A Description of Baselines

- Empirical risk minimization (ERM) [49] 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 [51]. For ISIC, we use the implementation from [16].
- Deep CORAL [45] 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 [51]. 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 [64] 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 [51] 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) [27] over all training domains. For WILDS datasets, we pull baseline performance from [51]. As their implementation is not publicly-available, we implement it for ISIC.

B Description of Datasets

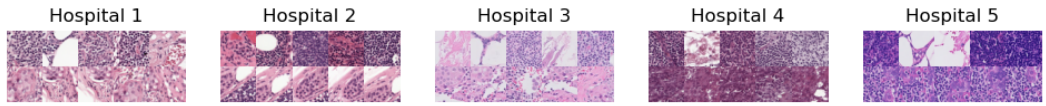
Representative examples of the 3 datasets are shown in Fig. 4.

- **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) archive[7]. Data from the archive are collected by different organizations at different points in time [7, 8, 9, 17, 41, 43, 47]. 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 collected [8]. 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

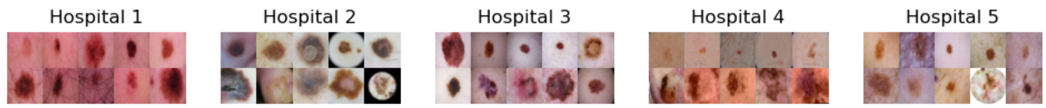
[7] https://www.isic-archive.com

[8] Hospitals 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.



(a) Camelyon-17



(b) ISIC

(c) FMoW

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 λ, we perform a grid-search over the values $\{0.01, 0.1, 1\}$. Fig. 5 depicts $\text{NW}_\text{e}^\text{B}$ performance vs N_c for Camelyon-17 and ISIC datasets. We find that performance is relativity insensitive to N_c above ~ 5 examples per class.

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	$N_c \times 2 = 16$	$N_c \times 2 = 16$	$N_c \times 62 = 62$
λ	0.01	0.01	0.1
k	3	3	3

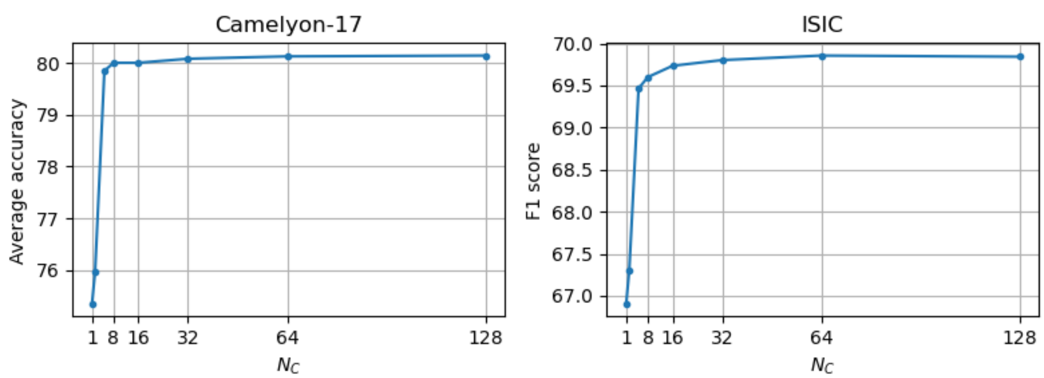


Figure 5: $\text{NW}_\text{e}^\text{B}$ performance vs N_c for Camelyon-17 and ISIC datasets. Full mode. Performance is relativity insensitive to N_c above ~ 5 examples per class.

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.

	Algorithm	*Camelyon-17*	*ISIC*	*FMoW*
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 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 prevalance 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.

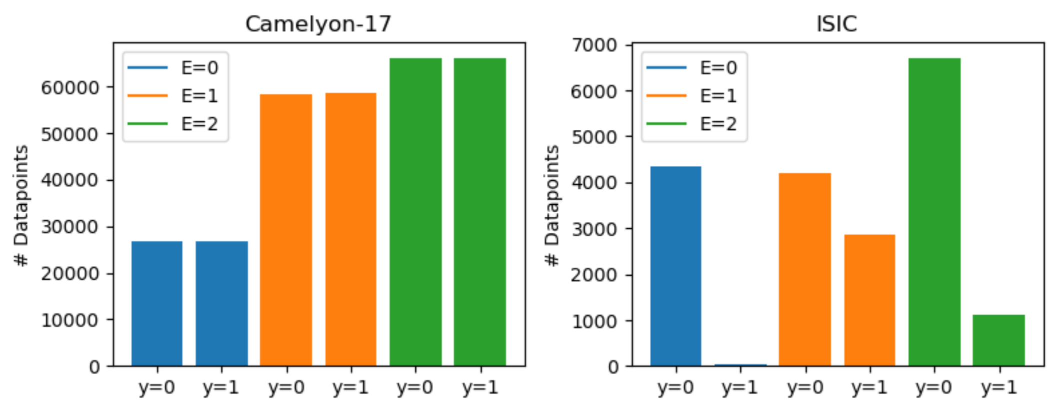


Figure 6: Number of datapoints separated by class for Camelyon-17 and ISIC datasets. There is significant label imbalance for the ISIC dataset.

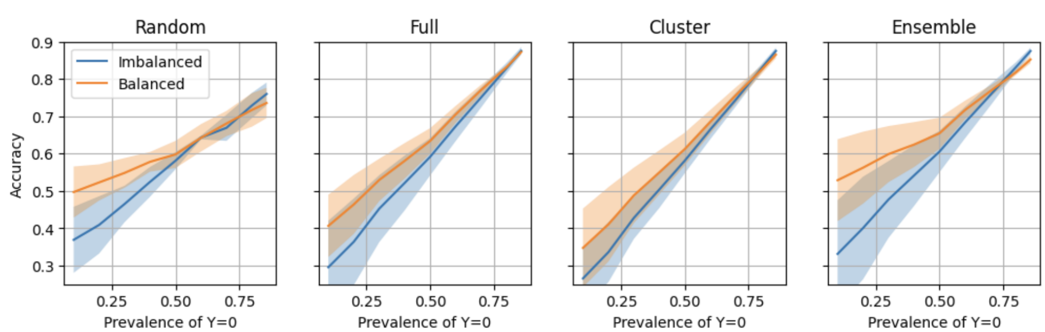


Figure 7: Accuracy of NW (imbalanced) and NW^{B} (balanced) models over varying prevalence of $y = 0$ for ISIC dataset. At low prevalence where the prevalence differs the most from training domains, we observe that model performance is higher for NW^{B}. The default prevalence is 6705/7818 = 0.8576,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