Deep Learning When Data is Scarce

Ana Pimenta Alves
anapimentaalves@tecnico.ulisboa.pt
Instituto Superior Técnico, Lisboa, Portugal

***Abstract*—Scarce data has recently been a challenge for deep learning models that predominantly rely on extensive datasets to achieve state-of-the-art results, for example, in the fields of computer vision and Natural Language Processing (NLP). This is not only computationally demanding, but also inefficient compared to human learning. In areas like medical imaging, collecting extensive labeled datasets is also unfeasible, due to the time consuming and expensive data annotation process. Few-shot learning (FSL) addresses this challenge and enables models to classify unseen data with limited reference samples. This paradigm derives from meta-learning, where the idea is to "learn to learn", allowing models to extrapolate from a few examples. This thesis presents an approach that integrates Maximum Mean Discrepancy (MMD) regularization with few-shot classification of histology images. It aims to enhance model generalization from limited examples, effectively learning invariant features instead of domain or dataset specific characteristics. We compare performance across different datasets and integrate the regularization term during pre-training on the BreakHis dataset. For instance, applying MMD regularization to align the distributions of BreakHis and NCT improves the 10-shot test accuracy on CRC-TP from 79.4%, obtained without regularization, to 83.0%. This result becomes significantly closer to the one obtained when pre-training the model, without regularization, on a dataset 35 times larger.**

***Index Terms*—Few-shot Learning; Invariant Feature Learning; Convolutional Neural Network; Domain adaptation; Histology.**

I. Introduction

State-of-the-art deep learning models rely on large-scale training datasets, particularly in fields such as computer vision and natural language processing. For instance, many large language models illustrate this problem. GPT-3 [19] and Gopher [27], for example, are both trained on 300 billion tokens. When comparing machines and humans in this matter, humans simply don't need such large quantities of data to learn. Additionally, recent studies [17] show that these models can still be severely undertrained. This means that the trend to keep increasing the training data of models to consequently improve their performance is not biologically reasonable.

Challenges with large-scale data extend beyond just computational demands. In fields like medical imaging, the availability of labeled data is often constrained by privacy concerns. Moreover, creating such datasets necessitates the expertise of radiologists, which is both time-consuming and expensive.

We focus on few-shot classification of histology images, using four different datasets. Rather than pre-training on the largest dataset, we pre-train the model on a significantly smaller dataset, but incorporate a Maximum Mean Discrepancy (MMD) regularization term into the loss function.

The main objective is to analyse the potential of MMD regularization to promote general, invariant features rather than dataset-specific characteristics, potentially leading to models with better generalization capabilities, which is particularly useful in medical imaging.

II. Background

A. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [25] have become fundamental in computer vision due to their proficiency in tasks like image classification, segmentation, and object detection. Unlike methods that rely on manual feature extraction, CNNs autonomously extract vital features from images. This capability stems from their design to process data with grid-like topologies, such as images. Typically, a CNN architecture alternates between convolution, activation functions, and pooling layers and ends in one or more fully connected layers.

The convolution is the primary operation performed by a CNN. Mathematically, the convolution operation involves taking two functions and producing a third function that represents how one function modifies or slides over the other. In a CNN, the first function is the input, the second is the kernel and the output is referred to as the feature map [14].

During convolution, the kernel, typically initialized randomly and refined through training, moves over the image, performing element-wise product between the input pixels and the filter weights and then summing the results. This results in the feature map, a matrix whose values indicate the presence or absence of the feature that the kernel was designed to detect.

Stride specifies how much the convolution filter is moved at each step. A larger stride causes reduced overlap between receptive fields, influencing specific CNN features. Also, to retain feature map dimensions, padding is often employed. This ensures consistency in image dimensions after the convolution operation and prevents potential data loss at image edges.

For the CNN to be able to learn and model more complex relationships between the input and the output, beyond linear ones, the result of the convolution operation is passed through an activation function. This introduces non-linearity to the model, allowing it to learn and represent more complex functions.

For instance, a popular choice for an activation function is the Rectified Linear Unit (ReLU) [16] [24].It provides a very simple, but successful, non-linear transformation,

$$ReLU(x) = \max(0, x). \tag{1}$$

When the input is negative, the gradient is 0, while for positive inputs, the gradient is 1. This helps mitigate the vanishing gradient problem because it has a constant gradient for positive inputs [2] [21].

It is common to apply a pooling layer next for dimensionality reduction of the output feature map without the loss of important information, by aggregating results over a window of values. This helps with shortening training time, avoiding overfitting and reducing the number of parameters, resulting in an improved performance and efficiency [14] [37].

Pooling involves sliding a window over the input feature map to get a summary statistic of the nearby outputs. Two of the most common pooling functions are max pooling and average pooling. In max pooling, the output is determined by selecting the maximum pixel value from within the specified window size and stride. Average pooling, instead, calculates the average of the pixel values.

The stack of features maps that come from the pooling layer are first flattened into a single vector of data before being fed into one or more fully-connected layers.

Fully-connected layers connect every input neuron to every output neuron. These are responsible for determining the probability distribution over different classes. The last fully connected layer is the output layer and its number of neurons correspond to the number of classes. Activation functions, commonly softmax for multi-class classification, are used in this layer to achieve the final output of the network.

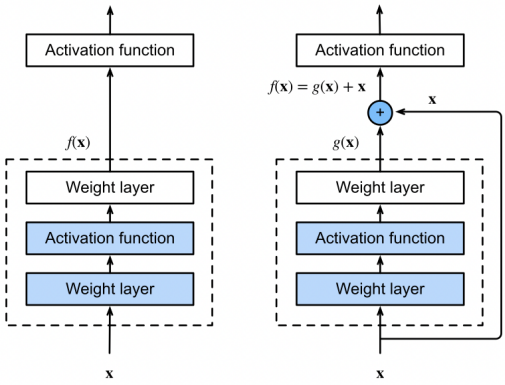

Fig. 1. Comparison between a traditional and a residual block [37]

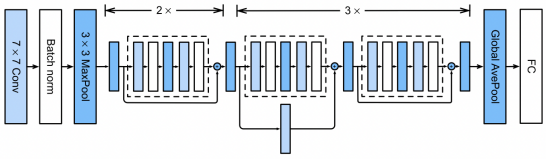

Fig. 2. ResNet-18 architecture [37]

B. Deep Learning Architectures

In 2012, a team from the University of Toronto, led by Alex Krizhevsky, won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [28] with an 8-layer CNN called AlexNet [21]. It significantly outperformed previous state-of-the-art methods and boosted the popularity of CNNs. Benefiting from the available ImageNet [10] dataset, released in 2009, along with an efficient GPU implementation, AlexNet achieved remarkable progress in deep learning. It also introduced the use of ReLU as the activation function, replacing the previously used sigmoid. It has since become a standard choice for many deep learning models [16].

Deeper CNNs face challenges like the vanishing gradient problem, that can compromise the network's learning capability. The Residual Neural Network (ResNet) [15], introduced in 2015, successfully addresses this issue.

The residual block, which incorporates skip connections is represented in Figure 1. These connections enable certain layers to receive not only outputs from the previous layer, but also outputs from earlier layers. This helps with gradient propagation during training and facilitates the learning of the identity mapping.

An illustration of the ResNet-18 architecture is provided in Figure 2, highlighting its sequence of residual blocks, global average pooling layer, and a final classification layer.

C. Scarce Data and Few-Shot Learning

Artificial Intelligence has outperformed humans in many tasks, such as the AlphaGo [30] that defeats (human) champions and the ResNet [15], that achieves higher performance than humans on ImageNet [10]. These successes, however, rely predominantly on large-scale datasets to learn and generalize effectively [35].

The challenge with the current data-driven approaches is that the required volume of data isn't always available. It is normally unreachable, for instance, in medical image datasets. Gathering data is feasible, since most patient diagnosis and care procedures involve medical image analysis. However, the labelling of medical images for research is an expensive and time-consuming process that requires the specialized knowledge of radiologists [35] and the use of specialized equipment.

Medical datasets, in addition to data scarcity, often face other challenges, such as varying data preparation methods and subtle cancerous tissue variations, which may cause labelling inconsistencies. Histology, for instance, involves the microscopic analysis of diseased tissues. Given the potential risks associated with misclassification, ensuring high accuracy in histology image classification is even more crucial than in other tasks. There is also a vast diversity in cancer-related tissue classification tasks not only across different cancer sites, but also within a single site. These challenges coupled with the risk of class imbalance in medical datasets, can lead to biased models that perform well for common conditions but may struggle with rare or atypical cases.

One topic where humans are still winning over AI is their ability to generalize from few examples, with little supervision [31].Humans have an innate ability to leverage what they have previously learned to rapidly master new tasks [22].

Few-shot Learning (FSL) addresses the challenges of needing large-scale data [36]. Specifically, few-shot classification involves classifying new examples based on a few reference samples of each class [11] [12]. It minimizes the necessity for vast data collection, leading to more efficient algorithms that can generalize their knowledge regardless of the amount of data available [36] [35].

FSL has been considered as an application of meta-learning

[33], which focuses on designing models that can adapt to new tasks with small data. The ultimate objective is not to learn about any training class in particular, but to "learn to learn". During training, the model experiences various tasks and learns to quickly adapt to new, unseen tasks. By adopting this strategy, the models can successfully classify new samples from few examples.

Let D_{base} be a large labeled dataset and Y_{base} its label space. Additionally, D_{test} refers to the target dataset with new classes Y_{test}, such that $Y_{base} \cap Y_{test} = \emptyset$. Formally, the problem can be described as using the knowledge acquired from the pre-training on D_{base} to learn from few unseen examples in D_{test}.

Normally, an episodic structure is followed after pre-training. In each episode, a N-way k-shot task is designed by sampling a subset of N classes from Y_{test} and selecting k examples from each class. This is the support set S and it has a total of $N \times k$ samples from D_{test} to help the model in quickly adapting to the novel classes. Additionally, the query set Q contains a separate unseen set of samples from the same N classes. The model's performance on Q gives an indication of its ability to generalize from the limited examples in S.

FSL minimizes the necessity for vast data collection, leading to more efficient algorithms that can generalize their knowledge regardless of the amount of data available [36] [35] and is of high importance in critical domains, like medical image classification.

III. Invariant Feature Learning

A. OOD Generalization

Empirical studies have shown that hidden stratifications and spurious associations found in real data, often cause machines to justifiably inherit these data biases. Rather than learning informative and robust features, deep learning models often rely on shortcuts to minimize training error, which makes them fail to generalize to Out-of-Distribution (OOD) data and misclassify samples with high confidence.

One study [4] showed an example where a husky was classified as a wolf because its picture was taken on a snowy background. Similarly, in a classification task between cows and camels [3], models wrongly classified cows on sandy beaches as camels, due to their reliance on background information and colors rather than on animal characteristics.

These results show that while such spurious correlations might work for some tasks, they can be dangerous in some domains, especially when models misclassify OOD samples confidently, such as in medical imaging. For instance, a recent study [9] underlined this idea by revealing that systems designed to detect COVID-19 in Chest X-Rays were not always relying on genuine medical indicators. Instead, these systems often detected non-relevant patterns, such as areas outside the lungs.

B. Causality

To address the OOD generalization failures caused by existing biases, confounders, and hidden stratifications within the training dataset, invariant risk minimization [3] and several other works were proposed [1] [8]. All of them rely on the principles of causality [26].

Causal inference has been discussed as a possible solution to improve the robustness of machine learning models to OOD data, ensuring successful and trustworthy results. For instance, the process of image annotation by radiologists, derived solely on visual assessment, is a causal task. For an anti-causal example, consider a skin lesion classification task where a set of dermoscopic images is used along with the biopsy results for melanoma [7].

In image classification, the relationship between the extracted features X from an input image I and the prediction Y, can be approached from two distinct directions:

- **Causal Task:**
  - Direction: $I \rightarrow Y$
  - Objective: Predict the effect (Y) from its cause (I).
- **Anti-Causal Task:**
  - Direction: $Y \rightarrow I$
  - Objective: Determine the cause (Y) from its effect (I).

C. Causal Task

Our work addresses classification of histology images. The annotation process of these images is done by pathologists relying on the assessment of the image's content, so the task can be considered causal.

Observed features X can be decomposed based on their relationships with environment E and label Y:

- $X_E^{\perp}$:
  - Features independent of E and that cause Y.
  - Relevant features for the classification task.
- $X_Y^{\perp}$:
  - Features caused by E and independent of Y .
  - Non-relevant features that might lead to shortcut learning.
- $X_{Y \wedge E}$:
  - Features influenced by E and that also cause Y.
  - A mix of both relevant and non-relevant features.

These causal connections are represented in the diagram of Figure 3. Solid arrows represent causal relationships and dashed lines denote non-causal ones. The diagram exemplifies the robustness of human annotators to OOD data, caused by their dependence on $X_E^{\perp}$ and $X_{Y \wedge E}$

D. Regularization for IFL

Invariant Feature Learning (IFL) aims to extract features that remain consistent across different data distributions, essential for enhancing model generalization in scenarios with potential domain shifts. Recent works [34] have suggested the minimization of distribution distances using the Maximum Mean Discrepancy (MMD), which measures the difference between two probability distributions by comparing their means in a feature space defined by a kernel. Minimizing this discrepancy helps reduce a model's dependence on dataset-specific,

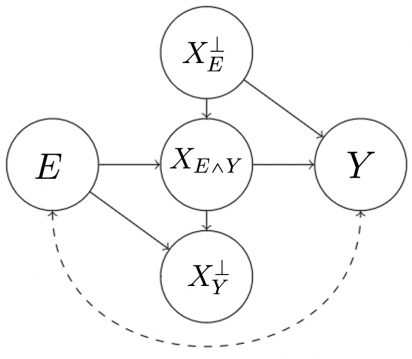

Fig. 3. Diagram of a causal task adapted from [34]

spurious features. The squared MMD [13], between samples X and X', is

$$\text{MMD}^2(X, X') = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) - \frac{1}{n'} \sum_{i=1}^{n'} \phi(x_i') \right\|^2, \quad (2)$$

with ϕ being a mapping to a Reproducing kernel Hilbert space (RKHS).

In the domain of IFL, the data representation function ϕ plays an important role. It is the function that maps the data into a feature space where invariant predictors across environments can be elicited. In the context of the network used in our model, a ResNet, ϕ is practically implemented as the high-level features extracted from the penultimate layer of the network, which is the last layer before the final classifier layer in the network architecture. This layer is chosen for its ability to capture high-level, abstract features of the input data. These features are central to the computation of MMD, which is utilized to measure the disparity in data distributions between two different sample sets. By focusing on these abstract features, MMD helps in assessing and minimizing domain-specific variances, steering the model towards learning invariant and robust representations.

The most common choice of kernel is the Gaussian (RBF) Kernel [13], given by

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \quad (3)$$

where σ is the bandwidth of the kernel.

A typical choice for σ is the median heuristic for balance and to avoid extreme cases [13], which is the median of all pairwise distances calculated between the points in the combined set of X and X'.

During the model training, in addition to the standard cross-entropy, an MMD term is integrated in the loss function,

$$Loss_{causal} = \mathcal{L} + \lambda \cdot MMD^2(X, X'), \quad (4)$$

where $\mathcal{L}$ is the standard cross-entropy loss for the task, λ is the regularization term that balances between the original task and domain adaptation, and $MMD^2(X, X')$ is the squared MMD.

As we backpropagate the MMD loss through the used network, a ResNet, the weights are updated in a way that encourages the penultimate layer to produce features that minimize the mean discrepancy between domains. This leads to a function ϕ, as described in equation 2, that is more adapted to extracting invariant features, which are less domain-specific, thereby enhancing the model's capacity to generalize across different domains. The overall loss, a combination of the standard task-specific loss, cross-entropy, and the MMD loss, becomes a more comprehensive measure of the model's performance. While the task-specific loss ensures that the model performs well on the main task, the MMD component of the loss ensures that the model does not overfit to domain-specific features of the training data, maintaining robust performance even on new, unseen data. Tuning the hyperparameter λ has a crucial role in model training. It balances the importance of the original task loss and the domain adaptation term. A higher value of λ forces the model to prioritize domain adaptation by reducing the difference between the distributions of the two sets. On the other side, with $\lambda = 0$, the model focuses solely on ERM, only computing the cross-entropy loss. An inadequate balance of the two tasks might lead the model to underperform: setting λ too high may cause it to underfit the primary task, while a value too low could compromise its ability to generalize effectively to the target domain. Previously, we have decomposed features into $X_E^{\perp}$, $X_Y^{\perp}$, and $X_{Y\wedge E}$. The new combined loss function $Loss_{causal}$ emphasizes the learning of features that are consistent across different environments, the causal features $X_E^{\perp}$ and $X_{Y\wedge E}$, as represented in Figure 3. By minimizing distributional differences, we ensure that the model isn't overly relying on features $X_Y^{\perp}$ that are environment-specific and might lead to shortcut learning. Overall, by focusing on these causal relations and minimizing reliance on environment-specific features, the model becomes better equipped to handle domain shifts, ensuring better robustness and generalization capabilities.

IV. Implementation

A. Datasets

This work's experiments were conducted on four datasets, following the FHIST benchmark [29]. They are briefly described below:

1) CRC-TP: The Colorectal Cancer Tissue Phenotyping (CRC-TP) dataset [18] focuses on tissue phenotyping in colorectal cancer, sourced from University Hospitals Coventry and Warwickshire (UHCW). It contains 280,000 non-overlapping patches from 20 Whole Slide Images (WSI) stained with Hematoxylin & Eosin (H&E). The patches, 150 × 150 pixels each, come from different patients and represent seven distinct phenotypes, each representing a different class.

2) NCT-CRC-HE-100K: NCT-CRC-HE-100K (NCT) [20] represents human colorectal cancer and normal tissue, with samples from the NCT Biobank (National Center for Tumor Diseases, Heidelberg, Germany) and the UMM pathology archive (University Medical Center Mannheim, Mannheim, Germany). It includes 100,000 non-overlapping image patches from 86 H&E stained slides, each 224x224 pixels. The images have been color-normalized and are classified into nine distinct tissue classes.

3) LC25000: The LC25000 dataset [5] showcases benign and malignant conditions of lung and colon tissues, collected from the James A. Haley Veterans' Hospital. From 750 lung tissue images and 500 colon tissue images, 25,000 images (768x768 pixels each) were created through augmentation. The dataset has five distinct classes.

4) BreakHis: The BreakHis dataset [32] originated from a study in Parana, Brazil and comprises 7,909 microscopic images from 82 patients, differentiated under four magnification factors considered as four super classes. Each super class contains eight sub classes representing four distinct types of benign breast tumors and four distinct types of malignant breast tumors.

B. Problem Setup

Our work follows a similar framework to FHIST benchmark [29]. The authors introduce three adaptation scenarios based on the domain shift from the CRC-TP base dataset: near-domain (NCT), middle-domain (LC25000) and out-domain (BreakHis), forming a cross-domain FSL approach, where domain shifts are all within histology images.

Alternatively to the FHIST approach, we analyse the performance of a model pre-trained on the BreakHis dataset using MMD regularization. Our work focused on their two best performing methods, TIM [6] and Finetune [23].

To ensure fair comparison, we mirror the methodologies of FHIST benchmark. The methods are pre-trained for three seeds up to 100,000 iterations and all experiments are performed on ResNet-18. Adam optimizer with an initial learning rate set to 5e-4, and cosine decay is also used. The data augmentations remain the same: random cropping, random flipping, and color jittering. TIM and Finetune undergo standard, non-episodic training with cross-entropy supervision. Training proceeds in batches of 100 samples.

In addition to the standard training procedures followed in our work, a crucial aspect of our experimental setup involves the use of MMD regularization. This is implemented by first extracting features from two separate batches of 100 samples each, one from the BreakHis dataset (source domain) and another from a selected target dataset (CRC-TP, NCT or LC25000). These features are normalized to ensure consistency in scale and distribution, enabling a meaningful comparison. The MMD loss is then computed between these normalized feature sets.

Every 1000 iterations the models are evaluated from 250 tasks randomly sampled in a 5-way 5-shot setting from the validation set. Meta-testing is done on 1000 randomly chosen tasks using standard 5-way tasks with 15 query samples per class, with 1-shot, 5-shot, and 10-shot settings and we report the average accuracy with 95% confidence intervals.

An example of the meta-testing in the NCT dataset is shown in Figure 4. The first row shows the support set made of 1 sample of each of the selected 5 classes (1-shot 5-way). The query set is made of 15 other samples of these classes, however, here, due to space constraints, we illustrate only part of the query set, made of 5 other samples of each class.

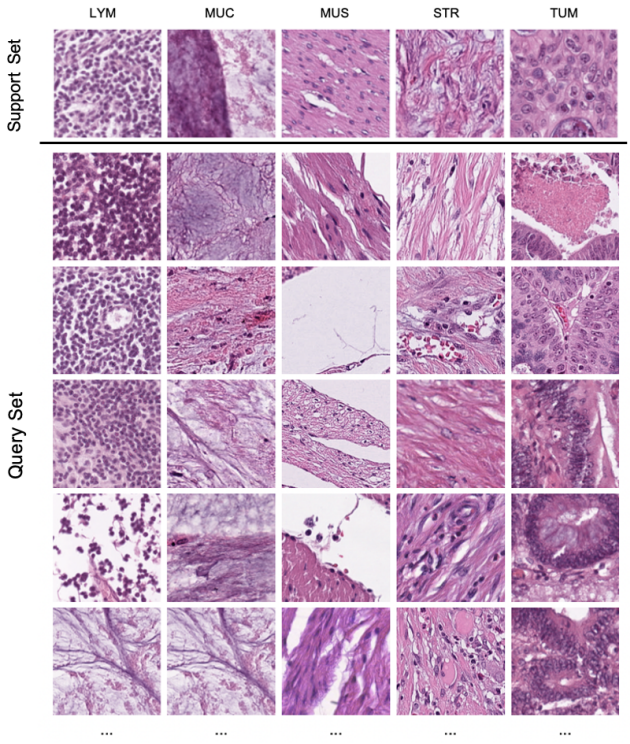

Fig. 4. Examples of a 1-shot 5-way testing in NCT dataset

C. Baseline Model

We start with a baseline model pre-trained on the BreakHis dataset that follows the previously outlined methodologies for training and meta-testing. It serves as a foundation point for evaluating the advantages of subsequent regularization techniques. An interesting comparison with prior results [29] emerges, since their base models are pre-trained on CRC-TP or NCT, datasets that are, approximately, 35 and 13 times smaller than BreakHis, respectively.

D. MMD Regularization Model

Our main interest is to correctly estimate the benefit of using MMD regularization, as explained in equation 2. For that, we incorporate it into our baseline model during the pre-training phase on the BreakHis dataset, where high-level features are extracted from the penultimate layer of our ResNet. This layer is adept at capturing complex, abstract representations of the input data, making it an ideal source for invariant feature extraction. The extracted features are then utilized in the computation of MMD, aligning the model's learning

process with the goal of invariant feature learning across diverse histology datasets.

Our experimental setup comprises three cases of MMD regularization. In each case, we pre-train the model on the Breakhis dataset with MMD regularization with respect to one of the remaining datasets (CRC-TP, NCT, LC25000). Once pre-trained, we follow the same training strategy described in the FHIST benchmark [29] and in our baseline model. We then evaluate the performance of each model on the remaining two datasets which haven't been seen during the regularization process. Training on a smaller and more challenging dataset, using MMD regularization with respect to another dataset, aims to extract invariant, general features from the data, rather than learning spurious dataset/domain specific features. When tested on new datasets, this model might exhibit enhanced robustness. Additionally, pre-training on BreakHis not only reduces computational costs, but also represents a more realistic scenario in medical imaging, where large labeled datasets might be scarce. Also, even though we use four labeled histology datasets, MMD compares the distributions of features from Breakhis and a selected dataset without relying on their labels. This implies that our approach can still be beneficial even if only one labeled dataset is available for training

V. Results

Each model is pre-trained on BreakHis using MMD regularization against one of the other datasets (CRC-TP, NCT, LC25000), followed by evaluations on the two remaining unseen datasets. Based on the dataset used for testing, we organize the results in three sections, CRC-TP, NCT and LC25000.

The first entry of each method corresponds to the results of baseline model pre-trained on BreakHis (without regularization). Following this, we integrate the MMD regularization between BreakHis and the other two datasets. Regularization strength λ from equation 4 is initially set to $\lambda = 1$. We report the average accuracy with 95% confidence intervals of TIM and Finetune methods. Meta-testing is performed as previously described.

A. Testing on CRC-TP

We evaluated models with MMD regularization on NCT and LC25000 and testing on CRC-TP. See Table I for details.

TABLE I
Test accuracy in the CRC-TP dataset with $\lambda = 1$.

BreakHis → CRC-TP				
Method	MMD	1-shot	5-shot	10-shot
TIM	–	37.3 ± 0.59	51.2 ± 0.49	56.8 ± 0.53
	NCT	$\mathbf{40.6 \pm 0.65}$	$\mathbf{54.3 \pm 0.5}$	$\mathbf{59.7 \pm 0.47}$
	LC25000	35.7 ± 0.59	50.1 ± 0.48	56.2 ± 0.48
Finetune	–	36.1 ± 0.52	49.9 ± 0.45	56.4 ± 0.44
	NCT	$\mathbf{36.6 \pm 0.50}$	$\mathbf{51.4 \pm 0.46}$	$\mathbf{57.4 \pm 0.45}$
	LC25000	33.8 ± 0.51	48.5 ± 0.46	55.2 ± 0.44

MMD regularization with the NCT dataset enhanced results for both methods across settings. This indicates NCT's potential in boosting model generalization. On the other hand, regularization with LC25000 yielded mixed results, suggesting either less meaningful invariant features compared to NCT or suboptimal generalization on CRC-TP.

B. Testing on NCT

We evaluated models with MMD regularization on CRC-TP and LC25000 and testing on NCT. See Table II for details.

TABLE II
Test accuracy in the NCT dataset with $\lambda = 1$.

BreakHis → NCT				
Method	MMD	1-shot	5-shot	10-shot
TIM	–	$\mathbf{67.0 \pm 0.85}$	$\mathbf{82.1 \pm 0.55}$	$\mathbf{86.4 \pm 0.44}$
	CRC-TP	63.1 ± 0.86	79.6 ± 0.57	85.1 ± 0.43
	LC25000	61.1 ± 0.87	78.3 ± 0.57	83.9 ± 0.46
Finetune	–	54.4 ± 0.69	73.3 ± 0.52	79.4 ± 0.47
	CRC-TP	$\mathbf{54.9 \pm 0.67}$	$\mathbf{76.6 \pm 0.50}$	$\mathbf{83.0 \pm 0.42}$
	LC25000	54.3 ± 0.69	75.2 ± 0.51	81.8 ± 0.44

For the TIM method without MMD , the 5-shot and 10-shot results nearly match the results proposed in FHIST [29], with pre-training on a dataset around 35 times larger than BreakHis. MMD regularization doesn't outperform the baseline, suggesting that invariant features from these regularizations don't effectively generalize to NCT.

The Finetune method experiences a boost in the 5-shot and 10-shot settings when applying MMD regularization with both datasets, but more visibly with CRC-TP. It suggests that some invariant features from the regularization to CRC-TP or LC25000 can be advantageous when adapting to NCT.

C. Testing on LC25000

We evaluated models with MMD regularization on CRC-TP and NCT and testing on LC25000. See Table III for details.

TABLE III
TEST ACCURACY IN THE LC25000 DATASET WITH $\lambda = 1$.

BreakHis → LC25000				
Method	MMD	1-shot	5-shot	10-shot
TIM	–	56.1 ± 0.63	74.8 ± 0.52	81.4 ± 0.48
	CRC-TP	54.1 ± 0.61	72.9 ± 0.45	80.9 ± 0.36
	NCT	$\mathbf{57.0 \pm 0.66}$	$\mathbf{76.5 \pm 0.44}$	$\mathbf{82.8 \pm 0.34}$
Finetune	–	$\mathbf{58.5 \pm 0.57}$	$\mathbf{75.7 \pm 0.35}$	$\mathbf{81.1 \pm 0.38}$
	CRC-TP	49.6 ± 0.47	71.3 ± 0.40	79.2 ± 0.33
	NCT	52.3 ± 0.52	73.7 ± 0.39	80.8 ± 0.32

TIM, when regularized with NCT, shows improved results, suggesting that NCT's domain knowledge contributes with invariant features for LC25000 testing. The 10-shot result is not so distant to the FHIST outcome considering their pre-training in the NCT dataset, which is a considerably larger dataset than BreakHis.

The Finetune baseline achieves impressive 5-shot and 10-shot results, surpassing the model in FHIST pre-trained on NCT. However, in this context, the MMD regularization doesn't enhance performance.

Overall, in certain cases, MMD has significantly enhanced the model's generalization across datasets, particularly when NCT is utilized for regularization during CRC-TP testing. However, in other scenarios it aligns or comes slightly below the baseline model in performance, emphasizing the need for dataset selection in regularization.

The achieved results in few-shot classification of histology images with pre-training on a much smaller dataset than previously considered, show that BreakHis probably relies on more invariant features that enhance generalization, rather than domain-specific ones.

Among the three cases, using NCT for MMD regularization consistently shows better results than the baseline, except for one case. This suggests that aligning the feature distributions of BreakHis and NCT yields more beneficial representations for the tasks at hand compared to when using LC25000 or CRC-TP for regularization

The results also showcase the potential of domain adaptation in medical imaging. The ability to leverage knowledge from one domain to enhance performance on a different, yet related domain is a good approach to tackle the data scarcity problem, while still achieving robust and efficient models.

D. Higher regularization strength

A regularization strength $\lambda = 10$ was tested on the scenarios that benefited from MMD regularization before, to determine whether it would further improve the model performance.

TABLE IV
TEST ACCURACY IN THE CRC-TP DATASET WITH $\lambda = 10$.

BreakHis → CRC-TP				
Method	MMD	1-shot	5-shot	10-shot
TIM	NCT	34.5 ± 0.48	50.1 ± 0.44	52.7 ± 0.44
Finetune	NCT	32.8 ± 0.46	44.9 ± 0.43	48.7 ± 0.43

TABLE V
TEST ACCURACY IN THE NCT DATASET WITH $\lambda = 10$.

BreakHis → NCT				
Method	MMD	1-shot	5-shot	10-shot
Finetune	CRC-TP	44.2 ± 0.62	58.7 ± 0.55	62.5 ± 0.54

TABLE VI
TEST ACCURACY IN THE LC25000 DATASET WITH $\lambda = 10$.

BreakHis → LC25000				
Method	MMD	1-shot	5-shot	10-shot
TIM	NCT	52.7 ± 0.49	66.0 ± 0.44	70.0 ± 0.44

Using the regularization strength with a value of 10.0 did not produce better results that when compared to using a value of 1.0. Tuning λ can be a difficult task. A large choice, like $\lambda = 10$, will prioritize the MMD regularization term at the expense of the primary classification task, which can decrease performance. However, in section VI, we discuss how more values of λ should be investigated as future work.

VI. CONCLUSIONS AND FUTURE WORK

Drawing inspiration from the inherent learning efficiency of humans, we proposed to address the challenge of data scarcity in histology datasets. We focused on cross-domain few-shot classification and tried to encourage the learning of more general and invariant features, through MMD regularization.

The results of this study challenge traditional approaches of large-scale data training, by still achieving good performances with models trained on a significantly smaller dataset. Although not all experiments benefit from MMD regularization in our tests, certain scenarios demonstrated that incorporating such regularization can indeed enhance generalization and promote IFL, which appears to be the right direction for handling data scarcity.

Regarding future improvements, the first limitation regards the regularization strengths λ from equation 4. In this work, only two values were analysed. As future work, the testing of additional values of λ could lead to interesting results. A range of values should be tested, including values below 1 and between the already tested 1 and 10. Also, a more complex approach could be considered. We could start with a large value to force the model to prioritize the regularization term

initially and then decrease it over time, allowing the model to still perform well in the main classification task. Or we could try the opposite, allow the model to achieve good classification performances with a small λ first and then increase it.

The regularization using MMD implies the selection of a kernel function and corresponding parameters. In this work, we only considered a Gaussian kernel with bandwith set to the median heuristic. Tuning parameters or testing additional kernels remains as future work, as well as experimenting with other metrics for regularization.

REFERENCES

[1] K. Ahuja, E. Caballero, D. Zhang, Y. Bengio, I. Mitliagkas, and I. Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Neural Information Processing Systems*, 2021.

[2] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, 2017.

[3] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

[4] P. Besse, C. Castets-Renard, A. Garivier, and L. J-M. L'ia du quotidien peut-elle être éthique? loyauté des algorithmes d'apprentissage automatique. *en ligne], https://hal. archives-ouvertes. fr/hal-01886699v2 (consulté le 9 septembre 2019)*, 2018.

[5] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[6] M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020.

[7] D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[8] M. Chevalley, C. Bunne, A. Krause, and S. Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

[9] A. DeGrave, J. Janizek, and S. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nat Mach Intell*, 3:610–619, 2021.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28:594–611, 05 2006.

[12] M. Fink. Object classification from a single example utilizing class relevance metrics. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, 12 2004.

[13] D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

[14] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[17] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.

[18] S. Javed, A. Mahmood, N. Werghi, K. Benes, and N. Rajpoot. Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping. *IEEE Transactions on Image Processing*, 29:9204–9219, 2020.

[19] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[20] J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018.

[21] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[22] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum. One shot learning of simple visual concepts. In L. Carlson, C. Hoelscher, and T. F. Shipley, editors, *Expanding the Space of Cognitive Science - Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, CogSci 2011*, Expanding the Space of Cognitive Science - Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, CogSci 2011, pages 2568–2573, Boston, United States, 2011. The Cognitive Science Society.

[23] X. Luo, H. Wu, J. Zhang, L. Gao, J. Xu, and J. Song. A closer look at few-shot classification again. *arXiv preprint arXiv:2301.12246*, 2023.

[24] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines vinod nair. In *International Conference on Machine Learning*, volume 27, pages 807–814, 06 2010.

[25] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.

[26] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:947–1012, Oct 2016.

[27] J. W. Rae, S. Borgeaud, T. Cai, and M. et al. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[29] F. Shakeri, M. Boudiaf, S. Mohammadi, I. Sheth, M. Havaei, I. B. Ayed, and S. E. Kahou. Fhist: A benchmark for few-shot classification of histological images. *ArXiv*, abs/2206.00092, 2022.

[30] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan. 2016.

[31] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017.

[32] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[33] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, and H. Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. *CoRR*, abs/1903.03096, 2019.

[34] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

[35] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016.

[36] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *CoRR*, abs/1904.05046, 2019.

[37] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. Dive into deep learning, 2023.