Names: Rahil Radia, Phil Leander, Adith Gopal, Ben House, Kyle Chen

Task & dataset & preprocessing:

Our chosen task was to develop a computer vision model that could classify MRI images of the brain based on stages of dementia. We hope that such a model could be used to detect early signs of Alzheimer's Disease in patients, and improve the care that they receive as a result of early diagnosis. The dataset we used was from Kaggle, which contained brain MRI images from four categories of dementia. These categories are "NonDemented," "VeryMildDemented," "MildDemented," and "ModerateDemented." There were a total of 6400 brain MRIs split between the four categories. Regarding data preprocessing, the MRIs needed to be rescaled and resized into a one dimensional vector. Because of there being a large class imbalance between the four classes, the SMOTETomek algorithm was used to randomly resample the data into an equal number of cases per class. 10% of the data was split before the rebalancing to be used as an independent and unbalanced test set. The remaining 90% of the data was split into 60-20-20 train, validation, and balanced test sets. A sample of the data can be seen in Figure 1.

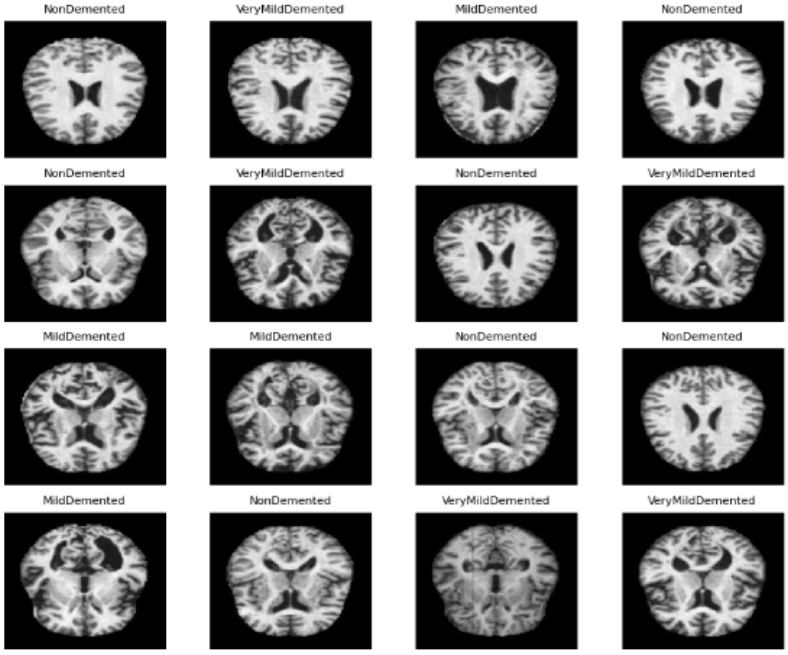


Figure 1. Randomly sampled MRIs from the dataset

The implementation & architectures of two deep learning systems:

To achieve our goal, we set out to build a Convolutional Neural Network, and Vision Transformer based model and see which of the two performed better at accurately classifying the MRI scans. Both models were developed using the Keras package and its corresponding libraries.

The CNN model is composed of four convolutional blocks that contain 2D convolutional, group normalization, ReLU activation, and average pooling layers. The outputs of the

convolutional blocks are then passed through a series of dropout, flatten, and dense layers before the final output is passed through a softmax layer which classifies the image as belonging to one of the four classes. Stochastic gradient descent was used while training the model, along with a callback function to reduce learning rate when loss began to converge. We chose to test a CNN model as they have been the most prevalent model used for computer vision tasks such as our own. The network's architecture is designed to effectively learn from the image data, capturing essential features necessary for distinguishing between different dementia stages through the convolutional layers. This model was trained using the aforementioned train set and took about 16 minutes to train at 12 epochs when loss converged.

Our second model was developed using the ViT package from keras that implements a prebuilt 16 patch vision transformer model. While attention transformers had commonly been used for NLP related tasks, they have gained momentum in the computer vision field as well for providing comparable or even superior accuracy to CNN models. A drawback of vision transformers is that they require large amounts of data to train, and due to the medical nature of our task it is difficult to get access to such a large datasets of MRIs labeled specifically for dementia. Because of this we opted to use the pretrained vit_b16 model from the package, and fine tuned using the same training set of MRI images. Our overall ViT model consisted of the vit_b16 block followed by the same series of flatten, dense, and dropout layers as the CNN model to keep testing as consistent as possible. Again stochastic gradient descent and the callback function for learning rate were defined. This model took significantly longer to train at around 4 hours for 8 epochs when loss converged.

Results and Conclusions:

Both models were tested on the balanced and unbalanced datasets for performance via metrics such as accuracy, precision, recall and F1 score. For our application, looking at recall was the most important as in a clinical setting it's important to minimize the number of false negative classifications made. Our CNN model on the balanced dataset had the best accuracy out of all testing at 98.5% and recall of 98.38%. When we tested the CNN model on the unbalanced dataset it performed slightly worse with a 96.7% accuracy and 96.53% recall.

Our ViT model had worse accuracy than our CNN model across the board. On the balanced dataset, the ViT model had an accuracy of 97.4% which is less than the 98.5% for CNN on the same dataset, and 97.40% recall. On the unbalanced dataset, the ViT model had an accuracy of 91.4% which is a good bit less than the 96.7% for CNN on the same dataset and a recall of 91.25%. As a result, we can see that the CNN model ended up being the best performing model overall at classification of MRI scans. Further performance of the CNN model on both balanced and unbalanced test data can be seen in Figure 4 at the end of this document.

There are a number of potential reasons for this, one being that the 16x16 patch size used for the ViT model was too large to differentiate between the often minute details in the MRI scans. Potentially using the 32x32 ViT would have produced more comparable results to the

CNN model. CNNs are also known for being able to find patterns in images through their augmentation of the data through convolutional layers. Without these layers, the ViT model which was primarily composed of attention transformer layers may have had trouble doing so. Further performance of the ViT model on both balanced and unbalanced test data can be seen in Figure 3.

Challenges and Obstacles:

One of the main challenges that we encountered was finding accessible data. Since the task involved using medical data, there are concerns around patient privacy so not many publically available datasets exist. The Kaggle dataset we did end up using worked for our purposes, but was only 6400 images which was somewhat limiting especially with the significant class imbalances that it had. We chose to rebalance the data because of this since few-shot classes would likely have poor predictive accuracy without it. We also wanted to make sure that the models still performed well on unbalanced datasets that would likely be seen in practical use so we decided to split off 10% of the dataset before rebalancing for independent data testing. Finally, the long training time of the ViT model was a challenge, so we had to limit the number of epochs to 8 when the loss started to converge to a near zero value. Given more time and better hardware, we would have attempted to train the ViT model further.

Figure 3: ViT Model Performance

On balanced dataset:

Testing Loss:	**0.082042**
Testing Accuracy:	**97.404003%**
Testing AUC:	**99.795556%**
Testing F1-Score:	**97.380304%**
Testing Precision:	**97.456712%**
Testing Recall:	**97.404003%**

On unbalanced dataset:

Testing Loss:	**0.251072**
Testing Accuracy:	**91.406250%**
Testing AUC:	**98.861933%**
Testing F1-Score:	**93.877947%**
Testing Precision:	**91.823900%**
Testing Recall:	**91.250002%**

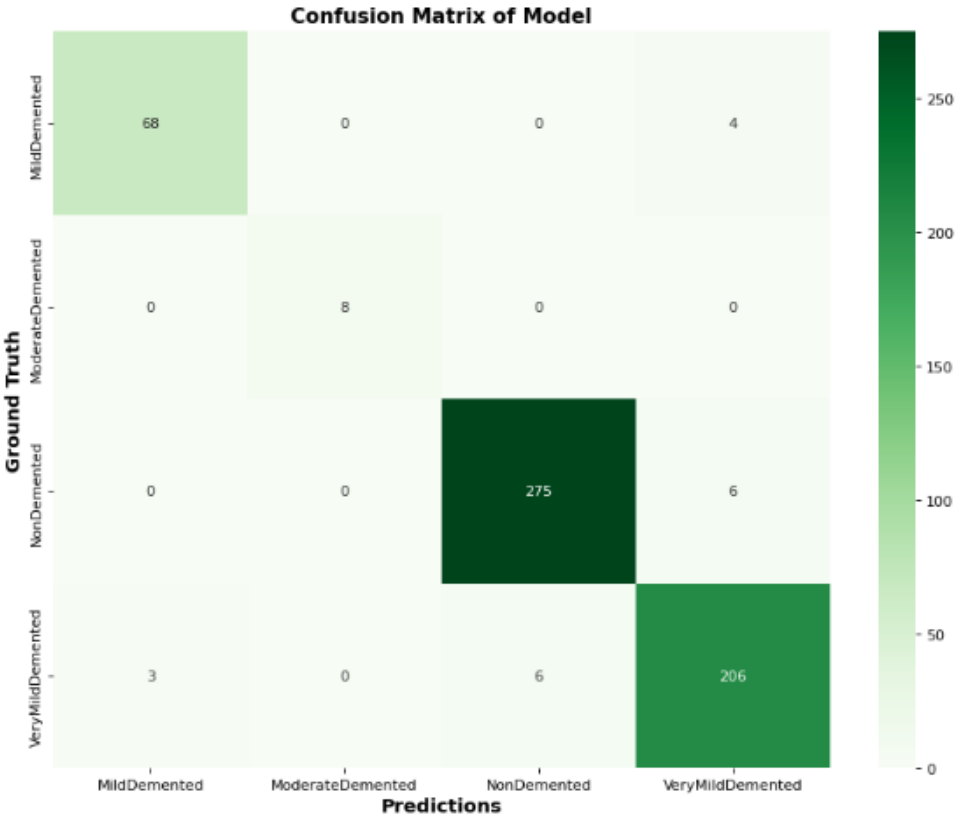
Figure 4: CNN Model Performance

On balanced dataset:

Testing Loss:	**0.049281**
Testing Accuracy:	**98.503888%**
Testing AUC:	**99.889153%**
Testing F1-Score:	**98.551488%**
Testing Precision:	**98.502100%**
Testing Recall:	**98.384202%**

On unbalanced dataset:

Testing Loss:	**0.098508**
Testing Accuracy:	**98.701390%**
Testing AUC:	**99.620926%**
Testing F1-Score:	**97.140324%**
Testing Precision:	**96.864110%**
Testing Recall:	**96.527779%**


Confusion Matrix of Model
Ground Truth
MildDemented
ModerateDemented
NonDemented
VeryMildDemented
68
0
0
4
0
8
0
0
0
0
275
6
3
0
6
206
MildDemented
ModerateDemented
NonDemented
VeryMildDemented
Predictions
250
200
150
100
50
0