January 2023

Improving Cancer Classification With Domain Adaptation Techniques

Juliana Veira

Improving Cancer Classification with domain adaptation techniques

A Thesis Submitted to the Yale University School of Medicine

in Partial Fulfillment of the Requirements for the Degree of Doctor of Medicine

By

Juliana Veira

2023

Abstract:

Improving Cancer Classification with domain adaptation techniques

Background: As the quantity and complexity of oncological data continue to increase, machine learning (ML) has become an important tool in helping clinicians better understand malignancies and provide personalized care. Diagnostic image analysis, in particular, has benefited from the advent of ML methods to improve image classification and generate prognostic information from imaging collected in routine clinical practice [1-3]. Deep learning, a subset of ML, has especially achieved remarkable performance in medical imaging, including segmentation [4, 5], object detection, classification [6], and diagnosis [7].

Despite the notable success of deep learning computer vision models on oncologic imaging data, recent studies have identified notable weaknesses in deep learning models used on diagnostic images. Specifically, deep learning models have difficulty generalizing to data that was not well represented during training. One potential solution is the use of domain adaptation (DA) techniques, which improve the generalizability of a deep learning model trained on one domain to better generalize to data of a target domain.

Techniques: In this study, we explain the efficacy of four common DA techniques – MMD, CORAL, iDANN, and AdaBN - used on deep learning models trained on common diagnostic imaging modalities in oncology. We used two datasets of mammographic imaging and CT scans to test the prediction accuracy of models using each of these DA techniques and compared them to the performance of transfer learning.

Results: In the mammographic imaging data, MMD, CORAL, and iDANN increased the target test accuracy for all four domains. MMD increased target accuracies by 3.6 - 45%, CORAL by 4- 48%, and iDANN by 1.5-49.4%. For the CT scan dataset, MMD, CORAL, and iDANN increased the target test accuracy for all domains. MMD increased the target accuracy by 2.0 – 13.9%, CORAL by 2.4 - 15.8%, and iDANN by 2.0 – 11.1%. in both the mammographic imaging data and the CT scans, AdaBN performed worse or comparably to transfer learning.

Conclusion: We found that DA techniques significantly improve model performance and generalizability. These findings suggest that there's potential to further study how multiple DA techniques can work together and that these can potentially help us create more robust, generalizable models.

Acknowledgments

Thank you to my thesis advisor and research mentor, Dr. Sanjay Aneja, who supported me throughout this project and mentored me throughout the past three years. He has taken time from his busy schedule to answer my questions about projects I have worked on and to advise me on various aspects of my career. I would also like to thank Yongfeng (Miles) Hui, whose work contributed greatly to this project and without whom this thesis would not have been possible.

I am also incredibly grateful for the support of my family and friends. My husband's endless support and encouragement have been my driving force throughout these four years. My parents' tremendous sacrifices have paved the way for my siblings and me to reach for our dreams. Thank you to my amazing siblings who have always pushed me to be a better person. I am lucky to have them, and many more, who helped me through this journey.

TABLE OF CONTENTS:

Introduction:

Artificial Intelligence in Medicine

Artificial intelligence (AI) refers to the ability of a machine or computer system to perform tasks that typically require human-like intelligence, such as learning, problem-solving, and decision-making. The use of AI in medicine has represented numerous new possibilities in the clinical management of patients. In particular, medical specialties that heavily rely on images, such as pathology, oncology, and radiology, show the greatest promise for the use of AI in clinical applications [8]. Most commonly in these medical specialties, AI is being used to aid physicians in diagnoses, disease prognoses, image segmentations, individualized treatment planning, and outcome predictions [9, 10]. Many of these tasks can be time-consuming and vulnerable to physician bias, but AI has allowed for automation that has reduced physician workload and increased accuracy and efficiency.

AI has opened new doors for early diagnosis and a better understanding and management of diseases. Predictive models used in internal medicine have been capable of making complex diagnoses. They have also been used to classify tumors as benign or malignant, and to predict continuous values such as survival time or treatment dose requirements from a set of input data [8, 11]. The performance of these models has shown accuracies equal to or greater than specialists in those respective fields. As a result, today we are closer than ever to having AI become a regular part of the whole medical field.

ML is a subset of AI that involves the use of algorithms and statistical models to allow a system to learn from and make predictions or decisions based on data, without being explicitly programmed to perform the task, mimicking characteristics of human intelligence. For example, an artificial neural network (NN) is a type of ML model that teaches computers how to process

data in a method similar to how the human brain functions. NNs have interconnected models of mathematical cells, or "neurons," that process and transmit information by collecting dendritic-like input into a weighted sum that triggers an axonal-like output through a nonlinear activation function. Input data is processed through the layers of neurons, and the output is a prediction or decision made by the model. NNs are trained using a large dataset, and the connections between neurons are typically adjusted through a process called backpropagation, in order to make more accurate predictions [12]. NNs start as generic predictive models that once given input data can learn to perform specific tasks and have been used for a wide variety of applications, including image and speech recognition, and natural language processing, among many others.

NNs can have different numbers of layers. A single-layer network it's called a multilayer perceptron while NNs that go beyond the classical shallow structure, are known as deep learning (DL) [8]. DL can stack multiple layers of simple, trainable features leading to a hierarchical learning model structure. These models can be used to approximate more complex functions and their capacity is roughly proportional to the number of synaptic weights or parameters [12]. DL is particularly well-suited for tasks that involve analyzing and interpreting complex and unstructured data, such as medical images.

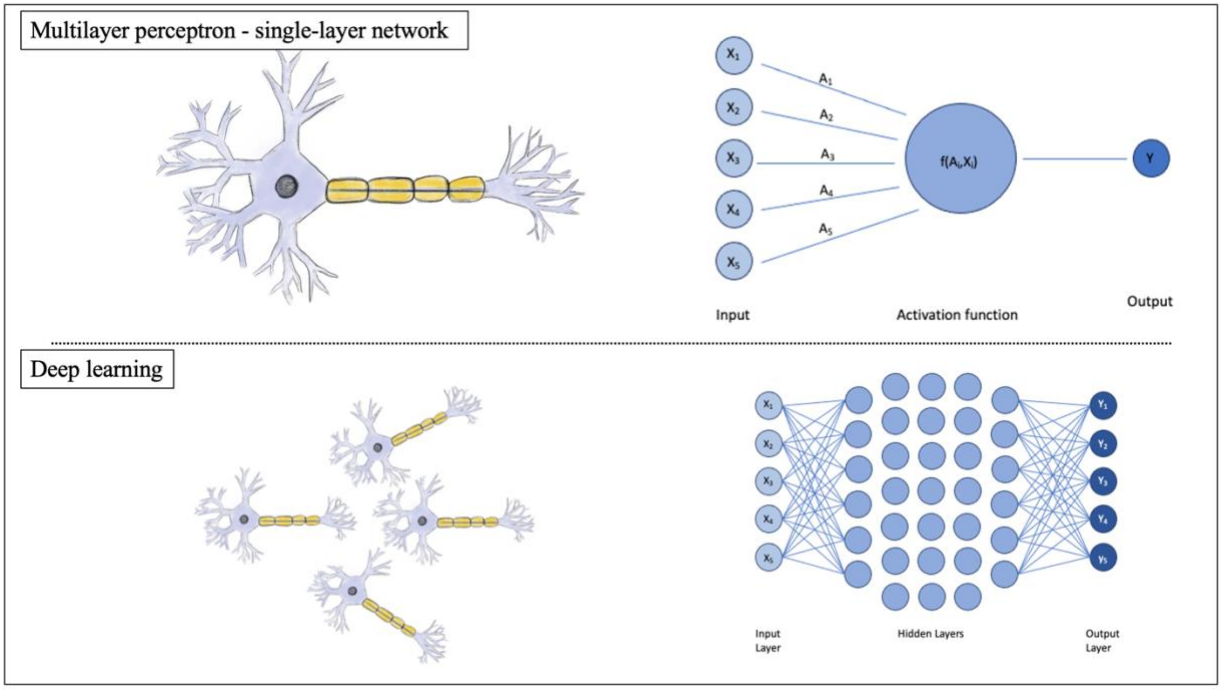

Fig. 1 Demonstration of a Multilayer perceptron and deep learning a neural network

Convolutional Neural Networks in Medicine

In medicine, the most successful type of model used to analyze images has been the convolutional neural networks (CNNs) [9]. CNNs are a type of neural network that are particularly well-suited for analyzing and interpreting complex and unstructured grid-structure, such as images and audio. They were inspired by the human visual system to use the spatial arrangement of data within medical images [12]. The principal difference between CNNs and other types of neural networks is the use of convolutional layers, which apply a set of filters to the input data to extract features and patterns.

In CNNs input data is converted into a numerical representation and passed into the first layer of neurons. Subsequent data is passed through a series of intermediate layers that each

learns to perform a specific task. The input for subsequent layers is the output of the previous layer. In this architecture, each neuron only responds to a specific area of the previous layer, forming an output known as an activation map [12, 13].

An activation map shows the effects that a given filter can have on the input data of a layer in a CNN. Each layer in a CNN is scanned by a convolutional kernel of a fixed size to create a features map. The output of the convolutional layers is passed through a series of fully connected layers, which combine the extracted features to make a prediction or decision. Training of these CNNs is accomplished by minimizing a loss function between the desired input and output. This is done using partial derivatives or gradients of the loss function with respect to the parameters being used. Each neural connection is then given a weight which, in the training phase, is continuously being optimized by the kernels [13].

One of the main advantages of CNNs is their ability to work with changing data which enables them to perform well on tasks such as object recognition, where the position and orientation of the object in the image may vary. CNNs have been proven to be useful in visualizing anatomical structures in different modalities. For example, Moeskops et al. showed that CNNs could be trained to perform multi-organ segmentation on brain MRI, breast MRI, and cardiac CTA [14-16]. Additionally, significant work has been done on using CNNs to segment organs at risk (OARs) during radiotherapy from CT images. OARs are healthy tissue or organs located near a targeted malignancy in a patient with cancer. Their proximity to the clinical target site makes them likely to suffer damage from irradiation when the patient is undergoing treatment. Accurate segmentation of OARs is critical in minimizing the toxicity to these healthy structures. The current standard of care is a manual delineation of the OARs, but thanks to the

implementation of CNNs on CT images, many OARs such as the spinal cord, mandible, parotid glands, submandibular glands, larynx, pharynx, eye globes, optic nerves, optic chiasm, and cardiac structures have been successfully segmented [17-19]. The CNN's architecture of multiple convolutions allows its networks to first extract simple features, such as shapes or edges, and incrementally gain complexity until it is able to identify full organs. For organs with poorly visible boundaries on CT, adding information from MR images can improve their segmentation.

Segmentation of medical images requires the differentiation of pixels belonging to organs and malignant lesions from those of background information found on CTs or MRIs [13]. To use CNNs on medical images, local features must first be extracted in an unsupervised method that aims to discover patterns within the data. Parameters learned from this data are then stacked to allow high-level features to become evident [12]. Because this method is done in an unsupervised fashion, some of the extracted features may be irrelevant or redundant and must therefore be assigned a null weight to signal it as non-important.

These advancements promise to drastically reduce the workload on radiologists and increase the accuracy of the results. For example, over five hundred thousand people are diagnosed with head and neck cancer around the world every year, and treatment of these tumors often puts patients at risk of adverse effects from irradiation [19, 20]. To prevent this, OARs are segmented manually by radiologists to avoid radiation in these areas. This task is not only time-consuming, often taking up to eight hours for cardiac structures and greater than four hours for most other cases, but can be highly variable in accuracy. Using CT scans and CNNs, researchers have been able to surpass the performance of experienced radiographers [20].

The Paradigms of Machine Learning

Within ML, there are three major types of learning paradigms - supervised learning, unsupervised learning, and transfer learning. These three approaches differ in how the model is trained and the type of data it is trained on. In supervised learning, a model is trained on labeled data and its main goal is to learn a function that maps input data to the correct output labels. The model can then apply what it has learned about the relationship between the input data and the corresponding labels to new, unseen examples and make predictions about this data. Most commonly, supervised learning is applied to classification tasks where a model is trained to label a given input, or to regression, where a model is trained to predict a continuous output given a certain input value. Because these models are trained to yield the desired output, they have a simpler, more constrained framework to guarantee these results [20]. Although supervised learning is the most commonly used type of ML, the training data used in these models need to be well annotated – including labels for input features and corresponding correct outputs - and is, therefore, more time-consuming, expensive, and requires expert knowledge.

Unsupervised learning, on the other hand, is a type of ML in which a model can use more data as it is not given any labels in training. Instead, the model must discover patterns and relationships in the data on its own. For instance, one class of unsupervised learning techniques aims to minimize the measure of distance between features extracted from the source and the target domains [8]. This is known as the maximum mean discrepancy (MMD), which creates a task-specific loss and learns domain-invariant and semantically meaningful features [12]. Unsupervised learning is most commonly used for clustering tasks, where a model is trained to group similar objects together, or for dimensionality reduction, where a model can be trained to

identify the most important features of a given set of data. Although supervised learning is easier to implement, unsupervised learning tends to be more flexible and adaptable. Most recently, unsupervised training has been used for well-known NLP models such as ChatGPT.

Transfer learning is another type of ML where pre-trained models can be fine-tuned and applied to a different task or data. This allows users to retain knowledge from different but related domains, which makes this technique particularly useful in situations where it may be difficult, time-consuming, or expensive to collect and label new large datasets for training a model. Instead of having to re-train a model, a pre-trained model can be used as a starting point and fine-tuned to perform a new task on a smaller dataset of interest. Another way in which transfer learning can be used is to use a pre-trained model as a feature extractor. The learned features are then used with a new model to perform a specific task [21].

The Challenges of Artificial Intelligence in Medical Imaging

The detection of breast cancer is another medical task in which ML has been demonstrated to have great diagnostic strength. The main reason for the common application of AI to mammograms is the widespread use of this modality all over the world. Mammography screenings have gained large popularity since the 1980s due to their ability to detect breast cancer and decrease mortality by 20-40% [9]. Despite the increase in people obtaining routine mammograms, over six hundred thousand people still die from breast cancer worldwide each year [8]. One of the explanations for this is the significantly high rate of false negative reads of mammograms. Although studies indicated that in a false negative read of a mammogram, 20-60% of the time an indicator of cancer was available to be found on the image, reading

mammograms can be difficult and time-consuming. With an increase in the volume of screenings each year to over 40 million, radiologists are pressured to read them faster, making it difficult to find small abnormalities found in only 0.5% of people screened [22].

Although the increase in the volume of mammographs has created many challenges, recent studies have shown promising results for the use of ML on the classification of these images as either benign or malignant. NNs outperformed five out of five full-time breast imaging specialists with an average increase in sensitivity of 14% [8]. Additionally, they were able to detect cancers in patients with prior negative mammogram reads and among populations with low screening rates.

Despite the notable success of deep learning computer vision models on medical imaging data, recent studies have identified notable weaknesses in deep learning models used on diagnostic images. Most deep networks are trained and tested on datasets with images of the same distribution. Although these models are effective when tested within these parameters, they underperform in settings where they must analyze a related sample from a different target domain such as those of a different site, with a different imaging protocol, a different imaging modality, or a different patient population. For example, studies have shown deep learning models of chest x-rays are unable to generalize to female patients when a substantial portion of their training data is composed of male patients [12, 23]. Similarly, a mathematical model using Bayesian statistics was created to recognize patterns in symptoms to diagnose abdominal pain and meningitis. Although it was successful with these diseases, when it was applied to other symptoms, it was not able to generalize to these new pieces of information [8]. To further investigate this phenomenon, Albadway et al. 2018 studied the generalizability of CNNs models

for outlining glioblastomas [24]. They tested three models - one on a dataset with patients from the same institution as the training data, one on a dataset with patients from a different institution, and one on a dataset with patients from both the same and different institutions to those of the training data. This study showed a very strong effect on performance due to differences in the training versus test datasets.

The reason why these results have been observed is due to poor model generalizability. Generalizability is an important consideration in the development and deployment of AI applications for medical imaging. It refers to the ability of a model to perform well on a range of different data types and samples, rather than just the specific data it was trained on. Medical data can be highly variable and commonly subject to bias. For example, this phenomenon can be seen in datasets of MRIs from different centers where there are differences in machine vendor, software, sequence parameters, and frequency of coils [9, 25]. Most frequently, source and target domains vary in brightness, resolution, or texture while maintaining high-level features like class, object types, numbers, etc. consistent [26, 27]. Differences in protocols and patient characteristics can also create bias. Dataset bias is especially common with deep learning models that are trained on single institutional datasets making them subject to local training biases.

If a model is not able to generalize well, it may perform poorly when applied to new data or situations that differ from the training data rendering it clinically irrelevant [18]. Given the relative heterogeneity of cancer patients and the wide variety of anatomical structures in medical images even across the same imaging modalities, deep learning models within oncology have shown similar limitations in their generalizability [28, 29]. The consequences of incorrect or unreliable results from AI models applied to medical imaging are misdiagnosis or incorrect

treatment plans that could have serious effects on patient care, disease prognosis, and survival [30].

Additionally, there are concerns regarding data distribution shifts that occur when there are differences in the parameters or protocols between the images that the model is learning on and those to which it is being applied [9, 12]. Although these may be of similar modality and on a common object, data distribution shifts can result, especially in medical imaging where it is rare to have labeled, available data from the same center and modality to both train and test on.

There are several ways to improve the generalizability of AI models for medical imaging. One approach is to use a diverse and representative dataset for training, to ensure that the model is exposed to a wide range of data types and characteristics. Another approach is to use techniques such as regularization and data augmentation to reduce overfitting [31, 32]. There have also been multiple efforts at data fusion techniques such as combining disease samples from different data sources. Unfortunately, these have not proven to be effective in solving dataset bias [1]. The major challenge in medical imaging is the limited availability of annotated data. Collecting and annotating medical images can be a time-consuming and expensive process, often requiring experts to define relevant features and to have knowledge of the domains being analyzed [33]. The performance of models is limited by the quality and quantity of datasets and as a result, many models can fail to obtain a large enough dataset to obtain a high level of accuracy. Collecting large enough datasets is especially difficult among cancer patients given known data silos that exist in healthcare. Since data collection automatization is still poor and the amount and quality of training data are dominant influences on an ML model's performance, there is an urgent need to find better ways to collect, annotate, and reuse medical imaging data as well as a

need to find better methods to mitigate local training biases and improve generalizability across different datasets [34].

A solution to both problems is the use of DA techniques. DA refers to the process of adapting an ML model that was trained on a larger, more diverse dataset to perform on a different but related, dataset, oftentimes with fewer annotations [26]. DA is also commonly used for domain shifts, where it can align high-level features between source and target domains while getting rid of low-level features. Additionally, DA in medical imaging has the ability to adapt a model to different imaging modalities. Different medical imaging techniques, such as CT, MRI, and ultrasound, all produce different types of images with distinct characteristics. A model trained on one modality may not perform well on images from another modality, even if the images depict the same underlying condition and anatomy. By using DA techniques, it is possible to adapt a model trained on one modality to perform well on images from a different modality.

Domain Adaptation Techniques

DA techniques offer the potential to train a model using both labeled synthetic data that is often abundantly available and unlabeled real data [25]. They have been effective at mitigating the bias of deep learning models in a variety of settings including chest x-ray analysis [35], diagnostic models [36], and speaker [37], facial [38], and location [39] recognition among many others. Similarly, there have been several proposed DA techniques that have shown some efficacy on computer vision tasks, but their efficacy on diagnostic images in oncology remains unknown [40].

There are several types of DA techniques that can be used to mitigate this problem. These techniques can be broadly classified into two categories: supervised DA and unsupervised DA. Supervised DA techniques assume that labeled data is available in both the source and target domains, and the goal is to use this labeled data to learn a model that can perform well on the target domain. Unsupervised DA techniques, on the other hand, do not assume that labeled data is available in the target domain. Instead, they assume that there is an underlying relation that is shared between the source and target domains. In this study, we focus on three unsupervised DA techniques because they are more general tools, and also include one supervised technique for proper comparison [41].

Four commonly used DA techniques are Maximum Mean Discrepancy (MMD), Instance-level Domain Adaptive Neural Network (iDANN), CORrelation Alignment (CORAL), and Adaptive Batch Normalization (AdaBN). MMD is an unsupervised DA technique that aims to reduce the difference between the source and target domains by minimizing the maximum mean discrepancy between them. To do this, a model is trained to minimize the difference between the mean feature vectors of the source and the target domains while also trying to maximize its classification accuracy on the target domain. By minimizing the distance between the distribution of the source and target domain, the model attempts to generalize well to the target domain.

iDANN is another type of unsupervised DA technique that aims to reduce the difference between the source and target domains by aligning the feature distributions at the instance level. This is done by training a model to transform the source domain data to have the same feature

distribution as the target domain data. At the same time, the model also tries to preserve a high classification accuracy in the source domain.

CORAL is also an unsupervised DA technique that aims to align the correlations between the source and target domains by minimizing the distance between their second-order statistics. To do this, a model is trained to transform the target domain data to have the same second-order statistics as the source domain data. Simultaneously, the model also attempts to preserve the classification performance on the source domain.

Lastly, AdaBN is a supervised DA technique that adapts the batch normalization layers of a neural network to the distribution of the target domain data. Batch normalization is a technique that is used to normalize the activations of a neural network across different batches of data. This helps to stabilize the training process and improve the generalization performance of the model. However, when the distribution of the target domain data differs significantly from the distribution of the source domain data, the batch normalization layers may not be effective, and the model may not perform well on the target domain. To address this problem, AdaBN adapts the batch normalization layers of the neural network to the distribution of the target domain data [42]. This is done by estimating the mean and standard deviation of the target domain data and using these estimates to adjust the batch normalization layers of the model. This allows the model to perform well on the target domain, even when the data distributions between the source and target domains differ significantly.

Statement of purpose:

We aimed to test the utility of applying DA techniques to deep learning models in order to improve their generalizability to a target domain different from which the model had been trained. With the implementation of these methods, we also hoped to be able to apply these models to target domains with insufficiently annotated data that more closely resembles real-life medical imaging applications. To do so, we created models from publicly available training datasets of mammograms from breast cancer screening and thoracic computed tomography (CT) scans for lung cancer screening, to classify lesions on both imaging modalities as benign or malignant. We tested four commonly used DA techniques on each of these datasets and analyzed their performance on four specific domains for each imaging modality – density, image view, mass margin, and mass shape for mammograms, and spiculation, subtlety, margin, and contrast for the CT scans. We expected the application of DAs would successfully improve our deep learning models' accuracies on target domains, as past studies have shown considerable improvements with models of widely used, non-medical imaging datasets. We also expected to find one specific DA that improved accuracies across both imaging modalities on all domains. However, although we did not find one DA that outperformed all others, we did find that three out of the four tested significantly increased the accuracy of the target domain compared to transfer learning. We also expected to obtain higher accuracies on combined datasets of images from both source and target samples, which proved to be consistent with our results. Lastly, we expected each DA to perform uniquely but to observe some consistency in task performance across those that performed best. We found this to be the case and observed sustained agreement across the best-performing DAs compared to transfer learning.

Specific Aims

1. Train a CNN on two datasets of medical imaging using transfer learning.
    a. Divide images into subcategories within four domains using their intrinsic properties.
    b. Randomly assign some to the training set and some to the test set.
    c. Train a CNN on our source training data and evaluated on source test data.
    d. Adjust model parameters and used obtained weights to test it on the target training set.
    e. Obtain transfer learning source validation accuracy (Val Acc), target training accuracy (Train Acc), and target test accuracy (Test Acc).
2. Train CNN with all each of the four DA techniques using saved weights
    a. Save the DA weights at 5 incrementing quantities of epochs to select for highest target training accuracy
    b. Apply weights to obtain Train Acc and Test Acc.
3. Test transfer learning and DA techniques on combined dataset.
    a. Randomly select a certain amount of source test data and target training data to create a combined dataset
    b. Train all four DA techniques and transfer learning on this dataset.
4. Obtain a statistical analysis of the performance of each model.
    a. Test the stability of DA techniques using bootstrap to calculate the confidence interval of each model's performance.

   b. Use a one-sided t-test to compare the accuracies of each DA technique to those of transfer learning.

5. Analyze agreement in performance between each DA technique.
   a. Create heatmaps to show what images that were incorrectly identified by transfer learning were correctly identified by each DA technique.
   b. Obtain quantification of agreement using Kappa statistics.

Methods:

Student contributions:

The experiments were performed using two publicly available datasets. Yongfeng (Miles) Hui, MPH developed CNN models and performed the statistical analysis. Data analysis with medical applications and the generation of figures and tables were conducted by the author (Juliana Veira) with guidance from Dr. Sanjay Aneja.

Ethics statement:

The research was conducted in accordance with the Declaration of Helsinki guidelines and approved by the Yale University Institutional Review Board (Protocol ID: HIC#2000027592).

Participants and Data Collection:

We studied the performance of four DA techniques on mammographic imaging and CT scans from two publicly available datasets– the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) and the Lung Image Database Consortium image collection (LIDC-IDRI) (Table 1). For each imaging modality, we used a separate CNN model. To do this, each dataset was divided into two groups: the training set and the testing set at a 2:1 ratio [43].

The mammography imaging from the CBIS-DDSM dataset contained 1,696 lesions from 1,566 patients at four clinical sites across the United States. Of these lesions, there were 753 cases of calcification and 891 cases of mass findings. The outcome of interest was the

pathological labeling of these lesions - whether malignant or benign – based on regions of interest that were verified by pathologic reports. For this imaging, we considered four domains: density, image view, mass margins, and mass shape. These are commonly found characteristics of mammographic lesions utilized by radiologists to determine whether they are malignant or benign.

CT imaging data from the LIDC-IDRI dataset made up of 1,018 thoracic CT scans from patients in 15 clinical sites across the United States. These scans contained a total of 2,600 lung nodules that were identified by experienced thoracic radiologists. The outcome of interest was the pathological labeling of these lesions from verified pathologic reports. In the case of patients without pathologic confirmation, malignancy was determined by radiologist interpretation. For these scans, we also considered four domains: spiculation, subtlety, margin, and contrast. These are commonly found characteristics of lesions on CT scans utilized by radiologists to determine whether they are malignant or benign.

CBIS-DDSM Dataset			LIDC-IDRI Dataset		
	Training	**Test**		**Training**	**Test**
Density **(N = 1696)**			**Spiculation** **(N = 6342)**		
<= 2	765 (45.11%)	329 (19.40%)	<= 2	4074 (64.24%)	1746 (27.53%)
>2	421 (24.82%)	181 (10.67%)	>=4	365 (5.76%)	157 (2.48%)
Image View **(N = 1696)**			**Subtlety** **(N = 6859)**		
Craniocaudal (CC)	548 (32.31%)	236 (13.92%)	<=3	1694 (24.70%)	726 (10.58%)
Mediolateral oblique (MLO)	638 (37.62%)	274 (16.16%)	>=4	3107 (45.30%)	1332 (19.42%)
Mass Margins **(N = 1638)**			**Margin** **(N = 6859)**		
Spiculated	284 (52.50%)	123 (22.53%)	=5	1981 (28.88%)	849 (12.38%)
Not Spiculated	860 (17.34%)	369 (7.51%)	<=4	2820 (41.11%)	1209 (17.63%)
Mass Shape **(N = 1692)**			**Contrast** **(N = 6859)**		

Regular	368 (21.75%)	158 (9.34%)	True	1637 (23.87%)	702 (10.23%)
Irregular	816 (48.23%)	350 (20.69%)	False	3164 (46.13%)	1356 (19.77%)

Table 1. This table shows the two datasets (CBIS-DDSM and LIDC-IDRI), their respective domains (in bold), and the number of images used in the training and test groups for each domain.

Method Description:

Deep Learning Models

All models used a convolutional neural network with the VGG19 architecture as a baseline. We also used an embedded TensorFlow data augmentation method to generate batches with real-time data augmentation, the parameters of which can be found in Table 2. Unless otherwise specified, we use a Stochastic Gradient Descent (SGD) optimizer, with a learning rate of 0.01 without momentum. Our implementation was developed using Keras and will be publicly available on GitHub. All the equations for these methods can be found in the Data Supplement.

Hyperparameters	
Batch size	64
Total epochs	500
Image size	(116,116,3)
Learning rate	0.01
Optimizer	stochastic gradient descent
momentum	0.0/0.9
regularization	Dropout (0.5)
Image Data Generator	
Rotation range	30
Zoom range	0.2
Width shift range	0.3
Height shift range	0.15
Shear range	0.3

Horizontal flip	True
Fill mode	nearest

Table 2. - Hyperparameter used in the experiments

Domain Adaptation Techniques

Typically, DA assumes access to two related datasets, a target domain (the domain of interest) that lacks labeled data and a source domain that has abundant labeled data (but differs from the target domain in some respect). DA techniques then train a network in a supervised fashion on the labels in the source domain, while simultaneously adjusting model parameters so that unlabeled data from the target domain 'matches' data from the source domain [44].

The DA techniques we study are designed to encourage the network to view images from the source domain and images from the target domain as coming from similar distributions. Two of these DA techniques, MMD and CORAL, use metrics that calculate the distance between two distributions. During training, they align the source and target domains by minimizing the distance metric on latent embeddings generated from the source and target domain respectively [5]. MMD is a kernel method that calculates distances on probability measures, while CORAL calculates the difference of second-order statistics. We apply the MMD loss and CORAL loss to the latent representations output by each linear layer (except for the final prediction).

Adaptive Batch Normalization (AdaBN), is another DA method that we consider. AdaBN makes use of batch normalization (BN) layers to ensure that images from the source and target domains are treated as coming from similar distributions [45]. It makes use of the fact that batch normalization layers capture statistics of their input distribution. In AdaBN, images from the

target domain are used to compute batch normalization statistics that are then applied to images of the source domain. To use AdaBN, we added a batch normalization layer following the VGG19 layers.

The final DA method we study is Domain Adaptation Neural Network (DANN). In DANN, a domain classification branch is added to the network, in parallel to the label classification branch. The domain classification branch is begun with a gradient reversal layer, which when combined with a domain classification loss, encourages the network to learn a representation that is not able to distinguish between samples from the source and target datasets [46]. We use a variant of this method, which we call incremental DANN (iDANN). In this variant, during the training, some samples from the target domain, for which the label classifier is confident, are added to the source domain. Following the original DANN paper, we use SGD with a learning rate of 0.01, decay of 10−6, and momentum of 0.9.

A visual explanation of the architecture of these four DA techniques can be found in Figure 2.

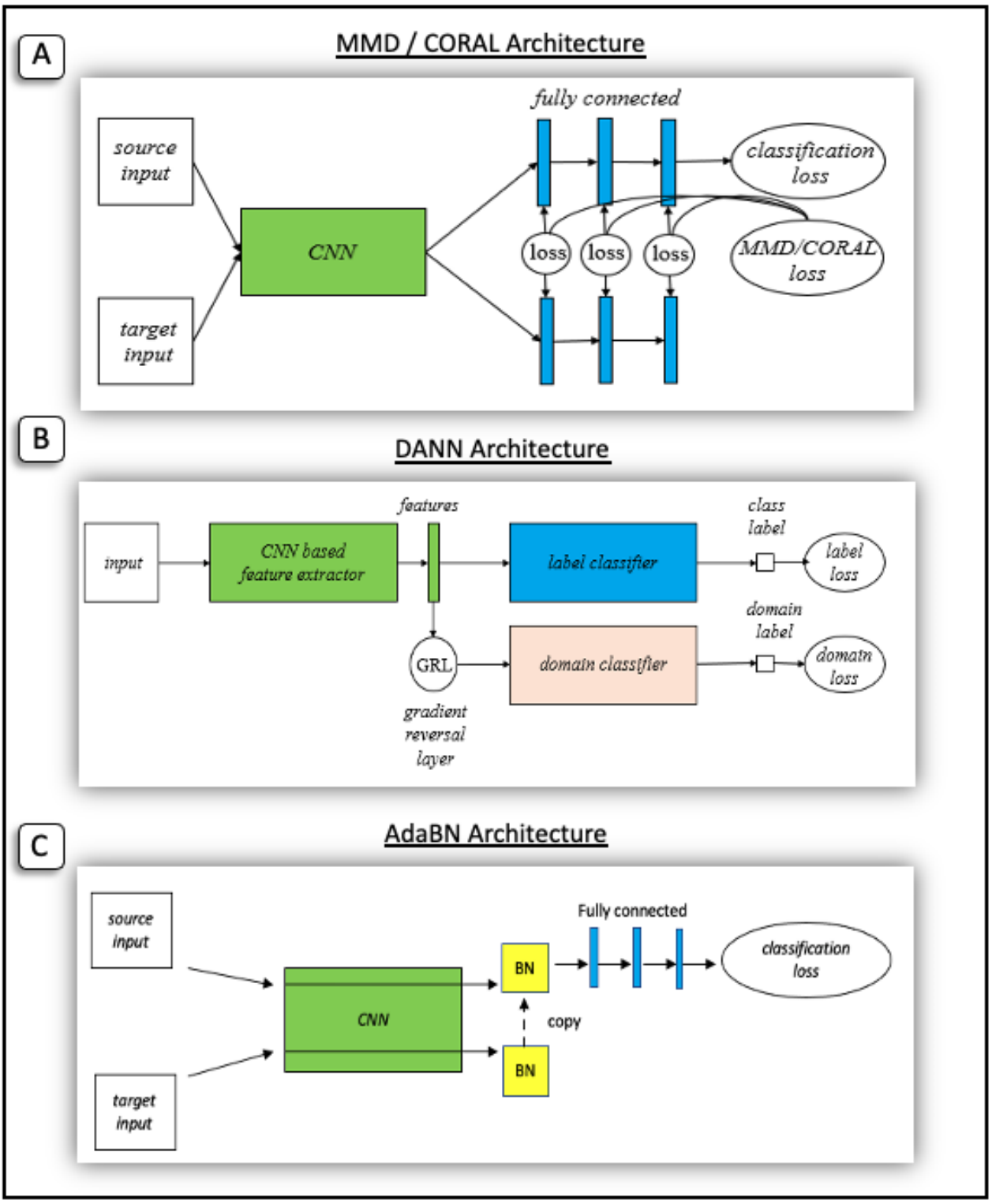

Fig. 2: Visual explanation of the architecture of each adaptation domain method. *Image A* shows the components of the MMD and CORAL architecture: input layers, a CNN block, and fully connected layers. *Image B* shows the components of the DANN architecture: a CNN-based feature extractor, a label classifier, and a domain classifier. *Image C* shows the components of the AdaBN architecture: input layers, a CNN block, and a batch normalization layer.

Domain Adaptation Performance

To evaluate how different DA techniques influence accuracy on target domain samples, we begin by randomly splitting each domain into train (70%) and test (30%) sets. First, we pretrain a model on the source-domain training set. We perform early stopping based on the source-domain test-set and save the model weights for the epoch that achieved the best source-domain test set accuracy. All our experimental variants load these saved weights at initialization.

We compare each DA method against a baseline transfer learning method. In transfer learning, the pre-trained model is further finetuned on the training set of the target domain under a supervised objective. The inputs for MMD, CORAL, and iDANN were from source training data, source training labels, and target training data. Although AdaBN can work using inputs from source training data and source training labels alone, target training data was also used to improve its performance. We saved the DA weights after 50, 100, 150, 200, and 500 epochs to select the highest target training accuracy. Then, we applied these weights to get Train Acc and Test Acc. Following this, we created a combined dataset by calculating what two-thirds of the source test data set and the target training data set would be. We picked the smaller number and then chose this number of items from each test dataset and combined them together.

Statistical Analysis:

To test the stability of DA techniques, we used the bootstrap method to calculate the 95% confidence interval (95% CI) for both the transfer learning and DA techniques. We then employed a one-sided t-test to check if each DA method showed statistically significant improvement in target domain accuracy compared to transfer learning.

Results:

We used two publicly available datasets of mammograms and CT scans, as described above, to compare the accuracy of predictions made by models using one of four different DA techniques to those of transfer learning. In our first set of experiments, we studied the target accuracy of these models on the four domains for each dataset. We found that the DA techniques MMD, CORAL, and iDANN, significantly outperformed the transfer learning on most tasks. AdaBN showed similar performance to transfer learning.

In the mammographic imaging data, MMD, CORAL, and iDANN increased the target test accuracy for all four domains - density, image view, mass margins, and mass shape compared to transfer learning. MMD increased target accuracies by 3.6 - 45%, CORAL by 4-48%, and iDANN by 1.5-49.4%. AdaBN performed comparably to transfer learning. These findings are shown in the boxplots of Figure 3 and Table 2.

For the CT scan dataset, MMD, CORAL, and iDANN increased the target test accuracy for all domains - spiculation, subtlety, margin, and contrast compared to transfer learning. MMD increased the target accuracy by 2.0 – 13.9%, CORAL by 2.4 - 15.8%, and iDANN by 2.0 – 11.1%. AdaBN performed comparably to transfer learning. These findings are shown in the boxplots of Figure 4 and Table 3.

The second set of experiments studied the performance of the four DA techniques on combination datasets. As explained above, these datasets were made by randomly combining images from the source test data and from the target training data. For the mammographic imaging combination dataset, MMD, CORAL, and iDANN all increased the target accuracies in density, mass margins, and mass shape. No one adaption domain caused any significant changes

in accuracy for image view. AdaBN only had a slight increase in accuracy for density. For the CT scan validation combination dataset, all four DA techniques increased the target accuracies in the margin and contrast domains. MMD and CORAL also significantly increased accuracy in spiculation and subtlety, and iDANN in spiculation.

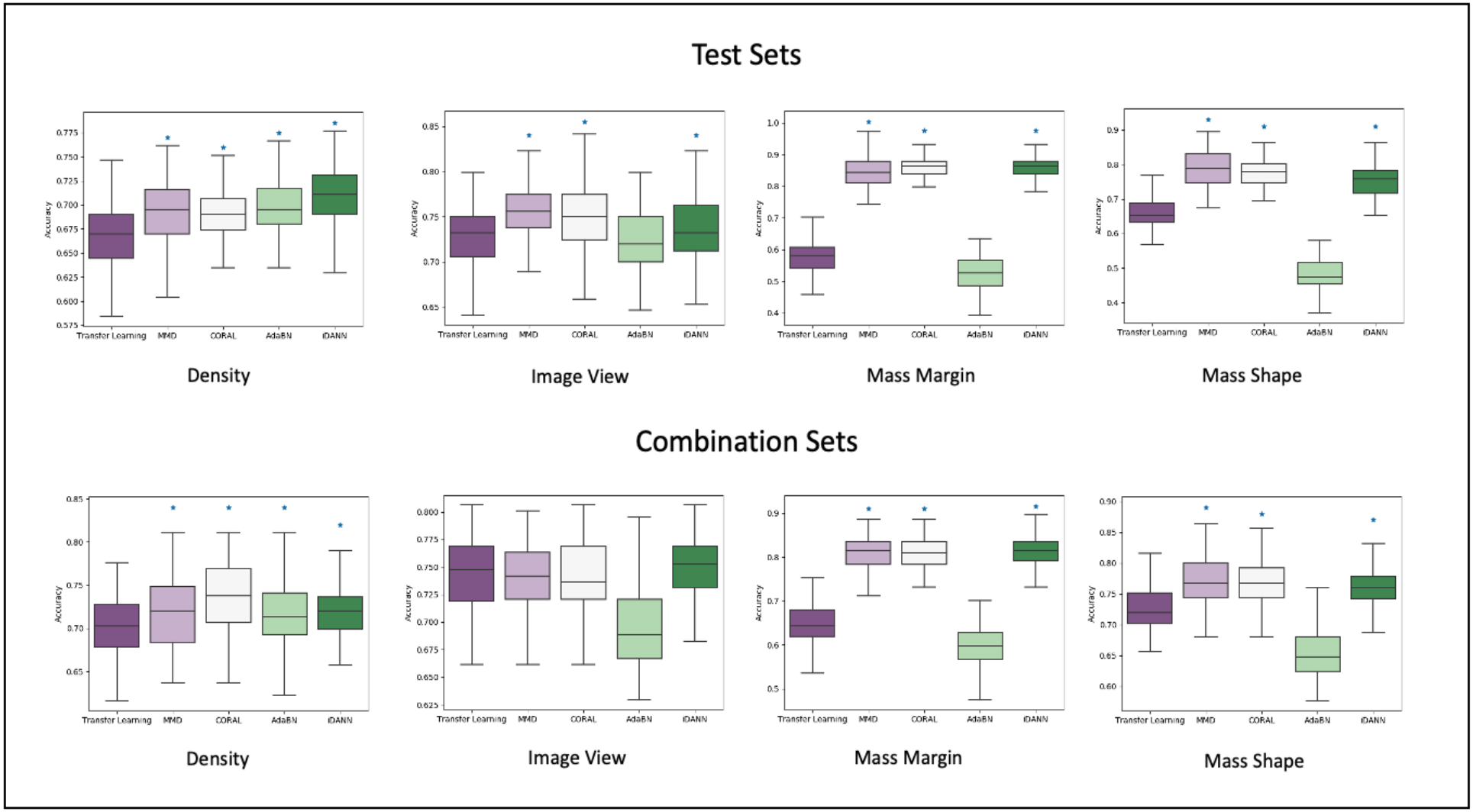

Fig. 3 These box plots, created using bootstrap calculations, show the accuracies achieved by transfer learning and each adaptation domain method on the CBIS-DDSM dataset. Each plot shows the performance of the four adaptation domain techniques on a given domain of interest. The top four plots show the accuracies on the test set and the bottom four plots show the accuracies on the combined validation set. Outliers were not included in these plots. A blue star was placed above a box plot if its results were significantly better than those of transfer learning ($p < 0.05$).

	Accuracy on Target Domain (% change)	Accuracy on Combination Test set (% change)
MMD	3.6 – 45.0	-0.4 – 23.7
CORAL	4.0 – 48.0	-0.8 – 23.7
AdaBN	-25.5 – 4.0	-10.0 – 1.7
iDANN	1.5 – 49.4	0.9 – 24.6

Table 2. Percent change in target accuracy on all domains compared to transfer learning for CBIS-DDSM Dataset

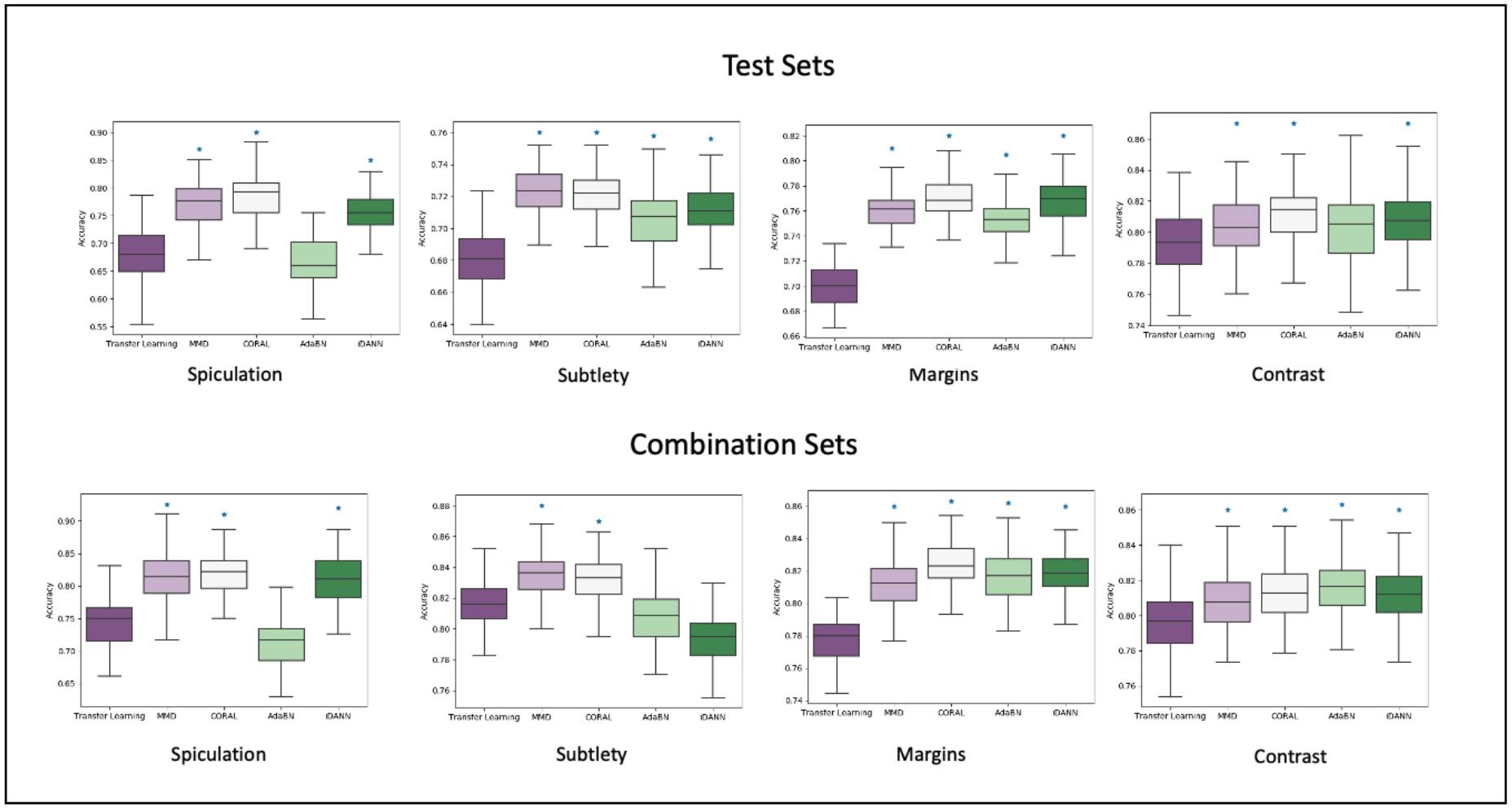

Fig. 4 These box plots, created using bootstrap calculations, show the accuracies achieved by transfer learning and each adaptation domain method on the LIDC-IDRI dataset. Each plot shows the performance of the four adaptation domain techniques on a given domain of interest. The top four plots show the accuracies on the test set and the bottom four plots show the accuracies on the combined validation set. Outliers were not included in these plots. A blue star was placed above a box plot if its results were significantly better than those of transfer learning ($p < 0.05$).

	Accuracy on Target Domain (% change)	Accuracy on Combination Test set (% change)
MMD	2.0 – 13.9	1.3 – 9.3
CORAL	2.4 – 15.8	1.9 – 9.8
AdaBN	-2.9 – 7.3	-3.9 – 4.4
iDANN	2.0 – 11.1	-2.1 – 9.2

Table 3. Percent change in target accuracy on all domains compared to transfer learning for LIDC-IDRI Dataset.

Lastly, we used heatmaps to study the agreement between the four adaptation domain techniques, as can be seen in Figure 5. We showed the images that had been incorrectly predicted by transfer learning and the DA techniques (red) and those that had been incorrectly predicted by transfer learning but correctly predicted by one or more adaptation techniques (green). These heatmaps show that different DA techniques can correctly predict many more images that transfer learning, and that many of these images are correctly predicted by multiple DA techniques. To quantify this agreement between the four tested techniques, we used Kappa statistics. In most of the experiments, the Kappa statistics between MMD, CORAL, and iDANN were over 0.61, except for the image view domain [47]. These findings show us that there's potential to further study the use of multiple DA at once to create more robust, generalizable models.

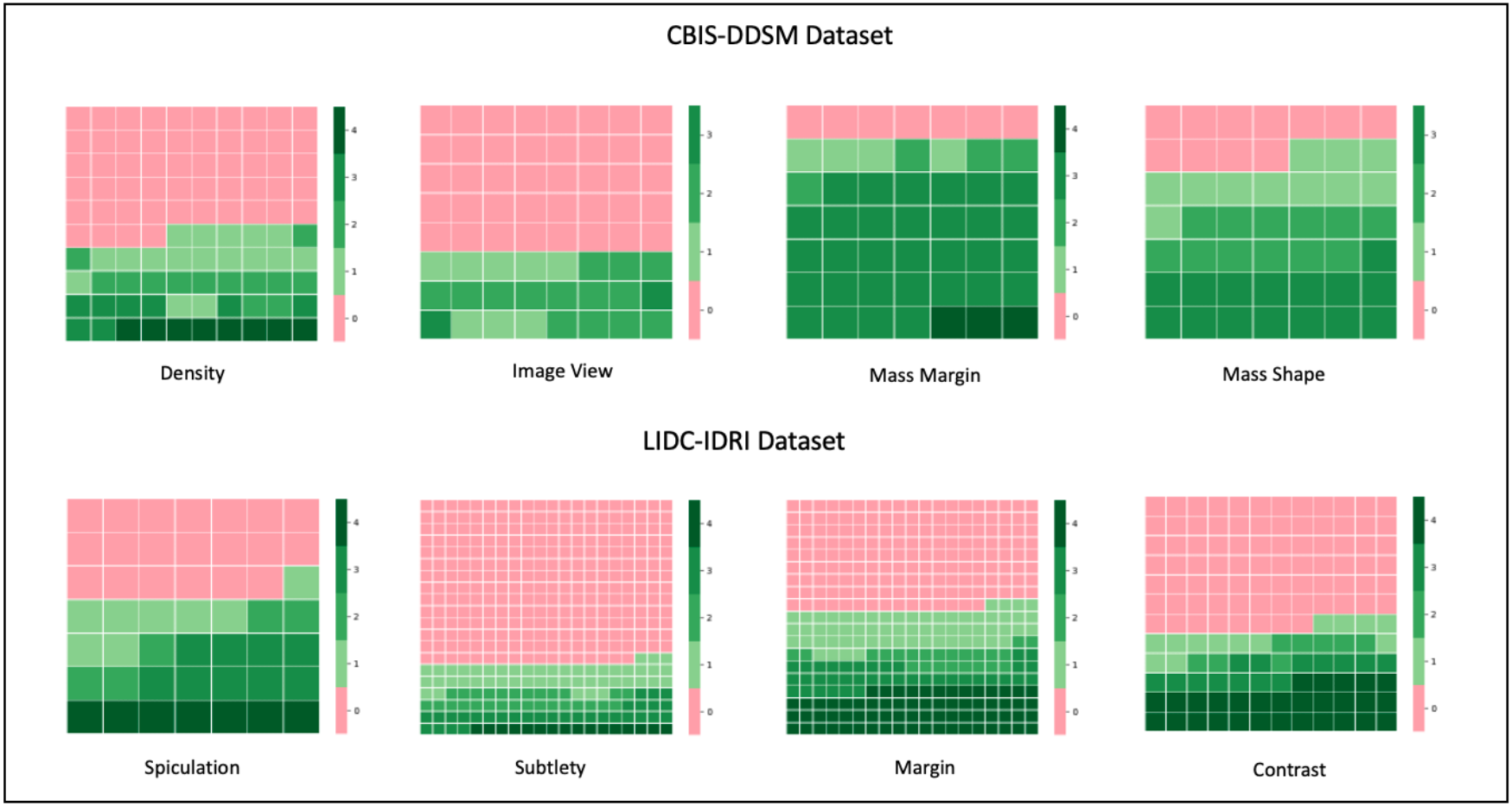

Fig. 5 Each square in the heatmaps represents an image predicted wrong by transfer learning. In red are those images incorrectly predicted by both transfer learning and all DA techniques. In the different shades of green are images that were incorrectly predicted by transfer learning but correctly predicted by one or more of the DA techniques. The light green boxes represent images predicted correctly by only one of the DA techniques. The darker green boxes represent images predicted correctly by multiple adaptation domain techniques. The top four heatmaps show the performance of the four adaptation domain techniques on each of the four domains for the CBIS-DDSM dataset. The bottom four heatmaps show the performance of the four adaptation domain techniques on each of the four domains for the LIDC-IDRI dataset.

Discussion:

The fast evolution of deep learning has led to significant advancements in medical diagnoses, prognoses, and treatment recommendations. Of special importance is the role of CNNs which have shown to be useful in image classification, segmentation, and localization and have been applied to detecting malignancies, cardiac structural anomalies, retinal diseases, and Alzheimer's among other illnesses [3]. Despite outperforming medical professionals in many clinical settings, deep learning networks have shown to have poor generalizability, and efforts to solve this problem have, up to now, proven ineffective.

In this study, we test four DA techniques on two publicly available datasets of mammograms for breast cancer screening and thoracic computed tomography (CT) scans for lung cancer screening. For each of these datasets, we examine four specific imaging domains. For mammograms, the domains are density, image view, mass margin, and mass shape. For CT scans the domains are spiculation, subtlety, margin, and contrast. Although we did not find one DA method that outperformed all others, we did find that using three out of the four tested – MMR, CORAL, and iDANN – lead to statistically significant improvements to the target domain accuracy compared to transfer learning. Moreover, we also found that these DA techniques could be used to improve target domain accuracy on combined datasets - images from both source and target samples - compared to transfer learning. Furthermore, we found that there was largely sustained agreement among the best-performing DA techniques, but they still retained unique properties that made their performance highly specific to the dataset and the target domain, suggesting the possibility of improvements in models with multiple incorporated DA techniques. These findings demonstrate that DA techniques can be of great use to create more robust models that will combat current problems of generalizability.

Our results corroborate past studies that show that the application of DA techniques to deep learning models successfully increased their accuracies on target domains. One of these investigations used the Office-31 dataset to evaluate DA techniques and achieved accuracies significantly higher than with transfer learning [48]. Similarly, Tzeng et al. showed that the accuracy of the target domain in the task from MNIST to USPS was increased by almost twenty percent using CoGAN [49, 50]. Despite these impressive results, these studies evaluate their techniques on widely used benchmarks like Office Dataset, MNIST, USPS, SVHN digits datasets, or other similar datasets composed of synthetic and real data. Our work expands on the findings of previous studies by evaluating the performance of DA techniques on commonly used, real-world clinical image diagnostic tasks – mammographic imaging and CT scans.

Although limited by the lack of training data, studies have proven that deep learning techniques can effectively model mammogram datasets. Carneiro et al. showed how Imagenet can produce an AUC of 0.97(±0.03) on DDSM [51]. Similarly, Li et al proposed DenseNet-II neural network model, with an average accuracy of 94.55% on mammogram datasets [52]. Wang et al. presented a deep learning model that can achieve a discriminative accuracy of 87.3% in breast lesions classification tasks [53]. In our experiment, we showed that by using VGG19 without fine-tuning, the deep learning model can achieve close to 75%-80% validation accuracy in each domain.

Similarly, there have been many studies that show how vulnerable deep learning models are to the effects of dataset bias, ultimately impacting their generalizability [54]. Among others, collecting new data [55] and using data augmentation [56] have all been proposed as methods to solve this problem without success. In our experiments, we observed a similar drop in accuracy

on the target domain of between 2% to 20% when using transfer learning alone but by using DA techniques, we were able to overcome dataset biases and yield more reliable classifiers.

Past studies report conflicting results on the performance of iDANN and deep CORAL, two of the four DA techniques we studied. Some have shown that iDANN performed much better than the Deep CORAL method on MNIST, SVNH, and Syn Num datasets [46] while others show that Deep CORAL outperformed DAN[57] . When we applied these methods to medical images, both produced similar results and significantly outperformed baseline transfer learning. Overall, we did not find a method that performed generally better in both datasets across all four tested domains although CORAL was the best method as it had the highest improvement in accuracy in 5 experiments. Second and third in performance were iDANN, which had the highest accuracy improvement in 2 experiments, followed by MMD, with the highest accuracy improvement in 1 experiment.

AdaBN is a technique that modulates all the batch norm layers' statistics to make each layer receive data from a similar distribution. Other studies that evaluated the performance of this method on the Office-31 dataset and Caltech-Bing dataset obtained favorable results [52]. In our study, AdaBN performed comparable to or worse than transfer learning. This is likely due to this technique being more general than the other three. AdaBN does not require the use of any labeled data from the target domain, which makes it more versatile and applicable to a wider range of scenarios, but in situations where labels are available, it underperforms compared to those techniques that do make use of this additional information.

Challenges and limitations:

There are several limitations to our study. First, we only used one CNN architecture, VGG19, for all our experiments. Previous studies have shown that Deep CORAL and DAN have worked well using VGG16 or AlexNet network, an architecture consisting of five convolutional layers and three fully connected layers. Although we do not have enough information to prove that VGG19 is the best CNN model, it has proven to have overall better transfer learning performance than many other architectures [58]. Additionally, this architecture is widely recognized for its application on medical imaging data [39, 59, 60].

Secondly, we only calculated the MMD loss and CORAL loss between two domains for a single feature layer, and for AdaBN we only added a single BN layer in the network [57]. Although past studies have demonstrated advantages to applying CORAL loss to multiple layers we wanted to ensure that the network structures of all methods were similar [48]. We also tried to mimic real-world settings where CNN models being used have already trained on clinical images and where the weights can be applied to DA network architectures. As BN layers, by design, change the statistics of the network activations between two layers, it can be challenging to add BN layers to a pretrained network without dramatically affecting performance.

Lastly, we only used two types of medical images – mammograms and CT scans. Thus, our findings might not generalize to other many medical diagnostic problems. Despite this, mammograms and CT scans are two of the most widely used medical imaging tools and improving the accuracy of how models classify these types of images will be of great importance.

In conclusion, we used two datasets of mammographic imaging and CT scans to test the prediction accuracy of models in the classification of lesions as malignant or benign. We added one of four DA techniques to these models and compared its performance to that of transfer learning. We found that DA techniques significantly improve model performance and generalizability, specifically unsupervised DAs like MMD, CORAL, and iDANN. We also found that these techniques also helped increase prediction accuracy among combined datasets and that while each DA technique had its unique properties, there was significant agreement between the best-performing techniques in the images they were able to correctly classify. These findings suggest that there's potential to further study how multiple domain adaptations can work together and that these can potentially help us create more robust, generalizable models that will continue to facilitate and improve the medical care we can provide to our patients.

Dissemination:

The results of this study have been submitted to the journal, Frontiers of Oncology. This paper is currently under review and there is therefore not yet a decision on the publication of these results.

References

1. Gu, Y., et al., *Progressive Transfer Learning and Adversarial Domain Adaptation for Cross-Domain Skin Disease Classification.* IEEE J Biomed Health Inform, 2020. **24**(5): p. 1379-1393.
2. Nagy, M., N. Radakovich, and A. Nazha, *Machine Learning in Oncology: What Should Clinicians Know?* JCO Clin Cancer Inform, 2020. **4**: p. 799-810.
3. Sarvamangala, D.R. and R.V. Kulkarni, *Convolutional neural networks in medical image understanding: a survey.* Evol Intell, 2022. **15**(1): p. 1-22.
4. Lai, M., *Deep learning for medical image segmentation.* arXiv preprint arXiv:1505.02000, 2015.
5. Rozantsev, A., M. Salzmann, and P. Fua, *Beyond Sharing Weights for Deep Domain Adaptation.* IEEE Trans Pattern Anal Mach Intell, 2019. **41**(4): p. 801-814.
6. Razzak, M., S. Naz, and A. Zaib, *Deep Learning for Medical Image Processing: Overview, Challenges and the Future*. 2018. p. 323-350.
7. Bakator, M. and D. Radosav *Deep Learning and Medical Diagnosis: A Review of Literature*. Multimodal Technologies and Interaction, 2018. **2**, DOI: 10.3390/mti2030047.
8. Barragan-Montero, A., et al., *Artificial intelligence and machine learning for medical imaging: A technology review.* Phys Med, 2021. **83**: p. 242-256.
9. Litjens, G., et al., *A survey on deep learning in medical image analysis.* Med Image Anal, 2017. **42**: p. 60-88.

10. Kortesniemi, M., et al., *The European Federation of Organisations for Medical Physics (EFOMP) White Paper: Big data and deep learning in medical imaging and in relation to medical physics profession.* Phys Med, 2018. **56**: p. 90-93.

11. Liang, X., et al., *Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy.* Phys Med Biol, 2019. **64**(12): p. 125002.

12. Dou, Q., et al., *PnP-AdaNet: Plug-and-Play Adversarial Domain Adaptation Network at Unpaired Cross-Modality Cardiac Segmentation.* IEEE Access, 2019. **7**: p. 99065-99076.

13. Hesamian, M.H., et al., *Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges.* J Digit Imaging, 2019. **32**(4): p. 582-596.

14. Moeskops, P., et al., *Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI.* Neuroimage Clin, 2018. **17**: p. 251-262.

15. Zhang, W., et al., *Deep convolutional neural networks for multi-modality isointense infant brain image segmentation.* Neuroimage, 2015. **108**: p. 214-24.

16. Kamnitsas, K., et al., *Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.* Med Image Anal, 2017. **36**: p. 61-78.

17. Prados, F., et al., *Spinal cord grey matter segmentation challenge.* Neuroimage, 2017. **152**: p. 312-329.

18. Lustberg, T., et al., *Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer.* Radiother Oncol, 2018. **126**(2): p. 312-317.

19. Ibragimov, B. and L. Xing, *Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks.* Med Phys, 2017. **44**(2): p. 547-557.

20. Nikolov, S., et al., *Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study.* J Med Internet Res, 2021. **23**(7): p. e26151.

21. Falconí, L.G., et al., *Transfer Learning and Fine Tuning in Breast Mammogram Abnormalities Classification on CBIS-DDSM Database.* Advances in Science, Technology and Engineering Systems Journal, 2020. **5**: p. 154-165.

22. Lotter, W., et al., *Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach.* Nat Med, 2021. **27**(2): p. 244-249.

23. Larrazabal, A.J., et al., *Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis.* Proc Natl Acad Sci U S A, 2020. **117**(23): p. 12592-12594.

24. AlBadawy, E.A., A. Saha, and M.A. Mazurowski, *Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing.* Med Phys, 2018. **45**(3): p. 1150-1158.

25. French, G., M. Mackiewicz, and M. Fisher, *Self-ensembling for visual domain adaptation.* arXiv preprint arXiv:1706.05208, 2017.

26. Baffour, A.A., et al., *Generic network for domain adaptation based on self-supervised learning and deep clustering.* Neurocomputing, 2022. **476**: p. 126-136.

27. Torralba, A. and A.A. Efros, *Unbiased look at dataset bias.* CVPR 2011, 2011: p. 1521-1528.

28. Gilson, A., et al., *Abstract PO-074: The impact of phenotypic bias in the generalizability of deep learning models in non-small cell lung cancer.* Clinical Cancer Research, 2021. **27**(5_Supplement): p. PO-074-PO-074.

29. Perone, C.S., et al., *Unsupervised domain adaptation for medical imaging segmentation with self-ensembling.* Neuroimage, 2019. **194**: p. 1-11.

30. Shen, J., et al., *Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review.* JMIR Med Inform, 2019. **7**(3): p. e10010.

31. McLaughlin, N., J.M.D. Rincon, and P. Miller. *Data-augmentation for reducing dataset bias in person re-identification*. in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2015.

32. Gupta, A., et al., *Robot learning in homes: Improving generalization and reducing dataset bias.* Advances in neural information processing systems, 2018. **31**.

33. Kohli, M.D., R.M. Summers, and J.R. Geis, *Medical Image Data and Datasets in the Era of Machine Learning-Whitepaper from the 2016 C-MIMI Meeting Dataset Session.* J Digit Imaging, 2017. **30**(4): p. 392-399.

34. Harvey, H. and B. Glocker, *A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology*, in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, E.R. Ranschaert, S. Morozov, and P.R. Algra, Editors. 2019, Springer International Publishing: Cham. p. 61-72.

35. Calli, E., et al., *Deep learning for chest X-ray analysis: A survey.* Med Image Anal, 2021. **72**: p. 102125.

36. Xiao, D., et al., *Unsupervised machine fault diagnosis for noisy domain adaptation using marginal denoising autoencoder based on acoustic signals.* Measurement, 2021. **176**: p. 109186.

37. Bai, Z. and X.L. Zhang, *Speaker recognition based on deep learning: An overview.* Neural Netw, 2021. **140**: p. 65-99.

38. Saxena, S. and J. Verbeek. *Heterogeneous Face Recognition with CNNs*. in *Computer Vision – ECCV 2016 Workshops*. 2016. Cham: Springer International Publishing.

39. Fernando, B., T. Tommasi, and T. Tuytelaars, *Location recognition over large time lags.* Computer Vision and Image Understanding, 2015. **139**: p. 21-28.

40. Loghmani, M.R., et al., *Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition.* IEEE Robotics and Automation Letters, 2020. **5**(4): p. 6631-6638.

41. Chen, C., et al., *Unsupervised Bidirectional Cross-Modality Adaptation via Deeply Synergistic Image and Feature Alignment for Medical Image Segmentation.* IEEE Trans Med Imaging, 2020. **39**(7): p. 2494-2505.

42. Li, Y., et al., *Adaptive Batch Normalization for practical domain adaptation.* Pattern Recognition, 2018. **80**: p. 109-117.

43. Joel, M.Z., et al., *Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology.* JCO Clin Cancer Inform, 2022. **6**: p. e2100170.

44. Wang, M. and W. Deng, *Deep visual domain adaptation: A survey.* Neurocomputing, 2018. **312**: p. 135-153.

45. Li, Y., et al., *Revisiting batch normalization for practical domain adaptation.* arXiv preprint arXiv:1603.04779, 2016.

46. Gallego, A.J., J. Calvo-Zaragoza, and R.B. Fisher, *Incremental Unsupervised Domain-Adversarial Training of Neural Networks.* IEEE Trans Neural Netw Learn Syst, 2021. **32**(11): p. 4864-4878.

47. Viera, A.J. and J.M. Garrett, *Understanding interobserver agreement: the kappa statistic.* Fam Med, 2005. **37**(5): p. 360-3.

48. Long, M., et al. *Learning transferable features with deep adaptation networks*. in *International conference on machine learning*. 2015. PMLR.

49. Tzeng, E., et al. *Simultaneous deep transfer across domains and tasks*. in *Proceedings of the IEEE international conference on computer vision*. 2015.

50. Tzeng, E., et al. *Adversarial discriminative domain adaptation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

51. Carneiro, G., J. Nascimento, and A.P. Bradley. *Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models*. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015. Cham: Springer International Publishing.

52. Li, H., et al., *Benign and malignant classification of mammogram images based on deep learning.* Biomedical Signal Processing and Control, 2019. **51**: p. 347-354.

53. Wang, J., et al., *Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning.* Sci Rep, 2016. **6**: p. 27327.

54. Tommasi, T., et al., *A deeper look at dataset bias*, in *Domain adaptation in computer vision applications*. 2017, Springer. p. 37-55.

55. Zhang, J., W. Li, and P. Ogunbona, *Transfer Learning for Cross-Dataset Recognition: A Survey.* arXiv: Computer Vision and Pattern Recognition, 2017.

56. Long, M., et al. *Deep transfer learning with joint adaptation networks*. in *International conference on machine learning*. 2017. PMLR.

57. Sun, B. and K. Saenko. *Deep coral: Correlation alignment for deep domain adaptation*. in *European conference on computer vision*. 2016. Springer.

58. Shaha, M. and M. Pawar, *Transfer Learning for Image Classification.* 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018: p. 656-660.

59. Wu, E., et al., *Conditional infilling GANs for data augmentation in mammogram classification*, in *Image analysis for moving organ, breast, and thoracic images*. 2018, Springer. p. 98-106.

60. Zhang, Q., W. Wang, and S.-C. Zhu. *Examining CNN representations with respect to Dataset Bias*. in *AAAI Conference on Artificial Intelligence*. 2017.

Data Supplement

CBIS-DDSM Dataset

A Density

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.669	0.661 - 0.675	-	-	0.702	0.693 - 0.708	-	-
MMD	0.699	0.687 – 0.700	+ 4.5%	< 0.001	0.714	0.707 – 0.723	+ 1.7%	0.004
CORAL	0.696	0.686 - 0.698	+ 4.0%	< 0.001	0.739	0.732 – 0.748	+ 5.3%	< 0.001
AdaBN	0.696	0.689 - 0.701	+ 4.0%	< 0.001	0.714	0.710 – 0.725	+ 1.7%	< 0.001
iDANN	0.714	0.706 - 0.718	+ 6.7%	< 0.001	0.727	0.712 – 0.726	+ 3.6%	< 0.001

B Image View

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.726	0.719 - 0733	-	-	0.748	0.738 – 0.751	-	-
MMD	0.752	0.749 – 0.761	+ 3.6%	< 0.001	0.745	0.734 – 0.746	- 0.4%	0.860
CORAL	0.755	0.742 – 0.757	+ 4.0%	< 0.001	0.742	0.732 – 0.746	- 0.8%	0.883
AdaBN	0.726	0.716 – 0.731	0 %	0.708	0.694	0.686 – 0.700	- 7.2%	1.000
iDANN	0.737	0.728 – 0.743	+ 1.5%	0.030	0.755	0.744 – 0.756	+ 0.9%	0.127

C Mass Margins

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.577	0.569 – 0.592	-	-	0.654	0.636 – 0.656	-	-
MMD	0.837	0.835 – 0.853	+ 45.0%	< 0.001	0.809	0.802 – 0.816	+ 23.7%	< 0.001
CORAL	0.854	0.851 – 0.867	+ 48.0%	< 0.001	0.809	0.802 – 0.818	+ 23.7%	< 0.001
AdaBN	0.520	0.512 – 0.536	- 9.9%	1.000	0.599	0.584 – 0.606	- 8.4%	1.000
iDANN	0.862	0.857 – 0.869	+ 49.4%	< 0.001	0.815	0.809 – 0.824	+ 24.6%	< 0.001

D Mass Shape

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.646	0.646 – 0.664	-	-	0.721	0.719 – 0.733	-	-
MMD	0.785	0.777 – 0.797	+ 21.5%	< 0.001	0.769	0.762 – 0.777	+ 6.7%	< 0.001
CORAL	0.778	0.770 – 0.786	+ 20.4%	< 0.001	0.769	0.761 – 0.777	+ 6.7%	< 0.001
AdaBN	0.481	0.470 – 0.490	- 25.5%	1.000	0.649	0.641 – 0.658	- 10.0%	1.000
iDANN	0.753	0.744 – 0.762	+ 16.6%	< 0.001	0.764	0.754 – 0.766	+ 6.0%	< 0.001

Table 4 (A-D). This table shows the accuracies of transfer learning and four domain adaptation methods on the training set of the target domain and the combination training set. Included are also the confidence interval and p-values. These results are on the mammography dataset, CBIS-DDSM.

LIDC-IDRI Dataset

A Spiculation

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.682	0.679 - 0.697		-	0.743	0.738 - 0.754		-
MMD	0.777	0.764 – 0.780	+ 13.9%	< 0.001	0.811	0.805 – 0.819	+ 9.2%	< 0.001
CORAL	0.790	0.775 - 0.792	+ 15.8%	< 0.001	0.816	0.808 – 0.820	+ 9.8%	< 0.001
AdaBN	0.662	0.649 - 0.667	- 2.9%	1.000	0.714	0.712 – 0.729	- 3.9%	1.000
iDANN	0.758	0.750 - 0.766	+ 11.1%	< 0.001	0.811	0.797 – 0.812	+ 9.2%	< 0.001

B Subtlety

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.683	0.680 – 0.687		-	0.813	0.811 – 0.817		-
MMD	0.719	0.715 – 0.721	+ 5.3%	< 0.001	0.834	0.830 – 0.837	+ 2.5%	< 0.001
CORAL	0.720	0.718 – 0.724	+ 5.4%	< 0.001	0.832	0.832 – 0.838	+ 2.4%	< 0.001
AdaBN	0.703	0.700 – 0.706	+ 2.9%	< 0.001	0.807	0.806 – 0.813	- 0.7%	0.974
iDANN	0.713	0.711 – 0.717	+ 4.4%	< 0.001	0.796	0.792 – 0.798	- 2.1%	1.000

C Margin

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.701	0.698 – 0.705		-	0.779	0.777 – 0.783		-
MMD	0.760	0.758 – 0.764	+ 8.4%	< 0.001	0.813	0.808 – 0.815	+ 4.3%	< 0.001
CORAL	0.772	0.772 – 0.778	+ 10.1%	< 0.001	0.825	0.821 – 0.827	+ 5.9%	< 0.001
AdaBN	0.752	0.747 – 0.753	+ 7.3%	< 0.001	0.813	0.810 – 0.816	+ 4.4%	< 0.001
iDANN	0.770	0.766 – 0.771	+ 9.8%	< 0.001	0.818	0.815 – 0.822	+ 5.0%	< 0.001

D Contrast

	Accuracy on Target Domain	CI	Percent change	P-Value	Accuracy on Mixed Domain	CI	Percent change	P-Value
Transfer Learning	0.792	0.788 – 0.796		-	0.796	0.793 – 0.799		-
MMD	0.808	0.804 – 0.812	+ 2.0%	< 0.001	0.807	0.801 – 0.808	+ 1.3%	< 0.001
CORAL	0.811	0.806 – 0.814	+ 2.4%	< 0.001	0.811	0.806 – 0.813	+ 1.9%	< 0.001
AdaBN	0.805	0.801 – 0.809	+ 1.6%	< 0.001	0.814	0.812 – 0.819	+ 2.3%	< 0.001
iDANN	0.808	0.803 – 0.810	+ 2.0%	< 0.001	0.809	0.805 – 0.812	+ 1.6%	< 0.001

Table 5 (A-D). This table shows the accuracies of transfer learning and four domain adaptation methods on the training set of the target domain and the combination training set. Included are also the confidence interval and p-values. These results are on the CT scan dataset, LIDC-IDRI.

CBIS-DDSM Dataset

A

Density

	CORAL	AdaBN	iDANN
MMD	0.618	0.676	0.817
CORAL		0.612	0.649
AdaBN			0.737

B

Image View

	CORAL	AdaBN	iDANN
MMD	0.593	0.593	0.693
CORAL		1	0.545
AdaBN			0.545

C

Mass Margins

	CORAL	AdaBN	iDANN
MMD	0.876	0.080	0.714
CORAL		0.079	0.700
AdaBN			0.062

D

Mass Shape

	CORAL	AdaBN	iDANN
MMD	0.833	-0.016	0.449
CORAL		0.046	0.577
AdaBN			0.168

Table 6 (A-D). The accuracies of transfer learning and four domain adaptation methods on the training set of the target domain and the combination training set, as well as the confidence interval and p-values.

LIDC-IDRI Dataset

A

Spiculation

	Accuracy on Target Domain	CI	P-Value	Accuracy on Mixed Domain	CI	P-Value
Transfer Learning	0.704	0.693 - 0.704	-	0.763	0.759 - 0.769	-
MMD	0.767	0.765 – 0.777	< 0.001	0.788	0.784 – 0.793	< 0.001
CORAL	0.784	0.779 - 0.790	< 0.001	0.825	0.819 – 0.828	< 0.001
AdaBN	0.666	0.662 - 0.674	1.000	0.813	0.809 – 0.818	< 0.001
iDANN	0.767	0.760 - 0.773	< 0.001	0.821	0.813 – 0.822	< 0.001

B

Subtlety

	Accuracy on Target Domain	CI	P-Value	Accuracy on Mixed Domain	CI	P-Value
Transfer Learning	0.686	0.683 - 0.687	-	0.811	0.808 - 0.812	-
MMD	0.723	0.722 – 0.726	< 0.001	0.838	0.836 – 0.840	< 0.001
CORAL	0.724	0.722 - 0.726	< 0.001	0.838	0.837 – 0.841	< 0.001
AdaBN	0.695	0.694 - 0.698	< 0.001	0.844	0.842 – 0.846	< 0.001
iDANN	0.709	0.708 - 0.712	< 0.001	0.834	0.831 – 0.836	< 0.001

C

Margin

	Accuracy on Target Domain	CI	P-Value	Accuracy on Mixed Domain	CI	P-Value
Transfer Learning	0.713	0.710 - 0.715	-	0.787	0.785 - 0.789	-
MMD	0.768	0.768 – 0.771	< 0.001	0.822	0.821 – 0.824	< 0.001
CORAL	0.782	0.780 - 0.784	< 0.001	0.827	0.826 – 0.829	< 0.001
AdaBN	0.754	0.753 - 0.756	< 0.001	0.815	0.814 – 0.818	< 0.001
iDANN	0.755	0.752 - 0.756	< 0.001	0.823	0.820 – 0.824	< 0.001

D

Contrast

	Accuracy on Target Domain	CI	P-Value	Accuracy on Mixed Domain	CI	P-Value
Transfer Learning	0.781	0.779 - 0.785	-	0.800	0.794 - 0.800	-
MMD	0.807	0.803 – 0.809	< 0.001	0.825	0.822 – 0.826	< 0.001
CORAL	0.817	0.812 - 0.817	< 0.001	0.833	0.832 – 0.836	< 0.001
AdaBN	0.804	0.800 - 0.806	< 0.001	0.829	0.827 – 0.831	< 0.001
iDANN	0.811	0.807 - 0.812	< 0.001	0.832	0.832 – 0.836	< 0.001

Table 7 (A-D). The accuracies of transfer learning and four domain adaptation methods on the training set of the target domain and the combination training set, as well as the confidence interval and p-values.

CBIS-DDSM Dataset

	Density	Image View	Mass Margins	Mass Shape
MMD	15 (4.6%)	11 (4.0%)	7 (5.7%)	15 (9.5%)
CORAL	36 (10.9%)	19 (6.9%)	6 (4.9%)	13 (8.2%)
AdaBN	2 (7.0%)	19 (6.9%)	15 (12.2%)	31 (19.6%)
iDANN	18 (5.5%)	17 (6.2%)	7 (5.7%)	9 (5.7%)

LIDC-IDRI Dataset

	Spiculation	Subtlety	Margin	Contrast
MMD	0 (0%)	34 (2.5%)	34 (2.8%)	33 (4.7%)
CORAL	3 (1.9%)	40 (3.0%)	46 (3.8%)	37 (5.3%)
AdaBN	24 (15.3%)	35 (2.6%)	30 (2.5%)	36 (5.1%)
iDANN	9 (5.7%)	76 (5.7%)	93 (7.7%)	45 (6.4%)

Table 8. Images correctly identified by transfer learning but incorrect by domain adaptation

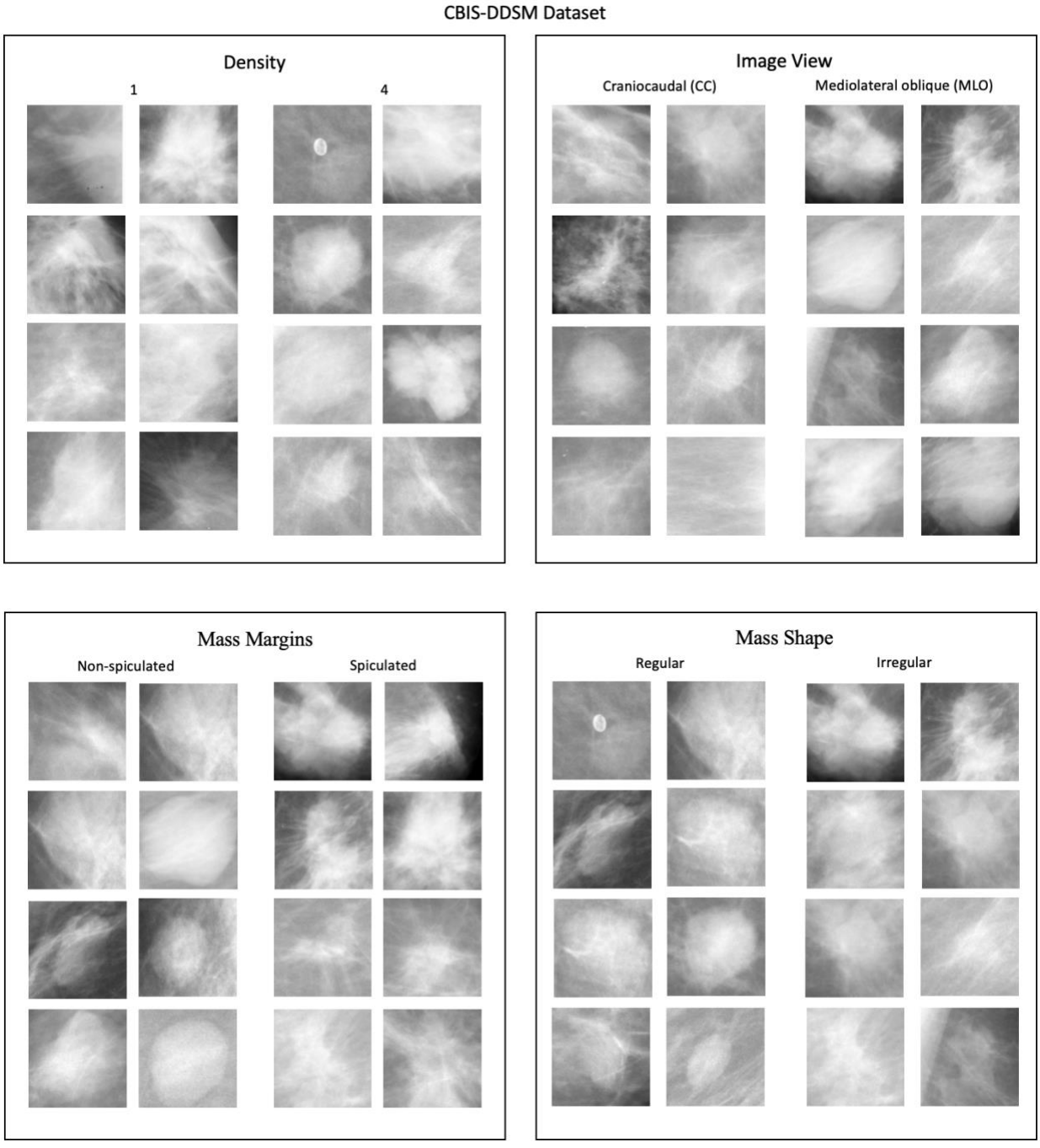

Figure 6. This figure shows mammogram images from different domains.

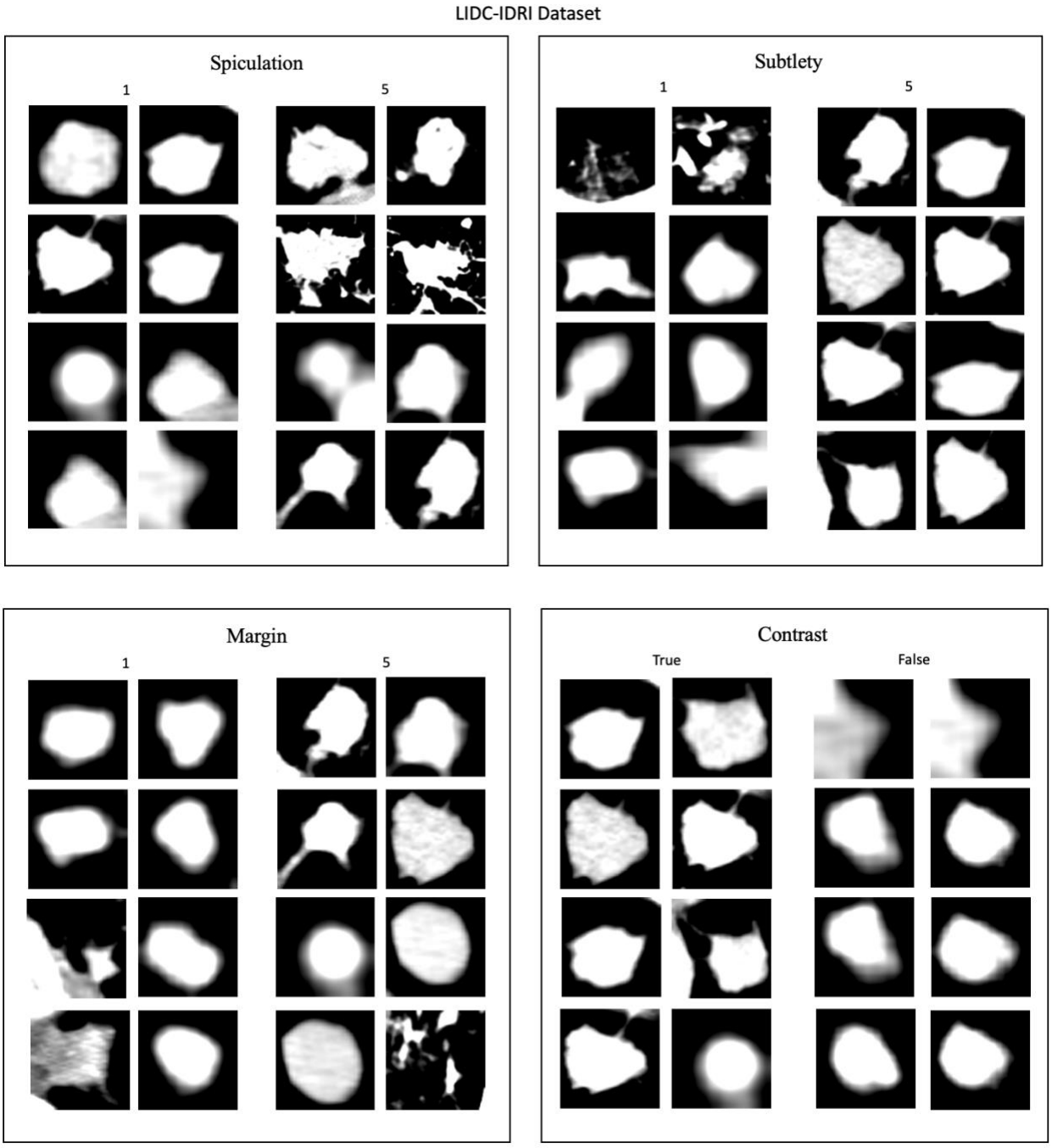

Figure 7. This figure shows CT scan images from different domains.

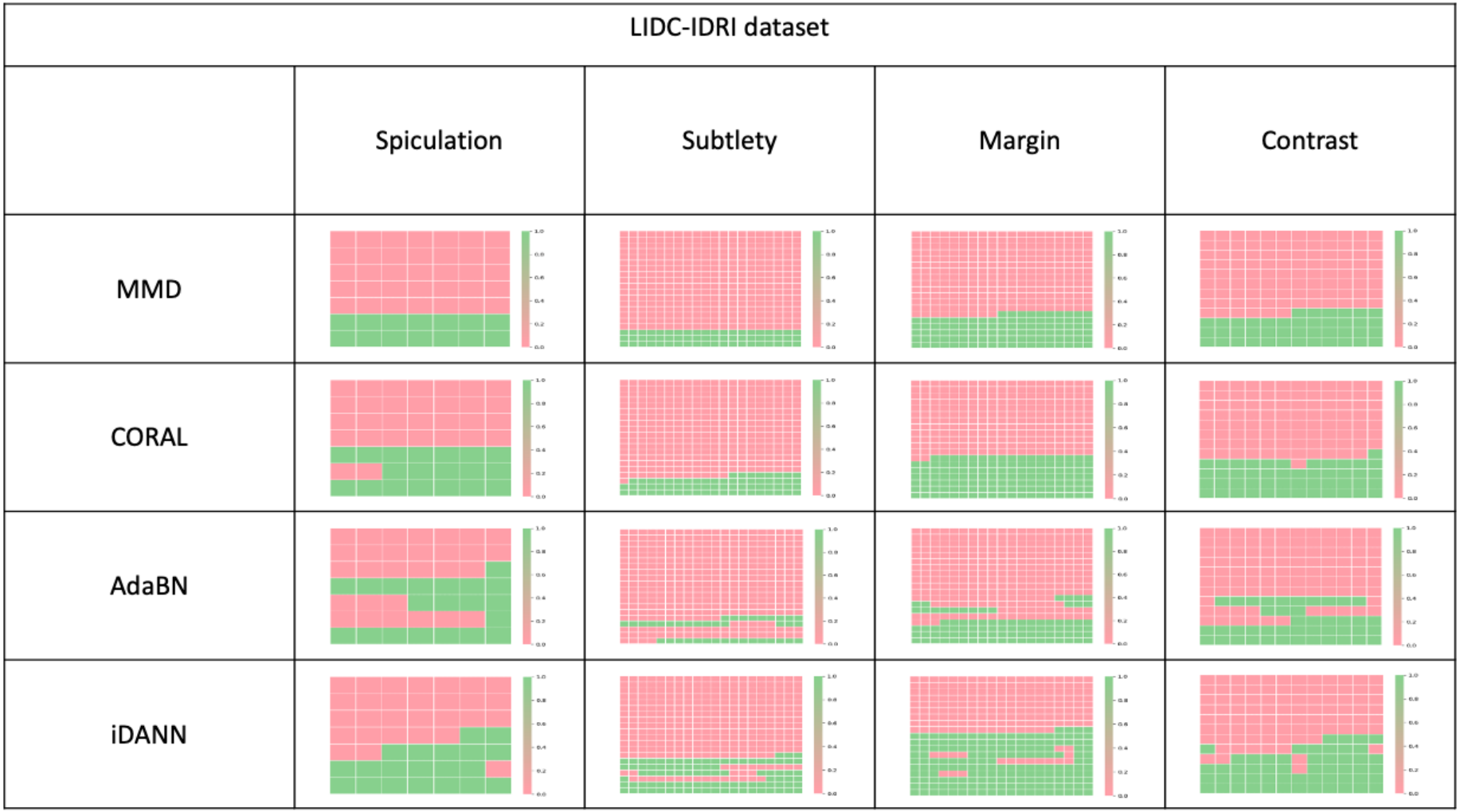

Figure 8. Includes all mammogram images that are predicted wrong by transfer learning. Each plot uses one of the domain adaptation methods. Boxes in green represent mammographic images that a particular domain adaptation method predicts correctly. Boxes in red represent mammographic images a particular domain adaptation method predicts wrong.

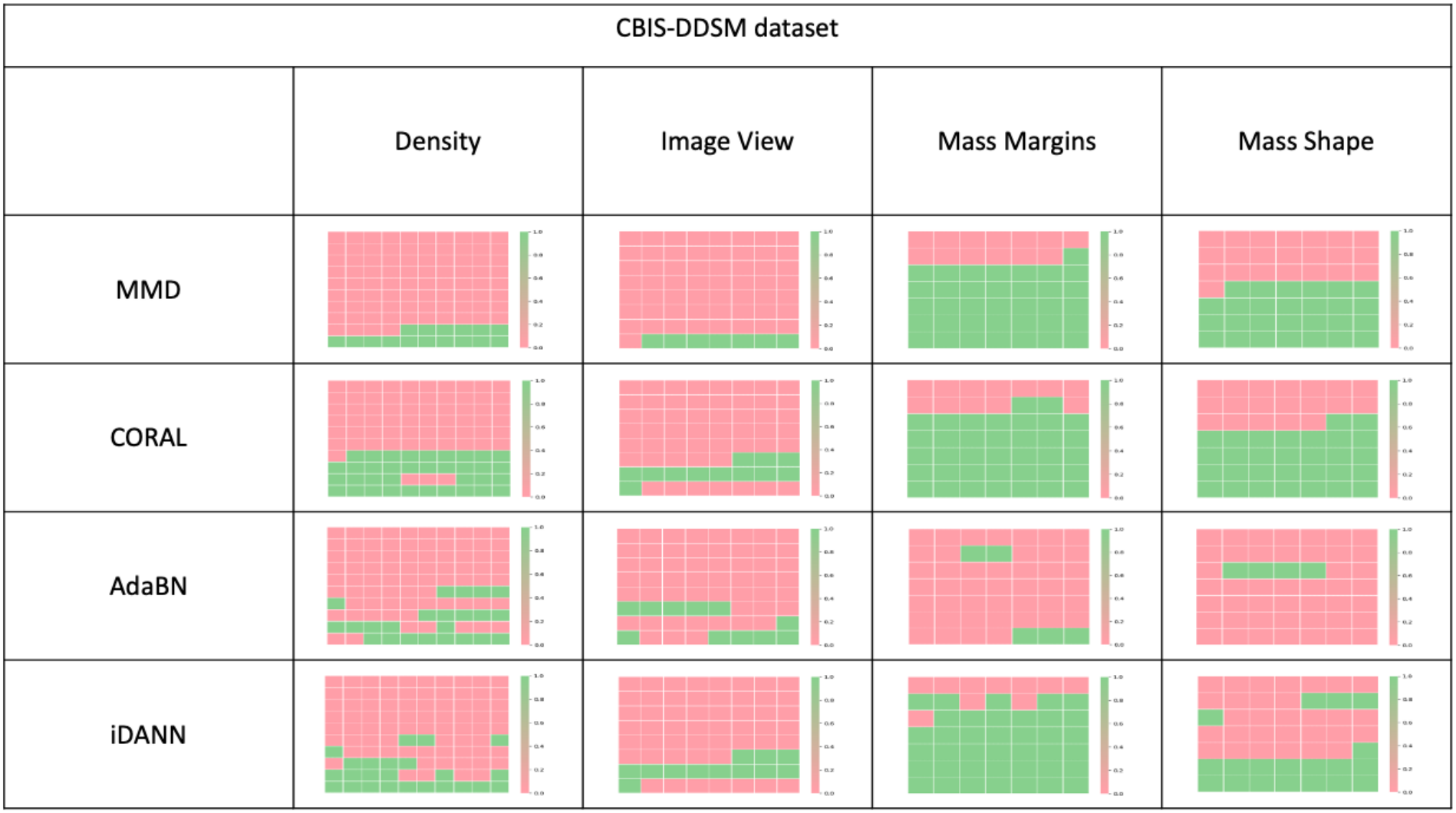

Figure 9. Includes all CT scan images that are predicted wrong by transfer learning. Each plot uses one of the domain adaptation methods. Boxes in green represent CT images that a particular domain adaptation method predicts correctly. Boxes in red represent CT images that a particular domain adaptation method predicts wrong.

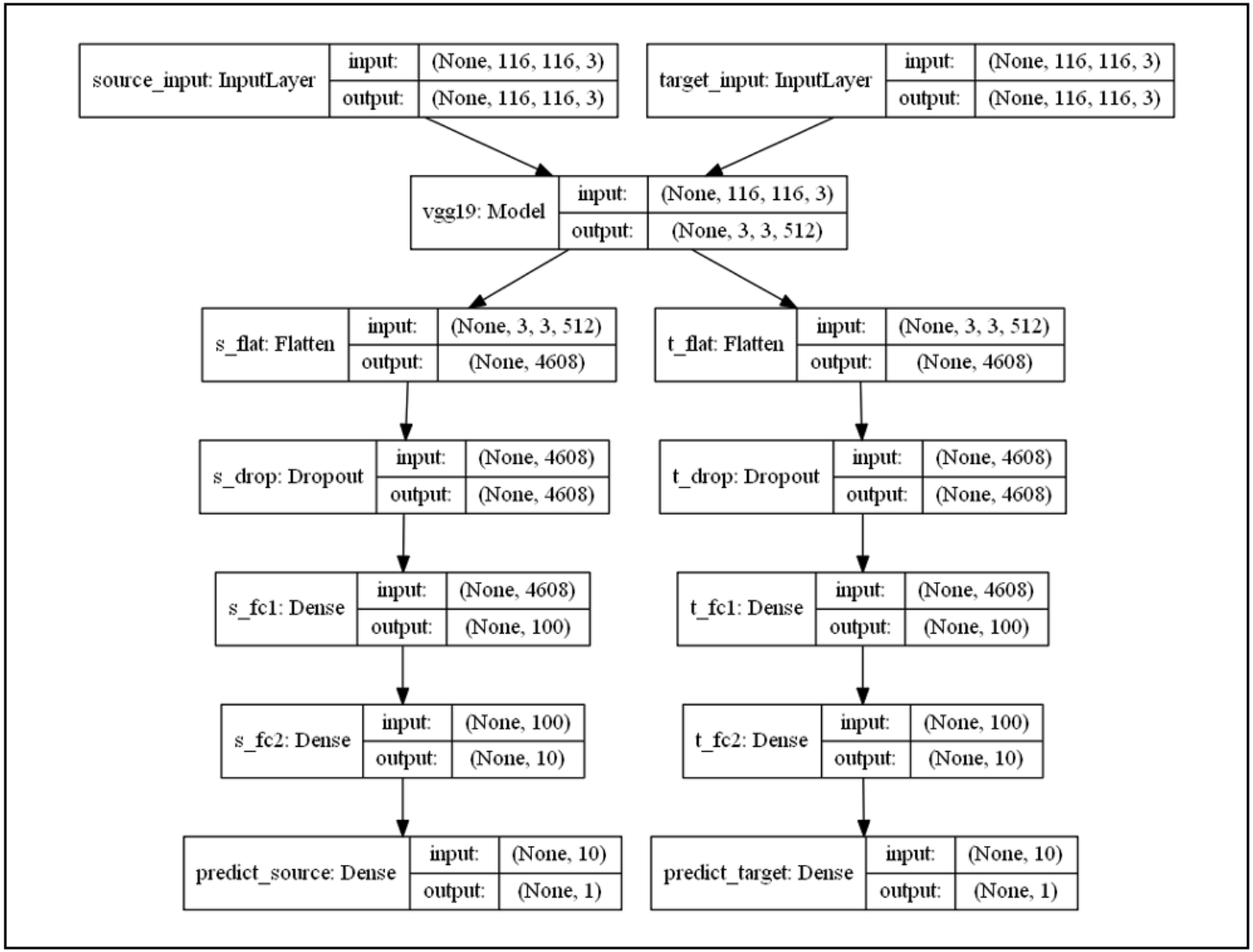

Figure 10. The deep learning network architectures show every layer in the networks.

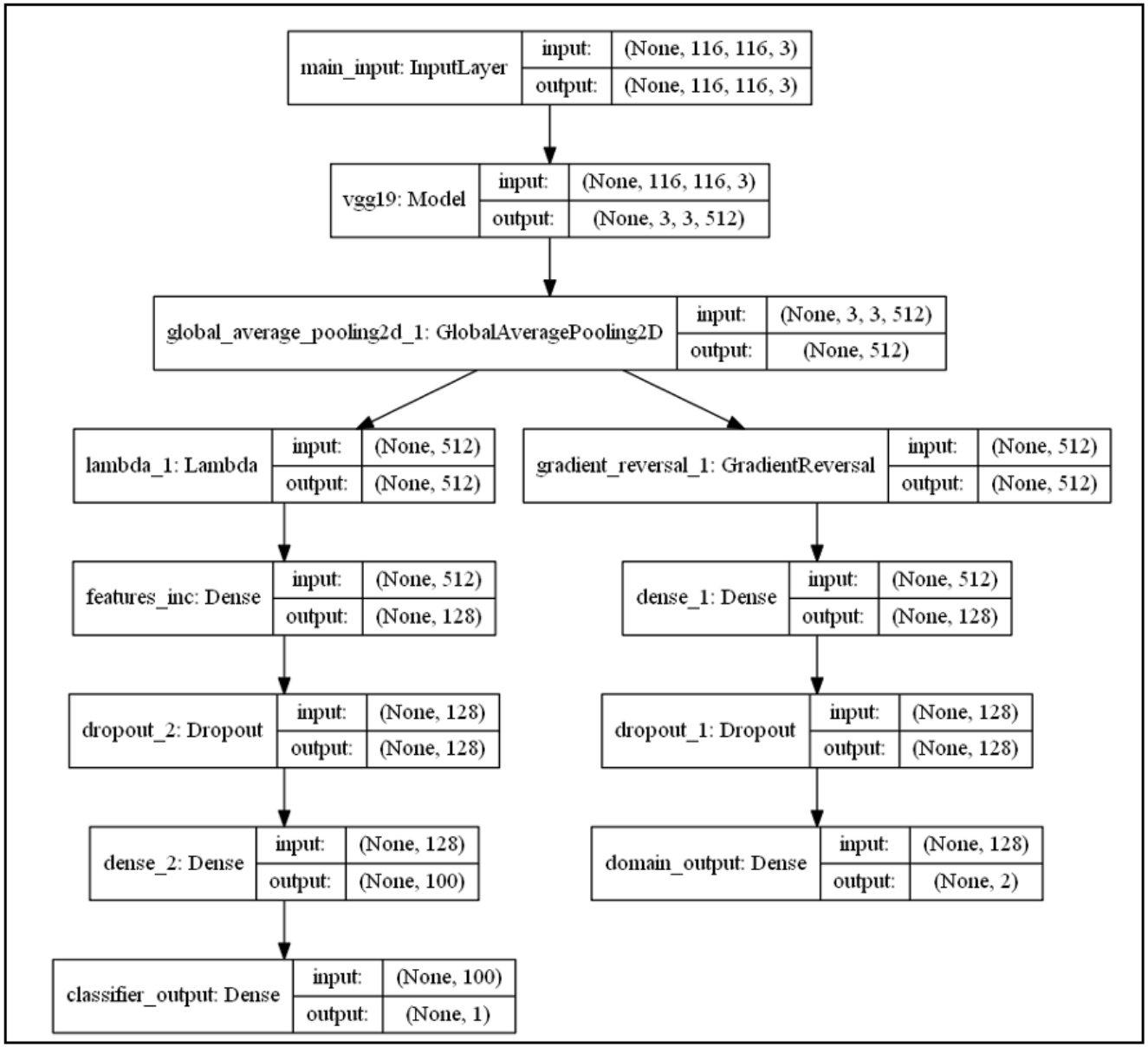

Figure 11. The deep learning network architectures show every layer in the networks.