On the Robustness of Pretraining and Self-Supervision for a Deep Learning-based Analysis of Diabetic Retinopathy

Vignesh Srinivasan, Nils Strodthoff, Jackie Ma,
Alexander Binder, Klaus-Robert Müller, Wojciech Samek

Abstract

Objective: There is an increasing number of medical use-cases where classification algorithms based on deep neural networks reach performance levels that are competitive with human medical experts. To alleviate the challenges of small dataset sizes, these systems often rely on pretraining. In this work, we aim to assess the broader implications of these approaches. Methods: For diabetic retinopathy grading as exemplary use case, we compare the impact of different training procedures including recently established self-supervised pretraining methods based on contrastive learning. To this end, we investigate different aspects such as quantitative performance, statistics of the learned feature representations, interpretability and robustness to image distortions. Results: Our results indicate that models initialized from ImageNet pretraining report a significant increase in performance, generalization and robustness to image distortions. In particular, self-supervised models show further benefits to supervised models. Conclusion: Self-supervised models with initialization from ImageNet pretraining not only report higher performance, they also reduce overfitting to large lesions along with improvements in taking into account minute lesions indicative of the progression of the disease. Significance: Understanding the effects of pretraining in a broader sense that goes beyond simple performance comparisons is of crucial importance for the broader medical imaging community beyond the use-case considered in this work.

*VS, NS, JM and WS are with the Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany (email: firstname.lastname@hhi.fraunhofer.de). NS and WS are also with BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany. AB is with the Department of Informatics, Oslo University, 0373 Oslo, Norway. KRM is with the Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany, and also with the Department of Artificial Intelligence, Korea University, Seoul 136-713, South Korea, the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany, and BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany. (e-mail: klaus-robert.mueller@tu-berlin.de).

1 Introduction

The role of computer vision algorithms based on deep learning in medical imaging in the form of decision support systems has increased steadily in the past few years [1–7]. There is an enormous amount of data that is being produced on a daily basis from different areas using different imaging modalities such as MRI, CT, microscopy, etc., leading to an unprecedented potential for machine learning algorithms. However, while there exists a lot of data, it is usually not prepared to be used for research in machine learning. In particular, it is often unlabeled as the labeling process is expensive and time-consuming, or sometimes medical experts may not agree on the appropriate label.

A practitioner using Deep Neural Networks (DNN) for the task of medical imaging, is faced with a plethora of options when it comes to the training methodology for the DNN. Several factors can influence the decision making process including, but not limited to the size, noise level and quality of the dataset at hand, computational resources available and robustness of the trained DNN. Transfer learn-

ing for medical imaging from models trained on natural images have been found to be beneficial for improvements in performance along with speeding up convergence [1, 8]. A straightforward way of utilizing transfer learning is to finetune a model that is initially trained on ImageNet [9] on the medical dataset.

Other common state-of-the-art methods in machine learning are *supervised-learning* methods, i.e. models that are trained with labeled data, opposed to other methods that require only some or even no labeled data such as *semi-supervised* or *self-supervised learning*. Fortunately, the field of self-supervised learning has recently advanced significantly [11–14], which gives rise to hope for a successful deployment of machine learning in medical applications without relying on overly large amounts of labeled data. A first result in this regard was obtained in [6, 15, 16] where the authors showed that pretraining using self-supervision helps to improve the models for chest x-ray classification [17], dermatology condition classification [18] and Covid-19 deterioration prediction [16].

With widespread adoption of transfer learning in medical imaging, it becomes essential to explore the differentiating features of the various training methodologies—supervised or self-supervised. While [1] observe the effects of pretraining in supervised learning on the speed of convergence and feature representations, [8] study the effects on the performance of pretrained models from ImageNet providing improvements on diverse datasets and the quality of the features learned. Despite the benefits of transfer learning, it has, however, remained unclear what transfer learning, especially with self-supervised learning actually exploits when making a prediction. For this (as we will see) simply looking at performance metrics like classification accuracy or area under the operating curve (AUC) is not sufficient. The potential advantages of using self-supervised methods over supervised methods for medical imaging beyond such performance metrics thus remain a challenging object of study.

In this contribution, we demonstrate for diabetic retinopathy (DR) as a particular medical imaging use case, that going beyond metrics of predictive performance is mandatory. We further analyze robustness to statistical variations of the data. Furthermore we validate previous results on smaller data sets which are of ubiquitous interest to practitioners in medical data science.

To this end, we perform a detailed study of what is being learned by the different training methodologies available to train a DNN for medical imaging. Broadly, the training methodologies will be categorized into two types:

- Fully supervised (FS)
- Self-supervised with contrastive learning (CL)

along with two types of initialization of the weights before training on the medical dataset:

- Initialization with no external data (IWNE)
- Initialization from ImageNet (IFI)

The focus of this paper is to study the effects of training the DNN using these strategies and evaluate the benefits. Our contributions are as follows:
1) We evaluate the performance of the four different training strategies: supervised and self-supervised models using models trained with or without using external data for pretraining for detecting diabetic retinopathy in retinal images. We find that IFI helps in achieving significant gain in performance, especially when a limited amount of the downstream (medical) labeled dataset is used. IFI-CL provides a further increase in performance.
2) Given that IFI is beneficial in terms of performance, we investigate what makes them better by analyzing the eigenvalue spread of the activations on the hidden layers. We find that the redefined conditioning number for the IFI models is lower than that of IWNE models for the initial layers that are important for learning diverse and effective feature representations from the input. IFI makes the eigenvalue spread of the activations of the first hidden layer broader implying wider range kernels firing for a given input. In both IWNE as well as IFI models, we show that CL achieves broader eigenvalue spread compared to its supervised counterparts.
3) Using explainability of DNNs, we investigate what the different models look at in the input for making a decision. With the help of ground-truth

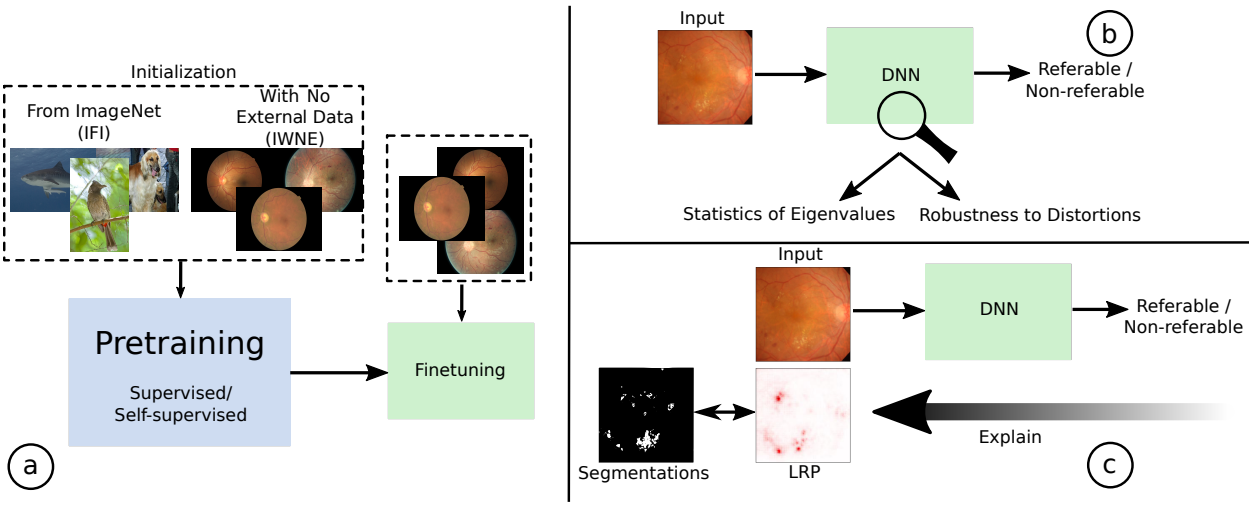


Figure 1: Overview of the experiments presented in this work. a) shows the different training strategies which includes pretraining and finetuning. b) investigates the statistics of the eigenvalues of the feature representations learned by the different methods which lead to increased robustness to distortions. c) shows the experiments we perform using the Indian Diabetic Retinopathy Image Dataset (IDRiD) challenge data [10] to quantitatively evaluate the cues learned.

segmentation maps available for diabetic retinopathy on the IDRiD challenge [10], we study in a quantitative manner what information was used by the models to make the prediction. We find that IWNE-FS overfits to large lesions like hard exudates and ignores smaller lesions to predict the disease. IFI models show significantly reduced tendency to overfit to one particular type of lesions. Especially IFI-CL is able to consider a wider range of lesions to make an accurate prediction for the disease.

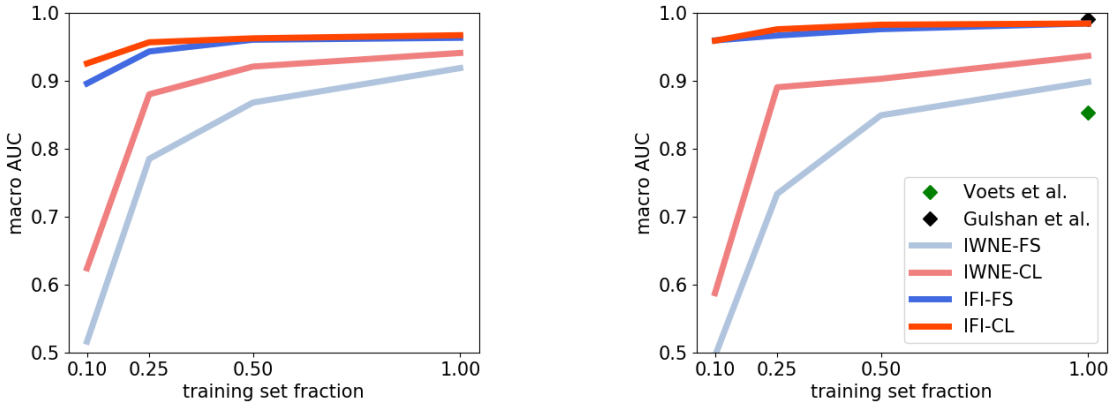
Dataset	# instances	# patients
EyePacs-1 (EyP) [40]	88,702	44,351
Messidor-2 [41]	1,744	872
IDRiD [10]	80	-

Table 1: Diabetic retinopathy datasets used for this study.

2 Related Work

Diabetic Retinopathy: DNN has seen wide adoption for the task of DR in [2, 3, 19–36] among others. While some methods train their model from scratch [19, 20, 34, 35, 37], IFI models have predominantly achieved higher performance [2,3,22,26,32,36]. Some methods also perform their training on large private data [2, 19, 23, 29, 30]. A reproduction study of [2] was performed by [3] showing difficulty in achieving similar performance for DR when trained on publicly available datasets. Systematic study of using uncertainty measures for DR were also conducted by [37,38]. While [21] studied the probability maps with ground-truth segmentation maps to ascertain what the DNN prediction was looking for, [39] studied a computer-assisted setting with explanation methods for deep learning models in grading for DR. There is, however, no dedicated study on the implications of different training methodologies.

Supervised vs. Self-supervised Learning: Self-supervised learning has been utilized in a wide range of biomedical applications including chest x-



(a) Training and evaluation on Eyepacs-1

(b) Training on Eyepacs-1 and evaluation on Messidor-2

Figure 2: Classification performance on Eyepacs-1 and Messidor-2 dataset for referrable DR as a function of fraction of the downstream training set used for training for four different training procedures. The state-of-the-art method for DR—Voets et al. [3] and Gulshan et al. [2] are shown as green and black diamonds for training with the full dataset for the Messidor-2 dataset.

rays [4–6, 16], diabetic retinopathy [42, 43], covid detection [16] etc. In spite of the improvements shown by self-supervised learning, [44] find that self-supervised models behave quite similarly to their supervised counterparts in many aspects of robustness. Self-supervised models report a slightly higher performance gain over their supervised counterparts on medical imaging [4, 6]. Recent works show the generalizing capabilities of self-supervised learning on chest x-rays [45]. The improvements and benefits still need to be rigorously investigated to ascertain the limits of using self-supervised learning on real-life healthcare applications.

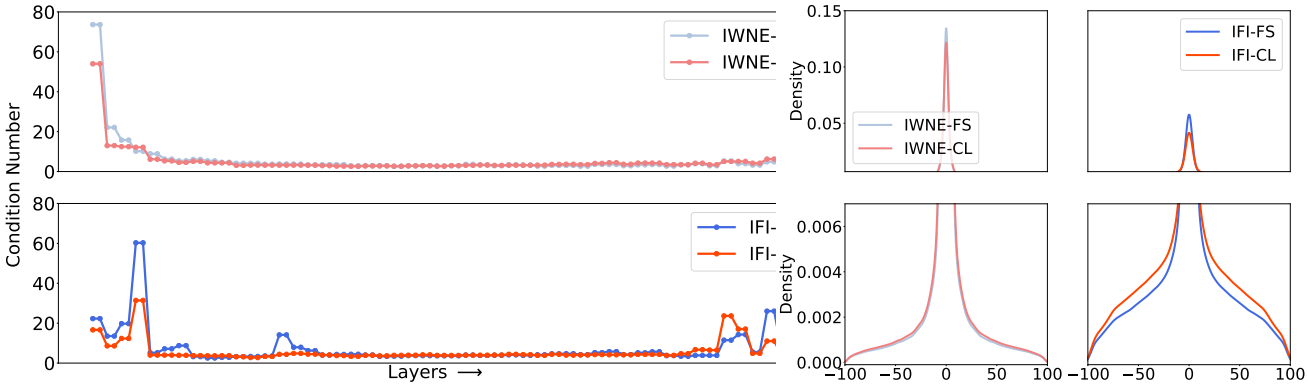
IWNE vs IFI Pretraining on ImageNet dataset (i.e. IFI), either supervised or self-supervised, is considered an effective strategy [4–6, 8, 46–51]. Several benefits have been attributed to pretraining including robustness [8,46–49], to generalization [52,53] to finding sparser subnetworks from the original [54] and to speed up convergence on the downstream task [1, 8]. Using IFI for DR has been widely adopted owing to benefits in performance [1–3, 22, 26, 35, 55]. The performance benefits of pretraining have been observed even on diverse datasets which seem distant from the ImageNet dataset [8]. The benefits of pretraining can be attributed to effective feature extracting capability of pretrained models in the lower layers [1,8]. Although, it is unclear how this translates to a DNN being used for a downstream task. While the above mentioned methods investigate supervised learning, we make a comparative study of IWNE vs IFI along with FS vs CL and their combinations to understand their differentiating features.

3 Materials & Methods

3.1 Datasets

We focus on diabetic retinopathy (DR) as a use case for our investigations and solely work on publicly available datasets, which are summarized in Table 1.

We make use of the Eyepacs-1 dataset [40], which is available from a former Kaggle challenge. The images are graded from a scale of 0 to 4 (0: no DR, 1: mild DR, 2: moderate DR, 3: severe DR, 4: proliferative DR) according to the International Clinical Diabetic Retinopathy (ICDR) severity scale. DR advances from a healthy eye to a proliferate one slowly and may also take years. However, this transition is



(a) Condition number for each layer of ResNet50 and for the different models.

(b) Eigenvalues of the activations of the first convolutional layer made symmetrical around 0 and plotted in the form of density for better visualization.

Figure 3: The statistics of the eigenvalues are shown here. a) shows the condition number of all the layers and b) shows the eigenvalues of the activations of the first convolutional layer.

discrete and often goes undetected to worsen into a proliferate DR. Hence, it is essential that this progression is detected and a timely medical diagnosis is performed. In our experiments, we train the models to perform the quinary classification using all the five grades. During inference, we modify it to a binary classification problem by considering classes $[0-2]$ as *healthy* and classes $[3-4]$ as *disease*. This binary class formulation is consistent with referable DR (rDR) classification in [2, 3].

The Eyepacs-1 dataset [40] consists of 35216 images in the training set and 53576 in the test set. We utilize non overlapping set of around 15% of the training set as the validation set. We train all our different methods on the training set of Eyepacs-1 dataset and evaluate the performance of the models on two datasets—test set of Eyepacs-1 and Messidor-2 [41]. Messidor-2 dataset [41] is a benchmark dataset consisting of 1744 images that are 100% gradable. Since the dataset is not used for training and was collected under different conditions at a different geographical location and with different hardware, the evaluation on the Messidor-2 dataset is supposed to measure the generalization performance of the algorithms. Hence, we use all the images of this dataset for testing. We report the AUC for the binary rDR classification task on the respective test sets of each dataset.

Method	Distribution	Parameters
IWNE-FS	Pareto	$\alpha = 1.45$
IWNE-CL	**Pareto**	$\boldsymbol{\alpha = 1.28}$
IFI-FS	Pareto	$\alpha = 0.87$
IFI-CL	**Pareto**	$\boldsymbol{\alpha = 0.73}$

Table 2: Distribution fitting for the eigenvalues of the activations of the first layer. For all the four models, the eigenvalues are best parametrized by a Pareto distribution. We also find that the self-supervised models show smaller value for the shape parameter of the Pareto distribution.

3.2 Models & Training Procedures

We compare the four training setups which are eventually trained on the DR target dataset.

- *Initialization With No External Data (IWNE)*
  - **FS**: supervised training on the DR dataset starting from randomly initialized weights.

- **CL**: self-supervised pretraining on the target domain and finetuning also on the same dataset using labeled data.

- *Initialization From ImageNet Data (IFI)*
  - **FS**: supervised training on the DR dataset starting from supervised ImageNet-pretrained weights.
  - **CL**: self-supervised pretraining on ImageNet dataset and finetuning on the DR dataset using labeled data.

For comparability, we fix the architecture and use a Resnet50 [56] model for all of our experiments. In the self-supervised setting, we pretrain the models using MoCoV2 strategy [57]. For the supervised pretraining, we use the ImageNet-pretrained model provided by torchvision. The IWNE models are trained for 500 epochs with a learning rate of 10^{-4}. Pretrained models have shown to be faster at convergence than the models trained from scratch [1,8]. Hence, we finetune the IFI models starting from ImageNet-pretrained weights for 50 epochs with a learning rate of 10^{-3}. The AdamW optimizer [58] with weight decay was used in all the settings. The best models in each training run was chosen based on the maximum AUC score achieved on the validation set and this model was used for inference on the test.

4 Experiments & Results

4.1 Quantitative performance

We evaluate the performance of the different methods discussed in Section 3.2 in terms of AUC. Each model was trained on the full dataset and on various fractions of the training set down to a fraction of 10% labeled samples. Figure 2 shows the final AUC of the binary classification for rDR. We find largely consistent results in terms of the ranking and overall behavior of the different training procedures between evaluation on a subset of the Eyepacs-1 dataset used for training and an evaluation on the external Messidor-2 dataset, which is a reassuring sign that our results generalize across datasets. The best-performing method across all training set fraction is IFI-CL, i.e. finetuning a model that was trained in a self-supervised fashion on ImageNet data, closely followed by IFI-FS, corresponding to the standard training methodology in medical imaging, where a model pretrained on ImageNet is finetuned on the target dataset. The results for the IWNE-CL model, i.e. self-supervised pretraining in target (DR) domain are weaker than the former two results. This trend is again followed at lower training set fractions where the model is trained with reduced fractions of the labeled dataset. While IWNE models deteriorate in performance, IFI models show only a marginal drop as shown in Figure 2.

The results clearly advocate the use of IFI models as opposed to not using external data, which is in line with most part of the medical imaging literature but at first sight contradicts [1], who found no improvements from IFI as compared to direct training on a considerably larger closed source DR dataset. The inferior results of IWNE-CL compared to IFI-CL can potentially be attributed to two factors: the size of Eyepacs-1 as pretraining is with around 30k samples, very small compared to large natural image datasets, such as ImageNet with 1.2M images, where self-supervised contrastive methods were demonstrated to work really well. In addition, for IWNE-CL we used the same set of transformations proposed for ImageNet in [12], which certainly represents a suboptimal choice for the DR images that differ qualitatively from natural images and the pretraining algorithm is rather sensitive to this choice.

4.2 Statistics of Eigenvalues

4.2.1 Condition number

To better understand what makes the IFI models achieve higher performance, we study the activations of the hidden layers. In particular, we compute the eigenvalues of the activations of each layer in the four models we considered. Using the eigenvalues, we plot the condition number [59] as shown in Figure 3a. To prevent the condition number from having very large values due to division by the minimum of the eigen-

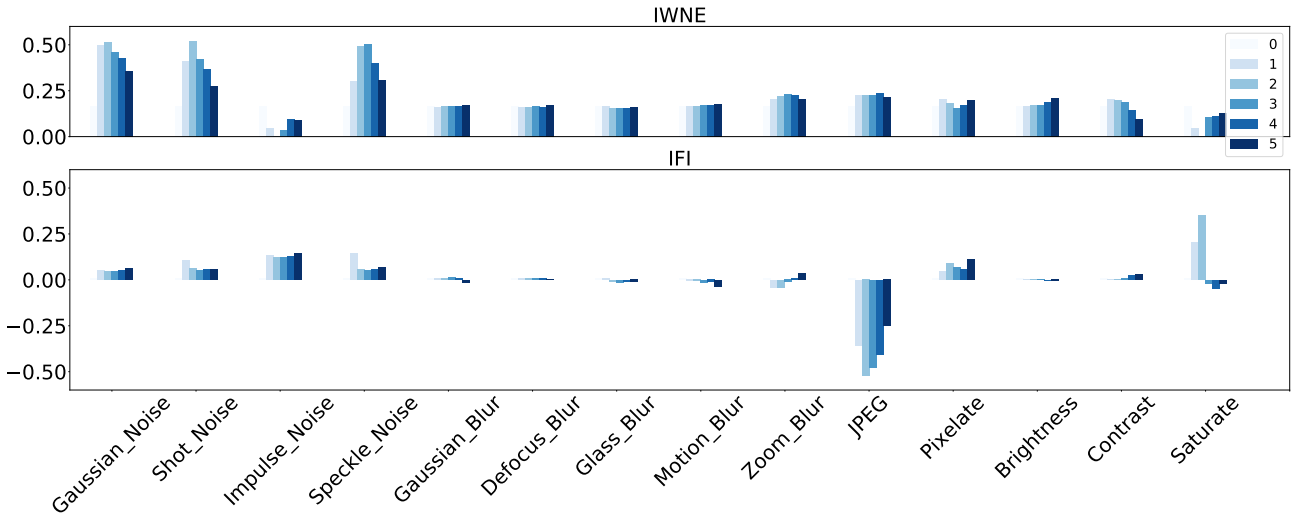


Figure 4: This figure shows the robustness to distortions for the different models. The difference in the softmax probabilities of the output between the CL and FS model is plotted here. The intensity of the color indicates the severity of the distortions. Top row shows the difference for IWNE models. Bottom row shows the difference for IFI models. In case of IWNE, the difference is consistently positive, implying that the self-supervised model has a higher prediction score than the plainly supervised model and thus exhibits a higher robustness to distortions. See Section 4.3 for a detailed discussion.

values, we define the condition number as follows:

$$\kappa(A) = \frac{|\lambda_{p_{99.9}}(A)|}{|\lambda_{p_{90}}(A)|} \tag{1}$$

where A are the activations of a hidden layer, $\kappa(A)$ is the condition number and $\lambda_{p_i}(A)$ is the eigenvalue corresponding to the ith percentile of the eigenvalues. While the top row in Figure 3a shows the condition numbers of the IWNE models, the bottom row shows the condition number of the IFI models. The x-axis in both the figures corresponds to the layers of ResNet50.

We find in Figure 3a that the condition number for IFI models is much lower than that of IWNE implying significantly more diverse features learned. Also, in both versions of initializations, we find that the condition number for self-supervised learning is lower than that of supervised learning in the initial layers. This indicates that self-supervised learning extracts more diverse features than its supervised counterparts. We also find in Figure 3a that for all the different models, the condition number is flattened out and becomes indistinguishable for the latter layers. The initial layers form the crux of the learning process extracting effective and diverse feature representations while the latter layers learn to aggregate these features.

4.2.2 Spread of Eigenvalues

To investigate the distinctive aspects of the initial layers, we plot the eigenvalues of the first layer for all four models in Figure 3b. The eigenvalues are made symmetrical around 0 and plotted in the form of density to make for better visualization. The bottom row in Figure 3b also zooms in on the tails. We find that the IWNE models obtain high and peaked eigenvalues. On the other hand, the IFI models, show lower peak values. Similar to the findings in the experiments on the condition number, self-supervised learning in contrast to supervised learning shows a slightly lower peak value. Additionally, in both versions

Lesions	Method	Pooling	RMA				RRA			
			Random		LRP-$\alpha_1\beta_0$		Random		LRP-$\alpha_1\beta_0$	
			Mean	Median	Mean	Median	Mean	Median	Mean	Median
Microaneurysms	IWNE-FS	sum_pos	0.0073	0.0064	0.0076	0.0072	0.9885	0.9895	0.3447	0.4146
		l2_norm_sq	0.0074	0.0067	0.0041	0.0025	0.9894	0.9900	0.4888	0.5047
	IWNE-CL	sum_pos	0.0074	0.0064	**0.0093**	**0.0077**	0.9879	0.9901	**0.4370**	**0.5224**
		l2_norm_sq	0.0073	0.0064	**0.0075**	**0.0042**	0.9882	0.9912	**0.5777**	**0.5736**
	IFI-FS	sum_pos	0.0073	0.0061	0.0172	0.0143	0.9913	0.9922	0.5097	0.5551
		l2_norm_sq	0.0074	0.0068	0.0374	0.0218	0.9891	0.9900	0.5705	0.5713
	IFI-CL	sum_pos	0.0073	0.0061	**0.0198**	**0.0186**	0.9896	0.9897	**0.5831**	**0.6251**
		l2_norm_sq	0.0073	0.0067	**0.0595**	**0.0381**	0.9902	0.9917	**0.6357**	**0.6366**
Haemorrhages	IWNE-FS	sum_pos	0.0234	0.0130	0.0251	0.0165	0.9902	0.9911	0.3845	0.4547
		l2_norm_sq	0.0233	0.0126	0.0139	0.0056	0.9880	0.9905	0.5414	0.5565
	IWNE-CL	sum_pos	0.0232	0.0126	**0.0602**	**0.0458**	0.9889	0.9904	**0.4971**	**0.6076**
		l2_norm_sq	0.0234	0.0125	**0.1063**	**0.0525**	0.9881	0.9892	**0.6357**	**0.6371**
	IFI-FS	sum_pos	0.0233	0.0126	0.0711	0.0578	0.9896	0.9895	0.5840	0.6127
		l2_norm_sq	0.0233	0.0125	0.1438	0.1243	0.9891	0.9904	0.6571	0.6551
	IFI-CL	sum_pos	0.0234	0.0127	**0.0765**	**0.0722**	0.9897	0.9911	**0.6898**	**0.7194**
		l2_norm_sq	0.0234	0.0125	**0.1874**	**0.1808**	0.9873	0.9911	**0.7403**	**0.7405**
Hard Exudates	IWNE-FS	sum_pos	0.0409	0.0195	**0.1954**	**0.1734**	0.9897	0.9906	0.5086	**0.6959**
		l2_norm_sq	0.0409	0.0190	**0.4201**	**0.4921**	0.9898	0.9903	**0.7114**	**0.7435**
	IWNE-CL	sum_pos	0.0409	0.0200	0.1206	0.1018	0.9889	0.9898	**0.5136**	0.5887
		l2_norm_sq	0.0409	0.0191	0.2338	0.1652	0.9892	0.9898	0.6656	0.7038
	IFI-FS	sum_pos	0.0408	0.0195	**0.1103**	**0.0861**	0.9895	0.9905	**0.5480**	**0.6258**
		l2_norm_sq	0.0410	0.0194	**0.2125**	**0.1659**	0.9905	0.9915	**0.6125**	**0.6561**
	IFI-CL	sum_pos	0.0409	0.0188	0.0725	0.0533	0.9890	0.9896	0.5425	0.5762
		l2_norm_sq	0.0412	0.0193	0.1195	0.0858	0.9888	0.9903	0.6088	0.6260
Total	IWNE-FS	sum_pos	0.0710	0.0558	**0.2266**	**0.2000**	0.9899	0.9905	0.4459	0.5655
		l2_norm_sq	0.0711	0.0563	**0.4363**	**0.5104**	0.9893	0.9898	0.6151	**0.6503**
	IWNE-CL	sum_pos	0.0710	0.0565	0.1887	0.1619	0.9884	0.9891	**0.5015**	**0.6083**
		l2_norm_sq	0.0711	0.0565	0.3457	0.3330	0.9884	0.9890	**0.6459**	0.6334
	IFI-FS	sum_pos	0.0709	0.0561	**0.1969**	**0.1886**	0.9893	0.9897	0.5479	0.5941
		l2_norm_sq	0.0711	0.0557	**0.3905**	**0.3964**	0.9893	0.9896	0.6150	0.6245
	IFI-CL	sum_pos	0.0711	0.0576	0.1671	0.1724	0.9893	0.9896	**0.5847**	**0.6144**
		l2_norm_sq	0.0713	0.0569	0.3625	0.3650	0.9895	0.9897	**0.6428**	**0.6463**

Table 3: Relevance mass accuracy (RMA) and relevance rank accuracy (RRA) on the LRP-$\alpha_1\beta_0$ explanation heatmaps of images of the IDRiD dataset. The results show that while supervised models overfit on the hard exudates, the self-supervised models look at diverse set of input features (lesions). On the other hand, we also find that IFI models show higher accuracies when compared to IWNE models.

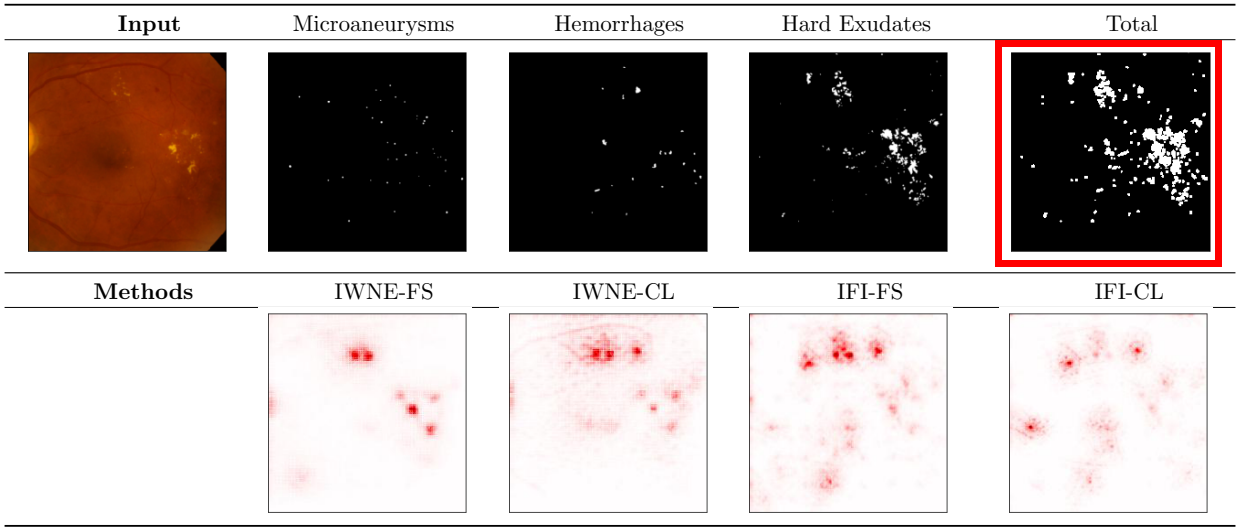


Figure 5: Top left image in the figure shows the input followed by the segmentation maps from the IDRiD dataset. Top right image is the total which we compute by combining the segmentation maps of different lesions. Bottom row shows the explanation heatmaps for the given input. Each explanation heatmap is correlated with the total image marked in red to evaluate the effectiveness of the model towards making the prediction for the disease. We find that IWNE-FS overfits on the hard exudates and also fails to pick up on cues related to microaneurysms. We also find that explanation heatmaps of IFI models show reduced signs of overfitting to a single lesion when compared to IWNE.

of the initialization, self-supervised learning models show more heavy tailedness.

The results indicate that IWNE models learn kernels in the first convolutional layer that are activated for some very specific patterns. On the contrary, IFI models learn kernels that activate for a broader range of input features. The superior performance of IFI models can be attributed to this effect while this may lead to several other benefits including increase in generalization and robustness.

4.2.3 Distribution Fitting

In this section, we fit the eigenvalues of the first convolutional layer to the parameters of several distributions and report the distribution that fits best [60]. Among a wide range parameterized distributions, we find in Table 2 that all the four models fit best to the *Pareto* distribution, though the parameters vary. Pareto distribution with the shape parameter $\alpha = 1.16$ corresponds to the $80 - 20$ rule, implying that 80% of the results come from 20% of the causes [61]. IWNE models show α values higher than 1.16. This indicates that the overall result comes from less than 20% of the activations. In other words, the kernels learned by the IWNE models extract small number of, yet highly curated set of features from the input. In contrast, we find that IFI brings down the value of α for the Pareto distribution implying a wider range of feature representations learned by the first convolutional layer. Additionally, in both versions of initializations, CL shows reduced value of α when compared to FS indicating that the kernels learned by CL methods fire on a further broader range of input.

Our studies show that pretraining and self-supervised learning is beneficial for the downstream medical imaging task to be able learn kernels that fire broadly and in turn extract more diverse and effective features from the input.

4.3 Robustness to Distortions:

The heavy-tailed activation statistics in combination with ReLU-thresholding in Section 4.2 showed that a larger number of neurons are capable of detecting structures in the input when the input data is varied according to sampling from the dataset. One can expect that this also may translate to an increased detection capability when input samples are varied by data augmentation parameters towards zones of lower data density. We have performed this experiment for the IWNE and IFI models by distorting the input with a set of predefined distortions as shown in [62].

One can see from Figure 4 that for the majority of distortion cases, the score for the self-supervised model is higher, indicating a higher robustness to the respective distortions. There is a marked difference between IWNE and IFI models. In the former case CL always provides an increase in robustness in comparison to FS. Using IFI in the latter case is known to provide good generalization for finetuning with respect to a wide range of target datasets. This improved generalization levels the difference between FS and CL. However IFI-CL still improves robustness for different noise types, pixelation and lower levels of saturation changes. Note the conspicuous outlier in IFI for JPEG compression.

4.4 Quantitative Analysis of Learned Cues

Explainability for DNN reveals what the model looks at on the image to make the prediction [63–74]. Using ground-truth segmentation masks, explanations have been evaluated to show quantitatively if what the model is looking at, is relevant for making the decision [75]. In the case of DR, a reasonable expectation is that the trained model looks at lesions in the retina that is indicative of the disease in order to make its decision. In order to evaluate the explanation heatmaps, we use the dataset of IDRiD [10] containing detailed pixel-wise annotation of the different lesions that contribute to the disease. The dataset consists of 80 images[1] with segmentation masks for microaneurysms, haemorrhages and hard exudates. To obtain explanation heatmaps, we utilize Layer-wise Relevance Propagation (LRP) with $\alpha_1\beta_0$ rule [65, 69].

Figure 5 shows the input followed by the segmentation maps for different lesions in the top row. The final image in the top row combines the different lesions to form the total. The bottom row shows the explanation heatmaps by using the different training methods. By comparing each result to the total marked in red in Figure 5, we can evaluate the effectiveness of the model in looking at the lesion to make the prediction. We find that explanation heatmaps from IWNE overfit on the hard exudates and show minimal correlation with the other lesions. On the other hand, explanation heatmaps from IFI models are significantly more outspread correlating better with different lesions.

The correlation of explanation heatmaps to the ground-truth segmentation maps also helps us make a quantitative evaluation of how accurately the models relies on the disease to make its prediction. We follow the evaluating strategies adopted in [75] including relevance mass accuracy and relevance rank accuracy. Given input $\boldsymbol{x}$, relevances R_i determining the importance of the input features x_i and $S \subseteq [0, 1]$ the ground truth segmentation mask, relevance mass accuracy is defined as:

$$\text{RMA} = \frac{\sum_{i \in S} R_i}{\sum_i R_i} \tag{2}$$

where the numerator corresponds to the sum of relevances where the ground truth segmentation maps exists and the denominator is the sum of all relevances. The relevance rank accuracy is defined as:

$$\text{RRA} = \frac{|R_{p_i} \cap S|}{|S|} \tag{3}$$

where R_{p_i} is the relevances in the top i^{th} percentile. While RMA corresponds to the precision, RRA corresponds to the recall. For pooling the relevances across the channels, we utilize the two pooling strategies followed by [75], although the findings here are agnostic to the pooling strategy utilized:

[1] The IDRiD dataset also contains segmentation maps for soft exudates for a smaller subset of images which we excluded from our quantitative evaluation.

- sum_pos: $R_{pool} = max(0, \sum_{i=1}^{C} R_i)$
- l2_norm_sq: $R_{pool} = \sum_{i=1}^{C} {R_i}^2$

where C is the number of channels.

Table 3 shows the results for RMA and RRA for the explanation heatmaps correlated on the ground-truth segmentation maps from the IDRiD challenge. We report the accuracies for each lesion—microaneurysms, haemorrhages and hard exudates and a total, where we combine the above mentioned lesions. The heatmaps for each of the methods are computed by backpropagating from the output neuron corresponding to severe DR. The heatmaps are evaluated using the two pooling strategies mentioned above for each lesion. As a control, we also report the results by replacing explanation heatmaps with random variables from Gaussian distribution. Any method which shows similar results to the control indicates that the heatmaps are just random, i.e. the model looks at random set of input features to make its prediction. In each category (lesion), the best result among the different training strategies are marked in bold for each pooling method.

We find in Table 3 that in the case of microaneurysms, random explanations achieve a mean accuracy of 0.0073 for RMA. Here, the model IWNE-FS achieves results that is very close to the results for the random explanations. On the other hand, all the other models report accuracies that are higher than the corresponding control value. This indicates that IWNE-FS may be ignoring microaneurysms for making its decision. The RMA results in Table 3 show that for the IWNE models, CL achieves better results. IFI models, in general report higher accuracies than that of IWNE models. Similar to IWNE, we find for IFI models that CL reports better RMA than FS using both the pooling strategies. This is confirmed again with results of RRA in the same table, where models with CL achieves the best results. Microaneurysms are the smallest lesions and it is vital for a method to base its decision on them for detecting progressive cases of DR. Our results indicate that IFI models and CL in particular are better equipped at including microaneurysms to make their predictions.

Haemorrhages are lesions that are slightly larger than microaneurysms. We find in Table 3 that here again IWNE-FS reports similar accuracies to that of the control indicating that this model may be ignoring the haemorrhages as well. Among IWNE models, CL clearly achieves higher RMA as well as higher RRA. This is again the case on the IFI models where CL achieves higher RMA and RRA indicating that the explanations using this model are better correlated with the ground-truth than their supervised counterpart FS.

In contrast to the smaller lesions, the hard exudates are large yellowish white deposits with sharp gradients. Here for RMA, the supervised models achieve better results than the self-supervised models as shown in Table 3. The results on RRA for hard exudates show that on majority of the cases, for both IWNE and IFI models, the supervised models show higher accuracies than the self-supervised models.

The *total* which measures the sum of the all the different lesions, we find here again that the supervised models achieve better results with RMA as shown in 3. With RRA, the IWNE models do not clearly outperform each other in the case of total. However, for IFI, the self-supervised model clearly outperforms the supervised model for the total of all the lesions.

The results of RMA and RRA in Table 3 reveal that the supervised models overfit on the hard exudates in both versions of initializations. IWNE-FS in particular fails to base its decision on microaneurysms and haemorrhages that may be highly relevant for the prediction of the disease. The results on the total are skewed by the results on the hard exudates. In alignment with our observations in Section 4.2, we find that the IFI models look at diverse set of input features (lesions) and report consistently higher accuracies than their IWNE counterparts. Among IFI, the results of CL correlates better with the explanation heatmaps for a variety of lesions indicating that they look at more diverse set of input features than any other method.

5 Summary and conclusions

Deep learning-based methods for the diagnosis of diabetic retinopathy have shown remarkable performance. In our paper, we study the important ques-

tion of the robustness of different training strategies — namely initialization from ImageNet pretraining and self-supervised learning. Our findings are three-fold: Firstly, we show the performance gains obtained by self-supervised learning in diabetic retinopathy. Secondly, we demonstrate the advantage of self-supervised learning along with initialization from ImageNet pretraining for diabetic retinopathy by analyzing the statistics of the eigenvalues of the feature representations learned. We also show improvements in robustness to distortions for self-supervised learning in comparison to purely supervised training. Finally, we use interpretability methods to gain quantitative insights into the patterns exploited by models trained using the different training schemes. In particular, we find that initialization from ImageNet pretraining significantly reduces overfitting to large lesions along with improvements in taking into account minute lesions which are indicative of the progression of the disease.

With our study, we try to convey that a more holistic view on the benefits of pretraining and self-supervision in medical imaging along the lines of the present study is important. To summarize, in absence of large unlabeled domain-specific data that would allow for self-supervised pretraining, we see numerous benefits in favor of using self-supervised pretrained models on Imagenet as starting point for finetuning on domain-specific data, which we put as a general recommendation.

Acknowledgments

Models were trained in Pytorch [76] building on the code kindly provided by [11]. This work was supported in part by the German Ministry for Education and Research (BMBF) under grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18056A, 01IS18025A and 01IS18037A. It is also supported in part by the Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-001779), as well as by Math+, EXC 2046/1, Project ID 390685689 through the German Research Foundation (DFG). Correspondence to WS, KRM.

References

[1] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in neural information processing systems*, 2019, pp. 3347–3357.

[2] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[3] M. Voets, K. Møllersen, and L. A. Bongo, "Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *PLOS ONE*, vol. 14, no. 6, p. e0217541, 2019.

[4] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," *arXiv preprint arXiv:2010.05352*, 2020.

[5] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 22 243–22 255.

[6] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, "Big self-supervised models advance medical image classification," *arXiv preprint arXiv:2101.05224*, 2021.

[7] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 355–366, 2021.

[8] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" in *Advances in Neural Information Processing*, 2020.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao *et al.*, "Idrid: Diabetic retinopathy–segmentation and grading challenge," *Medical Image Analysis*, vol. 59, p. 101561, 2020.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[13] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, 2020.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*, 2020.

[15] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," in *Medical Imaging with Deep Learning*, 2021.

[16] A. Sriram, M. Muckley, K. Sinha, F. Shamout, J. Pineau, K. J. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, and W. Moore, "Covid-19 deterioration prediction via self-supervised representation learning and multi-image prediction," *arXiv preprint arXiv:2101.04909*, 2021.

[17] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[18] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, no. 6, pp. 900–908, 2020.

[19] H. Takahashi, H. Tampo, Y. Arai, Y. Inoue, and H. Kawashima, "Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy," *PLOS ONE*, vol. 12, no. 6, p. e0179790, 2017.

[20] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[21] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *Investigative ophthalmology & visual science*, vol. 59, no. 1, pp. 590–596, 2018.

[22] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA summits on translational science proceedings*, vol. 2018, p. 147, 2018.

[23] Z. Gao, J. Li, J. Guo, Y. Chen, Z. Yi, and J. Zhong, "Diagnosis of diabetic retinopathy using deep neural networks," *IEEE Access*, vol. 7, pp. 3360–3370, 2018.

[24] X. Zeng, H. Chen, Y. Luo, and W. Ye, "Automated diabetic retinopathy detection based on

binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30 744–30 753, 2019.

[25] X. Wang, Y. Lu, Y. Wang, and W.-B. Chen, "Diabetic retinopathy stage classification using convolutional neural networks," in *International Conference on Information Reuse and Integration*, 2018, pp. 465–471.

[26] S. Wan, Y. Liang, and Y. Zhang, "Deep convolutional neural networks for diabetic retinopathy detection by image classification," *Computers & Electrical Engineering*, vol. 72, pp. 274–282, 2018.

[27] H. Chen, X. Zeng, Y. Luo, and W. Ye, "Detection of diabetic retinopathy using deep neural network," in *International Conference on Digital Signal Processing*, 2018, pp. 1–5.

[28] M. H. Johari, H. A. Hassan, A. I. M. Yassin, N. M. Tahir, A. Zabidi, Z. I. Rizman, R. Baharom, and N. Wahab, "Early detection of diabetic retinopathy by using deep learning neural network," *International Journal of Engineering and Technology*, vol. 7, no. 4, pp. 198–201, 2018.

[29] X. Xu, J. Lin, Y. Tao, and X. Wang, "An improved densenet method based on transfer learning for fundus medical images," in *International Conference on Digital Home*, 2018, pp. 137–140.

[30] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowledge-Based Systems*, vol. 175, pp. 12–25, 2019.

[31] A. Grzybowski, P. Brona, G. Lim, P. Ruamviboonsuk, G. S. Tan, M. Abramoff, and D. S. Ting, "Artificial intelligence for diabetic retinopathy screening: a review," *Eye*, vol. 34, no. 3, pp. 451–460, 2020.

[32] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, pp. 158–164, 2018.

[33] A. V. Varadarajan, R. Poplin, K. Blumer, C. Angermueller, J. Ledsam, R. Chopra, P. A. Keane, G. S. Corrado, L. Peng, and D. R. Webster, "Deep learning for predicting refractive error from retinal fundus images," *Investigative Ophthalmology & Visual Science*, vol. 59, no. 7, pp. 2861–2868, 2018.

[34] M. N. Bajwa, Y. Taniguchi, M. I. Malik, W. Neumeier, A. Dengel, and S. Ahmed, "Combining fine-and coarse-grained classifiers for diabetic retinopathy detection," in *Medical Image Understanding and Analysis*, 2019, pp. 242–253.

[35] A. Rakhlin, "Diabetic retinopathy detection through integration of deep learning classification framework," *bioRxiv*, p. 225508, 2018.

[36] C. A. Ludwig, C. Perera, D. Myung, M. A. Greven, S. J. Smith, R. T. Chang, and T. Leng, "Automatic identification of referral-warranted diabetic retinopathy using deep learning on mobile phone images," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 60–60, 2020.

[37] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific Reports*, vol. 7, no. 1, 2017.

[38] A. Filos, S. Farquhar, A. N. Gomez, T. G. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal, "A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks," *arXiv preprint arXiv:1912.10481*, 2019.

[39] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a deep learning

algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642018315756

[40] "Kaggle. diabetic retinopathy detection challenge," https://www.kaggle.com/c/diabetic-retinopathy-detection.

[41] "Messidor 2," http://www.adcis.net/en/third-party/messidor2/.

[42] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3d self-supervised methods for medical imaging," in *Advances in Neural Information Processing*, 2020.

[43] O. G. Holmberg, N. D. Köhler, T. Martins, J. Siedlecki, T. Herold, L. Keidel, B. Asani, J. Schiefelbein, S. Priglinger, K. U. Kortuem *et al.*, "Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 719–726, 2020.

[44] R. Geirhos, K. Narayanappa, B. Mitzkus, M. Bethge, F. A. Wichmann, and W. Brendel, "On the surprising similarities between supervised and self-supervised models," *arXiv preprint arXiv:2010.08377*, 2020.

[45] F. Navarro, C. Watanabe, S. Shit, A. Sekuboyina, J. C. Peeken, S. E. Combs, and B. H. Menze, "Evaluating the robustness of self-supervised learning in medical imaging," 2021.

[46] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *International Conference on Machine Learning*, 2019, pp. 2712–2721.

[47] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 15 663–15 674.

[48] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," in *Association for Computational Linguistics*, 2020, pp. 2744–2751.

[49] J. Djolonga, J. Yung, M. Tschannen, R. Romijnders, L. Beyer, A. Kolesnikov, J. Puigcerver, M. Minderer, A. D'Amour, D. Moldovan *et al.*, "On robustness and transferability of convolutional neural networks," *arXiv preprint arXiv:2007.08558*, 2020.

[50] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Robust pre-training by adversarial contrastive learning," in *Advances in Neural Information Processing*, vol. 33, 2020, pp. 16 199–16 210.

[51] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, "Adversarial robustness: From self-supervised pre-training to fine-tuning," in *Computer Vision and Pattern Recognition*, 2020, pp. 699–708.

[52] A. Y. Peng, Y. S. Koh, P. Riddle, and B. Pfahringer, "Using supervised pretraining to improve generalization of neural networks on binary classification problems," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018, pp. 410–425.

[53] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[54] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, M. Carbin, and Z. Wang, "The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models," *arXiv preprint arXiv:2012.06908*, 2020.

[55] I. Kandel and M. Castelli, "Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review," *Applied Sciences*, vol. 10, no. 6, p. 2021, 2020.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

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

[60] E. Taskesen, "distfit," https://github.com/erdogant/distfit, 2019.

[61] V. Pareto, *Cours d'économie politique*. Librairie Droz, 1964, vol. 1.

[62] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of the International Conference on Learning Representations*, 2019.

[63] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[64] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[65] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[66] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, vol. 1, no. 1, pp. 39–48, 2018.

[67] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017, pp. 3319–3328.

[68] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017, pp. 3145–3153.

[69] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[70] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019, vol. 11700. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-28954-6

[71] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Scientific reports*, vol. 10, no. 1, p. 6423, 2020.

[72] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[73] A. Holzinger, R. Goebel, M. Mengel, and H. Müller, *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-art and Future Challenges*. Springer Nature, 2020, vol. 12090.

[74] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai," *Information Fusion*, vol. 71, pp. 28–37, 2021.

[75] A. Osman, L. Arras, and W. Samek, "Towards ground truth evaluation of visual explanations," *arXiv:2003.07258*, 2020.

[76] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.