MRI-based Diagnosis of Alzheimer's Disease using Deep Learning with CycleGAN for Data Augmentation

Sunny Wang

Abstract

Alzheimer's disease is a progressive disease causing deterioration of neurons in the brain, leading to dementia and eventually death. Diagnosis of Alzheimer's conventionally consists of a combination of neuropsychological tests and laboratory tests, and clinical diagnosis accuracy lies at around 77%, As Alzheimer's is associated with loss in brain mass, which can be discerned from MRI scans, it is a suitable task for deep learning and computer vision. An accurate and efficient machine learning model could be of great assistance to physicians as it could reinforce their diagnosis. However, deep learning typically requires large amounts of data, and medical data is often scarce. A recent breakthrough in machine learning, the generative adversarial network (GAN), allows for generation of realistic images, providing a potential solution to lack of data. In this study, we construct ResNet50-based convolutional neural networks to perform Alzheimer's disease classification using MRI scans, achieving an F-1 score of 89%. Furthermore, by generating samples using CycleGAN, we demonstrate that GANs can significantly improve classification accuracy when used for data augmentation, achieving an F-1 score of 95%.

1 Introduction

1.1 Background

Alzheimer's disease (AD) is a progressive disease characterized by the loss of cognitive ability and is the sixth leading cause of death in the United States. The progression can be categorized by severity, consisting of mild, moderate, and severe AD. These stages are typically classified and diagnosed based on a variety of factors, including cognitive tests, interviews with family members, and laboratory tests. Early diagnosis of AD is critical, as it can greatly improve patients' quality of life and in some cases halt or slow the rate of progression. According to a study conducted by Beach et. al. in 2012, there was wide variation in AD diagnosis accuracy, but the overall accuracy was 77% [1]. This is far from perfect, as many AD cases are misdiagnosed, with a low true negative rate. This raises the demand for a computer assisted tool to reinforce physicians' diagnosis.

As AD causes the breakdown and death of neurons, these changes in brain mass can be observed through technology such as magnetic resonance imaging (MRI), computerized tomography (CT), and positron emission tomography (PET). These scans are suitable for computer vision and deep learning algorithms, particularly in image classification. Recent advancements in machine learning, such as the convolutional neural network (CNN), have achieved results in classification that even outperform humans in some cases [2]. Diagnosis of AD using machine learning can serve as a powerful tool for physicians, supplying an additional metric for diagnosis.

Convolutional neural networks typically require large datasets to perform effectively. However, medical data is often scarce and limited in size. This is largely due to the high standards of consistency and organization required for medical data, and the cost and time required for data collection. For example, the ADNI dataset, one of the largest datasets created for Alzheimer's disease neuroimaging, only consisted of 800 individuals. As a result, the lack of data in medical imaging is a prominent obstacle preventing more widespread use of machine learning.

This raises a demand for data augmentation techniques to improve medical machine learning models. One recently introduced technique is the generative adversarial network (GAN) [3], one of the most influential milestones in machine learning. By having two neural networks, a generator and a discriminator, compete against each other, GANs achieve promising results in image generation, super resolution, and data augmentation, among many other applications. The lack of large amounts of medical data leads to significant potential for the use of GANs for data augmentation. Using a relatively small dataset, GANs can generate similar but original images, as opposed to image modifications used in classical data augmentation. In this study, we investigate the potential for using deep learning in Alzheimer's disease classification by creating a convolutional neural network model. We also test the feasibility of using GANs for data augmentation, specifically using the CycleGAN architecture.

1.2 Literature Review

1.2.1 Convolutional Neural Networks.

A convolutional neural network (CNN) is a type of neural network that specializes in dealing with pattern recognition in images. Input usually consists of a three-dimensional array containing width, height, and the pixel values in the image. A diagram of the architecture of a convolutional neural network is shown in Figure 1.

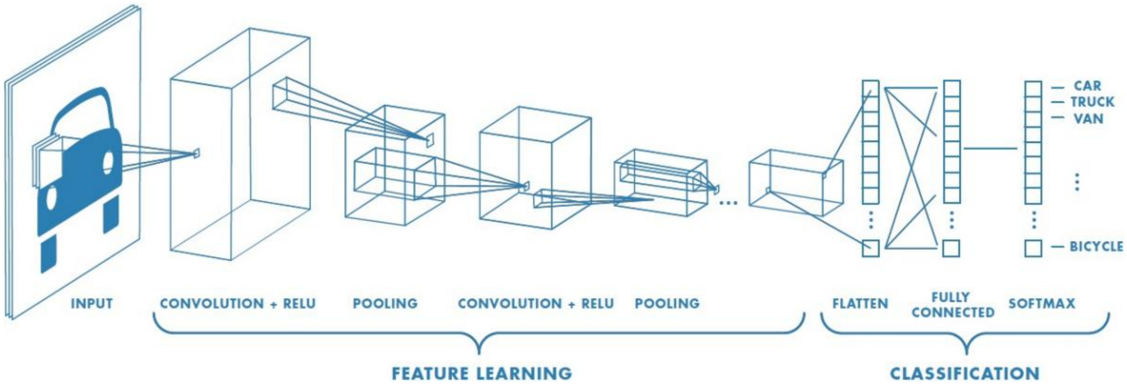

Fig. 1: CNN Architecture. Adapted from [4].

CNNs use layers of convolutions, where a filter or kernel is used to create a feature map, allowing the network to sample context within each frame as neighboring pixels are included. These feature maps allow the model to detect low level features within regions. Pooling layers are also added, which create subsamples of the previous layer. In classification tasks, a standard fully-connected feed forward neural network is commonly applied to the flattened final feature map output by the convolutional layers.

1.2.2 Transfer Learning.

A common method of training image classification networks is through transfer learning, which is the process of using pretrained models to operate on a different task, essentially transferring the knowledge stored in the original network. Common pretrained models include VGG, Inception, or ResNet, all of which have performed well on datasets such as ImageNet [5]. Transfer learning can often allow for quicker training times along with high accuracy, and is also less prone to overfitting.

1.2.3 Generative Adversarial Networks.

Generative adversarial networks were first proposed in 2014 by Ian Goodfellow [3]. The paper suggests the simultaneous training of two adversarial networks in the style of a zero-sum minimax game. The goal of the generator is to create images that trick the discriminator into classifying them as real, and does so by taking in random noise and upsampling it through convolutional layers. In contrast, the goal of the discriminator is to correctly classify what is fake data from the generator and what is real data from the training set. This is also usually done through a convolutional neural network, returning a probability that the image is real.

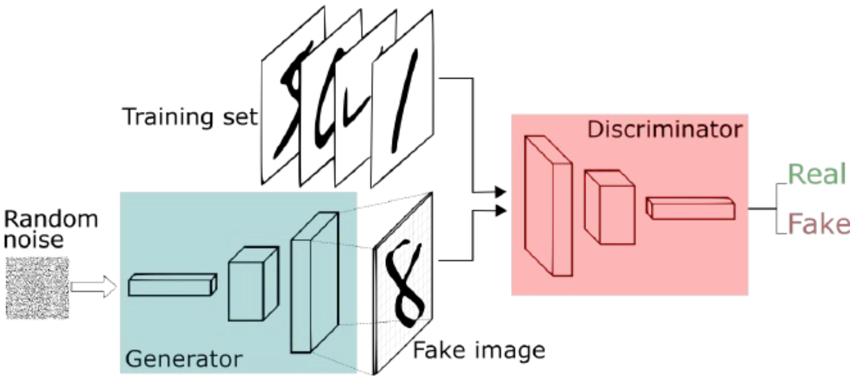

Fig. 2: GAN Architecture. Adapted from [6].

The GAN uses the following loss function, where x represents real samples and z represents generated samples.

$$Loss = E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))]$$

The discriminator attempts to maximize D(x), which is the likelihood of a correct sample being classified correctly. The generator wants D(G(z)), the likelihood of the discriminator classifying its generated image as real, to be as high as possible, and thus wants to minimize 1-D(G(z)). This means that the desired G will be minimized and the desired D will be maximized. During each training iteration, the discriminator is updated through gradient ascent in order to maximize the loss function, and the generator is then updated through gradient descent.

1.2.4 CycleGAN

Zhu et al. [7] proposed a cycle consistent GAN network that allows for unpaired image to image translation.

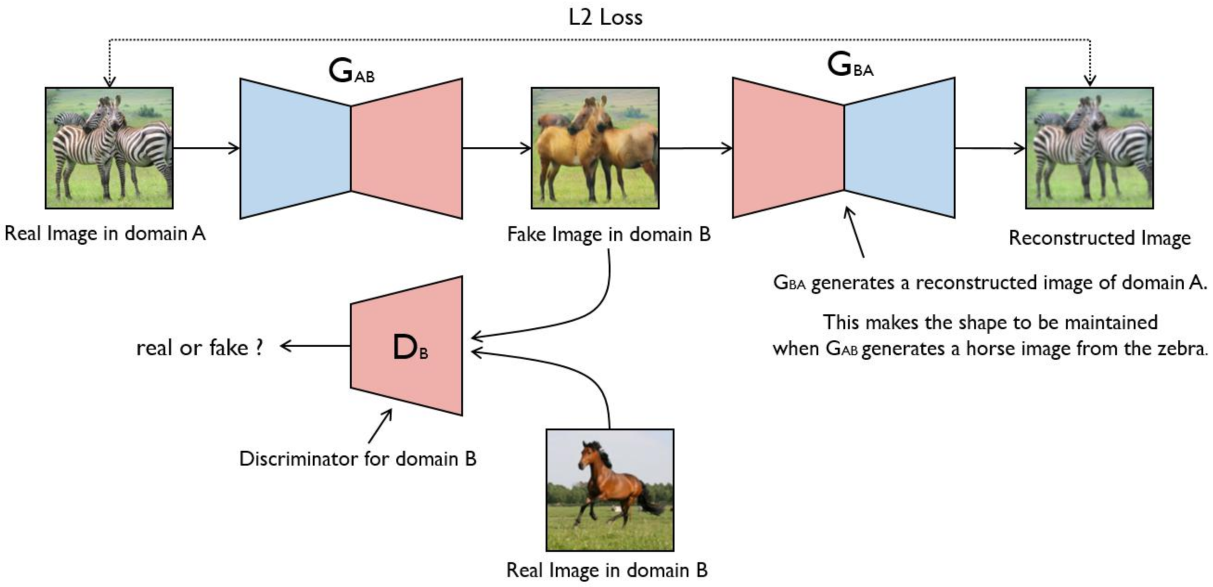
L2 Loss
G_AB
G_BA
Real Image in domain A
Fake Image in domain B
Reconstructed Image
GBA generates a reconstructed image of domain A.
This makes the shape to be maintained when GAB generates a horse image from the zebra.
real or fake ?
D_B
Discriminator for domain B
Real Image in domain B

Fig. 3: CycleGAN architecture. Adapted from [8].

Consisting of two generators and two discriminators, CycleGAN allows for discriminatory verification in both directions. Standard GAN loss is used to determine the realisticness of the generated images by comparing the generated image with a real image in the other domain. To ensure that the transformations are cycle consistent, cycle consistency loss is also introduced. When an image that is passed through both generators, the resulting image is compared with the original image, as shown in Figure 4. Using the loss functions together essentially allows the generators to learn a spatial transformation from one class to the other.

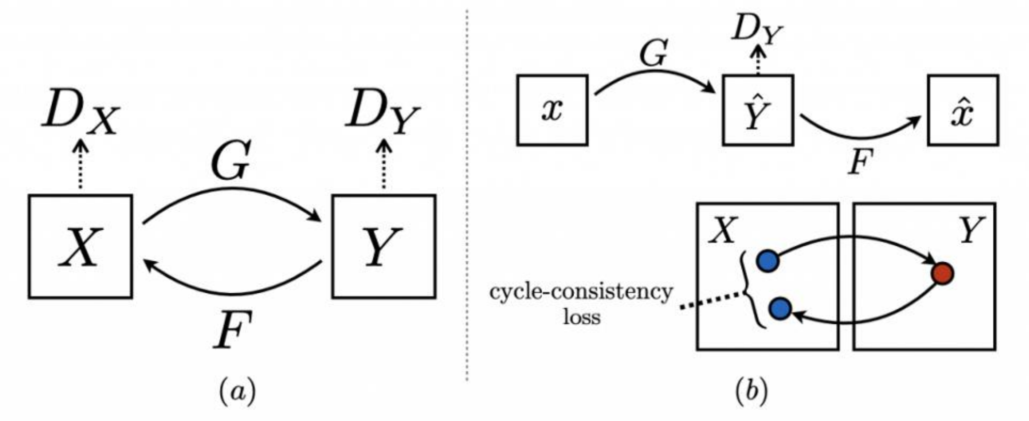

Fig. 4: CycleGAN Loss. Adapted from [9].

The generators G and F learn the spatial mappings from X to Y and vice versa, and are verified by discriminators D_Y and D_X. CycleGAN has been applied to topics such as style transfer and object transfiguration and produces impressive realistic results.

1.2.5 Data Augmentation

As stated before, GANs have great potential in data augmentation. Traditional data warping augmentation consists of techniques such as geometric transformations, filters, and random erasing [10]. This can reduce overfitting and improve accuracy. In contrast, GANs are a method of oversampling, as they are able to extrapolate beyond the training set to generate synthetic data, rather than solely modifying existing data within the training set. This is substantially useful in fields lacking large amounts of data, such as in medicine. One instance of GANs being used in augmentation is a study conducted by Frid-Adar et al. proposing the use of synthetic images in liver lesion classification [11]. GAN generated data was used with a CNN for image classification on a dataset of size 182. Standard DA techniques resulted in 78.6% sensitivity and 88.4% specificity, while the addition of GAN synthesized data yielded 85.7% sensitivity and 92.4%.

1.2.6 Alzheimer's and Machine Learning

As MRI scans are an important aspect of the diagnosis of Alzheimer's, there exists substantial literature regarding machine learning methods. The most common model is a convolutional neural network, as they are ideal for image processing. Farooq et al. [12] used a four way CNN classifier between the following classes: normal cognition (NC), early mild cognitive impairment (MCI), late mild cognitive impairment, and Alzheimer's disease (AD). The model was trained on the Alzheimer's disease Neuroimaging Initiative (ADNI) dataset. The only data augmentation performed was flipping images due to the symmetrical nature of MRIs, and specific slices were selected to exclude those without gray matter information. The proposed network based on GoogLeNet and ResNet outperformed other studies done on the same dataset, with about 98% accuracy. Hosseini-Asl et al. [13] used a deeply adaptive 3D convolutional neural network (DSA 3D-CNN) on the CADDementia dataset, which could then be generalized to the ADNI dataset, which they used for validation. The model achieved a 94.8% accuracy in task specific classification. Glotzman et al.

[14] proposed a network of 2D CNNs, applied to each of three images extracted from each sample in the ADNI dataset. This allowed for the use of two dimensional CNNs on a three dimensional dataset while preserving features. Both MRI scans and PET scans were used, and both two way classification (NC vs AD) and three way classification (NC vs MCI vs AD) were tested. Two way classification with PET-AV 45 scans performed the best, with 83% accuracy.

There have been relatively few applications of generative adversarial networks in classifying AD. A study by Bowles et al. [15] used GANs to model the progression of the disease using MRI data. Using image arithmetic, the model could predict changes in the brain over time and results were comparable to longitudinal examination data. Another study conducted by Pan et al. [16] used GANs to synthesize PET scans from MRI scans in order to fill in gaps in data as many AD patients do not have both due to the high cost of PET scans. A cycle consistent generative adversarial network was used as the first step, which was used in order to learn mappings between the two image domains. The features are then fed into a landmark based multimodal multi-instance learning classifier for diagnosis. Kim et al. [17] conducted a feasibility study on using GANs for slice selective learning on PET scans. Using a BEGAN, they showed that double slices over the posterior cingulate cortex achieves the best performance, and that two slices performed significantly better than using one slice.

It is also important to consider the practical applications of machine learning in AD diagnosis. A survey conducted by R. Bryan [18] on applying machine learning to AD diagnosis notes that machine learning models are biased by the original population, as it is limited to the dataset that it was trained upon. The overall conclusion reached was that although it is unlikely that machine learning models can replace the skills of radiologists, they can serve as useful tools to complement human skills.

The remainder of this paper is organized as follows. Section 2 describes the methodology and pipeline for model construction. Section 3 presents and discusses the results obtained from using different model architectures and displays the significance of using GAN augmentation. Section 4 and 5 report conclusions and future work respectively.

2 Methodology

Figure 5 outlines the model pipeline, consisting of dataset acquisition, preprocessing (including GAN augmentation), and classification using a CNN.

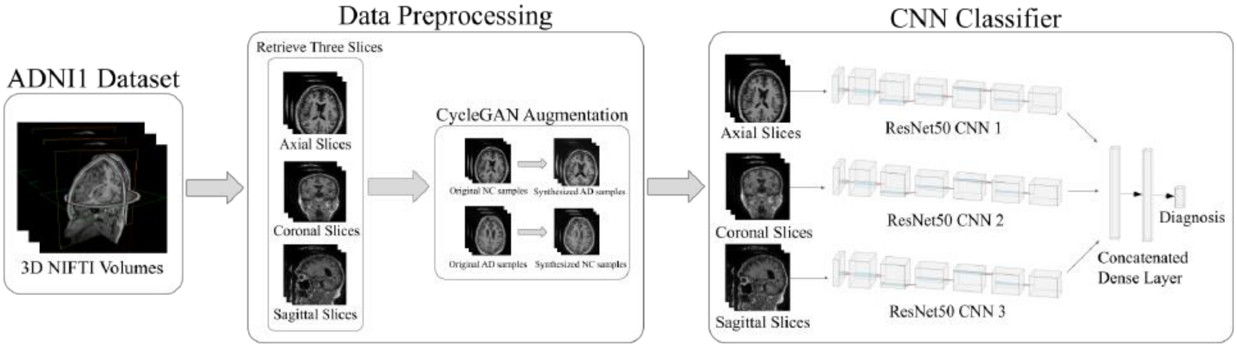

Fig. 5: Overview of the model pipeline.

2.1 Data Acquisition

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The specific dataset used for training was the ADNI1 standardized MRI dataset, categorized by severity: AD, MCI, and normal cognition (NC). For the purposes of this study, we trained a network to classify between AD and NC. The dataset contained 705 samples labeled as NC and 476 samples labeled as AD.

2.2 Data Preprocessing

The data was stored in NIFTI files, which were converted into three dimensional numpy arrays using nibabel. A csv file containing information about the scans, patients, and ground truth diagnosis labels was also downloaded from the ADNI database. As the NIFTI image data is three dimensional, slicing was required to prepare samples for training. In order to capture as much information from the original image, we extract three slices, one each from the axial, coronal, and sagittal orientations. The slices are taken by retrieving the midpoint of each axis length, and were resized to 224 x 224.

Two different methods of preprocessing were considered and tested:

1. Skull stripping was applied, which is the process of removing the skull from the MRI images. This isolates the brain tissue, allowing for more consistency among samples. This was done using the Extractor function from the deepbrain library.
2. RAS + ISO transforms and histogram normalization were performed using the TorchIO library [19]. These transforms change the orientation of the MRI image to also improve consistency.

2.3 GAN based Data Augmentation

We constructed the CycleGAN models using the implementation from [7]. The model architecture is shown in Figure 6.

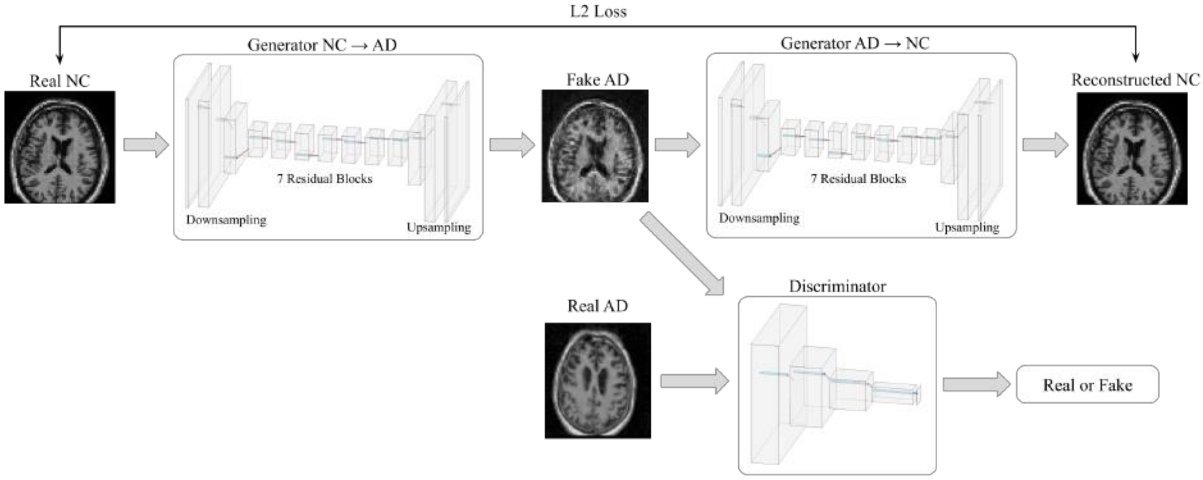

Fig. 6: CycleGAN Architecture.

The model consists of two generators, where one is trained to convert NC samples to AD samples and the other is trained to convert AD samples to NC samples. During training, a real NC image is passed through the first generator, and the resulting fake AD image is compared with an separate real AD image through the discriminator, computing GAN loss. The loss equation, as displayed in equation (1), is the same as the one proposed in the original GAN paper [3]. G represents the generator, D_Y represents the discriminator for AD samples, x represents a real NC image, and y represents a real AD image.

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y\sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D_Y(G(x)))]$$

This process is repeated, starting with a real AD image, resulting in another GAN loss, represented in equation (2) F represents the generator, D_X represents the discriminator for NC samples, x represents a real NC image, and y represents a real AD image.

$$\mathcal{L}_{GAN}(F, D_X, Y, X) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(F(y))]$$

The fake images are also passed through the second generator, returning a reconstructed version of the original image. Cycle consistency loss is computed by summing the losses from comparing the original NC image x with its reconstructed image $F(G(x))$ and comparing the original AD image y its reconstructed image $G(F(y))$. This is represented in equation (3), adapted from [7].

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[||F(G(x) - x||_1] + \mathbb{E}_{y \sim p_{data}(y)}[||G(F(y) - y||_1]$$

The overall loss function incorporates both GAN losses and the cycle consistency loss, and is represented in equation (4). λ is a constant representing how much weight is placed on the cycle consistency loss, and $\lambda = 10$ is used as described in the paper [7]. The objective of this loss function is to minimize G and F, which represent loss for the two generators, and maximize D_x and D_y, which are the two discriminators.

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda\mathcal{L}_{cyc}(G, F) \tag{4}$$

The generator is based on the ResNet architecture, and consists of downsampling, 9 residual blocks, and upsampling. Instance normalization and reflection padding is used as described in [7]. The tanh activation function is used in its last layer to scale the output image between -1 and 1. The generator architecture is shown in Figure 7.

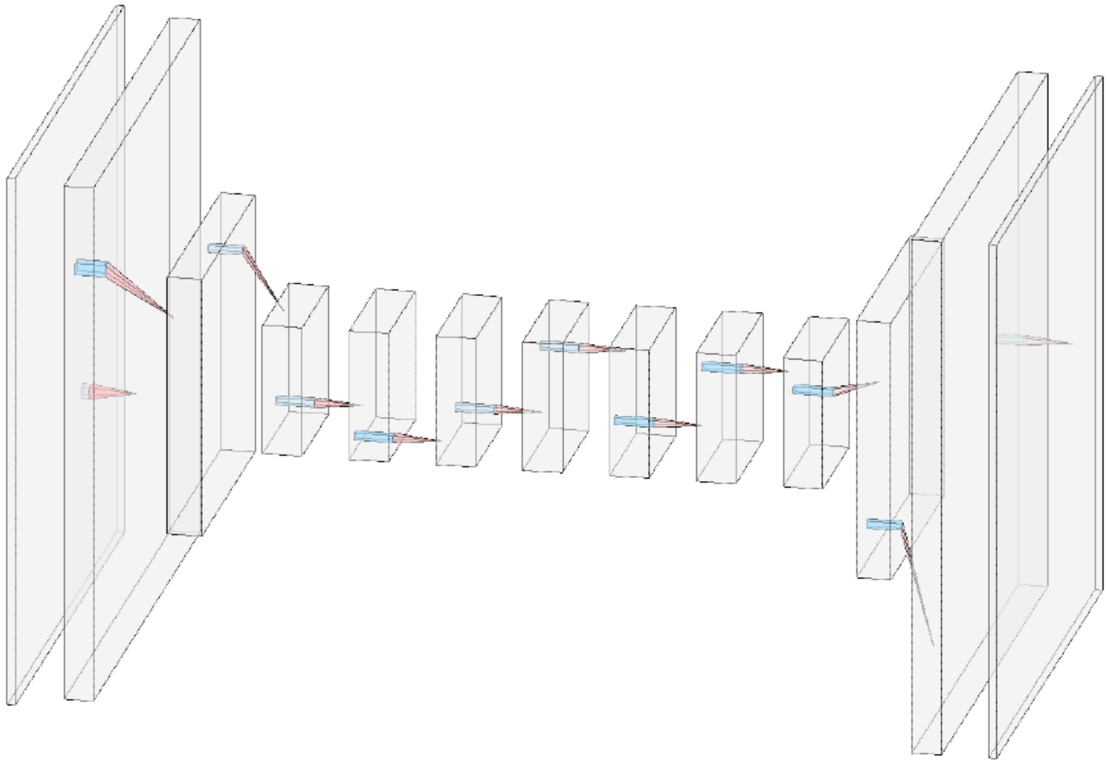

Fig. 7: CycleGAN Generator Architecture.

The discriminator is a CNN using PatchGANs, which classify whether an image is real or fake based on patches. This decreases the amount of parameters needed and is effective for images with high resolutions. The model also uses LeakyReLU as its activation function and utilizes instance normalization. The discriminator architecture is shown in Figure 8.

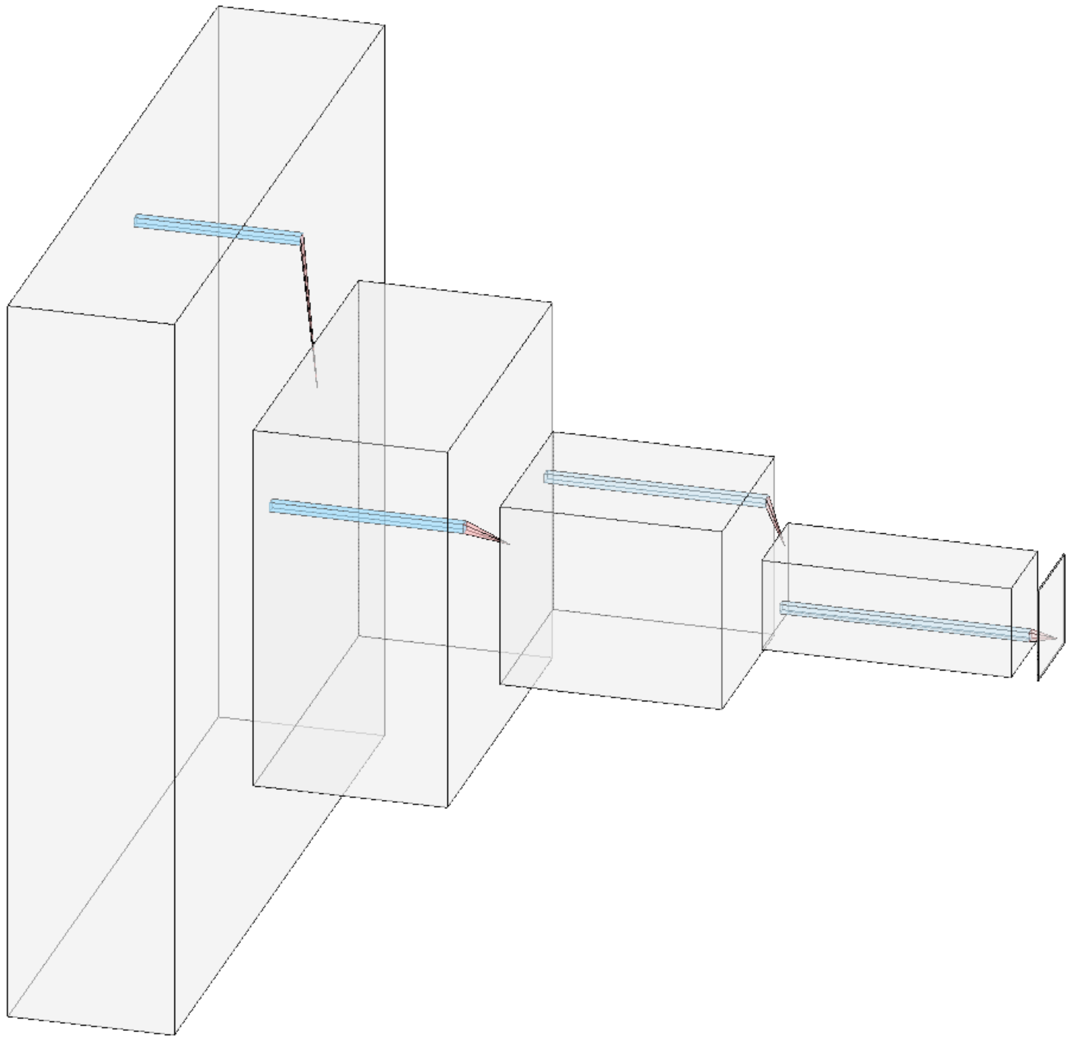

Fig. 8: CycleGAN Discriminator Architecture.

The original dataset was split according to the label and randomly paired. Three individual CycleGAN models were created, where each one was trained on data from a different MRI slice. Each model used the Adam optimizer with a learning rate of 2e-4, and were trained for 100 epochs with a batch size of 1, as specified in the CycleGAN paper.

The trained model was used to generate sufficient samples to create a balanced dataset. An AD version of each NC sample was generated and vice versa. A total of 705 AD samples and 476 NC samples of each orientation were generated, for a total of 1181 images of each class, as shown in Table 1.

	Normal Samples	Alzheimer's Samples	Total
Dataset	705	476	1181
GAN generated	476	705	1181
Total	1181	1181	2362

Tab. 1: Dataset sizes after GAN augmentation.

2.4 Convolutional Neural Network Classifier

We used a transfer learning approach to create the model architecture as it would save training time and is generally effective when datasets are small. We used the ResNet50 convolutional neural network (CNN) as our pretrained model, as shown in Figure 9.

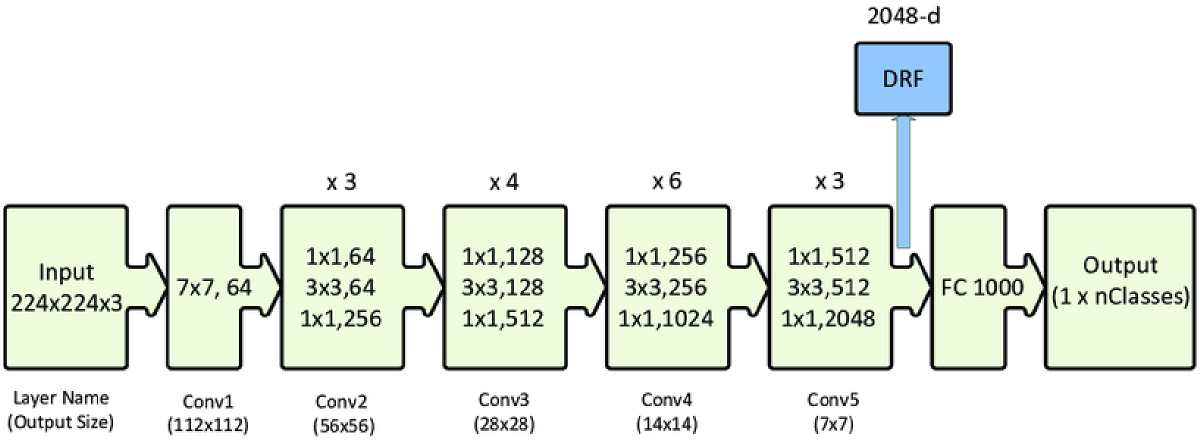

Fig. 9: ResNet50 Architecture. Adapted from [20].

The pretrained ResNet50 architecture takes in 3-channel RGB images while the MRI scans are grayscale. To match the network, the one channel images were transformed to three channels by stacking the tensor three times across dimension 0. The last layer was also modified to become a binary classifier. We used a modified CNN with multiple inputs in order to better encapsulate volumetric data. The model architecture consists of three ResNet50 CNN models, where outputs from each individual CNN are concatenated and passed through fully connected layers, which returns the diagnosis group. The model architecture is shown in Figure 10.

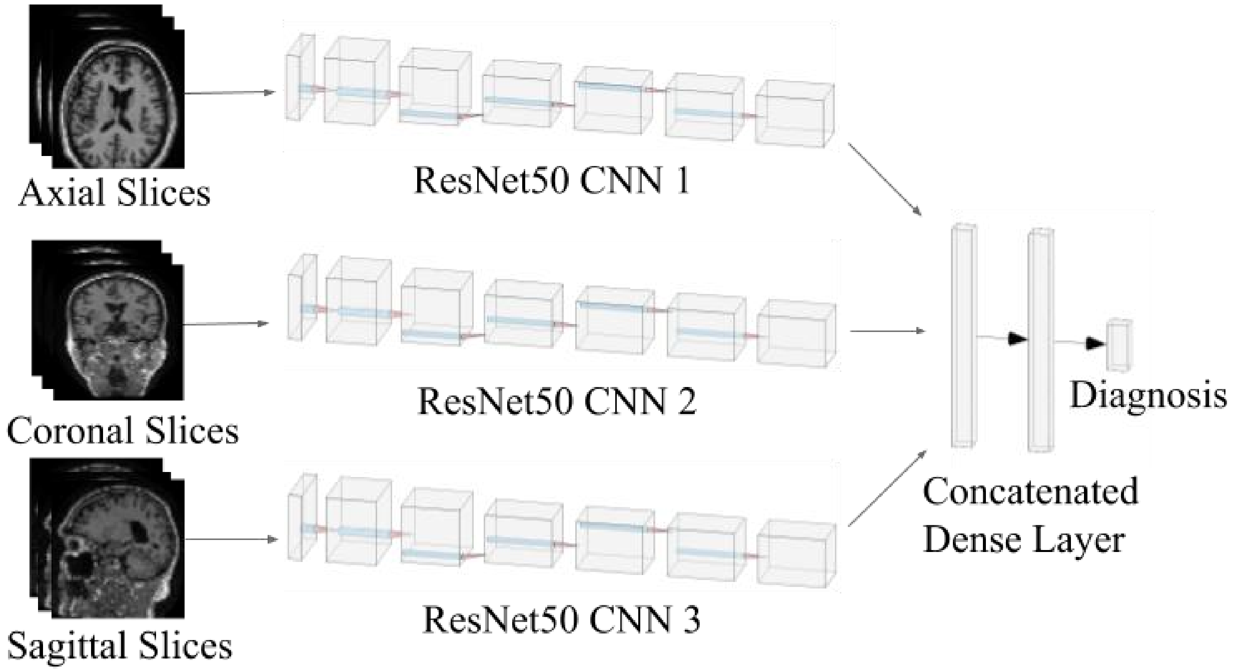

Fig. 10: Proposed Multiple CNN architecture.

The neural network was fine-tuned using the Adam optimizer with a learning rate of 1 e-4 and trained for 50 epochs with a batch size of 32. A training, validation, and testing split of 80%-10%-10% was used.

2.5 Model Implementation

All discussed networks were implemented using Python 3.7 with the PyTorch library. All training was done on a personal computer with an AMD 3700X CPU and an NVIDIA RTX 2070 GPU.

2.6 Model Evaluation

CNN models were evaluated using accuracy, precision, recall, and F1 score, with F1 score being the primary indicator for classification performance. The metrics are detailed in equations (5) - (8), where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$PRE = \frac{TP}{TP + FP} \tag{6}$$

$$REC = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{PRE \times REC}{PRE + REC} \tag{8}$$

3 Results and Discussion

3.1 Comparison of Preprocessing Methods

Two methods of preprocessing were tested, skull stripping and TorchIO transforms. Figure 11 displays the resulting images after applying preprocessing.

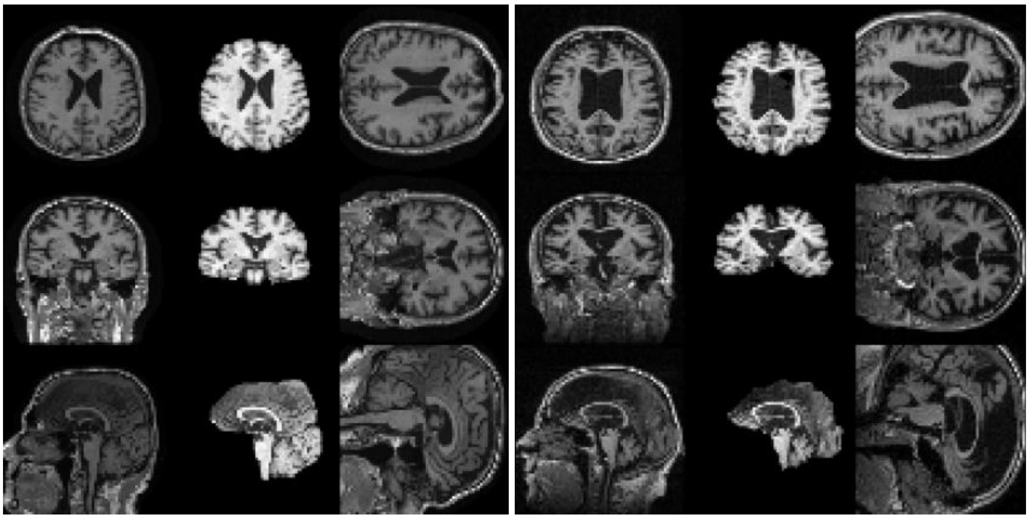

Fig. 11: Normal sample (left) and Alzheimer's sample (right). The original slices are displayed in the first column, the skull stripped samples are displayed in the second, and the TorchIO transformed samples are displayed in the third.

Table 2 compares the results from applying different methods of preprocessing on the three input ResNet50 network as shown in Figure 6. The different transforms were not compatible with each other, so results were obtained separately.

Metric	ResNet50	ResNet50 + Skull Stripping	ResNet50 + TorchIO
Accuracy	0.891	**0.908**	0.907
Precision	**0.932**	0.882	0.897
Recall	0.804	**0.900**	0.854
F1 Score	0.863	**0.891**	0.875

Tab. 2: Comparison of ResNet50 networks with different preprocessing.

The model utilizing TorchIO transforms improved upon the unmodified model, increasing the F1 score from 86.3% to 87.5%. However, the model utilizing skull stripping outperformed both models. This is likely because it improves the consistency among samples in the dataset, which makes it easier for the model to extract important features. The model with skull stripping was kept for the remainder of the study.

3.2 CycleGAN Generation Results

Examples of CycleGAN generated images are shown in Figure 12.

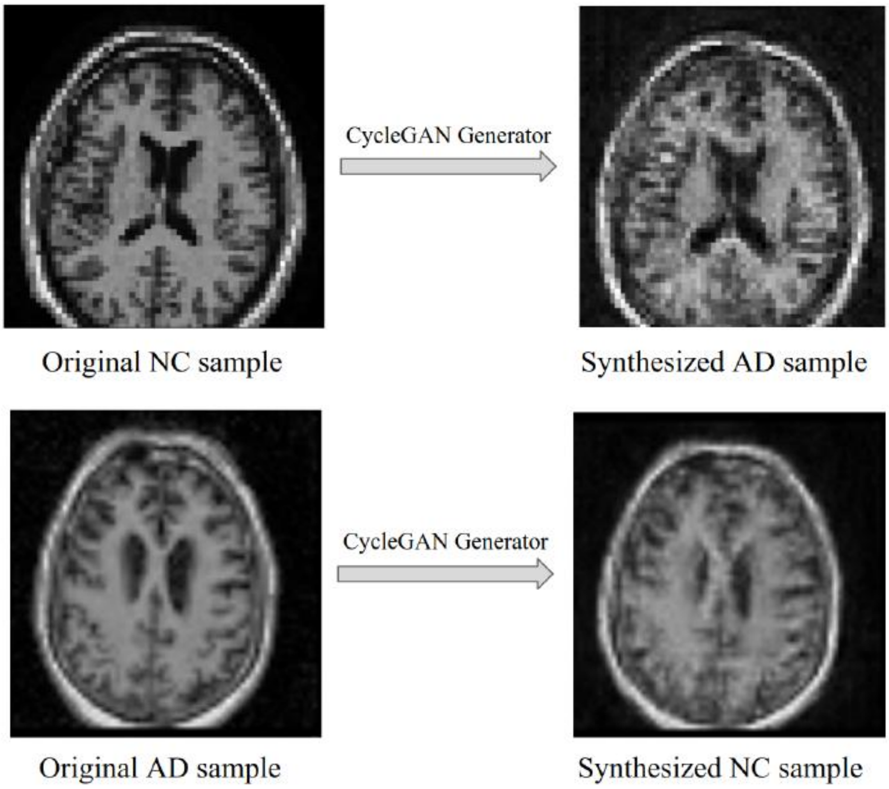

Fig. 12: CycleGAN image synthesis.

By observation, the synthesized Alzheimer's sample displays more dark space throughout the brain when compared to the normal sample that it was transformed from, which is an indication of loss of brain mass and a characteristic of Alzheimer's disease. On the contrary, the synthesized normal sample on the bottom right has much less dark space. While the quality of our synthetic images has not been verified by experts, they exhibit many characteristics of real MRI images.

3.3 Comparison with GAN Augmentation

When using CycleGAN for augmentation, an additional 705 AD samples and 476 NC samples of each orientation were generated, for a total of 1181 images of each class. Table 3 shows that there was a substantial increase in performance to the CNN model when GAN augmentation is applied. The F1 score for the ResNet50 model increased from 0.863 to 0.946, an 9.6% increase. The F1 score for the ResNet50 using skull stripping increased from 0.891 to 0.951, an 6.7% increase.

Metric	ResNet50	ResNet50 + GAN	ResNet50 + SS	ResNet50 + SS + GAN
Accuracy	0.891	**0.954**	0.908	0.949
Precision	0.932	**0.951**	0.882	0.944
Recall	0.804	0.942	0.900	**0.959**
F1 Score	0.863	0.946	0.891	**0.951**

Tab. 3: Comparison of CNN models with GAN augmentation.

These results indicate that the addition of CycleGAN improves CNN classification performance. From this, it is reasonable to infer that the synthesized images had meaningful features that benefited the model.

The increased size and balance among classes in the CycleGAN augmented dataset are also factors that are potentially responsible for the increase in performance. Overall, these results demonstrate the effectiveness of GANs in data augmentation.

4 Conclusions

In this study, we constructed convolutional neural network models utilizing the ResNet50 architecture to diagnose Alzheimer's disease using MRI scans, with variants using one input and three inputs. We also address the problem of size limitations in medical datasets with the use of generative adversarial networks (GANs). Using the ADNI1 dataset, we demonstrated that the addition of GANs can greatly improve deep learning classification accuracy for Alzheimer's disease diagnosis. Specifically, we used CycleGAN to generate images of one class using the other, balancing the dataset and increasing its overall size. Our results show that classification accuracy improved substantially, with F1 scores increasing from 0.863 to 0.946 for the standard model and 0.891 to 0.951 for the model utilizing skull stripping. Due to the lack of large datasets in many medical fields, the results obtained in this study can be generalized to many other fields as well. Overall, with promising results in data augmentation, GANs have potential to significantly improve upon classification tasks across a wide variety of applications.

5 Future Work

5.1 Preprocessing

Upon inspection, the current methods of preprocessing yield some inconsistency between images, particularly in axial scans, due to subjects having slight variation in brain shape. In the future, other methods of preprocessing could be used to improve consistency across each slice taken. Software such as statistical parametric mapping could also be applied to perform preprocessing such as gray matter segmentation and modulation to reduce variation between images.

5.2 Intermediate Stages

The current model serves to classify between MRI scans that either have or don't have Alzheimer's disease. The intermediate stage, mild cognitive impairment (MCI), could also be incorporated into the classifier. Augmented CycleGAN could be implemented to generate images of all three classes.

5.3 MRI and PET Fusion

Diagnosis of Alzheimer's Disease often uses both MRI scans and PET scans. In contrast to MRIs, which provide identification of abnormalities in the brain, PET scans can show areas of low metabolism, allowing for differentiation between Alzheimer's and other types of dementia. Using a fused image with both scans in conjunction would contain more features and will likely improve classification accuracy.

5.4 External features

A full Alzheimer's disease diagnosis requires many elements other than neuroimaging, such as mental status tests and physical exams. Incorporating all of these elements would greatly improve the accuracy of a machine learning system.

References

[1] Thomas G Beach, Sarah E Monsell, Leslie E Phillips, and Walter Kukull. Accuracy of the clinical diagnosis of alzheimer disease at national institute on aging alzheimer disease centers, 2005–2010. *Journal of neuropathology and experimental neurology*, 71(4):266–273, 2012.

[2] Christopher Bowles, Roger Gunn, Alexander Hammers, and Daniel Rueckert. Modelling the progression of alzheimer's disease in mri using generative adversarial networks. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741K. International Society for Optics and Photonics, 2018.

[3] R Nick Bryan. Machine learning applied to alzheimer disease, 2016.

[4] Antoine Buetti-Dinh, Vanni Galli, Sören Bellenberg, Olga Ilie, Malte Herold, Stephan Christel, Mariia Boretska, Igor V Pivkin, Paul Wilmes, Wolfgang Sand, et al. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports*, 22:e00321, 2019.

[5] Ammarah Farooq, SyedMuhammad Anwar, Muhammad Awais, and Saad Rehman. A deep cnn based multi-class classification of alzheimer's disease using mri. In *2017 IEEE International Conference on Imaging systems and techniques (IST)*, pages 1–6. IEEE, 2017.

[6] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[7] Tanya Glozman and Orly Liba. Hidden cues: Deep learning for alzheimer's disease classification cs331b project final report, 2016.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Ehsan Hosseini-Asl, Robert Keynton, and Ayman El-Baz. Alzheimer's disease diagnostics by adaptation of 3d convolutional network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 126–130. IEEE, 2016.

[10] Han Woong Kim, Ha Eun Lee, Sangwon Lee, Kyeong Taek Oh, Mijin Yun, and Sun Kook Yoo. Slice-selective learning for alzheimer's disease classification using a generative adversarial network: a feasibility study of external validation. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–10, 2020.

[11] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Tao Zhou, Yong Xia, and Dinggang Shen. Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimer's disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 455–463. Springer, 2018.

[12] Fernando Pérez-García, Rachel Sparks, and Sebastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv preprint arXiv:2003.04696*, 2020.

[13] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[14] Rajat Vikram Singh. Imagenet winning cnn architectures–a review.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

A Supplemental Figures

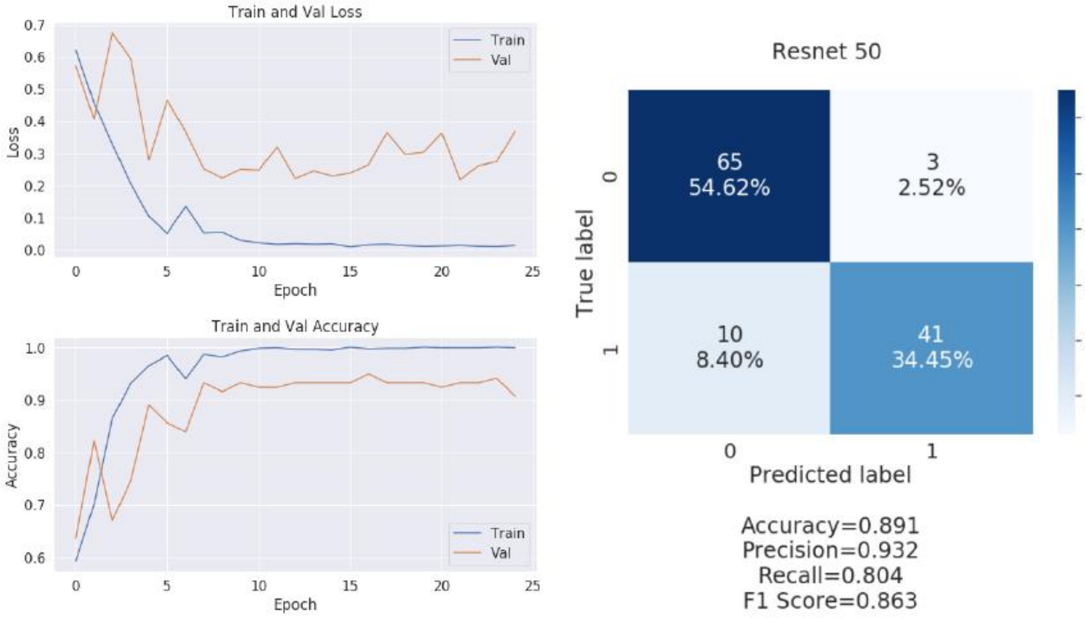

Fig. 13: Results for ResNet50.

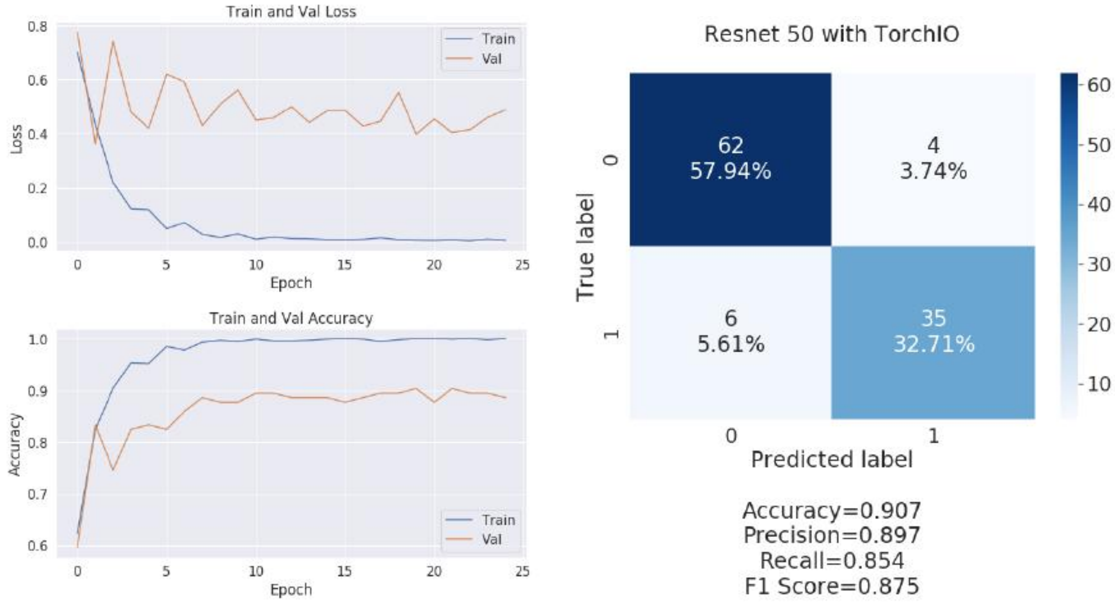

Fig. 14: Results for ResNet50 with TorchIO preprocessing.

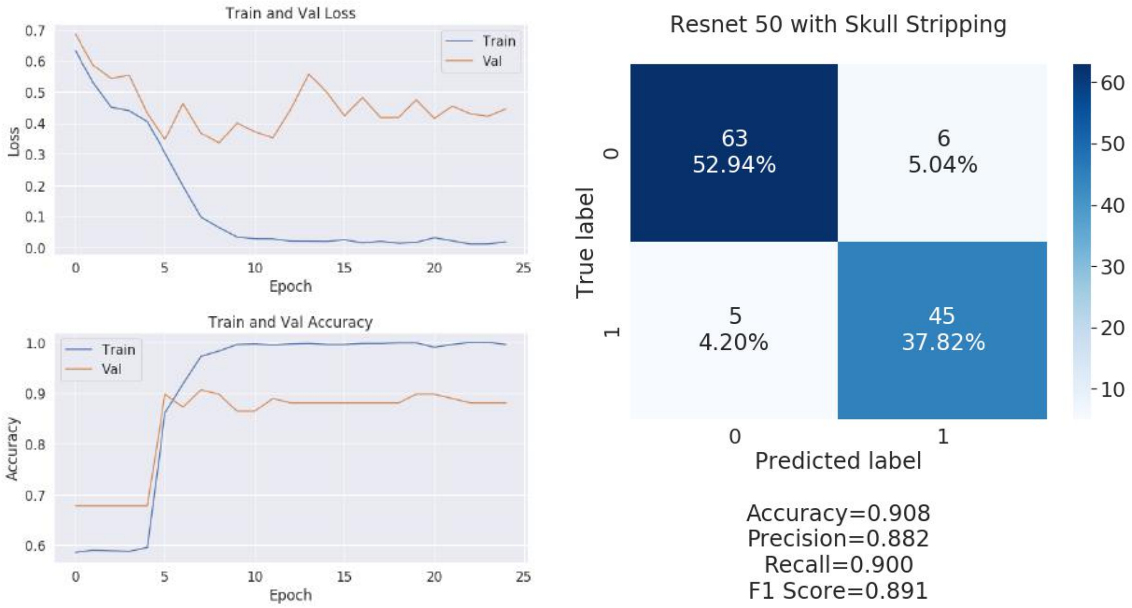

Fig. 15: Results for ResNet50 with skull stripping preprocessing.

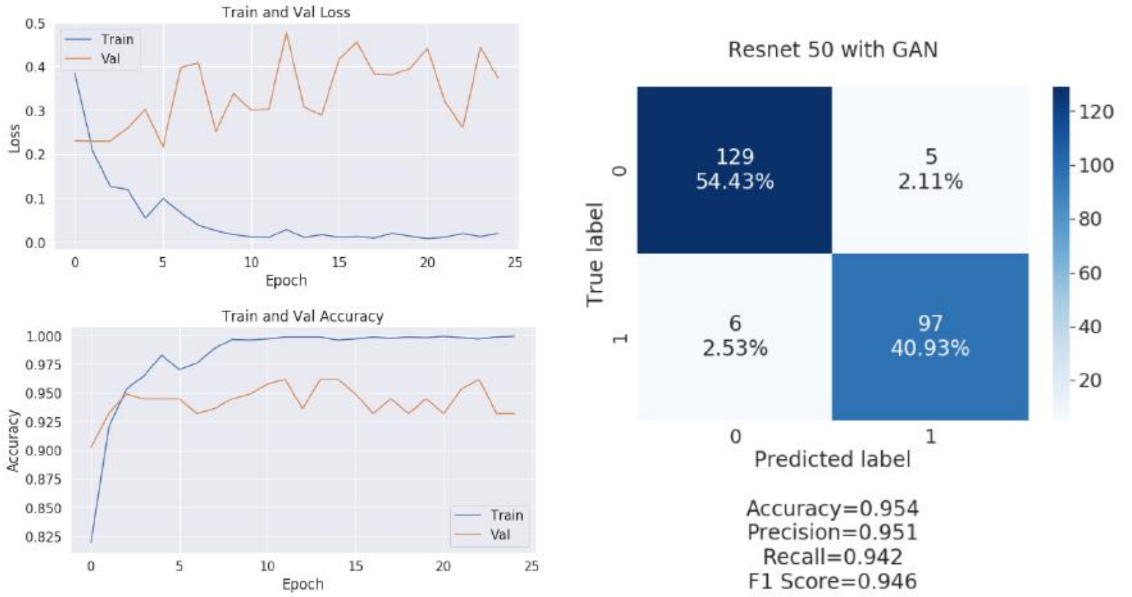

Fig. 16: Results for ResNet50 with GAN data augmentation.

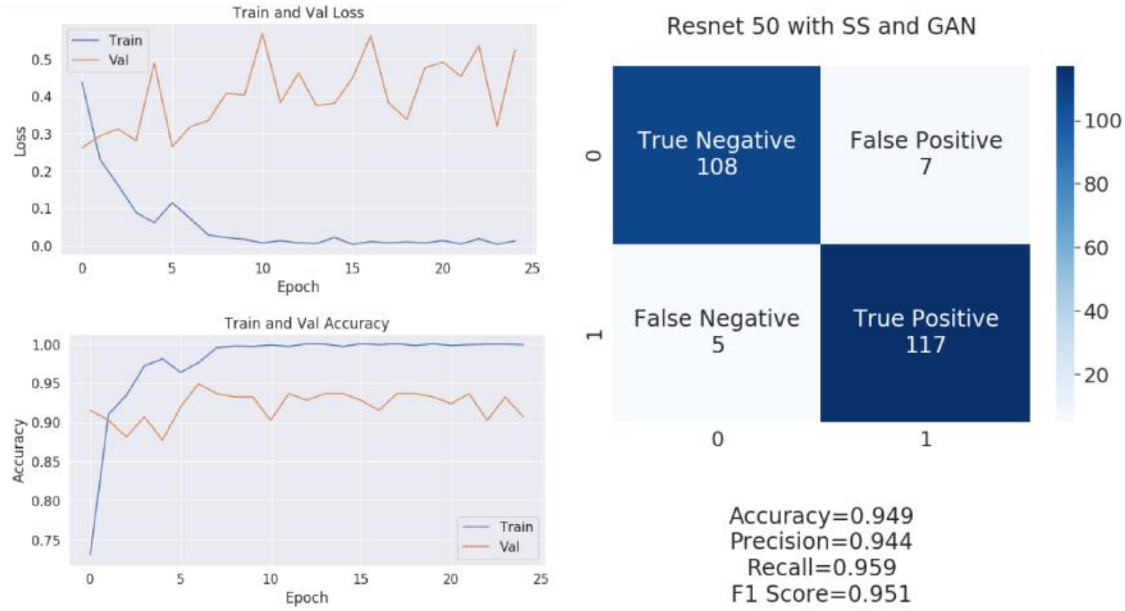

Fig. 17: Results for ResNet50 with skull stripping preprocessing and GAN data augmentation.