Transfer Learning of Histology Slides Improved CNN Performance on Lung Cancer by Pretraining on Colon Cancer

Wesley Hu

A fully automated digital pathology workflow will save 49.4% of time spent on a histology case study and training of machine learning models is necessary for automation. A major limiting factor for the training of machine learning models is the lack of large datasets because datasets in the medical field are private and confidential. To solve this problem I utilized transfer learning to share knowledge between two similar oncology datasets of histology images for colon cancer and lung cancer. Transfer learning reduces the risk of overtraining by exposing the model to different datasets. In my study, I took a look at how transfer learning effects model performance between lung cancer and colon cancer in different dataset sizes. Remarkably, transfer learning on 3,750 lung cancer images outperformed a scratch model trained on twice the dataset. The lowest validation loss the scratch model achieved was about 0.35 while transfer learning achieved a validation loss of about 0.125 which is around a 280% improvement in validation loss. Transfer learning on extremely small dataset sizes (1,000 images for colon cancer and 1,500 images for lung cancer) showed no performance improvements and even performance degradation. All models trained on the extremely small datasets overtrained regardless whether the model was pretrained or not.

Keywords: Neural Networks, Machine Learning, Transfer Learning, Colon Cancer, Lung Cancer

1. INTRODUCTION

Lung and colon cancers are the most invasive types of cancers and the two most common sources of death by cancer in the United States [1]. Lung cancer is the leading cause of cancer deaths globally with only 14 18% of patients surviving after diagnosis in 5 years [1]. The American Cancer Society estimates 235,760 new cases and 131,880 deaths from lung cancer in 2021 [2]. There are two types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [3] with NSCLC accounting for roughly 85% of all lung cancer diagnoses [4]. NSCLC diagnoses are subcategorized as adenocarcinoma (ACA), squamous cell carcinoma (SCC), or large cell carcinoma. ACA and SCC are the two most predominant forms of NSCLC lung cancers with ACA making up 30% of all lung cancers and NSCLC containing 40% ACA, 30% SCC, and 10% large cell carcinoma [5] [3]. ACA lung cancer is moderately linked to smoking but is the most common type of lung cancer in people who have never smoked. ACA is found commonly in a younger patient population and is more common in women. ACA often results in weakness and secretion of mucus in the lungs [6]. Lung SCC starts in flat cells that line the lungs called squamous cells; SCC is more strongly linked to smoking than any other type of NSCLC. SCC has a very aggressive phenotype and can spread across the body to the spine or even the brain if left untreated.

My research studies use machine learning to help identify lung SCC, lung ACA, and colon ACA from tissue histology images. Convolutional neural networks (CNN) are widely accepted as the best method for image processing in machine learning. Unlike other machine learning models, CNN can study image data as a 2 dimensional tensor instead of 1 dimensional vectors. By viewing images as a 2 dimensional tensor, CNN models are able to detect patterns and make connections that other models can't by preserving the pixel's locality. CNN works by scanning an image throughout its pixels in 2 dimensional batches known as a kernel and moves across the image (Fig. 1).

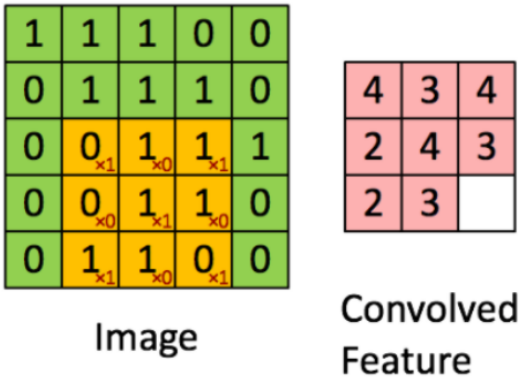

FIGURE 1. An image showing how CNN models scan through images.
The kernel size is the size of the orange square and the model iterates through. A single neuron is connected to a kernel and a neuron is a single square in the convolved feature. (Fig. Source: [7]).

Transfer learning is a method of machine learning where a neural network is pretrained on a specific dataset and is then trained on another dataset. As a neural network trains on a dataset, the parameters get changed over time to fit the training data better. By implementing a model that has been first trained on another dataset, the preferred outcome is for the model to have better parameter initialization compared to a model training from scratch with initialization from the pretrained ResNet18 model on the ImageNet dataset [8]. With the development of ImageNet, transfer learning has become popular with the rise of many models such as AlexNet and GoogLeNet pretrained on ImageNet. Transfer learning has been cited as being consistently beneficial in a study looking at lung nodules [9] [10]. My studies compared transfer learning models in different sizes of datasets between lung and colon histology images as they are similar oncology outcomes.

Histopathology is a key technique in studying and identifying key features of certain cells. Histopathology is most notably used to identify and distinguish between different types of cancers and benign tumors and is the current standard in cancer diagnosis. Histology analysis is conducted by taking a thin sample slice of tissue which is then stained to make patterns more visible when under examination. On average 36% of the time spent on a histology case study is used to review slides and 13.4% of the time spent can be automated [11]. The stained slide is then analyzed under either an optical or electron microscope. Figure 2 shows an example of a histology image.

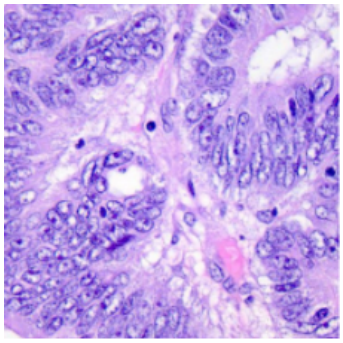

FIGURE 2. Histology image of tissue with colon adenocarcinoma cancer.
Figure 2 shows how histology images are able to help bring out details in tissue samples. (Fig. Source: [1])

2. METHODS

2.1. Histology Dataset

The dataset used in this study are histology images taken from the Veterans Health Administration (VHA), the largest US healthcare system where annually almost 50,000 cases of cancer are diagnosed. 50 cases of lung squamous cell carcinoma (SCC), 50 cases of lung adenocarcinoma (ACA), and 50 cases of colon ACA samples were taken from VHA's molecular database. Histological images of the cells were extracted from the samples by using Leica Microscope MC190 HD Camera connected to an Olympus BX41 microscope. The histological images were taken at a resolution of 1024 x 768 using a 60x dry objective lens which was then cropped to a resolution of 768 x 768 pixels. A total of 750 unique lung images were recorded (250 SCC, 250 ACA, and 250 benign) and a total of 500 unique colon cancer images were recorded (250 ACA and 250 benign) [12]. These images were then augmented by using a combination of left and right 25 degree rotations, horizontal flips, and vertical flips (Fig. 3). The original 1,250 image dataset was expanded to 25,000 images consisting of 5,000 lung ACA images, 5,000 lung SCC images, 5,000 lung benign images, 5,000 colon ACA images, 5,000 colon benign images [12].

In my study, I created two independent CNN models, one for colon cancer and another for lung cancer. All models were trained for 10 epochs with no time limit. The dataset was split into 80% for training and 20% for validation, all random seeds were set to 0 to make the split of data into training and validation, and generation of training batches repeatable. Due to hardware limitations, the resolution of the entire dataset's resolution was halved, from 768 x 768 to 384 x 384.

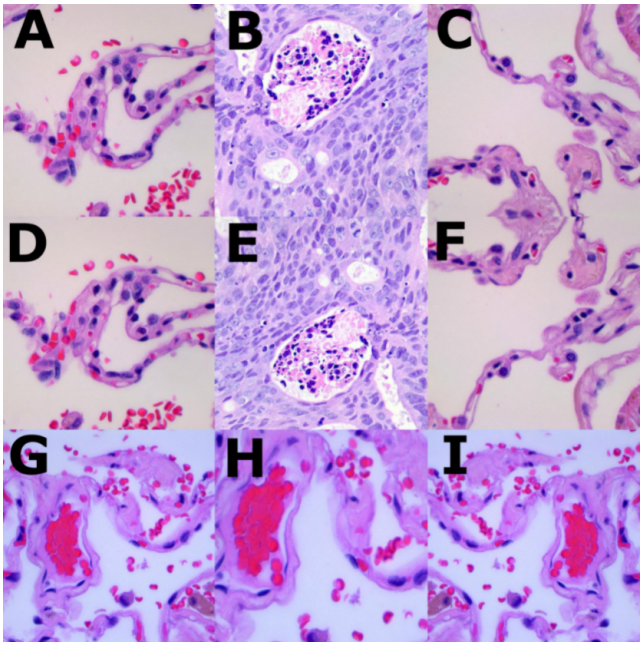

FIGURE 3. Histology images taken from both the lung and colon cancer dataset.
A rotational augmentation can be visualized with benign lung histology images A and D. Colon adenocarcinoma histology images B and E shows an image that has been augmented with both a vertical and horizontal flip. Benign lung histology images C and F show a vertical flip augmentation. Image H was augmented to zoom into the benign lung histology image G and image I was augmented with a horizontal flip from image G. (Fig. Source: [1]).

2.2. Study Design and Model Training

Study Design and Model Training To utilize transfer neural networks I trained each program independently and saved the model parameters where the moving average validation loss was at its lowest across 10 iterations [13]. The drawing of batches is stochastic and moving averages reduce the stochastic effect by taking the last 10 records of losses and averaging them out. Since my CNN programs validate the model every 10 iterations, the moving average for validation loss will keep a record of the last 10 losses and the moving average for training loss will keep a record of the last 100 losses. The saved model was then imported for further training in the other outcome to compare the performance of the model pretrained on histology images to a generic model pretrained on the ImageNet dataset.

To emphasize the influence transfer learning has on model performance four different studies were conducted (Table 1).

	Lung Dataset	Colon Dataset
Study One	15,000 images	10,000 images
Study Two	7,500 images	5,000 images
Study Three	3,750 images	2,500 images
Study Four	1,500 images	1,000 images

TABLE 1: **Table showing how transfer learning was conducted across the four studies.**
Table 1 shows how many datasets of each outcome were used for training transfer learning models in all four studies.

The first study viewed the performance of a model pretrained on histology images against a generic model pretrained on the ImageNet dataset with both models utilizing the entire dataset. ImageNet is a large database of over 14 million images composed of over 1,000 classes including cars, animals, insects, fish, planes, and rockets [8]. The transfer learning model was pretrained on 10,000 colon cancer images for lung cancer and pretrained on 15,000 lung cancer images for colon cancer. The second study also compared a histology pretrained model against a generic model but instead, the dataset was artificially halved for both colon and lung cancer models. The transfer learning model was pretrained on 5,000 colon cancer images for lung cancer and pretrained on 7,500 lung cancer images for colon cancer. The third study had the dataset artificially quartered for both colon and lung cancer models where the transfer learning model was pretrained on 3,750 lung cancer images for colon cancer and 2,500 colon cancer images for lung cancer. Finally, the fourth study tests the model when the dataset is exceptionally small with the dataset only containing 500 images per class (1,000 images for colon cancer and 1,500 images for lung cancer). Model performance would also be compared between studies to understand how a model using transfer learning would compare to a generic model despite having less training data.

All of my studies utilized a pretrained ResNet18 network with the same hyperparameters. ResNet18 is a model designed in 2015 to tackle the ImageNet classification task and has been trained successfully to predict a number of outcomes including, cars, boats, planes, and rockets [8]. Resnet models utilized residual models. Before residual modules, the more layers a model contains, the harder it is to train and gradient values will either explode or disappear [14]. Several different versions of Resnet were created to tackle this problem by using residual models. Residual models reduced the effect of this phenomenon by adding jump connections. The pretrained ResNet18 model was highly regarded as “an efficient tumor detector” and when compared to other pretrained networks, ResNet18 had short training times, great classification performance, and significantly lower parameters which reduce the chance of overfitting in a similar histology study [15]. The final ResNet18 model’s performance was not only high but also the most consistent [15].

Due to the successes ResNet18 had with identifying and classifying gastrointestinal cancer, I decided to also use the pretrained ResNet18 model with the default hyperparameters in my studies. The Adam optimizer with a learning rate of 0.001 was also utilized across all studies. Adam was first published in 2014 looking to solve the problems of the Stochastic Gradient Descent (SGD) optimizer [16]. Adam can be regarded as a combination of Root Mean Squared Propagation (RMSprop) and SGD. It utilizes squared gradients to scale the learning rate like RMSprop and uses momentum like SGD but with a moving average instead [17]. In an experiment in 2020, different optimizers' performances were compared against each other, Adam and different extensions of Adam always came at the top of performance metrics and always learned the quickest. The original Adam optimizer from time to time would place at the top but even when Adam's performance was not at the top, Adam was consistently close to the top. Adam was utilized mainly since it required little tuning and is computationally efficient compared to other optimizers [18]. The learning rate was set to 0.001, β_1 was set to 0.9, and β_2 was set to 0.999 for Adam.

2.3. Related Works

My study closely follows two studies from 2021: Lu. et al., 2021 [19] and Benhamida. et al., 2021[20]. Lu. et al., compares the outcomes and performance results of popular pretrained CNN models including ResNet-18, GoogLeNet, AlexNet, and VGG-19 on lung tissue. The experiment utilized 15,419 images of CT scans of lung tissue from the RIDER dataset. The results shows that, apart from their proposed CNN model, the pretrained Resnet-18 model yielded the highest metrics with an accuracy of 89.3 and had the lowest error of 6.6 [19].

A study conducted by Benhamida. et al., 2021 used histopathological colon cancer images to train and compare different popular CNN models including AlexNet, VGG-19, ResNet-18, DenseNet, Inception, Ensemble DNN, Texture Analysis, and SqueezNet. The dataset used were 100,000 histology images of colon tissue. The research found that Resnet-18 yielded the highest accuracy score across all CNN models with an accuracy of 96.98%. This result was also surprising considering that the Resnet model had low computational cost relative to other models.

A separate study conducted in 2020 took a look at a pretrained model's ability to analyze histology images using the same dataset in my study. Pretrained VGG16, ResNet50, InceptionV3, Inception-ResNetV2, MobileNet, Xception, NASNetMobile, and DenseNet169 models were trained on the lung and colon cancer dataset separately. The results of the study indicates that all of the pretrained models performed exceptionally well with no models showing significant differences in recall and precision metrics. For lung cancer all models had a perfect F1 scores in exception to NASNetMobile (0.97), ResNet50 (0.96), and VGG16 (0.98). All models had perfect performance metrics in exception to NASNetMobile with an F1 score of 0.98 [34].

2.4. Performance Metrics

Model performance was measured with outcome-specific precision, recall, and F1 score metrics while model accuracy was calculated across all outcomes. All of the performance metrics use true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in their equation (Fig. 4).

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

FIGURE 4. A grid combining prediction and actual results to give TP, FP, FN, and TN.
The TP, FP, FN, and TN numbers calculated from figure 4 were used extensively in developing different performance metrics.
(Fig. Source: [21])

Accuracy is an elementary metric and is simply the percentage in decimal form of correct predictions by dividing the total correct predictions by the entire dataset. Accuracy is a useful metric as it is very simple to understand and read, as seen in equation 1.

$$\frac{TP + FN}{TP + TN + FP + FN} \tag{1}$$

Precision provides a ratio describing the number of positives correctly identified compared to what the model identified as positives and measures the FP rate (equation 2).

$$\frac{TP}{TP + FP} \tag{2}$$

Recall provides a ratio showing how many positives were correctly labeled as positives compared to the total amount of positives and measures the FN rate (equation 3).

$$\frac{TP}{TP + FN} \tag{3}$$

Precision and recall are important as low precision scores mean the model has a high bias in labeling images as positives. A low recall score shows that the model has a high bias towards labeling images as negatives. Thus it is important to keep track of the precision and recall of a model despite having a low loss or high accuracy score. The F1 score takes account of both FP and FN

rates and thus is a balance of recall and precision. The equation to calculate an F1 score is shown in equation 4.

$$\frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

Equation 4 also simplifies to equation 5:

$$2 * \frac{recall \cdot precision}{recall + precision} \tag{5}$$

F1 score is a great performance metric to gauge overall model performance as it is comparable to accuracy but offers more insight on FP and FN rates. Metrics were visualized in real-time during training across both time and iterations with a smoothing of 0.995 [22].

3. RESULTS

3.1. Transfer learning has a significant impact on model performance on 7,500 and 3,750 lung cancer images.

3.1.1. Transfer learning showed continuous rapid improvement when training on 3,750 lung cancer images

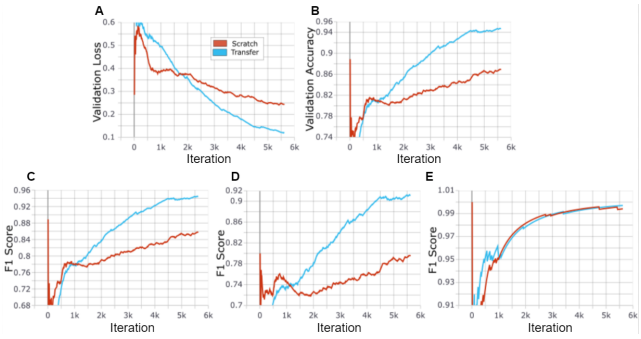

FIGURE 5. Validation loss, F1, and accuracy curves for 3,750 images of the lung cancer dataset.
A) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (blue) or trained from scratch (red). B) Validation accuracy curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (blue) or trained from scratch (red). C) Validation F1 score curves for the training of ResNet18 on the ACA outcome either pretrained on lung cancer (blue) or trained from scratch (red). D) Validation F1 score curves for the training of ResNet18 on the benign cell outcome either pretrained on lung cancer (blue) or trained from scratch (red). E) Validation F1 score curves for the training of ResNet18 on the SCC outcome either pretrained on lung cancer (blue) or trained from scratch (red). A, B, C, D and E shows that transfer learning improved model performance significantly in almost all performance metrics and reduced loss by 50%.

When the dataset was quartered from 15,000 to 3,750 images, the transfer learning model, pretrained on colon cancer, started off with a much higher loss than the scratch model but the transfer learning model improved significantly faster than the scratch model, eventually surpassing the scratch model (Fig. 5, Table 2).

	Iteration	Loss	ACA F1	SCC F1	Benign F1	Accuracy
Scratch	5,240	0.24	0.85	0.79	1.00	0.87
Pre-trained	5,586	0.12	0.95	0.91	1.00	0.95

TABLE 2: **Validation performance metrics and iteration taken from the lowest moving average validation loss for both pretrained and scratch models on 3,750 images from the lung cancer dataset.**
All values are rounded to 2 decimal places. Table 2 shows that transfer learning was able to significantly outperform the scratch model and perfecting all performance metrics.

Figure 6 also shows overtraining in both the transfer learning and scratch model, however since the transfer learning's validation metrics were rapidly improving, by the end of 10 epochs, the transfer learning model was no longer overtrained.

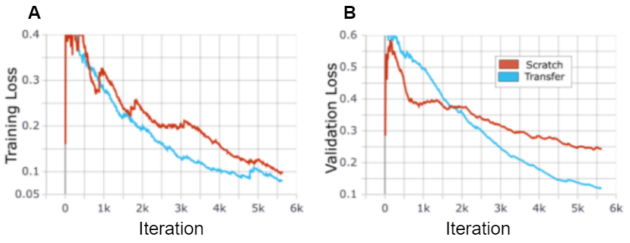

FIGURE 6. Validation and training loss curves for 3,750 images of the lung cancer outcome.
A) Training loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (blue) or trained from scratch (red). B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer images (blue) or trained from scratch (red). A and B shows that transfer learning was able to surpass the scratch model in validation performance and was able to fix the overtraining issue on its own, unlike the scratch model.

The scratch model overtrained from the beginning to end and became more severe as validation loss was decreasing at a slower rate compared to its training loss (Fig. 6)

3.1.2. Transfer learning on 3,750 lung cancer images was able to significantly outperform a scratch model trained on 7,500 lung cancer images

When comparing the transfer learning model trained on 3,750 images against a scratch model trained on 7,500 images, the transfer learning model significantly outperforms the scratch model despite the scratch

model utilizing two times the number of images than the transfer learning model (Fig. 7).

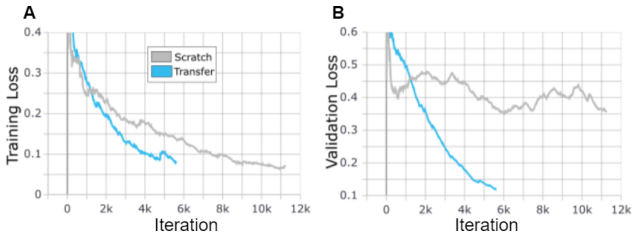

FIGURE 7. Validation and training loss curves for the comparison of a transfer learning model trained on 3,750 lung cancer images against a scratch model trained on 7,500 lung cancer images.
A) Training loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on 2,500 colon cancer images (blue) or trained from scratch with 7,500 lung cancer images (gray). B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on 2,500 colon cancer images (blue) or trained from scratch with 7,500 lung cancer images (gray). A and B shows that transfer learning was so effective it was able to outperform a scratch model training on twice the dataset size at a significant scale.

Transfer learning on 3,750 lung cancer images ended with a validation loss of 0.125 compared to the scratch model's validation loss of 0.35 trained on 7,500 images which is an improvement of 64%. Transfer learning showed constant improvements and was not significantly overtrained unlike the scratch model (Fig. 7).

3.1.3. Transfer learning showed sudden improvements halfway through training on 7,500 lung cancer images.

Training on small lung cancer datasets resulted in transfer learning significantly improving model performance. When the dataset was artificially halved from 15,000 to 7,500 images, the transfer learning model pretrained on colon cancer had a lower loss than the scratch model thanks to a better initialization of the parameters (Fig. 8).

Figure 9 shows significant overtraining in both the transfer and scratch models. The transfer model had a sudden drop in the validation loss midway through training and significantly helped to combat the overtraining issue. The scratch model never dropped as significantly compared to the transfer model. The validation accuracy and F1 scores help support the claim of sudden performance improvement at 5,500 iterations of training and eventually improving model performance with transfer learning (Fig. 9, Table 3).

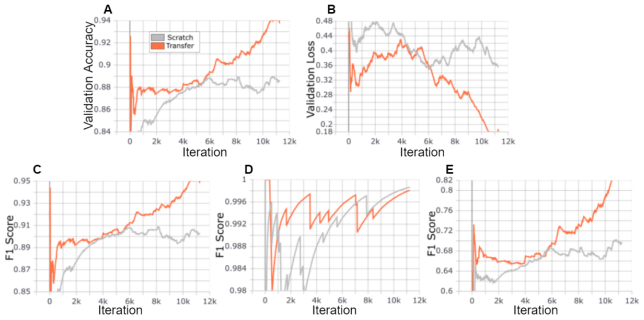

FIGURE 8. Validation loss, F1, and accuracy curves for 7,500 lung cancer images.
A) Validation accuracy curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (orange) or trained from scratch (gray). B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (orange) or trained from scratch (gray). C) Validation F1 Score curves for the training of ResNet18 on the ACA outcome either pretrained on colon cancer (orange) or trained from scratch (gray). D) Validation F1 Score curves for the training of ResNet18 on the benign outcome either pretrained on colon cancer (orange) or trained from scratch (gray). E) Validation F1 Score curves for the training of ResNet18 on the SCC outcome either pretrained on colon cancer (orange) or trained from scratch (gray). A, B, C, D, and E show that transfer learning improved model performance significantly in almost all performance metrics.

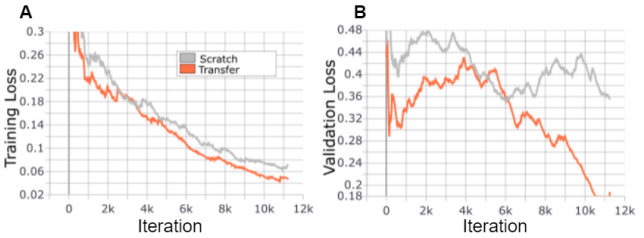

FIGURE 9. Training and validation loss curves for 7,500 lung cancer images.
A) Training loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (orange) or trained from scratch (gray). B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (orange) or trained from scratch (gray). A and B show that the validation loss was not keeping up with the training loss and identifies significant overtraining but transfer learning dropped significantly midway through training.

	Iteration	Loss	ACA F1	SCC F1	Benign F1	Accuracy
Scratch	6,189	0.35	0.68	0.9	0.99	0.89
Pre-trained	10,996	0.16	0.84	0.96	1.0	0.95

TABLE 3: **Validation performance metrics and iteration taken from the lowest moving average validation loss for both pretrained and scratch models on 7,500 images from the lung cancer dataset.**
All values are rounded to 2 decimal places. Table 3 shows that transfer learning was able to outperform the scratch model.

3.1.4. Transfer learning on 7,500 lung cancer images was unable to compete with a scratch model trained on 15,000 lung cancer images within 10 epochs of training

Despite having significant improvements in performance, the transfer learning model trained on 7,500 lung cancer images was unable to compete with the scratch model trained on 15,000 lung cancer images within 10 epochs. If both experiments ran for more epochs, the transfer learning model would have probably converged and caught up to the scratch model trained on 15,000 lung cancer images; however, 10 epochs were the limit of my experiment (Fig. 10).

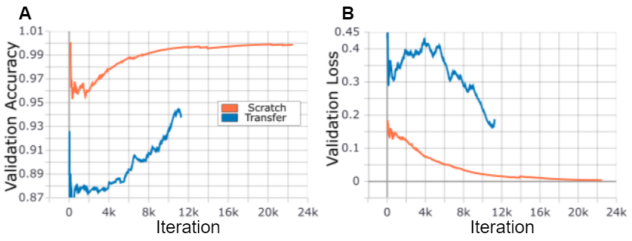

FIGURE 10. Validation accuracy and loss curves for the comparison of transfer learning on 7,500 images against a scratch model on 15,000 images on the lung cancer outcome.
A) Validation accuracy curves for the training of ResNet18 on 7,500 images of the lung cancer outcome either pretrained on 5,000 images of colon cancer (blue) or trained from scratch with 15,000 images of lung cancer (orange). B) Validation loss curves for the training of ResNet18 on 7,500 images from the lung cancer outcome either pretrained on 5,000 colon cancer images (blue) or trained from scratch with 15,000 lung cancer images (orange). A and B shows that transfer learning applied on 7,500 lung cancer images was very effective but it was unable to reach the performance metrics the scratch model trained on 15,000 lung cancer images did.

3.2. Transfer learning improves model performance on 10,000 colon cancer images

When comparing a model trained on scratch for 10,000 colon cancer images to a model pretrained on 15,000 lung cancer images, the pretrained model resulted in an instant improvement from transfer learning (Fig. 11).

The pretrained model not only started out with a better loss, but the model improved much more rapidly over time compared to the scratch model with eventual convergence. This is due to transfer learning having better initial parameters from training on the lung cancer dataset compared to preset parameters generated from the ResNet18 model. The initial improvement is more pronouncedly seen in the validation F1 scores and validation accuracy graphs (Fig. 12).

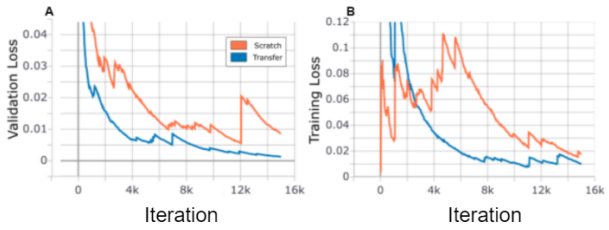

FIGURE 11. Training and validation loss curves for 10,000 colon cancer images.
A) Training loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on colon cancer (blue) or trained from scratch (orange). B) Validation loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (blue) or trained from scratch (orange). The results of A and B indicate that transfer learning provided an immediate improvement in performance but the models converged over time.

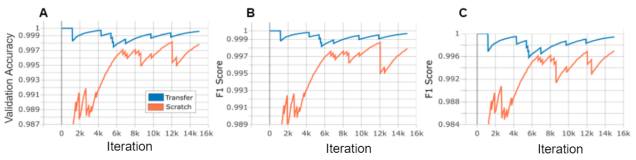

FIGURE 12. Validation accuracy and F1 curves for 10,000 colon cancer images.
A) Validation accuracy curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (blue) or trained from scratch (orange). B) Validation F1 score curves for the training of ResNet18 of benign colon cells for both pretrained on lung cancer (blue) and trained from scratch (orange).C) Validation F1 score curves of ACA colon cancer cells for the training of ResNet18 of both pretrained on lung cancer (blue) and trained from scratch (orange). The results of A, B, and C indicate that transfer learning provided an immediate improvement in performance and the models converged over time.

As shown in figure 12, transfer learning on 10,000 colon cancer images is considered to be effective as the transfer learning model had good initialization of the parameters, and table 4 shows that the scratch model was never able to catch up to the transfer learning model despite convergence near the end of

training. Figure 11 shows that transfer learning reached a validation loss of 0.005 at 7,000 iterations while it took the scratch model 12,000 iterations to reach a validation loss of 0.005. Convergence with the scratch model is considered normal and expected.

	Iteration	Loss	ACA F1	Benign F1	Accuracy
Scratch	12,044	5.55e-03	1.00	1.00	1.00
Pre-trained	15,000	1.26e-07	1.00	1.00	1.00

TABLE 4: **Validation performance metrics and iteration taken from the lowest moving average validation loss for both pretrained and scratch models on 10,000 colon cancer images.**
Table 4 shows that the transfer learning model was able to outperform the scratch model with eventual convergence between the scratch and transfer model.

3.3. Transfer learning does not improve model performance on extremely small datasets and overtrains

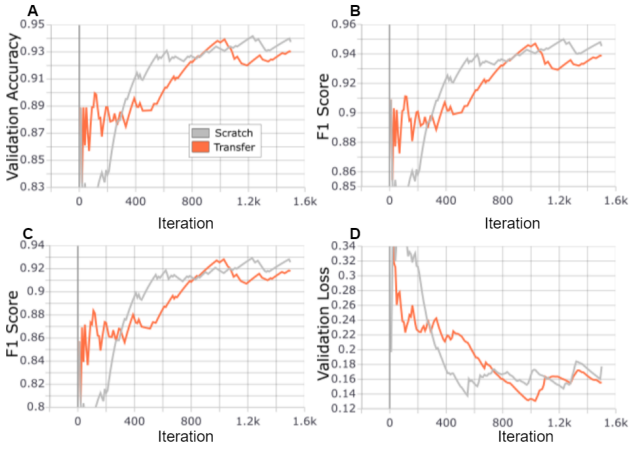

FIGURE 13. Validation loss, F1, and accuracy curves for 1,000 colon cancer images.
A) Validation accuracy curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (orange) or trained from scratch (gray). B) Validation F1 score curves for the training of ResNet18 on the ACA cell outcome either pretrained on lung cancer (orange) or trained from scratch (gray). C) Validation loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (orange) or trained from scratch (gray). D) Validation F1 score curves for the training of ResNet18 on the benign cell outcome either pretrained on lung cancer (orange) or trained from scratch (gray). A, B, C, and D show that the transfer learning model did not result in any major improvement compared to the scratch model.

When the dataset was brought to the extreme (500 images per class) with only 1,000 images for colon cancer and 1,500 images for lung cancer, transfer learning was ineffective. The transfer learning model for 1,000 colon cancer images pretrained on 1,500 lung cancer images closely followed the scratch model in all performance metrics with no effective initialization of parameters on the transfer learning model. (Fig. 13).

Though the transfer learning model trained on 1,000 images was able to compete with the scratch model (Fig.13), overtraining has become evident when trained on 1,000 images (Fig. 14).

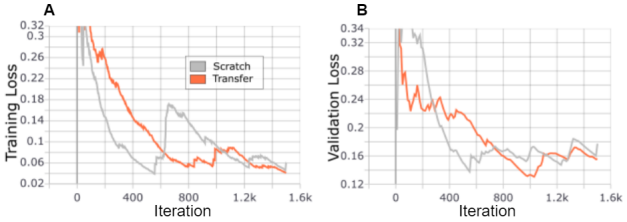

FIGURE 14. Training and validation loss curves for 1,000 images from the colon cancer dataset.
A) Training loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (orange) or trained from scratch (gray). B) Validation loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (orange) or trained from scratch (gray). A and B show that the validation loss was not keeping up with the training loss which signifies overtraining.

Overtraining in the extremely small dataset was consistent across lung cancer as well, with 1,500 lung cancer images. The fourth study with lung cancer shows extreme overtraining, with transfer learning at its best having a validation loss of 0.32 at 800 iterations while having a training loss of 0.25 at 800 iterations; the scratch model at its best at 1,470 iterations having a validation loss of 0.364 while having a training loss of 0.19 at 1,470 iterations (Fig. 15).

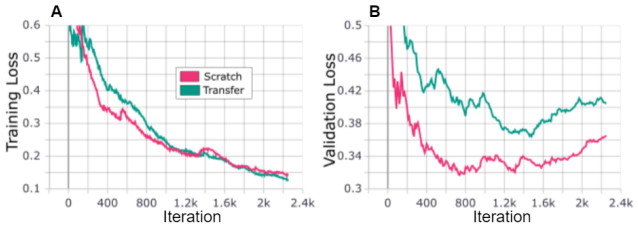

FIGURE 15. Training and validation loss curves for 1,500 images from the lung cancer dataset.
A) Training loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (green) or trained from scratch (pink). B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (green) or trained from scratch (pink). A and B show that the validation loss was not keeping up with the training loss and identifies significant overtraining.

The transfer learning model trained on lung cancer had actually significantly diminished the model's

performance. The transfer learning model had worse parameter initialization than the scratch model and did not improve as fast as the scratch model either. Across the whole training period of 10 epochs, the model never converged with transfer learning constantly performing worse (Fig. 16).

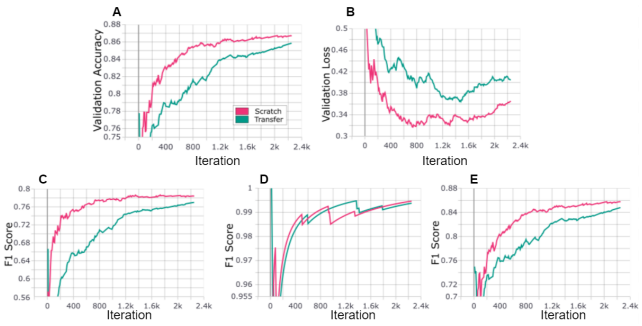

FIGURE 16. Validation loss, F1, and accuracy curves for 1,500 lung cancer images.
A) Validation accuracy curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (green) or trained from scratch (pink). B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (green) or trained from scratch (pink). C) Validation F1 score curves for the training of ResNet18 on the SCC outcome either pretrained on lung cancer (green) or trained from scratch (pink). D) Validation F1 score curves for the training of ResNet18 on the benign cell outcome either pretrained on lung cancer (green) or trained from scratch (pink). E) Validation F1 score curves for the training of ResNet18 on the ACA outcome either pretrained on lung cancer (green) or trained from scratch (pink). A, B, C, D and E show that the transfer learning model did not improve model performance but instead had lower performance metrics.

3.4. Transfer learning improves model performance on 5,000 colon cancer images

	Iteration	Loss	ACA F1	Benign F1	Accuracy
Scratch	7,500	9.32e-04	0.99	1.00	1.00
Pre-trained	4,070	4.76e-06	1.00	1.00	1.00

TABLE 5: **Validation performance metrics and iteration taken from the lowest moving average validation loss for both pretrained and scratch models on 5,000 colon cancer images.**
Values are rounded to 2 decimal places. Table 5 shows how transfer learning allowed the model to perform better than the scratch model at its best while still reaching that optimum 2,950 iterations earlier.

When the dataset was halved from 10,000 images to 5,000 images, a model pretrained on 7,500 images of the lung cancer dataset did not show the same instantaneous improvement that was visible when the model was trained on the entire 10,000 images, but rather a more mild instantaneous improvement as the transfer learning model had a consistent 94% improvement in validation loss through the first 1,000 iterations but improvements slowing down as a 50% improvement were seen in validation loss from 1,000 to 2,000 training iterations (Fig. 17). Though the model utilizing transfer learning finished with higher performance metrics (Table 5), Figure 17 shows that the model pretrained on lung cancer took leads in performance metrics but started converging over time with the scratch model in F1, accuracy, and loss.

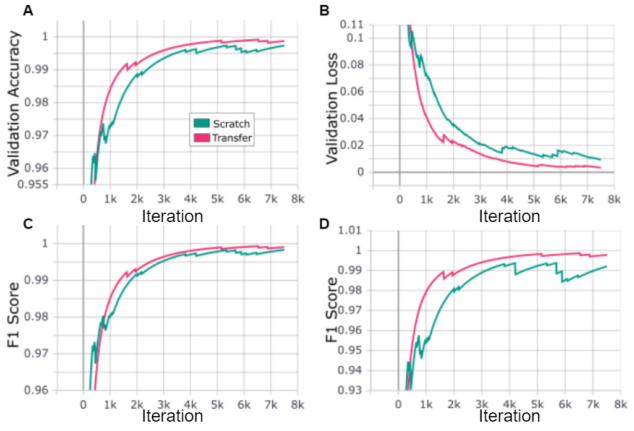

FIGURE 17. Validation loss, F1, and accuracy curves for 5,000 colon cancer images.
A) Validation accuracy curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (pink) or trained from scratch (green). B) Validation loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (pink) or trained from scratch (green).C) Validation F1 score curves for the benign outcome for the training of ResNet18 either pretrained on lung cancer (pink) or trained from scratch (green). D) Validation F1 score curves for adenocarcinoma on the training of ResNet18 either pretrained on lung cancer (pink) or trained from scratch (green).

3.5. Transfer learning does not improve model performance on 2,500 colon cancer images

Training on smaller datasets was consistently ineffective for colon cancer. When the dataset was quartered from 10,000 images to 2,500 images for the third study, transfer learning on 2,500 colon cancer images failed to improve model performance. When trained on 5,000 images, transfer learning on colon cancer was able to match up with the scratch model and insignificantly surpassing the scratch model in performance metrics. However, transfer learning trained on 2,500 images was unable to keep up with the scratch model and was falling behind in validation accuracy, loss, and F1 scores. The Transfer learning model was constantly behind the scratch model by 0.04 in validation loss

on average. The transfer learning model caught up to the scratch model in validation accuracy and F1 scores at 750 to 1,000 iterations but then transfer learning's validation accuracy and F1 deviated, falling behind the scratch model (Fig. 18).

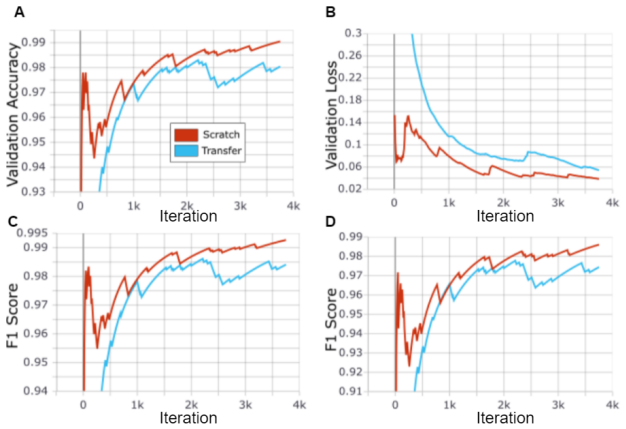

FIGURE 18. Validation loss, F1, and accuracy curves for 2,500 colon cancer images.
A) Validation accuracy curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (blue) or trained from scratch (red). B) Validation loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (blue) or trained from scratch (red).C) Validation F1 score curves for the ACA outcome for the training of ResNet18 either pretrained on lung cancer (blue) or trained from scratch (red). D) Validation F1 score curves for the benign outcome on the training of ResNet18 either pretrained on lung cancer (blue) or trained from scratch (red).

Through comparing the training loss with validation loss, overtraining is revealed with the transfer learning model (Fig. 19).

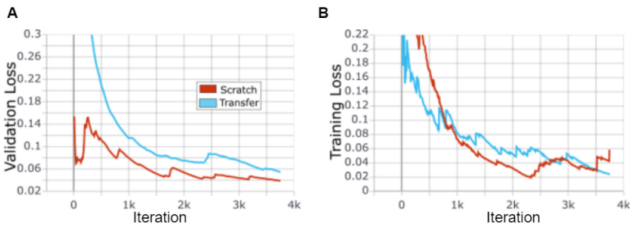

FIGURE 19. Training and validation loss curves for 2,500 colon cancer images.
A) Validation loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (blue) or trained from scratch (red). B) Training loss curves for the training of ResNet18 on the colon cancer outcome either pretrained on lung cancer (blue) or trained from scratch (red). A and B show that the validation loss was not keeping up with the training loss which signifies overtraining.

Table 6 shows the final results from the moving average and presents how at its best, the transfer learning model pretrained on lung cancer was unable to significantly surpass the scratch model.

	Iteration	Loss	ACA F1	Benign F1	Accuracy
Scratch	2,220	3.87e-02	0.99	0.99	0.99
Pre-trained	3,750	5.44e-02	0.98	0.97	0.98

TABLE 6: **Validation performance metrics and iteration taken from the lowest moving average validation loss for both pretrained and scratch models on 2,500 images from the colon cancer dataset.**
All values are rounded to 2 decimal places. Table 6 shows that transfer learning was unable to make a significant impact on model performance compared to the scratch model.

3.6. Transfer learning does not improve model performance on 15,000 lung cancer images

When the whole 15,000 lung cancer image dataset was utilized, the transfer learning model pretrained on the entire colon cancer dataset showed instant improvements in training loss and the transfer learning model outperformed the scratch model and eventually converged. However, this pattern was not shown in validation loss. The transfer learning model started out with a higher loss and never truly outperformed the scratch model (Fig. 20).

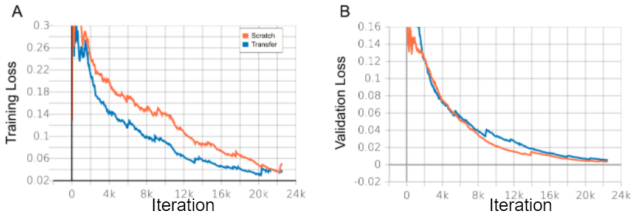

FIGURE 20. Training and validation loss curves for 15,000 lung cancer images.
A) Training loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (blue) or trained from scratch (orange). The results indicate that transfer learning provided an immediate improvement in performance but the models converge over time. B) Validation loss curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (blue) or trained from scratch (orange)

Figure 21 with validation F1 and accuracy metrics shows the same patterns shown in validation loss in figure 20. Transfer learning did not have better parameter initialization than the scratch model but the transfer learning model was able to constantly keep up with the scratch model without ever significantly surpassing the scratch model in any validation performance metrics.

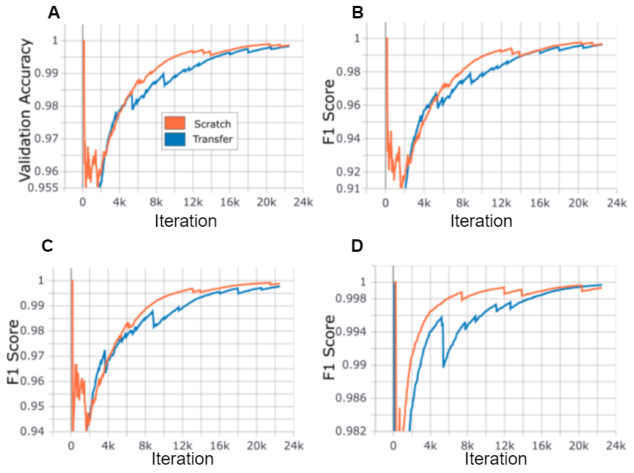

FIGURE 21. Validation F1, and accuracy curves for 15,000 lung cancer images.
A) Validation accuracy curves for the training of ResNet18 on the lung cancer outcome either pretrained on colon cancer (blue) or trained from scratch (orange). B) Validation F1 score curves for the training of ResNet18 on the ACA outcome either pretrained on colon cancer (blue) or trained from scratch (orange). C) Validation F1 score curves for the training of ResNet18 on the SCC outcome either pretrained on lung cancer (blue) or trained from scratch (orange). D) Validation F1 score curves for the training of ResNet18 on the benign cell outcome either pretrained on lung cancer (blue) or trained from scratch (red). A, B, C, and D show that transfer learning did not result in major performance uplifts compared to the scratch model.

4. DISCUSSION

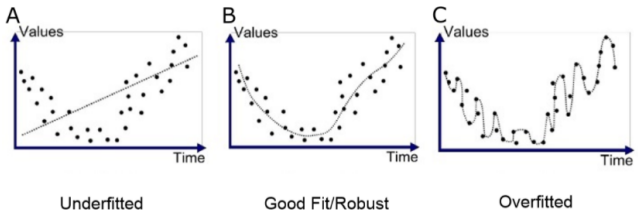

FIGURE 22. Three hypothetical model training scenarios on the same dataset.
A shows an underfitted model where the parameters are not updated to their fullest extent. B shows a well-trained model where the parameters are a good fit for the training dataset. C shows overtraining, where the model parameters are trained for too long. Even though the model performance looks near perfect, it's only perfect for the specific training dataset and the model will have trouble with any other data points outside of the training dataset.

Transfer learning on lung cancer for 7,500 and 3,750 images improved model performance immensely as transfer learning on 7,500 resulted in a 54% decrease in loss and 50% decrease in loss on 3,750 images. These increases happened due to the scratch models being overtrained significantly. Overtraining happens when a model's parameters have been updated too well and become localized to the training dataset. Essentially when a model's training metrics are better than the validation or testing metrics. Figure 22 visualizes overtraining and how a model leads to overtraining over time through three different scenarios.

The scratch model overtraining can be explained through the dataset size. Since the datasets were significantly smaller, the datasets were also less diverse. Models that are trained on datasets that are not diverse are prone to overtraining as localizing parameters become significantly easier. As for why the scratch models overtrained significantly when utilizing 7,500 and 3,750 images for lung cancer but not on 5,000 and 2,500 colon cancer images can be explained through the number of classes. Lung cancer has 3 classes, ACA, SCC, and benign while colon cancer only has 2 classes, ACA and benign. The more classes a model needs to identify the number of datasets needed exponentially grows as it becomes significantly harder to identify more and more classes. Though the datasets for lung cancer were increased, the amount increased was only linear compared to the exponential increase the models needed. The transfer learning models were able to counter overtraining as they not only had better initial parameters but halfway through training the transfer learning models had sudden performance improvements that were not shown in the scratch model. The transfer learning model trained on 7,500 images seemed to have been caught at a local minima making it seem at 5,000 iterations that the transfer learning and scratch model were identical, but that was only a local minima and the momentum of the Adam optimizer was able to get out of the local minima and continue. The transfer learning model for the third study on lung cancer was so effective it was able to surpass the scratch model trained on 7,500 images by 280%. Both transfer learning models never showed signs of diminishing returns in the 10 training epochs.

The first study in colon cancer when the dataset was at 10,000 images, showed abnormal results. Transfer learning started out with perfect accuracy and F1 scores which is extremely surprising but also worrisome. Though it's true that transfer learning is supposed to have better parameter initialization, it does not make sense for the transfer learning model to start out with perfect metrics when introduced to another dataset. This could be caused by the stochastic nature of choosing batches and the random seeds used to split the dataset.

The fourth study took a look at an extreme scenario where the dataset was extremely small with only 500 images per class meaning 1,000 colon cancer images and 1,500 lung cancer images. For lung cancer, transfer learning did not result in performance increases, and all models overtrained significantly. Transfer learning was consistently performing worse in validation metrics. The transfer learning model

was never able to compete with the scratch model and showed no signs of convergence in validation loss. A theory for this activity can be explained through transfer learning itself and overtraining. Since the dataset is already extremely small at 1,500 images, models being trained are susceptible to overtraining which is why all models trained in the fourth study overtrained. The transfer learning model has more updated parameters which also means it is closer to overtraining than the scratch model. Transfer learning in lung cancer thus overtrained much earlier than the scratch model. A study in 2019 utilizing neural networks pretrained on the ImageNet dataset to identify medical images found the same results. When the dataset is significantly limited, transfer learning does not improve the model's performance metrics and sometimes perform significantly worse compared to the scratch model. This study also noted that all models significantly and rapidly overtrained with disregard to whether the model was pretrained or not [23].

The next step for my research is to do further experimentation on the four studies: train for more epochs and use different combinations of transfer learning. It will be interesting to see what would happen if I took transfer learning from an entire dataset and applied it to smaller datasets. For example, I would take the trained model on 10,000 colon cancer images and utilize transfer learning on 7,500, 3,750, and 1,500 lung cancer images. Doing so will allow us to analyze two things: transfer learning from a large dataset to a small dataset and remove the possibility of using an overtrained model for transfer learning as many times I used an overtrained model for transfer learning.

Another issue faced during the four studies was the random splitting of data. Random splits were used two times for most studies: the first split was used to artificially cut the datasets for the second, third, and fourth study; the second split was used to split the data for training and validation. Many random seeds perform better or worse than others and can affect a model's stability [24]. Examples of this is with the study of colon cancer on 10,000 images where the transfer learning model was performing unreasonably well and with the study of lung cancer with 15,000 images when instant improvement was seen in training loss but not validation loss. To resolve this issue, instead of using only one seed to generate the splits, I should repeat experiments with different set seeds, such as 0 - 9, and averaging all their results for final analysis. A proposed method of further reducing the standard deviation of different random seeds is the usage of Norm-filtered Aggressive Stochastic Weight Averaging (NASWA) which was shown to reduce standard deviations of random seeds by 72% on average [24].

Another shortcoming of my research is not having a test set, specifically a test set from a separate source. Histology images can vary drastically between sources due to different sampling and histology methods, including the size of samples, dye used, and preparation methods such as paraffin or semithin methods. The next step would be contacting a local hospital or histology lab to obtain histology images.

REFERENCES

[1] Borkowski, A. A., Bui, M. M., Brannon Thomas, L., Wilson, C. P., DeLand, L. A., and Mastorides, S. M. Lung and colon cancer histopathological image dataset (LC25000). , **?**

[2] Torre, L. A., Siegel, R. L., and Jemal, A. Lung cancer statistics. *Adv. Exp. Med. Biol.*, **893**, 1–19.

[3] Wang, R., Li, Y., Hu, E., Kong, F., Wang, J., Liu, J., Shao, Q., Hao, Y., He, D., and Xiao, X. S100A7 promotes lung adenocarcinoma to squamous carcinoma transdifferentiation, and its expression is differentially regulated by the Hippo-YAP pathway in lung cancer cells. *Oncotarget*, **8**, 24804–24814.

[4] Wang, B.-Y., Huang, J.-Y., Chen, H.-C., Lin, C.-H., Lin, S.-H., Hung, W.-H., and Cheng, Y.-F. The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients. *J. Cancer Res. Clin. Oncol.*, **146**, 43–52.

[5] Minna, J. D., Roth, J. A., and Gazdar, A. F. Focus on lung cancer. *Cancer Cell*, **1**, 49–52.

[6] Tanoue, L. T. Women and lung cancer. *Clin. Chest Med.*, **42**, 467–482.

[7] Nouri, D. Using convolutional neural nets to detect facial keypoints tutorial. https://danielnouri.org/notes/2014/12/17/using-convolutional-neural-nets-to-detect-facial-keypoints-tutorial/. Accessed: 2021-8-29.

[8] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June, pp. 248–255.

[9] Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., and Tekinerdogan, B. Analysis of transfer learning for deep neural network based plant classification models. *Comput. Electron. Agric.*, **158**, 20–29.

[10] Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. Deep convolutional neural networks for Computer-Aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging*, **35**, 1285–1298.

[11] Ho, J., Ahlers, S. M., Stratman, C., Aridor, O., Pantanowitz, L., Fine, J. L., Kuzmishin, J. A., Montalto, M. C., and Parwani, A. V. Can digital pathology result in cost savings? a financial projection for digital pathology implementation at a large integrated health care organization. *J. Pathol. Inform.*, **5**, 33.

[12] Borkowski, A. A., Wilson, C. P., Borkowski, S. A., Thomas, L. B., Deland, L. A., Grewe, S. J., and Mastorides, S. M. Comparing artificial intelligence platforms for histopathologic cancer diagnosis. *Fed. Pract.*, **36**, 456–463.

[13] Hu, W. Cancer-transfer-learning.

[14] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. , **?**

[15] Kather, J. N., et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.*, **25**, 1054–1056.

[16] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. , **?**

[17] Bushaev, V. Adam — latest trends in deep learning optimization. https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c. Accessed: 2021-8-7.

[18] Soydaner, D. A comparison of optimization algorithms for deep learning. , **?**

[19] Lu, X., Nanehkaran, Y. A., and Karimi Fard, M. A method for optimal detection of lung cancer based on deep learning optimized by marine predators algorithm. *Computational Intelligence and Neuroscience*, **2021**, 3694723.

[20] Ben Hamida, A., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G., and Wemmert, C. Deep learning for colon cancer histopathological images analysis. *Computers in Biology and Medicine*, **136**, 104730.

[21] Dilmegani, W. b. C. and Dilmegani, C. Machine learning accuracy: True vs. false positive/negative. https://research.aimultiple.com/machine-learning-accuracy/. Accessed: 2021-8-12.

[22] Abadi, M., et al. TensorFlow: Large-Scale machine learning on heterogeneous distributed systems. , **?**

[23] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging.

[24] Madhyastha, P. and Jain, R. On model stability as a function of random seed.

[25] What is lung cancer? https://www.cancer.org/cancer/lung-cancer/about/what-is.html. Accessed: 2021-8-8.

[26] Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., and Jemal, A. Colorectal cancer statistics, 2020. *CA Cancer J. Clin.*, **70**, 145–164.

[27] GIPHY. Blog daniel GIF - find & share on GIPHY.

[28] Types of lung cancer: Common, rare and more varieties. https://www.cancercenter.com/cancer-types/lung-cancer/types. Accessed: 2021-8-8.

[29] Bhande, A. What is underfitting and overfitting in machine learning and how to deal with it. https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76. Accessed: 2021-8-12.

[30] Types of lung cancer: Common, rare and more varieties. https://www.cancercenter.com/cancer-types/lung-cancer/types. Accessed: 2021-8-8.

[31] Bloice, M. D., Stocker, C., and Holzinger, A. Augmentor: An image augmentation library for machine learning. , **?**

[32] Evolution history of ResNet network - programmer sought. https://www.programmersought.com/article/37851671778/. Accessed: 2021-8-6.

[33] Types of lung cancer: Common, rare and more varieties. https://www.cancercenter.com/cancer-types/lung-cancer/types. Accessed: 2021-8-7.

[34] Garg, S. and Garg, S. Prediction of lung and colon cancer through analysis of histopathological images by utilizing pre-trained cnn models with visualization of class activation and saliency maps. *2020 3rd Artificial Intelligence and Cloud Computing Conference*, New York, NY, USA AICCC 2020 38–45. Association for Computing Machinery.