On Teacher-Student Semi-Supervised Learning for Chest X-ray Image Classification

Roberto Augusto Philippi Martins and Danilo Silva
Universidade Federal de Santa Catarina
Department of Electrical and Electronic Engineering
Florianópolis, Brazil
Emails: robertophi@gmail.com, danilo.silva@ufsc.br

Abstract—**The lack of labeled data is one of the main prohibiting issues on the development of deep learning models, as they rely on large labeled datasets in order to achieve high accuracy in complex tasks.**

Our objective is to evaluate the performance gain of having additional unlabeled data in the training of a deep learning model when working with medical imaging data. We present a semi-supervised learning algorithm that utilizes a teacher-student paradigm in order to leverage unlabeled data in the classification of chest X-ray images.

Using our algorithm on the ChestX-ray14 dataset, we manage to achieve a substantial increase in performance when using small labeled datasets. With our method, a model achieves an AUROC of 0.822 with only 2% labeled data and 0.865 with 5% labeled data, while a fully supervised method achieves an AUROC of 0.807 with 5% labeled data and only 0.845 with 10%.

Index Terms—**Semi-Supervised learning, Teacher-Student, Chest X-Ray, Medical Image Classification**

I. Introduction

Deep learning models rely on optimizing parameters for specific tasks, requiring large labeled datasets for producing sufficiently accurate image classification models. Many tasks require thousands of images to train a model for each class, increasing with the necessary accuracy.

With the development of Computer Assisted Diagnostic (CAD) tools with the aid of deep learning models, we face the challenge of obtaining sufficiently large labeled datasets for the specific pathology we intend to classify. Some tasks require high accuracy in order to be useful in the diagnostic process, as wrong predictions can have great impact in the patient's health.

The lack of labeled data is one of the first prohibiting issues on the development of many deep learning models, even more so for pathology classification and detection. Acquiring accurate labels for medical images requires the work of medical professionals, which are often difficult to obtain or expensive.

Although labeled data is difficult to obtain, it is often possible to have access to a larger, but unlabeled dataset. This kind of data has no explicit information we can use in the supervised training of a deep learning model, but it still carries information in the context of the image. This motivates us to search for a method that can use both the information from the labeled and unlabeled data in the training of deep learning models.

This class of algorithms is called Semi-Supervised Learning (SSL), in which we leverage unlabeled data as well as labeled data in order to increase the model accuracy when compared to an exclusively supervised training method.

One approach for semi-supervised learning is the teacher-student pipeline, in which two models are used in a multi-step training algorithm in order to make use of the unlabeled data. In this framework the unlabeled data is leveraged in the training process by the student model, using a set of pseudo labels produced by the teacher model.

Our objective in this paper is to evaluate the performance gain of having additional unlabeled data in a medical imaging dataset when training a deep learning model, using a teacher-student approach. We want to measure the value of labeling data when compared to using unlabeled data in semi-supervised training.

We have chosen to use the dataset ChestX-ray14 for our tests, as it is one of the largest medical imaging datasets available. With over 100.000 frontal view X-ray images with 14 classified diseases (see Fig. 1), it has allowed the development of large, accurate models that match the radiologist-level performance for pneumonia detection [1].

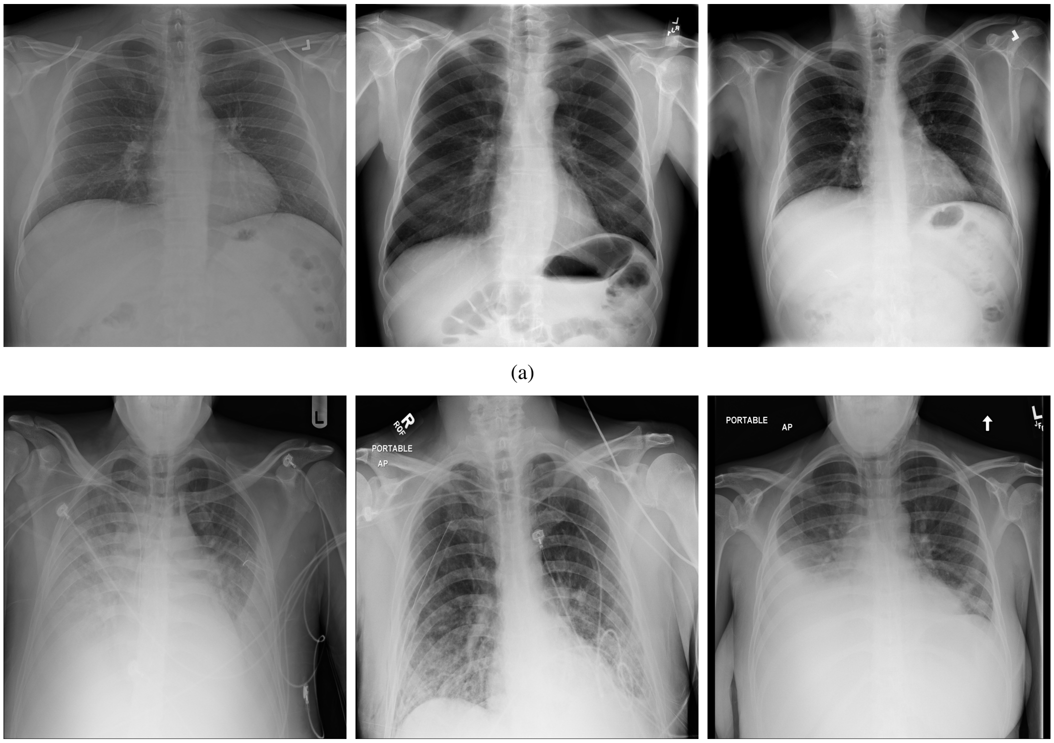
In this paper we use a teacher-student semi-supervised learning algorithm to train a convolutional neural network for classification of chest X-ray images from the ChestX-ray14 dataset—which, to the best of our knowledge, had not been done before. We evaluate the performance of deep learning models with only partially labeled datasets, considering several proportions of labeled data.

The main contributions of this paper are:

- Evaluation of a teacher-student pipeline for semi-supervised learning in the ChestX-ray14 medical imaging dataset;
- Pseudo labels processing and balancing for filtering out low confidence predictions;
- Performance comparison for different proportions of labeled data.

II. Related works

There are multiple approaches for semi-supervised learning on convolutional neural networks, such as adversarial learning [2], consistency learning [3], [4], contrastive learning [5], [6],



(a)

(b)

Fig. 1: Chest x-ray images from the ChestX-ray14 dataset. (a) Images with no identified pathology. (b) Images labeled with pleural effusion.

the use of pretext tasks [7], [8], teacher-student pipeline [9], [10] and others [11].

In this section we review some of these approaches, as well as some important concepts related to our work.

A. Self-supervised Learning

Self-supervised learning is a form of unsupervised learning which has many similarities with semi-supervised algorithms [10]. It has the objective of learning good feature representation by utilizing information contained in the images themselves, and not in the labels.

This methods generally involves the use of pretext tasks, meaning that it works by optimizing functions that may not be directly related to the main objective [12]. Ultimately, a model trained this way can be adapted to a supervised task by fine-tuning the model with extra labeled data, in such a way that the self-supervised training works as a pre-training step to the algorithm [13].

Multiple pretext tasks have been developed for use with image classification, such as colorization [8], tracking [14] and Jigsaw puzzle reassembly [7].

B. Contrastive Learning

SimCLR [6], [15] and MoCo [12] are both popular frameworks that make use of a contrastive loss for learning visual representations effectively. They work with unlabeled data by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. The model

Altough this methods achieve promising results, they require a large amount of computational power in order perform well because they fundamentally rely on large batch sizes to work. This problem becomes even worse when working with medical images, as they often have larger resolution and smaller points of interest.

C. Teacher-Student and pseudo labels

Our paper is based on the work of Yalniz et al. [10], which introduces a teacher/student based pipeline for semi-supervised learning.

This algorithm makes use of a two step process similar to information distillation [10], in which a teacher model is first trained with the labeled data and the student model is trained to reproduce the best outputs of the teacher. Yalniz et al. have shown that this method can be adapted to leverage large amounts of unlabeled data into an image classification task.

Our objective is to use this method on the ChestX-ray14 dataset, adapting it to work on a much smaller dataset.

A similar approach is taken by Shaw et al. [16] for histology image classification, extending the approach to a chain of teacher-student models, where the student becomes the teacher to the next student. This chain of teacher/student models shows improvement over the single teacher-student models when training for multiple iterations with less than 1% of labeled data

Xiao Qi et al. [17] also use a teacher-student pipeline for classifying COVID-19 diagnosis on a different chest X-ray dataset. In addition to the semi-supervised training, they utilize a model and data augmentation methods specific to the classification of COVID-19, such as local phase features. They also perform a filtering step in the pseudo labels generated for the unlabeled dataset, although the results obtained differ from ours. This difference can be explained by the distinct dataset and task, as even though they both contain chest x-ray images, many differences can occur in the quality of images and labels.

III. Methods

This section describes the steps required for training the model, as well as some important details in the execution of the pipeline.

A. *Teacher/student pipeline*

Our algorithm is inspired by the training pipeline shown by Yalniz et al. [10] which is based on a teacher/student method. This algorithm makes use of the unlabeled dataset by creating pseudo labels for the data based on a preliminary training done solely on the labeled dataset. This means we first train the teacher model on the labeled dataset, and use this model to produce labels for the unlabeled data.

This algorithm is agnostic to the model used, and is flexible with training of both the teacher and student models. In this paper we will be considering the teacher and student model to have the same architecture.

For our experiments, we divide the original training dataset of size N into two parts of different size using a ratio λ. This results in a labeled dataset $\mathcal{D}$ of size $N_l = \lambda N$ and a unlabeled dataset $\mathcal{U}$ of size $N_u = (1 - \lambda)N$.

Given that we have a teacher model θ^t and a student model θ^s, the training steps are as follow:

1) We use the labeled dataset $\mathcal{D}$ to train the teacher model θ^t using the standard cross entropy loss.
2) Classify the images from the unlabeled dataset $\mathcal{U}$ using the teacher model θ^t, producing a set of labels $\mathcal{L}$.
3) Filter the set of pseudo labels $\mathcal{L}$, keeping only the high confidence predictions and maintaining class balance.
4) Train the student model θ^s using only the unlabeled dataset $\mathcal{U}$ and the filtered labels.
5) Fine-tune the student model using the labeled dataset $\mathcal{D}$.

B. *Prediction filtering*

One of the possible problems with the pseudo labels created for the unlabeled dataset is that we may be producing incorrect labels for the data. Although our algorithm is tolerant to noisy labels [9], they can decrease the ability for our model to learn. In order to avoid this issue, we consider only the high confidence predictions, filtering out the remaining data.

In this process, we also guarantee that the class balance remains the same between the labeled dataset and the new set of labels. We match the proportion of classes in the set with the statistics obtained in the labeled dataset.

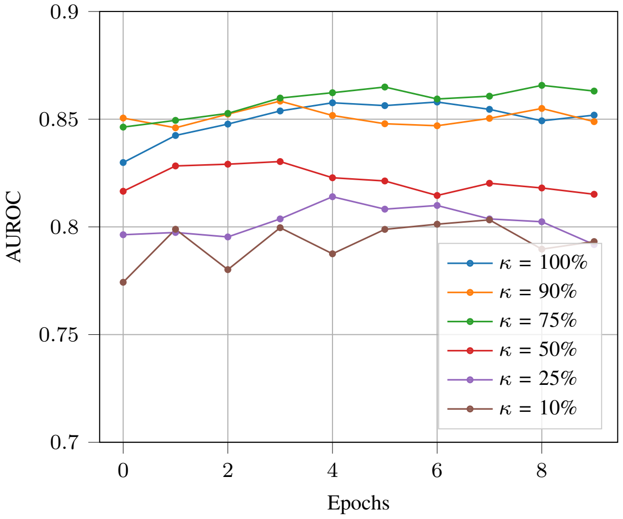


Fig. 2: Effects of the parameter κ in the training of the student model. This test is done in the case $\lambda = 10\%$. We can see both higher and lower values of κ can be detrimental to the model's performance, and that the highest AUROC value is obtained with $\kappa = 75\%$.

In order to do this, we define the threshold κ as the proportion of images that will be kept for training. This means that only the best $K = \kappa N_u$ predictions on $\mathcal{U}$ will be kept for the intermediate training of the student model, distributed for each class proportionally to the class distribution seen on the labeled dataset.

As per Fig. 2, we optimize the value of κ in order to maximize the AUROC (Area Under the Receiver Operating Characteristic curve) [18] of the student model. We found that, for our dataset and task, a value of $\kappa = 75\%$ is optimal for our setup. We can conclude that a small κ results in poor performance caused by the lack of data, while using too many images can deteriorate the performance by including low-confidence and flipped labels in the training.

C. *Data augmentation*

Data augmentation is a very important aspect of training deep learning models, as it decreases overfitting effects on the training data and ultimately increases the model's performance [19]. The use of a small labeled dataset in semi-supervised learning makes powerful models like DenseNet121 even more susceptible to overfitting, meaning we must pay more attention to the augmentation pipeline used.

We have tested the effects of data augmentation in the training of the teacher and student models by evaluating the training results with multiple different augmentation pipelines. As per Fig. 3, we can see that the student model achieves better results when using strong augmentation functions that modify the image in a meaningful way.

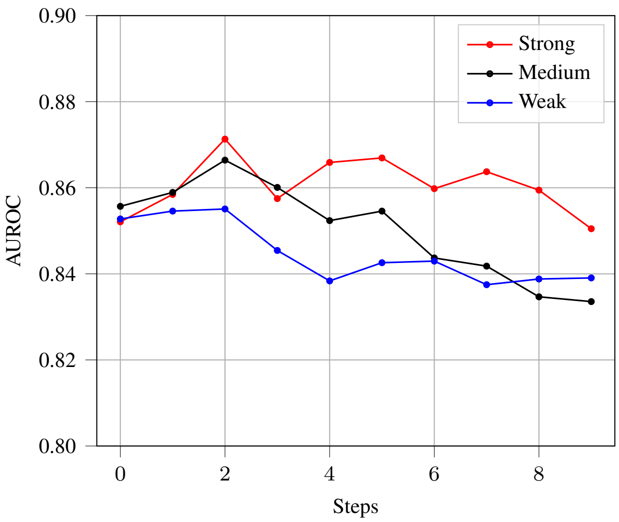


Fig. 3: AUROC values on the validation set during the fine-tuning of the student model on the labeled dataset, given different augmentation pipelines. We see that a strong augmentation results in better performance for the student model. Tests were done with $\lambda = 10\%$, with all models starting from the same pre-trained model.

The same behavior was not observed while training the teacher model or in the initial training of the student model with the dataset $\mathcal{U}$. In these cases, little difference was observed on the performance of the models, and we use the same augmentation as the student model in the remaining tests.

For the remaining experiments in this paper, we use the following augmentations:

- Random horizontal flip (with 50% probability)
- Random rotation (max 25°)
- Random translation (max 15% in any direction)
- Random scaling (in range $(0.9, 1.1)$)
- Random perspective transform (max 0.2 distortion)

IV. Experimental Setup

A. Models

We use for both the student and teacher model a DenseNet [20] network. Our algorithm allows us to choose any architecture for the models, as the training pipeline is transparent to this choice.

We have decided to use DenseNet121 models for our tests based on the results of Rajpurkar et al. [21], and it shows to be a good baseline for this problem.

Starting with a DenseNet121 model pretrained on the ImageNet dataset [22], we replace the last layer with a fully connected layer with sigmoid activation.

B. Dataset

For both the labeled and unlabeled datasets we use the ChestX-ray14 dataset released by Wang et al. [1], which contains 112120 frontal-view X-ray images. In total, the dataset contains 14 different pathologies, but we restrict our tests to the subset of the data that either contains the pathology Effusion or no pathology. From the total of 73719 images, we use $N = 51666$ images for training, 7371 for validation and 14682 for testing.

We constructed the labeled dataset $\mathcal{D}$ and unlabeled dataset $\mathcal{U}$ by randomly picking images from the full training dataset and hiding the labels for $\mathcal{U}$.

As our objective is to analyze the effects of different proportions of labeled to unlabeled data in the accuracy of the models, we divided the original dataset into multiple cases. We have created subsets with $\lambda = \{20\%, 10\%, 5\%, 2\%\}$, containing $N_l = \{10333, 5166, 2581, 1033\}$ labeled images and $N_u = \{41333, 46500, 49085, 50633\}$ unlabeled images respectively.

C. Training Details

For training, we re-scale the images to a resolution of 256x256, training with a batch-size of 16. We use an Adam optimizer with learning rate starting from 10^{-4}, multiplying the learning rate by a factor of 0.9 at each epoch. We train the teacher model for 25 epochs, the student model for 10 epochs with the pseudo labels and 15 epochs with the labeled data.

The experiments are run on a single GTX1660 with 6GB of memory and 16GB of RAM.

For evaluating the validation and test datasets we use TTA (Test-time Augmentation), which show to increase the accuracy of predictions especially for smaller datasets [23]. We produce ten 224x224 different crops from the original image, taking the four corners and the central crop, plus the horizontally flipped version of these, and making an independent prediction for each crop. The final prediction for the given image is the average of the ten crop predictions.

For training the teacher and student model, we have used strong affine augmentations for the images, including random rotations, scaling and translation, as well as random perspective transforms.

V. Results

We evaluate our algorithm on validation and test datasets, both derived from the ChestX-ray14 dataset. We measure the AUROC values for both the teacher and student models at every epoch with the validation dataset (see Fig. 4) and with the test dataset (see Table I) once training is finished.

From Fig. 4 we can see that using our method of semi-supervised learning the student model has better accuracy and stability during training, with more significant results for smaller values of λ.

We can see from Table I a gain in performance at every value of λ tested. Even though the intermediate step for training the student model uses only pseudo labels, it still achieves a higher AUROC than the teacher model itself, except for the case $\lambda = 20\%$.

The increase in AUROC is more prominent at lower proportions of labeled data, as the model benefits proportionally more from the extra information gained from the unlabeled dataset. The largest increases in AUROC occurs at $\lambda = 2\%$,

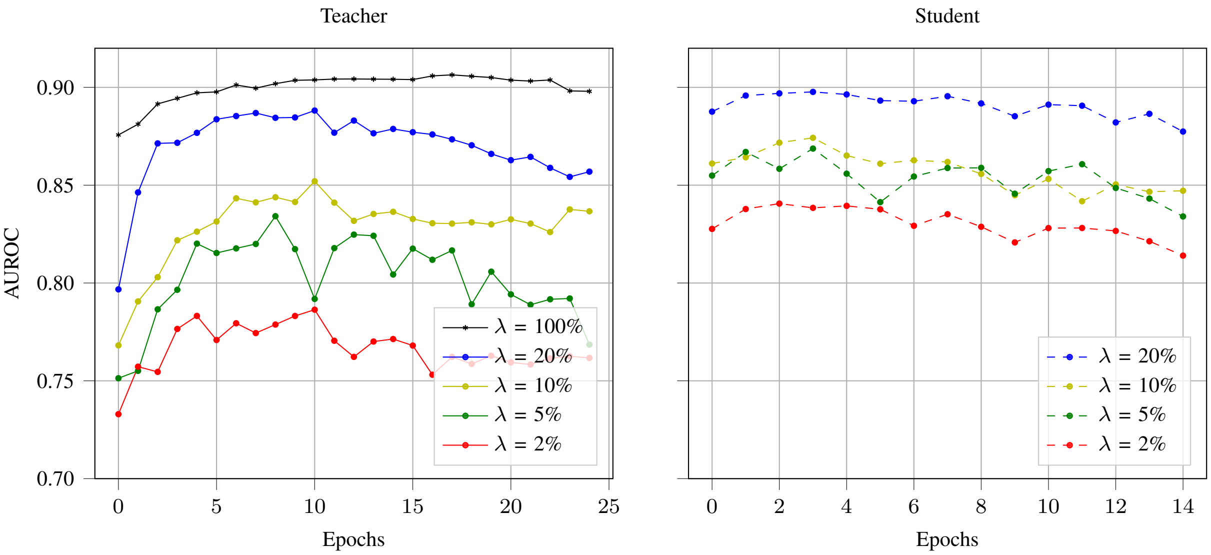


Fig. 4: AUROC values on the validation set during the training of the teacher and student models, given different proportions of labeled data. We see that the performance for both models gets better as the number of labeled data increases. Our student model trained with 2% of the labeled data (AUROC=0.841) almost matches the AUROC values from the teacher model with 10% of labeled data (AUROC=0.852) in the validation set. We also show the results for training the teacher model with $\lambda = 100$, meaning a fully supervised model with the entire labeled dataset.

where the AUROC increases from 0.750 to 0.822, while the smallest one occurs at $\lambda = 20\%$, with a small increase from 0.887 to 0.893.

From Table I we can also see that the student model trained with $\lambda = 2\%$ has better results than the teacher model trained in a supervised method with $\lambda = 5\%$, even though it has less than half the amount of labels to train. The same can be observed for the student trained with $\lambda = 5\%$, which has a better result than the teacher at $\lambda = 10\%$.

TABLE I: Comparison results from the baseline model (teacher) and the student model. The intermediate results show the AUROC values for the student model trained only in the unlabeled dataset.

Labeled data	Teacher	Intermediate	Student
2%	0.750	0.795	0.822
5%	0.807	0.823	0.865
10%	0.845	0.869	0.879
20%	0.887	0.867	0.893
100% (fully supervised)	0.906	-	-

VI. Conclusion

In this paper we have considered a teacher-student algorithm for leveraging unlabeled data in the training of deep learning models and evaluated its performance on the ChestX-ray14 image classification dataset.

We have shown that this technique can have a substantial benefit to the performance of models, when compared to the purely supervised counterparts.

We can see that there is a diminishing return in labeling data beyond a certain point, because a combination of labeled and unlabeled data can be sufficient to reach the performance of a fully supervised training method.

This kind of algorithm can be of great benefit for developing CAD (Computer-Assisted Diagnostic) tools, as it decreases the reliance on labeled data. Deep learning techniques have shown promising results in the area of medical imaging, and we hope that semi-supervised learning can help the development of practical tools for medical professionals.

Future works include developing a multi-label framework, allowing the easier training of models for multiple pathologies. It is also possible to include an iterative training pipeline [9], [16] by replacing the teacher model with the current student and restarting the training steps, and using probability distributions for the pseudo labels instead of binary values.

References

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *CoRR*, vol. abs/1705.02315, 2017. [Online]. Available: http://arxiv.org/abs/1705.02315

[2] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, apr 2018. [Online]. Available: http://arxiv.org/abs/1704.03976

[3] Q. Xie, Z. Dai, E. H. Hovy, M. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," *CoRR*, vol. abs/1904.12848, 2019. [Online]. Available: http://arxiv.org/abs/1904.12848

[4] A. Tarvainen and H. Valpola, "Mean teachers are better role models : Weight-averaged consistency targets improve semi-supervised deep learning results," *CoRR*, vol. abs/1703.01780, 2017. [Online]. Available: http://arxiv.org/abs/1703.01780

[5] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[6] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” *CoRR*, vol. abs/2002.05709, 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[7] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” *CoRR*, vol. abs/1603.09246, 2016. [Online]. Available: http://arxiv.org/abs/1603.09246

[8] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” *CoRR*, vol. abs/1603.08511, 2016. [Online]. Available: http://arxiv.org/abs/1603.08511

[9] Q. Xie, E. H. Hovy, M. Luong, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” *CoRR*, vol. abs/1911.04252, 2019. [Online]. Available: http://arxiv.org/abs/1911.04252

[10] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, “Billion-scale semi-supervised learning for image classification,” *CoRR*, vol. abs/1905.00546, 2019. [Online]. Available: http://arxiv.org/abs/1905.00546

[11] A. Chaudhary, “Semi-supervised learning in computer vision,” 2020, https://amitness.com/2020/07/semi-supervised-learning/.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” *CoRR*, vol. abs/1911.05722, 2019. [Online]. Available: http://arxiv.org/abs/1911.05722

[13] A. Newell and J. Deng, “How useful is self-supervised pretraining for visual tasks?” pp. 7343–7352, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9157100

[14] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, “Tracking emerges by colorizing videos,” *CoRR*, vol. abs/1806.09594, 2018. [Online]. Available: http://arxiv.org/abs/1806.09594

[15] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, “Big self-supervised models are strong semi-supervised learners,” *CoRR*, vol. abs/2006.10029, 2020. [Online]. Available: https://arxiv.org/abs/2006.10029

[16] S. Shaw, M. Pajak, A. Lisowska, S. A. Tsaftaris, and A. Q. O’Neil, “Teacher-student chain for efficient semi-supervised histology image classification,” *CoRR*, vol. abs/2003.08797, 2020. [Online]. Available: https://arxiv.org/abs/2003.08797

[17] X. Qi, J. L. Nosher, D. J. Foran, and I. Hacihaliloglu, “Multi-feature semi-supervised learning for covid-19 diagnosis from chest x-ray images,” 2021. [Online]. Available: https://arxiv.org/abs/2104.01617v2

[18] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, 1997.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” vol. 25, 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[20] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[21] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” *CoRR*, vol. abs/1711.05225, 2017. [Online]. Available: http://arxiv.org/abs/1711.05225

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, “Imagenet large scale visual recognition challenge,” *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[23] D. Shanmugam, D. W. Blalock, G. Balakrishnan, and J. V. Guttag, “When and why test-time augmentation works,” *CoRR*, vol. abs/2011.11156, 2020. [Online]. Available: https://arxiv.org/abs/2011.11156