Exposing and addressing the fragility of neural networks in digital pathology

Joona Pohjonen$^{1,*,†}$, Carolin Stürenberg$^{1,†}$, Atte Föhr$^{1}$, Esa Pitkänen$^{2,3,4,‡}$, Antti Rannikko$^{1,5,‡}$, and Tuomas Mirtti$^{1,6,‡}$

$^1$Research Program in Systems Oncology, University of Helsinki
$^2$Institute for Molecular Medicine Finland (FIMM), HI LIFE, University of Helsinki
$^3$Research Program in Applied Tumor Genomics, University of Helsinki
$^4$iCAN Digital Precision Cancer Medicine Flagship, Finland
$^5$Department of Urology, Helsinki University Hospital
$^6$Department of Pathology, Helsinki University Hospital

†Authors contributed equally
‡Senior author

Abstract

Neural networks have achieved impressive results in many medical imaging tasks but often perform substantially worse on out-of-distribution datasets originating from different medical centres or patient cohorts. Evaluating this lack of ability to generalise and address the underlying problem are the two main challenges in developing neural networks intended for clinical practice.

In this study, we develop a new method for evaluating neural network models’ ability to generalise by generating a large number of distribution-shifted datasets, which can be used to thoroughly investigate their robustness to variability encountered in clinical practice. Compared to external validation, shifted evaluation can provide explanations for why neural networks fail on a given dataset, thus offering guidance on how to improve model robustness. With shifted evaluation, we demonstrate that neural networks, trained with state-of-the-art methods, are highly fragile to even small distribution shifts from training data, and in some cases lose all discrimination ability.

To address this fragility, we develop an augmentation strategy, explicitly designed to increase neural networks’ robustness to distribution shifts. StrongAugment is evaluated with large-scale, heterogeneous histopathology data including five training datasets from two tissue types, 274 distribution-shifted datasets and 20 external datasets from four countries. Neural networks trained with StrongAugment retain similar performance on all datasets, even with distribution shifts where networks trained with current state-of-the-art methods lose all discrimination ability. We recommend using strong augmentation and shifted evaluation to train and evaluate all neural networks intended for clinical practice.

1 Introduction

Through digitalisation and neural network based solutions, pathology is experiencing its third revolution [1]. Neural networks have been applied to tissue diagnostics with impressive results, often surpassing human counterparts in consistency, speed and accuracy [2, 3]. Despite promising results, many recent works have demonstrated neural networks performing substantially worse on datasets not used during training [4, 5, 6, 7, 8, 9, 10]. The two main challenges facing neural networks intended for clinical practice are how to identify networks failing to generalise to these out-of-distribution datasets without having access to a large number of such datasets, and how to create networks not suffering from this problem [11]. Otherwise, neural network predictions can only be trusted with in-distribution datasets, thus requiring training or fine-tuning a neural network for each medical centre, imaging equipment and sample processing pipeline.

Clinical grade neural networks should be robust to all data distribution shifts between in-distribution and out-of-distribution datasets possibly encountered in clinical practice [12]. This robustness of neural networks is hard to assess and has been commonly evaluated by performance on external validation datasets. Although an important part of neural networks’ path to clinic [11], external validation is limited to exposing a failure to generalise in the particular dataset, and cannot establish clinical usefulness, robustness or generalisation ability of a network [13]. Even if a neural network achieves good performance on multiple external datasets, it may fail on another dataset or a small subset of samples. Thus, there is a crucial need for evaluating neural networks’ robustness to data distribution shifts.

Large multi-cohort datasets, which are representative of all types of data encountered in clinical practice, are often seen as the solution to the generalisation problem [11]. Although important, larger datasets often have lower data quality [14] and different cohorts introduce their own biases, which makes training harder as neural networks overfit these biases easily [10, 12, 15]. Still, even a large high-quality multi-cohort dataset could not fully represent all types of data encountered in clinical practice, and

*Correspondence: joona.pohjonen@helsinki.fi
Code: https://github.com/jopo666/StrongAugment

arXiv:2206.15274v1 [eess.IV] 30 Jun 2022
other complementary methods for improving neural networks’ robustness are needed.

Several solutions have been suggested for increasing neural networks’ robustness to data distribution shifts encountered in clinical practice. Removing variability (mainly colour) by normalising each image to a well-defined common standard \[6][16][17][18][19][20]\, does allow networks trained on less representative samples to perform more consistently, but can be computationally costly. The recently proposed method of spectral decoupling \[21\] can be used to avoid overfitting unwanted biases, for example in multi-cohort datasets, while also increasing robustness to data distribution shifts \[12\]. Recent works have also shown that introduction of artificial variability to images during neural network training, also known as augmentation, can be used to explicitly improve robustness to certain data distribution shifts \[12][20][22\]. For example, the sharpness of the scanned pathology slide images varies significantly between medical centres, scanning equipment and slide areas \[12][20\]. By randomly adjusting the sharpness of the images during training, neural networks can be made robust to sharpness variations. Automatic augmentation strategies \[23][24][25\] have become a standard in training neural networks \[20\]. However, augmentation has been traditionally seen as a way to increase the amount of training data, and the development of augmentation strategies has not focused explicitly on improving neural networks’ robustness to data distribution shifts.

In this study, we tackle both of the main challenges facing neural networks intended for clinical practice \[11\]. First, we develop an evaluation method, shifted evaluation, which can be used to exhaustively evaluate neural networks’ robustness to data distribution shifts, and thus also its generalisation ability in clinical practice. We use shifted evaluation to demonstrate that neural networks trained with the current state-of-the-art methods are extremely fragile to even small shifts from the training data distribution. Second, we develop an automatic augmentation method, StrongAugment (SA), which is designed to explicitly improve neural networks’ robustness to data distribution shifts. We demonstrate that neural networks trained with SA perform consistently despite large data distribution shifts, even when networks trained with the current state-of-the-art methods display no better than random performance. Importantly, SA also improves robustness for data distribution shifts, which are not included in the augmentation strategy, such as shifts caused by different haematoxylin and eosin stain intensities. Finally, we validate the results by training neural networks on four different datasets, evaluating each on five external datasets. Networks trained with SA achieve better or comparable performance in each external dataset, whereas networks trained with the current state-of-the-art methods fail to generalise to several datasets.

2 Materials and methods

2.1 Shifted evaluation with distribution-shifted dataset

External validation datasets are often used for evaluating neural networks. This evaluation method is limited to exposing a failure to generalise in the specific dataset used, and cannot establish clinical usefulness, robustness or generalisation ability of a network \[13\]. Additionally, when a neural network fails to generalise to an external dataset, there is no practical way of knowing why. For these reasons, we define distribution-shifted datasets which can be used to evaluate neural networks’ robustness to known data distribution shifts.

To create a distribution-shifted dataset \(X_f\) from dataset \(X\), function \(f\) is applied to each image \(x \in X\). As \(f\) completely describes the data distribution shift, any performance differences between the datasets \(X_f\) and \(X\) is guaranteed to be due to the distribution shift defined by \(f\). For example, using function \(f(x) = 255 - x\) creates a new dataset, where colours in each image have been inverted. This distribution-shifted dataset can then be used to evaluate neural networks’ robustness to the distribution shift caused by colour inversion. A more useful example would be to create several datasets with \(jpeg(x, q)\), where \(jpeg\) denotes the JPEG compression algorithm \[27\] and \(q \in \{0.1, 0.2, \ldots, 1.0\}\) the quality level. These distribution-shifted datasets can then be used to evaluate whether a neural network is robust to artefacts created during image compression.

Figure 1 demonstrates the usefulness of shifted evaluation. The training dataset \(X_{train}\), evaluation dataset \(X_{eval}\) and several distribution-shifted datasets \(X_{shift}\) are represented in the space of all possible data distributions. Gray arrows represent the known function \(f\) used to shift the data distribution in each \(X_{shift}\). The red arrow indicates the unknown \(f\) which represents the transformation from the training to evaluation distribution. With shifted evaluation, it is easy to evaluate neural networks’ robustness to different known data distribution shifts.
2.1.1 Common image transformations

To evaluate neural networks’ overall robustness to data distribution shifts, 12 commonly used image transformations are used to create 94 distribution-shifted datasets. These datasets are denoted by \( X(t, m) \), where \( t \) denotes the used transformation and \( m \) its magnitude. By increasing or decreasing magnitude \( m \), the data distribution can be incrementally shifted further from the original distribution of dataset \( X \). Figure 2 presents all selected transformations with examples of transformed images with different magnitudes.

The distribution-shifted datasets presented in Figure 2 are used to evaluate neural networks’ overall robustness to data distribution shifts. As the transformations and their magnitudes were selected to measure overall robustness to distribution shifts, some of the distribution-shifted datasets may not be directly encountered in a clinical setting. For example, it is unlikely that a neural network will ever encounter a dataset similar to \( X(\text{Posterize}, 1) \), but image sharpness and colours vary within and between datasets. Nevertheless, it is reasonable to expect a neural network to perform similarly on all of the examples in Figure 2 and any potential failures will give important insights to make the network more robust.

We use the transformations implemented by the torchvision (1.11.1) library, except for the colour channel intensity transformations (Red, Green, Blue) which have been implemented by us.

2.1.2 Haematoxylin and eosin stain intensities

Colours of the digital slide images vary widely due to differences in the staining processes and the choice of imaging equipment and its settings, especially between medical centres [20]. Several transformations presented in Figure 2 evaluate robustness to data distribution shifts caused by changes in the colour of the images, but none explicitly measure robustness to data distribution shifts caused by differences in the staining process. For this reason, and to demonstrate the flexibility of distribution-shifted datasets, 180 datasets are created by adjusting the intensity of the haematoxylin and eosin stains.

To create a distribution-shifted dataset \( X(\text{M}, h, e) \), the Macenko method [16], denoted by \( M \), is used to separate the haematoxylin and eosin stains, and then concentrations of each stain are multiplied with magnitudes \( h, e \), respectively. For example, \( h = 0.0, e = 1.0 \) would remove eosin stain completely from the image and \( e = 1.0, m_h = 2.0 \) would double the haematoxylin stain intensity. Figure 3 shows example images from all distribution-shifted datasets \( X(\text{M}, h, e) \). We use the mean stain vector of the dataset to separate the stains.

2.2 StrongAugment

SA is a novel automatic augmentation method, explicitly designed to increase neural networks’ robustness to data distribution shifts from the training data. In this study, SA is compared to RandAugment (RA) [24], one of the most commonly used automatic augmentation methods, and TrivialAugment (TA) [25] which represents the current state-of-the-art in automatic augmentation. Additionally, implementations for both methods...
Table 1: Augmentation spaces of StrongAugment, RandAugment \cite{24} and TrivialAugment \cite{25}. No effect column represents the magnitude which does not change the image where applicable. StrongAugment includes 6 additional transformations, Gaussian blurring and wider magnitude ranges for most of the common transformations. Examples for some of the transformations can be seen in Figure 2.

| Transformation | StrongAugment | RandAugment | TrivialAugment | No effect |
|----------------|---------------|-------------|----------------|-----------|
| Identity       | included      | included    | included       |           |
| Shear \(X,Y\)  | \([-145, 145]\) | \([-17, 17]\) | \([-145, 145]\) | 1.0       |
| Translate \(X,Y\) | \([-32, 32]\) | \([-72, 72]\) | \([-32, 32]\) | 0         |
| Rotate         | \([-135, 135]\) | \([-30, 30]\) | \([-135, 135]\) | 0         |
| Saturation     | \([0.0, 2.0]\) | \([0.1, 1.9]\) | \([0.01, 1.99]\) | 1.0       |
| Brightness     | \([0.1, 1.9]\) | \([0.1, 1.9]\) | \([0.01, 1.99]\) | 1.0       |
| Contrast       | \([0.1, 1.9]\) | \([0.1, 1.9]\) | \([0.01, 1.99]\) | 1.0       |
| Sharpness      | \([1.0, 2.0]\) | \([0.1, 1.9]\) | \([0.01, 1.99]\) | 1.0       |
| Gaussian blur  | \([0.0, 2.0]\) | 0.0         |               |           |
| Solarize       | \([0, 255]\)  | \([0, 255]\) | \([0, 255]\)  | 256       |
| Posterize      | \([1, 8]\)    | \([4, 8]\)  | \([2, 8]\)    | 8         |
| Equalize       | included      | included    | included       |           |
| Auto contrast  | included      | included    | included       |           |

Grayscale included
Gamma \([0.1, 1.9]\) 1.0
Hue \([-0.5, 0.5]\) 0.0
Red \([0.01, 1.99]\) 1.0
Green \([0.01, 1.99]\) 1.0
Blue \([0.01, 1.99]\) 1.0

Figure 3: Examples of image \(x\), where the haematoxylin and eosin stain intensities have been transformed with \(M(x, h, e)\), where \(M\) denotes the Macenko method of separating stains, and \(h\) and \(e\) magnitudes are denoted by the y-axis and x-axis labels.

are offered through the torchvision library and are thus likely to be used by other researchers.

SA applies a varying number of transformations sequentially to an image \(x\). Each transformation \(t\) and its magnitude \(m\) is sampled uniformly from an augmentation space \(A\), and after each transformation \(t(x, m)\), the transformation \(t\) is removed from \(A\). After \(x\) has been transformed two times, each consecutive transformation is applied with probability \(p\), or until \(x\) has been transformed five times. Only one affine transformation is allowed per image, to allow significantly higher number of transformations than previous state-of-the-art methods \cite{24, 25, 29}. The augmentation space \(A\) contains a specified list of transformations \(t\) and a range of possible magnitudes \(m\) for the transformation. RA, TA and transformations implemented in the torchvision (1.11.1) library are used, except for the colour channel intensity transformations (Red, Green, Blue) which have been implemented by us. For RA, the magnitude and number of operations are set to \(m = 10\) and \(n = 2\) \cite{26}, and for TA there are no tunable hyper-parameters.
2.3 Training details

ResNet-RS50 [20] neural networks are used for all the experiments in this study. Dropout [30] and stochastic depth [31] are set to 0.2 for all networks. For training, a random area of the image is cropped and resized to $160 \times 160$ pixels with either bi-linear, bi-cubic or nearest interpolation, and flipped horizontally and/or vertically with a probability of 0.5. Then the image is transformed with either SA, RA or TA, a random area of the image is erased with an 0.2 probability, and either Mixup [32] or Cutmix [33] is applied for pair of images in a batch with an alpha value of 0.2. For testing, the images are resized to $224 \times 224$ pixels with bi-linear interpolation.

Adaptive sharpness aware minimisation [34] is used with $\rho = 0.05$ and Adam [35] with $\beta_1 = 0.9, \beta_2 = 0.999$ as the base optimiser. Learning rate is increased linearly from $10^{-6}$ to $0.0004 \times \text{batch\_size}/256$ during the first 10 epochs, and then reduced back to $10^{-6}$ with a cosine schedule. Label smoothing is set to 0.1 [36]. All networks are trained for 150 epochs with a batch size of 1024, and each experiment is repeated 5 times.

Spectral decoupling [21] is used instead of weight decay, which has been shown to increase neural networks’ robustness to data distribution shifts [12]. Spectral decoupling coefficient is set to 0.0001 and weight decay to 0. For experiments, where spectral decoupling is intentionally left out to assess its effect on neural networks’ robustness, we set weight decay to 0.00004.

PyTorch (1.10.0) [37] is used for training the neural networks, and ResNet-RS50, Mixup and Cutmix implementations are from the torchvision (0.5.4) library [38].

2.4 Evaluation metric

To evaluate the performance of a neural network in a given dataset, the area under the receiver operating characteristic curve (AUROC) is reported for each experiment. AUROC measures the ability to discriminate between positive and negative samples. The AUROC value corresponds to the proportion of positive and negative samples. We denote this dataset as Helsinki60.

A total of 30 prostate cancer patients’ full glass slide sets from surgical specimens are annotated for classification into cancerous and benign tissue, where the cancerous areas were annotated in consensus by two observers (C.S., T.M.). All patients have undergone radical prostatectomy at the Helsinki University Hospital between 2014 and 2015. Each case contains 14 to 21 tissue section slides. Tissue sections have a thickness of 4 $\mu$m and were stained with hematoxylin and eosin in a clinical-grade laboratory at the Helsinki University Hospital Diagnostic Center, Department of Pathology. Two different scanners are used to obtain images of the tissue section slides at 20x magnification. Larger macro slides (whole-mount, 2x3 inch slides) are scanned with an Axio Scan Z.1 scanner (Zeiss, Oberkochen, Germany), and the normal size slides with a Pannoramic Flash III 250 scanner (3DHistech, Budapest, Hungary). From the 30 patient cases, seven are set aside for a validation set. Digital slide images are cut into tiles with $1024 \times 1024$ pixels and 20% overlap, resulting in 4.7 million tiles with 10% containing cancerous tissue. For training, we selected all cancerous tiles and sample randomly the same amount of benign tiles to have equal amounts of positive and negative samples. We denote this dataset as Helsinki30.

Another set of 60 radical prostatectomy slides is also annotated into cancerous and benign tissue by one of six experienced pathologists. All patients have undergone radical prostatectomy at the Helsinki University Hospital between 2019 and 2020. Each case contains 10 to 21 normal and macro tissue section slides of the prostate. Tissue sections have a thickness of 4 $\mu$m and are also stained with hematoxylin and eosin in a clinical-grade laboratory at the Helsinki University Hospital Diagnostic Center, Department of Pathology. All slides are scanned with an Axio Scan Z.1 scanner. From the 60 patient cases, seven are set aside for a validation set. Digital slide images are cut into $1024 \times 1024$ pixel tiles with 20% overlap, resulting in 13.1 million tiles with 16% containing cancerous tissue. For training, we selected all cancerous tiles and sample randomly the same amount of benign tiles to have equal amounts of positive and negative samples. We denote this dataset as Helsinki60.

As external datasets, four freely available prostate cancer datasets are used. These publicly available datasets yield five external datasets from three different countries.

The PANDA development dataset contains 10616 prostate biopsy slides from 2113 patients from Radboud University Medical Center between 2012 and 2017 and Karolinska Institutet (Stockholm, Sweden) between 2012 and 2014 [28]. The biopsy slides from Radboud University Medical Center are scanned with a Pannoramic Flash II 250 (3DHistech, Budapest, Hungary) scanner and the slides from Karolinska Institutet with either Hamamatsu C9600-12 (Hamamatsu Photonics, Hamamatsu, Japan) or Aperio ScanScope AT2 (Leica Biosystems, Wetzlar, Germany) scanners. The annotations for the biopsy slides from Radboud University Medical Center are determined on the pathology reports from routine clinical practice, and the biopsy slides from Karolinska Institutet from routine clinical workflow of one uropathologist. The digital slide images are cut into $512 \times 512$ pixel tiles with 20% overlap resulting in 609180 (48% cancer) and 780694 (23% cancer) tiles from Radboud University Medical Center and Karolinska Institutet, respectively. Tile images were labelled as cancerous if there was any overlap with the annotations. Tiles with pen markings were excluded.

2.5 Datasets

Seven different datasets from two tissue types, originating from four countries are used in this study. A basic summary of the datasets is presented in Table 2.
Validation splits are omitted for both datasets as patient identifiers are missing from the publicly available data. Thus, the validation split might contain slides from patients also in the training split. We denote the datasets from Radboud University Medical Center and Karolinska Institutet as Radboud and Karolinska, respectively.

The PESO dataset contains tissue section slides from patients who have undergone a radical prostatectomy at the Radboud University Medical Center (Nijmegen, the Netherlands) between 2006 and 2011 \([39, 40]\). The dataset contains images with 2500 \(\times\) 2500 pixels annotated by a uropathologist as either cancerous or benign. Slides are scanned with a Pannoramic Flash II 250 scanner (3DHistech, Budapest, Hungary) at 20x magnification but later reduced to 10x magnification. These images are cut into 512 \(\times\) 512 pixel tiles with 20% overlap, resulting in 5655 tiles with 45% containing cancerous tissue.

The Gleason2019 dataset contains 333 prostate tissue microarray spots from 231 patients who had undergone radical prostatectomy at Vancouver General Hospital between 1997 and 2011 \([41]\). The slides are scanned with a SCN400 Slide Scanner (Leica Microsystems, Wetzlar, Germany). All slides are annotated by four to six pathologists with 1 to 27 years of experience. There are considerable disagreements between the pathologists, and thus a pixel-wise majority vote is used as the ground truth label. These digital slide images are cut into 224 \(\times\) 224 pixel tiles with 20% overlap resulting in 94230 tiles with 88.5% containing cancerous tissue.

To evaluate augmentation strategies on training data from a different tissue type, a dataset was compiled from 167 patients with renal cell carcinoma (clear cell). All patients had undergone radical nephrectomy at the Helsinki University Hospital between 2006 and 2013. Tissue sections were stained with hematoxylin and eosin in a clinical-grade laboratory at the Helsinki University Hospital Diagnostic Center, Department of Pathology. A total of 698 cancerous and 172 benign tissue microarray spots were gathered from the sections and scanned with a Pannoramic Flash III 250 scanner (3DHistech, Budapest, Hungary). Of the 167 patients, 24 were set aside for a validation set. Scanned tissue microarray spots were cut into tiles with 384 \(\times\) 384 pixels and 20% overlap, resulting in 2.0 million tiles with 21% containing cancerous tissue. We denote this dataset as HelsinkiRCC.

All digital slide images are processed with the HistoPrep library \([42]\) (1.0.7).

### 2.5.1 Label noise

Each dataset used in this study has been annotated with a different strategy, and thus contains different amounts of label noise. For this reason, the network performances on different datasets are not directly comparable. For example, a perfect classifier would achieve 100% accuracy on a clean dataset, but only 80% accuracy on a dataset with 20% label noise. In reality, there are no perfect classifiers \([43]\) and the decreased performance on an external dataset may also be caused by the fact that the classifier cannot generalise to the dataset. For this reason, the amount of label noise in a dataset should be estimated.

To estimate the amount of label noise in a given dataset, a neural network is trained on the dataset, and the training discrimination performance is reported. Due to overfitting, the estimate of label noise in the dataset can easily be to be deflated. To minimise overfitting, we use strong regularisation outlined in Section 2.5.5 with SA. To evaluate the accuracy of this estimation method, label noise is introduced to the Helsinki30 dataset, which contains negligible amounts of label noise. After flipping 0, 5, 10 and 20% of the labels in the Helsinki30 dataset, the estimated label noise is 0.03, 5.5, 10.21 and 20.16%, respectively.

The estimated amount of label noise in the Helsinki30, Helsinki60, Karolinska, Radboud, Gleason2019 and HelsinkiRCC datasets are 0.03%, 17.0%, 9.6%, 3.9%, 14.7% and 8.1%, respectively.

### 3 Results

#### 3.1 Robustness to data distribution shifts

In this section, we evaluate neural networks’ robustness to data distribution shifts with 193 distribution-shifted datasets created from the PESO dataset. For each experiment, five neural networks are trained on the Helsinki30 with either SA, RA or TA while keeping all other hyper-parameters fixed.

First, the general robustness to data distribution shifts is evaluated using the 94 distribution-shifted datasets, described in Section 2.1.1 with example images presented in Figure 2. Discrimination performances of the trained neural networks on these datasets are presented in Figure 4. Neural networks trained with either RA or TA are highly sensitive to even small data distribution shifts, which can be seen by the large performance differences between the original and distribution-shifted datasets. Neural networks trained with SA retain the discrimination performance on all datasets, even in situations where networks...
Second, neural networks’ robustness to data distribution shifts caused by differences in the haematoxylin and eosin stain intensities are evaluated using the 81 distribution-shifted datasets described in Section 2.1.2, with example images presented in Figure 3. Neural networks trained with either RA or TA are again highly sensitive to distribution shifts caused by changes in the haematoxylin and eosin stain intensities, and the discrimination performance quickly degrades with even small distribution shifts. Neural networks trained with SA retain their discrimination performance on all distribution-shifted datasets, even with several datasets where networks trained with RA or TA have degraded to random discrimination. The only exception is the dataset $X_{\text{brightness}, 0.05}$, which contains almost black images. The performance decreases for networks trained with RA and TA are unacceptable with most of the datasets, especially when considering that the decrease does not happen with networks trained with SA. We recommend comparing the discrimination performances in Figure 4 to the example images in Figure 2.

3.2 Performance on real-world datasets

To assess, whether better performance on the distribution-shifted datasets translates to better performance on real-world datasets, we evaluate the trained neural networks on the Helsinki30, Helsinki60, Karolinska, Radboud [28], PESO [39] and Gleason2019 [41] datasets. To demonstrate that SA has not just been over-fitted to the Helsinki30 dataset used in Section 3.1, we also train and evaluate neural networks on the Helsinki60, Karolinska and Radboud datasets.

The discrimination performances of neural networks trained with SA, RA or TA are presented in Figure 6. Each trained neural network is evaluated with five external datasets. Columns denote the training dataset and rows the evaluation dataset. Performances are measured on the validation split for Helsinki30 and Helsinki60, and on the whole dataset with Karolinska, Radboud, PESO and Gleason2019. Metrics where the training and evaluation datasets are the same, have been denoted by a red box.
Figure 5: Robustness of neural networks, trained on the Helsinki30 dataset with either SA, RA or TA, to distribution shifts caused by differences in the haematoxylin and eosin stain intensities. Each cell represents a given distribution-shifted dataset \(X_{(M, h, e)}\), where \(h\) and \(e\) denote the magnitudes for haematoxylin and eosin stains, respectively. Cells are annotated with AUROC×100 value, rounded to the closest integer. Neural networks trained with SA retain performance on all datasets, even when networks trained with RA or TA have degraded to random discrimination. Supplementary Figure S2 presents the same results with the use of weight decay instead of spectral decoupling.

Figure 6: Discrimination performance of neural networks trained with SA, RA or TA on four different training datasets. Each network is evaluated with five external datasets and the training dataset, denoted by a red box. Neural networks trained with SA achieve better or comparable performance on every evaluation dataset, whereas RA and TA demonstrate a significant lack of generalisation ability with several evaluation datasets. Please note that the y-axes of Helsinki60, Karolinska and Gleason2019 evaluation datasets have been truncated to account for the label noise in these datasets.
Neural networks trained with SA achieve better or comparable results on every external validation dataset than current state-of-the-art methods. Networks trained with RA or TA outright fail to generalise to several of the external datasets, and in many cases, these same networks achieve good results on other external validation datasets. For example, networks trained on the Helsinki30 dataset with RA perform only slightly worse than SA on the Helsinki60, Karolinska, Radboud and PESO datasets, but then achieve unacceptably low discrimination performances of 0.628 to 0.692 on the Glaesone2019 dataset. Without the Glaesone2019 external validation dataset, the failure to generalise would have been missed, whereas the network fragility to data distribution shifts was easily discovered with shifted evaluation in Section 3.1.

Neural networks trained on the datasets with more label noise achieve noticeably worse performance with all augmentation methods. Networks trained with SA still maintain an acceptable performance on all evaluation datasets, whereas several networks trained with the other methods degrade to no better than random discrimination performance. This is most apparent with the Karolinska and Radboud training datasets, where SA improves the mean discrimination performance over RA and TA by 0.116 to 0.196 on the Helsinki30 dataset, and by 0.099 and 0.181 on the Helsinki60 dataset.

In addition to improved performance, with SA there is a significantly lower variance in the discrimination performances across the five trained networks, whereas the difference between the best and worst-performing networks trained with RA and TA is as high as 0.132 and 0.121, respectively.

3.3 Effect of spectral decoupling

As spectral decoupling has been shown to increase robustness to data distribution shifts \[12\], five neural networks are trained on the Helsinki30 dataset with weight decay instead of spectral decoupling. The neural networks are then evaluated on the 99 distribution-shifted datasets created in Section 2.1.2.

The discrimination performances of the trained neural networks are presented in Supplementary Figure S2. Networks trained with SA achieve consistent results even without spectral decoupling, although the result is not as good as in Figure 5. Spectral decoupling is much more crucial for RA and TA, where the use of weight decay decreases the AUROC values by up to 0.25 AUROC. Based on the results, spectral decoupling is complementary with SA and crucial for training robust neural networks.

3.4 Other tissues

As all of the above results have been trained on scanned images of prostate sections, five neural networks are trained with SA, RA or TA on the HelsinkiRCC training split to evaluate whether SA achieves similar results on other tissue types. The neural networks are evaluated on 81 distribution-shifted datasets created from HelsinkiRCC validation split by adjusting haematoxylin and eosin stains similarly to Section 2.1.2.

The discrimination performances of the trained neural networks are presented in Supplementary Figure S3. Neural networks trained with SA achieve consistent performance on 80 of the 81 distribution-shifted datasets, whereas the performance of networks trained with RA and TA degrades quickly with even small distribution shifts. We see no reason why SA could not achieve similar results on other tissue types as well.

4 Discussion

Neural network evaluation is crucial for understanding the clinical usefulness, safety and accuracy of neural networks intended for clinical practice \[11\]. Nevertheless, there are no methods which could exhaustively evaluate the developed neural networks. Currently, neural networks are typically validated in two steps. During model development, neural network training is monitored with a held-out split of the training data, the validation split. Differences in performance between the training and validation split may indicate overfitting and suggest using for example stronger regularisation. After model development, the performance is evaluated with internal and/or external validation, where performance is measured on another dataset from the same or different source. This is done to assess the performance of new situations, and possibly uncover problems with generalisability.

There are several limitations to the currently used validation methods. During model development, the training and validation split often originate from the same dataset, and thus there can still be significant overfitting to the underlying dataset. The same is true with internal validation after model development, where overfitting to the data source would not be discovered. Finally, external validation is limited to uncovering potential problems, and good performance cannot establish clinical usefulness, robustness or generalisation ability of a model \[13\].

Shifted evaluation offers an easy solution to all these limitations. By shifting the data distribution of a dataset to known directions, neural networks’ robustness to these distribution shifts can be assessed. This way, it is possible to "expand" internal or external datasets, which can be then used to more thoroughly validate the network. For example, in Figure 6 all networks trained on the Radboud dataset all achieve above 0.95 AUROC values with the training dataset and the external PESO dataset. Without access to the other external datasets, such as Helsinki30 and Helsinki60 where networks trained with RA and TA achieve unacceptably low discrimination performances, a researched would have no way of knowing which of the networks would fail to generalise. Alternatively, by “expanding” the PESO dataset by creating distribution-shifted datasets described in Section 2.1.1, the networks can be validated more thoroughly. When exploring performance on these distribution-shifted datasets, demonstrated in Supplementary Figure S4, it is easy to conclude that networks trained with RA and TA are extremely fragile to data distribution shifts, and thus would not most likely generalise to other external datasets. Most importantly, shifted evaluation tells us exactly what the reason is for the decreased performance, and the reason can be addressed.

We believe that due to the lack of adequate evaluation methods, the role of augmentation as a method of increasing neural networks’ robustness to data distribution shifts has been largely unexplored. In medical imaging, some work has been done with
training stain-invariant neural networks through augmentation \cite{22}. A comparison of different stain augmentation and normalisation methods concluded that augmentation is crucial for achieving good performance, but found slightly better results with light rather than strong augmentation \cite{20}. As the results were based on external validation performance and augmentation was coupled with stain normalisation, it is likely that the fragility of neural networks was simply not discovered. With natural images, the development of augmentation methods has focused on improving performance on the validation split of the ImageNet \cite{44} dataset \cite{23,24,25,29}. Thus, light augmentation has been preferred over strong, which can reduce performance on in-distribution datasets, although the reduction in performance is often dwarfed when compared to the performance decrease on external datasets \cite{14} and Figure 6.

Neural networks trained with SA achieve consistent results with out-of-distribution datasets, even in cases where current state-of-the-art methods have degraded to random discrimination. Often the decrease in performance with current state-of-the-art methods is unacceptably high, especially when considering that the drop in performance does not happen with SA. In a clinical setting, it should be reasonable to expect that a neural network performs consistently, for example with blurry images, and not experience a 0.276 and 0.324 drop in AUROC as happens respectively with RA and TA (X_{Blur,7} dataset in Figure 7).

Although all augmentation methods contain several of the transformations used to create the distribution-shifted datasets in Section 2.1.1 there are large differences in discrimination performance as seen in Figure 4. TA even has a wider augmentation space for Brightness and Contrast transformations (Table 1), but the networks trained with SA achieve up to 0.45 better AUROC values in Figures 4 and 4H. We speculate that the combination of a wide augmentation space coupled with a significantly higher number of transformations per image has made neural networks trained with SA robust to all data distribution shifts, not just the ones included in the augmentation space. This is supported by the fact that networks trained with SA achieve consistent results when varying the haematoxylin and eosin stain intensities in Figure 5 although these distribution shifts are not included in the augmentation space.

Although neural networks trained with SA achieve unprecedented results, it does not solve the generalisation problem on its own. Other methods which increase neural networks’ robustness to data distribution shifts should be used. For example, spectral decoupling has been shown to increase robustness to data distribution shifts \cite{12}, and it is highly complementary with SA as seen in Supplementary Figure S2. Also, high-quality and representative datasets are crucial for training neural networks intended for clinical practice. This is demonstrated in Figure 6 where datasets with more label noise achieve lower discrimination performance on external validation. With SA this gap is significantly smaller, emphasising the usefulness of strong augmentation on lower quality datasets. With SA, the collection of representative training datasets can now focus on biological representativeness, instead of using multiple different scanners or staining pipelines.

5 CONCLUSIONS

In this study, we have exposed the fragility of neural networks using distribution-shifted datasets. Shifted evaluation offers an easy way to thoroughly evaluate a neural network, while also providing an exact reason in case of a potential failure. To address the exposed fragility of neural networks, we developed an augmentation strategy to explicitly train neural networks to be robust to data distribution shifts. StrongAugment achieves consistent results on out-of-distribution datasets, even when previous state-of-the-art methods have degraded to no better than random performance. We recommend using strong augmentation and distribution-shifted datasets to train and evaluate all future neural networks intended for clinical practice.

ACKNOWLEDGEMENTS

This work was supported by Cancer Foundation Finland [304667, 191118], Jane and Aatos Erkko Foundation [290520], Academy of Finland [322675] and Hospital District of Helsinki and Uusimaa [TYH2018214, TYH2018222, TYH2019235, TYH2019249]. The authors also wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

ETHICS STATEMENT

The images of the tissue slides are applied in this study based on national legislation and a research permission from the Helsinki University Hospital (§105).

DECLARATION OF COMPETING INTERESTS

The authors have no interests to declare.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Joona Pohjonen: Conceptualization; Formal analysis; Visualization; Software; Data curation; Roles/Writing - original draft
Carolin Stürenberg: Data curation; Writing - review & editing.
Atte Föhr: Data curation, Formal analysis Esa Pitkänen: Funding acquisition; Resources; Supervision; Writing - review & editing.
Antti Rannikko: Data curation; Formal analysis
Aatos Erkko Foundation [290520], Academy of Finland [322675] and Hospital District of Helsinki and Uusimaa [TYH2018214, TYH2018222, TYH2019235, TYH2019249]. The authors also wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

ACKNOWLEDGEMENTS

This work was supported by Cancer Foundation Finland [304667, 191118], Jane and Aatos Erkko Foundation [290520], Academy of Finland [322675] and Hospital District of Helsinki and Uusimaa [TYH2018214, TYH2018222, TYH2019235, TYH2019249]. The authors also wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

ETHICS STATEMENT

The images of the tissue slides are applied in this study based on national legislation and a research permission from the Helsinki University Hospital (§105).

DECLARATION OF COMPETING INTERESTS

The authors have no interests to declare.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Joona Pohjonen: Conceptualization; Formal analysis; Visualization; Software; Data curation; Roles/Writing - original draft
Carolin Stürenberg: Data curation; Writing - review & editing.
Atte Föhr: Data curation, Formal analysis Esa Pitkänen: Funding acquisition; Resources; Supervision; Writing - review & editing.
Antti Rannikko: Funding acquisition; Resources; Supervision; Writing - review & editing.
Tuomas Mirtti: Data curation; Funding acquisition; Resources; Supervision; Writing - review & editing.

REFERENCES

[1] Manuel Salto-Tellez, Perry Maxwell, and Peter Hamilton. Artificial intelligence-the third revolution in pathology. Histopathology, 74(3):372–376, 2019.
[2] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from...
medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

[3] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflo, Vitor Wernack Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[5] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):1–10, 2019.

[6] Thomas de Bel, Meyke Hermsen, Jesper Kers, Jeroen van der Laak, and Geert Litjens. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *International Conference on Medical Imaging with Deep Learning–Full Paper Track*, 2018.

[7] Yun Liu, Timo Kohlberger, Mohammad Norouzi, George E Dahl, Jenny L Smith, Arash Mohtashamian, Niels Olson, Lily H Peng, Jason D Hipp, and Martin C Stumpe. Artificial intelligence-based breast cancer nodal metastasis detection: Insights into the black box for pathologists. *Archives of pathology & laboratory medicine*, 143(7):859–868, 2019.

[8] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[9] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[10] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.

[11] Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[12] Joona Pohjonen, Carolin Stürenberg, Antti Rannikko, Tommaso Mirtti, and Esa Pitkänen. Spectral decoupling for training transferable neural networks in medical imaging. *iScience*, page 103767, 2022.

[13] Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John PA Ioannidis, Gary S Collins, and Mahiben Maruthappu. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Bmj*, 368, 2020.

[14] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1):1–8, 2022.

[15] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[16] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. pages 1107–1110, 2009.

[17] Andrew Janowczyk, Ajay Basavanahalli, and Anant Madabhushi. Stain normalization using sparse autoencoders (stanosa): application to digital pathology. *Computerized Medical Imaging and Graphics*, 57:50–61, 2017.

[18] Hyungjoo Cho, Sungbin Lim,Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using gan for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.

[19] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Dingyi Hu, Shujiao Sun, Jun Shi, and Chenghai Xue. Stain standardization capsule for application-driven histopathological image normalization. *IEEE journal of biomedical and health informatics*, 25(2):337–347, 2020.

[20] David Tellez, Geert Litjens, Péter Bándi, Wouter Bul ten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[21] Mohammad Pesheski, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Whole-slide mitosis detection in h amp:e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, 2018.

[23] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[24] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.
[25] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 774–782, 2021.

[26] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. Advances in Neural Information Processing Systems, 34, 2021.

[27] Digital compression and coding of continuous-tone still images: Requirements and guidelines, 2 1992.

[28] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuan-nan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine, pages 1–10, 2022.

[29] Tom Ching LingChen, Ava Khonsari, Amirreza Lashkari, Mina Rah Nazari, Juspreet Singh Sambee, and Mario A Nascimento. Uniformaugment: A search-free probabilistic data augmentation approach. arXiv preprint arXiv:2003.14348, 2020.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958, 2014.

[31] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In European conference on computer vision, pages 646–661. Springer, 2016.

[32] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.

[33] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[34] Wouter Bulten, Péter Bándi, Jeffrey Hoven, Rob van de Loo, Johannes Lotz, Nick Weiss, Jeroen van der Laak, Bram van Ginneken, Christina Hulsbergen-van de Kaa, and Geert Litjens. Peso: Prostate epithelium segmentation on h&e-stained prostatectomy whole slide images, 2018. Accessed: 05.03.2021.

[35] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett De-lahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. The American journal of surgical pathology, 40(2):244–252, 2016.

[36] Davood Karimi, Guy Nir, Ladan Fazli, Peter C Black, Larry Goldenberg, and Septimiu E Salcudean. Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. IEEE journal of biomedical and health informatics, 24(5):1413–1426, 2019.

[37] Joona Pohjonen and Valeria Ariotta. Histoprep: Preprocessing large medical images for machine learning made easy! https://github.com/jopo666/HistoPrep, 2021.

[38] David H Wolpert and William G Macready. No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1):67–82, 1997.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.
**Supplementary Figures**

**Figure S1:** Example of augmented images using SA with $p = 0.4$ (A), RA with $m = 10$ and $n = 2$ (B), and TA with (C). Images augmented with SA differ significantly more from each other than with either of the other methods.

**Figure S2:** Robustness of neural networks, trained on the Helsinki30 dataset with weight decay instead of spectral decoupling, to distribution shifts caused by differences in the haematoxylin and eosin stain intensities. Each cell represents a given distribution-shifted dataset $X(m_h, m_e)$, where $m_h$ and $m_e$ denote the magnitudes for haematoxylin and eosin stains, respectively. Even without spectral decoupling SA achieves good results on all datasets (A), whereas spectral decoupling is more important to RA and TA (B-C).
Figure S3: Robustness of neural networks, trained on the HelsinkiRCC dataset to distribution shifts caused by differences in the haematoxylin and eosin stain intensities. Each cell represents a given distribution-shifted dataset $X_{(m_h, m_e)}$, where $m_h$ and $m_e$ denote the magnitudes for haematoxylin and eosin stains, respectively. SA achieves consistent discrimination performance on 80 of the 81 datasets (A), whereas the performance of RA and TA quickly degrade with even small distribution shifts (B-C). The results with scanned sections of kidney are similar to ones of prostate, indicating that SA can achieve similar results on any tissue type.
Figure S4: Robustness of neural networks, trained on the Radboud dataset with either SA, RA or TA, to distribution shifts caused by common image transformations. The lines show the mean AUROC values and shaded regions one standard deviation around the mean from five trained networks. The distribution-shifted datasets $X_{(t,m)}$ are represented by dots, where the title denotes the transformation $t$ and x-axis the magnitude $m$. The original PESO dataset $X$ is represented by dotted red lines. It is clear that the neural networks trained with RA or TA are extremely fragile to distribution shifts, especially to colour changes, which is not apparent from the performance on the original PESO evaluation dataset, denoted by the dotted red lines.