Agree to Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations

Matthew Watson, Bashar Awwad Shiekh Hasan, Noura Al Moubayed
Durham University
Durham, UK
{matthew.s.watson,bashar.awwad-shiekh-hasan,noura.al-moubayed}@durham.ac.uk

Abstract

Deep Learning of neural networks has progressively become more prominent in healthcare with models reaching, or even surpassing, expert accuracy levels. However, these success stories are tainted by concerning reports on the lack of model transparency and bias against some medical conditions or patients' sub-groups. Explainable methods are considered the gateway to alleviate many of these concerns. In this study we demonstrate that the generated explanations are volatile to changes in model training that are perpendicular to the classification task and model structure. This raises further questions about trust in deep learning models for healthcare. Mainly, whether the models capture underlying causal links in the data or just rely on spurious correlations that are made visible via explanation methods. We demonstrate that the output of explainability methods on deep neural networks can vary significantly by changes of hyper-parameters, such as the random seed or how the training set is shuffled. We introduce a measure of explanation consistency which we use to highlight the identified problems on the MIMIC-CXR dataset. We find explanations of identical models but with different training setups have a low consistency: $\approx$ 33% on average. On the contrary, kernel methods are robust against any orthogonal changes, with explanation consistency at 94%. We conclude that current trends in model explanation are not sufficient to mitigate the risks of deploying models in real life healthcare applications.

1. Introduction

Deep Learning (DL) applications in healthcare have recently enjoyed a series of successes, with DL models performing on par with human experts leading to the US Food & Drugs Administration (FDA) to approve 64 DL based medical devices and algorithms as summarised in a recent survey [5]. Whilst these results demonstrate that the trained models are able to perform well on the selected performance criteria, this is not enough for models to reach widespread adoption in practice. This is particularly true in the healthcare domain where it is imperative that the DL models used must be both transparent and explainable, in order to ensure that the relevant stakeholders (patients, medical practitioners) can place their trust in the model, and to help prevent "catastrophic failures" [7, 16].

The ultimate aim of a DL model in highly sensitive applications, such as healthcare, is to capture the underlying causal inter-relationships that medical professionals learn through experience to use for classification. Such a model would be robust to spurious correlations and changes in model training perpendicular to the classification task. Without this level of robustness there will be no trust for its use in the real-world. Current DL training methods often fail to satisfy this requirement, as robustness/trust is yet to be an intricate part of the evaluation and optimisation of said models [8, 24]. An egregious recent example can be seen in certain pneumonia diagnosis models, where it has been shown that the models learned to detect regions (e.g. a metal token placed by the radiologists) of the chest x-ray (CXR) image that indicated which hospital the sample was from, rather than the regions of the image that were causally linked to pneumonia. Despite this, the model still achieved a reasonable ROC-AUC of 0.773 as, incidentally, some hospitals had higher rates of pneumonia than others and so image origin was a good predictor of pneumonia [39]. Since the model relied disproportionately on spurious correlations that are not causally linked to pneumonia, it was unable to generalise to unseen data outside of training hospitals.

Recent theoretical and experimental work has demonstrated the challenge of generalisation for DL models and their vulnerability to small changes in the data [10]. Ensemble models, where multiple, slightly different models work together to make a final prediction, have been proposed to alleviate these issues [15, 26]. However, while these techniques can improve the robustness of models, they are rarely inherently explainable and do not necessarily capture causal

relationships. Additionally, a fundamental requirement of trustworthy models is the interpretability of their decisions. The development of explainable DL techniques to date use either model agnostic post-hoc or model specific approaches. However, the quality of explainable methods is still very difficult to quantify and is designed to be truthful to the model not the data [18, 37].

This paper explores the limits of explainable machine learning which highlights fundamental problems in the training and generalisation of neural networks. In particular, we demonstrate how the noise learned by a deep learning model can change significantly when factors such as the random seed, initial weights or even training set order are changed (whilst all other variables remain the same). We propose a measure of the consistency of explanations to quantify the problem and discuss its impact on the interpretation of the explainable output in relation to the input features importance. We show that even the current state-of-the-art ensemble models present with the same issues, and discuss the implications of these findings on the viability of deploying machine learning models in sensitive application domains [1, 2, 12].

2. Generalisation and Underspecification

With the increased use of ML in general and DL in particular, we are becoming increasingly aware of the limitations of DL models. For example, deep neural networks have been shown to be susceptible to imperceptible changes in the the input [34], or rely on unexpected parts of the input when making their decisions [4]. There is also an increasing number of concerning scenarios wherein a neural network makes biased decisions, such as face detection models reporting high error rates for faces from ethnic minority groups [6, 38].

There is a growing concern of applications with profound difference between the training dataset and that used in practice, so much so that the differences in the underlying causal structure of the data leads to the poor performance of the trained model [8]. Even when models are able to generalise well, there is a lack of understanding of why, for example, SOTA vision models converge and generalise even when trained on unstructured noise [40]. The picture gets even more complex with recent work suggesting neural networks are immune to the bias-variance trade-off with over-parameterised networks demonstrating a striking absence of the classic U-shaped test error curve [25, 36]. Additionally, shortcut learning [14], or decision rules which work well on standard benchmarks but fail to generalise to more complex situations, has recently been shown to be prevalent across many different machine learning domains. Post-hoc explainable methods have gained traction recently to mitigate the issues with model training by opening, albeit rather partially, the black box of a neural network. However, the quality of explanations produced by these methods is difficult to quantify [37]. In [9], the authors demonstrated the susceptibility of explainable methods to the same type of adversarial attacks to that of the original models. We demonstrate here that the generated explanation can be unstable and inconsistent due to variations in model training that are irrelevant to the classification task.

From their inception, ensemble models that incorporate many, diverse sub-models have been proposed to address the problems of robustness and generalisability [32, 26, 35]. However, as we will demonstrate they also fail to mitigate the low consistency problem of model explanations. We argue that the lack of understanding of exactly how these deep learning models work [11] and generalise is ultimately preventing us from addressing the aforementioned issues. Understanding how the stochastic nature of the training process affects what properties of the data is captured by the model is fundamental. But recent theoretical and experimental studies to understand the generalisation of neural networks concluded the inadequacy of current measures of generalisation [10, 20].

A closer look at explainable outputs of DL models allows us to understand how the randomness introduced during the training significantly affects the explanation of the model's decisions despite consistent accuracy levels. This raises important questions around the robustness of these models. On the contrary kernel methods, namely SVM, are robust against these changes, suggesting that it is the stochastic nature of deep learning model training that may be causing these issues to arise. We argue that these issues significantly impede our ability to confidently suggest DL models for use in healthcare, as they imply that the models might be relying on spurious correlations in the data leading to models producing inconsistent explanations upon retraining.

3. Measuring Explanation Consistency

We argue that consistency of the explanations produced by a model regardless of orthogonal changes to hyper parameters is a strong surrogate to model robustness. Fidelity of explanations on the micro level, i.e. input features, is the basis to quantify explanations [37, 28]. Here, we are validating explainability on the macro level, i.e. the robustness of the produced explanation regardless of changes to model training that are orthogonal to the model architecture, data content, and classification task. Intuitively speaking, the consistency of explanations across model variations engender trust in these models as the end user does not expect changes in the explanation due to an incremental model update. Existing similarity metrics of different model outputs(e.g. cosine similarity, root mean squared error) are ill-suited to this task as they are unable to accurately quantify the small (yet important) changes that are particularly of interest here. The separability of a binary classifier, i.e. training accuracy, is an established measure of changes in model output [13] which we adapt here to form the basis to

measure consistency within the framework defined next.

3.1. A Measure of Consistency

Given a dataset $X = \{x_1, ..., x_N\} \subset R^d$, where $d \in \mathbb{N}$ is the dimension of the sample data, we have a classification task $Y(x_i) \in R^n$, where n is the number of classes in a classification setting. We want to evaluate the consistency of explanation method E, where $E(Y(x_i)) \in R^d$ assigns a weight to every input feature based on its influence on $Y(x_i)$.

Assume we have V variations of the model Y, which we will indicate as $Y^v, v \in \{1, \ldots, V\}$, then we define the explanation separability of any two of these variations as:

$$S_{(a,b)} = \mathbb{E}_i\Big[D\Big(E(Y^a(x_i)), E(Y^b(x_i))\Big)\Big] \tag{1}$$

where $i \in \{1, \ldots, N\}$, and D is a similarity measure between the two explanations provided by E of the output of the two models Y^a and Y^b, and $\mathbb{E}_i$ is the expected value. The larger $S_{(a,b)}$ is then the more distinct the explanations produced by the same model architecture under the training conditions, a and b. Without loss of generality we assume $S_{(a,b)}$ to be normalised in the range $[0, 1]$ and we define consistency as:

$$C = 1 - \frac{\sum_{(a,b)} S_{(a,b)}}{\alpha} \tag{2}$$

where α is the number of comparisons made between variations of the trained model. The separability metric $S_{(a,b)}$ should be defined such that when the explanations are completely separable (i.e. $S_{(a,b)} = 1$) then the consistency $C = 0$, and vice-versa.

3.2. Choosing a Suitable Separability Metric

The definition of $S_{(a,b)}$ should be determined based on the characteristics of X, e.g. data dimension and sparsity, and as such it makes sense that different definitions may be appropriate in different scenarios, as long as it is monotonic in the range $[0, 1]$. Multiple definitions could be chosen ranging from information-theoretic measures to statistical metrics of similarity (note that similarity metrics can be modified to fit our definition of $S_{(a,b)}$ by "flipping" their output to ensure that $S_{(a,b)} = 0$ when a, b are identical). Throughout this paper we use the training accuracy of a binary model, $M_{(a,b)}$, trained to classify between $E(Y^a(x_i))$ and $E(Y^b(x_i))$ for $i \in 1, \ldots, T$, where T is the size of the testing set. Eq.2 can then be re-written as:

$$C = 1 - \frac{\sum_{(a,b)} 2 * |M_{(a,b)} - 0.5|}{\alpha} \tag{3}$$

where $|.|$ is the absolute operator. $S_{(a,b)}$ is set to $2*|M_{(a,b)} - 0.5|$ to normalise the classification accuracy and make it more meaningful as separability by measuring its distance from theoretical random baseline. An accuracy $M_{(a,b)} = 1$ means the two explanations are completely separable with $S_{(a,b)} = 1$ and $C = 0$, and on the other extreme an accuracy $M_{(a,b)} = 0.5$ means that there is perfect agreement between a and b resulting in $S_{(a,b)} = 0$ and $C = 1$. However, while we have chosen to use the training accuracy of a binary classifier to measure the distance, D, between the explainability values, as noted earlier different distance measures could be used and it may be the case that different distance metrics are suited better to different applications and datasets. When choosing a separability metric, it is important to determine whether the chosen distance metric is sensitive enough to detect the small changes in the explanations that we wish to detect. Each possible consistency metric will have various advantages and disadvantages, and it may be that some are better suited to different scenarios; one of the reasons we have chosen to use a binary classifier is its wide range applicability and intuitive interpretation.

Table 1 contains the values of different divergence measures that we have tested on 4 CNNs (of identical architecture) trained on MNIST with different random seeds. Jensen-Shannon divergence (JSD) is based upon Kullback-Leibler (KL) divergence, and is a method of measuring the similarity between two probability distributions; making it common in machine learning applications, and a prime candidate for use here. JSD is better suited for measuring separability as it is normalised in the range $[0, 1]$. Its main disadvantage is that it measures the divergence between probability distributions, and not samples drawn from a distribution. This requires us to estimate the distribution of the explainability values for the two models under test. This adds an extra layer of complexity to the calculation, and could lead to errors where differences in the techniques and assumptions used to estimate the probability functions. For our experiments reported in Table 1 we used Kernel Density Estimation (KDE), a method of estimating an unknown probability density function using a kernel function [27], which has produced good results, however this would be entirely problem-dependent, whereas the binary classifier method (e.g., Linear Regression(LR)) discussed in the previous section is more generalisable.

Statistical hypothesis tests that are designed to test whether two sets of samples are drawn from the same distribution are other candidates. The 2 sample Kilmogorov-Smirnov (KS) test is a two-sided test for the null hypothesis that the 2 sets of samples are drawn from the same continuous distribution [29]. Using the KS test as a separability measure has the benefit of having a solid statistical underpinning, but we encounter problems when carrying out the test. While we can accurately compute the test statistic (reported for a small set of model in Table 1), we cannot compute the associated p-values, preventing us from accurately completing the hypothesis test. In all of our experiments (except

M1 Seed	M2 Seed	JSD	KS	Wilcoxon	LR
1	1	0	0	0	0.5
1	12303	0.8062	0.9744	7.877e+09	0.973
1	15135	0.8012	0.9690	1.738e+10	0.978
1	16959	0.7346	0.8890	2.464e+11	0.975
12303	12303	0	0	0	0.5
12303	15135	0.8228	0.9913	4.350e+08	0.979
12303	16959	0.7900	0.9567	3.316e+10	0.974
15135	15135	0	0	0	0.5
15135	16959	0.8122	0.9810	6.611e+09	0.975

Table 1: Table reporting the Jensen-Shannon divergence, 2 sample Kilmogorov-Smirnov and Wilcoxon signed-rank *test statistics* on the SHAP values from a small subset of the MNIST CNNs tested. The p-values for all hypothesis tests were calculated as 0. Kernel Density Estimation was used before calculating the Jensen-Shannon divergence of the explanations. LR is the accuracy of Logistic Regression classifiers trained on the SHAP values, as used throughout this paper as $M_{(a,b)}$.

those where we were testing a model against itself, where we calculated a test statistic of 0 and p-value of 1), our calculations returned a p-value of 0 (due to technical limitations, we cannot calculate precise enough p-values and so they are rounded down to 0). A similar issue arises when we use the Wilcoxon signed-rank test, which is a non-parametric alternative to the paired t-test which can work on highly non-normal data that works on the null hypothesis that the median differences between pairs of samples are 0. While these results (i.e. calculating a p-value of 0) highlight that our results are highly statistically significant (and hence we can reject the null hypothesis and conclude the explanations are drawn from different distributions), we cannot use results from hypothesis tests to quantify to what degree the explanation's from two models are separable (i.e. we will be unable to infer if one architecture produces more consistent explanations than another), whereas our results with a binary LR classifier allow us to do so. This is not to say that JSD or KS/Wilcoxon hypothesis tests are entirely unsuited to use as a basis for the consistency measure. In this work we have focused our experiments on image data, where input contains a large number of features; applications where fewer features are used might alleviate the technical issues mentioned above. In these cases, it may be appropriate to use one of these measures. However, our choice of a binary classifier is easy to use in any scenario, to any dataset and is easy to interpret and quantify.

4. Experimental Setup

We use two publicly available datasets. MNIST is used for efficient baseline tests, and we then extend our experiments to use the MIMIC-CXR-JPG [21]. We investigate a wide breadth of different model architectures, explanation methods, and training variations[1]. For both datasets, we use the recommended train/test/val splits. For reproducibility, the specific hyperparameters used for each experiment can be found in the Supplementary Material.

MNIST Experiments: We experimented with the following variations: **1) MLP** with two hidden layers of sizes 412 and 512 respectively and a dropout layer, **2) Small-CNN**, a convolutional neural network of 1 convolutional layer with kernel size 3, followed by a max pooling and fully connected layer, **3) CNN** two convolutional layers with kernel size 3, using max pooling and fully connected layers in between, **4) GaborNet**, a Small-CNN network with the first convolutional layer restricted to use Gabor filters (the exact parameters of these filters are learned by the network) [3], **5) ResNet18** [17] with the first convectional layer modified to take 1 channel inputs and the final output layer to have an output size of 10, and **6) SVM** with RBF kernel. We also train two ensemble models: **1) ADP ensemble** [26] using the default hyperparameters and consisting of 10 ResNet sub-models, and **2) Hyperensemble** a hyper-batch ensemble [35] using the default hyperparameters with 3 sub-models.

MIMIC-CXR-JPG Experiments: The dataset contains 377,110 chest x-rays (CXRs) images from 227,827 studies [21]. Each study has up to 14 associated labels denoting the disease(s) which are present in the CXR images. For our purposes, we focus only on images with the Edema label; this gives us a subset of 77,483 images of which 47.2% present with the disease (have a positive label) and the remaining 52.8% do not (have a negative label). We use the labels as presented in the MIMIC-CXR-JPG dataset: these have originally been extracted from free-text radiology reports via the CheXpert tool [19, 21]. We use the MIMIC-CXR-JPG dataset to demonstrate the issues raised here on a real-life healthcare application. We focus on the Edema label as otherwise we are left with a multi-label classification problem (as one CXR image may show multiple diagnoses), which would make isolating the source of variation very difficult to guarantee. We chose the Edema label specifically as it provides a large number of images whilst also having largely balanced classes. The scope for experimentation with MIMIC-CXR-JPG is necessarily more limited than that with MNIST, as the data requires more complex networks to gain optimal performance. We follow the same process as CheXNet [31], fine tuning a pre-trained Densenet-121 model. We also train a voting ensemble consisting of 3 pre-trained Densenet-121 models trained on subsets of the training dataset.

On both datasets, we train the models repeatedly. For each run we change the hyperparameters that can lead to variations in the randomness used during training without

[1]Code to reproduce our experiments can be found at https://github.com/mattswatson/agree-to-disagree

Model Architecture	Dataset	Shuffle	Random Seed	Dropout
MLP	MNIST	98.195 ± 0.9550	98.18 ± 0.94	98.25 ± 0.8292
SVM	MNIST	93.825 ± 0.7746	94.218 ± 0.3943	n/a
Small-CNN	MNIST	98.385 ± 0.0250	98.345 ± 0.015	98.3267 ± 0.0330
ADP Ensemble	MNIST	98.5 ± 0.14	99.0875 ± 0.2573	n/a
CNN	MNIST	97.5 ± 0.5	99.2170 ± 0.0443	99.1580 ± 0.0595
GaborNet	MNIST	95.031 ± 0.2769	95.034 ± 0.2742	95.054 ± 0.2934
ResNet18	MNIST	99.083 ± 0.2514	99.471 ± 0.0438	n/a
Densenet-121	MIMIC-CXR	76.005 ± 0.8363	75.4535 ± 1.2539	n/a
Densenet-121 Ensemble	MIMIC-CXR	81.98 ± 0.34	80.8533 ± 0.5311	n/a
Hyperensemble	MNIST	n/a	99.32 ± 0.0082	n/a

Table 2: Table reporting mean model accuracy ($\pm$ standard deviation) across variations on the base classification task.

changing the architecture of the model. We change: **1)** the random seed used during training, **2)** the dropout rate used in the networks (where applicable), and **3)** the order of the training data. It is important to note that the train/test/val splits remain the same, rather it is the order in which the training data is passed to the model during training which changes. The accuracy of the models on the base classification task (i.e. MNIST or MIMIC-CXR) are summarised in Table 2. To inspect the consistency of decision explanations as a result of changing these hyperparameters, we use two state-of-the-art explainability techniques: SHAP [22] and Integrated Gradients (IG) [33]. These two techniques were chosen as they represent the most commonly used state of the art feature-attribution explanation methods: I) SHAP is a permutation-based model-agnostic approach, so can be applied to the output of any model II) IG is gradient based making it applicable for all neural networks architectures. We calculate the explanation consistency for each explanation technique per model and dataset taking into account every training variation. A Logistic Regression (LR) classifier is used as the binary model to classify between $E(Y^a(x_i))$ and $E(Y^b(x_i))$ as per Eq. 3. This LR model takes the explanation values (i.e. SHAP values, IG values) of the two models as input, and is trained to classify which model the values originated from. The average training accuracy from 10-fold cross validation of the LR model is used. The higher the accuracy of the LR models, the more separable the explainability values are, suggesting that the two models are placing importance on significantly different parts of the input.

To confirm that the underlying problem lies in the models themselves, and not the explainability techniques used, we calculate three different explanation quality metrics that are designed to ensure the explanations produced accurately represent the models: I) (In)fidelity: is the mean squared error between the explanation multiplied by a (meaningful) change in the input and the difference between the model output when given the original and perturbed inputs. II) Sensitivity: measures the change in explanations when the input is slightly perturbed, calculating this using a Monte Carlo sampling based approximation [37]. III) Explanation Accuracy: is the accuracy of a model on the base task (of the same architecture the explanations were produced from) trained on the produced explanations (for example, for MNIST, can a model be trained on the explanations to classify each explanation into one of the 10 digit classes) [23].

5. Results and Discussion

Through visualisation of the explanation differences, we are able to discern whether the lack of consistency between variations is a cause for concern when deploying deep learning models to real-world scenarios. Figure 1 demonstrates the change in explanations between two variations of the same Densenet-121 model using SHAP. We see two main sets of differences in the images: **1)** areas of the image that are clinically significant (e.g. the lungs and the heart), and **2)** areas in background portions of the image. Those differences that are in clinically relevant to diagnosis can result in significantly reduced trust in the model, as we ideally want a model which has learnt the entire set of causal links present in the data (whereas these differences show that the two models have learnt to look at different sets of causal features). The remaining differences are in the background noise of the images, which suggests that the models are potentially picking up spurious correlations, with each model learning different sets of spurious correlations. Neither of these scenarios is desirable. Examples on Small-CNN trained on MNIST are shown in Figure 1 in the Supplementary Material - similarly to the CXR samples, we can see that the changes in the SHAP values are mainly centered around the areas of the image that are critical for number classification. These results are significant - it suggests both that variations in the training setup of a model changes the importance of the fun-

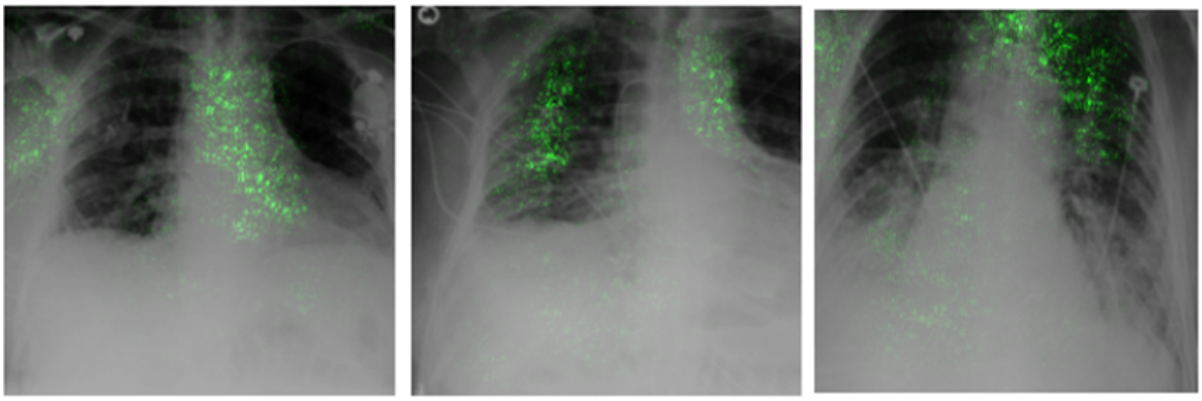

Figure 1: 3 random samples from the MIMIC-CXR-JPG dataset overlayed (in green) with the difference between the normalised SHAP values from two Densenet121 training variations.

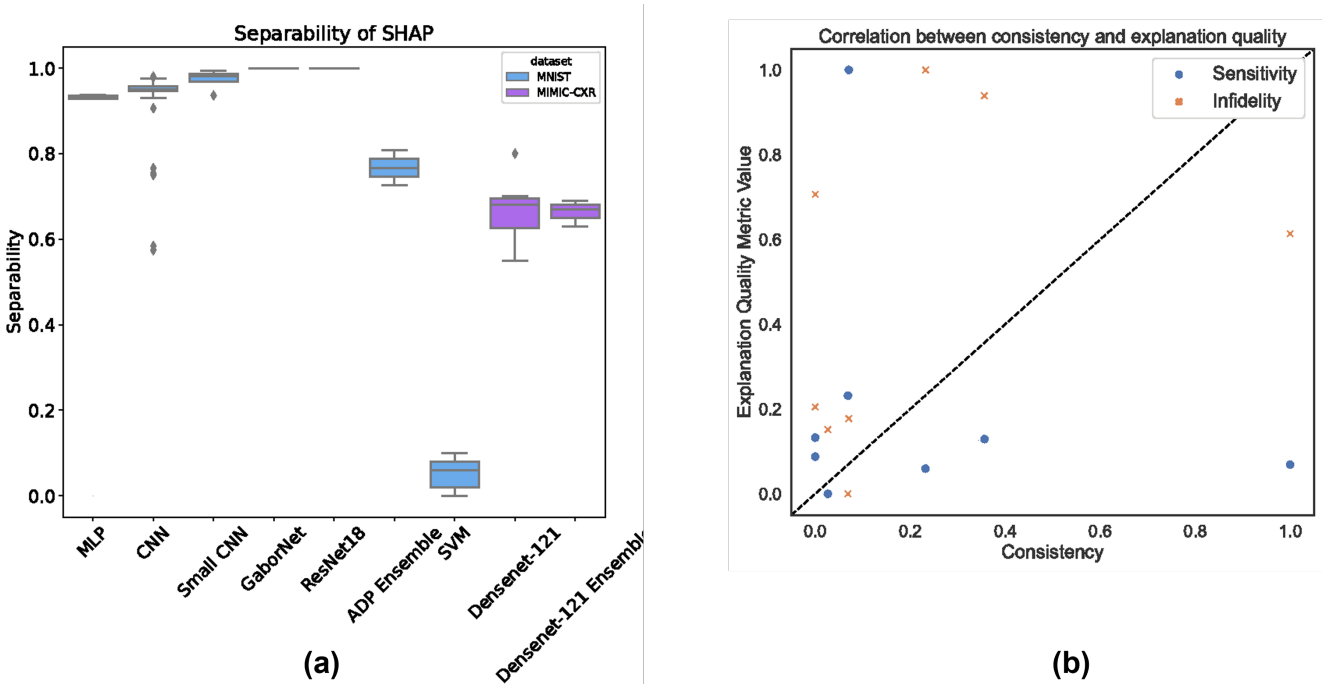

Figure 2: (a) Box plot of $S_{(a,b)}$ for SHAP across all training variations (a, b), for all model architectures tested. (b) Plot of SHAP explanation consistency of model architectures vs. SHAP infidelity and sensitivity of the same models across both MNIST and MIMIC data.

damental features that we would expect to be causally linked to the final classification, and on more complex tasks are also changing the spurious correlations learned by models.

Following, we report the accuracy of all models tested on MNIST and MIMIC-CXR-JPG, and the consistency of the explainability methods per model/dataset. Table 3 contains each model architecture's consistency, and a further breakdown of the consistency for the different types of training variation tested. For all model architectures, the degree of consistency is similar regardless of which hyperparameters is changed; this suggests that deep learning models are sensitive to all training hyperparameters, and not just a select few. Figures 2(a) and 3 further demonstrate the variation in the separability measure ($S_{(a,b)}$) used to calculate consistency across all models/datasets. These figures show that there is very little consistency of either SHAP and IG for any training variation when used with deep learning models. By contrast, we find that SVMs do not suffer from the same issue as deep learning models, achieving very high levels of consistency across both random seed and training shuffle

Model Architecture	Dataset	α	Consistency				Accuracy
			Overall	Shuffle	Random Seed	Dropout	
MLP	MNIST	6	0.0668	0.062	0.066	0.0687	98.125 ± 0.9270
SVM	MNIST	10	0.9444	0.96	0.94	n/a	94.0556 ± 0.6213
Small-CNN	MNIST	6	0.0252	0.018	0.06	0.034	98.3486 ± 0.0360
GaborNet	MNIST	12	0	0	0	0	95.038 ± 0.2824
ResNet18	MNIST	10	0	0	0	n/a	99.425 ± 0.0626
ADP Ensemble	MNIST	6	0.2193	0.192	0.233	n/a	99.083 ± 0.2514
CNN	MNIST	12	0.0652	0.052	0.0564	0.0914	98.9976 ± 0.5756
Densenet-121	MIMIC-CXR	6	0.3329	n/a	0.3329	n/a	75.6723 ± 1.1379
Densenet-121 Ensemble	MIMIC-CXR	4	0.3367	n/a	0.3667	n/a	80.8 ± 0.7483
CNN (IG)	MNIST	12	0	0	0	0	98.9976 ± 0.5756
Hyperensemble (IG)	MNIST	2	0	n/a	0	n/a	99.32 ± 0.0082
Densenet-121 (IG)	MIMIC-CXR	6	0.168	0.115	0.2033	n/a	75.6723 ± 1.1379

Table 3: Table reporting the consistency between training variations for the models tested and the average accuracy of the model architecture on the base classification task. The Shuffle, Random Seed and Dropout columns report the consistency of models when *only* the respective hyperparameter was changed. The Overall column reports the overall consistency of that architecture, taking an average of the consistency across all hyperparameters. α refers to the number of models tested for the overall architecture consistency (see Eq. 2). Please refer to Table 1 in the Supplementary Material for α values for the shuffle, seed and dropout consistencies.

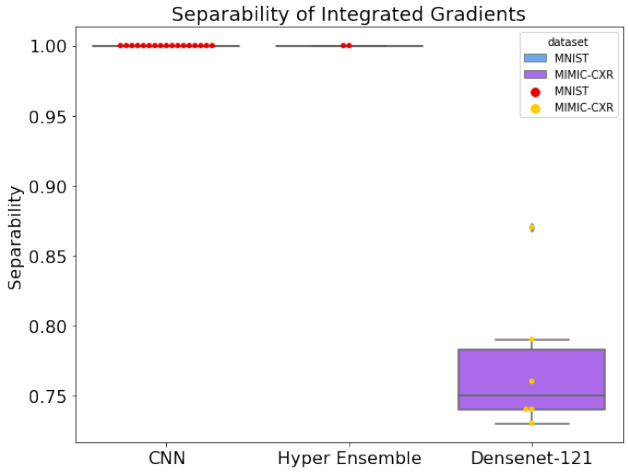

Figure 3: Boxplot of the separability $S_{(a,b)}$ of the Integrated Gradients explanations.

variations. This provides evidence for our hypothesis that it is the stochastic nature of deep learning model training that may be causing these issues to arise. Figure 3 shows the boxplot for IG, with even more pronounced separability, which can likely be attributed to how IG is calculated based on the weights of the network. Figure 2(a) does not show any real link between the size/depth of a network architecture and the separability/consistency.

Interestingly, both GaborNet and ResNet18 are highly inconsistent. The purpose of Gabor filters in CNNs is to more accurately simulate our biological understanding of human vision, with these filters picking up low-level features. Our results show that the features picked up are inconsistent - intuitively this makes sense, with lower-level features being more prone to smaller changes in the model. The purpose of testing the ResNet18 architecture was to investigate whether overparameterised networks also suffer from this inconsistency problem; as can be seen in Figure 2(a) and Table 3, they do. This implies that even models which have many more times the number of parameters than data points are converging to slightly different points on the loss landscape when small hyperparameter changes are made. It also suggests that even high capacity networks, which we would expect to be able to learn the entire set of meaningful features, are in fact either not able to do so. We hypothesise that, although ResNet18 is most likely learning (to some degree) all of the features present in the model, it is applying different weights to the noise present in the model when training hyperparameters are changed. This is not surprising as the overparameterised model has more chance of picking up spurious correlations. Figure 2(b) shows the correlation, or lack thereof, between explanation consistency and (in)fidelity and sensitivity as measures of the explanation's quality across all experimental settings. Both measures show weak Pearson correlation (0.4 for (in)fidelity and -0.3 for sensitivity). This is not surprising as those metrics are designed to be faithful to the model, not to the underlying data. This disparity between explanation consistency and quality highlights the problem with the use of explanation methods as a surrogate to model transparency. A lower consistency model is less robust and can lead to misinterpretation of model output, hence

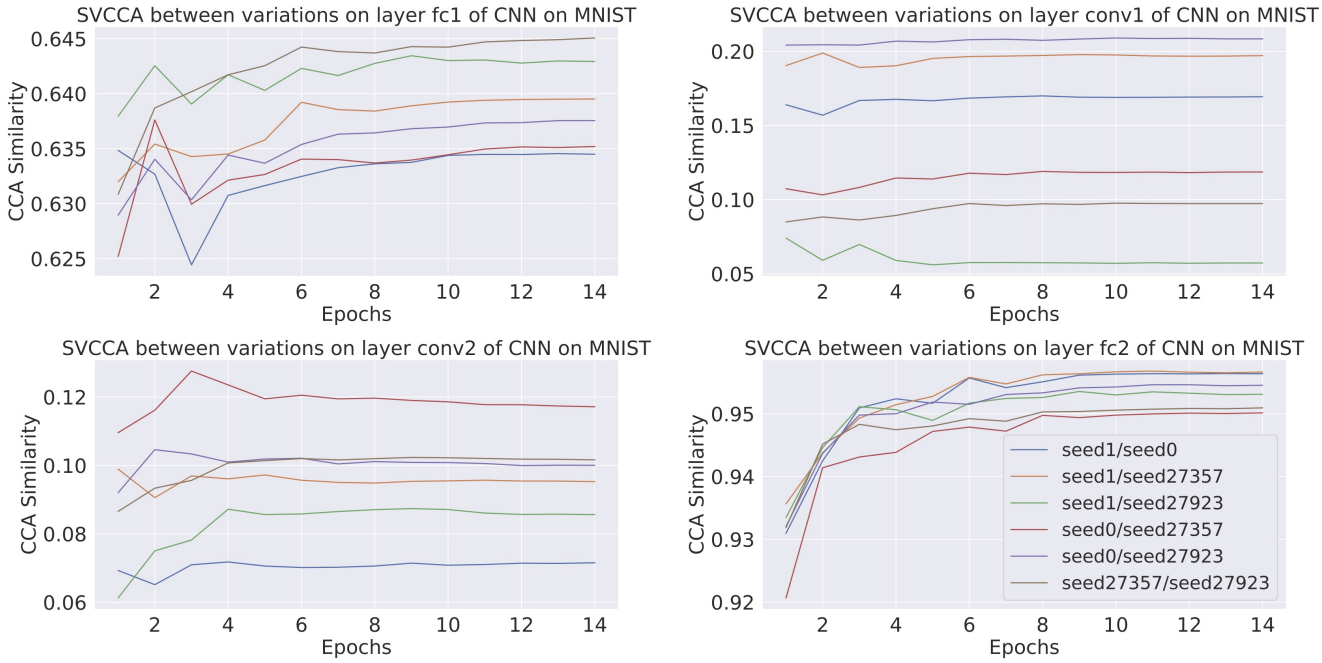

Figure 4: Figures showing the CCA similarity as training progresses between layer parameters. Each coloured line is a separate training variation pair of a CNN trained on MNIST.

damaging the confidence of using the model in sensitive domains. To further measure the quality of the SHAP and IG explanation, we also calculate the explanation accuracy for each model, i.e. the accuracy of the model trained on the explanation output of the sample data: a higher accuracy suggests more representative explanations. We report each model's individual explanation infidelity, sensitivity max and accuracy in Table 1 in the Supplementary Material. The weak correlation between the quality metrics and consistency lead to two conclusions: **1)** explanation quality metrics are unable to detect inconsistency in the models, and **2)** if the explanations are indeed faithful to the model, then the only remaining source of inconsistency is the trained model itself, or more precisely the training approach of these models. Ensemble approaches seem to have higher consistency but it is still significantly lower than that of SVMs. We use SVCCA [30] to inspect the similarity of layer parameters between two training variations, and how these change as training progresses. SVCCA views neurons as their activation vectors, and uses an amalgamation of Singular Value Decomposition and Canonical Correlation Analysis to analyse these representations: we encourage the interested reader to peruse [30] for a more thorough explanation. Figure 4 shows the SVCCA similarity between layers of CNNs trained on MNIST with different random seeds. It shows a high degree of similarity for the final layer, whereas the middle layer (`conv2`) shows a significant difference. This corroborates our explainability consistency results; the final layers (`fc2`) are similar and so the models will produce similar outputs, resulting in similar performance levels. Conversely, all other layers are significantly different and so the explanations, which take into account the whole model, are different. In addition, the two convolutional layers show an extremely low degree of similarity between the two models, hence the feature maps learned by these two models are likely also not similar resulting in lower consistency.

6. Conclusion

In this paper we introduced a consistency measure of explainable machine learning and demonstrated that deep learning models converge to learn different features when the same model is trained with different random seeds, training set orders and dropout rates. By validating the quality of the explanation techniques used, and using both gradient-based and perturbation-based techniques, we have shown that this is a fundamental problem with deep learning models rather than an issue with the explanations. Additionally, we verified that SVMs are immune to this problem. We argue that there is still significant work that need to be done to build robust trustworthy deep learning solutions in real-life healthcare applications.

Acknowledgements

This work is supported by grant 25R17P01847 from the European Regional Development Fund and Cievert Ltd.

References

[1] Zakhriya Alhassan, David Budgen, Riyad Alshammari, Tahani Daghstani, A Stephen McGough, and Noura Al Moubayed. Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 541–546. IEEE, 2018.

[2] Zakhriya Alhassan, A Stephen McGough, Riyad Alshammari, Tahani Daghstani, David Budgen, and Noura Al Moubayed. Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models. In *International Conference on Artificial Neural Networks*, pages 468–478. Springer, 2018.

[3] Jun Bai, Yi Zeng, Yuxuan Zhao, and Feifei Zhao. Training a v1 like layer using gabor filters in convolutional neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision - ECCV 2018*, volume 11220 of *Lecture Notes in Computer Science*, pages 472–489. Springer, 2018.

[5] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8, 2020.

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

[7] Danton S. Char, Michael D. Abràmoff, and Chris Feudtner. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics*, 20(11):7–17, 2020.

[8] Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020.

[9] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *Proceedings of the Workshop on Artificial Intelligence Safety, SafeAI@AAAI 2020*, volume 2560, pages 63–73. CEUR-WS.org, 2020.

[10] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[11] Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CoRR*, abs/2009.10713, 2020.

[12] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[13] John Q Gan, Bashar Awwad Shiekh Hasan, and Chun Sing Louis Tsui. A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space. *International Journal of Machine Learning and Cybernetics*, 5(3):413–423, 2014.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020.

[15] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, 1990.

[16] Joshua James Hatherley. Limits of trust in medical ai. *Journal of Medical Ethics*, 46(7):478–481, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017.

[19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019.

[20] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

[21] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr-jpg - chest radiographs with structured labels (version 2.0.0). *PhysioNet*, 2019.

[22] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 4765–4774, 2017.

[23] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Lulu, Morisville, North Carolina, 2019.

[24] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *CoRR*, abs/2010.15775, 2020.

[25] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. arxiv 2018. *arXiv preprint arXiv:1810.08591*, 2018.

[26] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 4970–4979. PMLR, 2019.

[27] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962.

[28] Gregory Plumb, Denali Molitor, and Ameet Talwalkar. Model agnostic supervised local explanations. *arXiv preprint arXiv:1807.02910*, 2018.

[29] John W. Pratt and Jean D. Gibbons. *Kolmogorov-Smirnov Two-Sample Tests*, pages 318–344. Springer New York, New York, NY, 1981.

[30] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6076–6085, 2017.

[31] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.

[32] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8(4), 2018.

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[35] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[36] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

[37] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 10965–10976, 2019.

[38] Seyma Yucer, Samet Akcay, Noura Al-Moubayed, and Toby P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[39] 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):1–17, 11 2018.

[40] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.