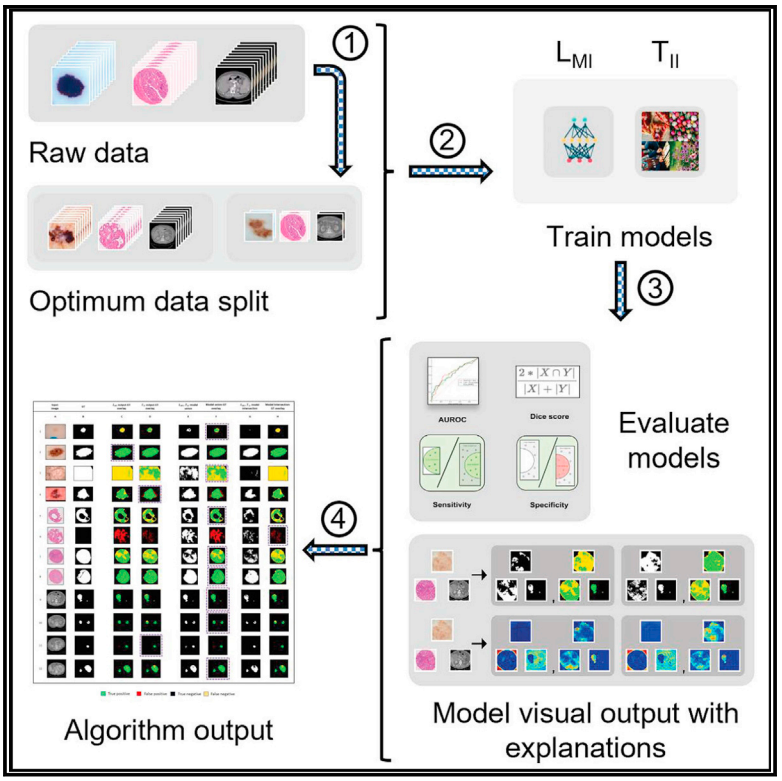
Cell Reports
Methods

Article

A deep-learning toolkit for visualization and interpretation of segmented medical images

Graphical abstract



Authors

Sambuddha Ghosal, Pratik Shah

Correspondence

pratiks@mit.edu

In brief

Ghosal and Shah present a deep-learning toolkit for high-accuracy segmentation of tumors and organs. The toolkit includes a range of analytical techniques for statistical evaluation and visual interpretation of pixel-level segmentation of medical images for reproducibility and performance improvement of deep-learning models.

Highlights

- Python toolkit for visual interpretation of medical image segmentation by deep learning
- Comparison between transfer learning and randomly initialized models
- Algorithm for computing high-accuracy ensemble disease segmentations
- Methods for analyzing datasets and reproducibility of deep-learning models

Ghosal & Shah, 2021, Cell Reports Methods *1*, 100107
November 22, 2021
https://doi.org/10.1016/j.crmeth.2021.100107

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

A deep-learning toolkit for visualization and interpretation of segmented medical images

Sambuddha Ghosal[1] and Pratik Shah[1,2,*]
[1]Program in Media, Arts, and Sciences and Media Lab, Massachusetts Institute of Technology, 20 Ames Street, Cambridge, MA 02139, USA
[2]Lead contact
*Correspondence: pratiks@mit.edu
https://doi.org/10.1016/j.crmeth.2021.100107

MOTIVATION Reliable segmentation of cells or tissues from medical images is a critical need for generalizable deep learning. Random initialization or transfer of model weights from natural-world images, gradient-based heatmaps, and manifold learning have been used to provide insights for image classification tasks. An automated workflow based on statistical reasoning that achieves reproducible medical image segmentation is lacking. The toolkit reported in this study identifies training and validation data splits, automates the selection of medical images segmented with high accuracy, and describes an algorithm for visualization and computation of real-world performance of deep-learning models.

SUMMARY

Generalizability of deep-learning (DL) model performance is not well understood and uses anecdotal assumptions for increasing training data to improve segmentation of medical images. We report statistical methods for visual interpretation of DL models trained using ImageNet initialization with natural-world (T_{II}) and supervised learning with medical images (L_{MI}) for binary segmentation of skin cancer, prostate tumors, and kidneys. An algorithm for computation of Dice scores from union and intersections of individual output masks was developed for synergistic segmentation by T_{II} and L_{MI} models. Stress testing with non-Gaussian distributions of infrequent clinical labels and images showed that sparsity of natural-world and domain medical images can counterintuitively reduce type I and type II errors of DL models. A toolkit of 30 T_{II} and L_{MI} models, code, and visual outputs of 59,967 images is shared to identify the target and non-target medical image pixels and clinical labels to explain the performance of DL models.

INTRODUCTION

Applications of artificial intelligence for automated classification and segmentation of images require large amounts of well-annotated data. Over the past ten years, thousands of research studies (Suzuki, 2017) have reported the optimization of small amounts of available data to conduct research with deep neural networks (DNNs) for medical image classification and segmentation tasks to generate prototypes for real-world applications (Yauney et al., 2017; Rana et al., 2020; Javia et al., 2018). Researchers routinely leveraged the learned weights of pre-trained natural-world deep-learning (DL) models such as AlexNet to fine-tune the use of transfer learning (TL) for medical image segmentation to improve performance (Krizhevsky et al., 2012; Bayat et al., 2021). Visual Geometry Group-16 (VGG16) (Simonyan and Zisserman, 2014), U-Net (Ronneberger et al., 2015), and other DL architectures are routinely used to fine-tune DL models (Raghu et al., 2019). Anecdotal assumptions of superior performance of TL compared with models trained exclusively using medical images (termed as trained from scratch [TFS]) are prevalent despite being poorly understood (Huynh et al., 2016; Näppi et al., 2016). For example, when images were similar and even if the classes were not the same, an augmented or larger dataset or TL is assumed to train DL models with superior accuracy than TFS (Esteva et al., 2017). A recent study reported that TL from natural-world images did not increase the DL model performance for medical image classification tasks (Raghu et al., 2019). Medical images have complex biological textures and several high-gradient regions leading to underspecification of DL models for small perturbations by previously unseen images (Ghorbani et al., 2020; D'Amour et al., 2020). Variances of illumination, image sensor optics, clinical labels, and out-of-training data examples have also been reported to dramatically reduce DL model performance (Finlayson et al., 2019; Paschali et al., 2018). State-of-the-art DNNs designed for large-scale natural image processing are often overparameterized for medical imaging tasks and remain vulnerable to adversarial attacks (Finlayson et al., 2019). Additionally, 90%–95% of TFS and TL models demonstrate underspecification, low sensitivity, and low specificity for medical-grade segmentation precluding their clinical

utility (Chen et al., 2019; Kelly et al., 2019; D'Amour et al., 2020). The use of random 80:20 splits for generating training and validation image splits leads to low-data regimes and over- or under-specification of underlying clinical labels and medical images, resulting in poor reproducibility and replicability of DL models (Collins and Moons, 2019; Shie et al., 2015). Generating interpretable DL models from small amounts of medical image and clinical label data remains challenging for the computational medicine and biological image-processing communities (Chen et al., 2019; Kelly et al., 2019; Shah et al., 2018, 2019). In addition, DL models often memorize rather than learn information and perform poorly when tested with out-of-distribution (OoD) but related classification and segmentation tasks (Zech et al., 2018; Finlayson et al., 2019; D'Amour et al., 2020). Neural network explanation methods (Ghosal et al., 2018, 2019; Simonyan et al., 2013; Pokuri et al., 2019) and image-based visualizations using gradient-weighted class-activation mapping (Grad-CAM) (Selvaraju et al., 2017) have been reported for visualizing and interpreting DL model performance and mechanisms. Thus, methods for generating statistically significant segmentation and visual Grad-CAM explanations for interpretation and generalization of TL and TFS models benchmarked for segmentation of different medical images are valuable for computational rigor and clinically meaningful inferences.

In this study, we report a process and methods for generalizing DL image segmentation using natural-world (T_{II}) models (trained with supervised TL with ImageNet initialization and fine-tuned on medical images) or learning with medical images (L_{MI}) models (trained with supervised learning trained natively on only medical images) for the binary segmentation of tumors and organs. Statistical estimation and visual explanation of the training data and DL model outputs for segmentation of three unique medical image subtypes widely used for computational medicine research were set as targets. Our goal was to support researchers working with small numbers of images and labels in multiple fields to train high-performance DL models and provide resources for their interpretability. We report (1) detailed validation of a process for maximizing available clinical labels and medical images for synergistic use of T_{II} and L_{MI} models; (2) a description of parametric and non-parametric statistical methods to test significance of data distributions; (3) model performance estimation by using the area under the receiver operating curve (AUROC), Dice (F1) score, sensitivity, and specificity of segmentation of target features; (4) estimation and identification of the least numbers and types of individual images and available clinical labels for improvement of model performance; (5) Grad-CAM (Selvaraju et al., 2017) and uniform manifold approximation and projection (UMAP)-based (McInnes et al., 2018) visualization and interpretation of L_{MI} and T_{II} DL models; and (6) an algorithm that automates the comparison of Dice scores by multiple DL models to compute the highest possible accuracy of segmentation of medical image pixels contributing to type I and type II errors. We release 30 (and a larger set of 270 derived from five replicates) fully trained and validated DL models and 11,892 output images from T_{II} or L_{MI} models trained under high- and low-data regimes. This work also communicates specific use cases when L_{MI} and T_{II} models can be used synergistically or in ensemble configurations to improve performance or reduce false-positive and -negative clinical diagnosis. To our knowledge, this is the largest repository of detailed characterizations of medical image segmentation and performance evaluation of DL models, which can be a valuable resource for the community.

RESULTS

After estimation of the optimal 80:20 data split and clinical label distributions for each of the three datasets, the grand median, mean, and standard deviations of distributions of AUROC, Dice scores, sensitivity, and specificity were evaluated for performance estimations of T_{II} and L_{MI} models (Figure 1; Table 1) used in this study. Pairwise differences between medians (Δ_m) of individual performance metrics achieved by either T_{II} or L_{MI} models were calculated for all individual test images in this study (Table 1). We report numbers of images in each dataset that achieved a metric value greater than or equal to 0.9 (a common threshold to indicate superior performance), with 1 indicating a perfect score (Table 1). Approximately 13,000 red/green/blue (RGB) images with binary, benign (n = 12,668), and malignant (n = 1,118) clinical labels (Table S2) of skin cancer diagnoses were used in this study. Transfer-learned T_{II} models achieved higher mean and median Dice, AUROC, sensitivity, and specificity scores for segmentation of both benign and malignant skin cancer lesions from RGB images (Table 1). The T_{II} models also achieved higher mean AUROC (86%) and Dice scores (78%) across all skin cancers (Table 1). Greater numbers of skin images were segmented with 0.9 or higher mean AUROC (n = 1,651 images), Dice scores (n = 1,177), and sensitivity (n = 1,301) by T_{II} models (Table 1). The T_{II} models consistently achieved scores higher than $\approx$0.9 (AUROC) for 1,651 images (59% of total images) compared with 984 images by L_{MI} models for segmentation of any skin cancer (Table 1). Additionally, 40% (1,177) of all skin cancer images were segmented with Dice scores greater than 0.9 by T_{II} models compared with 671 images (22%) by L_{MI} (Table 1). T_{II} models outperformed L_{MI} models with higher and statistically significant (Mood's median test, $p < 0.05$) differences for the segmentation of all benign and malignant skin cancer lesions across 2,758 test images (Figure 2Ai and Aii; Table 1). Binary segmentation using L_{MI} models performed equally well in detecting non-target pixels as transfer-learned models (Table 1). For the data depletion experiments, starting with the selected 80:20 (train:test) split, the training data was depleted to 60:40, 40:60, 20:80 and 10:90 ratios (Figure 1; Table 2 following randomization to enrich different splits.

Visual explanations were used to perform image-based analysis to compare and interpret the differences in L_{MI} and T_{II} model performances reported in Table 1. The T_{II} models demonstrated lower false-negative and higher true-positive regions than the corresponding outputs from L_{MI} models for segmentation of benign (Figure 2A) and malignant skin cancer lesions (Figure 2B). The majority of L_{MI} outputs in column d in Figures 2Ai and 2Biii, on the other hand, demonstrated higher false-negative (yellow) and lower true-positive detection (green) compared with T_{II} models (column f of Figures 2Ai and 2Biii). Grad-CAM analysis showed that L_{MI} models were less capable of distinguishing and had lowered activation between non-target skin pixel areas surrounding boundaries of benign (Figure 2Aii, columns b and c)

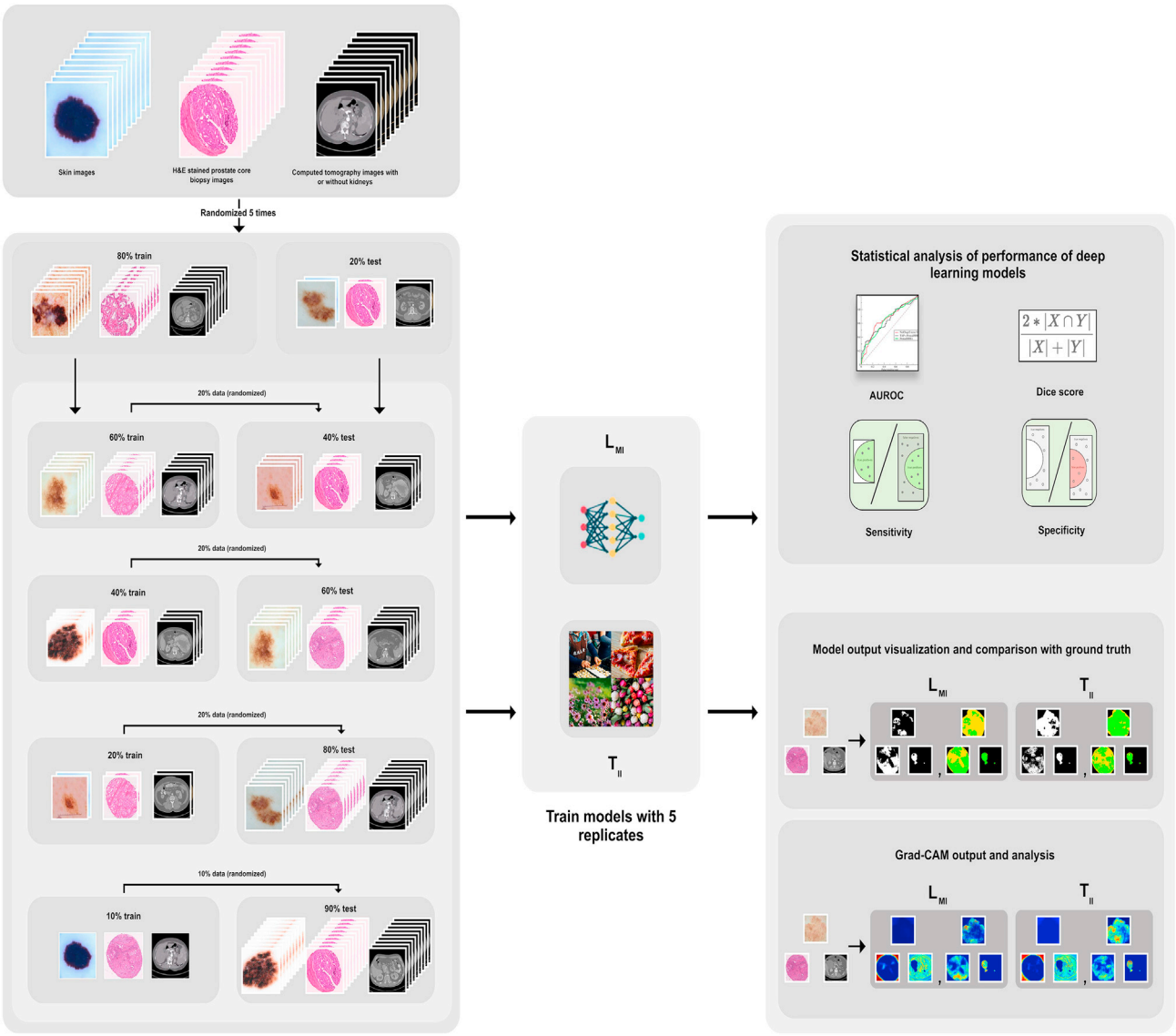


Figure 1. Overview of study design and deep-learning models

Three datasets consisting of macroscopic optical skin (n = 13,786), microscopic RGB prostate core biopsy (n = 244), and CT DICOM images (n = 45,937) were randomized and split five times into different percentages of 80 (training) and 20 (validation). Each of the five 80:20 splits (sets) for these three individual image types were then used for training a VGG-UNet deep-learning model with internal 5-fold repeats by transfer learning with pretrained weights from 14 million natural-world images with ImageNet initialization (T_{II}) or training with only medical images (L_{MI}) in a particular dataset. The resulting five sets of T_{II} (n = 25) and L_{MI} (n = 25) models were then compared using statistical testing of the pixel-by-pixel mean, median, and standard deviations of AUROC, Dice scores, sensitivity (true positive rate), and specificity (true negative rate) of the associated segmentation masks. This process was repeated for estimating deep-learning model performance following depletion of training data to smaller proportions. See also Figure S1; Tables S1 and S2.

and malignant skin lesions (Figure 2Biv, columns b and c). Thus, the L_{MI} model detected and segmented fewer target pixels but did not seem to lower its specificity values (non-target detection capability). Additionally, Grad-CAM outputs showed that the higher segmentation accuracy of T_{II} models was based on their higher and more precise activation and ability to distinguish benign (Figure 2Aii, column e) or malignant (Figure 2Biv, column e) skin lesions from non-target pixels. For larger and diffuse malignant lesions, T_{II} models were not deceived by the red coloration of the non-target skin pixels surrounding the cancer moles (Figure 2Biii, images 1a and 1f) compared with L_{MI} models that had higher false-negative rates (Figure 2Biii, images 1a and 1d) for such tasks. The T_{II} models' better performance can be explained by the higher activation and precise learning of key features of moles and lesions causal for superior binary segmentation. Correspondingly, Grad-CAM activation profiles revealed that T_{II} models were more likely to focus or preferentially activate regions with brighter colors or explicit shapes in benign and

Table 1. Grand median, mean of medians, and standard deviations (SD) calculated from the distributions of area under the receiver operating curve (AUROC), Dice score, sensitivity, and specificity from five replicates achieved by transfer learning (T_{II}) and learning from medical images (L_{MI}) deep-learning models for binary segmentation of RGB images with skin cancer, microscopic H&E-stained prostate core biopsy, and CT of kidneys

	Skin (n_{test} = 2,758)			Prostate core biopsy (n_{test} = 49)			Kidney CT (n_{test} = 9,085)		
	L_{MI}	T_{II}	Percentage (%)	L_{MI}	T_{II}	Percentage (%)	L_{MI}	T_{II}	Percentage (%)
AUROC									
Median	0.8544*	0.9282*†	–	0.9359	0.9337	–	0.9980*	0.9978*	–
Mean	0.8120	0.8826	–	0.9083	0.8991	–	0.9876	0.9871	–
SD	0.1514	0.1277	–	0.0831	0.0852	–	0.0588	0.0587	–
Value > 0.9	984	1,651†	–	34	29	–	3,167	3,163	–
$\Delta_m > 0$	–	–	86	–	–	31	–	–	9
$\Delta_m = 0$	–	–	4	–	–	4	–	–	70
$\Delta_m < 0$	–	–	10	–	–	65	–	–	21
Dice score									
Median	0.8273*	0.8857*†	–	0.9476	0.9325	–	0.9597	0.9598	–
Mean	0.7483	0.8250†	–	0.8858	0.8733	–	0.9509	0.9502	–
SD	0.2169	0.1786	–	0.1536	0.1512	–	0.0657	0.0622	–
Value > 0.9	671	1,177†	–	32	29	–	3,086	3,074	–
$\Delta_m > 0$	–	–	79	–	–	31	–	–	18
$\Delta_m = 0$	–	–	3	–	–	4	–	–	67
$\Delta_m < 0$	–	–	18	–	–	65	–	–	15
Sensitivity									
Median	0.7156*	0.8891*†	–	0.9520*	0.9059*	–	0.9985*	0.9979*	–
Mean	0.6331	0.7944†	–	0.8988	0.8507	–	0.9772	0.9761	–
SD	0.3080	0.2541	–	0.1438	0.1599	–	0.1178	0.1177	–
Value > 0.9	642	1,301†	–	36	27	–	3,116	3,043	–
$\Delta_m > 0$	–	–	86	–	–	4	–	–	9
$\Delta_m = 0$	–	–	5	–	–	4	–	–	70
$\Delta_m < 0$	–	–	2	–	–	92	–	–	21
Specificity									
Median	0.9999*	0.9986*	–	0.9572	0.9761	–	1	1	–
Mean	0.9922	0.9722	–	0.9108	0.9490	–	0.9993	0.9993	–
SD	0.0341	0.0945	–	0.1056	0.0777	–	0.0011	0.0011	–
Value > 0.9	2,707	2,576	–	36	42	–	9,085	9,085	–
$\Delta_m > 0$	–	–	2	–	–	96	–	–	15
$\Delta_m = 0$	–	–	21	–	–	2	–	–	78
$\Delta_m < 0$	–	–	77	–	–	2	–	–	7

The numbers of test images that achieved metric values of 0.9 and higher for each model are shown. The subtraction of $\Delta_m > 0$, < 0, and = 0 for each image were used to calculate the numbers of test images that achieved higher median values by either T_{II} and L_{MI} models. The total number of test images in each dataset is indicated by n_{test}. Statistically significant (Mood's median test, $p < 0.05$) differences between nonparametric distributions are indicated by *. Differences in metric values greater than 5% are indicated by † for the better performing model. A value of 1 indicated a perfect score. Percentage denotes percentage numbers of n_{test}. See also Figures S2 and S3.

malignant lesions (Figure 2Biv, columns a, c, and e). Interestingly, despite less than 100% coverage for the segmentation of malignant skin lesions (Figure 2B), the T_{II} models still achieved higher AUROC and Dice scores (Tables 1 and S3). The lower performance of the T_{II} models for segmentation of malignant tumors may be attributed to the few coarse ground-truth clinical annotations that may have included pixels without tumors. Alternatively, the trained T_{II} models could not comprehensively distinguish malignant moles from the surrounding skin pixels that resembled other cancer lesions in the training data. The presence of artifacts (e.g., stickers) (image 1a of Figure 2A) were identified as non-target pixels by both T_{II} and L_{MI} models across the dataset. Additionally, both models demonstrated comparable specificity and lower false-positive segmentation of the majority of the

Table 2. Low-data regimen experiments for the segmentation of medical images

	Skin		Prostate core biopsy		Kidney CT	
	L_{MI}	T_{II}	L_{MI}	T_{II}	L_{MI}	T_{II}
AUROC						
10%	0.8187*	0.8769*†	0.8730*	0.9158*	0.9967*	0.9963*
20%	0.8160*	0.9126*†	0.9185	0.9237	0.9972*	0.9973*
40%	0.8613*	0.8932*	0.9215	0.9183	0.9976	0.9976
60%	0.8542*	0.9044*†	0.9153	0.8995	0.9980*	0.9976*
80%	0.8544*	0.9282*†	0.9359	0.9337	0.9980	0.9978*
Dice score						
10%	0.7818*	0.8437*†	0.9246	0.9250	0.9533*	0.9541*
20%	0.7855*	0.8743*†	0.9268	0.9392	0.9551*	0.9543*
40%	0.8346*	0.8624*	0.9390	0.9302	0.9571	0.9567
60%	0.8274*	0.8721*	0.9262	0.9030	0.9570*	0.9577*
80%	0.8273*	0.8857*	0.9476	0.9325	0.9597	0.9598
Sensitivity						
10%	0.6571*	0.7769*†	0.9523*	0.9394*	0.9959*	0.9950*
20%	0.6449*	0.8593*†	0.9567	0.9559	0.9968*	0.9971*
40%	0.7341*	0.8195*†	0.9726*	0.9535*	0.9976	0.9975
60%	0.7174*	0.8321*†	0.9407*	0.8919*	0.9983*	0.9974*
80%	0.7156*	0.8891*†	0.9520*	0.9059*	0.9985*	0.9979*
Specificity						
10%	1*	0.9995*	0.9223	0.8960	1	1
20%	1*	0.9990*	0.9203	0.8768	1	1
40%	0.9999*	0.9995*	0.8987	0.8971	1	1
60%	0.9999*	0.9993*	0.9547	0.9657	1	1
80%	0.9999*	0.9986*	0.9572	0.9761	1	1

AUROC, Dice score, sensitivity, and specificity from five replicates of transfer learning (T_{II}) and learning from medical images (L_{MI}) deep-learning models for binary segmentation of RGB images with skin cancer, microscopic H&E-stained prostate core biopsy, and CT of kidneys are reported. Images were depleted into different proportions (indicated by %). Statistically significant (Mood's median test, $p < 0.05$) differences between nonparametric distributions are indicated by *. Differences greater than 5% are indicated by † for the better performing model. A value of 1 indicates a perfect score.

non-target tissue pixels as tumors. The shared features and RGB pixel intensities common between classes of ImageNet natural-world and skin cancer images and the sufficient availability of examples of the majority of target clinical labels in the skin image dataset may have played an important role in achieving higher performance by the T_{II} model. Thus, a transfer-learned T_{II} model could leverage previously learned features from natural-world images to segment skin cancer from RGB images and exhibit lower false negatives and higher discriminatory ability for tumors and robustness against non-target artifacts in the images.

The L_{MI} and T_{II} models were trained with ≈250 whole-slide hematoxylin and eosin (H&E) images of prostate core biopsy images using 80% of the available data (224 images with any tumor labels and 20 without tumors). The L_{MI} models achieved higher median AUROC and Dice scores compared with T_{II} models for the test dataset (Table 1; Figure S3Bvii and Bviii). Greater numbers of prostate core biopsy images segmented by L_{MI} models had higher mean AUROC, specificity, and Dice scores for tumor segmentation (Table 1). Approximately 65% of images were segmented with higher AUROC and Dice scores by L_{MI} models (Table 1). The L_{MI} models also achieved 0.9 or higher AUROC and Dice scores for 34 test images (61%) compared with 29 for T_{II} (Table 1). The higher performance of metric distributions for L_{MI} models, save for sensitivity (Table 1), did not reach statistical significance of $p < 0.05$ (Mood's median test) (Table 1; Figure S3Bvii and Bviii). Figure 3A shows that both T_{II} and L_{MI} outputs had false-negative (yellow colors) regions and comparable true-positive regions for images with prostate tumors (green). Grad-CAM analysis showed that Gleason grade tumor regions were activated by L_{MI} models with higher intensity (Figure S4Cvi, column c) for making predictions compared with the activation profiles of T_{II} models (Figure S4Cvi, column e). The L_{MI} models were thus more accurate and more precise at making predictions and demarcating boundaries between tumor tissue and non-tissue (non-target) pixels. The T_{II} models could not segment all tumor tissue pixels and exhibited higher false-negative errors (Figure S4Cv and Cvi, images 3 and 4). Supervised learning using native prostate images improved segmentation (<5% [see Table 1]) and achieved statistically significant differences (Mood's median test, $p < 0.05$) in sensitivity.

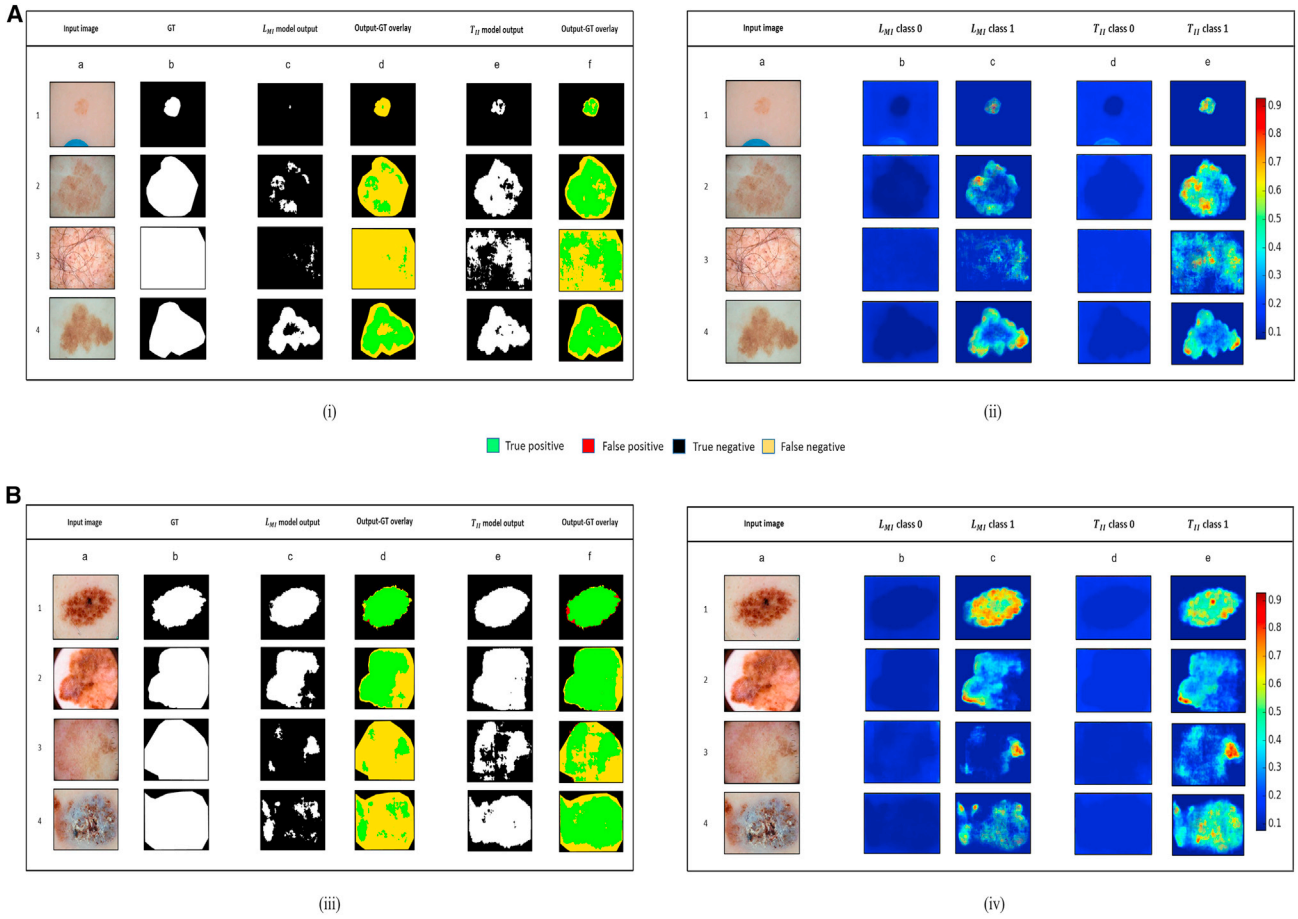


Figure 2. Visualization and explanation of transfer learning (T_{II}) and learning from medical images (LM_{II}) models for segmentation of skin cancers

(A) shows benign skin cancer and (B) shows malignant skin cancer images. A(i) and B(iii), left to right columns: a, input RGB image; b, binary mask of the clinical ground-truth label; c, mask of the output image after binary segmentation by the L_{MI} model; d, overlay of the clinical ground truth and L_{MI} model binary output masks; e, mask of the output image after binary segmentation by the T_{II} model; f, overlay of the clinical ground truth and the T_{II} model binary output masks. A(ii) and B(iv), target class-based Grad-CAM outputs. Class 0 represents the non-target or non-tumor pixel regions of the skin; class 1 represents pixels with benign tumors, moles, and lesions. Color bars represent the degree of model attention and importance, with deeper red indicating the most importance and deeper blue indicating the least important. Green, true positive (TP); black, true negative (TN); red, false positive (FP); yellow, false negative (FN); GT, clinical ground truth. See also Figures S4 and S5; Tables S1–S3.

The test dataset had two prostate core biopsy images without any ground-truth tumor signatures, and both L_{MI} and T_{II} models correctly predicted one image (image 3 in Figures 3A and 3B) with true-negative pixels. The Grad-CAM analysis for this image (Figure 3B, image 3) when queried with target class 0 correspondingly highlighted the non-target pixels but did not highlight any pixels when queried with target class 1 (which corresponds to tumor pixels). In contrast, both models incorrectly segmented benign tissue pixels for image 4 in Figures 3A and 3B as tumor and showed corresponding false-positive Grad-CAM activation for the non-target pixels. Thus, visual explanations from the image analysis of microscopic RGB images of prostate core biopsy interpreted that L_{MI} models achieved slightly better segmentation accuracy than T_{II} models (Table 1). We reason that this was based in part on the necessity and availability of the higher complexity of microscopic histology images for initializing accurate binary segmentation of prostate tumors by L_{MI} models. Moreover, the L_{MI} models trained using randomly initialized weights were well suited for differentiating between the non-target pixels and target tumor pixels for the prostate core biopsy image dataset. Conversely, for images that did not have tumors or with only benign tissue pixels, the transfer-learned T_{II} models performed better with lower false-positive outputs. We reason that DL models trained on natural-world images such as the ImageNet database may be slightly less efficient for complex tumor segmentation from microscopic pathology but may be used in conjunction with L_{MI} models for achieving superior segmentation of non-target classes. A corollary is that TL from a larger, more heterogeneous data and model weights of natural-world images to a smaller dataset may have optimal parameterization for learning the non-target pixels. Additionally, training using only pathology biopsy images (L_{MI} models) might be beneficial for

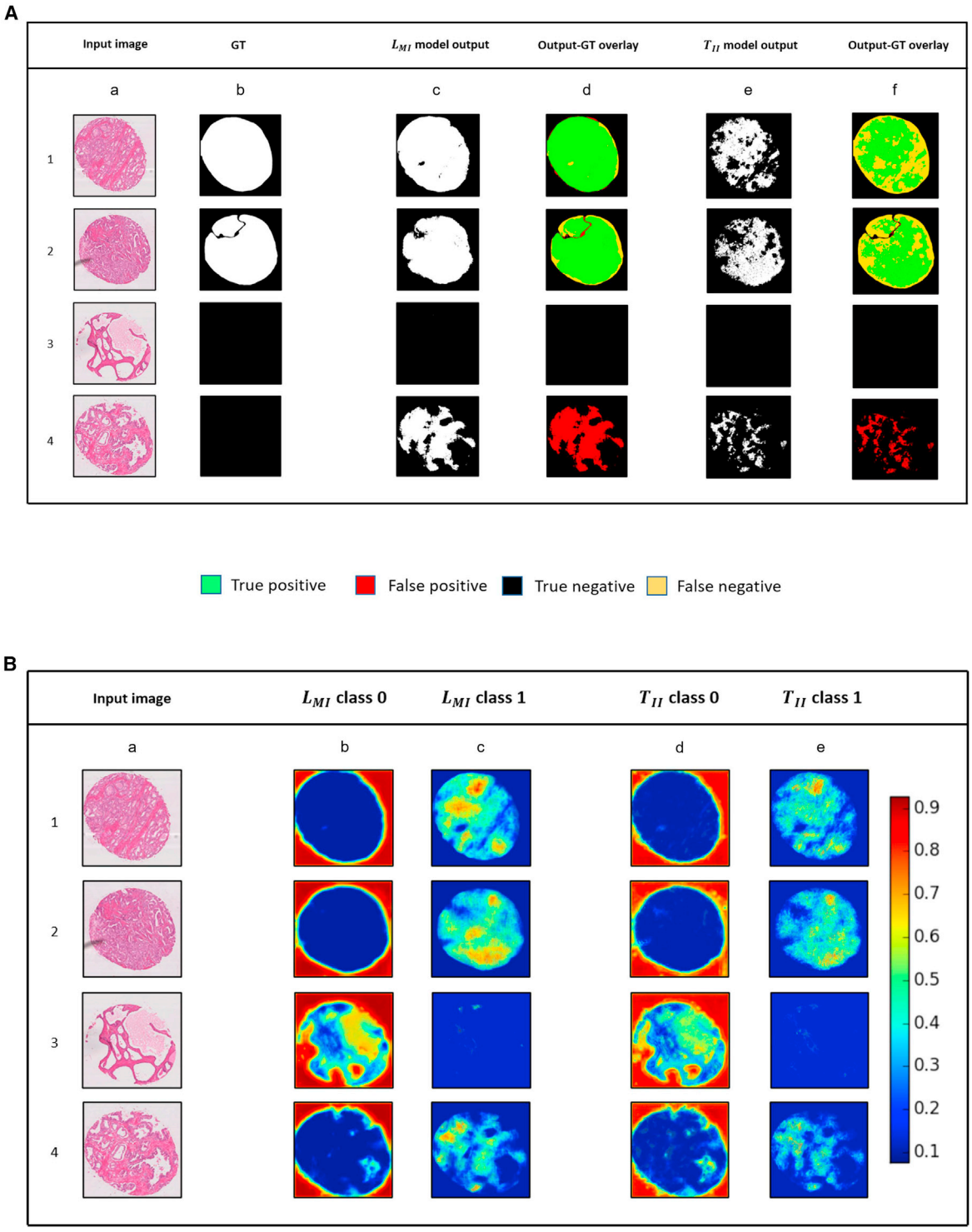


Figure 3. Visualization and explanation of transfer learning (T_{II}) and learning from medical images (L_{MI}) models for the segmentation of prostate core biopsy images

(A) Left to right columns: a, input RGB image; b, binary mask of the clinical ground-truth label; c, mask of the output image after binary segmentation by the L_{MI} model; d, overlay of the clinical ground truth and the L_{MI} model binary output masks; e, mask of the output image after binary segmentation by the T_{II} model; f, overlay of the clinical ground truth and the T_{II} model binary output masks.

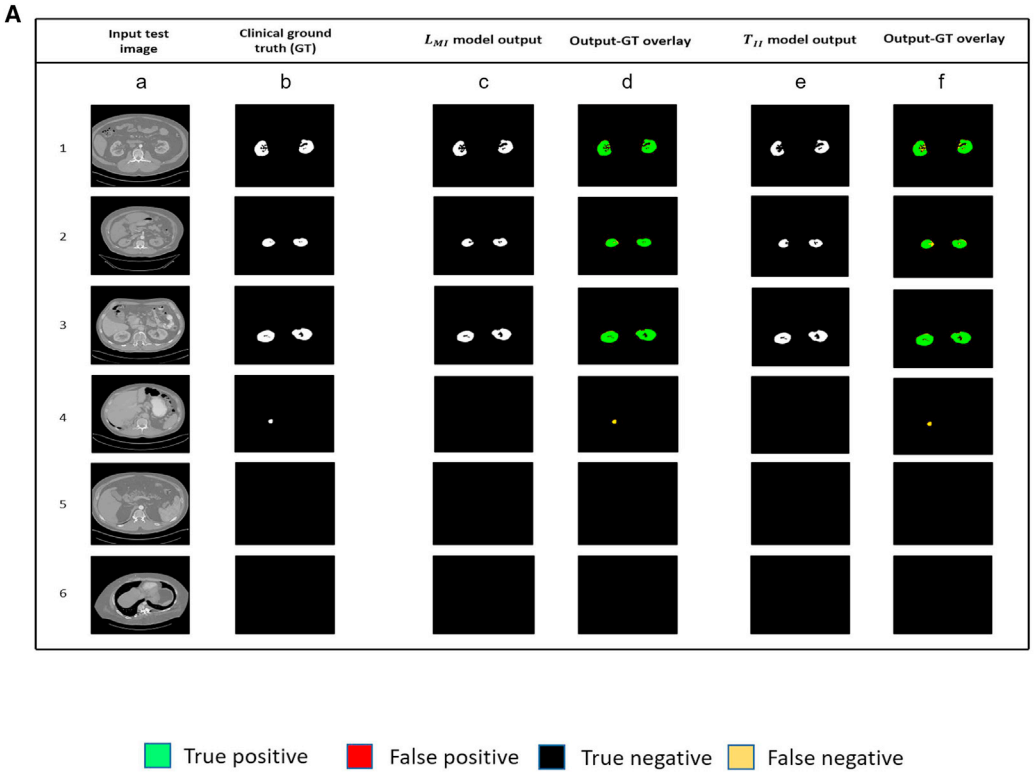
(legend continued on next page)

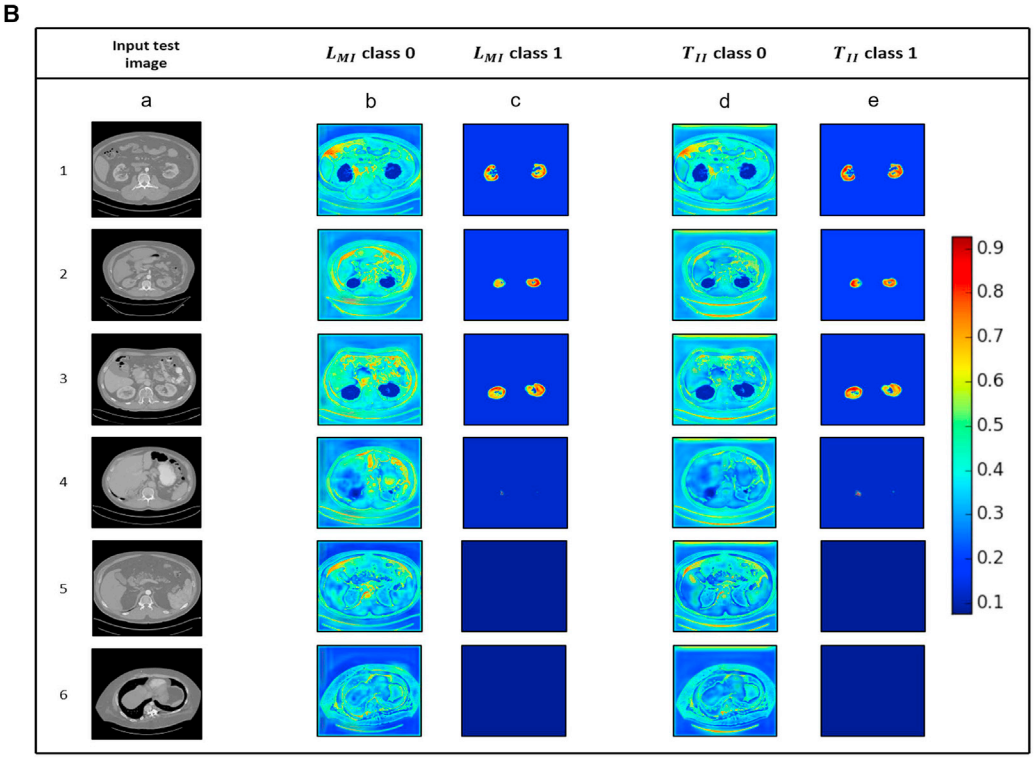
the comprehensive and accurate learning of detailed features for the demarcation of boundaries of tumors from non-tumor tissues and other non-target pixels.

45,000 (approximately) abdominal computed tomography (CT) digital imaging and communications in medicine (DICOM) images with 16,336 kidney tissue and 29,088 non-target clinical labels were used in this study (Table S2). Differences in the distributions of AUROC and sensitivity were statistically significant (Mood's median test, $p < 0.05$) between models, with L_{MI} performing slightly better but with less than 5% gain in performance (Table 1; Figure S3Cix and Cx). A statistically significant difference (Mood's median test, $p < 0.05$) between L_{MI} model sensitivity of 0.9985 and 0.9979 for T_{II} was calculated. The L_{MI} models performed better, with 1,937 test images achieving higher AUROC scores ($\approx$21% of total test images or $\approx$60% of images with kidneys). The images segmented by T_{II} models, on the other hand, showed higher Dice scores ($\approx$18% of all test images or $\approx$50% of images with kidneys) (Tables 1 and S4). This dataset contained significant numbers of images without kidneys (5,840 out of 9,085) (i.e., corresponding ground-truth mask labels were black pixels). Both models achieved the perfect median score of $\approx$1 for specificity and comparable Dice scores for segmentation of kidneys from non-target pixels and other organs (Table 1). Inherent embedded features and properties of gray-scale CT images are unlike natural-world images in the ImageNet database. Values for the AUROC and Dice scores indicated that both models performed comparably with marginal performance ($<$5%) gains by L_{MI}. Representative CT images with (Figures 4 and S4D) and without kidneys or with other organs (Figure 4) are shown for performance evaluation. For most test images, both models could segment out kidney tissue pixels with reasonable accuracy (Figure 4A, images 1, 2, and 3). In a few test images, L_{MI} models segmented larger areas of kidney pixels (green regions in Figure S4Dvii, columns d and f) while T_{II} models demonstrated higher false-negative (yellow regions) rates. Grad-CAM analysis also showed that both L_{MI} and T_{II} models could distinguish non-target pixels and exhibited the least activation for pixels without kidneys (class 0) (Figure 4B, columns c and e). In particular, the T_{II} models had lower sensitivity in demarcating the boundaries between kidneys and non-target pixels from other organs (Table 1). For image 4 in Figure 4A, both models showed false negatives or absence of kidney segmentation, and corresponding Grad-CAM outputs (Figure 4B, images c4 and e4 for L_{MI} and T_{II}) did not activate target class 1 in the final model outputs. Other images in the test dataset had similar outcomes, with models failing to segment kidney tissue pixels. In summary, training L_{MI} models from scratch for targeted segmentation was more optimal for achieving higher sensitivity when the target class of kidney tissue pixels was available in lower numbers than pixels of a non-target region, when there were other organs, or when the presence of other organs and tissue pixels outnumbered the presence of kidney tissues. In all other scenarios, L_{MI} and T_{II} models performed equally well, and Grad-CAM explanations suggest they may be used synergistically and interchangeably.

UMAP has been used to visualize and explain learning manifolds of DL models (McInnes et al., 2018). For skin cancer images, the target class 0 was assigned to presence of cancer lesion or tumors, and target class 1 indicated a non-target region or the absence of a lesion. Both L_{MI} and T_{II} models learned the clinical features within the input skin cancer images without being provided or trained with explicit clinical labels (Figure S5A) and could distinguish between benign and malignant lesions (Figure S5Ai for L_{MI} and Aiii for T_{II}). We also note greater separation of the two UMAP clusters for Figure S5Aiii compared with Ai, indicating that the T_{II} model distinguished the clinical differences in the skin cancer data more effectively than L_{MI} and supported the previous observation that T_{II} is a better performing model for skin cancer segmentation (Figure S5Aiii compared with Ai). For the non-target class (Figure S5Aii for L_{MI} and Aiv for T_{II}) models, we did not observe major differences in the UMAP clustering of the two clinical classes. However, the malignant tumors formed a tighter grouping than the non-target skin tissue pixels for both models. We hypothesize that these differences in image-derived features may help the DNN model layers to classify pixels of benign and malignant skin tumors and other non-target classes. The dataset utilized for digital pathology (Gleason, 2019) had a total of 244 H&E-stained core images with binary tumor or no-tumor segmentation masks associated with them and was smaller in comparison with the other datasets. For images of prostates, the target class 0 indicated the presence of tumor tissue pixels and 1 indicated the presence of non-target pixels. Analysis of UMAP mappings for prostate core biopsy images showed distinct groups for target class 0 and 1 for both models (Figure S5B). Clinical class labels such as presence or absence of tumors, however, did not show distinct UMAP clusters. This could be attributed to the fact that there were only two images without tumors in the test dataset. The possibility of different Gleason grade 2, 3, and 4 scores of tumors being responsible for the discrete arrangement of clusters seems promising and can be used as the input for training DL models. However, for this study, we simply focused on the binary tumor and no-tumor classification/segmentation task. In the CT data, target class 1 indicated the presence of kidney tissues, and target class 0 indicated the presence of non-target and or non-kidney tissue pixels. From Figure S5C for target class 0, the L_{MI} model showed a different manifold than T_{II} models, indicating a unique learning process for kidney tissue structures. Both models showed separation between images without kidneys (yellow cluster in Figure S5Cx and Cxii) and those with kidneys (violet cluster in Figure S5Cx and Cxii) and were able to distinguish between the non-target and/or non-kidney tissue pixels. Thus, learning of non-target pixel manifold was crucial for high sensitivity and specificity of segmentation of small organs or tissue pixels such as kidneys from a large CT image dataset.

(B) Target class-based Grad-CAM output. Class 0 represents the non-target or non-tumor pixel regions of the prostate core biopsy; class 1 represents Gleason grade 3, 4, or 5 tumors. Color bar represents the degree of model attention and importance, with deeper red indicating the most importance and deeper blue indicating the least important. Green, true positive (TP); black, true negative (TN); red, false positive (FP); yellow, false negative (FN); GT, clinical ground truth. See also Figures S4 and S5; Table S2.





(legend on next page)

DISCUSSION

Previous studies have reported that CNN DL models trained using TL from large ImageNet or medical datasets compared with smaller numbers of target images of limited scales can achieve better performance and reduce the computational training cost and overfitting on small training data. Performance gains for classification following TL between medical images of the same modalities but from different clinical tasks—for example, between magnetic resonance images (MRI and MRI)—have been reported (Ghafoorian et al., 2017; Van Opbroek et al., 2014). Cross-modality learning between MRI and CT (Dou et al., 2018) or between natural-world images and medical images also improved classification (Raghu et al., 2019; Van Ginneken et al., 2015; Bar et al., 2015; Ciompi et al., 2015; Shie et al., 2015). Another study claimed that fine-tuning the pretrained AlexNet CNN model to classify RGB-fused images achieved better performance on MRI images (Banerjee et al., 2018). For satellite image segmentation, TL from ImageNet provided 30% savings in computational costs while achieving accuracy levels comparable with TFS (Giorgiani do Nascimento and Viana, 2020). Similarly, transfer between similar image modalities (from legacy MRI to MRI) achieved a 0.63 Dice score by fine-tuning on only two images from the target domain (Ghafoorian et al., 2017). A domain adaptation protocol was also utilized to adapt a CNN trained with MRI images to unpaired CT data for cardiac structure segmentation (Dou et al., 2018).

However, emerging literature is skeptical of the true effects and benefits of TL from natural-world images, such as ImageNet for specialized tasks in the medical imaging domain. For example, a study reported that random initialization was surprisingly robust even in the low-data regime (10% available training data), and ImageNet pretraining speeds up convergence early in training but does not necessarily provide regularization or improve the final target task accuracy in their study (He et al., 2019). Another study inspected the effects of transfer from natural-world images for two large-scale medical imaging classification tasks from chest X-rays and retinal fundus photographs (Raghu et al., 2019). They found that transfer does not significantly aid performance, and the model performance on the ImageNet database did not translate to the medical domain. They also reported that transfer from ImageNet did not significantly aid performance compared with smaller, simpler convolutional TFS models that use only medical images to classify retinal fundus and chest X-ray images. Results from another work showed that learned features from ImageNet do not transfer well for fine-grained medical image classification tasks (Kornblith et al., 2019). These studies suggest that some of the shortcomings of DL models trained natively using medical images may be attributed to overparameterization rather than sophisticated feature reuse advantage thought to be provided by TL.

The role of TL and training natively using medical images for segmentation tasks is not well characterized and understood. The detailed characterization of the value of available DL methods for medical image segmentation across different image modalities has also not been reported. In this study, we report several interesting findings for the use of statistical methods and visual explanation of DNN models for interpretation of high-accuracy segmentation of clinical information from medical images. Several commonly held practices and anecdotal concepts were put to the test, such as randomization and proportioning of training and test data, DL model selection, and performance under low-data regimes. We report that although TL from natural-world images showed benefits, it was not a universal solution for improving binary segmentation of medical images. In fact, rigorous statistical significance testing and Grad-CAM evidence showed that supervised learning with medical images provided unique gains in sensitivity and specificity and can be used synergistically with TL. We recommend splitting the available images into at least five different proportions and repeating the training and validation of each split five times to identify the optimal distribution of clinical labels to prevent skew and memorization. If clinical label distributions are non-Gaussian (i.e., one class heavily outnumbers the other), randomization is not optimal. For example, the skin cancer dataset had fewer (n = 1,118) malignant compared with benign tumor (n = 12,668) images, and the clinical label distributions were uneven between training and testing splits. DL models trained using these splits demonstrated higher false-negative errors (Figure 2B). We recommend checking for the normalcy of performance metric (AUROC, Dice score, sensitivity, and specificity) distributions and Yeo-Johnson transformations of such distributions to select the appropriate parametric Gaussian or non-parametric statistical testing to establish significance. For the non-parametric distributions reported in this study, medians served as better measures of central tendencies than means. Another finding from this study was instances in which both training schemes of L_{MI} and T_{II} could be used synergistically or as ensemble models to improve higher-level morphological and fine-grained segmentation (discussed in detail below). This is an important distinction from prior studies that reported anecdotal selection of larger datasets and pretrained L_{MI} or T_{II} models and discarding the inferior model.

For several target image classes, the L_{MI} and T_{II} models may be synergistically used to achieve a mutually beneficial increase in segmentation accuracy. For skin images, T_{II} models

Figure 4. Visualization and explanation of transfer learning (T_{II}) and learning from medical images (L_{MI}) models for the binary segmentation of CT images with kidneys

(A) Left to right columns: a, input RGB image; b, binary mask of the clinical ground-truth label; c, mask of the output image after binary segmentation by the L_{MI} model; d, overlay of the clinical ground truth and the L_{MI} model binary output masks; e, binary mask of the output image after segmentation by the T_{II} model; f, overlay of the clinical ground truth and the T_{II} model binary output masks.

(B) Target class-based Grad-CAM output. Class 0 represents the non-target or non-kidney class or region pixels of the CT image; class 1 represents the kidney tissue pixels. Color bar represents the degree of model attention and importance, with deeper red indicating the most importance and deeper blue indicating the least importance. Green, true positive (TP); black, true negative (TN); red, false positive (FP); yellow, false negative (FN); GT, clinical ground truth.

See also Figure S1 and Table S2.

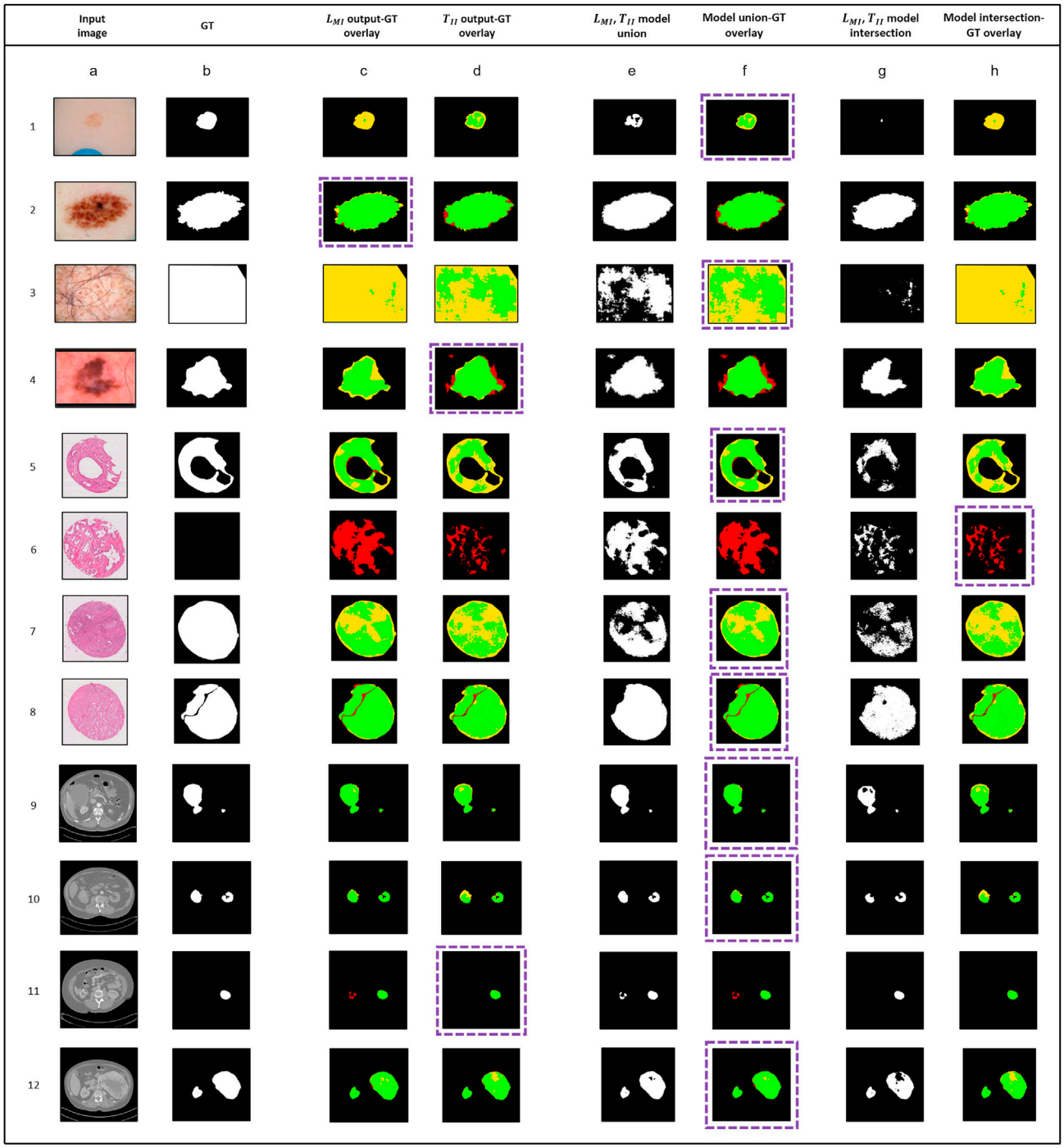


Figure 5. Visualization of synergistic outputs of the unions and intersections of TL (T_{II}) and learning from medical images (L_{MI}) models. Left to right columns: a, input RGB image; b, binary mask of the clinical ground-truth label; c, overlay of the clinical ground truth and the L_{MI} model binary output masks; d, overlay of the clinical ground truth and the T_{II} model binary output masks; e, binary mask of the output image after computing the union of binary segmentations by L_{MI} and T_{II} models; f, overlay of the clinical ground truth and the binary union output masks; g, binary mask of the output image after computing the intersections between binary segmentation by the L_{MI} and T_{II} models; h, overlay of the

(legend continued on next page)

segmented non-target pixels as disease signatures (false positive) in more images than L_{MI} models (higher specificity in Table 1). Thus, the synergistic use of T_{II} models to learn cancer lesions from skin images and L_{MI} models to distinguish non-target pixels can result in desirable segmentation (Figure 5). This was reversed for prostate core biopsy and kidney CT images, for which T_{II} models predicted target tumor or organ signatures as non-target (false negative). Thus, for prostate core biopsy and kidney CT images, the T_{II} models were better at learning non-target pixels while L_{MI} models were superior at segmenting tumors or organs (Figure 5). We also report images where L_{MI} and T_{II} models segmented the target pixels at a different location on the same image. The union of output masks from both models increased the segmentation of higher proportions of ground-truth pixels for such images (images 7, 9, and 10 in Figure 5). One limitation for this method arose when one model (T_{II}) demonstrated high false positives while the other (L_{MI}) resulted in high false negatives (images 4 and 8 in Figure 5). The union of model output masks for these images identified false-positive regions, and the intersection enriched the false-negative pixels for desired optimizations.

Automated segmentation of one image by both models and the use of individual segmentation output masks in various combinations were most optimal for achieving the desired performance across all datasets (Figure 5, columns e and g). In Figure 5, columns c and d show the individual model performance for L_{MI} and T_{II}, respectively. Applying a union operation that combined results from both L_{MI} and T_{II} output segmentation masks (Figure 5, columns e and f for images 1, 3, 5, 7, 9, 10, and 12) improved the performance when one model exhibited high false negatives and the other did not. Union operations, however, decreased the performance when one model exhibited high false positives (Figure 5, columns e and f for image 6). Operations that generated intersections between segmented region outputs from both models (Figure 5, columns g and h for images 2, 6, and 11) improved the performance when one model revealed high false positives. However, intersection operations were not useful when one model (L_{MI} for image 3 in Figure 5) exhibited high false negatives. Our overall results indicated that it was most optimal to use intersection of segmentation outputs when both models show high false positives and using the union of segmentation outputs when both models show high false negatives. An automated algorithm (Algorithm 1 in STAR Methods) was developed that compared the individual segmentation Dice scores achieved either by the L_{MI} and/or T_{II} models on medical images or their intersection and unions and generated a final image (Figure 5, purple bounding boxes) with the best possible segmentation output. False-positive rates were compared when Dice scores were not defined or available (Figure 5, image 6). As seen in Figure 5, the algorithm iteratively selected individual L_{MI} or T_{II} outputs in images where their unions and intersections did not improve segmentation performance (Figure 5, images 2c, 4d, and 11d). In the majority of instances, the algorithm selected union masks from both models (images in Figure 5, column f) to reduce false positives and negatives, underscoring the value of using L_{MI} and T_{II} models synergistically for achieving higher accuracy.

Findings from the UMAP visualizations corroborated our previous findings from statistical analysis and Grad-CAM visualizations. For the skin cancer data, T_{II} performed better than L_{MI} models for segmenting skin lesions. The T_{II} model also grouped benign and malignant class labels more efficiently than the L_{MI} model. For the prostate core biopsy data, we report similar performance for both models in terms of clinical class grouping under UMAP approximation, which can be attributed to the fact that one class (presence of tumor tissues) heavily outnumbered the other (absence of tumor tissue). Thus, when false positives are to be kept at a minimum, DL strategies can alternate synergistically for higher sensitivity and specificity for the detection and segmentation of disease regions. This was especially important under low-data regimes and when individual models were inadequate for achieving high sensitivity and specificity. Statistically significant (Mood's median test, $p < 0.05$) differences for estimating the performance, reproducibility, and replicability of L_{MI} and T_{II} models can also be used for situations where sparsity promotes better learning. Thus, even with the purported benefits, TL can result in suboptimal outcomes if model outputs and clinical labels are not carefully examined to provide information causal for medical-grade performance. This can be attributed to the tendency of the TL model to gravitate toward the original dataset manifold when sufficiently large and when diverse new data are not used for fine-tuning. For example, the L_{MI} model weights were easier to update for two medical image types in this study and outperformed TL models for specificity of binary segmentation. On the other hand, TL models (T_{II}) demonstrated significant performance gains in sensitivity for larger and complex segmentation of multicolored skin cancer RGB images, which possibly shared features and complexity with natural-world images. Training and generating ensemble T_{II} and L_{MI} models with automated computation of segmentation masks with high sensitivity and specificity for the segmentation of target features can therefore serve as a highly effective strategy for medical images.

The DL and statistical methods communicated in this study and findings derived from them can be used for selecting the best training and validation data splits, using the least numbers of images required for high-performance segmentation, and automated selection of best-performing models and correctly segmented images. The Dice score computation algorithm also automated screening for robust segmentation performance from multiple models. The approach described in our study may also alleviate under specification and stress testing challenges precluding real-world deployment of DL models when tested with OoD data (D'Amour et al., 2020). For example, we recommend training T_{II} and L_{MI} models using domain-specific data of choice and amalgamation of their desired performance to

clinical ground truth and the binary intersection output masks. Images with dashed purple bounding boxes indicate the segmentation output generated by an automated rule-based algorithm that uses synergistic output generation from columns c, d, f (union), and h (intersection) to select images and models with highest Dice score and lowest false-positive and -negative pixels. Green, true positive (TP); black, true negative (TN); red, false positive (FP); yellow, false negative (FN); GT, clinical ground truth.

achieve high-grade segmentation. During deployment, a previously unseen image can be sequentially segmented by both models, and independent segmentation masks are generated (see examples in Figure 5). These output masks can then be fed to the automated algorithm (Algorithm 1 in STAR Methods) to compute and rank-order (using the numbers of correctly segmented pixels) the best Dice scores in the union and intersection masks. The methods, code, and models from this study can be used as starting points for custom applications and to benchmark the performance of datasets and explainable DL models of choice. The open-access GitHub repository hosting (1) 30 fully trained and validated DL models (15 models each for L_{MI} and T_{II} translating to 10 models for each of the three image modalities), (2) Grad-CAM results and the associated software code and performance estimation from more than 10,000 test images, and (3) a separate statistical analysis package for calculating the AUROC, Dice score, and sensitivity/specificity performance of DL models will be a valuable resource for biomedical and computer science researchers.

Limitations of the study

Although benchmarked datasets and widely used DNN architectures and models were used, the findings from this study were optimized for the medical images, clinical labels, and DNN specifically used for this research. Skin cancer and prostate tumor regions annotated by pathologists can be coarse, contain non-relevant tissue and skin, and, in some cases, be inaccurate, which can increase disagreements with DL segmentation performance. Additional fine-grained clinical image annotation tools and labeled images may be needed for extremely precise analysis of the results generated by DL models. However, in view of the reproducible results and robust human-enabled labeling process for the datasets used in this study (Gleason, 2019; ISIC, 2019; Heller et al., 2019), we reason that scenarios such as noisy or mislabeled clinical features may manifest in the minority (subset) of the entire data and will not change the conclusions. As previously reported by us (Rana et al., 2020; Javia et al., 2018) and others, DL models often match and learn the most distinguishing class-based features from medical images following both supervised and unsupervised training. In congruence, this study showed that semi-supervised Grad-CAM activation and UMAP mappings were indeed in agreement with the results obtained from Dice score, AUROC, sensitivity, and specificity analysis for clinical labels. Performance metrics for less frequently used DL model architectures suitable for classification and generative tasks were not evaluated in this study. Multi-class segmentations (e.g., different prostate tumor grades and skin cancer types) are future growth areas from this research. Based on the shared representations between medical images and unified learning mechanisms of deep CNN architectures, the findings of this study should, however, generalize to other macro- and microscopic images and clinical segmentation tasks.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100107.

ACKNOWLEDGMENTS

Students from the MIT Undergraduate Research Opportunity Program for data preprocessing and discussions.

AUTHOR CONTRIBUTIONS

S.G. and P.S. designed the study and computational framework and analyzed data and results. S.G. made figures and tables and implemented the research. S.G. and P.S. wrote the manuscript. P.S. led conception, planning, and supervision of the research.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 14, 2021
Revised: July 23, 2021
Accepted: October 15, 2021
Published: November 8, 2021

REFERENCES

Athanasiou, L.S., Fotiadis, D.I., and Michalis, L.K. (2017). Atherosclerotic Plaque Characterization Methods Based on Coronary Imaging (Academic Press).

Banerjee, I., Crawley, A., Bhethanabotla, M., Daldrup-Link, H.E., and Rubin, D.L. (2018). Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma. Comput. Med. Imaging Graphics *65*, 167–175.

Bar, Y., Diamant, I., Wolf, L., and Greenspan, H. (2015). Deep learning with non-medical training used for chest pathology identification. In Medical Imaging 2015: Computer-Aided Diagnosis SPIE Proceedings, *Vol. 9414* (International Society for Optics and Photonics), p. 94140V. https://doi.org/10.1117/12.2083124.

Bayat, A., Anderson, C., and Shah, P. (2021). Automated end-to-end deep learning framework for classification and tumor localization from native non-stained pathology images. In Medical Imaging 2021: Image Processing SPIE Proceedings, *Vol. 11596* (International Society for Optics and Photonics), p. 115960A. https://doi.org/10.1117/12.2582303.

Bradski, G. (2000). The OpenCV library. Dr. Dobb's J. Softw. Tools Prof. Programmer *25*, 120–123.

Chen, D., Liu, S., Kingsbury, P., Sohn, S., Storlie, C.B., Habermann, E.B., Naessens, J.M., Larson, D.W., and Liu, H. (2019). Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digital Med. *2*, 43.

Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., and van Ginneken, B. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Med. Image Anal. *26*, 195–202.

Collins, G.S., and Moons, K.G. (2019). Reporting of artificial intelligence prediction models. Lancet *393*, 1577–1579.

Dice, L.R. (1945). Measures of the amount of ecologic association between species. Ecology *26*, 297–302.

Dou, Q., Ouyang, C., Chen, C., Chen, H., and Heng, P.-A. (2018). Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. arXiv, 1804.10916. http://arxiv.org/abs/1804.10916.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv, 2011.03395. http://arxiv.org/abs/2011.03395.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., and Kohane, I.S. (2019). Adversarial attacks on medical machine learning. Science *363*, 1287–1289.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.-E., Tempany, C.M., Van Ginneken, B., et al. (2017). Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 516–524.

Ghorbani, A., Natarajan, V., Coz, D., and Liu, Y. (2020). Dermgan: synthetic generation of clinical skin images with pathology. In Machine Learning for Health Workshop, pp. 155–170.

Ghosal, S., Blystone, D., Singh, A.K., Ganapathysubramanian, B., Singh, A., and Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. Proc. Natl. Acad. Sci. U S A *115*, 4613–4618.

Ghosal, S., Zheng, B., Chapman, S.C., Potgieter, A.B., Jordan, D.R., Wang, X., Singh, A.K., Singh, A., Hirafuji, M., Ninomiya, S., et al. (2019). A weakly supervised deep learning framework for sorghum head detection and counting. Plant Phenomics *2019*, 1525874.

Ghosal, S., Xie, A., and Shah, P. (2021). Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation. arXiv, 2109.00115. http://arxiv.org/abs/2109.00115.

Van Ginneken, B., Setio, A.A., Jacobs, C., and Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI) (IEEE), pp. 286–289.

Giorgiani do Nascimento, R., and Viana, F. (2020). Satellite image classification and segmentation with transfer learning. In AIAA Scitech 2020 Forum, p. 1864.

Gleason. (2019). Gleason 2019 Homepage. https://gleason2019.grand-challenge.org/.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with numpy. Nature *585*, 357–362.

He, K., Girshick, R., and Dollár, P. (2019). Rethinking ImageNet pre-training. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4918–4927.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al. (2019). The Kits19 Challenge Data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv, 1904.00445. http://arxiv.org/abs/1904.00445.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Huynh, B.Q., Li, H., and Giger, M.L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J. Med. Imaging *3*, 034501.

Iglovikov, V., and Shvets, A. (2018). TernausNet: U-Net with Vgg11 encoder pre-trained on ImageNet for image segmentation. arXiv, 1801.05746. http://arxiv.org/abs/1801.05746.

ImageNet (2012). ImageNet Archive. http://www.image-net.org/.

ISIC (2019). ISIC Archive Homepage. https://www.isic-archive.com/.

Javia, P., Rana, A., Shapiro, N., and Shah, P. (2018). Machine learning algorithms for classification of microcirculation images from septic and non-septic patients. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (IEEE), pp. 607–611.

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Med. *17*, 195.

Kornblith, S., Shlens, J., and Le, Q.V. (2019). Do better ImageNet models transfer better? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2661–2671.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pp. 1097–1105.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. arXiv, 1312.4400. http://arxiv.org/abs/1312.4400.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: uniform manifold approximation and projection. J. Open Source Softw. *3*, 861.

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, pp. 56–61.

Mood, A., Graybill, F., and Boes, D. (1963). Introduction to the Theory of Statistics (Mc-Graw Hill Book Company.).

Näppi, J.J., Hironaka, T., Regge, D., and Yoshida, H. (2016). Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography. In Medical Imaging 2016: Computer-Aided Diagnosis (International Society for Optics and Photonics), p. 97852B.

Van Opbroek, A., Ikram, M.A., Vernooij, M.W., and De Bruijne, M. (2014). Transfer learning improves supervised image segmentation across imaging protocols. IEEE Trans. Med. Imaging *34*, 1018–1030.

Paschali, M., Conjeti, S., Navarro, F., and Navab, N. (2018). Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 493–501.

Pokuri, B.S.S., Ghosal, S., Kokate, A., Sarkar, S., and Ganapathysubramanian, B. (2019). Interpretable deep learning for guided microstructure-property explorations in photovoltaics. NPJ Comput. Mater. *5*, 95.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: understanding transfer learning for medical imaging. In Advances in Neural Information Processing Systems, pp. 3342–3352.

Rana, A., Lowe, A., Lithgow, M., Horback, K., Janovitz, T., Da Silva, A., Tsai, H., Shanmugam, V., Bayat, A., and Shah, P. (2020). Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. JAMA Netw. Open *3*, e205111.

Razali, N.M., and Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal. *2*, 21–33.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 234–241.

Van Rossum, G. (1995). Python Tutorial, May 1995 (CWI), CWI Technical Report CS-R9526.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Shah, P., Yauney, G., Gupta, O., Patalano Ii, V., Mohit, M., Merchant, R., and Subramanian, S. (2018). Technology-enabled examinations of cardiac rhythm, optic nerve, oral health, tympanic membrane, gait and coordination evaluated jointly with routine health screenings: an observational study at the 2015 Kumbh Mela in India. BMJ Open *8*, e018774.

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., and Schork, N. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digital Med. *2*, 69.

Shie, C.-K., Chuang, C.-H., Chou, C.-N., Wu, M.-H., and Chang, E.Y. (2015). Transfer representation learning for medical image analysis. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE), pp. 711–714.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv, 1409.1556. http://arxiv.org/abs/1409.1556.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv, 1312.6304. http://arxiv.org/abs/1312.6034.

Suzuki, K. (2017). Overview of deep learning in medical imaging. Radiological Phys. Technol. *10*, 257–273.

Taha, A.A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging *15*, 29.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Weisberg, S. (2001). Yeo-Johnson Power Transformations (Department of Applied Statistics, University of Minnesota).

Yauney, G., Angelino, K., Edlund, D., and Shah, P. (2017). Convolutional neural network for combined classification of fluorescent biomarkers and expert annotations using white light images. In 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE) (IEEE), pp. 303–309.

Zar, J.H. (2013). Biostatistical Analysis: Pearson New International Edition (Pearson Higher).

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., and Oermann, E.K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. *15*, e1002683.

Zeiler, M.D. (2012). Adadelta: an adaptive learning rate method. arXiv, 1212.5701. http://arxiv.org/abs/1212.5701.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Processed data set with binary segmentation masks used in this study for skin cancer segmentation	https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main	https://doi.org/10.5281/zenodo.5570844
Processed data set with binary segmentation masks used in this study for prostate core biopsy segmentation	https://gleason2019.grand-challenge.org/	https://doi.org/10.5281/zenodo.5570844
Processed data set with binary segmentation masks used in this study for kidney computed tomography binary segmentation	https://kits19.grand-challenge.org/	https://doi.org/10.5281/zenodo.5570844
Trained L_{MI} and T_{II} model outputs from all three data sets	This paper	https://doi.org/10.5281/zenodo.5570844
Trained model Grad-CAM outputs from all three data sets	This paper	https://doi.org/10.5281/zenodo.5570844
Software and algorithms		
Code repository for the study	This paper	https://doi.org/10.5281/zenodo.5570844
Trained image segmentation L_{MI} and T_{II} model repository	This paper	https://doi.org/10.5281/zenodo.5570844
MATLAB R2019B	MathWorks, 2019	https://www.mathworks.com/products/new_products/release2019b.html
Python 3.6.10	Van Rossum, 1995	https://www.python.org/
Numpy 1.19.1	Harris et al., 2020	https://github.com/numpy/numpy
Pandas 1.1.3	McKinney, 2010	https://github.com/pandas-dev/pandas
Matplotlib 3.3.4	Hunter, 2007	https://github.com/matplotlib/matplotlib
Opencv 4.5.1	Bradski, 2000	https://github.com/opencv/opencv
SciPy 1.5.2	Virtanen et al., 2020	https://github.com/scipy/scipy
Keras 2.1.6 (GPU version)	https://keras.io/	https://github.com/keras-team/keras
Tensorflow 1.12.0 (GPU version)	https://www.tensorflow.org/	https://github.com/tensorflow
keras-explain-0.0.1	https://pypi.org/project/keras-explain/	https://github.com/primozgodec/keras-explain
CUDA 10.1	NVIDIA corporation	https://developer.nvidia.com/cuda-10.1-download-archive-base

RESOURCE AVAILABILITY

Lead contact

Further information and requests for codes, models, and other resources should be directed to and will be fulfilled by the lead contact, Dr. Pratik Shah Ph.D, pratiks@mit.edu.

Material availability

This study generated fully trained deep learning models and outputs for binary segmentation of skin cancer, prostate core biopsy and kidney CT images. The trained models are publicly available at Zenodo https://doi.org/10.5281/zenodo.5570844. This repository hosts 2,758 test images for skin, 49 test images for prostate core biopsy and 9,085 images for the kidney CT and their segmentation and Grad-CAM outputs from the L_{MI} and T_{II} models described in this study.

Data and code availability

- The three data sets used in this study can be obtained from: ISIC-Archive database: https://challenge.isic-archive.com/data/ for the skin images, Gleason-2019 database: https://gleason2019.grand-challenge.org/Register/ for the prostate core biopsy

images and Kits-19 database: https://kits19.grand-challenge.org/data/ for the kidney CT images. These URLs are also listed in the key resources table. Specific data obtained from these hyperlinks used for this study are deposited and publicly available at Zenodo repository: https://doi.org/10.5281/zenodo.5570844 under the open source Apache License 2.0 (https://opensource.org/licenses/Apache-2.0).

- The software code used for preprocessing raw data, training and evaluation of deep learning models reported in this study are deposited and publicly available at Zenodo repository: https://doi.org/10.5281/zenodo.5570844 under the open source Apache License 2.0 (https://opensource.org/licenses/Apache-2.0). The code, processed data, figures and documentation are available at Zenodo repository: https://doi.org/10.5281/zenodo.5570844as of the date of publication under the open source Apache License 2.0 (https://opensource.org/licenses/Apache-2.0). The output images from the models are available as of the date of publication at Zenodo repository: https://doi.org/10.5281/zenodo.5570844 under the creative commons public domain dedication version 1.0 or later. These URLs are also listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data and preprocessing

Three data sets consisting of macroscopic optical skin (13,786), microscopic RGB prostate core biopsy (244), and CT Digital Imaging and Communications in Medicine (DICOM) images (45,937) were used for this study. Images were represented by an n-by-3 data array that defines the red, green, and blue (RGB) color components for each pixel. The pixel colors were determined based on combinations of the RGB intensities stored in individual color planes. Skin images from the ISIC (The International Skin Imaging Collaboration) archive consist of 13,786 3-channel RGB images of resolution ranging from (7000 × 4500) to (1024 × 720) (W × H) with available ground truth (GT) binary mask associated with each image as its label (ISIC, 2019). Each binary mask had white pixels denoting regions of the image where skin moles and or cancer signatures were present (class 1). Black pixels in the binary masks represented non-target and or non-disease regions (class 0). Figure 2 shows examples of the input images and the corresponding ground-truth mask for the ISIC data. Prostate core biopsy RGB images from Gleason2019 (Grand Challenge for Pathology at MICCAI 2019) consisted of 244 tissue microarray (TMA) 3-channel RGB images of resolution 5120 × 5120 (W × H) (Gleason, 2019). Each TMA image was annotated in detail by several expert pathologists for assigning a Gleason tumor grade of 1 to 5 and a GT segmentation mask. Gleason grades 1 and 2 were clinically rare and not associated with tumors, and were designated as benign (class 0) with black pixels and regions of Gleason Grades 3, 4 and 5 were considered non-benign or tumors and represented by white pixels (class 1) (Rana et al., 2020). Figure 3 shows examples of input images and the corresponding ground truth masks of Gleason2019 data. The KiTS19 database had images with CT volumes of patients who underwent partial or radical nephrectomy for one or more kidney tumors at the University of Minnesota medical center between 2010 and 2018 (Heller et al., 2019). Each pixel represented a tissue or non target region and had an assigned gray scale value between 0 and 255, which represented the x-Ray beam attenuation to the tissue (Athanasiou et al., 2017). For this study, CT volumes were processed into 45,424 512 × 512 (W × H) 2D axial slices either with or without kidneys (along with kidney tumors). Kidneys (along with the tumor) were associated with ground truth multi color masks provided in KiTS19 data. Image slices and pixels in which kidneys were not present were associated with black pixels, and were considered to be the non target regions (class 0) for this study and the white pixels that corresponded to regions with kidney tissues (with or without tumors) were the target regions (class 1). Figures 4 and S4D show examples of input images and corresponding ground truth masks for the kidney CT image data.

The performance evaluation of the T_{II} and L_{MI} models trained using 80:20 data splits for different clinical classes on the medical image subtypes used in this study is shown in Figure 1. These classes were region-based, and a class was determined by the presence or absence of certain clinical features or signatures on an image and the severity of those features in the context of the data set. For example, for skin images lesions were clinically labeled as benign (12,668 images) or malignant (1,084 images) moles (lesions or tumors), and 19 images were intermediate. For H & E stained prostate core biopsy images the non target or benign regions of the prostate core were considered healthy or without tumors. Gleason grade 3, 4 or 5 regions were combined to represent a separate clinical label of tumors. This process resulted in 22 images with prostate tumors and 20 images without tumors. For the kidney CT DICOM images, the presence or absence of kidneys were considered as two explicit clinical classes. Before feeding into neural networks, the input images were resized to 608 × 416 (W × H) for skin and 608 × 416 (W × H) for prostate core biopsy data sets to reduce training time. Kidney CT images were used in their original resolution of 512 × 512 (W × H).

Ethics approval

All three data sets used in this study are publicly available, deidentfied and were exempt from Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects review. Skin images can be downloaded from the ISIC archive. Prostate core biopsy RGB images from Gleason2019 grand challenge for pathology website. Computed tomography images from the 2019 kidney tumor segmentation grand challenge website.

Deep learning

We evaluated several architectures suitable for segmentation tasks and selected the well-established and widely used VGG-UNet architecture (Iglovikov and Shvets, 2018; Ronneberger et al., 2015). The VGG-UNet is an encoder-decoder type framework in which the encoder is the VGG-16 (Simonyan and Zisserman, 2014) architecture without the fully-connected (FC) layers, and the decoder has subsequent upsampling of previously encoded layers output leading to the final class-dependent predicted segmentation mask. Using the VGG-UNet architecture as the backbone we trained DL models using ImageNet initialization (T_{II}) for training with pretrained weights from natural world images from the ImageNet database (ImageNet, 2012). Model random initialization (L_{MI}) used the same architecture but was trained without utilizing pretrained weights (i.e., it was initialized with random weights). All T_{II} and L_{MI} models were trained with optimization of the chosen loss function and convergence of learning and small loss values, and were capped at 40 epochs for the skin and prostate core biopsy data and 20 epochs for the kidney CT images. The categorical crossentropy loss function for pixel-wise binary classification and the Adadelta optimizer (with the default initial learning rate of 1.0) were used (Zeiler, 2012). An optimal batch size of two was used for all three data sets during training. A NVIDIA GeForce GTX 1080Ti with 12 GB of video memory was used to perform the training and performance evaluation. Five random 80:20 (train:test) splits for each of the three individual image types were trained and tested with internal five-fold repeats using either TL (total of 25 T_{II}-models) or L_{MI} (25 L_{MI}-models). This random five-fold splitting of the data sets was used for the reallocation of images and associated clinical labels to evaluate overfitting and underspecification (Table S2). The resulting five sets of T_{II}-L_{MI} models were then compared using statistical testing of the pixel-by-pixel mean, median and standard deviations of AUROC, Dice scores, sensitivity (true positive rate), and specificity (true negative rate) of the associated segmentation masks. A model performance threshold was defined using differences in metric values greater than 5% and or those that reach statistical significance (Mood's median test, $p < 0.05$). This process was repeated with 25 L_{MI} and T_{II} models (Figure 1; Table S1 for skin image data). Thus for each of the three individual medical image subtypes, 50 DL models were trained and analyzed for segmentation of tumors and organs (Figures 1 and S3). For the data depletion experiments, starting with the selected 80:20 (train:test) split the training data was depleted to 60:40, 40:60, 20:80 and 10:90 ratios (Figure 1; Table 2) by randomization. For example, the training data (80%) from the initial 80:20 training data set was split to generate two subsets - one representing 60% of the complete data set and the remaining that represents 20%. This 20% was then pooled with the initial 20% test split (from the 80:20 set) to generate the 40% test set and the new depleted 60:40 split. The subsequent 40:60 (train:test) split was similarly generated from the 60:40 proportion. The depletion was stopped after the train:test split reached the ratio of 10:90.

Data depletion and randomization methods

Randomization of available data into 80:20 proportions is standard practice in DL research. Researchers may also use repeats of individual model runs to fine tune and estimate the reproducibility of their results. Randomization can result in imbalance due to a skewed or non-Gaussian distribution of labels and clinical data available in smaller numbers. We employed a mixture of K-fold (K = 5) cross-validation and random sub sampling to generate five sets of data splits (Figure S1A), to train and test the L_{MI} and T_{II} models. The randomization scheme resulted in 20% images in common between at least two of the individual test sets and no common test data between all five sets. Approximately, 40% of the data was common across the generated training sets (Figure S1A). The median values for clinical label distribution remained consistent across the five randomized sets, and approximately equal proportions (90%) of the most frequent and target clinical labels ended up in training and testing data (Figure S1A). Less prevalent clinical labels (e.g. malignant skin cancers) were also proportionately distributed following randomization. Five-fold randomization of CT images resulted in proportional distribution of images without and with kidneys in training and test data (Figure S1A). The randomization and data depletion thus maintained a good balance between exploration of all available data and clinical labels, while optimizing high-accuracy segmentation for desired and most prevalent target classes. After medical images were split into five 80:20 proportions, the DL model, image category, and distribution of clinical labels were causal for segmentation performance (Figure 1). For the skin cancer images (n = 13,786), five-fold random shuffling and 80:20 splits resulted in approximately 11,000 training and 2,700 test images. Benign or malignant moles and lesions were randomized equally across five training ($\approx$ 91%) and test ($\approx$ 7%) splits for individual data (Table S2). Performance of the L_{MI} models from one split matched with that of the L_{MI} models from other data splits, but a > 5% difference was calculated for data split 2. And all the T_{II} model performance remained similar across all five splits (Table S2). In the prostate core biopsy data (n = 244), five-fold randomization into 80:20 splits resulted in approximately 195 training and 44 test images. Deep learning models for segmentation of regions with tumors were trained using approximately 177 training and were evaluated using 49 test images. This resulted in $\approx$ 8% (17 images) of training and 4-8% (4 images) of test data without tumors, and 90% of training and 91-93% of test data images with tumors. The computed tomography DICOM data set contained 45,424 2D images and L_{MI} DL models were trained using approximately 13,000 images with kidneys (36%) and 23,000 (64%) without kidneys. Evaluations were performed with approximately 3,200 (36%) images with and 5,800 (64%) images without kidneys in the test data (Table S2). For this data set, large numbers of data with other organs or only non-target pixels did not seem to impact model performances significantly (Figure S3C; Table S4). These results indicated that blind and random 80:20 splitting did not seem to impact model performance for simple segmentation tasks with medical image data sets with limited heterogeneity (as shown in Figure S3C; Table S4.

The data depletion experiments (Figure S1B) used restrictive bootstrapping, where a single 80:20 split was further processed by randomly moving 20% data to test and using the remaining for training. A similar trend of proportional distribution of clinical labels and

images were calculated in the data depletion experiments (Figure S1B). Performance trends calculated by 80:20 split DL experiments were then compared to models trained under lower data availability. Table 2 shows the median values calculated from the depletion experiments for non-Gaussian distributions of performance metrics for each image modality. Both L_{MI} and T_{II} models exhibited a gradual decrease in performance as training data were depleted across all three image types. Differences between L_{MI} and T_{II} model performance were statistically significant (Mood's median test, $p < 0.05$) for all skin cancer training data regimens, underscoring the impact of heterogeneity and Gaussian distribution of clinical labels on deep learning. Under low data availability (10-20% training data), the transfer-learned T_{II} models outperformed L_{MI} with higher Dice and AUROC scores for skin cancer segmentation, and both DL models achieved similar Dice and AUROC scores when 40 and 60% of the training data were available. The T_{II} models achieved higher sensitivity (higher true positive pixel segmentations for skin cancer) for all data splits, while both models performed equally well for specificity in reducing false positives by not segmenting non-target pixels as tumors. The decrease in training data availability for prostate core biopsy images, both T_{II} and L_{MI} models demonstrated an overall gradual decrease in performance of AUROC and sensitivity scores with statistically significant differences (Mood's median test, $p < 0.05$) calculated at 10% and 60% (Table 2). For the CT images, all DL models exhibited an increase for AUROC, Dice and sensitivity scores with increase in training data size. For this data set as stated earlier, except at the 10% training data mark for the Dice score, L_{MI} marginally outperformed T_{II} models (Table 2). The data depletion experiments can also be used to find out the least numbers of images required to train a DL model with satisfactory performance.

State-of-the-art (SOTA) and grand challenge results from the data sets used in this study

- Top-5 ISIC image segmentation results (Tumor lesion boundary segmentation) (link to ISIC Live results) (ISIC, 2019): The objective of this challenge was to submit automated predictions of only lesion segmentation boundaries of both malignant and benign skin cancers from dermoscopic images. The top scoring learning methods used an ensemble of deep learning methods and achieved Dice scores of 0.915, 0.914, 0.911, 0.906 and 0.904 respectively. While we optimized for binary segmentation of the full volume of pixels in the benign and malignant skin cancers rather than boundaries, and achieved median Dice coefficients that were comparable at 0.8273 and 0.8857 for L_{MI} and T_{II} models.
- Top-5 Gleason 2019 challenge results (link to Gleason-2019 results) (Gleason, 2019): The objective of this challenge was multi-class segmentation of prostate core biopsy Gleason grades from task 1: Pixel-level Gleason grade prediction and then task 2: Core-level Gleason score prediction. While we focused on a modified version of task 1 by segmenting core images based on prediction and classification of any of the Gleason-grade 3, 4 or 5 tumor labels (all or any of these were considered as tumors). The top 5 ranked submissions reported 0.9594, 0.8295, 0.9096, 0.8832 and 0.8896 for benign pixel classification accuracy, and Gleason grade accuracies of 0.3083, 0.4375, 0.4450, 0.4547 and 0.4023 were reported. The average accuracies were 0.6339, 0.6335, 0.6773, 0.6690 and 0.6460 for the top 5 ranked submissions. Our median benign tissue or non-tissue segmentation accuracy (true negative rate) were higher at 0.9572 (for L_{MI}) and 0.9761 (for T_{II}), and median Gleason grade tissue segmentation accuracy (true positive rate) were also higher at 0.9520 (L_{MI}) and 0.9059 (T_{II}).
- Top-5 results in the KiTS 2019 kidney-segmentation challenge results (link to KiTS-2019 results) (ImageNet, 2012): The goal of this challenge was to match ground-truth semantic segmentation for arterial phase abdominal CT scans of 300 unique kidney cancer patients who underwent partial or radical nephrectomy, and the task was to semantically segment kidneys and kidney tumors from these abdominal CT scan slices. We focused only on kidney segmentations from these images where binary segmentations may also kidney tumors as we both tumor and non-tumor kidney ground-truth masks were combined. The top five kidney-only segmentation Dice score results in the competition were: 0.9794, 0.9793, 0.98, 0.9791 and 0.9772. The median Dice scores for kidney segmentation calculated using five fold repeats in this study were 0.9597 (for L_{MI}) and 0.9598 (for T_{II}) and $\approx$ 0.96 for both models and matched the SOTA numbers.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation metrics

Four metrics – AUROC, Dice (F1) score, sensitivity (true positive rate), and specificity (true negative rate) that are frequently used to evaluate performance of DL models for image segmentation were used in this study (Taha and Hanbury, 2015; Rana et al., 2020). A segmentation was considered to be a true positive (TP) when tumor tissue pixels were correctly segmented. When the non tumor tissues or non target pixels were detected as tumors, the segmentation was considered to be a false positive (FP) or a type I error. When non-tumor tissues or non target pixels were not segmented it was considered to be a true negative (TN). The inaccurate segmentation of non tumor tissues or non target pixels was considered to be a false negative (FN) or a Type II error. Sensitivity or true positive rate (TPR) or recall was calculated as $TP/TP+FN$. In the context of the current study, this denoted the percentage of ground truth pixels with the target class segmented correctly as lesion/mole/tumor (skin images), tumor (prostate core biopsy), or organ (kidney CT) by the DL models. The specificity or true negative rate (TNR) = $TN/TN+FP$ was calculated using the percentage of the DL models detection that matched the non target class pixels from the ground truth clinical masks for the three data sets. The AUROC for estimating the overall model performance across all test images was calculated from the area of the curve created by plotting TPR against FPR at various threshold (FPR values of 0, 0.5 and 1) settings. The Dice (F1) score was defined as the

harmonic mean of precision and sensitivity, where the precision = $TP/TP+FP$. In the context of the current study, precision denoted the percentage of the DL models predicted segmentation pixels that matched the ground truth clinical labels. A Dice score = $2 * Precision * Recall/Precision + Recall$ was used to measure the segmentation accuracy of the DL models. Sensitivity and specificity together signified the TP and TN pixels detected by the model, which were confirmed by the color overlay maps to infer location specific correctness of the models. And segmentation overlay maps with detailed true and false positive, true and false negative regions were used to analyze the performance of the trained models in determining the correct location of the segmentation outputs. Dice score is a standard and widely accepted metric used for segmentation (Ghosal et al., 2021; Rana et al., 2020) and was used for the 2-D spatial information from images. We propose a metric based on AUROC, Dice score, sensitivity and specificity for calculating the differences in performances between the two models:

Given the same evaluation metric, m of choice, we define Δ_m as follows:

$$\Delta_m = \Delta_{T_{II}-L_{MI}}|_m = S_m^{T_{II}} - S_m^{L_{MI}} \quad \text{(Equation 1)}$$

where S represents one previously unseen test image or a distribution of images from a particular test data. m represents either AUROC, Dice score, sensitivity or specificity. Thus:

- $\Delta_m > 0 \Rightarrow T_{II}$ performs better than L_{MI} for "m" $\in$ [AUROC, Dice score, sensitivity and specificity].
- $\Delta_m = 0 \Rightarrow T_{II}$ performs as well as L_{MI} for "m" $\in$ [AUROC, Dice score, sensitivity and specificity].
- $\Delta_m < 0 \Rightarrow L_{MI}$ performs better than T_{II} for "m" $\in$ [AUROC, Dice score, sensitivity and specificity].

Statistical methods

Binary segmentation masks predicted by individual DL models were used to calculate the AUROC values and Dice scores for each image within a particular test data set. The mean, median and standard deviations of Dice scores and AUROC values associated with individual images segmented by either L_{MI} and T_{II} models were used for performance evaluations. The overall mean, median and standard deviation of sensitivity and specificity from a particular test data set were used to assess type I (false positive) and type II (false negative) errors for model comparisons used in this study. The Shapiro-Wilk test (Razali and Wah, 2011) showed that all four metric distributions deviated from the normality assumptions (Figure S2). Yeo-Johnson transformations (Weisberg, 2001) of the distributions also did not achieve normality assumptions (Figure S2). Non-parametric Mood's median test was used to test the significance between differences of medians of L_{MI} and T_{II} performance values (Mood et al., 1963; Zar, 2013). The null hypothesis for the Mood's test was that medians of the populations of test images segmented by L_{MI} and T_{II} models were equal (i.e., both models perform equally well for segmentation of a certain image and tumor type). The grand median (median of medians) of all the data was then computed. A contingency table was created by classifying the individual AUROC, Dice, sensitivity and specificity values for each test image from a particular data set as being above or below the grand median of the distributions being tested. This contingency table was used to compute the test statistic and p value. A p value of < 0.05 rejected the null hypothesis, and indicated that observations from L_{MI} and T_{II} did not come from the same distribution and showed statistically significant differences. This process was used to calculate and compare the means, medians and standard deviations of AUROC, Dice score, sensitivity and specificity values for the segmentation of skin, prostate cancer and kidneys by T_{II} and L_{MI} models across all data regimes and images. Results from these comparisons were then further processed using a five% threshold criterion to identify 80:20 data split and models. Corresponding output images, segmentation masks and associated L_{MI} and T_{II} models from this 80:20 data split were selected for further analysis. A non Gaussian distribution of AUROC and Dice scores from the T_{II} or L_{MI} models (Figure S3) was calculated for each of the three image data sets. A higher Dice score indicated superior segmentation performance for a particular image type and target pixels by the trained model (Dice, 1945; Taha and Hanbury, 2015). Differences between Dice scores were then used to rank order and compare the best and worst segmentation accuracy achieved by L_{MI} or T_{II} models for individual test images (Table 1).

Representative images from this process were analyzed by comparisons with their clinical ground-truth while Grad-CAM outputs were generated to explain the model performance. Suitability and comparisons of T_{II} and L_{MI} models for segmentation of images with benign and malignant skin cancer are shown in Figure 2A and 2B respectively and additionally in Figure S2A and S2B, the prostate core biopsy data set in (Figures 3 and S4C), and the kidney CT image data set in Figures 4 and S4D. Subsequently test images from the selected 80:20 split were assigned ground truth clinical diagnoses to evaluate the distribution and performance of the key metrics achieved by L_{MI} and T_{II} models for specific clinical outcomes (Figure S3; Tables S3 and S4). Figure S4A illustrates examples for the skin cancer data that could not be assigned a clinical diagnosis and were denoted as intermediate.

Algorithm 1

The unique aspect of the Algorithm 1 is that it is not limited to the data from this study, and can be utilized with other data sets to enrich high-quality segmentation tasks. The standalone models and Algorithm 1 reported in this study when tested with out-of-distribution (OoD) data will not have access to true ground-truth segmentation to carry out the evaluation step. In these circumstances, historical performance and Dice scores, AUROC and sensitivity and specificity values on similar or related image modalities used for training (Tables 1 and 2) may be used. The trained models and their performance from this study closely matched the ground-truth data metrics, and if provided with a small amount of unlabelled OoD data, these models can also be retrained or fine-tuned and used with the Algorithm 1. In situations with models with high false positive outputs, the algorithm recommends discarding or retraining to obtain

Algorithm 1. Segment Medical Images

Step 1: ***Model Training***: Input: X_{train}: Raw input images from chosen medical image modality, Y_{train}: Input images labels (binary segmentation masks). Five-fold repeats of model training based on:

L_{MI} - Learning from medical images with random initialization,

T_{II} - Learning using ImageNet initialization and fine-tuning using medical images

Step 2: ***Generate Outputs***: Given L_{MI} and T_{II} trained models from five-fold repeat training runs and test data: X_{test}. Generate outputs:

O_{MI}: L_{MI} model segmentation output from average of 5 runs;

O_{II}: T_{II} model segmentation output from average of 5 runs;

O_{Uni}: $O_{MI} \cup O_{II}$ (union);

O_{Int}: $O_{MI} \cap O_{II}$ (intersection)

Step 3: ***Iterative Selection***: Choose a threshold beyond which a model output segmentation is considered acceptable. This can be AUROC and/or Dice score >= 0.9 (a commonly acceptable threshold).

Initialize: $List_{AUROC}$ = [] (an empty list), $List_{Dice}$ = [] (an empty list)

if ground-truth labels are available:

for output: output $\in$ [, O_{II}, O_{Uni}, O_{Int}], X_{test} and Y_{test}:

$AUROC_{eval}$ = AUROC (output)

$Dice_{eval}$ = Dice score (output)

$List_{AUROC}$.append($AUROC_{eval}$)

$List_{Dice}$.append($Dice_{eval}$)

if minimum($List_{Dice}$) >= 0.9:

Choose output corresponding to maximum($List_{AUROC}$).

else if maximum($List_{Dice}$) < 0.9:

Choose output corresponding to maximum($List_{Dice}$).

else if $Dice_{eval} \in List_{Dice}$ is NaN (not defined) [when there are no positive signatures in ground-truth]:

Evaluate FPR (False Positive Rate) for output $\in$ [O_{MI}, O_{II}, O_{Uni}, O_{Int}].

Choose output corresponding to lowest FPR value.

else if ground-truth labels are NOT available: Refer model outputs to human-experts for visual cross-examination:

if high false positives are observed, go back to ***Step 1***,

else accept output with lowest false positives and/or lowest false negatives depending on task.

models with lower false positives or within an acceptable threshold. The algorithm also accounts for the possibility that false positive (or false negative) errors can only be automatically estimated when ground-truth labels are present. On the occasion that ground-truth labels are not readily available such as during primary image acquisition and or initial analysis, the algorithm defers to human experts to review the outputs visually.

Visual explanations of binary segmentation

Grad-CAM, a recently introduced interpretability method, follows the class-activation mapping (CAM) approach for localization (Zhou et al., 2016), and enables the modification of neural network architectures performing classification or classification based on segmentation. During the Grad-CAM operation, the FC layers are replaced by convolutional layers. Grad-CAM uses a gradient corresponding to a certain target class that can be fed into the final convolutional layer of a network to produce an approximate localization (heat) map of the important regions in the image for each target class. The subsequent global-average pooling (Lin et al., 2013) yields class-specific feature maps. For both L_{MI} and T_{II} models, the input image was first passed through the DL model architecture with the trained weights. Grad-CAM then operated on this trained model and the input image to generate a heatmap. The default and widely used set of parameters for Grad-CAM tuning described in Selvaraju et al. (2017) where the layer being explained is set to the last layer of the trained neural network were used in this study. The Grad-CAM function was also queried with a set of target prediction classes that we wanted to explain. As there were five trained L_{MI} and T_{II} models resulting from the five-fold repeats used in the study, each image in the test data set had five corresponding Grad-CAM outputs for each target class. These five outputs were then averaged to generate the final Grad-CAM heatmap. The L_{MI} and T_{II} model outputs in Figures 2, 3, 4, 5, and S4 show the intersection of the five individual repeats. They were obtained as follows:

$$MO = \bigcap_{n=1}^{5} mo_{Run_n} \quad \text{(Equation 2)}$$

For both L_{MI} and T_{II} models, MO was the final model output and mo_{Run_n} was the output for the nth run, the averages of the Grad-CAM outputs of the five runs were calculated. The final Grad-CAM outputs from L_{MI} and T_{II} models, where GC_{Run_n} is the Grad-CAM output for Run_n, were calculated as follows:

$$GC_{Out} = \frac{1}{5}\sum_{n=1}^{5} GC_{Run_n} \quad \text{(Equation 3)}$$

Interpretability of model learning strategy

UMAP constructs a high dimensional graph representation of the data then optimizes a low-dimensional graph to be as structurally similar as possible (McInnes et al., 2018). The number of approximate nearest UMAP neighbors used to construct the initial high-dimensional graph were: five for skin cancer images, two for prostate core biopsy images, and 20 for kidney CT images. Optimization experiments using different neighbor values showed that when analyzing larger data manifolds in high dimensions, the lower numbers constrained the neighboring points and pushed UMAP to focus more on the local structures. While the higher numbers pushed the UMAP towards representing the larger structures. For tighter grouping of individual lower dimensional points, the minimum distance between points in the low dimensional space was kept at 0.01. Figure S5 shows the UMAP representations of the penultimate layer of the VGG-UNet architecture for L_{MI} and T_{II} models for skin cancer, prostate core biopsy and kidney CT test data. Figures S5Ai and S5Aii, S5Bv and S5Bvi, and S5Cix and S5Cx show the two dimensional UMAP embeddings for the feature map generated for target class 0 and 1 for L_{MI} models, and S5Aiii and S5iv, S5Bvii and S5Bviii, and S5Cxi and S5Cxii for the T_{II} models.