Medical Image: Alzheimer's Disease Pathological Image Recognition and Classification

Pu Sun
Electrical and Computer Engineering
University of California, Davis
Davis, CA, USA
psun@ucdavis.edu

Zheng Zou
Electrical and Computer Engineering
University of California, Davis
Davis, CA, USA
zgzou@ucdavis.edu

***Abstract*—Supervised learning (SL) techniques, including deep neural networks, have shown promising results in pathology image analysis. However, acquiring a comprehensive and well-annotated dataset can be expensive and labor-intensive, particularly when domain expertise from neuropathologists is required. To address the scarcity of labeled data, this research investigates the effectiveness of semi-supervised learning (SSL) and transfer learning (TL) on a dataset comprising Amyloid plaques, which are characteristic markers of Alzheimer's disease. Initially, a supervised learning model (ResNet) was employed to assess the upper limit of deep learning methods for this problem. Subsequently, a state-of-the-art SSL method (FixMatch) was adapted for multi-label classification. Lastly, transfer learning was explored using a limited labeled dataset to evaluate the transferability of a model trained on ImageNet to this pathology problem. Our experiments revealed that SL consistently achieved good results when the labeled dataset was sufficiently large, but faced overfitting issues with small datasets. Furthermore, FixMatch proved unsuitable for multi-label classification due to its sole focus on prediction consistency as the training loss. Training an effective model solely with single-labeled images was challenging. Finally, the transferability of a pre-trained model from standard vision databases was found to be severely limited in this problem, given the significant visual differences between natural images and pathological images.**

I. Introduction

Alzheimer's Disease (AD) is a neurological disease common in the elderly. Approximately one in nine people age older than 65 years old in U.S. suffer from AD dementia [1]. Such disease causes problems such as memory impairment, language disorders, and disorientation. With the development of medical technology, although doctors cannot completely cure this disease , medical methods can be used to reduce the symptoms and improve the patients' quality of life in the case of early detection. Neurologists typically assess histopathology by detecting and identifying extracellular plaques, an important pathological feature of AD. Therefore, the identification of AD pathological plaques has become a meaningful topic in the field of medical image recognition.

In previous studies, models trained with Supervised Learning using Convolutional Neural Network (CNN) have achieved excellent results. Related experts train very accurate models by labeling more than 60,000 patch images so that many medical image problems will abandon the use of machine learning for identification. Therefore, our goal is to avoid the need for a large number of labeled images as a training set for machine learning. We proposed three methods to achieve our goal: Semi-Supervised Learning, Transfer Learning, and Supervised Learning.

Semi-Supervised Learning (SSL) methods, especially FixMatch and its improved versions, have shown excellent results in many data sets. However, FixMatch is rarely used in medical data sets, and it cannot recognize multi-label images.

Transfer learning has the potential to use incomplete datasets to train and get even more impressive results, because the original model may have very high achievement in various image recognition tasks.

Directly using less images for Supervised Learning training is a way to reduce the number of labeled images. It may overfit the model, but there is currently no relevant experiment on this dataset to prove the feasibility of this method.

Our experiments were dedicated to completing different challenges in utilizing different Supervised Learning models, FixMatch improvements, and transfer learning. We were searching for effective ways to use fewer labeled images in model training, while achieving decent performance.

II. Related Work

Tang et al. designed a customized 6-layer shallow CNN model to detect and classify Cored, Diffuse, and CAA plaques. With learning rate of 8E-5 and Adam optimizer, we are able reproduce excellent results using our processed dataset, which agrees to their paper. However, they achieved the result with all the plaque images available. The 70, 000 images were adapted from 43 Whole Slide Images, which takes experts 15 hours to label all the plaque information. Our models aim to reduce the labeling costs by using less labeled data.

III. Methodology

A. Dataset

This dataset was collected by Tang et al. from "Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline". As shown in Figure 1, they performed a comprehensive scan of brain extracellular plaques and then performed WSI color normalization. They then cut out the plaque area using an automated segmentation system. Finally, it was handed over to experts for image annotation.

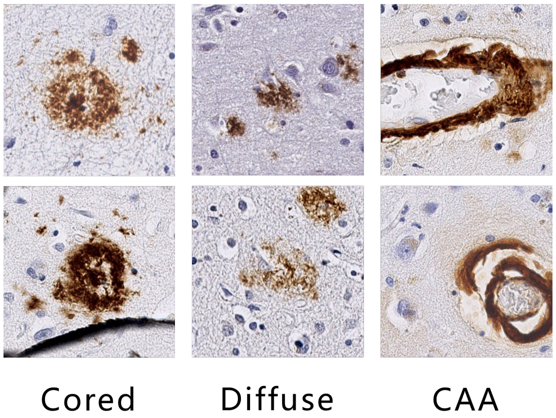

Fig. 1. Example of 3 classes in the dataset

Three different types of plaques, Cored, Diffuse, and CAA, were present in the entire data set. There may be 0 to 3 different plaques in each image. The overall image distribution is shown in Table 1.

TABLE I
NUMBER OF IMAGES IN ORIGINAL DATA SET

Plaque	Cored	Diffuse	CAA
Train	2140	48120	2222
Validation	381	7486	124
Hold-out	98	10480	6
Total	2619	66086	2352

B. Dataset Processing

We noticed that the train, validation, and hold-out sets have serious unbalancing problems. The minority classes are Cored and CAA, where Diffuse outnumbered them over twenty times. In addition, we also found that the number of Cored and CAA in the Hold-out set are inadequate in the actual hold-out testing. Therefore, we decided to mix all the images together and split them in 8:1:1 ratio for Train, Validation, and Hold-out sets. Also, to make the model train better, we up-sampled the minority class, Cored and CAA, in the Train set. Cored was up-scaled by 25 times, and CAA was up-scaled by 28 times. In the end, we have no class imbalance problem for training, and Validation and Hold-out dataset can be used to validate and test the model normally. The dataset information is in Table 2.

TABLE II
NUMBER OF IMAGES AFTER RE-BALANCING AND UP-SAMPLING

Plaque	Cored	Diffuse	CAA
Train	53107	74120	53150
Validation	332	6720	270
Hold-out	313	6691	272
Total	53752	87531	53692

C. Training Phase I: Supervised Learning With All Data

During the early phase of senior design project, we wanted to be familiar with the dataset and code. So, we did this Supervised Learning part and trained models using ResNet34 and RegNet02X. We were eager to figure out how the state-of-the-art models perform in medical image scenario like this. Then, we can compare them to the Author's model in terms of accuracy, precision-recall, and receiver operating characteristic. Also, the results from this phase can serve as a baseline for our next methods, Semi-Supervised Learning and Transfer Learning. We can compare how the model is trained with less labeled data using the baseline.

For ResNet34, we used the one provided in the PyTorch library.

D. Training Phase II: Semi-Supervised Learning With Fix-Match

The feature of FixMatch is that it performs both strong augment and weak augment processing on the image in the unSupervised learning part. When the model recognizes the image of weak augment, when the model's confidence in a certain class is higher than a threshold, it will mark this class as this image label. Then use this pseudo-label to compare the image processed by strong augment to calculate the loss value. Therefore, we also perform weak and strong processing on the image, as shown in Figure 2. Our weak augment processing includes random rotation, random cropping (around 10 pixels), and random horizontal and vertical flipping. Our strong augment processing uses the same AutoAugment as FixMatch. This processing is superimposed on the weak augment.

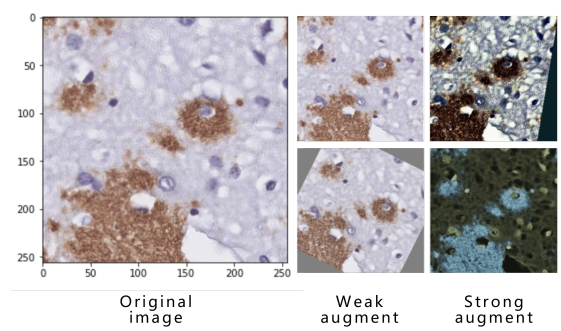

Fig. 2. Examples of Weak augment and Strong augment based on a image classified in Diffuse

The biggest problem with FixMatch not being able to be used directly on our dataset is that it is not suitable for the recognition of multiple labels. This algorithm uses the Softmax function when making pseudo-labels. This function causes predictions to be set to a probability distribution that sums to 1 for all classes. In this way, when there is only one label per image, the category with the highest probability can be directly selected as the pseudo label. However in multiple tabs this does not work. We try to replace this function with Sigmoid, so that the picture can have a probability between 0 and 1 for each category, that is, the confidence level of the model to judge each category. Then we created thresholds for each category separately. As shown in Figure 3, each category that exceeds the threshold will be selected as a pseudo-label.

At the same time, we also try to delete all the data with multiple labels in the dataset to verify whether the original FixMatch uses the plaque image.

E. Training Phase III: Transfer Learning With Limited Data

Since our goal is to try to use fewer images as possible when training, Transfer Learning can be an excellent method. Transfer learning relies on using existing pre-trained models, which are trained extensively on ImageNet dataset. Because the models are fully trained and learned images' characteristics extensively, we can leave the CNN layers untouched and change the final fully-connected layer for classification output. We believe that a pre-trained model will contain enough knowledge to recognize plaque's features so we freezed all the CNN blocks and started training.

Also, we changed the training dataset so that it only contains 10 percent of the total image data. We randomly selected 250 Cored, 6, 000 Diffuse, and 250 CAA images from our remade training set, and up-sampled Cored and CAA to 6, 000 to solve the unbalancing problem.

F. Training Phase IV: Supervised Learning With Limited Data

Lastly, we are interested in how the Supervised models will learn from limited amount of data. Using the 10 percent data in training phase III, we can figure out if ResNet34 or RegNet02X can learn enough information of Cored, Diffuse, and CAA plaque. If the model learns quite well, we are able to tell that AD pathological image detection and classification only requires a small amount of label data and less human labor.

IV. Evaluation and Results

The evaluation for our model's performance includes the testing metrics in the testing phase. The metrics include accuracy, loss, area under the curve for receiver operating characteristic, area under the curve for precision-recall, F1 score and weighted average. Other than loss, we want all other metrics as high as possible. Other than that, we also includes confusion matrix data to analyze the model's success rate for predicting each class.

A. Training Phase I: Supervised Learning With All Data

Supervised Learning results were strong and was comparable to the author's model. Using ResNet34 and RegNet02X with 5E-4 learning rate, GDM optimizer, batch size of 64, and 30 epoch runs, we were able to get data in table 3.

After listing all the information for accuracy, area under the curve for receiver operating characteristic (ROC), and Precision-Recall (P-R), we found that all the numbers are very close to each other. All the models achieved high accuracy and ROC and Precision-Recall performance, meaning that the models will have excellent capability of recognizing and classifying the three differen plaques. We also achieved

TABLE III
Results for Supervised Learning with All Data

	ResNet34	RegNet02X	Tang et al.'s model
ROC Cored	0.980	0.975	0.981
ROC Diffuse	0.988	0.977	0.986
ROC CAA	1.000	1.000	1.000
R-P Cored	0.754	0.748	0.742
R-P Diffuse	0.988	0.998	0.999
R-P CAA	0.999	0.984	0.987
Acc Cored	0.938	0.971	0.918
Acc Diffuse	0.963	0.972	0.947
Acc CAA	0.993	0.997	0.994

this result using less computational resources. Compared to author's running of 60 epoch with 8 minutes per epoch, we only ran 30 epoch with 7.5 min per epoch. Our model saved about half of the total training time. With the numbers achieved in this phase, we can conclude that the state-of-the-art model, ResNet and RegNet, is well-suited in detecting and classifying plaques. We can then use the models' result as a baseline and compare it to our later phases of training.

B. Training Phase II: Semi-Supervised Learning With FixMatch

In the training of fixmatch, we found that although the loss is decreasing, the actual accuracy and precision have not improved, especially the accuracy of Cored has been decreasing as shown in the figure. So we experimented with different thresholds. Table 4 shows our results, which show that the model does not substantially improve at different thresholds. We also tried different model: Resnet18, wide Resnet and turns out that the result is not good. Only the wide Resnet we use initially gives a relatively better result as shown in table 5. However, the results are still poor compared to Supervised learning.

For all the data in table 4 and 5 are trained based on the upsampled and rebalanced dataset with 32 epochs, 1024 iteration per epoch, image size: 64, batch size: 32, learning rate: 0.0001

We also perform semi-supervised learning training on images with only one label. The result is not good enough, most of the cored is recognized as diffuse, as shown in the confusion matrix in Figure 4.

C. Training Phase III: Transfer Learning With Limited Data

After Semi-Supervised Learning, we then implemented Transfer Learning with 10 percent of data, hoping that pre-trained models will achieve good result because they were trained on ImageNet and should be robust on classify different objects. In table 5.

However, the results were not promising, particularly on the Cored plaque classification. With only 0.088 and 0.222 Precision-Recall AUC score, the model simply could correctly recognize the cored image. The confusion matrix, with 25 percent error rates for Cored, also agrees with this. With also low in the receiver operating characteristic score and accuracy, we reached the same result as Semi-Supervised Learning

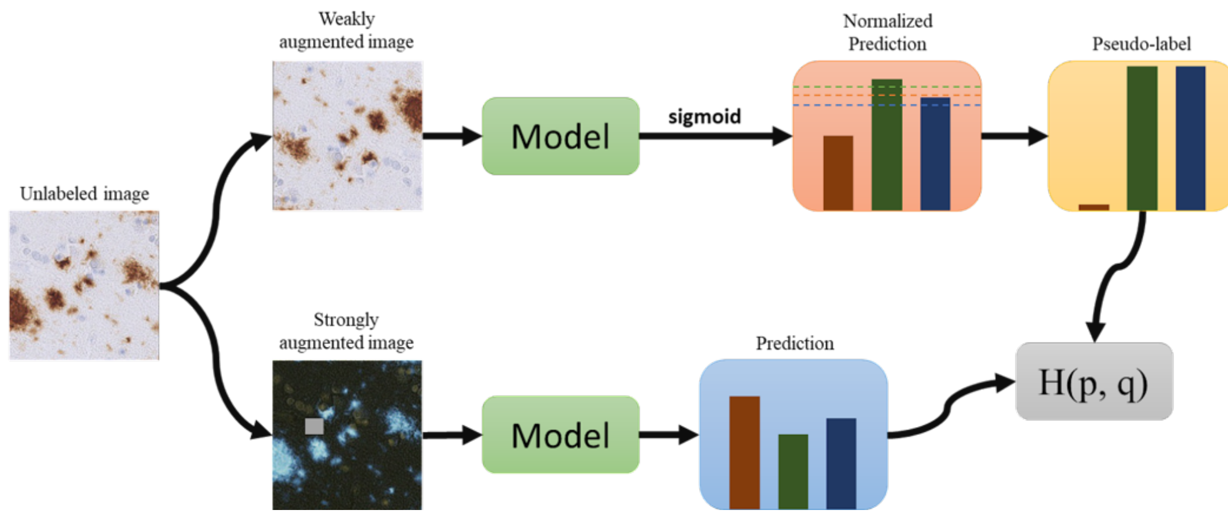

Fig. 3. Multi-label FixMatch algorithm

TABLE IV
RESULTS FOR DIFFERENT THRESHOLD(T) IN FIXMATCH

Cored t	Diffuse t	CAA t	Cored Acc	Diffuse Acc	CAA Acc	loss
0.9	0.95	0.85	0.0624	0.9639	0.9788	0.6281
0.8	0.95	0.85	0.0886	0.9639	0.9586	0.5799
0.7	0.95	0.75	0.1035	0.9640	0.9693	0.5794

TABLE V
RESULTS FOR DIFFERENT MODEL IN FIXMATCH

	ResNet18	wide Resnet	Tang et al.'s model
ROC Cored	0.546	0.589	0.981
ROC Diffuse	0.442	0.661	0.986
ROC CAA	0.276	0.928	1.000
R-P Cored	**0.047**	**0.058**	0.742
R-P Diffuse	0.927	0.963	0.999
R-P CAA	**0.024**	0.579	0.987

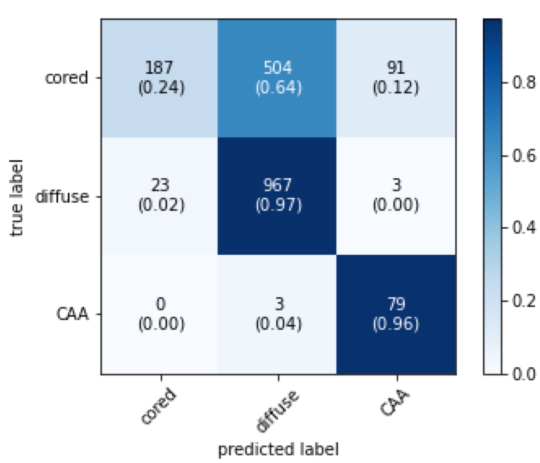

Fig. 4. Confusion Matrix for Transfer Learning

TABLE VI
RESULTS FOR TRANSFER LEARNING WITH 10 PERCENT OF DATA

	ResNet34	RegNet02X	Tang et al.'s model
ROC Cored	0.669	0.820	0.981
ROC Diffuse	0.675	0.987	0.986
ROC CAA	0.866	0.992	1.000
R-P Cored	**0.088**	**0.222**	0.742
R-P Diffuse	0.967	0.990	0.999
R-P CAA	**0.460**	0.894	0.987
Acc Cored	**0.538**	0.761	0.918
Acc Diffuse	0.674	0.750	0.947
Acc CAA	0.962	0.963	0.994

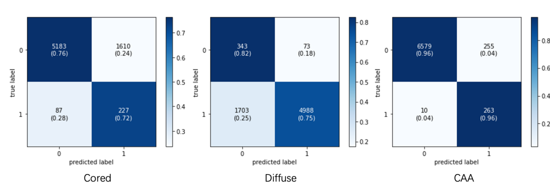

Fig. 5. Confusion Matrix for Transfer Learning

that Transfer Learning cannot provide enough information for model to identity Cored plaques. For Transfer Learning, we found one possible explanation for this: due to the "mismatch in learned features between the natural image, e.g., ImageNet, and medical images", medical imaging classification was proven to be ineffective [3]. Since the pre-trained models were focus on real-world objects such as animals, cars, plants, etc., the models do not have any knowledge about plaque images, which is extremely different from natural objects. Because of this, the pre-trained CNN layers cannot recognize the pattern of Cored plaque, resulting in poor classification result.

D. Training Phase IV: Supervised Learning With Limited Data

Finally, we planed to try the last method we can think of, which is using 10 percent data and fully re-train a Supervised models. This is very risky since with few data to learn, the models are likely to converge too fast and overfit. Also, it will not be robust because they did not cover too many varieties of plaques. The results are below in table 6.

TABLE VII
RESULTS FOR SUPERVISED LEARNING WITH 10 PERCENT OF DATA

	ResNet34	RegNet02X	Tang et al.'s model
ROC Cored	0.964	0.965	0.981
ROC Diffuse	0.951	0.947	0.986
ROC CAA	0.997	0.999	1.000
R-P Cored	**0.650**	**0.648**	0.742
R-P Diffuse	0.995	0.995	0.999
R-P CAA	0.953	0.973	0.987
Acc Cored	0.907	0.942	0.918
Acc Diffuse	0.935	0.964	0.947
Acc CAA	0.989	0.992	0.994

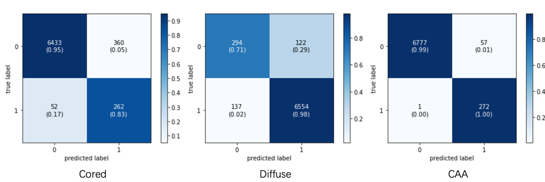

Fig. 6. Confusion Matrix for Supervised Learning

To our surprise, the models actually produced very good results. The recognition on Cored Plaque improved dramatically. The ROC and R-P score was comparable to the Author's model. The confusion matrix also shows that the classification accuracy is much better than transfer learning. We can conclude that Supervised Learning on classifying plaques is with achievable with limited data.

V. DISCUSSIONS AND FUTURE WORK

A. Training Phase II: Semi-Supervised Learning With FixMatch

One of the main factors that leads to poor performance during multi-label FixMatch training is the problem shown in the figure. This plaque image is marked as positive in both Cord and Diffuse categories. After weakly augmented, only Diffuse's confidence level exceeds the threshold, so that only Diffuse is positive in pseudo-labels and all others are negative. After completing the image prediction of Strong augment, the model makes the same judgment as weak augment, only Diffuse has a higher value of 0.88. This is exactly the same result set by the pseudo-label, so the loss is very low. However, the whole process did not label the correct label of the photo, but fell deeper and deeper into the wrong path. After that, the model made the same judgment for most of the photos that had both Diffuse and Cored, so the judgment of the cored of the whole model became worse and worse. Images with only cored tags may also change their judgment because of this, and consider them all as cored.

After going out with multi-label images, the model still did not train very well. We think this is because the graphics of Cored and Diffuse are too similar, as they themselves belong to a large class of patches. So these images are very difficult for machine models to recognize. In the process of Supervised training, we used 10 percent of the images to get better results, but in FixMatch we only used about 5 percent of the images for Supervised learning training (take FixMatch's training on CifAR10 as a reference set number).

B. Training Phase III: Transfer Learning With Limited Data

Transfer learning also doesn't give very good results. This is mainly because the models we are trying to transfer are all from ImageNet. Most of their pictures are from objects that are common in the real world, such as cars, airplanes, etc. The objects of these images are very different from our plaque in shape, color and type, so it is difficult for us to directly apply these models to the medical image data set. We found that there are few good plaque image models in the current research, so we believe that it is difficult to use existing models to make breakthroughs in new medical models in transfer learning.

C. Training Phase I and IV: Supervised Learning

We already knew that Supervised Learning will be extremely robust when a model is fed with a large amount of data. As we did in Phase I, with about 60, 000 images each class in training, the ResNet, RegNet, and the Tang et al.'s model all achieved excellent results. However, our goal for this senior design project is to explore the possibility of reducing the labeling effort while still maintaining model's high performance. With the Semi-Supervised Learning and Transfer Learning methods not worked as hoped, we took the risk to try use only 10 percent data to train Supervised model. The result was not bad at all, and even comparable to the model using all the images available. Our conclusion is that the state-of-the-art deep Supervised model is able to capture the characteristics of Cored, Diffuse, and CAA plagues effectively with small data sample. And in the end, we found that for AD plaque dataset, less human labor is needed for labeling, which accomplished our initial goal.

D. Future Works

If we have more time to continue working on our project, firstly, we can try more single label image effects in FixMatch, which may not be as good as the CIFAR data set, but can reduce human labeling effort to a certain extent. In the future, we can try to reduce human labeling effort through more means. For example we can use self-Supervised learning and Active learning. These methods have demonstrated excellent results on other data and they are able to train with less

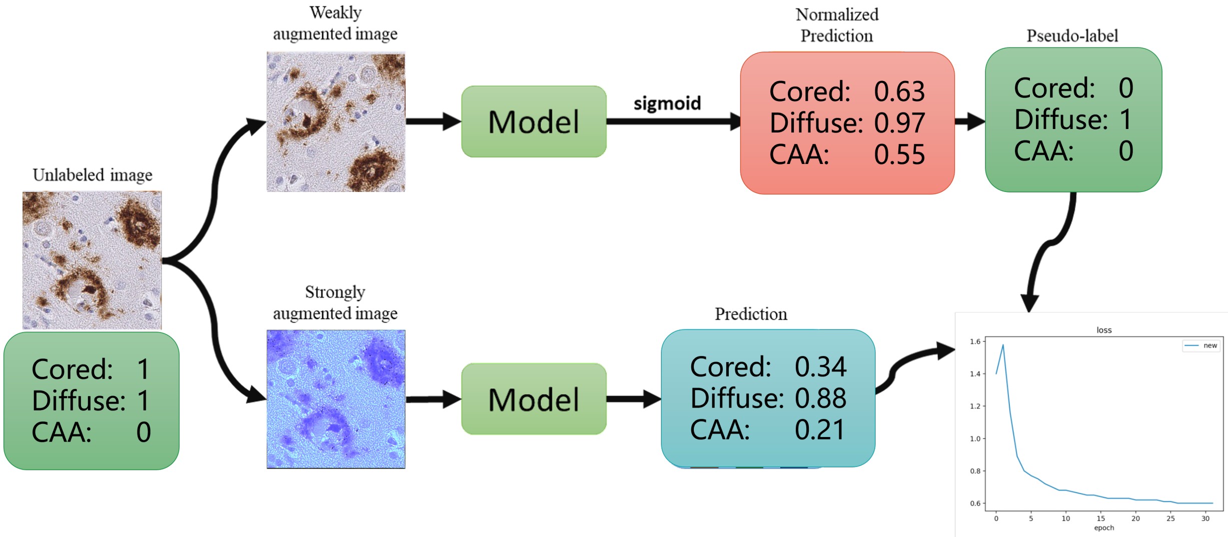

Fig. 7. Confusion Matrix for Transfer Learning

labeled images. At the same time, we can also introduce object detection into this field, so that we can achieve the effect of segmentation and classification at the same time. At the same time, Object Detection has also achieved certain results in Semi-Supervised learning, and most of their pictures also have the problem of multiple labels, so this is likely to be a new breakthrough direction.

VI. Acknowledgement

We would like to thank Professor Chen-Nee Chuah and TA ZhengFeng Lai for supporting us during our project. Without their suggestions and guidance in the weekly meetings and office hours, we might not have insight into our model's performance and methods to improve them. We also want to express our gratitude for providing powerful devices that made our training much easier.

We would also want to thank Tang and his team, PyTorch team and FrancescoSaverioZuppichini for providing pre-trained models and algorithms. Their work simplifies our set up and testing phases, making it possible for us to explore and compare different training methods.

References

[1] alz.org, "Facts and Figures," Alzheimer's Disease and Dementia. Accessed on: Mar. 13, 2022. [Online]. Available: https://www.alz.org/alzheimers-dementia/facts-figures.

[2] Z. Tang, K. V. Chuang, C. DeCarli, L.-W. Jin, L. Beckett, M. J. Keiser, and B. N. Dugger, "Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline," Nature Communications, vol. 10, no. 1, 2019.

[3] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and Y. Duan, "Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data," Cancers, vol. 13, no. 7, p. 1590, 2021.

Appendix

Pu was responsible for FixMatch implementation, and Zheng helped Pu tuning the FixMatch model. Zheng and Pu contributed equally to the Supervised Learning and Transfer Learning parts, as well as all the reports and presentation.