**Distribution Agreement**

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Fahd Hatoum April 8, 2025

# Examination of the Double Descent Phenomenon in Medical Imaging AI

by

Fahd Hatoum

Justin Clifford Burton
Adviser

Physics Department

Justin Clifford Burton

Adviser

Tankut Can

Committee Member

Ilya Nemenman

Committee Member

2025

# Examination of the Double Descent Phenomenon in Medical Imaging AI

By

Fahd Hatoum

Justin Clifford Burton

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Physics Department

2025

Abstract

Examination of the Double Descent Phenomenon in Medical Imaging AI

By Fahd Hatoum

One of the reasons behind the success of neural networks in radiology and other fields (e.g.: natural language processing) is that the models learn complex features from the training data and generalize well to new data. The flexibility and success of the models is due to over-parameterization. Generally, in order to produce an over-parameterized model, researchers use a simple heuristic: the number of parameters in the model should be much greater than the number of training samples. Given the differences between natural and medical images, this heuristic when applied by researchers in the medical imaging community, might lead them to develop critically parameterized or under-parameterized models that lead to worse performance. As such, in this thesis, we aim to investigate whether a commonly used model (Densenet 121 model) in medical imaging research is under-parameterized or over-parameterized for a certain combination of factors. These factors include transfer learning, data set size, data complexity and model width. We restricted the task of the model to a simple binary classification of disease from a patient chest radiograph. We find that for certain training sizes, the model is in the critically parameterized regime and a tenfold increase in the sample size does not yield better performance. We also find that diseases that are more challenging to diagnose (such as COVID-19) typically shift the interpolation threshold to the right and cause the model to become over-parameterized. Given these results, we find that it is important for researchers in the medical imaging field to not use heuristics as the ones researchers in computer vision use to develop deep learning models for natural images.

# Examination of the Double Descent Phenomenon in Medical Imaging AI

By

Fahd Hatoum

Justin Clifford Burton

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Physics Department

2025

Acknowledgements

I would like to first tank Dr. Tankut Can and Dr. Ilya Nemenman for taking the time to meet with me during my thesis write-up as well as for being on my committee. Second, I would like to thank Dr. Justin Burton for volunteering to be my advisor at Emory. His act of kindness allowed me to conduct the research that I am passionate about. I would also like to thank him for the guidance and patience he has extended to me in the process. Third, I would like to thank both Dr. Alfred Farris and Dr. Effrosyni Seitaridou without whom I would have never had the opportunity to conduct any sort of research. I am grateful to them for inspiring me to become a better scientist and a better person. Finally, I would like to thank Dr. Maryellen Giger and her lab group at the University of Chicago without whom this thesis would not have been possible. During the summer before my senior year, I interned in Dr. Giger's lab and under her guidance I discovered how one could bridge the gap between basic science and translational research. Working with her has inspired and shaped my current path to graduate school. However, above all, I would like to thank her for being a very kind and caring mentor. Ultimately, this thesis represents the culmination of my research journey under the tutelage of all the outstanding mentors I have had the privilege to learn and grow from. I owe all of them my sincere gratitude.

# Table of Contents

# List of Figures

# 1. Introduction

## 1.1. AI and its Impact on the Radiological Field.

Artificial intelligence (AI) is increasingly being integrated into various aspects of healthcare, decision support and medical imaging. The most recent estimate from the FDA states that there are 950 AI-enabled systems approved for use in the medical field with most (i.e.: 76%) being developed for the radiology field (medical imaging) [1-4]. The major uptake of AI in medical imaging is due to the increased workload of radiologists and the advances of machine learning in computer vision.

The use cases of such systems in radiology range from detecting disease regions in chest radiographs to assisting radiologists in the diagnosis and prognosis of diseases [5-9]. These systems can often outperform radiologists especially when identifying more pathologically challenging diseases [9-10].

## 1.2. The Schism

As stated earlier, many of the advances of the ML algorithms developed for medical images originate from advances in computer vision (transfer learning, adversarial networks, base models used…). However, there still exists a gap between the theoretical and empirical knowledge produced by the machine learning community and the subsequent applications in the medical imaging field. One such instance is the case of the double descent phenomenon in ML and deep learning (DL) algorithms (only one publication claims to have observed double descent for medical images) [11].

## 1.3. What is Double Descent?

In neural networks, the double descent phenomenon describes the fluctuation of the test error with respect to model complexity. Model complexity is affected by the model size (number of trainable parameters), optimization procedure (training time, presence or absence of regularization…), and data complexity (number of samples, intrinsic dimensions, class overlap…) [12-14]. Thus, a change in any of these factors leads to a change in model complexity. The double descent phenomenon posits that over-parameterized models (models that have a high complexity and that interpolate all points from the training data) perform better (lower test error) than under-parameterized models (models with lower complexity). In addition, with double descent, models in the under-parameterized regime present the bias-variance tradeoff as observed in classical statistics. In this regime, a model with an increased complexity begins to perform worse (larger test error) than models with lower complexity (a U-shaped curve is observed in the graph plotting test error against model complexity) [12-14]. The intersection of these two regimes is called the interpolation threshold and a peak is observed in the graph plotting test error against model complexity. The region around this threshold is the critically parameterized regime because any increase or decrease in the model complexity can cause either an increase or decrease in the test error [13].

## 1.4. Double Descent in Practice

In practice and in most of the current literature describing the double descent, the factor/hyper parameter that governs model complexity is the size of the model (number of model parameters). This choice was made to give a description that is aligned with the

current literature and because, in practice, the size of the model is the hyper parameter that one has the flexibility to adjust.

Here, it becomes important to note that regularization techniques such as lasso and ridge regularization, early stopping, and data augmentation have been found to mitigate the presence of double descent [15]. With these techniques, the test error generally continues to decrease with an increase in model complexity. However, even in the presence of these mitigation techniques, double descent might also occur [12]. As such, it is important that developed ML models are over-parameterized with respect to the training data distribution.

This is not currently an issue with radiology datasets as most are small (less than 10,000 images) and the models used are currently complex enough (in terms of size). However, with the increase of publicly available radiology datasets, it becomes important for researchers to scale the complexity of their models accordingly to observe an increase in performance (to escape the under-parameterized or critically parameterized regimes) [16-17]. However, the limits of scaling the number of parameters and the data size in tandem are not well documented for medical images. Furthermore, as mentioned earlier, since model complexity is affected by the data complexity, it is not assured that models built for natural images are also over-parameterized with respect to the distribution of radiology images.

## 1.5. Radiology vs. Natural Image Classification

Radiology datasets typically have intrinsic features that are more complex to learn than those of natural images due to the similarity between images (e.g., presence of the same background features in the images), which could make it harder for the DL algorithms to distinguish between classes [18-19]. Thus, for a family of DL models, the interpolation threshold for natural images might be lower than for radiology images because a higher number of parameters (a more complex model) are needed to learn the feature representations of the images. The occurrence of this shift for more complex datasets has already been observed in settings for natural image distributions that contain an increased percentage of label noise [12]. Here, label noise is the percentage of the data set labels that have been attributed to the wrong class, which just as in the case of radiology images increases the difficulty of DL models to attribute the features of an image to the correct class. In both these settings, label noise and data complexity might be considered cases of model misspecification as detailed in [12]. Model misspecification occurs when the model is unable to learn the feature space of the model. The reason this shift occurs in the double descent curves is due to classifiers with low complexity learning the feature space representation (i.e., the lower dimensional representation of images) of a certain class (that includes data with the noisy labels and correct labels), but not being able to correctly isolate the data with the noisy label from those with correct label within that same class [12&20]. However, over-parameterized models learn to differentiate between the noise and the correctly labeled data [12&20]. As such, it also might be the case that for radiology images, the less complex models will learn the salient features of a certain class which might show up in other classes as well (spots due to pneumonia vs. pleural effusion vs.

smoking damage). As such, a higher complexity model might be needed to isolate the correct labels and perform well on the test data (i.e. to differentiate the detailed differences).

## 1.6. Aims of Thesis

Consequently, the over-parameterization of models built for natural images might not guarantee that these models are also over-parameterized with respect to radiological image distributions [18-19]. In addition, the impact of transfer-learning strategies on the over-parameterization and interpolation threshold is also not yet well known. Transfer learning is a training strategy that is employed by researchers to increase the performance of their models. It involves training a model on a different task than the one intended and then using the trained weights of the base model to train a new model on a different task. This workflow is typical for radiology datasets, where in the weights of the base model trained to perform a task on natural images is then used to perform another relevant task on medical images [21-22]. Again, this workflow is typical due to small sizes of radiology datasets.

Consequently, given the above considerations and gaps in the current literature, the general aim of this thesis is to study the factors impacting whether a current pre-built DL model (built for use on natural images) that is used in the radiology literature is under-parameterized or over-parameterized with respect to its training data. Such factors include the impact of transfer learning, data set size, and disease type. To do so, we defined our task as being the binary classification (diagnosis) of a certain disease from chest

radiographs (i.e. the task is that of predicting from the presented radiographs whether the disease is present or absent). The types of diseases looked at were COVID-19 and pneumonia. These two disease were chosen due to the availability of a large number of labeled radiographs and due to the fact that they present different clinical relevancies. Diagnosing COVID-19 from chest radiographs is not considered a clinical standard while the gold standard tool for diagnosing chest radiograph is pneumonia [23-25]. In this manner, these different diseases are used as a proxy for studying the effect of the complexity of the features learned on the interpolation threshold (akin to the label noise discussed in Section 1.5). The model used in all experimental settings is Densenet121 (a convolutional neural network- CNN) due to its wide use and success in the literature for diagnosing diseases from chest radiographs [26-27].

## 1.7. Related Works

The main inspiration of this thesis was due to the work of Nakirran et al. in investigating the double descent phenomenon in deep neural networks. While much of the literature has investigated this phenomenon [12], the Nakirran et al. paper presented an experimental setup similar to ours (i.e., binary classification problem using CNNs on moderately sized datasets-CIFAR 10 and CIFAR 100) whilst also showing that the double descent occurs as a function of data set size as well training time (i.e., iteration or epochs). Furthermore, Nakirran et al. first showed that label noise increased the test error at the interpolation peak and also shifted it. This thesis builds upon their work to investigate the double descent phenomenon in terms of both model size and dataset size for the case of radiographs. In addition, we study the impact of transfer learning on double descent. It is

also pertinent to mention that the work done in this thesis is tangential to the work done by Konz et al. which showed that models tend to have a harder time learning the intrinsic features of medical images as opposed to natural images even when the intrinsic dimension of the datasets from the two domains were the same [18-19]. Finally, as mentioned in section 1.2, only one publication in medical imaging has observed the double descent curve. The double descent was observed with a dataset of only 20 MRI images whilst using a custom built CNN for image segmentation [11]. This observation of the double descent occurring with such a small sample size further motivates our research into identifying at what thresholds the over-parameterization and under-parameterization regimes occur for DL models used in medical imaging. As such, applying the heuristic that the number of data points D needs to be much larger than the number of parameters P (i.e., $D \gg P$) is certainly not strong enough of a condition to be considered when working with medical images.

## 1.8. Notes on the History of Double Descent

This section can be skipped without loss of continuity. In this section and for completeness, I aim to give a brief overview of the history of double descent in the literature. Much of the terminology used in the double descent literature stems from a 2019 paper written by Belkin et al. in which the terms “double descent” and “interpolation threshold” were first introduced [28]. In their paper, Belkin et al. attempted to reconcile the classical bias-variance tradeoff with the fact that over-parameterized DL models (models that achieve close to zero training error) still generalize well to unseen data. They attempted to do so by presenting a unified performance curve (the double descent curve) that includes both the U-shaped performance curve predicted by the classical bias-variance tradeoff (as observed

in the under-parameterized regime of neural networks) and the decreasing test error curve, as observed for over-parameterized neural networks [28]. However, the double descent curve had already been posited and observed by earlier papers in the 1990s and just a year earlier in 2018 [29-31]. Namely, a series of 2018 papers authored by Geiger et al. were the first to investigate the cause of the good generalization ability of over-parameterized DL models [30-31]. In their papers, Geiger et al. posited that over-parameterized models do not get stuck in local loss minima (minima of the cost function) but rather such models can converge during training to a global minimum of the loss [30-31]. They also posited that a decrease in the number of parameters leads to a "jamming transition" where there are many local minima for the loss and a training procedure that converges to these minima might lead to worse model performance. Here, Geiger et al. presented the analogy between this jamming transition and the physical phenomenon of jamming that is observed with repulsive ellipses [30-31]. In the physical phenomenon, a jamming transition occurs when the addition of repulsive particles in a container leads to these particles coming into contact and transitioning to a rigid glassy state [30-31]. Such glassy states have many local minima in their energy landscape. As such, in this analogy, the loss (cost) function of DL models is similar to that of the potential energy of particles and a jamming transition occurs when there are more samples than parameters (i.e., analogy: there are more particles than there is space to accommodate these particles without touching). More recent studies have attempted to explain the origin of double descent through the use of random matrix theory [32-34]. As mentioned earlier, although in this thesis we attempt to experimentally determine the conditions under which the double descent phenomenon occurs for DL models used in medical imaging, further

investigation into the theoretical considerations gleaned from our results might be of interest to both the DL and medical communities.

# 2. Introduction

## 2.1. Datasets

The COVID-19 dataset consists of 28656 chest radiographs. The chest radiographs are stored in png format and have a resolution of 256x256. The prevalence rate of COVID-19 in the dataset is 11%. The ground truth (label) for this dataset was provided by the gold standard for COVID-19 diagnosis: the reverse-transcriptase polymerase chain reaction (RT-PCR) test. This data was collected at the University of Chicago Medical Center. All images are gray-scale normalized.

The pneumonia dataset consists of 28656 chest radiographs. The chest radiographs are stored in png format and have a resolution of 256x256. The prevalence rate of pneumonia in the dataset is 11% (positive cases all contained lung opacities). The ground truth (label) for this dataset was provided by radiologists' examination of the radiographs. This data was sourced from the Radiological Society of North America (RSNA) grand challenge in 2018 [35]. Please note that the original data from the 2018 challenge contained 30000 chest radiographs and the prevalence rate was 15%. However, to make the prevalence rates between the COVID-19 and pneumonia dataset equal, 1343 pneumonia images were discarded. All images were gray scale normalized.

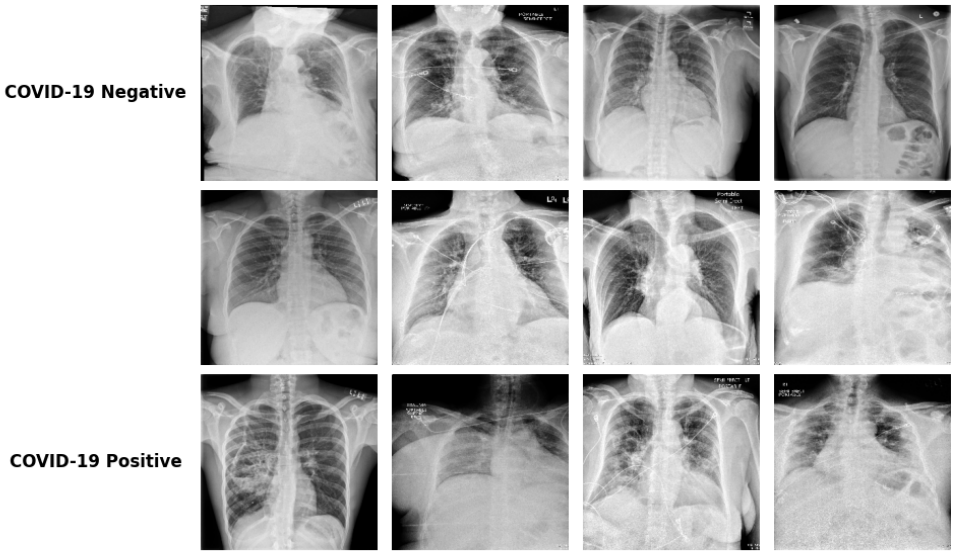


**Figure 1.** Radiographs from COVID-19 dataset. The first two rows shows chest radiographs with a negative diagnosis, while the last row shows the positive case.

As can be seen from Figure 1, it is visually difficult to distinguish between the COVID-19 positive cases and negative cases.

## 2.2. Model Archictecture

All models used in this work are based on the Densenet-121 architecture created by Huang et al. Densenets are convolutional neural networks (CNNs) that are widely used in the medical imaging field (and more broadly in computer vision tasks) due to their ability to detect complex features in images using a small number of parameters (as compared to traditional ResNets or VGGs) and less compute [26 & 36]. The Densenet models achieve a higher accuracy on common natural image datasets such as Imagenet, CIFAR-10 and CIFAR-100 due to the introduction of connections between all the convolutional layers within a dense block (a succession of 1x1 and 3x3 convolutional layers with RELU

activation, batch normalization and pooling layers) [36]. Thus, the feature map produced by any one convolutional layer of the dense block is then received by the subsequent layer in addition to the feature maps of all previous layers.

In the following, I will use Huang et al.'s notation to give a mathematical description of the information flow between layers [36]. For each non-linear transformation $H$, applied to layer $l$, the inputs are the feature maps of all $l-1$ layers. The growth rate k is defined as the number of feature maps outputted by each transformation $H$ [36]. The growth rate $k$ is the width of each layer akin to the number of channels used in CNN models with a linear architecture. $H$ is composed of three successive operations: batch Normalization, rectified linear unit and a convolutional layer [36]. In each dense block, the convolutional layer alternates between a 1x1 convolution and a 3x3 convolution. Each dense block contains a different number of transformations $H$. The Densenet-121 model is composed of 4 dense blocks with 6, 12, 24 and 16 transformations respectively. The growth rate of the base model is $k=32$ (i.e., the number of feature maps in the global average pooling layer before the classification layer would be $32\times k$).

In the following section, we aim to give an intuition on the differences in performance of DL/classifier for the two diseases.

## 3. Classifier Performance for COVID-19 and Pneumonia

To further establish the claim that COVID-19 classification necessitates a DL model to learn more complex intrinsic features than for pneumonia classification, we will construct the

ROC (operating charcteristic curve) for both diseases and compute the corresponding AUCROC (area under the operating charcteristic curve) for both diseases. In the subsequent sections, I will explain the significance of the AUCROC metric as well as all experimental details used to build the model.

## 3.1. Experimental Protocol

To do so, both datasets were split into a train/test (80%/20% split) using stratified sampling on the disease status to ensure that both train and test sets had the same prevalence of disease (i.e., 11%). We first pre-train both models on Imagenet database before freezing the base layers to train the classifier layer on each dataset and finally unfreezing all layers and fine-tuning on the corresponding dataset. Since the classes were imbalanced, each class was weighted so that there weight was inversely proportional to their frequency. The loss function used was binary cross entropy loss. Adam optimizer was used with an initial learning rate of 0.0001 and a learning schedule that decreased by 0.5 once each 5 epochs when the loss yielded no improvement. Both models were run for 250 epochs and early stopping was used as regularization (weight decay was not used for the Adam optimizer). Data augmentation was also applied before class balancing. Images were horizontally flipped and rotated with a range of ± 5 degrees. The batch size was 32 images. The small batch size is due to computational limitations. The last layer used for the binary classification utilizes a sigmoid activation function. All models in this thesis were trained using keras and tensorflow. The base model training procedure was adapted from [37].

## 3.2. AUCROC

AUCROC (the area under the operating charcteristic curve): As the name suggests, this a measure of the area under the curve that plots the true positive and false positive rate (evaluated on the test set) of a classifier.

The true positive rate TPR (also know as Sensitivity) is the number of true positives TR (samples that were correctly predicted to belong to the positive class) divided by the total number of positive cases P in the sample (alternatively, if we one were to express the TPR in terms of the number of the false negative (FN) cases, then one would simply replace P with TR + FN).

The false positive rate FPR (also know as Specificity) is the number of false positives FR (samples that were incorrectly predicted to belong to the positive class) divided by the total number of negative cases N in the sample (alternatively, if we one were to express the FPR in terms of the number of the true negative (TN) cases, then one would simply replace N with FR + TN).

In our experiment, the classifier outputs a prediction score between 0 and 1 that a patient's radiograph belongs to the positive class (presence of disease). To calculate the true positive rates (TPR) and false positive rates (FPR), a threshold needs to be chosen to separate the predictions into the two classes: positive and negative for a disease. The number of thresholds is one more than the number of unique prediction values (probabilities of classifier). For each of the thresholds, the FPR and TPR are calculated and plotted against

each other (ROC is a parametric curve). Thus, the ROC curve starts from the origin (i.e., TPR and FPR of 0) with the threshold chosen such that all samples with a prediction below the highest prediction value (i.e. all samples) would be assigned to the negative class. At this threshold the classifier predicts that all samples belong to the negative class. Conversely, the ROC curve ends at the point (1,1) with all samples being assigned to the positive class. The AUCROC is the area under the ROC curve. Thus, a perfect classifier would have an AUCROC of 1, since it would predict all positives with no false positive cases for all thresholds chosen. A random classifier would have an AUCROC of 0.50.

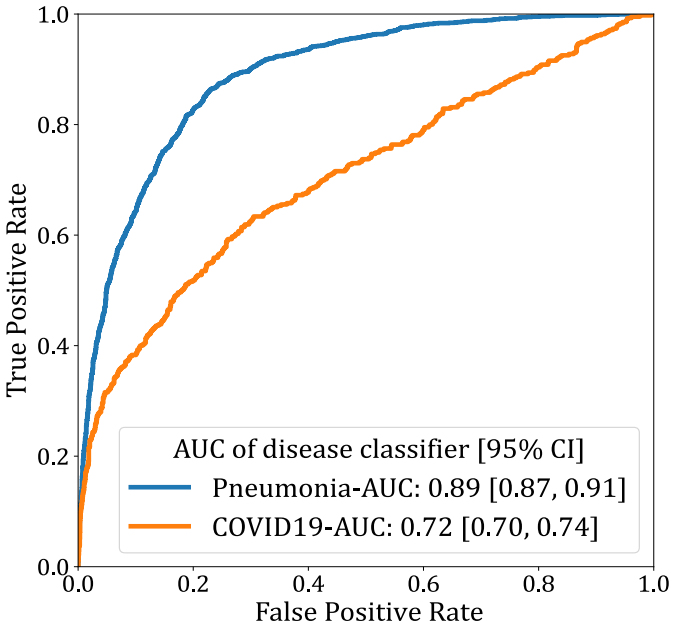


**Figure 2.** The ROC curve for both pneumonia and COVID-19. As expected, the classifier achieves a higher AUCROC on the Pneumonia dataset than on the COVID-19 dataset.

Figure 2 shows the ROC curves with the corresponding AUCROCs for both diseases. The 95% confidence intervals were calculated by bootstrapping 1000 samples. As expected the

mean AUCROC score of 0.89 for pneumonia is much higher than the AUCROC of 0.72 for COVID-19 (note the non-overlapping confidence intervals). These results indicate that the classifier is much better at learning the complex features of the training data for pneumonia than for COVID-19. This validates our intuition that datasets containing images with a high degree of similarity and overlap between classes (such as COVID-19) present a harder classification task than image datasets containing clear differences between images (such as the opacities in the pneumonia dataset). This finding motivates the subsequent sections in this thesis that compare the effect of feature complexity on the emergence of double descent phenomenon and the subsequent location of the interpolation threshold.

As such, in this next section of the thesis, we aim to uncover the impact of pretraining on Imagenet and the effect of feature complexity on the existence and location of double descent.

# 4. Pretraining and Double Descent

## 4.1. Experimental Protocol

**Datasets**: the same datasets for both pneumonia and COVID-19 as in section 3.1 were used. Datasets were split into a train/test (80%/20% split) using stratified sampling on the disease status.

**Models**: In all subsequent parts of this thesis, we will use the number of parameters in a model as a proxy for increasing model complexity. In practice, researchers in the medical imaging field in an effort to learn more complex features of the dataset would increase the

number of parameters by increasing the number of features maps. In a Densenet 121, this would amount to increasing the growth rate *k*.

However, due to the limited time and computational resources, we were not able to create Densenet models with different feature maps and train them on Imagenet. As such, we resorted to treating the base Densenet 121 model as a feature extractor whose inputs are then transmitted to fully connected layers placed after the base model (specifically after the global average pooling layer which combines all feature maps produced in the CNN). To do so, we introduced a new dense layer before the final classification layer. To change the number of parameters, we varied the number of neurons in the layer. The number of neurons added were: $4\times10, 10^4, 2\times10^4, 4\times10^4, 10^5, 4\times10^5$ and $7\times10^5$ . Each model corresponding to a specific number of neurons had the Imagenet weights loaded onto the base models layers. All models were later fine-tuned on each of the pneumonia and COVID-19 datasets separately. To compare the effect of pretraining, each model was also trained from scratch on both datasets. Furthermore, for each of these settings, the behavior of the model was assessed for both early stopping and no early stopping instances. The loss function used was again binary cross entropy loss. Adam optimizer was used with an initial learning rate of 0.0001 and a learning schedule that decreased by 0.50 once each 5 epochs when the loss yielded no improvement. All models were run for 230 epochs (or till early stopping). Data augmentation was also applied before class balancing was used. Images were horizontally flipped and rotated with a range of ± 5 degrees. The batch size was 32 images. Finally, to make sure that the behavior observed is not due to the random initialization of the layers, this process was repeated 3 times for all models with different initialization seeds.

Finally, the test loss (cross entropy loss) was plotted against the number of parameters to assess the presence or absence of double descent phenomenon. Here, the test loss was used instead of another metric such as the error or AUCROC due to the presence of an imbalanced dataset (which the 2 metrics above are affected by) and due to the intention of this work to show a general approach to the double descent phenomenon in medical imaging that is not dependent on downstream tasks or metrics. It is also pertinent to add that plotting the test loss against the number of parameters to investigate double descent is not uncommon and has been done by Nakirran et al. as well as in other studies.

## 4.2. Results and Discussion

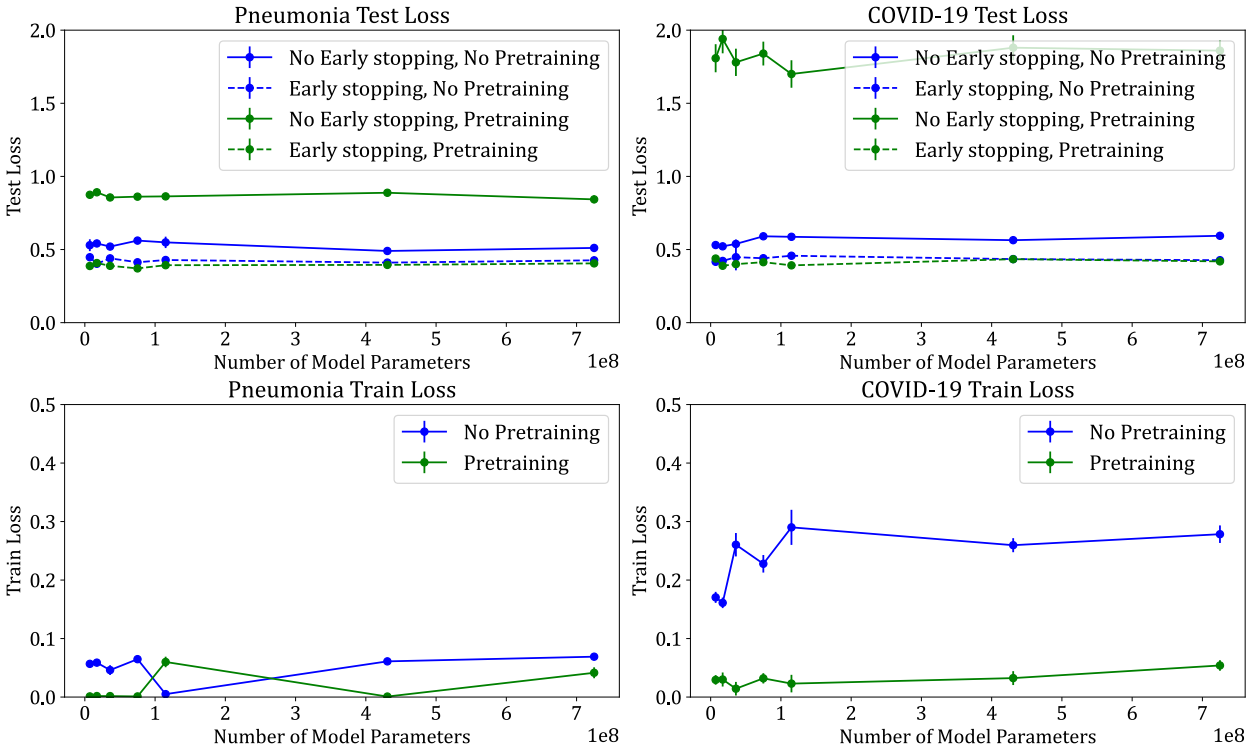


**Figure 3.** The average loss (test and train) as well as the standard deviation across three trials are plotted against the number of parameters for both COVID-19 and Pneumonia. The first rows show

the combination of conditions that include pre-training and early stopping. The last row shows the train loss for pre-training and no pretraining conditions.

As seen from Figure 3, no double descent phenomenon is observed for both diseases. All test loss curves except those pertaining to the no-pretraining case for COVID-19 achieve approximately zero training loss. This observation added to the fact that the test loss for these models fluctuates around a constant mean might suggest that these models are in the over-parameterization regime with respect to this model architecture and training procedure. The no-pretraining case for COVID-19 was not able to interpolate all training example. This might have due to the fact that the addition of hidden neurons only serves to increase the combination of features and not the number of features maps learned. Other reasons include the insufficient training time to convergence and the lack of a batch normalization layer. With the addition of pre-training the COVID-19 models are able to reach zero training loss, which might suggest that pre-training helps lower the interpolation threshold. As expected the addition of pre-training and early stopping does help to lower the test loss for all models. However, with no early stopping, the pre-training condition allows the models to overfit which leads to a higher test loss than the no-pretraining condition.

These results suggest that pre-training might reduce model complexity and the heuristic applied that the number of parameters N needs to be much larger than the number of samples D (i.e., $N \gg D$) trained on does not apply in this case of pre-training. When taking the notion of effective model complexity as in Nakirran et al. that considers the training procedure in its definition, this result is not surprising [9]. However, when considering the

notion of model complexity classically (classical statistics) such as is the case with the Rademacher complexity, which does not include the training procedure, this result might seem counter-intuitive. Other observations from Figure 3 include the expected result that early stopping leads to better generalization and a lower test loss. One observation is that the early stopping and no early stopping loss do not achieve the same loss as is expected when a model is over-parameterized. One of the reasons for this result might be due to the lower number of epochs trained in this setting. As the double descent phenomenon can also happen as a function of the number of epochs, our models might be in the critically parameterized regime with regard to training time/number of epochs. This behaviour was also observed in [9].

In the subsequent section, we will investigate how the double descent phenomenon scales with respect to dataset size and the number of parameters.

# 5. Impact of Model Width, Data Size on Double Descent

## 5.1. Experimental Protocol

**Datasets**:

For each disease, three training sets were sampled of sizes 303, 3026 and 22926. For all training set sizes considered in this section, all were tested on the same dataset of size 5730.

**Models**:

To increase the number of parameters we increase the number of feature maps. As discussed earlier, increasing the number of parameters in this manner would be more akin

to how researchers would increase the complexity of their models. Again, due to computational limitations, we were not able to train these models on Imagenet. 6 models were built with the following growth size: 1, 8, 12, 32, 48 and 58.

For each of these settings, the behavior of the model was assessed for both early stopping and no early stopping instances. The loss function used was again binary cross entropy loss. Adam optimizer was used with an initial learning rate of 0.0001 and a learning schedule that decreased by 0.50 once each 5 epochs when the loss yielded no improvement. All models were run for 200 epochs (or till early stopping). Data augmentation was also applied before class balancing was used. Images were horizontally flipped and rotated with a range of ± 5 degrees. The batch size was 32 images. Finally, to make sure that the behavior observed is not due to the random initialization of the layers, this process was repeated 3 times for all models with different initialization seeds.

## 5.2. Results and Discussion

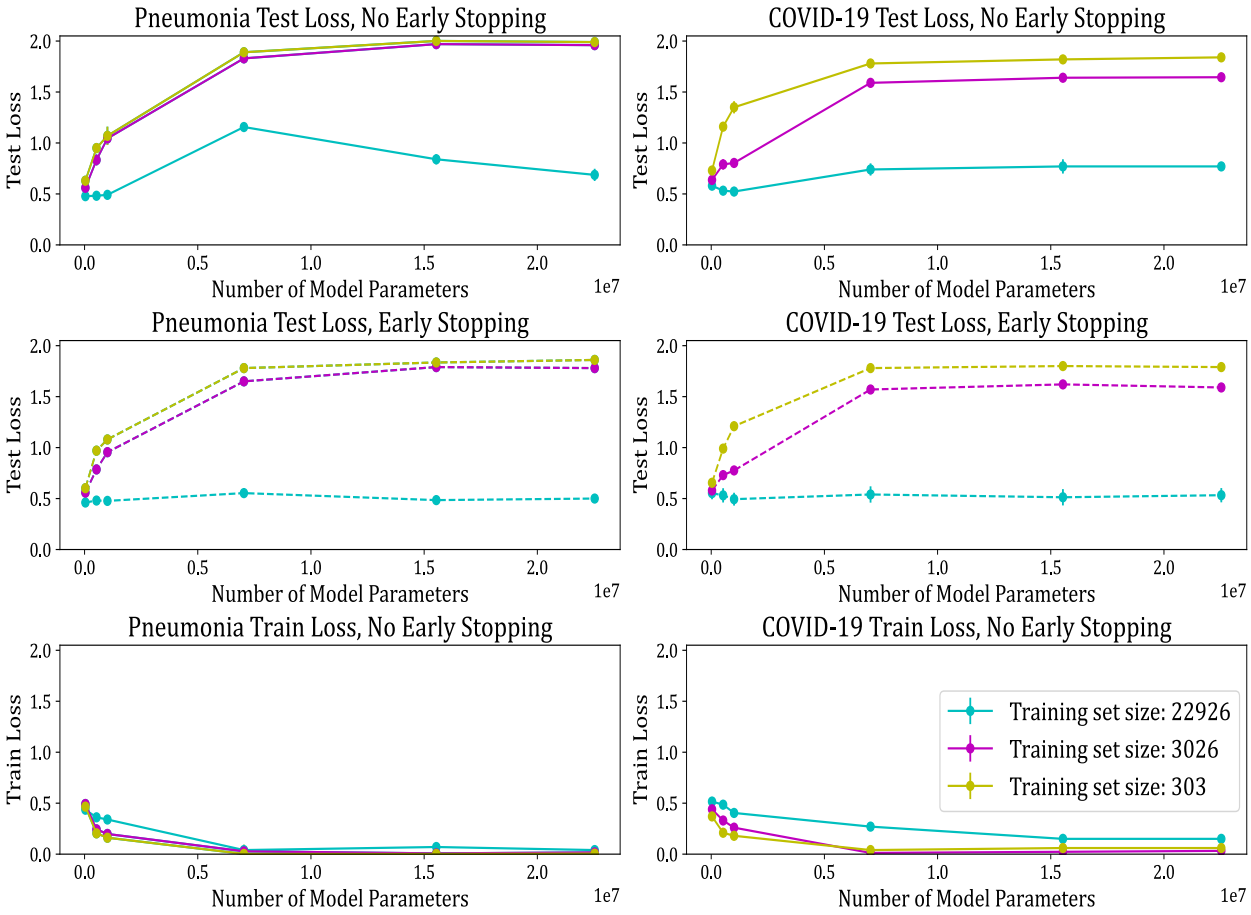


**Figure 4.** The average loss (and the standard deviation) across three trials is plotted against the number of parameters (varying growth rate) for both COVID-19 and Pneumonia. Each subplot shows either the test loss or train loss for different training set sizes: 22926, 3026, and 303. In addition, the early stopping behavior for all models tested is also reported.

As seen in figure 4, the double descent phenomenon is observed for the case of pneumonia with 22926 samples with the critically parameterized regime beginning at 1,001,964 parameters (k=12). This observation was made given that all models after 1,001,964 achieve approximately zero training loss while still not converging to the minimum measured loss of 0.45. As the training size decreased, the test loss increased and the

double descent is no longer observed (Figure 4). Instead, we observe a monotonic increase of the test loss for both training sample sizes (i.e., training set sizes of 303 and 3026). Since for models that have more than 1,001,964 parameters still achieve approximately zero training loss, it is suspected that the interpolation peak shifted to the left (towards models with lower complexity) and the corresponding test loss for that range of models was not sampled. However, since the test loss has still not converged to the minimum measured loss for both sample sizes, then we can state that models with more than 7,038,529 parameters (k=32) are in the critically parameterized regime. Consequently, increasing in the number of training size past 22926 samples might produce an increase in the test loss as opposed to a decrease as seen in [9]. This is because increasing the size of the training set shifts the interpolation threshold to models that are more complex. As such, even though the maximum test loss might decrease in amplitude, the test loss might still be higher than the measured test loss for a smaller sample size. Furthermore, from figure 4, we also observe that for the smaller sample sizes, the early stopping behavior matches the no early stopping condition, which might suggest that for these sample sizes further regularization is required (e.g., drop out layer, weight decay).

For the COVID-19 datasets, the training loss was approximately zero only for models with more than 7,038,529 parameters trained on the smallest sample sizes of 3026 and 303 images. As in the pneumonia case, double descent might not have been observed because the corresponding test loss for that range of models where the interpolation peak occurs was not sampled. As expected, with the increase of the training size, the interpolation threshold is pushed to higher model complexities and it takes more feature maps for a

model to approach zero training loss. As such, the observation of an increasing test loss and a non-zero training loss suggests that for these bigger samples, the models tested are located in the increasing test loss portion of the under-parameterized regime. However, problems in the training procedure (e.g., low number of epochs, small batch size…) might have led the models to not achieve close to zero training loss. In comparison with the case of pneumonia, this was to be expected since as datasets containing images with a high degree of similarity and overlap between classes (COVID-19) present a harder classification task than image datasets containing clear differences between images (pneumonia). Furthermore, from figure 4, as in the case for the pneumonia dataset, we also observe that for the smaller sample sizes, the early stopping behavior matches the no early stopping condition.

Finally, please note that the difference in the test loss between the 7,038,529 parameters models trained in this section and earlier might be due to the increased epoch number and the different architecture.

# 6. Conclusion and Future Work

In this thesis, we investigated how various factors such as pretraining, model width, data complexity and dataset size affect whether a prebuilt DL model (Densenet 121) that is frequently used in medical imaging is under-parameterized or over-parameterized with respect to its training distribution. We found that pre-training helped models learn more complex features of their training set which subsequently allowed them to become over-parameterized with respect to their training data and procedure. We also found that the use of images that have a high degree of similarity (COVID-19) presented a more challenging task for the classifier. The increased complexity of the data shifted the interpolation threshold to higher model complexity values and as such even for small datasets, classifiers trained with COVID-19 data were still in the under-parameterized regime. However, these same classifiers, when trained on the Pneumonia dataset, were either critically parameterized or over-parameterized with respect to their training procedure and data. This finding might have been surprising given that the number of parameters for some of the models was much bigger than the training set size. Consequently, we also found that these models in the over-parameterization regime do not improve their test loss (i.e., generalizing ability) even when their training set size is increased tenfold.

Given that the success of DL models in radiology is due to their over-parameterization, and given the above results, it becomes important for researchers in the medical imaging field to not use heuristics as the ones researchers in computer vision use to develop DL models for natural images. While the work presented in this thesis has aimed to support this claim,

further research that covers a greater scope of models (both in terms of architecture and number of model parameters), diseases, and different training procedures (e.g., impact of fine-tuning on medical image data on interpolation threshold, impact of self-supervised learning…) is needed. Furthermore, a more grounded theoretical approach that can explain the location of the interpolation threshold in the presence of factors such as pretraining, model width, data complexity and dataset size is also needed. For the latter, an interesting case for medical images would be to relate the interpolation threshold with the degree of similarity between images in a dataset. An approach to study such a case might apply methods from random matrix theory along with the use of random feature models and low-level statistics (e.g., use of entropy or cosine similarity to estimate the similarity of images within a distribution) to derive bounds within which the interpolation peak can be located.

# References


1. V. Muralidharan, B. A. Adewale, C. J. Huang, *et al.*, "A scoping review of reporting gaps in FDA-approved AI medical devices," *npj Digital Medicine*, vol. 7, p. 273, 2024.
2. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.
3. M. N. Flory, S. Napel, and E. B. Tsai, "AI in radiology: opportunities and challenges," in *Seminars in Ultrasound, CT and MRI*. WB Saunders, 2024.
4. M. G. Linguraru, S. Bakas, M. Aboian, *et al.*, "Clinical, cultural, computational, and regulatory considerations to deploy AI in radiology: perspectives of RSNA and MICCAI experts," *Radiology: Artificial Intelligence*, vol. 6, no. 4, p. e240225, 2024.
5. A. M. Rauschecker, J. D. Rudie, L. Xie, *et al.*, "Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI," *Radiology*, vol. 295, no. 3, pp. 626–637, 2020.
6. M. L. Giger, H.-P. Chan, and J. Boone, "Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM," *Medical Physics*, vol. 35, pp. 5799–5820, 2008.
7. M. L. Giger, N. Karssemeijer, and J. Schnabel, "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer," *Annual Review of Biomedical Engineering*, vol. 15, pp. 327–357, 2013.
8. B. Sahiner, A. Pezeshk, L. M. Hadjiiski, *et al.*, "Deep learning in medical imaging and radiation therapy," *Medical Physics*, vol. 46, no. 1, pp. e1–e36, 2019.

9. S. G. Armato, K. Drukker, and L. Hadjiiski, “AI in medical imaging grand challenges: translation from competition to research benefit and patient care,” *The British Journal of Radiology*, vol. 96, no. 1150, p. 20221152, 2023.
10. T. T. Tran, H. H. Pham, T. V. Nguyen, T. T. Le, H. T. Nguyen, and H. Q. Nguyen, “Learning to automatically diagnose multiple diseases in pediatric chest radiographs using deep convolutional neural networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3314–3323.
11. J. S. Baxter and P. Jannin, “Combining simple interactivity and machine learning: a separable deep learning approach to subthalamic nucleus localization and segmentation in MRI for deep brain stimulation surgical planning,” *Journal of Medical Imaging*, vol. 9, no. 4, p. 045001, 2022.
12. P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, “Deep double descent: Where bigger models and more data hurt,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.
13. X. Hu, L. Chu, J. Pei, W. Liu, and J. Bian, “Model complexity of deep learning: A survey,” *Knowledge and Information Systems*, vol. 63, pp. 2585–2619, 2021.
14. M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. Tax, “A brief prehistory of double descent,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 20, pp. 10625–10626, 2020.
15. P. Nakkiran, P. Venkat, S. Kakade, and T. Ma, “Optimal regularization can mitigate double descent,” *arXiv preprint arXiv:2003.01897*, 2020.

16. S. Tripathi, K. Gabriel, S. Dheer, *et al.*, "Understanding biases and disparities in radiology AI datasets: a review," *Journal of the American College of Radiology*, vol. 20, no. 9, pp. 836–841, 2023.
17. A. Jiménez-Sánchez, N. R. Avlona, D. Juodelyte, *et al.*, "Copycats: the many lives of a publicly available medical imaging dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 113383–404, 2024.
18. N. Konz and M. A. Mazurowski, "Pre-processing and compression: Understanding hidden representation refinement across imaging domains via intrinsic dimension," *arXiv preprint arXiv:2408.08381*, 2024.
19. N. Konz, H. Gu, H. Dong, and M. A. Mazurowski, "The intrinsic manifolds of radiological images and their role in deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2022, pp. 684–694.
20. Y. Gu, X. Zheng, and T. Aste, "Unraveling the enigma of double descent: An in-depth analysis through the lens of learned feature space," *arXiv preprint arXiv:2310.13572*, 2023.
21. H. E. Kim, A. Cosa-Linan, N. Santhanam, *et al.*, "Transfer learning for medical image classification: a literature review," *BMC Medical Imaging*, vol. 22, no. 1, p. 69, 2022.
22. P. Kora, C. P. Ooi, O. Faust, *et al.*, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 79–107, 2022.

23. I. El Naqa, H. Li, J. Fuhrman, *et al.*, "Lessons learned in transitioning to AI in the medical imaging of COVID-19," *Journal of Medical Imaging*, vol. 8, no. S1, pp. 010902–010902, 2021.
24. J. P. Kanne, H. Bai, A. Bernheim, *et al.*, "COVID-19 imaging: what we know now and what remains unknown," *Radiology*, vol. 299, no. 3, pp. E262–E279, 2021.
25. D. Wootton and C. Feldman, "The diagnosis of pneumonia requires a chest radiograph (x-ray)—yes, no or sometimes?," *Pneumonia*, vol. 5, no. 1, pp. 1–7, 2014.
26. T. Zhou, X. Ye, H. Lu, *et al.*, "Dense convolutional network and its application in medical image analysis," *BioMed Research International*, vol. 2022, p. 2384830, 2022.
27. H. Jia, J. Zhang, K. Ma, *et al.*, "Application of convolutional neural networks in medical images: a bibliometric analysis," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 5, pp. 3501, 2024.
28. M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15849–15854, 2019.
29. M. Opper, W. Kinzel, J. Kleinz, R. Nehl, On the ability of the optimal perceptron to generalise. *J. Phys. A Math. Gen.* Vol.23, pp.581–L586, 1990.
30. S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, "A jamming transition from under-to over-parametrization affects generalization in deep learning," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 47, p. 474001, 2019.

31. M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart, "The jamming transition as a paradigm to understand the loss landscape of deep neural networks," *arXiv preprint arXiv:1809.09349*, 2018.
32. Z. Liao, R. Couillet, and M. W. Mahoney, "A random matrix analysis of random Fourier features: Beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13939–13950, 2020.
33. F. Bach, "High-dimensional analysis of double descent for linear regression with random projections," *SIAM Journal on Mathematics of Data Science*, vol. 6, no. 1, pp. 26–50, 2024.
34. X. Meng, J. Yao, and Y. Cao, "Multiple descent in the multiple random feature model," *Journal of Machine Learning Research*, vol. 25, no. 44, pp. 1–49, 2024.
35. A. Stein, C. Wu, C. Carr, *et al.*, "RSNA pneumonia detection challenge," Mountain View: *Kaggle*, 2018.
36. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
37. Q. Hu, K. Drukker, and M. L. Giger, "Role of standard and soft tissue chest radiography images in deep-learning-based early diagnosis of COVID-19," *Journal of Medical Imaging*, vol. 8, no. S1, p. 014503, 2021.