**Self-Learning AI on Limited Labeled Image Data**

**Niraj Gupta1\*, Saniya Khalil2\*, Jolie Li3\*, Iris Ochoa4\*, Elisa Torres5\*, David Jose Florez Rodriguez6,7, Angela María Zapata Marin7**

1 Little Angels College, Kathmandu, Nepal

2 Yonkers High School, Yonkers, New York, USA

3 Ed W. Clark High School, Las Vegas, Nevada, USA

4 Mill Creek High School, Hoschton, Georgia, USA ­

5 Colegio Champagnat, Villa Alemana, Chile

6 SHTEM, Stanford University, California, USA

7 Papaya Institute, Phoenix, Arizona, USA

*\*These authors contributed equally to this work*

**Student Authors**

Niraj Gupta, Saniya Khalil, Jolie Li, Iris Ochoa, and Elisa Torres. All are in high school.

**SUMMARY**

This research explores self-supervised learning (self-learning) AI using Google Colab by pre-training a general TensorFlow-coded model on recognizing patterns in unlabeled image data. This is done following the example of Simple Siamese representation learning in the literature.  After this self-learning, the same model trains on labeled data. Separate general sequential machine-learning models are also trained on the same image data without prior self-learning. The authors hypothesize that self-learning will decrease how much we depend on extensively labeled data for the development of accurate AI models. Self-learning is not too helpful in many cases, though there is evidence in the results that self-learning may help certain models with extremely limited labeled data.

**INTRODUCTION**

The development of reliable visual artificial intelligence (AI) models usually requires a substantial quantity of high-quality, accurately labeled imaging data by professionals. Unfortunately, obtaining such data in the real world is typically limited or inaccessibly expensive. This poses one of the most significant challenges faced in the intersection of machine learning applications in the biomedical field. This hinders the development of potentially high-performing AI models that could help develop effective strategies for disease prevention, early detection, management, etc. In order to combat this, AI models can be trained using easily accessible, task-specific, unlabeled data at first. This allows the model to grasp basic visual patterns that may arise in the data in its unsupervised self-learning phase. These visual patterns and the model’s understanding of them can later improve the task to later train on. Then, labeled data can be utilized in order to train the model, as is usually done, and adjust the parameters that composed it in its self-learning phase.

This research focuses on testing this self-learning strategy on multiple datasets to gauge its general utility. The work herein developed models following the Simple Siamese representation learning strategy for self-supervised learning (He) and analyzed their results to answer the questions "How does self-learning affect the accuracy of AI models trained on limited, labeled data compared to regular supervised AI models?” and “How does self-learning’s impact vary across different datasets and thus use cases?”

**RESULTS**

While not all the datasets have coherent results in support or against the benefit of self-learning, they are best presented separately. Where accuracy and loss patterns tell the same story in terms of self-learning’s effect, only one is mentioned for brevity. The best and worst accuracy impacts of self-learning are also included to give the reader a tangible idea of how two models compare. All accuracies and accuracy improvements are presented as decimals between 0.0 and 1.0. Loss and accuracy, if unspecified, refer to training loss and training accuracy, respectively.

**Apple scab**

After analyzing the data, it is prevalent that self-supervised learning (hereafter ‘self-learning’) provided little to no improvement in the Apple scab dataset.

Validation loss for supervised learning and self-learning decreased as the labeled dataset sizes (hereafter ‘dataset sizes’) increased. The former went from 2.2 to 1.4, and the latter went from 2.2 to 1.8. It is evident that validation loss did not improve through the usage of self-supervised learning in most cases. Validation loss only decreased with self-learning for a labeled dataset size of 48 images, and this was a small improvement.

Furthermore, Figure 1 shows that training loss hardly decreased for self-learning with increasing dataset size. Values dropped quickly from over 2 and stayed around 1.8 after. For supervised learning, as dataset sizes increased, the training loss decreased from 2.2 to 1.4 as in the validation loss. The data suggests that self-learning worsened training loss for large dataset sizes and improved it for dataset sizes under 80 images.

In both scenarios, self-learning only improves the validation loss or training loss for some small dataset sizes. The loss and dataset sizes had a negative correlation in all cases. When impacting accuracy, self-learning dominates by up to +0.12 percentage points and hurts accuracy by as much as -0.07. In validation accuracy, displayed in Figure 2, the best and worst impacts are +0.075 and -0.11, respectively.

**Breast cancer**

This dataset was perhaps the worst performance improvement of self-learning over supervised learning. Nearly every model on every dataset size ran shows a better performance for supervised learning. In training accuracy, both models start at 1.0 and the supervised learning simply drops slower to .86 at the largest dataset tested, outperforming the self-learning model, which drops to .84. Both curves are extremely noisy. The losses tell a similar story, with the only change that losses increase as dataset sizes increase. The validation accuracy of the supervised model increases from .79 to .87, while the self-learning starts at .79 and largely doesn’t change. The validation loss is similar relation, though in the supervised learning, the validation loss fluctuates wildly between 1.8 and 2.05, and in the self-learning, the validation loss starts at slightly over 2.0 and drops to 1.9 in a more orderly fashion that the supervised learning curve. One singular value for self-learning validation loss outperforms the supervised learning loss at dataset size 160. It is a minute improvement of less than .05 in the loss function, however. The loss curves in Figure 3 follow similar patterns to the validation curves across dataset sizes, improving with more data, and are worse for self-learning than for supervised learning.

Self-learning’s impacts on accuracy were null or as low as -0.12. In validation accuracy, the improvement from self-learning ranged from -0.0 to -0.06.

**CIFAR10 10**

According to the validation accuracy data the self-learning model did not do better than the supervised learning model. Although the model did have very high accuracy, demonstrated in Figure 4 to be near 1.0 with small datasets of less than 50 images, once it was run with larger datasets (such as 600) the accuracy drastically decreased to under 0.5. However, the validation loss results tell a different story, showing that the model begins to learn as more images are given to be trained on. The supervised validation loss improves slightly, from nearly 3.9 to 3.6, in the self-learning model as the dataset size increases from less than 50 to over 200. Although the self-learning model improves on this with a constant loss of 1.5, it had no improvement in performance with increasing dataset size. This means that the self-learning model consistently did worse in accuracy than the supervised model but outperformed the supervised model in loss. Similar results are present in the training loss and accuracy data.

CIFAR10’s impact on training accuracy varies from -0.5 to -0.9. Surprisingly this pairs with an improvement in loss nearly half of the supervised training loss value. CIFAR10’s validation accuracy change after self-learning ranges from -0.02 to -0.14.

**COVID**

From the data for the COVID-19 dataset, the training loss and validation loss of self-learning did not often outperform supervised learning. For training loss, supervised learning always outperformed self-learning. Still, validation loss was better in self-learning than in supervised learning at some small dataset sizes with less than 72 images. The former hovered around 2.7, and the later began at nearly 2.8 and dropped non-linearly to 2.3 across the dataset sizes tested.

For training accuracy, shared in Figure 5, self-supervised learning outperforms supervised learning for larger datasets with over 150 images. The supervised model’s accuracy fell from .85 to near .60 with increasing dataset size, and the self-learning model fluctuated from .75 to .62.

For validation accuracy, results for both models were largely disorganized. The self-learning results were more stable staying at .56 for most dataset sizes, especially smaller than 116. For some dataset sizes, self-learning improves results, for others, it doesn’t.

COVID-19 accuracy improves by up to +0.05 in big data and falls by as much as -0.2. Validation accuracy improved by no more than +0.02 and fell by no more than -0.06.

**Glaucoma**

Self-learning shows almost no improvement over supervised learning with the Glaucoma dataset. The training accuracy in supervised learning fluctuates around .75, going as high as .87 and as low as .5; that of self-learning starts low at .66 and rises to just under .75 with an early spike to this high value. Supervised learning does best for every dataset size available for comparison. In the training loss curves of Figure 6, both supervised learning and self-learning improve with increasing dataset size. Supervised learning constantly dominates. However, self-learning shows improvement in validation loss around a dataset size of 16 and in validation accuracy for a dataset size under 50 images, the latter of which can be seen in Figure 7.

Glaucoma accuracy, with self-learning, falls between -0.02 and -0.12; however, the validation accuracy increases by as much as +0.10 and falls by up to -0.05.

**DISCUSSION**

Different datasets showed vastly different results. There is no implacable evidence for or against self-learning. There are, however, some overarching patterns. For one, self-learning does not contribute a consistent improvement over regular supervised learning for all datasets and dataset sizes across both accuracy and loss metrics. Some datasets see self-learning improving loss metrics, like CIFAR10, and some models see improved accuracy metrics, like apple scab. Moreover, the different sizes of unlabeled dataset sizes used in self-learning show no correlation with the impact of self-learning.

To begin, it’s important to note that only some models demonstrate increased learning with increased dataset size. The apple scab experiments are the only ones that show a clearly increasing performance on training and validation data with increasing dataset size. Meanwhile, the breast cancer experiments show a decreasing performance on training data and increasing performance on validation data. This suggests that overfitting was present in the breast cancer models. A similar pattern that may be the effect of overfitting is present in the CIFAR10 dataset’s and the COVID-19 dataset’s training curves. COVID-19’s validation loss curves do show improved performance with more data, however. And while glaucoma losses decrease with increased dataset size, the accuracies remain stagnant. This makes it difficult to say where a self-learning method, if any, could have improved a model’s learning if any learning was happening at all. When underperforming compared to supervised learning, it is not always clear if this is due to worse learning in general, or due to corrected overfitting, which can return lower training performance.

Outside the supervised learning and self-learning’s separate performances, their difference is the best available metric for self-learning’s effect. Leaving the questions of “What is proper learning?” and “What is overfitting?” aside, self-learning performance is only better than supervised learning performance in some select cases. Though self-learning does show some improved performance in the smaller dataset sizes in COVID-19's validation loss, larger dataset sizes in COVID-19's training accuracy and near everywhere in CIFAR10’s training and validation loss—a great contrast to the hugely negative self-learning impact on both training and validation CIFAR10 accuracies. Self-learning also improves performance in apple scab’s accuracy metrics in small datasets, especially in training data, and in COVID-19 for bigger datasets, never in glaucoma’s training, and for small values in glaucoma’s validation.

One must note that COVID-19 and CIFAR10 were multicategory datasets, unlike their counterparts, which call for binary labeling of images. It’s possible that the architecture, hyperparameters, and loss function used work best on binary decision problems. The mismatch between CIFAR10 losses and CIFAR10 accuracies, particularly in their change with self-learning, may be impacted by a loss function not best tailored to the multicategorical accuracy goal. Focusing on the studies with binary classification tasks, then, breast cancer’s supervised training curves show clear signs of overfitting as performance falls with increasing data. Self-learning is nearly never helpful. In the glaucoma experiments, where the supervised model’s training loss suggests learning, self-learning doesn’t help training performance but does improve validation performance in both loss and accuracy for small dataset sizes. Finally, in the apple scab models—where the clearest learning happens across training and validation on both metrics for the supervised model—the self-learning improves training accuracy and loss for small dataset sizes. Self-learning also improves validation loss at small dataset sizes and validation accuracy at seemingly random, scattered dataset sizes.

This could mean that self-learning can help visual AI models that have already shown improved performance when given increasing dataset sizes. It may be that self-learning is most helpful in some architectures and with certain loss functions. The research results also suggest that varying the unlabeled dataset size within the hundreds is not impactful. It may be the case that much more unlabeled data and computational power are needed to fully realize self-learning’s potential. In future work, greater computational power would allow the creation of more intricate and comprehensive models, which would allow for the production of more accurate results. Furthermore, an increased amount of data would allow the AI models created to have more reliable results and decrease the chances of overfitting. The quality of the datasets in question was not verified and could be a contributing factor to the impact of self-learning. Given some utility of self-learning, the tool could empower teams working in the medical field or other areas in need of computer vision to train powerful models even with limited labeled datasets for the desired application.

**MATERIALS AND METHODS**

**Datasets**

A sample of the datasetsin Figure 8 best shows what the models’ jobs were. All images were resized to 250 by 250 pixels due to the limited computation power.

**Breast cancer**

Breast cancer, a disease in which the cells in the breast grow uncontrollably and form tumors, is the most common type of cancer in the world. According to the World Health Organization, in 2020, over 2.3 million individuals worldwide were diagnosed with breast cancer. Furthermore, over 685,000 deaths occurred due to this fatal disease. The growing prevalence of breast cancer makes it imperative for new models and technology to be developed in order to classify possible tumors accurately.

The breast cancer dataset for this study was obtained from the “Breast Histopathology Images” dataset from Kaggle, a data science and AI platform under Google LLC. The images of the dataset were of patients with and without Invasive Ductal Carcinoma (IDC). IDC is the most common subtype of all breast cancers. The full 3 GB dataset consists of 198,738 IDC(-) images and 78,786 IDC(+) images. For the models of this investigation, 772 class 0 (non-IDC) images and 207 class 1 (IDC positive) images were used for training and validation purposes. The demographic information of the patients is unstated through the data source.

**Glaucoma**

Glaucoma is a leading cause of blindness, particularly for people 60 years and older. 3 million people in the United States have Glaucoma; the number rises to 80 million worldwide. This eye disease does not have noticeable symptoms early on, which makes it important to identify the disease easily at a low cost, something AI image classification can improve.

The glaucoma dataset used in this study was taken from the “Eye Disease retinal Images” dataset from Kaggle. This dataset consisted of images from sources such as Ocular recognition, IDRiD, and HRF. Other eye diseases such as Diabetic Retinopathy and Cataract are in the dataset, but for the purposes of this study only the glaucoma dataset was investigated. The Normal and Glaucoma classes of this dataset each contained approximately 1000 images. The demographic information of the patients is unstated through the data source.

**COVID-19**

COVID-19 is a highly contagious respiratory illness caused by the novel coronavirus, SARS-CoV-2. It spread globally, leading to a global catastrophe affecting tens of millions of people. Since the early symptoms of COVID-19 overlap with those of the flu and common cold, early detection of this disease can help control its widespread.

The COVID-19 dataset used in this study was taken from the ‘Chest X-Ray(Pneumonia, COVID-19, Tuberculosis) dataset from Kaggle. The dataset contains four categories(Normal, Pneumonia, COVID-19, Tuberculosis), a total of 7135 x-ray images are present.

**Apple scab**

Apple scab refers to a common disease caused by the fungus *Venturia inaequalis*, and is spread through airborne spores. This disease manifests as dark, scab marks which most commonly appear on an apple’s leaves, twigs, and the fruit itself. Leaves infected with this disease may fall prematurely, which weakens the tree and decreases the fruit yield. The scabs can lead to deformities, cracking, etc. which negatively impacts the quality of the fruit. This severely impacts the ability to grow and sell apples, therefore having a considerable impact on the produce industry, and efficient detection of this disease can aid farmers.

The apple scab data came from Kaggle’s New Plant Diseases data. The two categories are uninfected and infected. It is partially made with augmentation from an original dataset to reach its tens of thousands of images. Only a few hundred were used for this research.

CIFAR10 is a dataset full of 10 different classes representing airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. It was a popular dataset used in training machine learning projects at the start of the new age of machine learning.

The CIFAR10 dataset used in this study was downloaded through python. The University of Toronto hosts it. We downloaded a total of 8,000 images from this dataset, however only about 600 were used due to computation limits on Google Colab.

**Programming**

The AI models were coded on Google Colab notebooks using Python. Functions were imported from Tensorflow, Matplotlib, Numpy, Sci-kit learn and Pandas.

**Self-Supervised Learning**

The Self-Supervised Learning AI model used both a littleNtrain and Ntrain variable. They are so named for their relative magnitudes. LittleNtrain is the quantity of labeled data for training. This number directly influences the traditional fitting (overfitting or proper learning) in supervised learning. Ntrain is the quantity of unlabeled data used in the self-learning stage of the AI models. The self-learning can prepare the AI model to recognize visual patterns in the available data so that this recognition ability maybe useful in improving the results of supervised learning on the littleNtrain values later on.  Both values influence the and the validation accuracy and loss of the model

In most cases, the apple scab, breast cancer, CIFAR10, COVID-19, and glaucoma datasets use Ntrain values 400, 300, 600, 400, and 300 unlabeled images for self-learning respectively.

The self-learning model follows the simple Siamese representation strategy. Images go through a random sequence of cropping, rotating, and slight color shifting. Through this method, each original image can generate dozens of slightly edited images.

The model is compiled and the randomly modified images are fed to the model. The model uses a machine learning architecture to turn these input images into vectors representing them. The cosine loss function demonstrates the progress of the model as it measures how similar or different two inputs are. A well-trained model will achieve a low cosine loss when the two inputted images came from the same original image, and a large cosine loss then they did not. In this way, the model trains to recognize visual patterns in the input images. Note that this model never sees the labes for the input images.

After this self-learning, a short machine learning model is added on top of the vector representations of the previous self-learning model. This model is then trained in a traditional supervised learning setting, using littleNtrain original images and their labels.

The layers of the self-learning model include TensorFlow layers dropout, dense, and batch normalization.

Full code of Self Supervised Learning model is available at the github (Shtem). A lot of the self-learning code was adapted from the example on the Keras site (Paul).

**Supervised Learning**

Unlike the Self Supervised Learning AI model, the Supervised Learning AI model only trains on littleNtrain images. These models do not undergo self-learning beforehand and may then be at a disadvantage.

The supervised learning model does not train on any modified data. Instead, this model trains only on the original, labeled data source. Full code of the Supervised Learning model is available at the github (Shtem)

**ACKNOWLEDGMENTS (Optional)**

*This work would not have been possible without the structure of the SHTEM program at Stanford ran by Professor Tsachy Weissman and Sylvia Chin.*

**REFERENCES**

“SHTEM2023-SelfLearning”, The Papaya Institute -Github, August 2023

https://github.com/ThePapayaInstitute/SHTEM2023-SelfLearning

Kaggle: Your Machine Learning and Data Science Community, https://www.kaggle.com/. Accessed 20 July 2023.

Kaggle: Chest X-Ray (Pnemonia, COVID-19, Tuberculosis), 2 October 2022, https://www.kaggle.com/datasets/jtiptj/chest-xray-pneumoniaCOVID19tuberculosis?resource=download&select=train. Accessed 7 September 2023.

Kaggle: Eye Disease Classification, 2 October 2022, https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification?select=dataset. Accessed 7 September 2023.

He, Kaiming, et al. “[1512.03385] Deep Residual Learning for Image Recognition.” arXiv, 10 December 2015, https://arxiv.org/abs/1512.03385. Accessed 7 September 2023.

“Apple scab and pear scab.” RHS, https://www.rhs.org.uk/disease/apple-and-pear-scab. Accessed 7 September 2023.

“Apple Scab: Causes, Impacts, Prevention and Treatment Plan.” Rennie Orchards, https://rennieorchards.com/apple-scab/. Accessed 7 September 2023.

“Breast cancer.” World Health Organization (WHO), 12 July 2023, https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed 2 August 2023.

“Breast Histopathology Images.” Kaggle, https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images. Accessed 20 July 2023.

Chen, Xinlei. “[2011.10566] Exploring Simple Siamese Representation Learning.” arXiv, 20 November 2020, https://arxiv.org/abs/2011.10566. Accessed 7 September 2023.

“CIFAR10 Dataset.” Papers With Code, https://paperswithcode.com/dataset/CIFAR10. Accessed 7 September 2023.

“Glaucoma.” Johns Hopkins Medicine, https://www.hopkinsmedicine.org/health/conditions-and-diseases/glaucoma. Accessed 7 September 2023.

“New Plant Diseases Dataset.” Kaggle, https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset. Accessed 7 September 2023.

Paul, Sayak. “Self-supervised contrastive learning with SimSiam.” Keras, 19 March 2021, https://keras.io/examples/vision/simsiam/. Accessed 1 August 2023.

**Figures and Figure Captions**

A screenshot of a graph

Description automatically generated

**Figure 1. Apple scab training loss for supervised and self-learning.** Supervised learning curve improves with increased data, as does the self-learning curve. For small dataset sizes, self-learning improves results.

**A screenshot of a graph

Description automatically generated**

**Figure 2. Apple scab validation accuracy for supervised and self-learning models.** Both models show increasing accuracy with increased dataset sizes. The self-learning model only improves performance for a few select dataset sizes.

**A screenshot of a graph

Description automatically generated**

**Figure 3. Breast cancer training loss for supervised and self-learning models.** Both the supervised and self-learning models’ performances decrease—with increasing loss—as dataset sizes increase. Self-learning constantly provides a worse performance than the supervised model.

A screenshot of a graph

Description automatically generated

**Figure 4. CIFAR10 training accuracy for supervised and self-learning models.** While the supervised learning curve has diminishing performance with increasing dataset size, the self-learning performance increases with growing dataset size and later plateaus. Self-learning has worse performance throughout all dataset sizes.

A graph of a graph

Description automatically generated with medium confidence

**Figure 5. COVID19 training accuracy for supervised and self-learning models.** The supervised learning curve worsens in performance with increased dataset size while the self-learning curve has unstable performance. The self-learning model improves results for larger dataset sizes.

A graph of a number of points

Description automatically generated with medium confidence

**Figure 6. Glaucoma training loss for supervised and self-learning models.** Both curves show improving performance as dataset size increases. Self-learning model performance is worse throughout all dataset sizes tested.

A screenshot of a graph

Description automatically generated

**Figure 7. Glaucoma validation accuracy for supervised and self-learning models.** The supervised learning performance is rather steady across different dataset sizes. The self-learning performance is unstable. Self-learning outperforms supervised learning only in smaller dataset sizes.

**Breast cancer**

|  |  |
| --- | --- |
| Class 0 (non-IDC cells) | Class 1 (IDC positive cells) |
| A close up of a pink and purple background  Description automatically generatedA close up of a pink and blue speckled surface  Description automatically generatedA close up of a pink surface  Description automatically generatedA close up of a cell  Description automatically generatedA close up of a pink and white cell  Description automatically generated | A close up of a cell  Description automatically generatedA close up of a purple and white object  Description automatically generatedA close up of a red and white object  Description automatically generatedA purple and white background  Description automatically generatedA close-up of a purple and white background  Description automatically generated |

**Glaucoma**

|  |  |
| --- | --- |
| Normal eyes) | Glaucoma positive eyes) |
| A close-up of a human eye  Description automatically generatedA close-up of a human eye  Description automatically generatedA close-up of a human eye  Description automatically generatedA close-up of a red vein  Description automatically generated | A close-up of a red light  Description automatically generatedA close-up of a red sphere  Description automatically generatedA close-up of a human eye  Description automatically generatedA close-up of a human eye  Description automatically generated |

**COVID-19**

|  |  |  |  |
| --- | --- | --- | --- |
| Normal | COVID-19 positive | Pneumonia | Tuberculosis |
| X-ray of a child's chest  Description automatically generatedA x-ray of a person's chest  Description automatically generated | X-ray of a chest with cables  Description automatically generatedX-ray of a chest with a lung and lung pleural  Description automatically generated with medium confidence | A x-ray of a person's chest  Description automatically generated X-ray of a person's chest  Description automatically generated | A x-ray of a person's chest  Description automatically generated A x-ray of a chest  Description automatically generated |

**Apple scab**

|  |  |
| --- | --- |
| Normal leaves | Apple scab positive leaves |
| A close-up of a leaf  Description automatically generatedA close-up of a leaf  Description automatically generatedA close-up of a leaf  Description automatically generatedA close-up of a leaf  Description automatically generated | A close-up of a green leaf  Description automatically generatedClose-up of a green leaf  Description automatically generatedA close-up of a leaf  Description automatically generatedA close-up of a green leaf  Description automatically generatedA close-up of a leaf  Description automatically generated |

**CIFAR10**

|  |
| --- |
| Categories: planes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. |
| A collage of images of animals  Description automatically generated |

**Figure 8. All datasets used in learning.** The images above are samples of the thousands of images used in all the experiments. Examples of different categories in each dataset are present.