance models,” in Proc. Med. Image Comput.-Assisted Intervent. Conf. Prostate Segment. Challenge, 2012, pp. 75–81.
[14] N. Birkbeck, J. Zhang, M. Requardt, B. Keifer, P. Gall, and S. K. Zhou, “Region-specific hierarchical segmentation of MR prostate using discriminative learning,” in Proc. Med. Image Comput.-Assisted Intervent. Conf. Prostate Segment. Challenge, 2012, pp. 4–11.
[15] V. Melin, B. Strand, J. Kullberg, R. Nordenskjold, and E. Bengtsson, “Smart paint a new interactive segmentation method applied to MR prostate segmentation,” in Proc. Med. Image Comput.-Assisted Intervent. Conf. Prostate Segment. Challenge, 2012, pp. 1–8.
[16] J. Yuan, W. Qiu, E. Ukwwata, M. Rajehl, Y. Sun, and A. Fenster, “An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior,” in Proc. Med. Image Comput.-Assisted Intervent. Conf. Prostate Segment. Challenge, 2012, pp. 82–89.
[17] F. Maan and F. van der Heijden, “Prostate MR image segmentation using 3D active appearance models,” in Proc. Med. Image Comput.-Assisted Intervent. Conf. Prostate Segment. Challenge, 2012, pp. 44–51.
[18] Y. Liu, G. Yang, S. A. Mirak, M. Hosseiny, A. Azadikhah, X. Zhong, R. E. Reiter, Y. Lee, S. S. Raman, and K. Sung, “Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention,” IEEE Access, vol. 7, pp. 163626–163632, 2019, doi: 10.1109/ACCESS.2019.2952534.
[19] O. Ronneberger, P. Fischer, and T. Brox, “UNet: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI) ( Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Athens, Greece: Springer, 2016, pp. 234–421, doi: 10.1007/978-3-319-46723-8_49.
[20] G. Litjens, M. S. Litwin, and H. Tan, “ProstateX challenge data,” 2017, doi: 10.7937/K9TYC.2017.MRUS5CL.
[21] A. Castaldo, D. R. De Lucia, A. Yezzi, and M. Imbriaco, “Deep learning whole-gland and zonal prostate segmentation on a public MRI dataset,” J. Magn. Reson. Imag., vol. 54, no. 2, pp. 452–459, Feb. 2021, doi: 10.1002/jmri.27585.
[22] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A deep neural network architecture for real-time semantic segmentation,” 2016, arXiv:1606.02147. [Online]. Available: https://arxiv.org/abs/1606.02147.
[23] N. Aloj, F. Biavati, F. Michalek, S. Stober, and M. Dawey, “Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-Net,” Sci. Rep., vol. 10, no. 1, pp. 1–17, Aug. 2020, doi: 10.1038/s41598-020-71080-0.
[24] D. Nie, Y. Gao, L. Wang, and D. Shen, “ASDNet: Attention based semi-supervised deep networks for medical image segmentation,” in Medical Image Computing and Computer Assisted Intervention. Cham, Switzerland: Springer, 2018, pp. 370–378, doi: 10.1007/978-3-030-09937-7_38.
[25] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maer-Hein, “nuVet: A self-configuring method for deep learning-based biomedical image segmentation,” Nature Methods, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: 10.1038/s41592-020-01008-z.
[26] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfer learning: Understanding transfer learning for medical imaging,” 2019, arXiv:1902.07208. [Online]. Available: https://arxiv.org/abs/1902.07208.
SARA L. SAUNDERS received the master’s degree in engineering from Dartmouth College, Hanover, New Hampshire, in 2017. She is currently pursuing the Ph.D. degree in biomedical engineering with the University of Minnesota, Minneapolis. Her interests include image processing and application of deep learning methods to medical imaging, specifically application of MR imaging to the investigation of prostate cancer and non-alcoholic fatty liver disease.

ETHAN LENG received the Ph.D. degree in biomedical engineering. He is currently pursuing the M.D./Ph.D. degree with the University of Minnesota. He is interested in the development of fully-automated machine learning-based models for predicting prostate cancer presence and grade using quantitative mpMRI.

S. Motamed, I. Gujrathi, D. Deniffel, A. Oentoro, M. A. Haider, and F. Khalvati, “A transfer learning approach for automated segmentation of prostate whole gland and transition zone in diffusion weighted MRI,” 2019, arXiv:1909.09541. [Online]. Available: https://arxiv.org/abs/1909.09541

T. H. Sanford, L. Zhang, S. A. Harmon, J. Sackett, D. Yang, H. Roth, Z. Xu, D. Kesani, S. Mehlalvand, R. H. Baroni, T. Barrett, R. Girometti, A. Oto, A. S. Puryosko, S. Xu, P. A. Pinto, D. Xu, B. J. Wood, P. L. Choyke, and B. Turkbey, “Data augmentation and transfer learning to improve generalizability of an automated prostate segmentation model,” Amer. J. Roentgenol., vol. 215, no. 6, pp. 1403–1410, Dec. 2020, doi: 10.2214/AJR.19.22347.

L. Rundo, C. Han, J. Zhang, R. Hataya, Y. Nagano, C. Militello, C. Ferretti, M. S. Nobile, A. Tangherloni, M. C. Gialardi, S. Vitabile, H. Nakayama, and G. Mauri, “CNN-based prostate zonal segmentation on T2-weighted MR images: A cross-dataset study,” in Neural Approaches to Dynamics of Signal Exchanges (Smart Innovation, Systems and Technologies), vol. 151, A. Esposito, M. Faundez-Zanuy, F. C. Morabito, E. Pasero, and Eds. Singapore: Springer, 2020, pp. 269–280, doi: 10.1007/978-981-15-8950-4_25.

Q. Liu, Q. Dou, L. Yu, and P. A. Heng, “MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data,” IEEE Trans. Med. Imag., vol. 39, no. 9, pp. 2713–2724, Sep. 2020, doi: 10.1109/TMI.2020.2974574.

G. Litjens, J. Futterer, and H. Huisman, “Data from prostate-3T,” Cancer Imag. Arch., 2015, doi: 10.7937/K9/TCIA.2015.Q7V5L5.

B. N. Bloch, A. Jain, and C. C. Jaffe, “Data from prostate-diagnosis,” Cancer Imag. Arch., 2015, doi: 10.7937/K9/TCIA.2015.FQOUYVT.

K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” J. Digit. Imaging, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9627-7.

P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, “User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability,” NeuroImage, vol. 31, no. 3, pp. 1116–1128, Jul. 2006, doi: 10.1016/j.neuroimage.2006.01.015.

N. J. Tsitsiou, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, “N4ITK: Improved N3 bias correction,” IEEE Trans. Med. Imag., vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

B. C. Loweck, D. T. Chen, L. Ibañez, and D. Blezdek, “The design of SimpleITK,” Frontiers Neurinform., vol. 7, p. 45, Dec. 2013, doi: 10.3389/fninf.2013.00045.

D. G. Huynh, (2020). Augmented Volumetric Image Generator. [Online]. Available: https://github.com/dhuy228/augmented-volumetric-image-generator/blob/master/LICENSE

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

L. R. Dice, “Measures of the amount of ecologic association between species,” Ecology, vol. 26, no. 3, pp. 297–302, 1945, doi: 10.2307/1932409.

Ordog, “Ordog/segmentation_metrics: Seg-metrics (version v1.0.0),” Zenodo, Aug. 2020, doi: 10.5281/zenodo.3995075.