"Detection of Alzheimer's Disease from Brain MRI via Deep Learning" Comp 548 Medical Image Analysis, Spring 2023, Koc University Project Report

Pouya Pourakbarian Niaz [1 2] **Feras Kiki** [1 3] **Enes Ulas Dincer** [1 4]

Abstract

Alzheimer's Disease (AD) has become an important global health issue recently with the aging population. A mortal disease diminishing the life expectancy of the afflicted, AD severely affects the productivity of aging nations across the world. Consequently, it is of utmost importance to be able to detect it as early as possible. The present project uses deep learning techniques for end-to-end classification of brain MRI images into four classes, ranging from cognitively normal to moderately demented. Well-known Convolutional Neural Network architectures are trained from scratch, as well as being used in a transfer learning scenario, to find out how the best performance can be achieved.

1. Introduction

Alzheimer's disease (AD), an incurable neurodegenerative disease that can cause cognitive impairment and memory deterioration, afflicts millions of middle-aged and elderly people worldwide, decreasing their life expectancy to around 3-10 years (Chévez-Gutiérrez et al., 2012) as a result. AD is predicted to severely affect the development of nations with large elderly populations (Reitz & Mayeux, 2014). Countless researchers have proposed various classical or machine learning methods for detecting AD from medical images such as brain MRI.

Over time, image processing methods, classical machine learning models, and more recently deep learning models, have been proposed for the segmentation and classification of brain MRI images for AD diagnosis. Due to the emergence of very deep and high-performance convolutional neural networks in recent years, deep learning has become increasingly prevalent in the early detection and classification of AD. In some cases, such models are trained from scratch, and in some cases transfer learning methods are proposed. In such cases a model already trained on a large image classification dataset is taken, and its final classification layer is rebuilt, reinitialized, and retrained on the existing new dataset while keeping the rest of the model frozen and untouched (Buvaneswari & Gayathri, 2021; Ramzan et al., 2020; Ji et al., 2019).

In some studies, deep learning models are only used for image segmentation and feature extraction, before feeding the features to, e.g., an SVM or XGBoost classifier (Buvaneswari & Gayathri, 2021; Suh et al., 2020; Venugopalan et al., 2021). In other studies, the deep learning model is used in an end-to-end pipeline where the model is used for directly classifying the input image without a separate feature extraction module (Ramzan et al., 2020; Ji et al., 2019; Murugan et al., 2021).

In this project, an end-to-end pipeline is proposed wherein both training-from-scratch and transfer learning methods are explored and evaluated to perform the detection of AD using brain MRI images. In the former, two custom-designed CNN are used, while three famous pre-trained CNN models are used in the latter.

In this project, The Kaggle Alzheimer's Disease dataset is used. This dataset contains many brain MRI images, labeled into four classes: *not demented*, *very mildly demented*, *mildly demented*, and *moderately demented*. These are different stages of early AD divided into four classes. As a result, we are dealing with a 4-class image classification problem where inputs are brain MRI images, and outputs are class labels.

2. Training from Scratch

Here, we train a custom-designed CNN model from scratch without using transfer learning. The main challenge of training from scratch is that in order for a deep neural network

[1]Robotics and Mechatronics Laboratory & KUIS AI Center, Department of Mechanical Engineering, Koc University, Sariyer 34450, Istanbul, Turkey [2]ID: 0073978, pniaz20@ku.edu.tr [3]ID: 0066657, fkiki18@ku.edu.tr [4]ID: 0084519, edincer22@ku.edu.tr. Correspondence to: Pouya Pourakbarian Niaz <pniaz20@ku.edu.tr>.

Table 1. Training procedure hyperparameters

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	0.001
Patience	10
Maximum epochs	100
Minibatch size	64
Early stopping accuracy	0.99

to learn the mapping and relationship between the inputs and the outputs, a very large dataset is required. Otherwise, the model is bound to overfit (i.e., memorize) the training data, resulting in a subpar performance in unseen test data. Two custom-designed CNN architectures designated as "BasicNetwork" and "MoreDenseNetwork" in the file directory are proposed for this purpose. The architecture of the former can be seen in Fig. 1. In this figure, "sequential" refers to a CNN block containing two consecutive 2D convolution layers with 3×3 filters, *same* padding, and ReLU activation, followed by a batch normalization and a default max pooling layer. Also, "dense" in Fig. 1 refers to a fully connected block containing a dense layer with ReLU activation, followed by a batch normalization and a Dropout layer. The "MoreDenseNetwork" model is identical to the former, except that between the *sequential* block with 512 units and the next one with 128 units, there is also another *sequential* block with 128 units.

When training from scratch, the hyperparameter list shown in Table 1 is used for both models. The normalized test set confusion matrices for the *BasicNetwork* and *MoreDenseNetwork* can be seen in Fig. 2 and Fig. 3, respectively. Furthermore, classification reports for both training and test sets can be seen for both models in Table 2. This table contains class-balanced accuracy, micro-averaged F1 score, and class-specific precision and recall values for both models, for the training set and test set. An inspection of this table and the confusion matrices show slight overfitting due to some distance between the training set performance and test set performance, as well as some underfitting in the *MoreDenseNetwork* case, likely due to insufficient training or high learning rates.

3. Transfer Learning

One discernible obstacle to tackle when training a deep learning model from scratch is the size of the training data. When the number of data points in the training set is not quite on par with the number of features in every data point (pixels $\times$ channels) or the number of trainable parameters in the deep learning model (typically in the order of a few million or tens of millions), it can be difficult to train a model efficiently and reliably with good generalization performance.

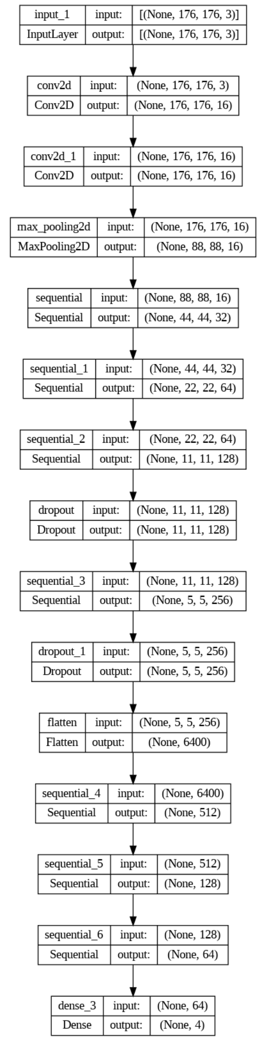

Figure 1. Architecture of "BasicNetwork", used when training from scratch

This can lead to overfitting, where the performance of the model over the training set is very good, but it is subpar or weak in the test set.

Transfer learning is a broad concept in the area of machine learning and deep learning, which allows us to utilize an existing large dataset (i.e., the source domain) where there are similar relationships to learn and similar features to extract and connect it to the existing limited dataset on which we want to train a model (i.e., the target domain). One way to perform transfer learning comes in the form of "domain adaptation", that is, learning transformations between source and target domains so data points in one feature space can be mapped to the other, allowing both source and target domain data to be expressed in a single common feature space thereby obviating the dataset size problem of the target domain. Another more frequent form of transfer learning is called "model adaptation", where

Table 2. Classification performance when training from scratch

Dataset	**Training Set**					**Testing Set**				
Metric	**Accuracy (%)**	**F1 Score**	**Class ID**	**Precision**	**Recall**	**Accuracy (%)**	**F1 Score**	**Class ID**	**Precision**	**Recall**
BasicNetwork	97.44	0.96	*Normal*	0.82	1.00	80.25	0.76	*Normal*	0.53	0.93
			Very Mild	0.75	1.00			*Very Mild*	0.26	0.83
			Mild	1.00	0.97			*Mild*	0.95	0.77
			Moderate	0.98	0.93			*Moderate*	0.75	0.68
MoreDenseNetwork	89.56	0.87	*Normal*	0.59	0.97	80.87	0.78	*Normal*	0.52	0.95
			Very Mild	0.87	0.94			*Very Mild*	0.33	0.83
			Mild	0.96	0.90			*Mild*	0.92	0.83
			Moderate	0.91	0.77			*Moderate*	0.83	0.63

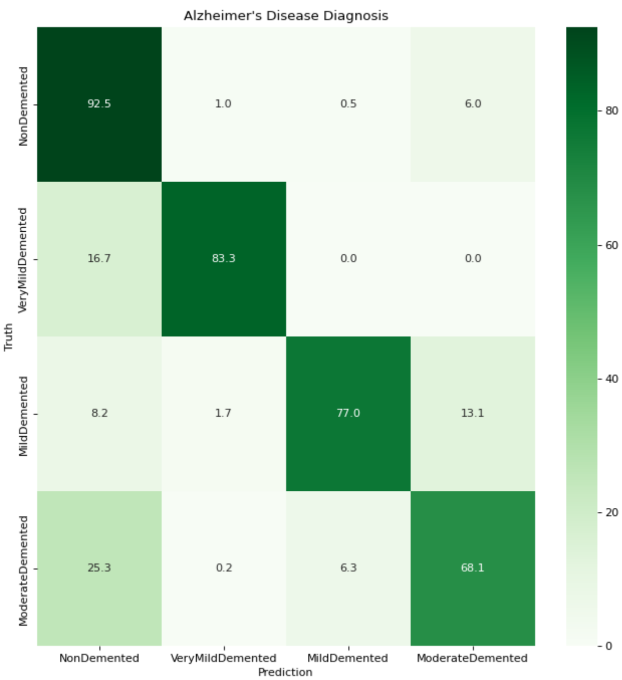

Figure 2. Normalized confusion matrix for the *BasicNetwork* model

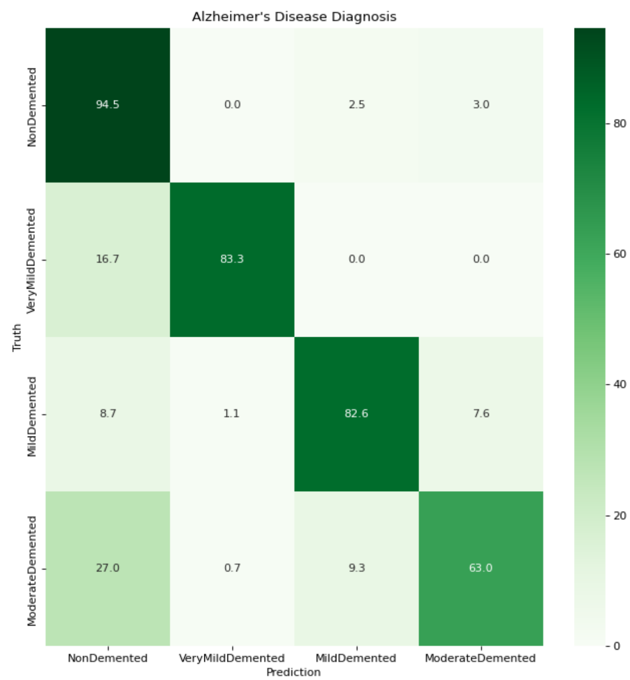

Figure 3. Normalized confusion matrix for the *MoreDenseNetwork* model

a deep learning model is trained on the large and diverse source domain, and then some of its layers (typically the last classifier layer or a few of the final fully connected layers) are retrained on the limited target domain data while keeping the rest of the layers frozen and untouched. Model adaptation is easier to implement and perform than domain adaptation, and is therefore much more frequent in the image classification realm.

It goes without saying that transfer learning only works well when the source domain and the target domain are sufficiently closely related. Similar features with similar joint distributions are required to be able to leverage the size and diversity of the source domain while extracting meaningful information from the target domain. If the high/low-level features and latent representations that can be extracted from the two domains are not sufficiently coincident, transfer learning might produce a model overfitting the source domain, and incapable of adapting itself to the limited data in the target domain. Transfer learning is typically done using well-known pre-trained models that are frequently used by researchers and practitioners. In this project, three models are used for this purpose: **VGG16**, **ResNet50**, and **InceptionV3**, all three of which are well-known image classification models trained on large and very diverse image datasets. In this project, transfer learning training is done using the same hyperparameters as the former case, shown in Table 1. Figures 4, 5, and 6 show the normalized confusion matrices for the VGG16, the InceptionV3, and the ResNet50 models, respectively.

The results shown in Table 2 and Figures 4, 5, and 6 show that the model that uses transfer learning has subpar results compared to the scenario when the CNN was trained from

Table 3. Classification performance when using transfer learning

Dataset	Training Set					Testing Set				
Metric	**Accuracy (%)**	**F1 Score**	**Class ID**	**Precision**	**Recall**	**Accuracy (%)**	**F1 Score**	**Class ID**	**Precision**	**Recall**
VGG16	71.36	0.65	*Normal*	0.42	0.95	69.46	0.58	*Normal*	0.42	0.92
			Very Mild	0.58	0.72			*Very Mild*	0.26	0.83
			Mild	0.95	0.67			*Mild*	0.92	0.54
			Moderate	0.56	0.52			*Moderate*	0.47	0.49
InceptionV3	97.80	0.97	*Normal*	1.00	0.99	62.15	0.60	*Normal*	0.72	0.67
			Very Mild	1.00	1.00			*Very Mild*	0.91	0.71
			Mild	1.00	0.92			*Mild*	0.70	0.39
			Moderate	0.93	1.00			*Moderate*	0.48	0.71
ResNet50	59.10	0.57	*Normal*	0.48	0.78	54.27	0.55	*Normal*	0.45	0.82
			Very Mild	0.93	0.63			*Very Mild*	0.78	0.47
			Mild	0.71	0.58			*Mild*	0.72	0.65
			Moderate	0.49	0.37			*Moderate*	0.41	0.23

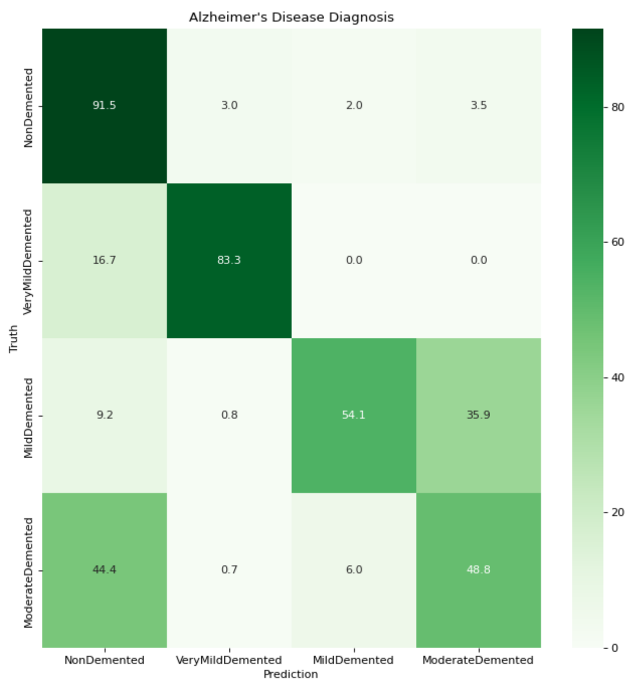

Figure 4. Normalized confusion matrix for the VGG16 model

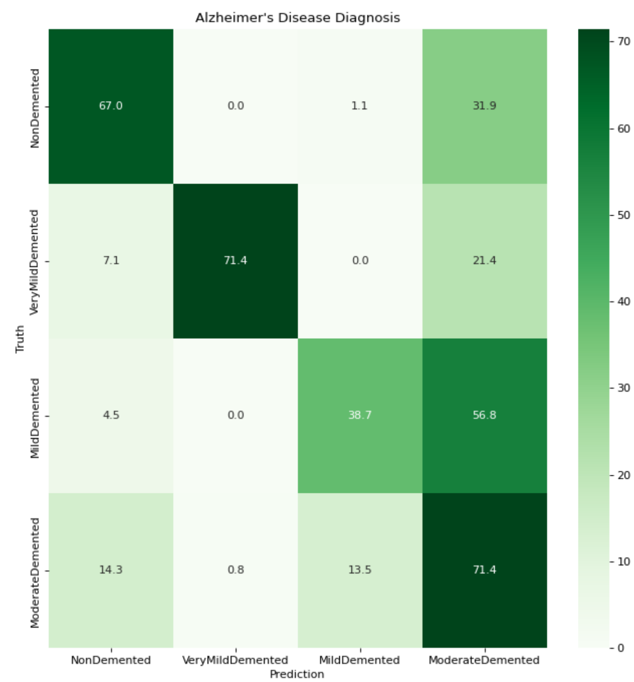

Figure 5. Normalized confusion matrix for the InceptionV3 model

scratch. There may be multiple reasons for this issue, some of which are explained below.

- **Overfitting**: By comparing training set performance and testing performance in Table 3 we can see that there is a lot of overfitting, meaning the generalization capability of the model is not on par with its learning capacity. To alleviate this, if possible, more data could be added to the training set. In addition, L2 regularization together with higher Dropout probabilities can help with overfitting. Also, more layers from the top of the model could be retrained, which would force more layers to adapt to the new dataset.

- **Data Mismatch**: As mentioned earlier, transfer learning only works when there is enough analogy and similarity between the source domain and the target domain in terms of features, and statistical distribution. If the brain MRI images are too different from the images that the base models were trained with, which is more than likely, it is possible that unseen new data causes a distribution mismatch between the source and target domains, leading to overfitting.

- **Data Quality**: The Kaggle Alzheimer Dataset is only one of a few AD brain MRI datasets. It is possible that the quality of the images in this dataset, along with the way they were labeled, is not sufficient for reliably

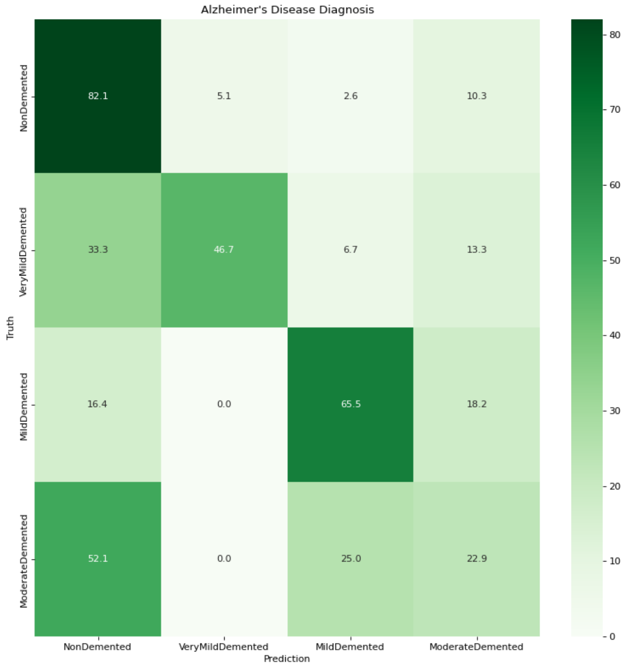

Figure 6. Normalized confusion matrix for the ResNet50 model

learning meaningful mappings from image data to class labels. By inspecting the confusion matrices we can see that one or two classes have the most confusion in the model, which can attest to this fact.

Depending on the circumstances, in cases where transfer learning methods fail due to the reasons mentioned above, a custom-designed network trained from scratch on the target domain data can achieve better performance than transfer learning, which is demonstrated in this study.

4. Conclusion

In this project, we attempt to classify brain MRI images into 4 categories representing 4 stages of AD-induced dementia. The Kaggle Alzheimer dataset is used, and two scenarios with deep learning are conducted. The first scenario involves training a custom-designed CNN from scratch on the AD dataset. The second scenario involves using a pre-trained image classification model in a transfer learning setting where the final classifier layer is retrained on the AD dataset and the rest of the layers are frozen. Our empirical results show that when training from scratch, performance is better than when using transfer learning, and the reason for that is not only the distinct differences between these brain MRI images and the generic images on which the base models were trained but also possibly a badly classified or badly labeled dataset, with not enough data to distinguish the classes from each other based on image data. With further hyperparameter tuning and better transfer learning methods, however, it is possible to improve model performance drastically.

References

Buvaneswari, P. R. and Gayathri, R. Deep learning-based segmentation in classification of alzheimer's disease. *Arabian Journal for Science and Engineering*, 46:5373–5383, 6 2021. ISSN 21914281. doi: 10.1007/S13369-020-05193-Z/TABLES/5. URL https://link.springer.com/article/10.1007/s13369-020-05193-z.

Chévez-Gutiérrez, L., Bammens, L., Benilova, I., Vandersteen, A., Benurwar, M., Borgers, M., Lismont, S., Zhou, L., Cleynenbreugel, S. V., Esselmann, H., Wiltfang, J., Serneels, L., Karran, E., Gijsen, H., Schymkowitz, J., Rousseau, F., Broersen, K., and Strooper, B. D. The mechanism of γ-secretase dysfunction in familial alzheimer disease. *The EMBO Journal*, 31:2261–2274, 5 2012. ISSN 1460-2075. doi: 10.1038/EMBOJ.2012.79.

Ji, H., Liu, Z., Yan, W. Q., and Klette, R. Early diagnosis of alzheimer's disease using deep learning. *ACM International Conference Proceeding Series*, pp. 87–91, 6 2019. doi: 10.1145/3341016.3341024. URL https://dl.acm.org/doi/10.1145/3341016.3341024.

Murugan, S., Venkatesan, C., Sumithra, M. G., Gao, X. Z., Elakkiya, B., Akila, M., and Manoharan, S. Demnet: A deep learning model for early diagnosis of alzheimer diseases and dementia from mr images. *IEEE Access*, 9:90319–90329, 2021. ISSN 21693536. doi: 10.1109/ACCESS.2021.3090474.

Ramzan, F., Khan, M. U. G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., and Mehmood, Z. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. *Journal of Medical Systems*, 44:1–16, 2 2020. ISSN 1573689X. doi: 10.1007/S10916-019-1475-2/FIGURES/10. URL https://link.springer.com/article/10.1007/s10916-019-1475-2.

Reitz, C. and Mayeux, R. Alzheimer disease: Epidemiology, diagnostic criteria, risk factors and biomarkers. *Biochemical Pharmacology*, 88(4):640–651, 2014. ISSN 0006-2952. doi: https://doi.org/10.1016/j.bcp.2013.12.024. URL https://www.sciencedirect.com/science/article/pii/S0006295213008083. Alzheimer's Disease – Amyloid, Tau and Beyond.

Suh, C. H., Shim, W. H., Kim, S. J., Roh, J. H., Lee, J. H., Kim, M. J., Park, S., Jung, W., Sung, J., and Jahng, G. H. Development and validation of a deep learning–based automatic brain segmentation and classification algorithm for alzheimer disease using 3d t1-weighted volumetric images. *American Journal of Neuroradiology*, 41:2227–2234, 12 2020. ISSN 0195-6108. doi: 10.3174/AJNR.A6848. URL https://www.ajnr.org/content/41/12/2227https://www.ajnr.org/content/41/12/2227.abstract.

Venugopalan, J., Tong, L., Hassanzadeh, H. R., and Wang, M. D. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific Reports 2021 11:1*, 11:1–13, 2 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-74399-w. URL https://www.nature.com/articles/s41598-020-74399-w.