
Learning Invariant Representations with a Nonparametric Nadaraya-Watson Head

Alan Q. Wang
Cornell University

Minh Nguyen
Cornell University

Mert R. Sabuncu
Cornell University

Abstract

Machine learning models will often fail when deployed in an environment with a data distribution that is different than the training distribution. When multiple environments are available during training, many methods exist that learn representations which are invariant across the different distributions, with the hope that these representations will be transportable to unseen domains. In this work, we present a nonparametric strategy for learning invariant representations based on the recently-proposed Nadaraya-Watson (NW) head. The NW head makes a prediction by comparing the learned representations of the query to the elements of a support set that consists of labeled data. We demonstrate that by manipulating the support set, one can encode different causal assumptions. In particular, restricting the support set to a single environment encourages the model to learn invariant features that do not depend on the environment. We present a causally-motivated setup for our modeling and training strategy and validate on three challenging real-world domain generalization tasks in computer vision.

1 Introduction

Machine learning models often fail when there is significant distribution shift. The goal of domain generalization is to be able to perform well with new distributions [21, 56, 71]. In this work, we are interested in settings where multiple domains/environments are available during training and we have access to environment indicators. A popular way to tackle domain generalization in this setting is to learn representations that are invariant across environments [20, 41, 60]. The hope is that such representations will work well in, or are transportable to, unseen environments. This invariance is often encoded via constraints on a learned predictor which aligns its behavior across environments; often, these conditions are derived using causal reasoning and/or by making assumptions about the data-generating process [39].

In a parametric setting, almost all existing methods enforce these constraints by training a single model and adding a regularizer on top of a standard predictive loss [1, 5, 18, 20, 27, 48, 54, 58]. Most notably, invariant risk minimization (IRM) enforces the representations to be such that the optimal classifier on top of those representations is the same across all environments. Other examples include enforcing the layer activations of the predictor to be aligned across environments [48], enforcing the predictor to be calibrated across environments [54], and enforcing gradients of the predictor to be aligned across environments [46]. Often, optimizing these constraints demand approximations or relaxations that undermine the efficacy of the approach [22].

In this work, we take a different approach using a nonparametric strategy based on a recently-proposed Nadaraya-Watson (NW) head [55]. Instead of computing the class probability directly from an input query, the NW head makes a prediction by comparing the learned representations of the query to the elements of a support set that consists of labeled data. Thus, the NW prediction is computed *relative to other real datapoints* in the support set, with the support set providing a degree of flexibility not

possible with parametric models. In particular, one can manipulate it during training in a way which restricts the types of comparisons that the model can make.

In this work, we manipulate the support set during training to encode causal assumptions for the purposes of learning invariant representations. Specifically, restricting the support set to be drawn from a single environment precludes the possibility of using environment-specific features to make a prediction for a given query. We show that this setup is causally-motivated and relates to existing causal frameworks. Furthermore, we show that this training strategy leads to competitive to superior results compared to state-of-the-art parametric baselines.

Our contributions are as follows:

- We present causally-motivated assumptions for domain generalization which justify our modeling and training strategy.
- We present a novel approach to invariant representation learning using the nonparametric Nadaraya-Watson head, which can account for causal assumptions by manipulating a support set. In particular, we propose a training strategy which, unlike competing baselines, has invariance hyperparameter to tune
- We validate our approach on several datasets and demonstrate competitive results compared to state-of-the-art parametric baselines.

2 Related Works

2.1 Domain Generalization and Invariant Representations

Domain generalization seeks to make models robust to unseen environments and is an active area of research [1, 56, 71]. One line of work augments or synthetically-generates additional training images to increase robustness of learned features to unseen environments [65, 70]. In particular, LISA uses a mixup-style [67] interpolation technique to generate augmented images, which the authors demonstrate improves out-of-distribution robustness. Another line of work broadly seeks to align features across distributions. Deep CORAL aligns correlations of layer activations in deep neural networks [48], and other works minimize the divergence of feature distributions with different distance metrics such as maximum mean discrepancy [51, 32], an adversarial loss [4, 30], and Wasserstein distance [69]. Still other works approach the problem from the perspective of the gradients and optimization [2, 29, 34, 46, 57]. For example, Fish aligns the gradients from different domains [46].

One can also achieve domain generalization via learning invariant representations, which often requires reasoning about the data-generating process from a causal perspective to arrive at appropriate constraints [39]. Invariant causal prediction (ICP) formulates the problem from a feature selection perspective, where the goal is to select the features which are direct causal parents of the label [40]. Invariant Risk Minimization (IRM) can be viewed as an extension of ICP designed for deep, nonlinear neural networks. The IRM objective can be summarized as finding the representation such that the optimal linear classifier's parameters on top of this representation is the same across all environments [1]. This bi-level program is highly non-convex and difficult to solve. To find an approximate solution, the authors consider a Lagrangian form, whereby the sub-optimality with respect to the constraint is expressed as the squared norm of each of the inner optimization problems. Follow-up works analyzing IRM have raised theoretical issues with this objective and presented some practical concerns [16, 22, 42]. Various flavors of IRM have also been proposed by introducing different regularization terms [27, 54, 58].

2.2 Nonparametric Deep Learning

Nonparametric models in deep learning have received much attention in previous work. Deep Gaussian Processes [50], Deep Kernel Learning [62], and Neural Processes [24] build upon Gaussian Processes and extend them to representation learning. Other works have generalized to nearest neighbors [37, 49], decision trees [68], density estimation [13], and more general kernel-based methods [5, 36, 66] to deep networks and have explored the interpretability that these frameworks provide. Closely-related but orthogonal to nonparametric models are attention-based models, most notably self-attention mechanisms popularized in Transformer-based architectures in natural language

Figure 1: Illustration of proposed approach. Support set of labeled datapoints (square/triangle) from 3 environments lie in 3 regions in the feature space. Black circle denotes query datapoint with unknown label. a) The NW head models $P(Y|X)$ by making predictions as a function of distances to labeled datapoints in the feature space (visualized as dotted arrows). b) Balancing comparisons across labels for all environments models $P^B(Y|X)$. c) Conditioning on a single environment models $P(E|Y|X)$.

processing [53] and, more recently, computer vision [1, 19, 38]. Nonparametric transformers apply attention in a nonparametric setting [24].

Recently, Wang et al. proposed the NW head [55], an extension of the classical NW model [35, 61] to deep learning. In the NW head, the prediction is a weighted average of labels from a support set. The weights are computed from distances between the query and support features. The NW head can yield better calibration and interpretability, with similar accuracy compared to the dominant approach of using a parametric classifier with fully-connected layers. In this work, we leverage the NW head to encode causal assumptions via the support set. The interpretability and explainability benefits of the NW head carry over in this work; while not of primary focus, we explore these properties in the Appendix.

3 Preliminaries

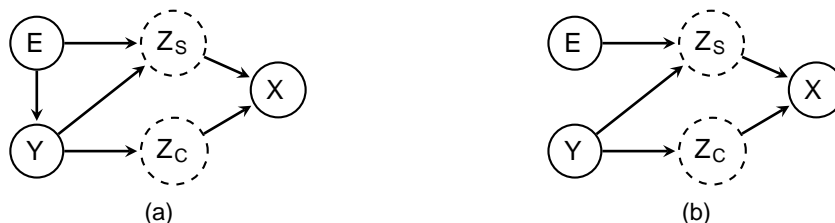


Figure 2: a) Causal Directed Acyclic Graph (DAG) we consider in this work. Solid nodes are observed and dashed nodes are unobserved. We assume an anti-causal setting where X is calculated by Z_s and Z_c , and X has 2 causal parents: “style” features Z_s , which are influenced by the environment E and environment-independent “content” features Z_c , which are causally influenced by the label Y . Both E and Y have direct influence on style features Z_s . b) Same DAG as a) with an intervention on Y . We note that $Y \perp\!\!\!\perp E \mid Z_c$ and $Y \perp\!\!\!\perp E \mid Z_s$.

Problem Setting. Let $X; Y$ denote a datapoint and its corresponding discrete class, and denote the environment (or domain) where $X; Y$ originates¹. That is, the elements of the training dataset $D_{tr} = \{x_i; y_i; e_i\}_{i=1}^N$ are drawn first by sampling the discrete random variable $P(E)$, and then sampling $x_i; y_i \sim P(X; Y \mid E = e_i) := P_{e_i}(X; Y)$. Our goal is to learn classifiers that will generalize to new, unseen environments.

Assumptions. We assume there exists a pair of latent causal parents: an environment-independent (“content”) factor Z_c and an environment-dependent (“style”) factor Z_s .² We as-

¹We assume the support of E , $\text{supp}(E)$, is finite.

²This is sometimes referred to as style-content or causal decomposition [21, 42, 60].

sume the causal mechanism that generates (Z_C, Z_S) is injective, so that, in principle, it is possible to recover the latent features from the observations; i.e. there exists a function that $g(X) = (Z_C; Z_S)$. We further assume that g can be disentangled into g_C and g_S , such that $(Z_C; Z_S) = g(X) = (g_C(X); g_S(X))$. The causal graph is depicted in Fig. 2a. Finally, we assume that if any $X = x$ has a non-zero probability in one environment, it has a non-zero probability in all environments.

Motivation. The motivating application in this work is image classification, where each realization of E might represent a different site, imaging device, or geographical region where they are collected. For example, in medical imaging, different hospitals (may collect images X) attempting to capture the presence of some disease Y , but may differ in their imaging protocols which lead to differences in the image style features Z_S (e.g. staining, markings, orientation). In addition, we allow Z_S to be influenced by the label itself (for example, positive examples are more likely to be marked by a doctor or have specific staining than negative examples). Finally, the prevalence may be influenced by E (for example, the prevalence of a disease may be higher in a certain hospital).

The goal is to find an estimator for Y which relies only on the direct causal links between E and X and not on any spurious associations between E and X , as these may change in a new, unseen environment. That is, we seek an estimator which relies only on Z_C and which is independent of Z_S .

First, we note the direct causal dependence $E \rightarrow Y$. For example, a model can exploit this association by learning to over-predict majority classes in a certain environment. One way to remove the direct dependence is by intervening on Y , thus removing incoming edges to Y . This essentially corresponds to matching the environment-specific prior $P(Y)$ between environments, and results in the intervened graph in Fig. 2b. Let us refer to any distribution which follows the DAG in Fig. 2b as $P_e^B(X; Y)$.

Second, we observe that there is a potential non-causal association between Y through the colliders X and Z_S , when either one of these are conditioned on (i.e. are observed). An estimator which relies on Z_S potentially leaks information from E , and this is unstable in new environments. Reading d-separation on this intervened graph, we infer that $E \perp Z_C = g_C(X)$ and $Y \not\perp E \mid Z_S = g_S(X)$, that is:

$$P_e^B(Y \mid g_C(X)) = P_{e_0}^B(Y \mid g_C(X)) \quad \forall e; e^0 \in E: \quad (1)$$

In words, this assumption states that the probability of Y given the environment-invariant parts of X is the same across any environment E .

Thus, we seek an estimator that 1) enables interventions such that the direct dependence $E \rightarrow Y$ can be removed, and 2) can further encode the assumption in Eq. (1).

4 NW Head for Invariant Prediction

Given a datapoint x , support set $S = \{x_i; y_i\}_{i=1}^{N_s}$, and parameters, the NW head estimates $P(Y = y \mid X = x)$ by outputting a prediction formulated as a weighted sum of support set labels, where the weights are some notion of similarity in the feature space [55]:

$$\hat{P}(Y = y \mid X = x; S) := f(x; S) = \frac{\sum_{i=1}^{N_s} \exp(s(x; x_i)) g_i}{\sum_{j=1}^{N_s} \exp(s(x; x_j)) g_j}. \quad (2)$$

Here, y is the one-hot encoded version of Y and $s(\cdot; \cdot)$ is a similarity/kernel function that captures the similarity between pairs of features. In this work, we use the negative Euclidean distance. A graphical depiction is shown in Fig. 3.

Manipulating the support set can encode certain causal assumptions. We consider two types of manipulations:

1. Balancing classes is denoted S^B (see Fig. 1b). This can be interpreted as an intervention on Y , and removes the dependence $E \rightarrow Y$, i.e.:

$$\hat{P}^B(Y = y \mid X = x; S) := f(x; S^B) \quad (3)$$

³This may be interpreted as making the model robust to label shift, see [31, 44].

Figure 3: A depiction of the NW head on a tumor detection task. The NW head computes Euclidean distances $\|x_i - x\|$ between query and support features, and uses the distances to weight the support labels. Colored squares represent labels. Diagram displays two different support sets. Top is unconditional support, where support data is drawn from the training data without knowledge of environment information. Bottom is an example of a manipulated support where all support data is drawn from a fixed environment (note similarity in color). Such a support set precludes the possibility of using environment-specific features to make a prediction.

- Conditioning S on a single environment, denoted S_e (see Fig. 1c). This can be interpreted as conditioning the probability estimate \hat{P}_e on e , i.e.:

$$\hat{P}_e(Y = y | X = x;) := f(x; S_e) \quad (4)$$

Note that both balancing and conditioning can be achieved simultaneously, which we denote

4.1 Objective and Optimization Details

Given a dataset of samples $\mathcal{D}_e = \{x_i; y_i; e\}_{i=1}^N$, we wish to leverage the NW head as a conditional estimator for Y conditioned on $Z_C = g_C(X)$, where $g_C(X)$ is characterized by Eq(1). This necessitates an optimization over both the space of functions f . Thus, we solve the following constrained maximum likelihood over g_C :

$$\begin{aligned} \operatorname{argmax}_{g_C} \quad & \sum_{i=1}^N \log \hat{P}_{e_i}^B(y_i | g_C(x_i);) \\ \text{s.t.} \quad & \hat{P}_{e_i}^B(y_i | g_C(x_i);) = \hat{P}_{e_i^0}^B(y_i | g_C(x_i);); \quad \forall i \in \{1, \dots, N\}; \quad e_i \in \mathcal{E} \end{aligned} \quad (5)$$

Note that Eq(1) implies that $\hat{P}_{e_i}^B(y_i | g_C(x_i)) = P^B(y_i | g_C(x_i))$. Thus, the objective is equivalent to unconstrained maximum likelihood under the assumption in Eq. (1).

Instead of solving for g_C explicitly, we let both f and g_C be related by the composition $f \circ g_C$, and set f to be the learnable mapping of the NW head, i.e. a neural network. Then, the objective becomes:

$$\begin{aligned} \operatorname{argmin}_{f, g_C} \quad & \sum_{i=1}^N L(f \circ g_C(x_i; S_{e_i}^B); y_i) \\ \text{s.t.} \quad & f \circ g_C(x_i; S_{e_i}^B) = f \circ g_C(x_i; S_{e_i^0}^B); \quad \forall i \in \{1, \dots, N\}; \quad e_i \in \mathcal{E} \end{aligned} \quad (6)$$

where L is the cross-entropy loss. To make the objective tractable, we consider two possible variants:

- Explicit. Solve the optimization problem explicitly via a Lagrangian formulation:

$$\operatorname{argmin}_{f, g_C} \sum_{i=1}^N L(f \circ g_C(x_i; S_{e_i}^B); y_i) + \sum_{e \in \mathcal{E}} \lambda_e \sum_{i=1}^N \|f \circ g_C(x_i; S_e^B) - f \circ g_C(x_i; S_{e^0}^B)\|_2^2 \quad (7)$$

where $\lambda_e > 0$ is a hyperparameter.

- Implicit. Relax the optimization problem into the following unconstrained problem:

$$\operatorname{argmin}_{f, g_C} \sum_{i=1}^N L(f \circ g_C(x_i; S_{e_i}^B); y_i) \quad (8)$$

In this formulation, the constraint will be approximately satisfied in the sense that the model will be encouraged to predict the ground truth for a given image, which is identical across all environments. In practice, how well the solution satisfies the constraint will depend on model capacity, the data sample, and optimization procedure.

4.2 Optimization Details

During training, the support set is drawn stochastically from the training set, and all queries and support datapoints are passed through the feature extractor. For computational efficiency, instead of sampling a unique support mini-batch at the query-level, we sample a unique support at the mini-batch level. Thus, N_q and N_s are the query and support mini-batch sizes respectively, the effective mini-batch size $N_q + N_s$, instead of $N_q N_s$. For the implicit variant, we sample one support set for a given mini-batch of queries, forward pass through the NW head, and compute the loss in Eq.(8). For the explicit variant, we sample two support sets for a given mini-batch of queries, perform two independent forward passes through the NW head for each support set, and compute the loss in Eq.(7). As discussed in prior work [5], the support batch size is a hyperparameter analogous and orthogonal to the query batch size.

A technical point is that the set of labels in the support mini-batch must cover the set of labels in the query mini-batch. Thus, in our implementation, ~~for~~, we cycle through all classes and randomly draw N_c examples per class to include in the support. For tasks with a large number of classes, one can subsample from the total number of classes, so long as the sampled classes cover the set of query classes.

4.3 Inference modes

Similar to how the support set can be manipulated during training, we can also design different inference strategies corresponding to different configurations of the support set at test-time. We explore several different inference modes which are possible under the NW framework:

1. Random. Sample uniformly at random over the dataset, such that each class is represented k times.
2. Full. Use the entire balanced training set.
3. Ensemble. Given the set of balanced features computed from Full mode, partition the computed features for all training datapoints by environment, compute the softmax predictions with respect to each environment balanced across labels, and average the predictions.
4. Cluster. Given the set of balanced features computed from Full mode, perform ~~cluster~~ means clustering on the features of the training datapoints for each class. ~~Cluster~~ centroids are then used as the support features for each class. This can be viewed as a distillation of the full training set for efficient inference, with the caveat that the support set no longer corresponds to observed datapoints.

While Full, Ensemble, and Cluster require computing features for the entire support set, in practice these features and centroids can be precomputed. In our experiments, we find that Cluster mode can be a sufficient replacement to Full mode, while being computationally cheaper. These inference modes can be used interchangeably and flexibly. As an example, consider a workflow which would involve using Cluster mode to perform efficient inference, and then using Full mode on a select few (potentially problematic) test queries to understand model behavior.

4.4 Connections to Prior Work

Our assumptions in Eq(1) are common across many works related to learning invariant predictors [27, 54, 41, 26]. Representatively, under the binary classification setting, the IRM objective finds a representation function which elicits an invariant predictor across environments such that for all h that has a non-zero probability $\text{Pr}(X = h)$ in any (and all) environment(s):

$$E_e[Y | X = h] = E_{e'}[Y | X = h]; \forall e, e' \in \mathcal{E}$$

Eq. (1) can be viewed as a generalization of this equality to multi-class settings.

⁴This equality has also been called "sufficiency invariance" [60].

Table 1: Summary of Datasets.

Dataset	# Classes	Env	# Envs	Architecture	Metric
Camelyon-17	2	Hospital	3	DenseNet-121	Average acc.
ISIC	2	Hospital	3	ResNet-50	F1-score
FMoW	62	Region	5	DenseNet-121	Worst-region acc.

Table 2: Metric average standard deviation for all datasets (%). Higher is best and underline is second-best. Implicit / Explicit.

Algorithm	Camelyon-17	ISIC	FMoW
ERM [52]	70.3 ^{6.4}	58.2 ^{2.9}	32.6 ^{1.6}
IRM [1]	70.9 ^{6.8}	57.9 ^{1.0}	31.3 ^{1.2}
CORAL [48]	72.4 ^{4.4}	59.1 ^{2.2}	31.7 ^{1.0}
Fish [46]	74.7 ^{7e-2}	64.4 ^{1.7}	34.6 ^{0.0}
LISA [64]	77.1 ^{6.5}	64.8 ^{2.3}	35.5 ^{1.8}
CLOvE [54]	<u>79.9</u> ^{3.9}	66.2 ^{2.2}	40.1 ^{0.6}
NW ^B _θ , Random	71.7 ^{5.3}	56.7 ^{1.4}	31.1 ^{0.8}
NW ^B _θ , Full	72.0 ^{6.7}	61.9 ^{3.5}	31.6 ^{0.9}
NW ^B _θ , Cluster	70.6 ^{6.9}	61.4 ^{2.3}	31.3 ^{0.9}
NW ^B _θ , Ensemble	71.9 ^{6.0}	63.9 ^{3.8}	32.2 ^{1.0}
NW ^B _θ , Probe	69.2 ^{7.4}	59.7 ^{2.5}	29.9 ^{1.5}
NW ^B _θ , Random	74.8 ^{8.4} / <u>75.3</u> ^{3.2}	57.5 ^{1.9} / <u>55.0</u> ^{0.9}	31.2 ^{0.7} / <u>30.9</u> ^{0.5}
NW ^B _θ , Full	80.0 ^{2.7} / <u>79.7</u> ^{1.9}	69.6 ^{2.3} / <u>70.0</u> ^{1.0}	35.0 ^{0.7} / <u>34.6</u> ^{0.4}
NW ^B _θ , Cluster	78.6 ^{2.5} / <u>79.0</u> ^{1.4}	71.1 ^{1.7} / <u>71.0</u> ^{1.0}	33.9 ^{0.6} / <u>34.0</u> ^{0.3}
NW ^B _θ , Ensemble	79.5 ^{2.6} / <u>79.6</u> ^{1.9}	69.5 ^{2.2} / <u>69.8</u> ^{0.8}	37.8 ^{0.9} / <u>38.2</u> ^{0.4}
NW ^B _θ , Probe	75.3 ^{7.3} / <u>75.8</u> ^{8.3}	61.4 ^{3.1} / <u>63.4</u> ^{2.8}	33.9 ^{1.5} / <u>32.7</u> ^{1.4}

Furthermore, note that given the feature extractor, the NW mechanism is a nonlearnable classifier, whereas is learned in the IRM setting. Thus, our proposed objective can be interpreted as learning invariant features, where the explicit classifier constraint is satisfied by construction. This avoids the need to approximate the complex bilevel optimization problem with a regularizer which assumes convexity and requires computing the Hessian. Essentially, it enforces invariance through the manipulation of the support set, providing a more intuitive and computationally simpler objective to optimize.

In the Experiments section, we compare IRM against a variant of our algorithm where we freeze the learned representations and tune a linear classifier on top using the same training data. We find that our algorithm performs better than IRM on all datasets, suggesting that it captures invariant representations better than IRM.

5 Experiments and Results

5.1 Baselines

We compare against several popular and competitive baseline algorithms: empirical risk minimization (ERM) [52], invariant risk minimization (IRM) [1], deep CORAL [8], Fish [46], LISA [64], and CLOvE [54]. When available, results on baselines are pulled from their respective papers. Details on baseline algorithms are provided in the Appendix.

5.2 Datasets

We experiment on 3 real-world domain generalization tasks. Two are from the WILDS benchmark [23], and the third is a challenging melanoma detection task. Details on the datasets are summarized in Table 1, and further information is provided in the Appendix.

1. The Camelyon-17 dataset⁴ comprises of microscopic images of stained tissue patches from different hospitals, where the label corresponds to presence of tumor tissue in patches and the environment is the hospital where the patch comes from.
2. The melanoma dataset is from the International Skin Imaging Collaboration (ISIC) archive⁵. The ISIC dataset comprises of dermoscopic images of skin lesions from different hospitals, where the label corresponds to whether or not the lesion is diagnosed as melanoma and the environment is the hospital where the image comes from. There is significant less positive examples and negative examples, with this label imbalance varying across environments (see Appendix).
3. The Functional Map of the World (FMoW) dataset⁶ comprises of RGB satellite images, where the label is one of 62 building or land use categories, and the environment represents the year the image was taken and its geographical region.

5.3 Experimental Setup

For each model variant, we train 5 separate models with different random seeds, and perform model selection on an out-of-distribution (OOD) validation set. For WILDS datasets, we follow all hyperparameters, use model selection techniques, and report metrics as specified by the benchmark. This includes using a DenseNet-121 backbone initialized with pretrained ImageNet weights as and no random augmentations for both datasets. Similarly for ISIC, we use a pretrained ResNet-50 backbone as with no augmentations, and perform model selection on an OOD validation set. Due to significant label imbalance, we report F1-score instead of average accuracy.

For NW algorithms, we refer to models which balance classes (i.e. modeling Eq.(3)) as NW^B , and models which additionally condition on environment (i.e. modeling both Eq.(3) and Eq.(4)) as NW_e^B . For NW_e^B models, we train explicit and implicit variants. For all NW algorithms, we perform evaluation on all inference modes. In addition, for completeness, we experiment on a variant where we freeze the feature extractor and tune a linear probe on top of the learned representations on the same training data \mathcal{D}_{tr} , which we refer to as "Probe". As an example, the implicit variant of NW_e^B is trained on Eq(8), where the support set is balanced across classes (B) and conditioned on an environment (e).

We set $N_c = 8$ for Camelyon-17 and ISIC and $N_c = 1$ for FMoW. An analysis of this hyperparameter is provided in the Appendix. The query batch size is set to 8 for all NW experiments. For Random and Cluster inference modes, we set $k = 3$. This was chosen based on prior work⁵, where $k = 3$ was shown to balance good error rate performance with computational efficiency. For explicit variants, we tune for all datasets via grid search on a held-out validation set. Full hyperparameter details are provided in the Appendix.

All training and inference is done on an Nvidia A6000 GPU and all code is written in PyTorch.

5.4 Results

Table 2 shows the main results. We find that on Camelyon-17 and ISIC datasets, NW^B Full mode outperforms all baselines and variants we consider. In addition, NW^B variants typically have lower variance across random seeds as compared to baselines. For FMoW, NW^B ensemble mode performs around 2% lower than the best performing baseline, CLoVE. We observe that the most computationally-efficient inference mode, Cluster, performs comparably to Full mode for NW models, and is in fact the highest-performing model for ISIC. Thus, we conclude that the support set can be an efficient replacement for Full.

For ISIC, we find that almost all NW^B modes (except Random) perform 3% better than ERM. This may be attributed to balancing classes across environments, which we suspect has added benefit for highly imbalanced tasks. In contrast, this boost is less apparent for Camelyon-17, which has relatively balanced classes. As an ablation, we compare NW^B against an NW variant without class-balancing in the Appendix. NW^B further improves over NW by 7%. Exploring further, we compare NW^B against an ERM variant with balanced classes per environment, which we denote ERM^B .

⁴<https://www.isic-archive.com>

⁶Our code is available at <https://github.com/alanqrwang/nwhead>.

achieves 63.0 ± 2.5, which is on-par with NW. This is expected as the theoretical assumptions are the same for both models.

Comparing implicit to explicit variants of NW we do not find much difference in explicitly enforcing Eq. (1), although we do observe significantly lower variances across model runs. Generally, we find the slight performance gain of explicit training to not be worth the computational overhead of doubling the number of support set forward passes per gradient step and tuning the hyperparameter

While not the highest-performing, we highlight the consistent 1.5% improvement of Probe models over IRM, indicating that NW may be better at capturing invariant features. However, other non-parametric inference modes still outperform Probe, possibly indicating that the learned features are more suitable for NW-style classifiers.

6 Discussion and Limitations

There are several advantages of the proposed NW approach over previous works. First, the implicit training strategy in Eq. (8) has no hyperparameter to tune, while remaining competitive with and often outperforming state-of-the-art baselines which all require tuning a hyperparameter coefficient in the regularized loss. Second, the NW head enables interpretability by interrogating nearest neighbors in the feature space. Since these neighbors directly contribute to the model's prediction, Eq. (2), interrogation enables a user to see what is driving the model's decision-making. This not only allows for greater model transparency, but also enables interrogating the quality of the invariant features. We explore this capability in Section H in the Appendix. Note this this degree of transparency is not present in parametric baselines. Lastly, from an intuitive standpoint, we believe our non-parametric approach to enforcing invariance across environments is more natural than baseline methods, since an environment is encoded by manipulating the support set to contain real samples only from that environment. Other baseline methods resort to proxy methods to enforce invariance [54, 48, 27].

One important limitation of our method is computational (see Appendix for analysis of runtimes). The proposed approach requires pairwise comparisons, which scales quadratically with sample size. Practically, this means passing a separate support mini-batch in addition to a query mini-batch at every training iteration. This limitation is compounded for explicit variants, in which two sets of support sets must be drawn independently. Future work may explore more computationally-efficient training strategies. At inference time, Full, Cluster, and Ensemble modes are expensive procedures which require computing features for the entire support set, although precomputing features can mitigate this. However, we argue that in high-risk, safety-critical domains like medical imaging, high inference throughput may not be as important as performance, interpretability, and robustness.

We expect the proposed approach to work well with tasks that have several (and diverse) sets of examples per label class in each environment. If this is not the case, as in the FMoW dataset, the resulting model will be sub-optimal. In particular, in the extreme case where no example is present for a specific class in a given environment, constructing a support set with labels that cover the ground truth label of the query images will not always be possible. This will, in turn, impact performance.

7 Conclusion

We presented a nonparametric strategy for invariant representation learning based on the Nadaraya-Watson (NW) head. In the NW head, the prediction is made by comparing the learned representations of the query to the elements of a support set that consists of labeled data. We demonstrated two possible ways of manipulating the support set, and demonstrate how this corresponds to encoding different assumptions from a causal perspective. We validated our approach on three challenging and real-world datasets.

We believe there are many interesting directions of further research. First, our treatment is restricted to classification tasks. Future work may explore an extension to the regression setting. Second, it can be interesting to explore adaptation to the test domain, given additional information. For example, reweighting the occurrence of samples per label could provide improved results given knowledge about the edge $\mathbb{E}[Y]$ in the test distribution. One can further envision implementing the proposed method in settings where there are previously unseen test time labels/tasks. Finally, we are interested in replacing the fixed similarity function with a learnable kernel.

Acknowledgements

Funding for this project was in part provided by the NIH grant R01AG053949, and the NSF CAREER 1748377 grant.

References

- [1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. URL <https://arxiv.org/abs/1907.02893>.
- [2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)* Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.
- [3] Jan Brabec, Tomáš Komárek, Vojtěch Franc, and Lukáš Machlička. On model evaluation under non-constant class imbalance, 2020.
- [4] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandeveld, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.
- [5] Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching, 2022.
- [6] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world, 2018.
- [7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kaloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). Eprint arXiv:1902.03368, 2019.
- [8] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kaloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Proceedings of ISBI* pages 168–172. IEEE, 2018.
- [9] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. Eprint arXiv:1908.02288, 2019.
- [10] Andreas Damianou and Neil D. Lawrence. Deep gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* volume 31 of *Proceedings of Machine Learning Research* pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL: <https://proceedings.mlr.press/v31/damianou13a.html>.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL <https://arxiv.org/abs/2010.11929>.
- [12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. A. Ché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems* volume 32. Curran Associates,

- Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2974788b53f73e7950e8aa49f3a306db-Paper.pdf.
- [13] Rasool Fakoor, Pratik Chaudhari, Jonas Mueller, and Alexander J. Smola. Trade: Transformers for density estimation, 2020. URL <https://arxiv.org/abs/2004.02441>.
- [14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016.
- [15] J. Ghosh and A. Nag. An Overview of Radial Basis Function Networks, pages 1–36. Physica-Verlag HD, Heidelberg, 2001. ISBN 978-3-7908-1826-0. doi: 10.1007/978-3-7908-1826-0_1. URL https://doi.org/10.1007/978-3-7908-1826-0_1.
- [16] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020.
- [17] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). Eprint arXiv:1605.01397, 2016.
- [18] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness, 2019.
- [19] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021. URL <https://arxiv.org/abs/2103.03206>.
- [20] Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts, 2022. URL <https://arxiv.org/abs/2207.01603>.
- [21] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems, 2022. URL <https://arxiv.org/abs/2206.15475>.
- [22] Pritish Kamath, Akilesh Tangella, Danica J. Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance?, 2021.
- [23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021.
- [24] Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Tom Rainforth, and Yarin Gal. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning, 2021. URL <https://arxiv.org/abs/2106.02584>.
- [25] Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. Nonparametric uncertainty quantification for single deterministic neural network, 2022.
- [26] Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem ?, 2021.
- [27] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex), 2021.
- [28] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 2805–2814. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/kumar18a.html>.

- [29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization, 2017.
- [30] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. doi: 10.1109/CVPR.2018.00566.
- [31] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors, 2018.
- [32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks, 2015.
- [33] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, 2018.
- [34] Lucas Mansilla, Rodrigo Echeveste, Diego H. Milone, and Enzo Ferrante. Domain generalization via gradient surgery, 2021.
- [35] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications* 9(1): 141–142, 1964. doi: 10.1137/1109020. URL <https://doi.org/10.1137/1109020>.
- [36] Timothy Nguyen, Zhoung Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression, 2021.
- [37] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, 2018. URL <https://arxiv.org/abs/1803.04765>.
- [38] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018. URL <https://arxiv.org/abs/1802.05751>.
- [39] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.
- [40] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, 2015.
- [41] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning, 2018.
- [42] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization, 2021.
- [43] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data* 8(1):1–8, 2021.
- [44] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning, 2012.
- [45] A Scope, AA Marghoob, CS Chen, JA Lieb, MA Weinstock, AC Halpern, and SONIC Study Group. Dermoscopic patterns and subclinical melanocytic nests in normal-appearing skin. *British Journal of Dermatology* 160(6):1318–1321, 2009.
- [46] Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization, 2021.
- [47] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017. URL <https://arxiv.org/abs/1703.05175>.
- [48] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016.

- [49] Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=UavQ9HYye6n> .
- [50] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* 5(1): 1–9, 2018.
- [51] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance, 2014.
- [52] V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* 10(5):988–999, 1999. doi: 10.1109/72.788640.
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL <https://arxiv.org/abs/1706.03762> .
- [54] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization, 2021. URL <https://arxiv.org/abs/2102.10395> .
- [55] Alan Q. Wang and Mert R. Sabuncu. A explainable nadaraya-watson head can offer explainable and calibrated classification. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=iEq6lhG4O3> .
- [56] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization, 2022.
- [57] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization, 2023.
- [58] Yixin Wang and Michael I. Jordan. Desiderata for representation learning: A causal perspective, 2022.
- [59] Yufei Wang, Haoliang Li, and Alex C. Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626, 2020. doi: 10.1109/ICASSP40776.2020.9053273.
- [60] Zihao Wang and Victor Veitch. The causal structure of domain invariant supervised representation learning, 2022. URL <https://arxiv.org/abs/2208.06987> .
- [61] G. S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A* (26):359–372, 1964.
- [62] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning, 2015. URL <https://arxiv.org/abs/1511.02222> .
- [63] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup, 2019.
- [64] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation, 2022.
- [65] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data, 2022.
- [66] Dequan Zhang, Ning Zhang, Nan Ye, Jianguang Fang, and Xu Han. Hybrid learning algorithm of radial basis function networks for reliability analysis. *IEEE Transactions on Reliability* 70(3):887–900, 2021. doi: 10.1109/TR.2020.3001232.
- [67] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

- [68] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees, 2018. URL<https://arxiv.org/abs/1802.00121> .
- [69] Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. Domain generalization via optimal transport with metric similarity learning. *Neurocomputing* 456: 469–480, oct 2021. doi: 10.1016/j.neucom.2020.09.091. URL<https://doi.org/10.1016%2Fj.neucom.2020.09.091> .
- [70] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation, 2020.
- [71] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pages 1–20, 2022. doi: 10.1109/tpami.2022.3195549. URL<https://doi.org/10.1109%2Ftpami.2022.3195549> .

A Description of Baselines

- Empirical risk minimization (ERM) [52] minimizes the sum of errors across domains and examples.
- Invariant risk minimization (IRM) [1] learns a feature representation such that the optimal linear classifier on top of that representation matches across domains. For WILDS datasets, we pull baseline performance from [54]. For ISIC, we use the implementation from [16].
- Deep CORAL [48] penalizes differences in the means and covariances of the feature distributions (i.e., the distribution of last layer activations in a neural network) for each domain. For WILDS datasets, we pull baseline performance from [54]. For ISIC, we use the implementation from [16].
- Fish targets domain generalization by maximizing the inner product between gradients from different domains. For WILDS datasets, we pull baseline performance from the original paper. For ISIC, we use the implementation from [16].
- LISA augments the set of training data by randomly performing two types of mixup-style [67] interpolations: intra-label (same label, different domain) and inter-label (same domain, different label). For WILDS datasets, we pull baseline performance from the original paper, and train their implementation on the ISIC dataset.
- CLOvE [54] finds an invariant classifier by enforcing the classifier to be calibrated across all training domains. While the original paper proposes several model variants leveraging this idea, we report their best-performing variant, which starts with a trained CORAL model and finetunes the weights using a regularized cross-entropy loss. The regularizer aggregates Maximum Mean Calibration Error (MMCE) [28] over all training domains. For WILDS datasets, we pull baseline performance from [54]. As their implementation is not publicly-available, we implement it for ISIC.

B Description of Datasets

Representative examples of the 3 datasets are shown in Fig. 4.

- **Camelyon-17.** We use Camelyon-17 from the WILDS benchmark [4, 23], which provides 450,000 lymph-node scans sampled from 5 hospitals. Camelyon-17 is a medical image classification task where the input x is a 96×96 image and the label y is whether there exists tumor tissue in the image. The environment denotes the hospital that the patch was taken from. The training dataset is drawn from the first 3 hospitals, while out-of-distribution validation and out-of-distribution test datasets are sampled from the 4th hospital and 5th hospital, respectively.
- **ISIC.** The melanoma dataset is from the International Skin Imaging Collaboration (ISIC) archive⁷. Data from the archive are collected by different organizations at different points in time [7–9, 17, 43, 45, 50]. There are about 70k data samples in total. In particular, the resized input image x is a 224×224 image and a binary target label y denotes whether the image exhibit is melanoma or not. The environment is the hospital from which the image was collected⁸. We follow a similar setup to Camelyon-17. The training dataset is drawn from the first 3 hospitals, while out-of-distribution validation and out-of-distribution test datasets are sampled from the 4th hospital and 5th hospital, respectively. For preprocessing, we filter out datapoints that are not specifically categorized as "benign" or "malignant" (e.g. "indeterminate"). The OOD validation dataset is from the "Barcelona1" site indicator and the OOD test dataset is the "Vienna1" site indicator.
- **FMoW.** The FMoW dataset is from the WILDS benchmark [6, 23], a satellite image classification task which includes 62 classes and 80 domains (16 years x 5 regions). Concretely, the input x is a 224×224 RGB satellite image, the label y is one of the 62 building or land use categories, and the environment represents the year that the image was taken as well

⁷<https://www.isic-archive.com>

⁸Hospitals are Hospital Clinic of Barcelona, Medical University of Vienna, University of Queensland Diamantina Institute, Memorial Sloan Kettering Cancer Center, University of Sydney Melanoma Diagnostic Centre, and University of Pittsburgh Medical Center.

as its corresponding geographical region – Africa, the Americas, Oceania, Asia, or Europe. The train/test/validation splits are based on the time when the images are taken. Specifically, images taken before 2013 are used as the training set. Images taken between 2013 and 2015 are used as the validation set. Images taken after 2015 are used for testing.

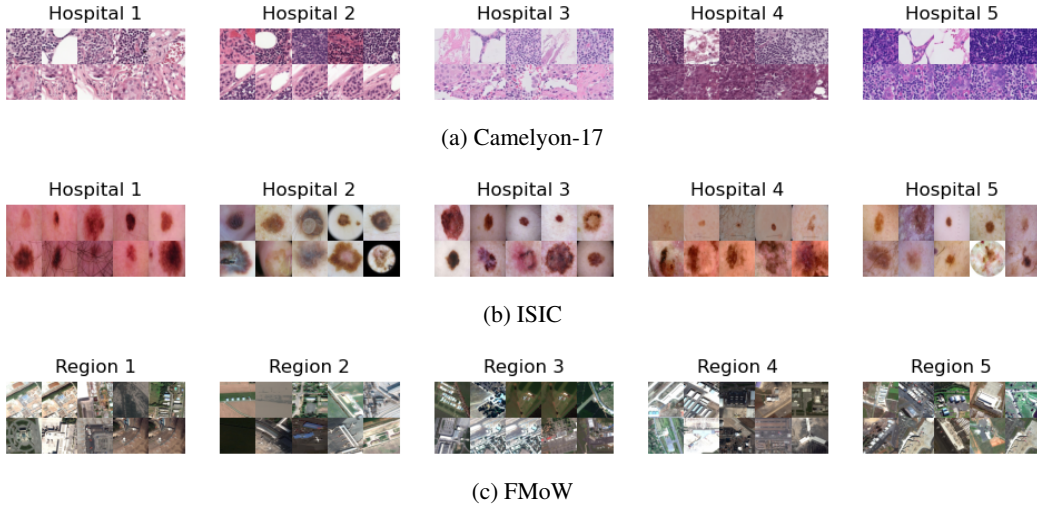


Figure 4: Representative images for datasets, separated by domain. Each row depicts a separate class. For FMoW, for simplicity, we show 2 classes out of 62 and only images before 2013.

C Hyperparameter Details

Table 3 shows hyperparameter settings for all datasets, where NW-specific hyperparameters are below the midline. For all models, we use pretrained ImageNet weights. For β , we perform a grid-search over the values $\{0.01, 0.1, 1\}$. Fig. 5 depicts NW_c^B performance vs N_c for Camelyon-17 and ISIC datasets. We find that performance is relatively insensitive to N_c above 5 examples per class.

In the original NW head paper [55], the authors experiment with a temperature (i.e. bandwidth) hyperparameter β . In this work, we set $\beta = 1$ for all experiments. The reason for this is that we optimize both the feature extractor and classifier end-to-end, and the kernel used in the classifier is dependent on the features that the feature extractor learns (unlike [25], e.g.). Thus, we let the feature extractor to optimize the bandwidth on its own. Note this same approach is taken in prior works [47].

Table 3: Hyperparameter settings for various datasets.

Hyperparameter	Camelyon-17	ISIC	FMoW
Learning rate	1e-4	5e-5	1e-4
Weight decay	1e-4	0	1e-2
Scheduler	None	None	StepLR
Batch size	32	8	8
Architecture	DenseNet-121	ResNet-50	DenseNet-121
Optimizer	SGD	Adam	Adam
Maximum Epoch	10	5	60
N_q	8	8	8
N_c	8	8	1
N_s	$N_c \cdot 2 = 16$	$N_c \cdot 2 = 16$	$N_c \cdot 62 = 62$
	0.01	0.01	0.1
k	3	3	3

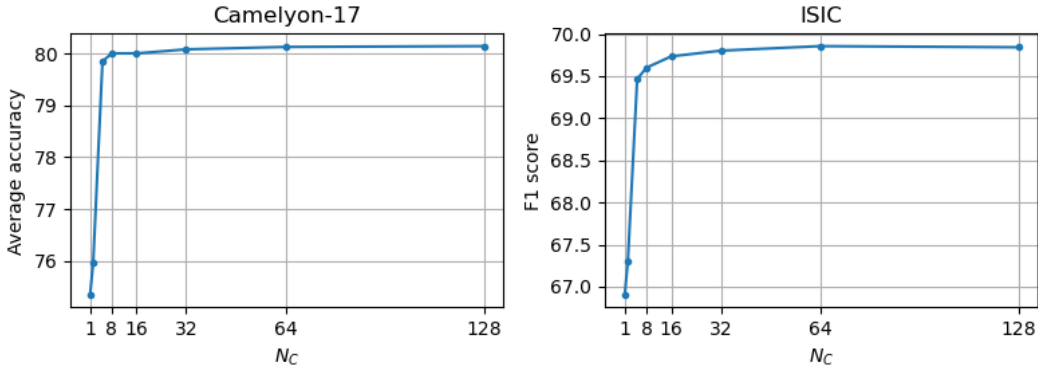


Figure 5: NW_c^B performance vs N_c for Camelyon-17 and ISIC datasets. Full mode. Performance is relatively insensitive to N_c above 5 examples per class.

D Table of Runtimes

Table 4 shows approximate runtimes for various datasets during training and inference. All experiments are performed on a GPU.

Table 4: Approximate runtimes for various algorithms. Training time is time to complete maximum epochs as specified in Table 3, and does not include validation. Inference time is time to evaluate the entire test set. Averaged over all training runs.

	<i>Algorithm</i>	<i>Camelyon-17</i>	<i>ISIC</i>	<i>FMoW</i>
Training	ERM	7 hr	1 hr	22 hr
	NW	14 hr	2 hr	40 hr
Inference	ERM	10 min	2 min	10 min
	NW, Random	15 min	3 min	20 min
	NW, Full	2 hr	15 min	1 hr
	NW, Ensemble	2 hr	15 min	1 hr
	NW, Cluster	2.2 hr	17 min	1.1 hr
	NW, Probe	10 min	2 min	10 min

E ID vs. OOD Performance

With most invariant learning methods, prior works observe a tradeoff between in-distribution and out-of-distribution generalization. In Table 5, we show in-distribution (ID) and out-of-distribution (OOD) results for Camelyon-17, which provides an ID validation set. We observe that there is a tradeoff between ID and OOD performance, similar to prior work.

Table 5: In-distribution vs. out-of-distribution generalization performance on Camelyon-17.

	<i>In-distribution</i>	<i>Out-of-distribution</i>
ERM	93.2 5.2	70.3 6.4
NW ^B , Full	96.1 1.0	72.0 6.7
NW _e ^B , Full	92.8 2.0	80.0 2.7

F Imbalanced ISIC Experiments

As ISIC exhibits significant label imbalance where the positive class is much less represented than the negative class (see Fig. 6), we experiment with an NW variant without support-set label balancing as an ablation. For both variants, we set $N_c = 8$. To train the imbalanced variant, we sample a mini-batch support set by first sampling one image per class (to guarantee both classes are represented in the support at least once), and then sampling the rest of the images randomly from the dataset.

To characterize the performance of both variants across class imbalances, we change the prevalence of $y = 0$ in the test set by removing negative class images until the desired prevalence is achieved. Note the default prevalence is 0.85. Then, we compute the accuracy over this manipulated test set (note, prior work has shown that F1-score is not a good metric for comparing classifiers with different label imbalances [3]).

Fig 7 shows results. We observe that at high prevalence where the proportion of negative classes matches in training and test, the imbalanced variant outperforms the balanced variant, whereas the opposite is true for low prevalence. This makes sense because at high prevalence, the class imbalance is similar for the training and test domains; thus, a model which overpredicts the negative class is usually right. On the other hand, this fails in test sets where the prevalence is flipped (i.e. low prevalence). These results suggest that NW^B is a more robust classifier in the presence of label shift.

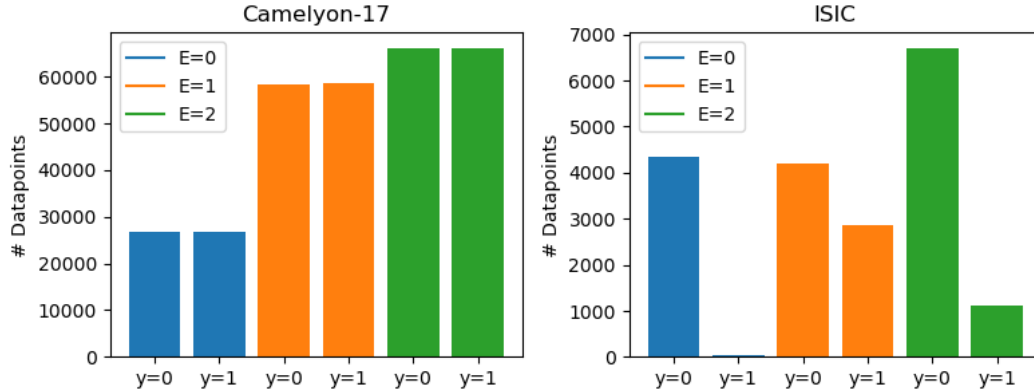


Figure 6: Number of datapoints separated by class for Camelyon-17 and ISIC datasets. There is significant label imbalance for the ISIC dataset.

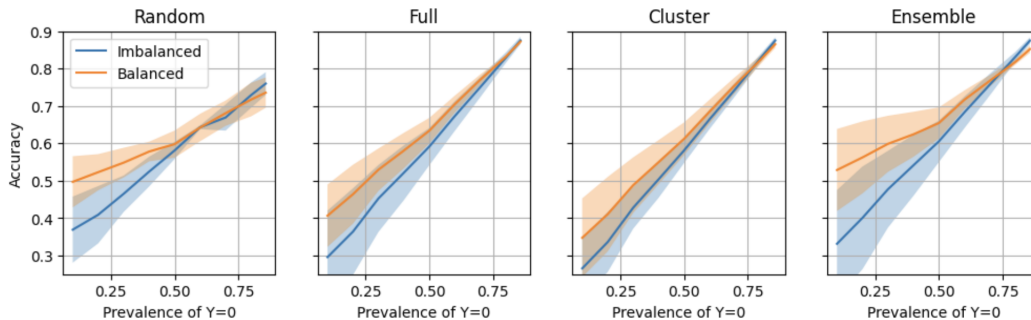


Figure 7: Accuracy of NW (imbalanced) and NW^B (balanced) models over varying prevalence of $y = 0$ for ISIC dataset. At low prevalence where the prevalence differs the most from training domains, we observe that model performance is higher for NW^B . The default prevalence is $6705/7818 = 0.8576$, which is the right-most value in the x-axis.

G Nearest-neighbor Inference Modes

In addition to Random, Full, Cluster, and Ensemble inference modes, we additionally experiment with two inference modes based on nearest neighbors: k-NN and HNSW. HNSW (Hierarchical Navigable Small Worlds) is a fast approximate nearest neighbor algorithm [33]. We choose $k = 20$ based on prior work [25]. HNSW is about two times faster in total runtime on a GPU than full k-NN.

Overall, we observe that k-NN and HNSW perform nearly identically for all the datasets and variants. Additionally, both modes perform better in terms of mean performance on Camelyon-17, and perform on par with the best-performing modes for ISIC (Cluster) and FMoW (Ensemble). However, they generally have higher variances across model runs. We suspect that less total samples used in the support (20 as compared to more than 1000 for Full) and the fact that not all classes are guaranteed to be represented in the support may lead to less stability across model runs.

Table 6: Metric average standard deviation for all datasets (%). Higher is better. Implicit / Explicit.

<i>Algorithm</i>	<i>Camelyon-17</i>	<i>ISIC</i>	<i>FMoW</i>
NW ^B , Full	72.0 _{6.7}	61.9 _{3.5}	31.6 _{0.9}
NW ^B , Cluster	70.6 _{6.9}	61.4 _{2.3}	31.3 _{0.9}
NW ^B , Ensemble	71.9 _{6.0}	63.9 _{3.8}	32.2 _{1.0}
NW ^B , k-NN	72.5 _{3.2}	64.2 _{2.6}	32.5 _{1.4}
NW ^B , HNSW	72.5 _{3.2}	64.2 _{2.6}	32.5 _{1.4}
NW _e ^B , Full	80.0 _{2.7 / 79.7} _{1.9}	69.6 _{2.3 / 70.0} _{1.0}	35.0 _{0.7 / 34.6} _{0.4}
NW _e ^B , Cluster	78.6 _{2.5 / 79.0} _{1.4}	71.1 _{1.7 / 71.0} _{1.0}	33.9 _{0.6 / 34.0} _{0.3}
NW _e ^B , Ensemble	79.5 _{2.6 / 79.6} _{1.9}	69.5 _{2.2 / 69.8} _{0.8}	37.8 _{0.9 / 38.2} _{0.4}
NW _e ^B , k-NN	80.9 _{10.3 / 80.2} _{4.9}	70.7 _{7.4 / 71.0} _{7.2}	37.9 _{2.3 / 38.3} _{1.1}
NW _e ^B , HNSW	80.9 _{10.3 / 80.2} _{5.0}	70.7 _{7.4 / 71.0} _{7.2}	37.9 _{2.2 / 38.2} _{1.1}

H Interpretability of NW Head

In this section, we provide both a visual and quantitative exploration of the interpretability capabilities of the NW head. In Fig. 8, we show several query images and its 8 nearest neighbors in the feature space, for both NW^B and NW_e^B variants. The colored border around each neighbor indicates which training environment the image comes from. Interestingly, we notice that the neighbors for NW_e^B come from a variety of environments (note a variety of colored borders), while the neighbors for NW^B are less diverse.

Fig. 9 quantifies this phenomenon, depicting a normalized histogram of the environments from which the top 20 nearest neighbors originate in the training dataset for Camelyon-17, averaged over all queries in the test set. From these results, it is clear that NW_e^B leverages support images from a wider variety of environments to make its prediction, suggesting that it captures more invariant representations.

