IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 April 2023, accepted 16 April 2023, date of publication 20 April 2023, date of current version 25 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3268704

TOPICAL REVIEW

Failure to Achieve Domain Invariance With Domain Generalization Algorithms: An Analysis in Medical Imaging

STEVEN KOREVAAR, RUWAN TENNAKOON, AND ALIREZA BAB-HADIASHAR, (Senior Member, IEEE)
Royal Melbourne Institute of Technology, Melbourne, VIC 3000, Australia

Corresponding author: Steven Korevaar (steven.korevaar@student.rmit.edu.au)

This research was supported by grants from NVIDIA and utilized an NVIDIA Quadro A6000 for running experiments.

ABSTRACT One prominent issue in the application of deep learning is the failure to generalize to data that lies on a different distribution to the training data. While many methods have been proposed to address this, prior work has shown that when operating under the same conditions most algorithms perform almost equally. As such, more work needs to be done to validate past and future methods before they are put into important scenarios like medical imaging. Our work analyses eight domain generalization algorithms across four important medical imaging classification datasets along with three standard natural image classification problems to discover the differences in how these methods operate in these different contexts. We assess these algorithms in terms of generalization capability, domain invariance, and representational sensitivity. Through this, we show that despite the differences between domain and content variations between natural and medical imaging there is little deviation in the operation of each method between natural images and medical images. Additionally, we show that all tested algorithms retain significant amounts of domain-specific information in their feature representations despite explicit training to remove it. Thus, revealing the failure point of all these methods is a lack of class-discriminative features extracted from out-of-distribution data. While these results show that methods that work well on natural imaging work similarly in medical imaging, no method outperforms baseline methods, highlighting the continuing gap of achieving adequate domain generalization. Similarly, the results also question the efficacy of optimizing for domain invariant representations as a method for generalizing to unseen domains.

INDEX TERMS Deep learning, domain generalization, medical image classification, natural image classification, model analysis, domain invariant representations, representational smoothness.

I. INTRODUCTION

In the medical field, deep learning is frequently used to aid radiologists in making diagnoses and speeding up treatment times. As such, ensuring that the diagnostic methods are based on fundamentally sound principles is a necessity. One issue in deep learning has garnered significant attention recently, domain shift, where the data used to train the deep learning models differs in some way from the data found during use. This issue can take many forms, but one of the most important is found in medical imaging, where a model may be trained on data from one hospital, which uses a certain method for capturing their scans, and when that model is applied to a new hospital with different data capture methodologies, the same high performing model would fail. The most obvious solution to this problem is to gather data from as many sources as possible to widen the variety of training data; however, in practice it is infeasible to collect data for every single variation possible. Instead, methods for training models that can generalize across variations without

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

access to them at training time have been developed. With the development of new methods though comes a necessity to properly validate them in ways that can provide a reliable benchmark for future users who want to put them into practice in new environments, and to ensure fair comparison between methods. As a result of how deep learning operates, where model architecture, parameter choice, and many other variables are learnt and tuned, it can be easy to accidentally overfit methods to specific datasets to achieve strong performance. To address this issue, there has been a recent push to standardise algorithm testing methods, one such work being the DomainBed framework [1]. Unfortunately, through this framework it was found that when properly tuned, even the most complex domain generalization methods very rarely outperform baseline methods. Calling into question how the methods developed differ from baseline methods and whether there is merit in their approaches.

The purpose of this work is three-fold: Firstly, to provide a framework for assessing methods in an important application of domain generalization, namely medical imaging. Secondly to analyze the differences between a variety of domain generalization methods in their operations on natural images compared to medical images, in terms of both raw performance and underlying function of the methods. Thirdly, to provide a framework for evaluating different properties of models based on theoretical approaches in the domain generalization field. Through the creation of this framework, we encountered results that question the narrative in how the domain generalization works in deep learning. This leads us to the main contributions of this work:

- We provide a set of benchmark performances on state-of-the-art algorithms on three commonly used natural imaging datasets as well as four important modalities in medical imaging datasets. We show that despite differences between the two contexts, algorithm performance on medical imaging datasets correlates strongly with performance on natural image classification datasets. Additionally, there is no significant discrepancy between the method's operations between the two contexts, despite their differences.
- We propose an initial framework (called Analysing Representational Domain Invariance, or ARDI in short) for analyzing domain generalization methods through the theoretical narratives of domain invariant and smooth representations. We also highlight the importance of further work in developing more methods for analysing generalization methods to encompass further generalization mechanisms, as well as answering the many questions regarding the importance of invariant representations that have been discovered within this work.
- We have used this framework to show that achieving domain invariant representations is a large scale problem, that persists across many different mechanisms. Both the failure to achieve domain invariance and the lack of correlation between model performance on OOD data and both domain invariant representations and representational smoothness highlights the limitations of current domain generalization methods and possibly the theories behind generalization in deep learning models.

A. BACKGROUND

1) WHAT IS DOMAIN GENERALIZATION?

As stated previously, domain generalization is a necessity when data undergoes a distribution shift between the training and testing environments. One such area where this distributional shift occurs frequently is in medical imaging.

AI is used extensively in medical imaging [2], [3], and is likely to continue to grow as more and more medical fields are looking to use AI to achieve more accurate diagnoses and to speed up treatment time.

While AI systems are often trained for diagnosis purposes; due to reliability and accountability concerns AI models are normally only used to aid a healthcare practitioner in making a diagnosis. Typically, these models' predictions will used in conjunction with other tests by a healthcare practitioner who will give the final diagnosis, hence the term ''computer aided diagnosis''.

One barrier in the way of implementing AI in medical imaging is domain shift. As such the importance of having a solid benchmark for various domain generalization methods on different modalities within medical imaging is growing.

To start with, it is important to discuss the ways in which medical data can vary. The most common shifts being: The modality of scan being captured (CT, MRI, X-ray, etc), source of the data (different hospitals/centres), scanning parameters (scanner manufacturer, resolution, contrast agents, reconstruction algorithms), difference scanning procedures (scans before vs after treatments, timing after contrast injection, etc). Due to the previously mentioned difficulties in gathering data for training, getting enough data that covers all these variations is extraordinarily difficult. As such, most training and testing are performed on only a single set of these varying parameters, thus the robustness problem is often not discovered until the methods are deployed into the real-world where these distributional shifts are encountered. This means that having a comprehensive set of benchmarks and a testing framework to analyse methods on varying tasks that can apply to many practical situations is becoming increasingly important.

The goal of domain generalization algorithms can be summarized succinctly: train models that can perform the same task across a wide variety of circumstances. Mathematically, this is defined as having N training/source datasets made up of inputs, x, and labels y, $S = \{S_n = \{(x^n, y^n)\}\}_{n=1}^{N}$, specifically each dataset has a different marginal distribution of inputs/labels, P_{XY}^n on $X \times Y$. The task is to train a predictive model, $f : X \rightarrow Y$, on the training data that achieves low prediction error on unseen testing data, $T = \{(x^t, y^t)\}$, with similar but different marginal distribution to the training datasets. While the goal is simple, it can be a very difficult task due to the differing distributions. Currently, there are

several leading theories as to why generalization fails, the most common are: relying on domain specific information, and learning non-robust models.

For example, in a full image cancer classification task, it is possible for a model to learn to recognize common side effects of cancer, such as inflammation of surrounding tissue, and rely on that for classification instead of identifying the root cause, the cancer. Likewise, non-robust models will fail when even very minor previously unseen shifts in domain occur due to poor feature extraction sensitivity.

B. RELATED WORK

This work does not aim to provide a full review of all prior work. For a full breakdown of domain generalization work the reviews by Zhou et al, and Wang et al, cover the field thoroughly [4], [5].

1) DOMAIN GENERALIZATION MECHANISMS

While there are many methods for domain generalization, most operate by attempting to address the first point of failure of generalization: Using domain specific representations [4], [5]. This can be addressed explicitly, with something as simple as data augmentation [6] that mimics the distributional shift between domains, or it can be learnt explicitly with adversarial learning [7], [8], [9], which directly optimizes a neural network to remove all domain identifiable information from the feature representations. In a similar vein, there are methods that use disentangled representations [10], [11] (like those used for style transfer networks) that separate domain variant and domain invariant representations, which are often learnt adversarially as well. It is also possible to regularize networks with information bottlenecks [12], or metric learning and statistical feature alignment [13] meta-learning [14], [15], [16], self-supervised learning [17], [18], [19], and gradient matching between domains [20]. Lastly, there are also methods that use an ensemble of models trained on different domains to find invariant features [4], [5], [21]. The common rationale behind all these methods is the goal to find "**domain invariant features**". Logic would assume that features that do not vary between the training domains are more likely to also be invariant when shifting to unseen domains, and thus if a model can learn to use only these features, the model should have stronger generalization. It could also be assumed that methods such as DANN [7] and MMD [8] which explicitly trains a model to remove all domain knowledge from their feature representations would lead to the most domain invariant representations. Whether this is the case is something our work explores.

In medical imaging, domain specific features are also likely to be confounding elements that are only present in fewer-than-all training domains. For example, in a cancer detection task, if one domain contains a high proportion of inflamed tissue surrounding the cancer while other domains do not, then domain invariant models are unlikely to learn to rely on inflammation as an indicator of cancer. It is currently unexplored whether this has a significant impact on performance or not.

Additionally, there is representational/distributional robustness. Shui et al. [22] proposes that models that have learnt only to be invariant to source domain shifts are not guaranteed to be invariant to possible future domain shifts. They propose that a model that has smooth representations (has more gradual shifts in its representations when the input changes) leads to more accurate interpolation between samples and domains, so new samples, which may fall between or close to the training data distribution, will be handled better. For a practical example, as a majority of domain shifts in the medical imaging space are small shifts in overall image style, noise, etc., a model that is robust to small shifts in the input space (i.e., different possible noise types) are less affected by future possible domain shifts and thus more likely to be better at generalizing to unseen domains. In this manner, a method such as inter-domain Mixup training [23], [24] which trains models on a linear mixture of samples should lead to smoother interpolated predictions/representations, and again, this is something our work explores.

The last form of domain generalization techniques revolve around the learning and exploitation of domain specific information. This is most commonly done by replicating parts of the network for each source domain: such as having domain-specific feature extractors [25] or novel masking layers [26]. This concept can be used further with domain-specific regularization parameters like batch-normalization, etc [27], [28].

These methods can be categorized into several different mechanisms [5]:

1) Data manipulation (e.g., data augmentation, Mixup, self-supervised learning),
2) Domain invariant representations (e.g., domain adversarial training, contrastive learning, invariant risk minimization),
3) Model regularization (e.g., meta-learning, linear-dependency regularization, information bottleneck, gradient matching),
4) Domain-specific models to leverage domain information.

This work focuses on the two single-model-centric methods: domain invariant representation learning, and model regularization.

2) DOMAIN GENERALIZATION METHODS FOR MEDICAL IMAGING

In medical imaging there are two main types of domain generalization, inter-modality and intra-modality generalization; where inter-modality generalization aims to generalize models between multiple modalities (MRI to CT and X-ray), whereas intra-modality aims to generalize between images of the same modality (CT machine to a different CT machine). In this work we are exploring intra-modality generalization. In terms of medical imaging, many methods have been explored, however most can fit into the above categories

with adjustments designed specifically for medical imaging contexts. Due to the well-understood differences between domains in medical imaging, data augmentation remains a strong contender [5]. Most methods either utilize a large number of transformations [29] or introduce targeted augmentations that mimic domain differences, with handcrafted or adversarial augmentation [9]. Other methods have been introduced that rely on regularizing deep learning models, either by regularizing the feature space itself (e.g., linear dependency regularization [30]) or by regularizing the model's parameters, done through explicit means [4] or meta-learning [31]. Meta-learning itself has numerous variations, with some opting for varying forms of task augmentation [16], and others introducing some medical-imaging-based prior [14], [15]. In a similar vein, methods have also been extended off self-supervised learning for domain generalization, and domain adversarial approaches with more medical imaging specific additions. In the case of self-supervised learning it is possible to incorporate 3D data transformations [32], [33], where 3D data is used in both MRI and CT scans. Likewise, privacy may also be a concern in medical imaging contexts, so some work has been done on modifying existing techniques, like MMD for preserving data privacy [34]. The final method uses a distinct learning paradigm of domain-oriented features, which learns a knowledge base of domain specific features which can be used to enhance the generalization of unseen samples that are close to previously seen samples [35]. All the above methods, apart from the final, can neatly be categorized into the previously noted sets, with novel additions or tweaks. While the tweaks may be important for medical imaging contexts, but the generalization mechanisms are the same.

3) VALIDATION OF DOMAIN GENERALIZATION METHODS

Once these methods have been developed, it becomes important to validate them on many problems to discover how well rounded the approaches are. Purely in terms of final model evaluation, there are three different methods (as given by [1]):

1) Training domain performance (performance on independent and identically distributed (IID) data).
2) Held-out training domain performance,
3) Real-world domain testing.

Having said that, when we are validating a methodology, another option presents itself: Train and validate methods using similar problems or datasets. When developing new methods within the deep learning field, the most common starting point is to test methods on standard datasets. The most common for domain generalization are: DigitsDG (a synthetic domain generalization dataset for handwritten digit classification), PACS (photos, art, cartoon, and sketches), and VLCS and OfficeHome datasets (natural images taken in different scenarios). Ideally, the out-of-distribution (OOD) performance of algorithms on these datasets would correlate strongly with the performance on future datasets and problems; and thus it is worth verifying that these standard datasets do perform their function in displaying where domain generalization methods work well, and where they do not. This could be a significant barrier for creating domain generalization methods in medical imaging, where data accessibility is limited, thus being able to rely on these standard datasets for development is crucial.

One important area where poor validation could cause significant issues is in medical imaging, where reliable results after deployment is a major concern. Given the limited supply of data it is common for researchers to test repeatedly on the same dataset, fine tuning their methods until they achieve strong results on the testing set but in turn possibly over-fitting their methods to said dataset. To address this most researchers also validate their methods on natural imaging datasets; however, given their innate differences may not hold a significant correlation to future medical imaging contexts. One aim of this work is to show that testing new methodologies on standard testing sets has a worthwhile correlation with medical imaging problems, showing that this validating medical imaging methods with natural imaging datasets is feasible.

4) BENCHMARKS AND ANALYSIS OF DOMAIN GENERALIZATION METHODS

There has already been some work in creating a framework for assessing domain generalization techniques by researchers at Facebook, who created the DomainBed framework [1]. The framework contains 10 datasets and 26 different domain generalization algorithms. One of the most important findings of this work was discovering that there was little difference in generalization between all the included algorithms. As such, understanding why these methods fail to generalize better than baseline ''do-nothing'' methods is a critical step in furthering domain generalization research in all contexts.

It is important to note that DomainBed only contains a single medical imaging dataset, WILDS Camelyon17, a histopathological dataset. As such, the framework may not be particularly useful for discovering which techniques will work well in different medical imaging modalities. In order to get a more accurate representation of algorithmic performance in the medical imaging field, it would be necessary to incorporate open-access domain generalization datasets for at least the most popular medical imaging modalities such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and positron emission tomography (PET). It may also be worthwhile to include medical diagnosis methods that use standard optical capturing, such as skin cancer photography.

There has been very limited works in creating benchmarks for medical imaging datasets, the most notable example being the work by Zhang et al. [36]. However, this work focuses specifically on methods that are modality agnostic and can be applied to both tabular data and medical imaging. As such, the methods used and analysed are limited. Additionally, only two datasets are used to test these chosen algorithms (a tabular

dataset ''In-Hospital Mortality (eICU)'', a combination of four open-access Chest X-Ray datasets, and several synthetic domain-shift augmentation methods for further testing and training on the two previous datasets), which does not include many key modalities mentioned earlier.

In addition to there being no benchmark for domain generalization algorithms in the most common medical imaging modalities, there is also another missing component in the literature: investigating the performance difference of algorithms between natural imaging and medical imaging datasets. The differences between natural and medical imaging are numerous, which gives rise to a concern that algorithms that perform well on one may not operate with the same level of performance on the other. On the surface level, medical images tend to have much lower subject matter variety, as they are almost always only taken of a single organ or biological component. Whereas natural images have a significantly larger amount of variety both in the subjects, but also the background and context in which the images are taken in. There is also a difference between the classes within each dataset. In medical images, the image classes are often determined by minute discrepancies in texture or small anomalies in the structure of the subject. Whereas in natural images, the entire structure of the image tends to change based on the class. For example: classifying cancer tends to be a task in finding small lesions or changes in the texture of an organ, as opposed to distinguishing a clock from a dog, which has much larger structural and textural changes.

There is also a difference in the capturing mechanisms themselves. In natural image photography, light comes from external sources and reflects off an object into the camera's lens. Thus what is captured is a product of the environment and the subject, not solely the subject itself. As opposed to medical imaging, where the external source of ''light'' is controlled significantly to ensure that the data being captured is as representative of the subject as possible. In medical images the values within the image often have a physical meaning, in the case of a computed tomography scan the units are Hounsfield units (HU), which represent the physical density of the subject being captured. Whereas in natural images, the values (often the intensity of red, green, and blue light for each pixel) have no direct physical meaning, beyond the amount of red, green, and blue light hitting the sensor at the time of the photograph, which may be inaccurate due to aforementioned external light sources (such as the sun) having a distinct colour tint. Another concern is with the purpose behind the capturing of each of the image types. Natural images are often captured for the purpose of human consumption, in modern photography there are significant post-capturing processing effects performed to alter the data captured to make the images more pleasing to the ''human eye'', and disregarding the actual reality of the subject being photographed. In contrast, medical images are taken for diagnosis purposes, as such must represent reality as closely as possible to aid the diagnosis procedure. While post-processing is performed in medical imaging, it is most often transformations like noise reduction which aim to reduce the effects of an imperfect capturing process, as opposed to making images more aesthetically pleasing. Lastly, there are also differences in the types of domain shifts seen between medical imaging and natural imaging. The differences in medical imaging often come down to the process used to capture the scans, of which every modality has different underlying physics that can cause different types of shifts [37], [38], [39], but in general the shifts are often quite small, due to the controlled nature of the scanning process. Meanwhile, in natural images the shifts between domains can be extremely large, with differences in background context, the colour of the surrounding light (night time vs day time photography) impacting how the subject is viewed, as well as the types of camera being used which can distort the perspective of the subject, along with many more possible shifts.

Overall, there are many differences between medical and natural images, that confirms the importance of affirming that methods that have been solely tested in natural imaging datasets will yield similar results when tested on real-world medical imaging problems.

Likewise, there may be more differences between how methods operate between the two contexts than just accuracy on out-of-distribution accuracy. As stated previously, the differences between domains in medical and natural contexts leads to different possibilities for how models may interact with them. For example: since the differences between domains are often smaller, with little stylistic differences, does this mean that domain invariant representations are likely to be more effective or easier to achieve? Due to the smaller differences between classes is the variation between feature representations likely to be smaller as well? These are just two possible ways in which the discrepancies between medical and natural imaging could display themselves. This work aims to explore the way these differences impact the model's feature representations, and how different generalization mechanisms operate on them.

This work aims to build on the foundation laid by the works [1], [40]. Guljrani et al. in the work ''In Search of Lost Domain Generalization'' were the first to discover and show that under controlled circumstances the difference in performance between domain generalization methods aggregated around the baseline method, ERM. Secondly, Galstyan et al. created a framework for analyzing domain generalization techniques and their domain invariance capabilities, while also showing most methods do not achieve adequate invariance on unseen domains. The goal of this work is to unify and expand on all of these issues. First by providing a series of benchmarks for eight algorithms on three standard natural image object classification and four real-world medical imaging tasks. We validate that domain generalization algorithm performance on natural imaging datasets continues to medical imaging datasets. As well as compiling datasets that can be used to directly assess models in the medical imaging domain. Secondly, we combine and improve on an

evaluation framework for domain generalization algorithms by analyzing created models in terms of domain invariance, distributional robustness, and representational smoothness. Lastly, through this framework we highlight several issues with current domain generalization methods and offer some insight into where future work could be dedicated to improving their performance.

II. METHODS

This section provides an overview for the experimental setup used for this work. This includes what datasets and algorithms were chosen and why, as well as the methods used to analyse the performance of the algorithms on each of the datasets.

A. DATASETS

Dataset selection process: In order to compare natural imaging to medical imaging in a domain generalization context, a representative set of datasets needs to be chosen from both fields. To reiterate, while the datasets are chosen to be representative of real-world applications, the goal of this work is not to provide state-of-the-art results and/or demonstrate strong performance on real-world problems for clinical use; instead, the goal in using these medical datasets is to assess domain generalization methods in a wide variety of situations that can be used to indicate trends in how these algorithms operate, and where future work is best to focus to improve outcomes.

1) NATURAL IMAGING DATASETS

One of the fundamental tasks of computer vision is object recognition: Take a photograph and try to predict what is the subject of the image. Despite being the most common task, it still remains difficult to do with high accuracy due to the seemingly large variation between objects within images and the context in which they appear, as well as things as simple as the camera used to take the images can have a noticeable impact [41]. Given the variety of work in this field, there are many open-access and varied datasets used to validate new methods, three that focus on domain generalization in natural imaging are: PACS, VLCS, and OfficeHome. The datasets used were included within the DomainBed framework [1].

The first dataset, PACS [1], [42] consists of four domains: Photos (1,670 images), Art Paintings (2,048 images), Cartoons (2,344 images) and Sketches (3,929 images). Each domain contains seven types of objects, and has a roughly equal distribution of samples for each class between domains. The images are of a standard resolution, 227 × 227 pixels, but all images are still resized to 224 × 224 with bilinear interpolation as part of the data augmentation strategy.

VLCS [1], [43] is an amalgamation of four other natural imaging datasets: Pascal 2007 (V), LabelMe (L), Caltech (C), SUN09 (S), each of which has five classes of images. The number of samples from each domain for each class varies significantly, for example: SUN09 only has 20 bird images, but has 1036 chairs, whereas Pascal 2007 has 330 birds but only 428 chairs. The range of sample sizes from each domain

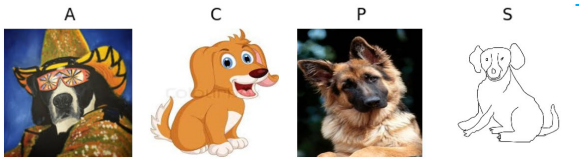

FIGURE 1. Sample images from each domain of the PACS dataset, from left to right: artistic rendition, cartoon, photograph, and sketch.

for each class has a minimum of 20 and a maximum of 1499; each domain contains a different distribution of classes as well. The images in VLCS do not have a standardized resolution, as such all images are resized to 224 × 224 pixels with bilinear interpolation.

FIGURE 2. Sample images from each domain of the VLCS dataset: Caltech, LabelMe, Sun09, and Pascal 2007.

The final natural imaging dataset is OfficeHome [1], [44] which is comprised of four domains: artistic images (A), clipart (C), product (P, images of objects without a background), and real-world (R, regular photographs of objects). In total there are 15,500 images, with each of the 65 classes ranging on average from 40 to 80 samples each. As with VLCS the images in OfficeHome do not have a standardized resolution, as such all images are, again, resized to 224 × 224 pixels with bilinear interpolation.

FIGURE 3. OfficeHome dataset sample images: artistic image, clipart, product, and real-world.

Ye et al [45], in their work on OoD-Bench, separate domain variations into two axes: correlational and diversity shifts. Correlational shifts are differences in non-causal elements of data which can be used as shortcuts to learn classification tasks. For example, in a horse vs camel classification task if all images of horses have plains of grass in the background while all camels images have a desert background, a model might learn that grass is strongly correlated with horses and deserts for camels rely on that for classification, as opposed to recognizing the actual animal. This can be seen in VLCS where all Caltech images are close-up photographs of dalmatians, thus looking for black and white spots would somewhat accurately predict the "dog" class for that domain. Whereas diversity shifts are differences that are changes in causal features, as can be seen in the PACS dataset, where different knowledge is required to identify the same objects between sketches and photographs.

The domain differences in PACS revolve around this diversity shift principle. The importance of edges vary between domains, being the most important in sketches, but are prevalent in cartoons as well, whereas artistic renditions and photographs rely more so on colour and texture along with general shapes to identify objects. Both artistic renditions and photographs have realistic backgrounds, whereas cartoons and sketches have a flat coloured backgrounds (mostly white). Art also tends to have additional textures from the media that created it, for example: paint brush strokes, which are unrelated to the content, whereas photos all textural information is related to the class.

The domains in VLCS are more difficult to differentiate compared to PACS, given all four datasets are photographs. However, as Torralba et al. [43] describes each class within the dataset contains biases, for example in the car class some datasets contain varying numbers of each type of car (sports, sedans, etc), or different viewing angles, and different background/foreground elements. Similarly, the dog class in the Caltech domain are almost entirely close-up photos of dalmatians, with other domains having significantly more diversity. These differences between domains do vary from class to class, but despite all being photographs, there are difference between them that are somewhat correlational and almost entirely unrelated to stylistic diversity.

Given the variety in differences in domain discrepancies between these three natural imaging datasets, while not covering every circumstance within natural imaging, cover a broad enough spectrum of contexts to provide a suitable benchmark for how well domain generalization for object classification can perform in real-world applications.

2) MEDICAL IMAGING DATASETS

While the focus of medical imaging is much more limited (images of different parts of the human body), the methods of taking the images have higher variation and the detail required to make classifications are more subtle, thus capturing all the variation within medical imaging is a difficult task. In an attempt to capture as much variation as possible, we have selected different datasets from the most common imaging modalities: CT, MRI, and histopathology. In order to keep a consistent format (and for use in the ResNet50 model) all 3D scans were split into individual slices, images were resized to a resolution of 224 × 224 with bicubic resampling and were normalized to a range of 0 to 1. While this is not useful for patient level diagnoses as required in medical imaging, it should still give an accurate representation for how well domain generalization methods can create generalizable models on medical imaging data.

a: CANCER METASTASES IN LYMPH NODES CHALLENGE (Camelyon17)

The first medical dataset is WILDS Camelyon 17 [46] (which is also included in the DomainBed framework) is a set of 96 × 96 histopathological images, with the task of predicting if the central 32 × 32 region contains any tumour tissue. The samples were collected from 5 different hospitals, each with an equal number of positive and negative cases, though each domain has varying numbers (ranging from 17452 to 73361).

The most common difference between capturing methods in histopathological images is the staining method for highlight specific regions or structures of cells within the sample [47]. To our eyes, these differences can be seen in the saturation and intensity of the colouring of those chosen regions/structures within the images. Generally these differences are similar across domains, given the procedures are the same, however there still remains some variation within domains [48]. Ye et al, [45], classifies Camelyon17 as entirely diversity shift, as there is little domain-dependent differences in samples conditional on classes.

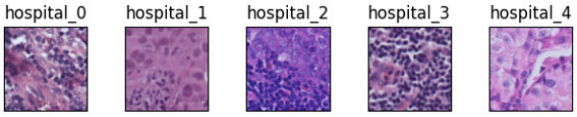

FIGURE 4. Sample histopathology images from each hospital within the WILDS Camelyon 17 dataset.

b: RETINAL OCT FLUID CHALLENGE (Retouch)

The Retinal OCT Fluid Challenge (Retouch) [49] was designed to compare the performance of detection and segmentation of various types of fluid in optical coherence tomography images between scanner manufacturers. For the original challenge there were 3 scanners included: Cirrus, Spectralis, and Topcon scanners each with approximately 23 3D volumes of retinas. As the volumes provided are 3D, to ensure the same model structure can be used between all datasets, the scans were split into slices using the provided segmentation masks to provide slice-level annotations for the presence of retinal fluid. The original challenge was also comprised of the independent detection of 3 separate fluid types (labeled L1, L2, and L3). As the labels are independent, the models were trained and tested on the detection of L1 fluids solely due to having the most balanced number of positive and negative samples across domains. All samples were also normalized to a range of 0 to 1. The original scans also contain large amounts of black space below the retinal layers, which was cropped out manually to a final image size of 512 × 512.

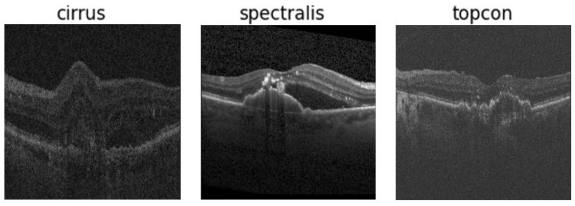

FIGURE 5. A sample slice from a scan from each manufacturer (Cirrus, Spectralis, and Topcon) within the Retouch dataset.

The differences in domain between each of the scanners is much more apparent in the Retouch OCT dataset, given

each has very different capturing parameters [50]. The most obvious difference being the intensity shift, Topcon samples have a higher average pixel intensity (after normalizing to ImageNet means and standard deviations) of −0.8 as opposed to Spectralis' −1.3 and Cirrus' −1.2, which leads to a much brighter appearance. Similarly, Spectralis samples have a higher contrast between the background and the retinal layers (Cirrus ranges from −2.12 to 1.51, Topcon from −1.72 to 1.77, and Spectralis from −2.12 to 2.64). These differences are highly consistent across samples within each domain, so there is little correlation shift found in the OCT dataset.

c: CARDIAC

MRI Anomaly Detection The Multi-Disease, Multi-View & Multi-Center segmentation challenge (MNM2) [51] was originally designed to test the segmentation of different sections of a human heart that has different ailments. The original dataset was comprised of three domains from different scanners: GE Medical Systems, Philips, and Siemens, with 53, 88, and 219 patients, respectively. In total there are 70 normal subjects, and 290 with other heart-based issues. Due to the distribution of illnesses across domains being unequal (some domains containing zero samples from some illnesses) the data was compressed to be a simpler anomaly detection task, using the normal subjects as one class, and all other illnesses combined into another class. Minor preprocessing was needed to convert the data into 2D slices, namely using the ground-truth segmentation masks of the heart to isolate slices to the heart, the scans were clipped between 0 and 2000 then normalized to a range of 0 to 1. As the challenge was also multi-view, only the end-systolic phase scans were used for training and testing.

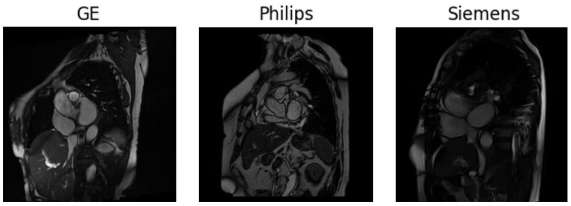

FIGURE 6. Sample slices from each domain within the MNM2 dataset: GE Medical Systems, Philips, and Siemens scanners from left to right.

MRI scans have significantly more standardization between manufacturers, given the possibility of using phantom scans for calibrating the machines. However there still remain slight differences in capturing resolution and field strength [51]. By the human eye however there is little noticeable difference between the scanner types, apart from the Siemens scanner having a very slight increase in average image intensity, of −1.56 compared to GE and Phillips' −1.69, though this varies on an image-by-image basis. There are also some slight differences in noise as well, with GE and Siemens appearing smoother in texture compared to Philips.

d: COMPUTED TOMOGRAPHY (CT) SCANS FOR COVID-19 DETECTION

The final medical imaging dataset is 4 separate datasets of lung CT scans comprised of COVID positive cases and normal cases. The first is a set from Mohammad Rahimzadeh et al. (called COVID-CTSet) [52], which contains 90 covid positive lung CT scans, and 285 ''normal'' lung CT scans. The Radiological Society of North America (RSNA) has released an open dataset of 120 COVID positive chest CT scans (MIDRC-RICORD-1a) [53], [54], along side another set of 120 ''normal'' (non-COVID positive and without symptoms) chest CT scans (MIDRC-RICORD-1b) [53], [55]. This data was released on the Cancer Imaging archive [56]. The third dataset released by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, released MosMedData [57] a dataset of 1110 chest CT scans: 856 scans having signs of COVID within the lung of varying severity, and 254 having no signs at all of COVID. The final dataset is a set of chest CT scans from the China Consortium of Chest CT Image Investigation (CC-CCII) [58] (called NCOV in this paper) containing 1480 COVID positive scans and 1101 control scans.

The Mosmed, MIDRC, and COVID-CTSet datasets had voxel values clipped to a range of -1000 to 2000 HU then normalized to a range of 0 to 1, whereas the NCOV dataset contained images that were already normalized to an unspecified range. Each dataset also required some specific preprocessing: MIDRC required its dimensions to be transposed in order for the slices to be along the axial plane, and the MosMed samples needed to be rotated 90 degrees. NCOV had many scans with the lungs masked already, which is a significant difference to the other domains and cannot be undone, but as our focus is on the internals of the lungs, the results should not be impacted.

Initially all slices that had no lung tissue were removed. Where possible COVID segmentation masks were used to get slice level labels, and where they were not provided the labelling was done manually. This process was not performed by a medical expert, as such there is room for some error, thus all accuracy results reported on this dataset is just an indication of general domain generalization performance. Ideally, native slice-level annotated datasets would be used; however, open access domain generalization datasets for medical imaging are rare, so some leniency is necessary in order to explore more modalities. These scans were also resized to 224 × 224 pixels with bicubic interpolation to keep as many variables constant as possible between datasets, as the purpose of this work was to analyse and compare methods, not to achieve the highest accuracy models for clinical use. This approach was validated using unseen source domain accuracy, as can be seen in figure 12. In all medical imaging datasets sufficient training domain performance is achieved which implies that 224 × 224 is suitable for classifying these datasets and thus

comparing methods to one another as is done in this work is valid.

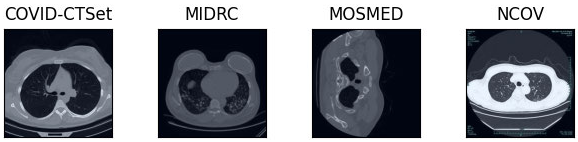

FIGURE 7. Sample slices from each of the chest/lung CTs for COVID detection from each chosen public dataset.

The differences between the COVID CT datasets are mostly small, with some notable exceptions that were previously mentioned: MOSMED were rotated 90 degrees, and NCOV was normalized to non-standard HU values and had the lung internals masked. Apart from these, the scans had minimal variation in contrast and average intensity and no noticeable discrepancy in textural details within the lungs.

B. ALGORITHMS

For this analysis eight different domain generalization algorithms were selected from DomainBed. The purpose of this work is to analyse a variety of domain generalization mechanisms, how they operate, and their differences in their final model's capabilities.

The model architecture was shared between all of the methods and can be described as a sequential pipeline of a feature extractor and a classifier. The feature extractor is a function (F) that reduces an input (X) to a vector of ''features'' (f) that ideally retains as much relevant information to the selected task (or tasks) as possible. In this work the feature extractor is the ResNet50 backbone, which transforms images of size $3 \times 224 \times 224$ to a vector of size 2048.

$$f_{x_i} = F(x_i) \tag{1}$$

The output of the feature extractor is then fed into a ''classifier'' function (C), which returns a prediction the class label, or set of labels, ($\bar{y}$) given a feature vector.

$$\bar{y_i} = C(f_{x_i}) \tag{2}$$

The aim of the selection was to capture a wide variety of both different generalization methodologies (as outlined in section I-B) and reported accuracy within the DomainBed framework. The chosen methods are: ERM, IRM, DANN, MMD, SagNet, Fish, Mixup, and SelfReg (as labeled in the DomainBed framework). These methods were chosen by selecting methods that operate using distinct generalization mechanisms. ERM is the baseline method for comparison that has no generalization mechanism. IRM, MMD, DANN, and SagNet operate similarly with domain invariance constraints but with different focuses and approaches. IRM enforces prediction invariance, whilst MMD and DANN constrain the feature representations, but using different tools to do so (statistical alignment vs adversarial learning). SagNet takes a different approach by attempting to separate domain invariant and domain specific knowledge (content and style). While these methods have a similar goal their mechanisms differ greatly. These specific methods were chosen to represent domain invariant learning based on their spread of average out-of-distribution accuracies from the original DomainBed paper, covering the worst performing method, MMD, to second best, SagNet. Three additional methods were selected, Fish, Mixup and SelfReg: Fish and Mixup use representational regularization to train models to learn smoother representations between domains. SelfReg uses both domain invariance and representational regularization. This selection therefore should give a sufficient overview of how different types of algorithms could perform in the medical imaging context, and lead to insight from the assessment of their failures.

All apart from one of the selected methods can be split up into two different categories based on their main mechanism for achieving domain generalization (either domain invariant representations, or representation regularization/smoothness):

Domain Invariance:

- Invariant Risk Minimization (IRM) [59]
- Domain Adversarial Neural Networks (DANN) [7]
- Maximum Mean Discrepancy (MMD) [8]
- Style Agnostic Network (SagNet) [11]

Representational Regularization:

- Gradient Matching for Domain Generalization (FISH) [20]
- Inter-domain Mixup (Mixup) [23]

Empirical risk minimization (ERM) is a baseline method, which performs no explicit training for generalization as such it does not fit into either of the above categories. It simply trains a model on all available training data and hence provides a useful point of comparison for other methods.

Invariant risk minimization [59] aims to learn representations that are invariant across training environments by constraining the learning to enforce that predictors are also invariant across training domains.

Domain adversarial neural networks [7] operate by using an adversarial domain discriminator to enforce a degree of domain invariance within the learnt feature representations. Additionally; the maximum mean discrepancy (MMD) [8] method acts as an extension of domain adversarial networks by introducing an additional MMD regularization term to align feature representations with a prior distribution to aid domain invariance.

Gradient matching for domain generalization (FISH) [20] aims to align the direction of gradients across training domains by maximizing the inner product of each domain's gradients, thus moving the model towards domain invariance.

Inter-domain Mixup [23] training uses a linear interpolation between images and their corresponding class labels from different domains to generate new mixed samples to learn from to enforce a linearity constraint between domains enabling stronger generalization.

Self-supervised Contrastive Regularization [19] trains feature representations by mapping samples of the same class

TABLE 1. A details summary table for the medical imaging datasets. For the COVIDCT, MNM2, and OCT datasets the cases represents a separate 3D volumetric sample which is split along one dimension to form individual images for training (as described in each datasets section).

Name	Image modality	Total # Cases	Total # Training Images	# Classes	Domains	# Cases	Resolution	Format
COVID CT	Lung CT	4306	141083	2	COVID-CTSet MIDRC MosMed NCOV	375 240 1110 2581	512x512	.tif 3D .dcm 3D NIFTI .png
MNM2	Cardiac MRI	360	4064	2	GE Philips Siemens	25 44 91	256x256	3D NIFTI
OCT	Retinal OCT	68	6936	2	Cirrus Spectralis Topcon	24 23 21	1024x512 496x512 650x512	3D MHD
Camelyon17	Histopathology	455954	455954	2	0 1 2 3 4	59436 34904 85054 129838 146722	3x96x96	.png

closer together and different classes further apart, thus leading to more smoother representations across the feature space and hopefully more domain invariant but still class discriminative features. Given this method uses similar concepts as inter-domain Mixup to lead to smoother representations, it also explicitly pulls representations closer together leading for domain invariant representations as well, as such it fits into both categories of domain generalization mechanisms.

The final method is the Style agnostic network [11] that learns to separate content and style knowledge by training a content-specific network through randomly swapping out the style of input images with other images using AdaIN. Through this separation, the SagNet reduces the model's reliance on textural information in input images, and becomes more domain invariant.

While there are many more algorithms and methods for domain generalization, these seven methods, plus the baseline ERM, cover a wide variety of generalization mechanisms and theories. We think the selection is wide enough for a substantive analysis under the failure mode framework for the different types of domain shift found in various medical imaging problems.

C. ALGORITHM ASSESSMENT - FAILURE MODES OF DOMAIN GENERALIZATION ALGORITHMS

To assess the capabilities of each algorithm a variation on the novel Failure Modes of Domain Generalization Algorithms framework by Galstyan et al [40], is employed. This framework assesses algorithms using 7 metrics in total the mathematical definitions and further visualizations and descriptions for which can be found in the original paper [40].

The first four are **classification based techniques:**

Training set underfitting, called e_0, measures the performance of the entire model on unseen training domain data, this metric highlights whether the model is underfitted and requires more training.

Test set inseparability, e_1, is a metric for how well the feature extractor can extract meaningful information from out-of-distribution test data, it is calculated by finding how strong the performance of a classifier is on fixed extracted features on the test data.

Training-test misalignment, e_2, aims to find how well a single classifier can be used to classify both training IID data and testing OOD data. If a classifier can perform well on both, then the training and testing distributions are aligned.

Classifier non-invariance, e_3, measures the performance of the entire model on unseen testing OOD data, i.e., the final generalization capability of the model.

The last three metrics revolved around predicting the domain of origin for each sample.

Domain prediction metrics:

Training domain distinguishability, d_0, measures the domain invariance of the training data representations by training a new classifier to predict the domain of each sample given the output of a fixed trained feature extractor on the training domain data.

Training-test domain distinguishability, d_1, measures how well domains can be distinguished on both the training and testing sets; This is found by training a new classifier to predict the domain given the fixed feature representations of both the training domain data and the OOD test data.

Training-test class-conditional domain invariance, d_2, is the final metric, which measures how well the representations from each class independently can be separated into their respective domains.

In the original work, these metrics were displayed as losses. However, in order to provide a more intuitive comparison between a larger number of models and methods, we report the top-1 accuracy for each of these metrics. While it is standard to report the area under the receiver operating characteristic curve (AUC-ROC) in medical imaging, it is used in binary classification problems. Given the natural imaging datasets are multi-class, we have opted to use top-1 classification accuracy to ensure valid comparison of accuracies between natural and medical imaging datasets. Reporting both metrics e_0 and e_3 require no further computations beyond the calculation of accuracy on the unseen in-domain validation set and the out-of-distribution set respectively with the final trained model. Whereas the calculation of each of

e_1, e_2, d_0, d_1, and d_2 requires a new classifier to be trained. For each of these metrics a new linear classifier was trained using categorical cross entropy loss with a learning rate of $1e-5$ for 50 epochs. The input for all of them is kept the same: The output of the (fixed) feature extractor trained for domain generalization (though which domain's features are included depending on the metric being calculated) and the label (which can either be the target class for e_n metrics or the domain for d_n metrics). For e_1, the classifier is trained on only the out-of-distribution test set features, and the accuracy is calculated on an unseen subset of OOD test features, where as e_2 uses features from both the training and test domains, and accuracy is calculated on unseen subsets of both. d_0, d_1, and d_2 all use the origin domain as the label instead of the class, thus training the classifier to predict the domain instead. d_0 uses only the training domain features as input, whereas d_1 uses both training and test domain features. d_2; however, trains a new domain predicting classifier for each class independently, and then averages the domain prediction accuracy across each class, leading to a class-conditional domain invariance accuracy.

The metrics e_1, e_2, d_0, d_1, and d_2 were calculated using a classifier as opposed to a simple distance metric as we are interested in measuring the separability of the features as opposed to the absolute distance between the representations. This approach was used by Galstyan et al. [40]. Additionally, the calculated metric is most applicable when the architecture used to calculate the discrepancy is the same architecture as what is used for the main task, given all our methods use linear classifiers then the most important metric is how well the same linear classifier can separate the features (both by class and by domain.

We also add another metric to the above list in the form of a maximum theoretical domain generalization accuracy, which does not need any additional calculations and can be found by calculating the average classification accuracy of the trained models on unseen data from each training domain. This values gives an upper bound on how well a domain generalization algorithm can perform; e.g., if a model trained on data cannot perform better than 80% accuracy then a model trained on out-of-distribution dataset should not perform better than 80%.

Metrics e_0 through e_3 and maximum theoretical accuracy were class-weighted top-1 accuracy scores to account for the variation in number of samples between classes, where as d_0 through d_2 were weighted by the number of samples from each domain for the same reason.

As with domain invariance: by having a low ability to predict the domain of origin from a set of features, the impact of domain shift should be lessened, thus leading to stronger predictions on novel and unseen domains.

D. REPRESENTATION SENSITIVITY

In addition to the failure mode framework, we also introduce a new model evaluation metric called feature/representation sensitivity. The purpose of this metric is to indicate how susceptible a model is to small shifts in the input space. In an ideal world, only when the qualities of the input that are directly causative to the class label are changed should the feature representation shift. Thus by measuring how much the features shift with respect to small shifts in the input, it is possible to quantify the robustness of a feature extractor. This would be useful in analyzing methods for domain generalization for two reasons. The first being that a model that is susceptible to very minor shifts in the input space is unlikely to also be robust to the larger shifts (those found between different domains). The second is that these very small shifts in the input are often seen in medical imaging. This representational sensitivity then should be useful as a part of a framework for understanding how generalizable models are. The following equation illustrates how the values were calculated in this work:

$$S_f = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\left|\frac{dF(x_i)}{dx_i}\right|\right|_2^2} \tag{3}$$

The above equation is used to calculate the root-mean squared deviation of the gradients of feature vector, $F(x_i)$, w.r.t. the input image x_i. Intuitively, this value indicates how much the feature vector varies when the input image is shifted. A high variance indicates that small shifts in the input image would lead to large changes in the learned representation of that image.

With representations that are more robust to changes unrelated to class information, the model should be more robust to future unseen domain shifts, leading to stronger out-of-distribution accuracy.

E. REPRESENTATIONAL DISTANCE BETWEEN DOMAINS

While the failure mode framework uses a classifier to calculate the distinguishability between domains, finding the actual distance between the representations from different domains may also give some insight into how domain generalization methods operate. As such, we have included a brief analysis of the average cosine distance between domains.

The average cosine distance used in this work was calculated as follows. Equation 4 shows the calculation of the cosine distance between two vector, A and B:

$$Dist_C(A, B) = 1 - \frac{A \cdot B}{||A||\,||B||} \tag{4}$$

Then the average distance between two domains (D_m and D_n), we find the sum of the pairwise distance between all feature representations from each domain.

$$Dist_{C_{avg}}(D_m, D_n) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} Dist_C(F(x_{m,i}), F(x_{n,j})) \tag{5}$$

For this work we use equation 5 to calculate three types of distances. **1)** OOD data to IID data, to find how far away the OOD representations are from IID training features. **2)** IID to IID, to see how far apart the different IID training domains

are. **3)** OOD domain features to the same OOD domain's features, to see how much variation there is within a single domain's representations.

As an extension of domain distinguishability, domain distance indicates the discrepancy between the feature representations of different domains. Thus, a lower distance between domains should indicate a general level of stronger domain invariance.

F. TRAINING METHODOLOGY

All domain generalization methods used a ResNet50 (pre-trained on ImageNet [60]) as a feature extractor with a linear classifier head trained for 50 epochs with a learning rate of $1e-5$ using the ADAM optimizer, with a mini-batch size of 64. The ResNet50 architecture was used due to it being open source, with numerous implementations and pre-trained weights available, as well as its strong performance in many deep learning applications, leading it to be applicable as a benchmark for a wide variety of problems. We use the same data augmentation strategy (as per the DomainBed framework) for all methods and datasets: Random crops and aspect ratio adjustment, resizing all images to 224 × 224 pixels, random horizontal flips, random colour jitter, random gray-scaling images, and normalizing data to the ImageNet mean and standard deviation. This resizing was kept consistent across all datasets to keep as many variables controlled between methods and datasets as possible. Regardless, 224 × 224 was shown to be adequate for classifying all datasets, as seen in figure 12 As not all of the chosen datasets have enough source domains to allow for a held-out domain for hyper-parameter tuning, all tuning and model selection was done using held-out training domain data (named IID Validation Accuracy in the DomainBed framework). While this is not optimal, it is rare for medical imaging datasets to contain enough domains to learn from effectively and to perform testing on. Likewise, given we aim to assess general performance (and not to generate state-of-the-art results), minimal parameter tuning/selection is required. All experiments were conducted on an NVIDIA Quadro RTX A6000 GPU and Intel I7-12700k.

III. RESULTS

A. OUT-OF-DISTRIBUTION PERFORMANCE

The first, and perhaps most important, set of results are the final out-of-distribution test set accuracy, which indicate the actual domain generalization performance each method achieved on each dataset. As can be seen in tables 2 and 3, all methods perform within a small margin of error (±0.015) of the baseline (ERM), which is consistent with prior work.

To remain consistent, the class-weighted top-1 classification accuracy was measured on the best performing model from each run (highest in-domain validation set accuracy). Each dataset was run across with each domain acting as the held-out final test set three times and their results averaged together.

TABLE 2. Average top-1 class-weighted classification accuracy on out-of-distribution data of chosen algorithms on natural imaging object classification datasets, across all domains.

Method	VLCS	PACS	OfficeHome
ERM	0.759 ± 0.02	0.825 ± 0.01	0.631 ± 0.02
IRM	0.763 ± 0.02	**0.842 ± 0.02**	0.622 ± 0.03
MMD	0.760 ± 0.03	0.830 ± 0.03	0.610 ± 0.04
DANN	0.752 ± 0.03	0.768 ± 0.01	0.611 ± 0.02
Fish	**0.769 ± 0.02**	0.829 ± 0.02	0.638 ± 0.02
Mixup	0.759 ± 0.03	0.823 ± 0.04	0.620 ± 0.04
SagNet	0.759 ± 0.02	0.837 ± 0.02	0.639 ± 0.03
SelfReg	0.758 ± 0.02	0.833 ± 0.03	**0.648 ± 0.04**

As can be seen in table 2, most methods perform equally with only minor variations, with outliers: DANN on PACS and MMD on OfficeHome, both of which had poorer performance across validation domains as well, which indicates poor training set fitting.

TABLE 3. Average top-1 class-weighted classification accuracy on out-of-distribution data of chosen algorithms on medical imaging datasets, across all domains.

Method	OCT	COVIDCT	MNM2	WILDSCamelyon
ERM	**0.911 ± 0.01**	0.636 ± 0.01	0.734 ± 0.05	0.921 ± 0.01
IRM	0.896 ± 0.01	0.638 ± 0.03	0.737 ± 0.04	0.915 ± 0.01
MMD	0.907 ± 0.02	0.633 ± 0.03	0.729 ± 0.10	0.906 ± 0.03
DANN	0.880 ± 0.02	0.627 ± 0.03	0.695 ± 0.12	0.896 ± 0.03
Fish	**0.911 ± 0.01**	0.615 ± 0.01	0.727 ± 0.02	0.922 ± 0.01
Mixup	0.896 ± 0.01	0.620 ± 0.02	0.701 ± 0.09	0.927 ± 0.01
SagNet	0.899 ± 0.01	0.625 ± 0.02	0.706 ± 0.06	**0.937 ± 0.01**
SelfReg	0.893 ± 0.01	**0.656 ± 0.04**	**0.741 ± 0.10**	0.929 ± 0.02

Similarly, table 3 shows the same pattern as seen in natural imaging, overall performance stays similar regardless of what algorithms are used. However, there is a slightly higher variation in accuracy for each dataset (though with no consistent pattern across methods) with differences between the highest and lowest accuracy ranging from 0.031 to 0.046 in medical datasets as opposed to 0.016 to 0.037 (ignoring DANN on PACS as an outlier) seen in natural imaging. The most notable concern is the poor performance in the COVID classification in CT datasets. There are many possible explanations for the poor performance though, as outlined in section II-A2.d.

Overall these findings are inline with expectations set by the original DomainBed framework, which stated that generalization performance rarely improves significantly over ERM.

B. CLASSIFICATION-BASED FAILURE MODE ANALYSIS

Next, we will go over the results of our implementation of the failure modes of domain generalization framework by Galystan et al. [40]. For the sake of brevity most figures have been placed in appendix B. Overall, minimal difference was found between any of the methods, this can be seen in figure 8, which shows the average classification metrics for all methods.

Metric e_0 (figure 12) shows all algorithms are successful in classifying the training domains in all datasets, with some

variation based on the dataset. The variation caused by algorithm selection is inconsistent, the only pattern that can be seen is DANN consistently has lower training set accuracy but only slightly. Likewise, there is little correlation between metric e_2, training-test alignment, and the domain generalization algorithm chosen, with all methods performing equally on all datasets within a small margin for error except for MMD and DANN which under-perform in several cases.

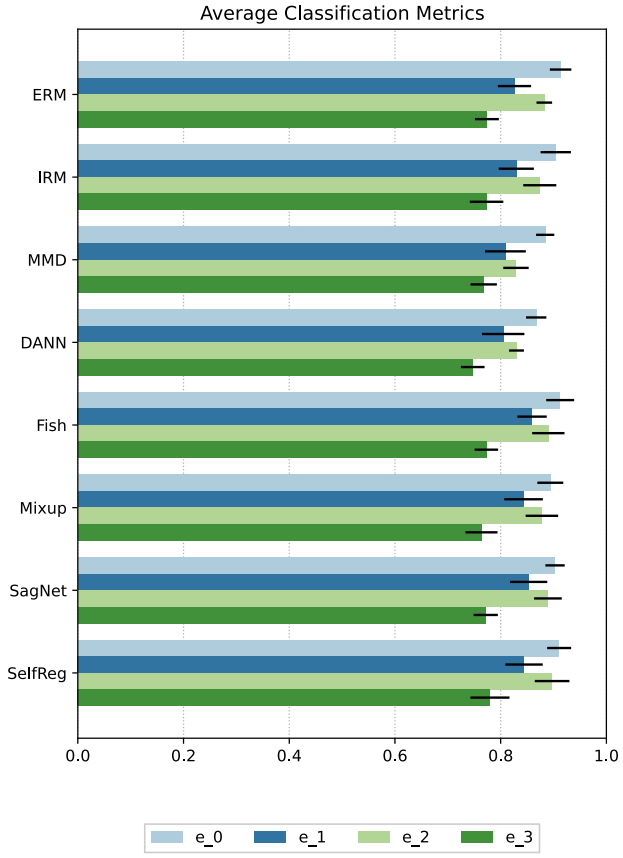

FIGURE 8. Each algorithms' e_0, e_1, e_2, and e_3 values averaged across all datasets.

The more important classification metrics within the framework are e_3 (classifier non-invariance) and e_1 (test set separability) (table 3 and 2 for e_3 and figure 13 for e_1). As can be seen in figure 15, the actual accuracy of the final created models for each dataset by each method (and as was seen earlier in tables 2 and 3) final generalization performance does not change noticeably between methods, there is no algorithm that consistently outperforms the baseline ERM, which again is inline with past research. In comparison to figure 13 (metric e_1, test set separability), which shows how well the test data can be separated given the trained feature extractor, the actual performance of the models is only slightly lower than the highest possible accuracy given the same feature extractor by 0.04 on average. This indicates that the test set separability of the features is the bottleneck in almost all of the tested methods, with one exception: COVIDCT. In the COVIDCT dataset there is a significantly greater difference between e_3 and e_1. With the average difference being approximately 0.2. This indicates that there is a misalignment in the classifier for COVIDCT, as opposed to the feature's generated not being generalizable. However, in order to improve the generalization capability on the rest of the datasets, future methods should focus on creating feature extractors that generate class discriminative features in unseen domains. This goal then points to the necessity of finding domain invariant features, which is explored in the next section (section III-C), and is a problem that requires further attention.

Figure 16 gives the highest achieved in-domain validation accuracy found by all methods on all datasets across all runs. While not a firm upper bound, it does give an indication for the best possible classification result these methods could achieve (i.e., with no domain shift present). On average the difference between the actual accuracy (figure 15) and maximum theoretical accuracy is 0.14, as seen in figure 17. The lowest being WILDSCamelyon with an average difference of around 0.025 and the highest being COVIDCT with approximately 0.23. Again, when looking at the methods, there is little correlation seen in the method averages, with all performing roughly equally to ERM, most of the variation can be explained by the dataset instead of the method chosen.

In all the above accuracy-based metrics, no method consistently outperforms any other method, which calls into question the efficacy of the newly developed methods and theories behind what enhances generalization capability. As such, further investigation is needed into what actually (if anything) separates these methods apart.

C. ANALYSIS OF DOMAIN-INVARIANCE

In order to discover differences between the chosen methods, we explore the invariance of representations created by each method. It is thought that methods that explicitly optimize for domain invariant representations (IRM, MMD, DANN, and SagNet) should have lower domain distinguishability than other methods, but this was not shown to be the case entirely. Overall, MMD, SagNet, and SelfReg showed noticeable and consistent improvement in domain invariance, with MMD being the most consistently improved. The average of all three domain invariance metrics can be seen in figure 9 which highlights this finding.

The first metric, d_0 as seen in figure 18, shows that in all methods on all datasets, the domain can be predicted from training domain representations to a high degree of accuracy. However, some methods do perform better than others. In the natural imaging datasets MMD, DANN, SagNet, and SelfReg all performed slightly better than ERM and the other methods; though still high enough that it cannot be said that they achieved domain invariance. In medical imaging though the story shifts. MMD, for instance, had high variability in how well it achieved domain invariance in the training data. Achieving much better domain invariant representations in OCT, MNM2, and COVIDCT, but then being on par with ERM in WILDSCamelyon. On the other side, DANN

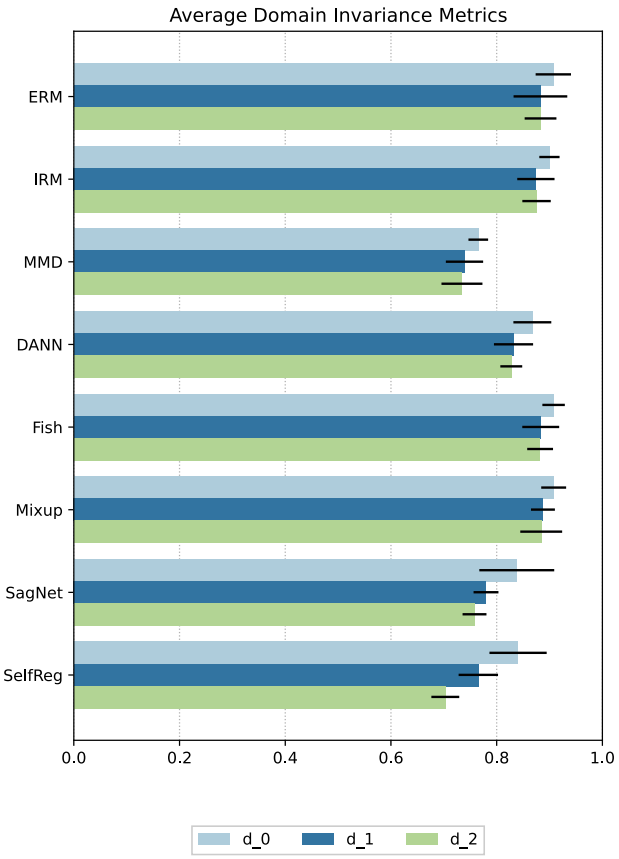

FIGURE 9. Each algorithms' d_0, d_1, and d_2 values averaged across all datasets.

performs on par with ERM in COVIDCT and MNM2, but lower in OCT and WILDSCamelyon. This trend continues in both d_1 and d_2, but all values are slightly lower, as we are including the test domain data as well in d_1, and then lower again when separating the training and calculation process by class in d_2.

In section II-B, we split up methods into those focused on domain invariance and those that aim for representational smoothness; specifically, IRM, DANN, MMD, and SagNet were the methods that explicitly trained for domain invariant representations. What we saw matched this expectation mostly. IRM and DANN only achieved marginally better domain invariance with an average of the three metrics, d_0, d_1, and d_2 of 0.88 and 0.84 respectively compared to ERM's 0.89. Whilst MMD, SagNet, and SelfReg all achieved much stronger domain invariance scores of 0.75, 0.79, and 0.77 respectively. These results also do not correlate at all with final OOD performance either as was expected to be the case.

D. ANALYSIS OF REPRESENTATIONAL SMOOTHNESS

In terms of representational smoothness: it was expected that the gradient matching, Mixup, and self-supervised contrastive regularization algorithms would produce the lowest feature variation, and this was seen in the results; though, as with domain invariant representations, to a lesser extent than was expected.

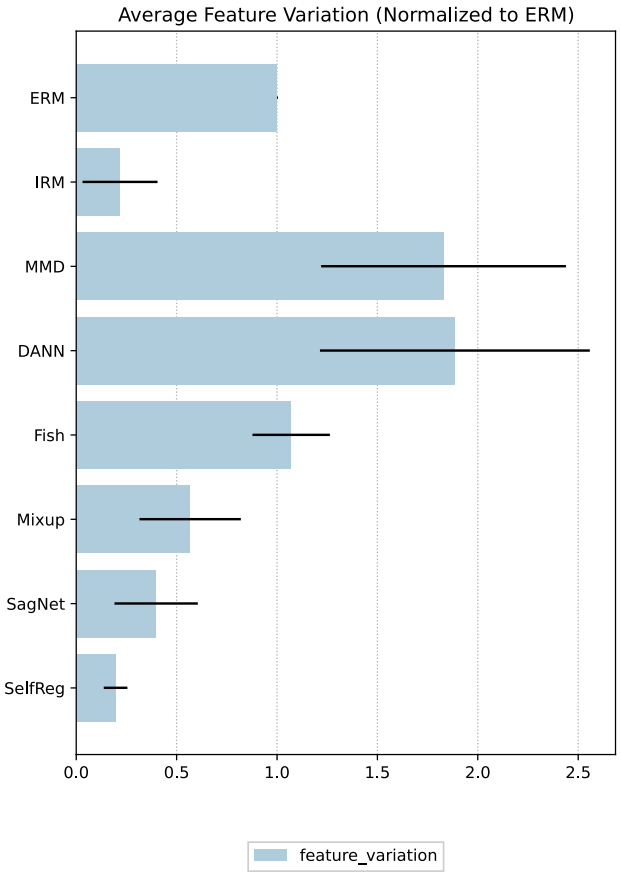

FIGURE 10. The average feature variation score (as calculated using equation 3) for each algorithm across all datasets normalized to ERM's feature variation to aid comparison.

Figure 10 shows large differences between the chosen algorithms. As our baseline method, ERM's normalized average feature sensitivity across all datasets was defined to be 1.0, and the other methods were normalized to be multiples of ERM by dividing their score by ERM's. The non-normalized values can be seen in figure 21 in the appendix. DANN and MMD have far higher feature representation sensitivities (at 1.83 and 1.88 times ERM's scores respectively). On the other side though, IRM and SelfReg have the lowest sensitivity on average (0.22 and 0.19 respectively) to minor changes in the input, followed closely by SagNet and Mixup at (0.40 and 0.57 respectively).

Overall these results follow as expected, Mixup and SelfReg obtaining lower feature variation than the baseline; however, IRM and SagNet, unexpectedly, also achieved lower feature variation, and FISH which was expected to lead to lower feature variation, had a higher variation than ERM. What was also not expected is that despite the high differences in representational smoothness, these differences have no obvious correlation with final OOD accuracy.

E. ANALYSIS OF DOMAIN DISTANCE

In this section we will be looking at the average cosine distances between the features from each domain as learnt by each of the methods. While not as useful as measuring the separability of the domains via a classifier, the distance can give some interesting insights into how the model is operating. For a truly domain invariant model, it would be expected that the average distance between domains would be low.

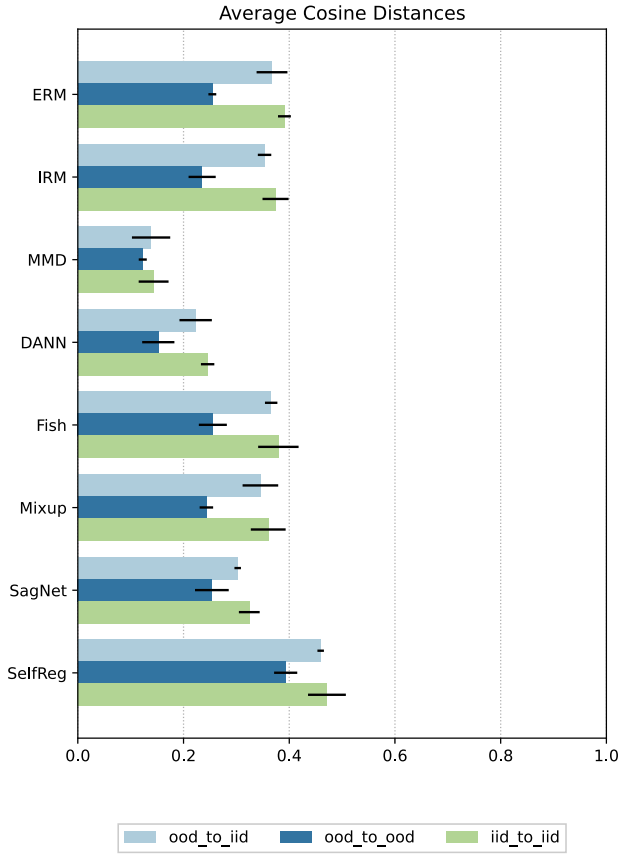

FIGURE 11. The average cosine distance between the pairings of OOD and IID domains. *"ood_to_iid"* shows the average cosine distance between features from an unseen OOD domain and each of the source domains. *"ood_to_ood"* shows how far are the features within the same OOD domain are spread apart. *"iid_to_iid"* shows how far apart each of the source domains are from each other.

When measuring the average cosine distance between each domain's features, it is possible to see some differences between how the methods operate in the feature space. Particularly, MMD and DANN both appear to place features closer together overall, whereas SelfReg appears to learn features that are further apart from each other.

In contrast to the classification-based domain invariance measures figure 5 shows how the methods organize their feature distributions. By both measures, MMD did achieve both the shortest distances between features and was somewhat inseparable; however, SelfReg also had separable features via the classification domain invariance measurement, but the distance between features is relatively large. This implies that the distance between features may not be a good measure of domain invariance. While the distance between each domain's features may be small, they still may be separable.

This analysis shows that these methods do operate and prioritize different things as their mechanisms would suggest. MMD appears to value features which are closer together in feature space; perhaps due to the distributional alignment mechanisms. But again, this has no real correlation to actual OOD performance questioning the efficacy of the algorithm's generalization mechanisms.

IV. DISCUSSION

A. DIFFERENCES IN ALGORITHM PERFORMANCE BETWEEN NATURAL IMAGING AND MEDICAL IMAGING TASKS

The results found through these experiments answers many questions, but also raises some more. From the outset, it is clear that medical imaging and natural imaging classification problems are similar enough in terms of performance, to the degree that if a method works well on natural image classification problems, unless the task is significantly more difficult, the method will also work well in medical imaging contexts. The consistent trend can be seen in all experiments carried out in this work: Accuracy, domain invariance, and feature sensitivity analysis all display a degree of consistency moving from natural to medical imaging problems.

However there are some differences to note. MMD is both more domain invariant and has lower feature variation in medical imaging than natural imaging. While DANN is has higher feature variation in natural images than in medical imaging. All other methods perform roughly equally between contexts. Additionally, the average feature distance is lower in medical imaging than natural imaging.

Our results show that the same trend that all tested domain generalization methods perform very similarly follows on medical imaging as well, reconfirming and extending what was discovered in the original DomainBed work, albeit on a slightly wider variance.

B. HOW DO THESE ALGORITHMS FAIL AT GENERALIZING?

While there are many observations one can make by analysing the results of the above failure modes framework, there are several that stand out: The impact of measuring overall domain invariance, and where specifically these algorithms fail.

The main purpose of this framework is to assess where focus needs to be placed in order to improve the performance for future applications. Overall, we see that the methods chosen all perform consistently on almost all metrics and the differences that are present are found mostly between datasets not algorithms. While there is no simple answer to why these methods aren't achieving stronger OOD accuracy, there are some clues. Firstly, metric e_0 (figure 12) shows some level of training set underfitting in VLCS, OfficeHome,

and the COVID CT datasets, all of which are the lowest performing datasets; so possibly training for longer periods of time to improve the fit to the training set could improve results. Secondly, there is only a small difference between the metrics e_1 (figure 13, test set separability) and e_3 (figure 15, classifier non-invariance) implying that the performance is being bottle-necked by the feature extractor not **extracting enough useful information for classifying the testing domains rather than a classifier that is over-fitted to the training domains**.

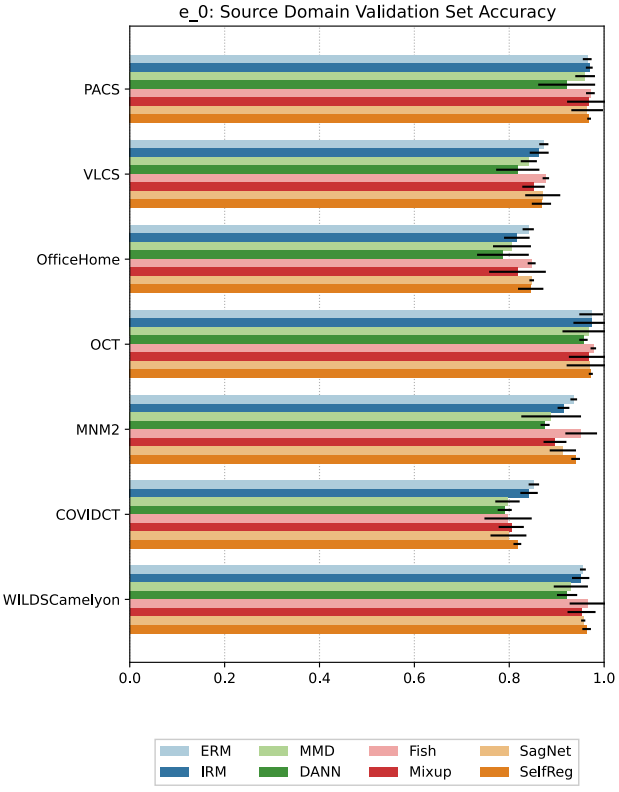

FIGURE 12. Failure mode metric e_0 (training set underfitting) for all algorithms and all datasets. The average classification accuracy of the whole model on unseen training domain samples.

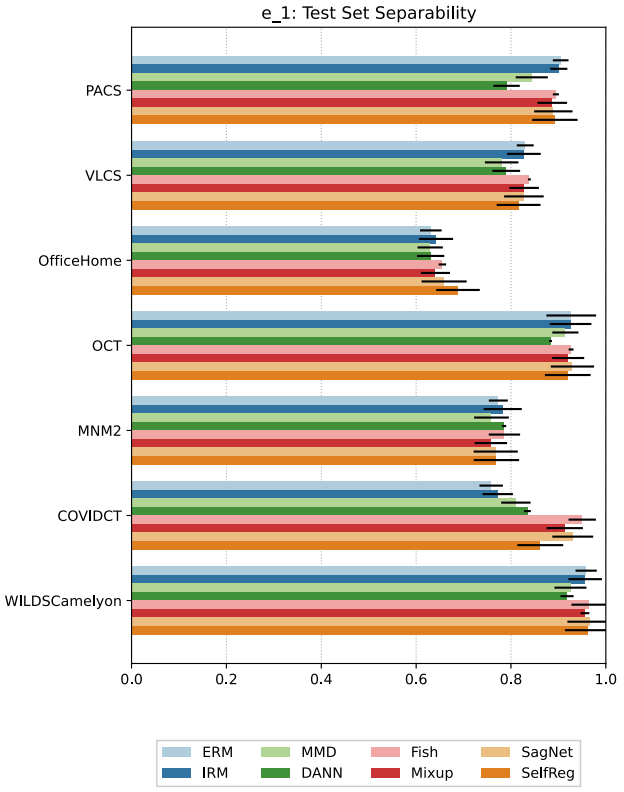

FIGURE 13. Failure mode metric e_1 (test set separability) for all datasets and methods. Average top-1 classification accuracy for a new classifier trained on a static feature extractor across all domains.

C. ADDRESSING THE POINTS OF FAILURE

Current research points to two main factors involved in creating generalizable model: Domain invariance, and representational smoothness. For domain invariance, the failure mode framework already has methods of quantifying the performance of algorithms in this respect, and with our additional metric to measure representational smoothness has allowed this work to understand the impact of both of these qualities in domain generalization.

In terms of domain invariance, we show that that all algorithms still retain a significant amount of domain specific information in their feature representations in both source and target domains. It would be expected that complete domain invariant representations would lead to a domain prediction accuracy of $1/n_d$, where n_d is the number of domains (for most datasets this would be $1/4 = 0.25$). However, even for the most domain invariant metric (as shown in figure 9, d_2, training-test class-conditional domain invariance), the average domain invariance accuracy across all datasets is 0.75 by MMD is much higher than the expected accuracy of 0.25 for a fully invariant representation.

This highlights a significant problem: Why are these methods not learning to be domain invariant? During the training process it is likely that domain information is exploited initially to achieve much stronger classification accuracy. Ideally, then a domain invariance based loss could prompt the network to remove domain-specific information. However, if the domain-specific information is closely entwined with classification-based information, it might not be possible to remove domain knowledge without severely harming the classification ability.

D. QUESTIONING THE EFFICACY OF DOMAIN GENERALIZATION MECHANISMS

Another question raised by these results is: Given all the methods perform relatively constantly across a dataset, and MMD, SagNet, and SelfReg have significantly greater domain invariance than other methods, what is the impact of domain invariant representations on final generalization performance? Although finding the exact answer to this question is outside the scope of this paper, there are some possible

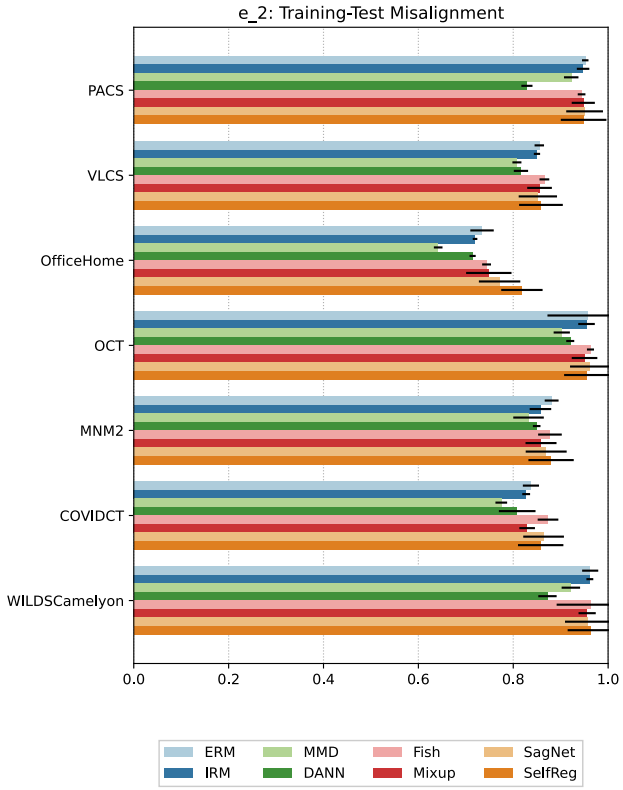

FIGURE 14. Failure mode metric e_2 (training-test misalignment) for all algorithms and datasets. Average classification accuracy of a new classifier on both training and unseen target domain features.

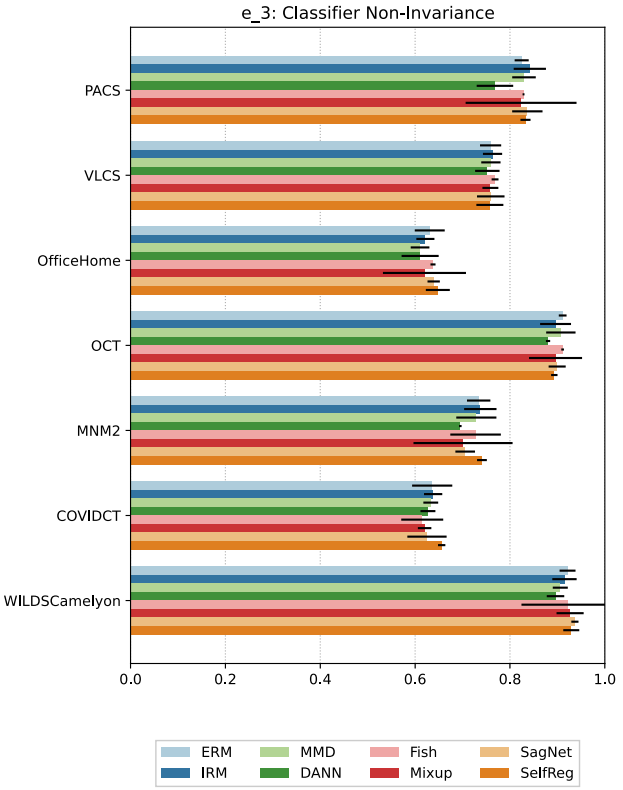

FIGURE 15. Failure mode metric e_3 (classifier non-invariance) for all algorithms and datasets. Average classification accuracy of the original model on unseen target domain data.

ideas to discuss. The large variation in domain invariance combined with minimal variation in accuracy points to complete domain invariant representations being less important than previously attributed. A large caveat to this however is the possibility that the features considered by a classifier to be important for classification are domain invariant, and all domain information is contained within features the classifier is ignoring. Likewise, a larger amount of domain invariance may need to be met before seeing significant improvements in accuracy as well, perhaps an entirely domain invariant feature extractor would have significantly improved accuracy. On the other side though, it is likely that learning to generate a wider variety of features to begin with will lead to a higher chance of discovering features that are class-discriminative on out-of-distribution data, as opposed to optimizing for domain invariance immediately, and possibly removing features that may operate well on test domains.

The results regarding representational smoothness are clearer than domain invariance given there is a much wider variety of values. As with domain invariance, there appears to be minimal correlation between smooth representation and final domain generalization performances. However, there still remains some cause for exploration. Firstly, the algorithm with the highest feature variation, DANN, had overall poorer training set fit (figure 12) and poorer final accuracy (figure 15), despite similar test set separability. Secondly, why does OfficeHome consistently generate models with higher feature variation, there may be a correlation between feature sensitivity and a low number of samples per class. As such, the benefit of representational smoothness remains under question. Theory proposes that smooth representations have benefits for generalization, so why is this not seen in practice? Unfortunately, we cannot propose answers as significantly more investigation must be undertaken. However a starting point would be to observe the differences between domains (as seen in Ye et al.'s work [45]) and how those differences, and how different the OOD domains are to the source domains, could all play a role in the importance of representational smoothness.

It was also found that there is likely a link between feature sensitivity and domain invariant representations, as the one method included in this analysis that utilizes both mechanisms, SelfReg, achieved both the most smooth representations as well as representations that were the most domain invariant of all tested methods. Again though, despite this, SelfReg still did not significantly improve the final OOD generalization performance over ERM.

V. CONCLUSION

Throughout this paper, we have explored the problem of domain generalization in several ways: Firstly to find the link

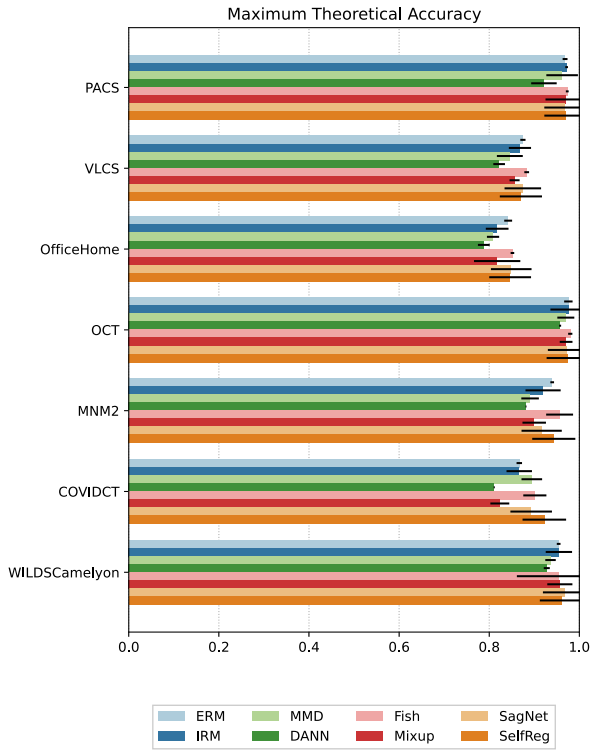

FIGURE 16. Average maximum theoretical accuracy for all algorithms and datasets. Highest IID accuracy achieved when training on each domain.

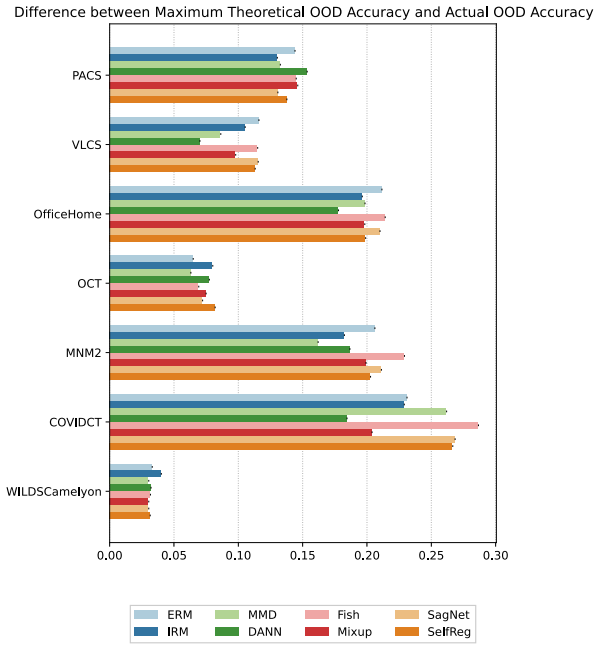

FIGURE 17. Difference between the maximum theoretical accuracy (IID validation accuracy) and the generalization accuracy (OOD validation accuracy) of the same domain.

between natural image classification problems and medical imaging problems, then continuing the exploration of why many novel algorithms fail to outperform the basic ERM

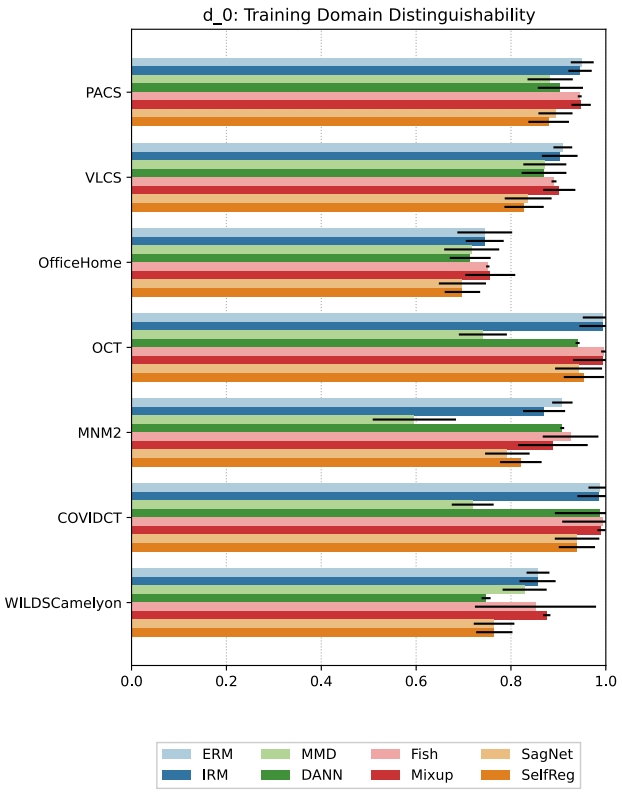

FIGURE 18. Failure mode metric d_0 (training domain distinguishability) for all datasets and methods. Average domain-weighted classification accuracy. Shows how accurately a new classifier trained on only source data can predict the domain of an unseen feature representation of training data features.

method through the lens of domain invariance and representational smoothness.

Through the results gathered of eight different domain generalization algorithms operating on seven different datasets (three natural imaging, and four medical imaging), it was confirmed that these algorithms do not lead to different generalization performance over the baseline algorithm (ERM), and similarly that performance of an algorithm on natural imaging datasets correlates with performance on medical imaging, despite the apparent differences in domain shifts. This would imply that methods designed for natural images ought to work on medical images as well, if not for the apparent lack of improvement in any method on any dataset over the baseline.

However, differences between methods were discovered when analyzing these models for domain invariant representations and representational smoothness. Given our understanding of how domain generalization should be achieved through domain invariant and smooth representations, when assessing models through these theories we should see a clear correlation between invariance, smoothness, and generalization capability, but this is not what we observed. Specifically, all methods created features that could lead to high domain prediction accuracies. Three methods (MMD, SagNet, and

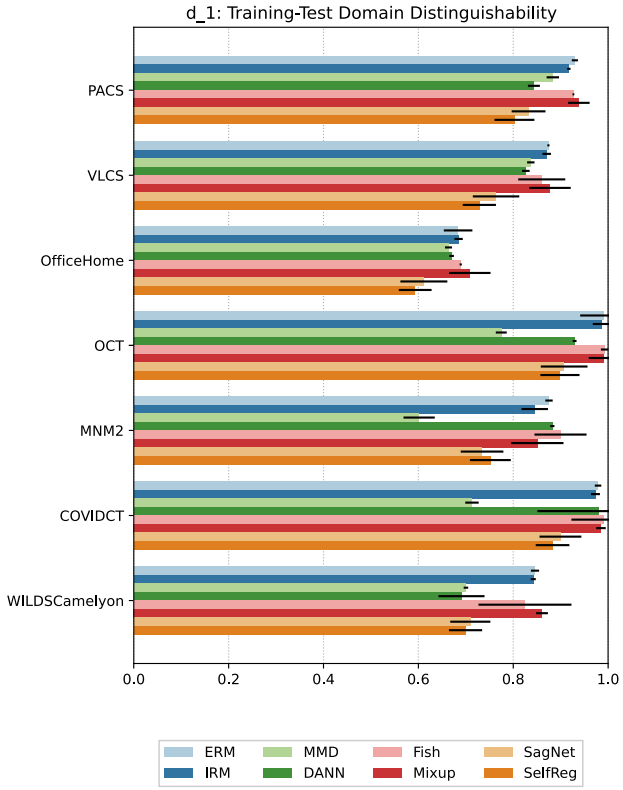

FIGURE 19. Failure mode metric d_1 (training-test domain distinguishability) for all datasets and methods. Average domain-weighted classification accuracy. Shows how accurately a new classifier trained on both source and target data can predict the domain of an unseen feature representation of features from both training and target domains.

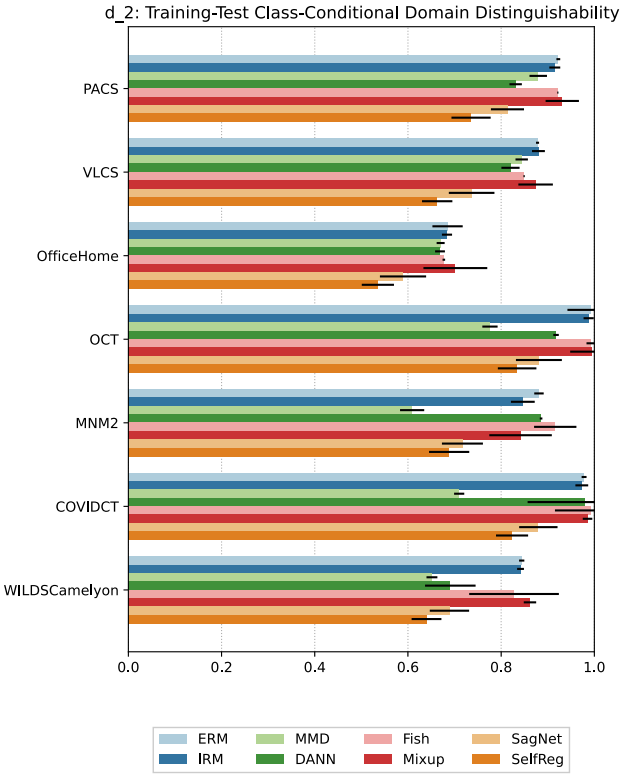

FIGURE 20. Failure mode metric d_2 (training-test class-conditional domain invariance) for all datasets and methods. Average domain-weighted classification accuracy. Shows how accurately a new classifier trained on samples from each class individually can predict the domain of an unseen feature representation of a sample from that class.

SelfReg) led to more domain invariant and more robust feature representations, but despite these differences there is minimal difference in final generalization performance. This calls into question the benefit of solely optimizing for domain invariance. In a similar vein, there is little impact of smooth representations on performance of these algorithms.

The findings of this work has a large implication for the use of deep learning domain generalization algorithms for medical imaging, specifically for clinical practice. If you encounter a problem with generalization in practice, none of the methods tested in this work would lead to a better outcome than the naive do-nothing approach (which is likely what was already used to discover the issue). This means that there is nothing to be gained from using domain-invariant-based or representational-smoothness-based domain generalization techniques, and in a clinical setting where gather more data from varied sources is an often insurmountable problem, the only option for improving results remains to use data-driven techniques such as handcrafted data augmentation methods.

Overall, this work shows there is little need for validating and designing medical imaging domain generalization methods separately from natural images. We also highlight two important gaps in the domain generalization literature: Firstly, if methods with stronger domain invariance do not perform better than methods with domain specific features then how important are domain invariant representations to domain generalization? Secondly, why do methods that aim to find domain invariant representations fail to do so? Both of which need to be answered if progress is to be made in the domain generalization field.

VI. FUTURE WORK

This work shows that in terms of medical imaging, there needs to be a deeper dive into the explicit differences between domains for each modality. The domain differences in CT and MRI scans appears to be more related to the image reconstruction process, which may alter fine details and textures structurally, as opposed to pure style. Additionally, there are many more medical imaging modalities, which requires an open-access dataset for verifying domain generalization algorithms, such as ultrasounds, PET scans, visible spectrum photographs (for identifying skin issues for example), and X-Rays. With the wider variety of datasets also leads to a wider variety of tasks, pure image classification is limited compared to the vast number of tasks medical practitioners require; image segmentation being a significant example. Possibly the largest factor that needs to be explored is the impact of these methods on 3D datasets. While this

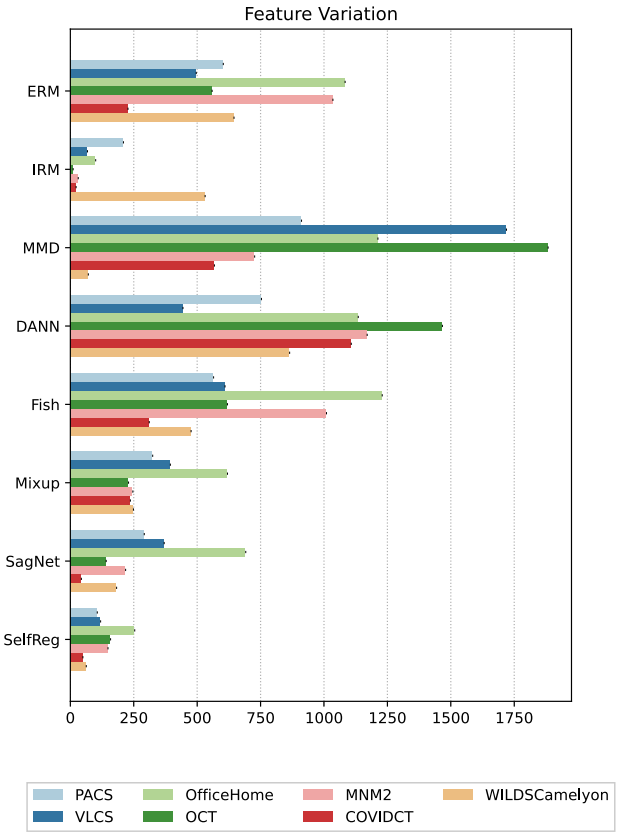

FIGURE 21. The average root mean squared deviation (as calculated using equation 3) on features from each method on each dataset.

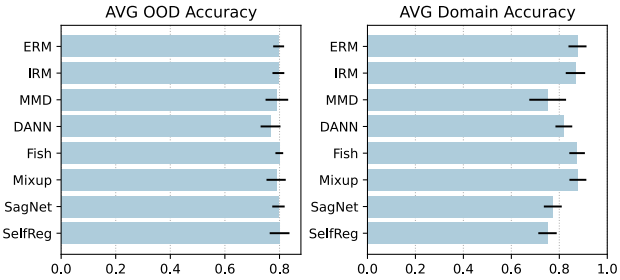

FIGURE 22. The average out-of-distribution accuracy and average domain classification across all datasets for all methods.

is not within the scope of this work (as a comparison of natural imaging to medical imaging using the same methods), a majority of important medical imaging modalities are natively 3D, as such understanding if algorithms require significant changes to adapt from 2D to 3D is necessary.

As stated in the previous section, this work has discovered that the most noticeable section of failure with current domain generalization algorithms appears to be the poor extraction of features that can be used to separate testing domain data, as opposed to ill-fitted classifiers. As such work could be focused on designing new methods which aim to generate a larger set of useful features. Likewise in terms of feature extraction, this work has raised questions regarding the importance of domain invariant features. As most methods appear to not be generating entirely domain invariant features to begin with, while those with fewer domain specific features perform equivalently regardless. Hence future work can be aimed at dissecting the importance of domain invariance in domain generalization applications.

Similarly, the purpose of this work is to prompt the creation of more analysis tools that can be used to inspect models to discover how and why they are or are not working. Current tools are limited in this aspect, as what differences between models that can be found using current understandings appear to be limited.

As stated in II-B there are more techniques which cannot be evaluated under the premise of domain invariant representations (such as those that leverage domain-specific components [25], [26], [27], [28]) and thus cannot be evaluated fairly under the framework proposed in this work. Future work should be performed in evaluating the generalization mechanisms of these methods, which may lead to insights into how important domain specific information is in the generalization process in contrast to this work's focus on domain invariant information.

APPENDIX A TRAINING HYPER-PARAMETERS

Shared hyper-parameters: Learning rate: 5e-5, dropout: 0.0, weight decay: 0.0.

IRM: IRM lambda: 100.0, IRM penalty anneal iterations: 500.

DANN: Discriminator steps per generator step: 1, discriminator gradient penalty: 0.0, lambda: 1.0, discriminator learning rate: 5e-5, generator learning rate: 5e-5, discriminator depth: 3, discriminator dropout: 0.0, discriminator width: 256, weight decay: 0.0, generator weight decay: 0.0.

MMD: MMD gamma: 1.0.

FISH: Meta learning rate: 0.5.

Mixup: Mixup alpha: 0.2.

SagNet: Adversarial loss weight: 0.1.

ERM and SelfReg have no method specific hyper-parameters.

APPENDIX B FAILURE MODE METRIC FIGURES

See Figures 16–22.

ACKNOWLEDGMENT

This research was supported by grants from NVIDIA and utilized an NVIDIA Quadro A6000 for running experiments.

REFERENCES

[1] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," Jul. 2020, *arXiv:2007.01434*.

[2] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Rev. Cancer*, vol. 18, pp. 500–510, May 2018.

[3] S. Wang, G. Cao, Y. Wang, S. Liao, Q. Wang, J. Shi, C. Li, and D. Shen, "Review and prospect: Artificial intelligence in advanced medical imaging," *Frontiers Radiol.*, vol. 1, Dec. 2021, Art. no. 781868.

[4] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[5] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," May 2022, *arXiv:2103.03097*.

[6] P. Li, D. Li, W. Li, S. Gong, Y. Fu, and T. M. Hospedales, "A simple feature augmentation for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8866–8875.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," May 2016, *arXiv:1505.07818*.

[8] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5400–5409.

[9] J. Lyu, Y. Zhang, Y. Huang, L. Lin, P. Cheng, and X. Tang, "AADG: Automatic augmentation for domain generalization on retinal image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3699–3711, Dec. 2022.

[10] C. Lin, Z. Yuan, S. Zhao, P. Sun, C. Wang, and J. Cai, "Domain-invariant disentangled network for generalizable object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 8751–8760.

[11] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," Mar. 2021, *arXiv:1910.11645*.

[12] B. Li, Y. Shen, Y. Wang, W. Zhu, C. Reed, D. Li, K. Keutzer, and H. Zhao, "Invariant information bottleneck for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 7399–7407.

[13] F. Zhou, Z. Jiang, C. Shui, B. Wang, and B. Chaib-Draa, "Domain generalization via optimal transport with metric similarity learning," *Neurocomputing*, vol. 456, pp. 469–480, Oct. 2021.

[14] P. Khandelwal and P. Yushkevich, "Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* (Lecture Notes in Computer Science), S. Albarqouni, S. Bakas, K. Kamnitsas, M. J. Cardoso, B. Landman, W. Li, F. Milletari, N. Rieke, H. Roth, D. Xu, and Z. Xu, Eds. Cham, Switzerland: Springer, 2020, pp. 73–84.

[15] Q. Liu, C. Chen, J. Qin, Q. Dou, and P. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 1013–1023.

[16] C. Li, X. Lin, Y. Mao, W. Lin, Q. Qi, X. Ding, Y. Huang, D. Liang, and Y. Yu, "Domain generalization on medical imaging classification using episodic training with task augmentation," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105144.

[17] I. Albuquerque, N. Naik, J. Li, N. Keskar, and R. Socher, "Improving out-of-distribution generalization via multi-task self-supervised pretraining," Mar. 2020, *arXiv:2003.13525*.

[18] S. Bucci, A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-supervised learning across domains," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5516–5528, Sep. 2022.

[19] D. Kim, S. Park, J. Kim, and J. Lee, "SelfReg: Self-supervised contrastive regularization for domain generalization," Apr. 2021, *arXiv:2104.09841*.

[20] Y. Shi, J. Seely, P. H. S. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," Jul. 2021, *arXiv:2104.09937*.

[21] S. Yang, K. Fu, X. Yang, Y. Lin, J. Zhang, and C. Peng, "Learning domain-invariant discriminative features for heterogeneous face recognition," *IEEE Access*, vol. 8, pp. 209790–209801, 2020.

[22] C. Shui, B. Wang, and C. Gagné, "On the benefits of representation regularization in invariance based domain generalization," May 2021, *arXiv:2105.14529*.

[23] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, "Improve unsupervised domain adaptation with mixup training," Jan. 2020, *arXiv:2001.00677*.

[24] S. Lee, Y. Lee, G. Lee, and S. Hwang, "Supervised contrastive embedding for medical image segmentation," *IEEE Access*, vol. 9, pp. 138403–138414, 2021.

[25] M. Mancini, S. R. Bulò, B. Caputo, and E. Ricci, "Best sources forward: Domain generalization through source-specific nets," in *Proc. 25th IEEE Int. Conf. Image*, Oct. 2018, pp. 1353–1357.

[26] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to balance specificity and invariance for in and out of domain generalization," Aug. 2020, *arXiv:2008.12839*.

[27] S. Seo, Y. Suh, D. Kim, G. Kim, J. Han, and B. Han, "Learning to optimize domain specific normalization for domain generalization," in *Computer Vision—ECCV 2020* (Lecture Notes in Computer Science), vol. 12367, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 68–83.

[28] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109115.

[29] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, and Z. Xu, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2531–2540, Jul. 2020.

[30] H. Li, Y. Wang, R. Wan, S. Wang, T.-Q. Li, and A. Kot, "Domain generalization for medical imaging classification with linear-dependency regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 3118–3129.

[31] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–12.

[32] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3D self-supervised methods for medical imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 18158–18172.

[33] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *PeerJ Comput. Sci.*, vol. 8, Jul. 2022, Art. no. e1045.

[34] C. Xing Tian, H. Li, Y. Wang, and S. Wang, "Privacy-preserving constrained domain generalization for medical image classification," May 2021, *arXiv:2105.08511*.

[35] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "DoFE: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4237–4248, Dec. 2020.

[36] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi, "An empirical framework for domain generalization in clinical settings," Apr. 2021, *arXiv:2103.11163*.

[37] O. Kilim, A. Olar, T. Joó, T. Palicz, P. Pollner, and I. Csabai, "Physical imaging parameter variation drives domain shift," *Sci. Rep.*, vol. 12, Dec. 2022, Art. no. 21302.

[38] J. Solomon, O. Christianson, and E. Samei, "Quantitative comparison of noise texture across CT scanners from different manufacturers," *Med. Phys.*, vol. 39, pp. 6048–6055, Oct. 2012.

[39] L. W. Goldman, "Principles of CT: Radiation dose and image quality," *J. Nucl. Med. Technol.*, vol. 35, no. 4, pp. 213–225, Dec. 2007.

[40] T. Galstyan, H. Harutyunyan, H. Khachatrian, G. V. Steeg, and A. Galstyan, "Failure modes of domain generalization algorithms," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19077–19086.

[41] Z. Liu, T. Lian, J. Farrell, and B. A. Wandell, "Neural network generalization: The impact of camera parameters," *IEEE Access*, vol. 8, pp. 10443–10454, 2020.

[42] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," Oct. 2017, *arXiv:1710.03077*.

[43] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1521–1528.

[44] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," Jun. 2017, *arXiv:1706.07522*.

[45] N. Ye, K. Li, H. Bai, R. Yu, L. Hong, F. Zhou, Z. Li, and J. Zhu, "OoD-Bench: Quantifying and understanding two dimensions of out-of-distribution generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7947–7958.

[46] P. Bándi et al., "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

[47] H. A. Alturkistani, F. M. Tashkandi, and Z. M. Mohammedsaleh, "Histological stains: A literature review and case study," *Global J. Health Sci.*, vol. 8, pp. 72–79, Mar. 2016.

[48] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstrom, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 325–336, Feb. 2021.

[49] H. Bogunović et al., "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1858–1874, Aug. 2019.

[50] J. Mazzarella and J. Cole, "The anatomy of an OCT scan," *Rev. Optometry*, vol. 152, no. 9, pp. 58–66, 2015. [Online]. Available: https://www.reviewofoptometry.com/article/the-anatomy-of-an-oct-scan

[51] V. M. Campello et al., "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3543–3554, Dec. 2021.

[52] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, "A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102588.

[53] E. B. Tsai et al., "The RSNA international COVID-19 open radiology database (RICORD)," *Radiology*, vol. 299, pp. E204–E213, Apr. 2021.

[54] E. Tsai et al., "Medical imaging data resource center—RSNA international COVID radiology database release 1a—Chest CT COVID+ (MIDRC-RICORD-1a)," Med. Imag. Data Resour. Center, USA, Tech. Rep., 2020. [Online]. Available: https://www.midrc.org/ and https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=80969742, doi: 10.7937/VTW4-X588.

[55] E. B. Tsai, "Medical imaging data resource center (MIDRC)—RSNA international COVID open research database (RICORD) release 1b—Chest CT COVID," Med. Imag. Data Resour. Center, USA, Tech. Rep., 2021. [Online]. Available: https://www.midrc.org/ and https://wiki.cancerimagingarchive.net/x/K4DTB, doi: 10.7937/31V8-4A40.

[56] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013.

[57] S. Morozov, A. Andreychenko, I. Blokhin, P. Gelezhe, A. Gonchar, A. Nikolaev, N. Pavlov, V. Chernina, and V. Gombolevskiy, "MosMedData: Data set of 1110 chest CT scans performed during the COVID-19 epidemic," *Digit. Diagnostics*, vol. 1, pp. 49–59, Jan. 2021.

[58] K. Zhang, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[59] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," Mar. 2020, *arXiv:1907.02893*.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

STEVEN KOREVAAR received the double degree (Hons.) in engineering (computer and network engineering) and computer science from the Royal Melbourne Institute of Technology, Australia, in 2019. During the degree, he worked on several research projects in the medical field which lead into a Doctorate degree upon graduation focused on investigating deep learning in medical imaging.

RUWAN TENNAKOON received the B.Sc. degree (Hons.) in electrical and electronics engineering from the University of Peradeniya, Sri Lanka, in 2007, and the Ph.D. degree in computer vision from the Swinburne University of Technology, Australia, in 2015. He was a Research Scientist with IBM Research, Australia. Since 2015, he has been a Research Fellow with the RMIT School of Engineering developing computer vision-based driver assist technologies for industrial vehicles. His research interests include computer vision, machine learning, and medical image analysis.

ALIREZA BAB-HADIASHAR (Senior Member, IEEE) received the B.Sc. degree from the University of Tehran, the M.Eng. degree from The University of Sydney, and the Ph.D. degree from Monash University, Australia. He has held various academic positions with Monash University, Swinburne University of Technology, and Copenhagen University. He is currently a Professor with RMIT University and leads the Intelligent Automation Research Group. He is an expert in the use of robust statistics methods and has a strong track record in developing robust vision-based industrial automation solutions.