Robustness to Input Distribution Shifts in Automatic MR Image Diagnosis

Editors: Under Review for MIDL 2023

Abstract

Magnetic Resonance (MR) imaging is an indirect process for generating images, in which the scanner first collects complex frequency space measurements of the electromagnetic activity within a human subject's body. Then radiologist renders the disease diagnosis by interpreting the magnitudes of the complex images that are reconstructed using the inverse Fourier transform of the acquired measurements. Recently, deep learning (DL) models have achieved state-of-the-art performance in diagnosing multiple diseases using these magnitude images as input. However, these models are susceptible to the change in the input data distribution between the training and testing phases, which can prevent their use in medical applications where safety is critical. In this work, we focus on improving the generalization performance of these models when they are used with low-field scanners, which can produce noisy images. We find that batch normalization, a common technique in deep learning, is a major cause of performance degradation in these changing environments. We show that using alternate normalizations, such as GroupNorm and LayerNorm, can improve the performance of the models by up to 10% in the diagnosis of various pathologies. To the best of our knowledge, this is the first study to investigate the use of these alternate normalizations for generalization to noisy MR images. We release the code online.

Keywords: List of keywords, comma separated.

1. Introduction

TBD

Magnetic Resonance Imaging (MRI) (Lauterbur, 1973) is non-invasive medical imaging technique that is considered the gold standard of diagnostic imaging because of the excellent soft-tissue contrast and non-ionising nature. It is an indirect imaging process in which the MR scanner collects measurements of the electromagnetic activity within the human subject's body after exposing the subject to a combination of magnetic field and radio frequency pulses. These measurements, which are complex valued, are collected in the frequency space called the *k*-space. The final gray scale volumetric image is generated from these measurements in two steps. First, a multi-dimensional inverse Fourier transform is used to generate an image in the complex space. Then the gray scale human interpretable image is generated by taking the magnitude of this complex image. These images (which we call "magnitude images") are then retained in order for the radiologist to render a diagnosis. The phase images, though used in quantitative applications in MRI, are typically not used in diagnostic capability.

In recent years, Deep Neural Networks (DNNs) have attracted significant attention for their role in enhancing MR imaging pipeline. The two primary areas where DNNs have shown great promise are 1) speeding up the measurement acquisition process by generating diagnostic quality images using under-sampled *k*-space data (Schlemper et al., 2018; Vellagoundar and Machireddy, 2015; Gözcü et al., 2018; Gözcü et al., 2019; Sanchez et al., 2020; Bakker et al., 2020), and 2) performing pathology segmentation or inferring the presence/absence of diseases from the volumetric gray scale

© 2023 .

images (Hammernik et al., 2018; Knoll et al., 2019; Liang et al., 2019; Putzky and Welling, 2019; Schlemper et al., 2018; Sriram et al., 2020). While DNNs have achieved state-of-the-art performance in accurately identifying a variety of diseases, across the board these models are often susceptible to the change in distributions.

In this research, we argue that blindly constraining the inputs of the DNN models to be the same as what a human would use to render a diagnosis is sub-optimal. We hypothesize that using the raw measurements as input to the DNNs, without regards to its human interpretability, will allow us to build systems that are more accurate at predicting diseases, by enabling the DNN models to automatically extract the most informative diagnostic signals.

Towards that end, we explore the utility of using alternate representations of the input signal to the DNN models and understand the role of various components of the complex data acquired by the MR scanners for disease identification. We accomplish this by dissecting the complex data into multiple components: *real, imaginary, magnitude*, and *phase* (see **??**). We then train DNN models with various combinations of these components on the publicly available fastMRI (Zbontar et al., 2018; Zhao et al., 2021) dataset, for the task of identifying multiple clinically significant pathologies from the MR scans. Our extensive experimental results show that it is beneficial to use alternate representations of the input to the DNN model as opposed to the reconstructed magnitude images. These results are particularly striking because until now the research community in this field has focused on training DNNs primarily using magnitude images. While MR physicists and researchers have used alternate representations, such as the phase, to deduce certain quantitative parameters of the MR scans (Funai et al., 2008; Poorter et al., 1995; Nielsen and Nayak, 2009; Chavez et al., 2002), we are not aware of any prior work that has used phase or other alternate signal representations as input for the disease identification task. We further establish the benefits of using alternate representations as input to DNN models, by measuring their classification performance on the noise degraded low signal-to-noise ratio (SNR) k-space data (one of the key characteristic of the data acquired by low-field inexpensive scanners). We show that DNN models trained using alternate representations are robust to input noise and in fact the benefits in terms of classification performance are only enhanced.

In summary, the contributions of our work are:

- We demonstrate the effectiveness of using alternate raw representations of the MR signal as input to DNNs for the task of disease prediction in comparison to the conventional MR prediction using the human interpretable magnitude images.
- We show that using the raw representations allows us to train DNNs that are more robust to noise by evaluating the learnt representations in the low SNR regime.
- We conduct qualitative analysis to demystify the usefulness of alternate representations showing that phase contains essential information as the hierarchical representations obtained using phase are similar to the representations learnt with magnitude and phase.

2. Related Work

Application of deep learning to MRI. Over the last years, DNNs have experienced significant success in solving image reconstruction and image analysis problems within the MR pipeline. Image reconstruction involves generating spatially resolved human interpretable images from the frequency

space data acquired by the scanner. DNNs have shown remarkable promise in improving the efficiency of this process by reconstructing diagnostic quality images from a fraction of the frequency space data, thereby significantly reducing the burden on the data acquisition process (Lustig et al., 2007; Haldar and Kim, 2019; Schlemper et al., 2018; Vellagoundar and Machireddy, 2015; Gözcü et al., 2018; Gözcü et al., 2019; Sanchez et al., 2020; Bakker et al., 2020). For image analysis, DNNs have been trained to perform tasks like tissue/organ segmentation or disease identification (Chen et al., 2016; Lakhani and Sundaram, 2017; Park et al., 2019; Moeskops et al., 2016; Milletari et al., 2016; Hwang and Kim, 2016; Liu et al., 2014; Pinaya et al., 2016; Hatakeyama et al., 2014; Schmah et al., 2008; Kim et al., 2019), with high accuracy. While successful, the training of these models is limited to using only the spatially resolved human interpretable gray scale images as input, where some information present in raw measurements captured by the MR scanner is discarded in favor of human eye interpretability.

Distribution shifts in medical imaging.

3. Magnetic Resonance Imaging (MRI)

In this section, we recap the data acquisition pipeline for MR Imaging and introduce some notations. The process of MR data acquisition involves exposing the human subject to varying magnetic field and radio-frequency pulses and capturing the resulting electromagnetic responses from the human body. These electromagnetic responses are measured by devices called *receiver coils* that are positioned in the vicinity of the tissue or organ being imaged. The measurements captured by the receiver coils are in the Fourier domain (a.k.a., *k*-space in the MR community). Measurements captured by a single coil can be represented as a 2D complex valued matrix.

More formally, let $\tilde{x} \in \mathbb{C}^{M\times N}$ denote the full *k*-space matrix associated with a single coil. We can then obtain a spatially resolved complex image $x \in \mathbb{C}^{M\times N}$ by taking the multi-dimensional inverse Fourier transform of the *k*-space data: $x = \mathcal{F}^{-1}(\tilde{x})$. That is:

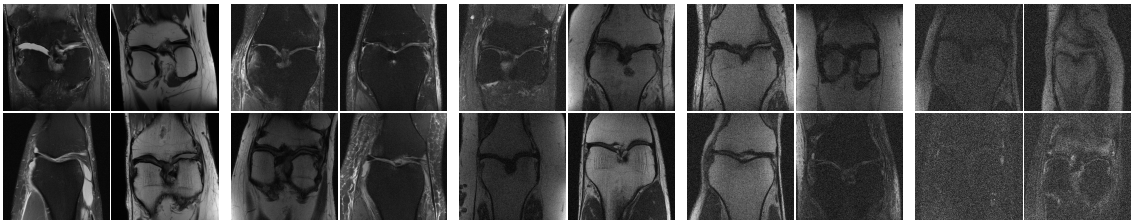
$$x_{hw} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=1}^{N-1}\tilde{x}_{mn}\left[\cos\left(2\pi\left(\frac{mh}{M}+\frac{nw}{N}\right)\right)+i\sin\left(2\pi\left(\frac{mh}{M}+\frac{nw}{N}\right)\right)\right], \quad (1)$$

where x_{hw} is the (h, w)-th element of the matrix x, $\tilde{x}_{mn}$ can be viewed as the Discreet Fourier Transform (DFT) of x_{mn}, $0 \le h \le M-1$, and $0 \le w \le N-1$. The magnitude of the resulting complex image x is denoted by $|x_{hw}| = \sqrt{\Re(x_{hw})^2 + \Im(x_{hw})^2}$ and $\angle x_{hw} = \arctan\left(\Im\left(x_{hw}\right)/\Re\left(x_{hw}\right)\right)$ represents the phase. Here $\Re(.)$ and $\Im(.)$ refer to the real and imaginary parts of a complex number.

MR scanners use multiple receiver coils in parallel to acquire measurements, where each coil produces a partial image with different spatial sensitivity across different parts of the volume. Specifically, the data-acquired by a single coil (denoted by c) can be represented as:

$$\tilde{x}^c = F\left(\mathcal{C}_c x^c\right) + \varepsilon_c, \quad c \in \{1, \ldots, n_c\}, \quad (2)$$

where $\tilde{x}^c$ are the *k*-space measurements acquired by the c-th coil, $\mathcal{C}_c$ is the sensitivity map of the coil, $\varepsilon \sim \mathcal{N}(0, I\sigma^2) + i\mathcal{N}(0, I\sigma^2)$ is the noise due the measurement errors by the scanner, and n_c is the number of receiver coils. The generation of ground truth images using the data from multiple coils consists of two steps. The first step involves an Inverse Fourier Transform of the *k*-space data from an individual coil using Equation 1 to generate coil specific images. In the second step, we generate



(a) Original (b) SNR 50 (c) SNR 20 (d) SNR 10 (e) SNR 5

(f) Original (g) Motion effect (h) Ghosting (i) Spike (j) Field

Figure 1: Visualization of simulated low-SNR MR images for the ground truth images having a Rician distribution (Rice, 1944; Gudbjartsson and Patz, 1995b) **(top)** and emulated single coil images with independent Gaussian noise for the real and imaginary channels **(bottom)**.

a single image by combining the magnitudes of the individual coil images voxel by voxel using the root-sum-of-squares method (Roemer et al., 1990) as follows:

$$x^{\text{RSS}} = \left(\sum_{c=0}^{n_c} |x^c|^2 \right)^{1/2}. \tag{3}$$

4. Disease Prediction with Distribution Shifts

4.1. Motivation

The acquisition of MRI data is often a combination of multiple signals, which introduces background noise (see Equation 2). In fact among various factors, noise is by-far the biggest factor that contributes towards image deterioration when going from a high-field scanner (scanner with high magnetic field strength) to a low-field scanner (scanner with low magnetic field strength). Despite being inexpensive these scanners have not seen widespread adoption because of the poor diagnostic quality of the images generated by them. Evaluation of any methodology on low signal-to-noise ratio (SNR) images is important for it gives some insights into how the methodology will generalize to low-field scanners, potentially leading to transformative impact in healthcare by facilitating the use of these scanners which are inexpensive and relatively accessible.

Due to the dearth of data collected from the low-SNR MR scanners, we make an attempt to simulate the low-SNR raw *k*-space data. In particular, we know that the noise in the data acquired in the raw *k*-space is Gaussian distributed, which leads to the ground truth images having noise that has a Rician distribution (Rice, 1944; Gudbjartsson and Patz, 1995b). More formally, the noisy signal

can be defined as:

$$\begin{aligned}\Re\left(\mathbf{x}_k^{\text{ESC}}\right) &= \Re\left(x_k^{\text{ESC}}\right) + \mathcal{N}_1(0,\sigma);\\ \Im\left(\mathbf{x}_k^{\text{ESC}}\right) &= \Im\left(x_k^{\text{ESC}}\right) + \mathcal{N}_2(0,\sigma),\end{aligned} \tag{4}$$

Furthermore, we consider the injection of noise to the raw emulated signal-coil data directly. More specifically, let $\mathbf{x}_k^{\text{ESC}} \in \mathbb{C}^{M\times N}$ denote the low-SNR complex images with SNR k obtained after the inverse Fourier transform of $\tilde{x}^{\text{ESC}}$. To degrade the signal, we add Gaussian noise independently to both the real and imaginary channels following Gudbjartsson and Patz (1995a). More formally, the injection of noise in this case can be defined as:

$$\begin{aligned}\Re\left(\mathbf{x}_k^{\text{ESC}}\right) &= \Re\left(x_k^{\text{ESC}}\right) + \mathcal{N}_1(0,\sigma);\\ \Im\left(\mathbf{x}_k^{\text{ESC}}\right) &= \Im\left(x_k^{\text{ESC}}\right) + \mathcal{N}_2(0,\sigma),\end{aligned} \tag{5}$$

where $\sigma = \text{Signal Strength}/k$. Here, we define the signal strength for each input sample as the maximum pixel intensity. Particularly, signal strength $= \max\left(|x^{\text{ESC}}|\right)$ and $\mathcal{N}_1$ and $\mathcal{N}_2$ are two distinct samples of the normal distribution. The SNR is directly proportional to the signal strength as higher field strength will consequently increase the number of protons that align with the magnetic field, which provides higher SNR and higher resolution images.

4.2. Problem Setup

We consider the task of disease prediction, where we are given the data distribution acquired in the frequency domain $\mathcal{D}^{\text{ESC}} = \{(\widehat{x}_i^{\text{ESC}}, y_i)_{i=1}^{\mathcal{N}}\}$ and ground-truth MR images $\mathcal{D} = \{(x_i^{\text{RSS}}, y_i)_{i=1}^{\mathcal{N}}\}$ with $\mathcal{N}$ examples and $y_i \in \mathbb{R}^K$ denote the corresponding K pathologies. For notation simplicity, we denote the input images as x for the rest of the paper. We consider a classifier $f_\Theta : \mathbb{R}^d \to K$ with L layers and Θ as parameters, which takes as input the real-valued MR images and generates as output the set of class conditional probabilities (the output of a Softmax function). We train this using the multi-class cross-entropy loss. Specially, the learning objective for disease classification with the cross-entropy loss $\mathcal{L}$ is given by:

$$\underset{\Theta}{\text{minimize}}\ \ \mathbb{E}_{(x,y)\sim\mathcal{D}_{tr}}\Big[\mathcal{L}(f_\Theta(x), y)\Big], \tag{6}$$

where $\mathcal{D}_{tr}$ denotes the training data distribution. In real-world applications, the input data distribution can change during the inference phase due various reasons. More specifically, the conditional distribution $p_{tr}(y|x) = p_{te}(y|x)$, but $p_{tr}(x) \neq p_{te}(x)$. This is called covariate shift (**?**Schölkopf et al., 2012).

4.3. Feature Normalization Layers

Following Wu and He (2018), the general formulation of feature normalization for features $a_i^\ell \in \mathbb{R}^{N\times C\times H\times W}$ defined as:

$$\widehat{a}_i^\ell = \frac{1}{\sigma_i}(a_i - \mu_i). \tag{7}$$

μ_i and σ_i denote the feature statistics:

$$\mu_i = \frac{1}{m}\sum_{k\in\mathcal{S}_i} a_k;\ \sigma_i = \sqrt{\frac{1}{m}\sum_{k\in\mathcal{S}_i}(a_k - \mu_k)^2 + \epsilon} \tag{8}$$

Specifically, different kinds of normalization can be defined as follows:

- **Batch normalization (BN) (Ioffe and Szegedy, 2015)** is a standard component in deep learning architectures, which normalizes the features of the intermediate layers using the mean and standard deviation computed across the training data: $\mathcal{S}_i = \{k|k_C = i_C\}$.

- **Adaptive Batch normalization (AdaBN) (Li et al., 2016)** recomputes the BN statistics using the exponential moving average across the test distribution. This improves the model generalization when the input distribution changes during testing (Schneider et al., 2020; Nado et al., 2020; Benz et al., 2021).

- **Layer normalization (LN) (Ba et al., 2016)** is a method that computes the mean and standard deviation across all the features in a layer: $\mathcal{S}_i = \{k|k_N = i_N\}$. It performs the same computation for the training and testing phases.

- **Group normalization (GN) (Wu and He, 2018)** divides channels into G groups and computes the mean and standard deviation across groups. $\mathcal{S}_i = \{k|k_N = i_N, \lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor\}$. Similar to layer normalization, group normalization is independent of the batch dimension.

To obtain the spatial real images for input to the deep neural network, we construct the real-valued images by considering the magnitude and phase as two individual channels. For notational simplicity, let $M \times N = P$, $x^{\text{ESC}} \in \mathbb{C}^P$. Then the equivalent real image can be defined as $\widehat{x}^{\text{ESC}} = (|x^{\text{ESC}}|, \angle x^{\text{ESC}}) \in \mathbb{R}^{2P}$. While it has been observed that information contained in phase is adequate for the reconstruction of the signal (Huang et al., 1975; Oppenheim et al., 1979; Oppenheim and Lim, 1981), it is still an open question whether phase is useful for automatic clinical diagnosis, since in $\widehat{x}^{\text{ESC}}$ is discarded in favor of human eye interpretability. In order to understand the usefulness of phase for DNNs, we compare networks learnt using only the magnitude or the phase components of the signal. Specifically, the training objective can be formulated as:

5. Experiments

5.1. Experimental Setup

Data Sets. The *fastMRI* (Zbontar et al., 2018) was the first large-scale dataset that consisted of raw *k*-space data and anonymized clinical MR images. The release of this dataset led to the development of various deep-learning based reconstruction approaches (Hammernik et al., 2018; Recht et al., 2020) for the acceleration of MRI acquisition. However, despite the scale of this dataset, it could not be utilized for automated diagnosis due to the absence of annotations until Zhao et al. (2021) released *fastMRI+*. Specifically, fastMRI+ consists of human-annotated bounding box labelled knee and brain pathology annotations. The release of these annotations encouraged us to investigate the expressiveness of individual components in the raw *k*-space data for automated diagnosis. We consider the *knee pathologies*, where we use the emulated single-coil *k*-space data with slice-level labels using fastMRI+ (Zhao et al., 2021) for the three most significant pathalogies as suggested by the clinicians for Knee. Concretely, we consider 1) *Anterior Cruciate ligament (ACL)* with 1,443 annotations of 254 subjects, 2) *Meniscus tear* with 5,658 annotations of 663 subjects, and 3) *Cartilage* with 3,600 annotations of 710 subjects. The slices were cropped to 320 $\times$ 320 and 15% of the dataset for used for validation and test sets.

Models and Metrics. For the baseline model, we conduct our experiments with PreactResNet-18 (He et al., 2016) and report the mean and standard deviation for all our experimental analysis

Table 1: **AUROC** of PreactResNet-18 on the fastMRI dataset. We report the mean and 95% CI across five runs. The network representations learnt with phase or real component in isolation achieve better performance than magnitude of the signal.

(a) ACL

NOISE	BN	ADABN	GN	LN
ORIGINAL	87.78 (0.59)	85.63 (1.71)	87.35 (0.55)	86.40 (1.69)
MOTION	64.14 (2.48)	55.29 (2.03)	**67.03 (2.13)**	65.44 (2.39)
GHOSTING	80.64 (1.14)	78.25 (1.37)	78.54 (0.73)	**82.45 (1.17)**
SPIKE	63.58 (2.09)	66.73 (4.11)	**81.20 (3.16)**	80.66 (1.89)
FIELD	62.71 (2.78)	61.41 (2.74)	**68.26 (2.41)**	64.99 (1.46)
MEAN	67.77 (3.71)	65.42 (2.76)	**73.76 (2.28)**	**73.39 (3.15)**
WORSE	62.71 (2.78)	55.29 (2.03)	**67.03 (2.13)**	64.99 (1.46)

(b) Menicus Tear

NOISE	BN	ADABN	GN	LN
ORIGINAL	92.04 (0.48)	91.02 (0.57)	91.82 (0.98)	91.67 (0.80)
MOTION	**78.66 (1.56)**	73.59 (1.52)	76.15 (1.36)	77.35 (1.40)
GHOSTING	87.16 (0.35)	86.86 (0.42)	83.80 (0.59)	**88.18 (1.99)**
SPIKE	63.72 (3.02)	71.53 (2.46)	**83.69 (3.91)**	83.62 (1.94)
FIELD	65.23 (0.45)	62.73 (2.26)	**71.15 (3.20)**	70.37 (2.63)
MEAN	72.94 (1.72)	73.68 (1.85)	78.70 (2.63)	**79.88 (2.04)**
WORSE	63.72 (3.02)	62.73 (2.26)	**71.15 (3.20)**	70.37 (2.63)

(c) Cartilage

NOISE	BN	ADABN	GN	LN
ORIGINAL	88.70 (0.50)	87.57 (0.76)	88.01 (1.36)	88.41 (0.77)
MOTION	**74.29 (1.93)**	71.18 (1.67)	71.00 (2.04)	74.12 (2.03)
GHOSTING	83.54 (1.02)	82.88 (1.24)	79.85 (0.49)	**85.27 (1.21)**
SPIKE	61.27 (3.95)	65.11 (1.13)	80.23 (2.02)	**81.57 (0.98)**
FIELD	61.36 (1.49)	60.17 (2.01)	66.72 (4.45)	**67.72 (2.98)**
MEAN	70.12 (2.38)	69.84 (2.20)	74.45 (2.96)	**77.17 (2.96)**
WORSE	61.27 (3.95)	60.17 (2.01)	66.72 (4.45)	**67.72 (2.98)**

(d) Others

NOISE	BN	ADABN	GN	LN
ORIGINAL	72.61 (0.83)	72.08 (1.23)	76.33 (1.75	76.44 (1.72)
MOTION	60.13 (1.43)	58.49 (0.99)	63.01 (0.93)	**63.03 (2.19)**
GHOSTING	67.84 (0.65)	68.25 (1.08)	66.04 (2.14)	**71.86 (2.01)**
SPIKE	63.14 (4.24)	62.71 (3.88)	68.80 (3.15)	**69.73 (1.57)**
FIELD	54.24 (1.42)	53.64 (2.45)	**58.58 (1.19)**	57.32 (1.71)
MEAN	61.34 (2.37)	60.78 (2.41)	64.11 (2.05)	65.49 (1.89)
WORSE	54.24 (1.42)	53.64 (2.45)	**58.58 (1.19)**	57.32 (1.71)

across five independent runs. The hyper-parameters and other experimental settings are provided in the supplementary material. We evaluate all our experiments with the two standard metrics:

1. **AUROC (BRA, 1997)** is a common metric used for the evaluation of medical problems solved in machine learning (Rajkomar et al., 2018; Rajpurkar et al., 2017; Solares et al., 2020; Tschandl et al., 2019; Teixeira et al., 2017; Dunnmon et al., 2019). It computes the area under the receiver operating characteristic curve and can be calculated by the area under the False Positive Rate (FPR) against the True Positive Rate (TPR) curve.

2. **Balanced Accuracy (García et al., 2009).** Due to the class-imabalance across pathologies in the knee data set, we use this metric as it gives equal weight to both majority and minority classes and is commonly used in the presence of class imbalance learning (Khan et al., 2019). It is computed as the average of the sensitivity and specificity, where sensitivity and sensitivity are the true positive rate and true negative rate respectively.

5.2. Quantitative Results

Evaluation on the knee dataset. Table 1 shows the results for fully-sampled k-space data with the knee pathologies. First, we observe that model trained with magnitude and phase can outperform the magnitude models across all the pathologies. Interestingly, the representations learnt with phase in isolation can surpass the performance obtained by the representations trained with the magnitude signal. Second, we demonstrate that the model trained with real and imaginary component as two separate channels show an absolute improvement of $2.21\%, 3.60\%,$ and 2.28% in comparison to magnitude on ACL, meniscus tear, and cartilage on the AUROC respectively showing that DNNs can extract clinically predictive information from the raw measurements. We show that real component

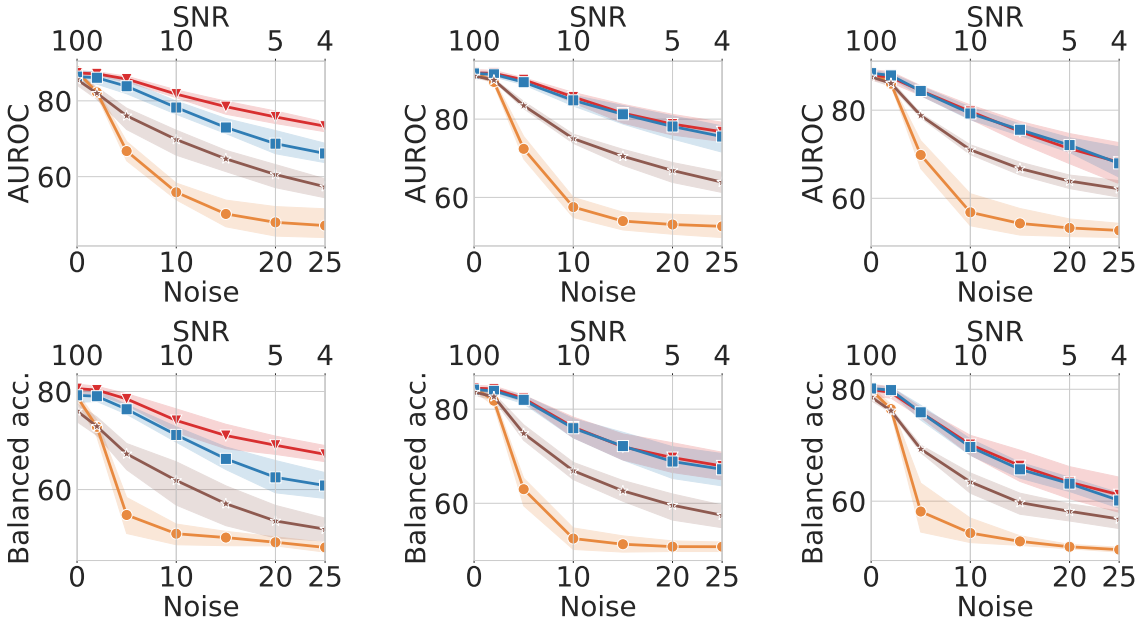


Figure 2: Comparison of PreactResNet-18 with **batch normalization**, **group normalization**, **layer normalization**, and **adaptive batch normalization** on ACL (**left**), Meniscus Tear (**middle**), and cartilage (**right**) root sum of square images on lower SNR images. We observe that models trained with group or layer normalization can obtain significantly better performance than magnitude for low-field MRI. The mean and standard deviation are over five independent runs.

in isolation can obtain comparable performance to the model trained with real and imaginary. Contrarily, imaginary component deteriorates the performance of the learnt representations. Third, we demonstrate that even-though x^{RSS} represents higher-SNR images, alternate representations can comparatively achieve better perform for all pathologies. This result is particularly striking because the high SNR of RSS images can be attributed to the multi-coil data used to generate these images. Whereas the alternate representations are derived by merging the data from multiple coils and emulating it to be coming from a single-coil, which inherently results in a relatively low-SNR.

Evaluation on low-SNR MRI. To validate the generalization of our empirical observations and show its scalability for mass population-level testing, we evaluate the trained models on low-SNR MR images simulated using Equation 5 in Figure 10. We observe that the representations learnt with magnitude deteriorate $\sim 20\%$ in AUROC on low-SNR signal when the signal is deterriorated from the standard acquired value to a value of four. While, the network learnt with the phase component and the real-valued image containing the magnitude and phase outperform the other methods for a high-SNR scanner. In contrast, phase in isolation and the combination of real and imaginary achieve state-of-the-art performance on simulated low-SNR scanners for all the pathologies. We defer the results for balanced accuracy and real, imaginary components evaluated on low-SNR MR images to the supplementary material.

6. Conclusion, Limitations, and Future Work

Despite the complex acquisition of the raw complex *k*-space data from the MR scanner, it has been a common practice to discard the phase information due to the lack of human interpretable

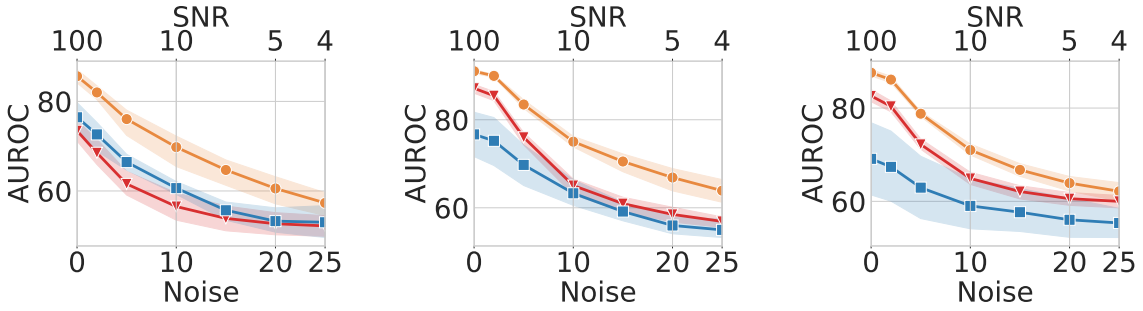


Figure 3: Comparison of PreactResNet-18 trained with adapting **both mean and variance**, **mean only**, and **variance only** on ACL (**left**), Meniscus Tear (**middle**), and cartilage (**right**) for grounth truth images (**top**) and emulated single-coil images (**bottom**). We observe that the adaptation of both mean and variance is essential to obtain better performance for noisy settings. The mean and standard deviation are over five independent runs.

signals when applying deep neural networks for clinical diagnosis from the MR images. However, in this work, we question that whether in this process we discard information that is predictive of the underlying diagnostic task. To address this question, we run an extensive set of quantitative experiments on the fastMRI knee pathologies, where we find that phase is almost all we need to build a highly predictive learner. Further, we demonstrate that the raw measurements containing the real and imaginary signal is also more beneficial than using the magnitude reconstructed image. To this end, we show the generalization of these alternate representations to low-SNR regime, where our result implies that one can use low SNR data acquired by low-field scanners and get similar diagnostic performance as the high-field scanner, so long as we use alternate representation of the input signal to the DNNs. To gain understanding into this observation, we run an exhaustive set of analysis, where we show that the representations learnt by the phase as input shows hierarchical structure similar to the network learnt with the combination of magnitude and phase and has higher overlap between the output predictions.

Therefore, our analysis implies that we were indeed throwing away quite a lot of diagnostic information. However, we do not claim that only the phase or the real component of the signal contains the necessary and sufficient diagnostic information for all pathologies or anatomical areas. The key takeaway from our research is that in medical imaging, especially within the MR pipeline there may be representations of the data which is readily available but not interpretable, but nevertheless may contain key diagnostic information that can be leveraged by a deep learning mode for disease identification or similar tasks. We show that in the particular case of knee, it seems to be the case that the alternate representations contain useful information. Blinding training DNNs on only human interpretable images might be a sub-optimal strategy.

While the results are interesting, there are several limitations of the proposed research. First, this is just a first step in this direction of exploring alternate signals for training DNNs for disease identification. The proposed work is far from being used in real clinical setting even for knees. In order to use it in real clinical setting we need to do rigorous testing including a thorough retrospective and prospective analysis. Second, we need to assess the generalizability of the results on other diseases from other anatomies. We intend to do that as part of our future work. Third, our analysis was restricted to using emulated single-coil data. In practice, the magnitude images are generated using data from multiple coils which have very high SNR. In the future, we plan to use alternate

representations composed of data from all the coils as opposed to first merging it to emulate single coil. Additionally, we believe that investigation of alternative tasks such as segmentation and detection would be a promising direction for future work.

References

The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 1997.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Tim Bakker, Herke van Hoof, and Max Welling. Experimental design for mri by greedy policy search. In *nips*, 2020.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. 2021.

Sofia Chavez, Qing-San Xiang, and Li An. Understanding phase maps in mri: a new cutline phase unwrapping method. *IEEE transactions on medical imaging*, 2002.

Hao Chen, Qi Dou, Lequan Yu, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, 2016.

Jared A. Dunnmon, Darvin Yi, C. Langlotz, Christopher Ré, D. Rubin, and Matthew P. Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 2019.

Amanda Funai, Jeffrey Fessler, Desmond Yeo, Valur Olafsson, and Douglas Noll. Regularized field map estimation in mri. *IEEE transactions on medical imaging*, 2008.

Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. In *IbPRIA*, 2009.

Baran Gözcü, Rabeeh Karimi Mahabadi, Yen-Huan Li, Efe Ilıcak, Tolga Cukur, Jonathan Scarlett, and Volkan Cevher. Learning-based compressive mri. *IEEE transactions on medical imaging*, 2018.

Hakon Gudbjartsson and Samuel Patz. The rician distribution of noisy mri data. *Magnetic Resonance in Medicine*, 1995a.

Hákon Gudbjartsson and Samuel Patz. The rician distribution of noisy mri data. *Magnetic resonance in medicine*, 1995b.

Baran Gözcü, Thomas Sanchez, and Volkan Cevher. Rethinking sampling in parallel mri: A data-driven approach. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

Justin P Haldar and Daeun Kim. Oedipus: An experiment design framework for sparsity-constrained mri. *IEEE transactions on medical imaging*, 2019.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 2018.

Yutaka Hatakeyama, Shinichi Yoshida, Hiromi Kataoka, and Yoshiyasu Okuhara. Multi-voxel pattern analysis of fmri based on deep learning methods. In Keon Myung Lee, Seung-Jong Park, and Jee-Hyong Lee, editors, *Soft Computing in Big Data Processing*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

T. Huang, J. Burnett, and A. Deczky. The importance of phase in image processing filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1975.

Sangheum Hwang and Hyo-Eun Kim. Self-transfer learning for weakly supervised lesion localization. In *International conference on medical image computing and computer-assisted intervention*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

Salman Hameed Khan, Munawar Hayat, Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Hyun-Chul Kim, Peter A. Bandettini, and Jong-Hwan Lee. Deep neural network predicts emotional responses of the human brain from functional magnetic resonance imaging. *NeuroImage*, 2019.

Florian Knoll, Kerstin Hammernik, Chi Zhang, Steen Moeller, Thomas Pock, Daniel K Sodickson, and Mehmet Akcakaya. Deep learning methods for parallel magnetic resonance image reconstruction. *arXiv preprint arXiv:1904.01112*, 2019.

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 2017.

P.C. Lauterbur. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. In *Nature*, 1973.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Dong Liang, Jing Cheng, Ziwen Ke, and Leslie Ying. Deep mri reconstruction: Unrolled optimization algorithms meet neural networks. *arXiv preprint arXiv:1907.11711*, 2019.

Siqi Liu, Sidong Liu, Weidong Cai, Sonia Pujol, Ron Kikinis, and Dagan Feng. Early diagnosis of alzheimer's disease with deep learning. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 2014.

Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 2007.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International conference on 3D vision (3DV)*, 2016.

Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Jon-Fredrik Nielsen and Krishna S. Nayak. Referenceless phase velocity mapping using balanced ssfp. *Magnetic Resonance in Medicine*, 2009.

A. Oppenheim, Jae Lim, G. Kopec, and S. Pohlig. Phase in speech and pictures. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979.

A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 1981.

Beomhee Park, Yongwon Cho, Gaeun Lee, Sang Min Lee, Young-Hoon Cho, Eun Sol Lee, Kyung Hee Lee, Joon Beom Seo, and Namkug Kim. A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-pa x-ray screening for pulmonary abnormalities. *Scientific Reports*, 2019.

Walter Pinaya, Ary Gadelha, Orla Doyle, Cristiano Noto, André Zugman, Quirino Cordeiro, Andrea Jackowski, Rodrigo Bressan, and João Sato. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports*, 2016.

John De Poorter, Carlos De Wagter, Yves De Deene, Carsten Thomsen, Freddy Ståhlberg, and Eric Achten. Noninvasive mri thermometry with the proton resonance frequency (prf) method: In vivo results in human muscle. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 1995.

Patrick Putzky and Max Welling. Invert to learn to invert. In *nips*, 2019.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc V. Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam Haresh Shah, Atul Janardhan Butte, Michael D. Howell, Claire Cui, Greg Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 2018.

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

Michael P Recht, Jure Zbontar, Daniel K Sodickson, Florian Knoll, Nafissa Yakubova, Anuroop Sriram, Tullie Murrell, Aaron Defazio, Michael Rabbat, Leon Rybak, et al. Using deep learning to accelerate knee mri at 3 t: results of an interchangeability study. *AJR. American journal of roentgenology*, 2020.

Stephen O Rice. Mathematical analysis of random noise. *The Bell System Technical Journal*, 1944.

Peter B Roemer, William A Edelstein, Cecil E Hayes, Steven P Souza, and Otward M Mueller. The nmr phased array. *Magnetic resonance in medicine*, 1990.

Thomas Sanchez, Baran Gözcü, Ruud B van Heeswijk, Armin Eftekhari, Efe Ilıcak, Tolga Çukur, and Volkan Cevher. Scalable learning-based sampling optimization for compressive dynamic mri. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony N. Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 2018.

Tanya Schmah, Geoffrey E Hinton, Steven Small, Stephen Strother, and Richard Zemel. Generative versus discriminative training of rbms for classification of fmri images. In *Advances in Neural Information Processing Systems*, 2008.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

José Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H. Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, Nathalie Conrad, Kazem Rahimi, and Gholamreza Salimi Khorshidi. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of biomedical informatics*, 2020.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

Pedro L. Teixeira, Wei-Qi Wei, Robert Michael Cronin, Huan Mo, Jacob P. VanHouten, Robert J. Carroll, Eric R. LaRose, Lisa A. Bastarache, Samuel T. Rosenbloom, Todd L. Edwards, Dan M. Roden,

Thomas A. Lasko, Richard A. Dart, A Maleev Nikolai, Peggy L. Peissig, and Joshua Charles Denny. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *Journal of the American Medical Informatics Association*, 2017.

Philipp Tschandl, Cliff Rosendahl, Bengu Nisa Akay, Giuseppe Argenziano, Andreas Blum, Ralph Peter Braun, Horacio Cabo, J Y Gourhant, Jürgen Kreusch, Aimilios Lallas, Jan Lapins, Ashfaq A. Marghoob, Scott W Menzies, Nina Maria Neuber, John Paoli, Harold S. Rabinovitz, Christoph Rinner, Alon Scope, H. Peter Soyer, Christoph Sinz, Luc Thomas, Iris Zalaudek, and Harald Kittler. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatology*, 2019.

Jaganathan Vellagoundar and Ramasubbareddy Machireddy. A robust adaptive sampling method for faster acquisition of mr images. *Magnetic resonance imaging*, 2015.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W Lui, Michael S Hansen, and Matthew P Lungren. fastmri+: Clinical pathology annotations for knee and brain fully sampled multi-coil mri data. *arXiv preprint arXiv:2109.03812*, 2021.

Supplementary Material
Importance of Alternate Signal Representations in Automatic MR Image Diagnosis

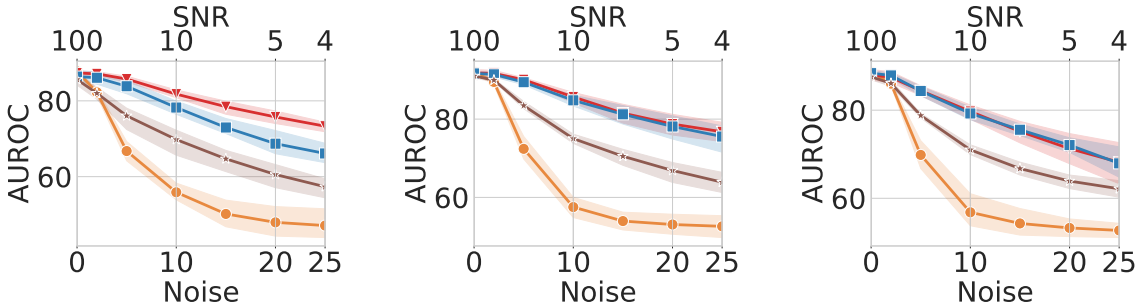


Figure 5: Noise

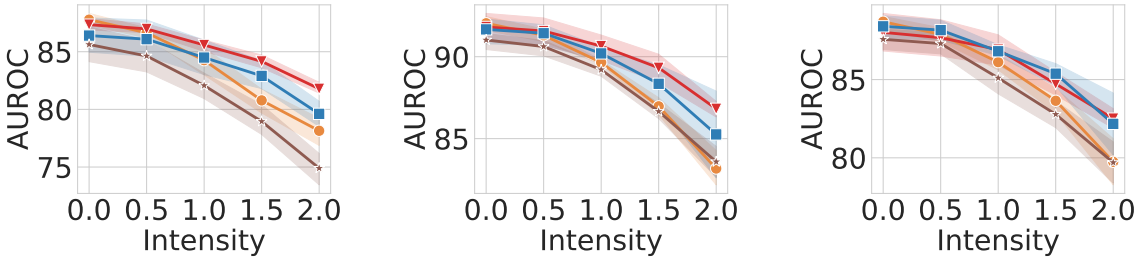


Figure 6: Ghosting

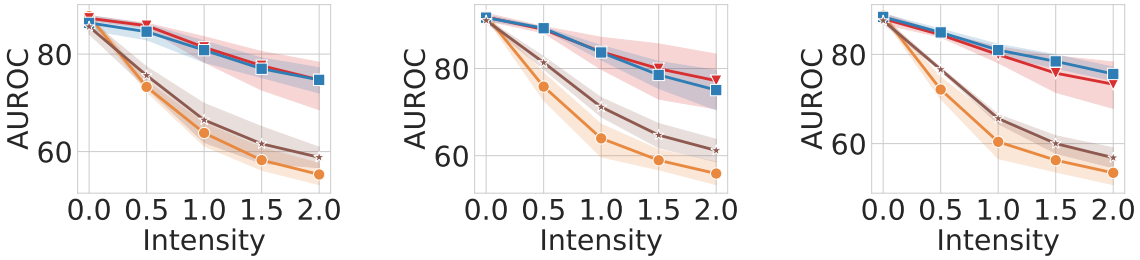


Figure 7: Spike

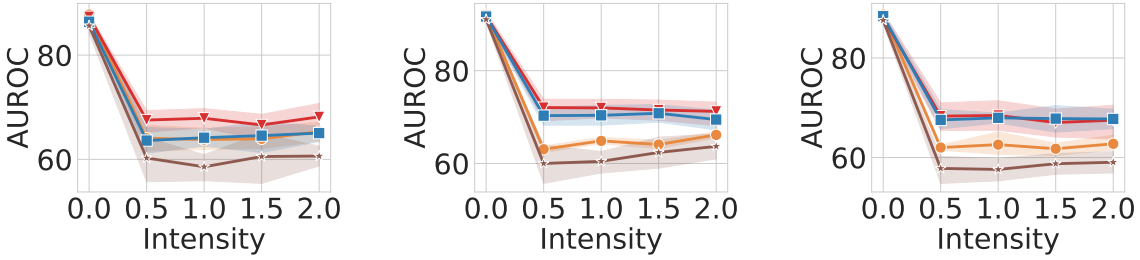


Figure 8: Field

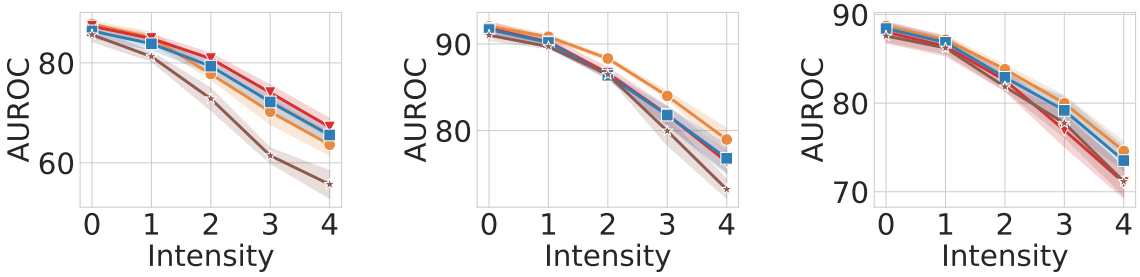


Figure 9: Motion, degree = intensity * 5, translation=intensity * 2

Figure 10: Comparison of PreactResNet-18 trained with **batch normalization**, **group normalization**, **layer normalization**, and **adaptive batch normalization** on ACL (**left**), Meniscus Tear (**middle**), and cartilage (**right**). We observe that phase in isolation can obtain significantly better performance than magnitude for low-field MRI. The mean and standard deviation are over five independent runs.