BT5151 Advanced Analytics & Machine Learning

Group Project Report

Multi-class Image Classification of MRI Brain Scans for Diagnosis of Alzheimer's Disease

1. INTRODUCTION

1.1 What is the situation today?

Dementia, a debilitating condition, progressively impairs an individual's work performance, daily activities, and social interactions. As the population ages, the risk of dementia increases. In Singapore, where there is a rapidly aging population, one in four individuals will be aged 65 and above by 2030 (Sajan, 2023). Currently, approximately 1 in 10 elderly individuals aged 60 years and above suffer from dementia. This translates to an increase from about 82,000 in 2018 to 152,000 by 2030 (*Your Guide to Understanding Dementia*, n.d.). Notably, Alzheimer's disease is the most common form of dementia, accounting for 60-70% of all cases (*Alzheimer's Disease - Symptoms and Causes - Mayo Clinic*, 2024). Despite extensive research, no definitive cure for Alzheimer's exists. However, early diagnosis is crucial as it opens a wider array of treatment options. This includes medications that can slow disease progression and are most effective when initiated at an early stage (*Treatments*, n.d.). This allows the patient to better manage symptoms and enhance quality of life.

Presently, the diagnostic approach for dementia involves amalgamating data from medical examinations, mental ability tests, laboratory results, and brain scans (*OASIS Alzheimer's Detection*, 2023). MRI scans are instrumental in diagnosing Alzheimer's disease by enabling physicians to identify brain abnormalities and rule out other causes of memory loss such as hydrocephalus, brain tumours, or strokes (Heerema, 2014).

1.2 What are we solving for?

Singapore's aging population and the rising concern of dementia and Alzheimer's disease will have a significant impact not only on the persons living with dementia and their caregivers, but also on Singapore's healthcare system, in terms of cost, morbidity and mortality.

Therefore, our client (Ministry of Health Singapore) is looking to leverage machine learning algorithms for image classification to build a solution that can analyse structural and functional brain scans to detect subtle abnormalities associated with Alzheimer's disease to assist healthcare professionals in identifying potential indicators of Alzheimer's from MRI images.

1.3 What has already been done?

The literature on Alzheimer's disease consistently points to early detection as a critical component in managing the disease effectively. Alzheimer's disease, particularly noted for its impact on the hippocampus, a brain region vital for memory functions, is marked by observable atrophy in MRI images, indicating rapid tissue loss (Bobinski et al., 1999). This neurodegeneration in the hippocampus forms a primary diagnostic marker in identifying the progression of Alzheimer's through MRI scans.

Recent advancements in machine learning, especially deep learning techniques, have significantly enhanced the diagnostic capabilities for Alzheimer's disease. Techniques like image segmentation have been employed to isolate and analyse specific brain structures such as the hippocampus (Thyreau et al., 2018). Ensemble methods incorporating multiple classifiers, such as support vector machines, decision trees, and logistic regression, further refine the diagnostic accuracy by integrating various computational approaches (Diogo et al., 2022; Sato et al., 2018).

The application of convolutional neural networks (CNNs) in analysing brain MRI scans has shown promising results. Networks like DenseNet have been specifically highlighted for their efficiency and robust performance in image classification tasks. These networks leverage deep learning architectures to handle complex image data and extract pertinent features that are crucial for accurate diagnosis (Huang et al., 2017).

However, the literature also notes several challenges associated with the use of machine learning in medical imaging. One significant issue is the potential for overfitting, where a model might perform well on training data but poorly on unseen data. This is a critical concern in medical applications where the stakes are high. Another challenge is the interpretability of machine learning models; understanding how decisions are made by these complex models is crucial for clinical acceptance and ethical considerations in healthcare (Castelvecchi, 2016).

While deep learning presents a revolutionary tool in the fight against Alzheimer's, ensuring the reliability and interpretability of these tools remains a priority. Future research is expected to focus on refining these models to enhance their diagnostic accuracy and utility in clinical settings.

1.4 What is the benefit of early Alzheimer's early detection?

Early detection of Alzheimer's will benefit both the patient and the healthcare system. Economic modeling shows that early diagnosis of Alzheimer's can greatly reduce healthcare costs, particularly through reduced hospitalisations and better management of chronic conditions — potentially up to $64,000 per person with dementia (*Advancing Early Detection*, n.d.).

2. OUR PROPOSED SOLUTION

2.1 Data Preparation & Processing

- **OASIS MRI Dataset as Source**

For this project, we utilised the Open Access Series of Imaging Studies (OASIS) MRI dataset from Kaggle, which consists of 86,437 brain MRI images (*OASIS Alzheimer's Detection*, 2023). These brain MRI images were obtained from 347 unique patients. For each patient, the MRI images included 61 2-dimensional slices taken from 4 different Multiplanar Reconstructions (MPR)[1]. The images are divided into four classes based on Alzheimer's disease progression – healthy control (77%), very mild dementia (16%), mild dementia (6%), and moderate dementia (1%).

- **Identify and Group by Unique Patients**

We first identified the 347 unique patients from unique identification in the file name – for example, file name "OAS1_0308_MR1_mpr-1_100.jpg" referred to unique patient id "0308"; then grouped the images of different patients in one set for ease of splitting into train, validation, test datasets in later stages.

[1] Multiplanar reconstruction (MPR) is a technique used in medical imaging, including MRI, to create images in different planes (such as axial, sagittal, and coronal) from the original acquired data. This allows for better visualisation and understanding of anatomical structures from different perspectives.

- **Group "Mild" and "Moderate" Classes into One Class**

After identifying the unique patients, we found that there are 266 "Healthy" class patients, 58 "Very Mild Dementia" class patients, 21 "Mild Dementia" class patients and 2 "Moderate Dementia" class patients. Due to the very small sample size of "Moderate Dementia" class, we decided to group "Mild Dementia" and "Moderate Dementia" into one class to form 3 classes of "Healthy", "Very Mild Dementia" and "Mild_Moderate Dementia".

- **Select MPR 1 Images**

Next, we used the MPR 1 (amongst the 6 different MPRs in the original dataset) because MPR 1 represents the "axial plane" which is the standard imaging plane used in many diagnostic scans and provides a fundamental view which many radiologists and clinicians are accustomed to reviewing. Furthermore, neurological structures (such as the brain and spinal cord) are best evaluated in this plane (Ghahnavieh et al., 2021). Given limited computational resources and RAM, this approach also enabled us to reduce the dataset size from 86,437 to 22,265 images.

- **Organise into Train / Validate / Test datasets**

The dataset is split into 70:20:10 ratio for train:validate:test split (15677, 4514 and 2074 respectively). In addition, we made sure that the MPR1 images of the same patient belong to the same split. This was done as the adjacent MRI image slices for each patient may have many similarities. Having all the slices from the same patient within the same split thus ensures that there will not be any data leakage or overfitting due to adjacent image slices in both the training and validation/testing split. Codes for the organisation of the dataset into train:validate:test sub-folders can be found in the separate attachment "organise_train_test_split.ipynb". The final dataset can be found in a separate zip file attachment "OASIS_dataset_organised_short.zip".

- **Image Transformation and Normalisation**

Before loading the images into dataloaders, we transformed and normalized the images:

- Original image in (3, 496, 248) is transformed into:
  - (3, 224, 224) to be applied to the DenseNet-121 model
  - (3, 299, 299) to be applied to the Inception-v3 model
- Normalisation along their means and standard deviations of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225] respectively.

2.2 Machine Learning Model

Our solution focuses on leveraging deep learning techniques to analyse MRI images and augment the diagnosis of Alzheimer's disease by healthcare professionals. CNNs are well-suited for image-based tasks and have shown promising results in medical imaging (Zhang & Qie, 2023). Specifically, we employed two pre-trained models – (i) DenseNet-121 and (ii) Inception-v3. which have been shown by

Bianco et al. (2018) to have relatively high accuracy of at least 75% and low computational complexity with manageable number of parameters for ImageNet dataset.

- **Using Pre-Trained DenseNet-121 Model**

Dense Convolutional Network (DenseNet) establishes direct connections between each layer and every other layer in a feed-forward manner. Unlike traditional convolutional networks that have 'L' connections between each layer and its subsequent layer, DenseNet has 'L(L+1)/2' direct connections. Each layer in DenseNet utilises the feature-maps from all preceding layers as inputs and provides its own feature-maps as inputs to all subsequent layers.

Advantages: DenseNet has been evaluated on various object recognition benchmark tasks, including CIFAR-10, CIFAR-100, SVHN, and ImageNet, and has consistently demonstrated significant improvements over the state-of-the-art approaches while requiring less computation (G. Huang et al., 2016). Overall, DenseNet offers a powerful and efficient architecture for deep learning tasks, providing improved performance, addressing redundant layers, alleviating the vanishing-gradient problem, enhancing feature propagation, encouraging feature reuse, and reducing the number of parameters.

Limitation: As DenseNet is not specifically used for the medical domain, it may not perform optimally as feature extractors for MRI brain scans. As we have a sizable dataset, we also attempted to fine-tune the later convolutional blocks of the DenseNet model, as these layers typically learn specialized features of images.

- **Using Pre-Trained Inception-v3 Model**

The Inception-v3 model is based on Convolutional Neural Networks (CNNs) and consists of 42 layers, and designed to be computationally efficient while maintaining high accuracy. One of the key optimisations is the use of auxiliary classifiers, which act as regularisers and help improve the convergence of deep neural networks.

2.3 Model Experimenting & Settings

- **General Training Model Parameters**

Batch size: 256

Epoch: 20 (max) with early stopping criteria of minimum delta of 0.001 (for validation AUROC metric) and patience of 3 (i.e., stopping if no improvement after 3 epochs)

- **Experiment Procedures**

We conducted five different experiments sequentially (as detailed in Table 1):

- Three experiments involve multi-class classification (Healthy, Very Mild, Mild_Moderate labels) with training/fine-tuning of different layers of DenseNet-121 model to evaluate the effects of various modifications on model prediction outcomes.
- Two experiments involve binary classification (Healthy, Dementia labels without differentiating by severity) through transfer learning of the DenseNet-121 and Inception-v3, in order to evaluate the efficacy of different pre-trained models.

	No. of Class(es)	Layers for Transfer Learning / Fine-tuning	Pre-Trained Model	Learning Rate
1	3 (Healthy, Very Mild, Mild_Moderate)	Classifier only	DenseNet-121	1×10^{-3}
2	3 (Healthy, Very Mild, Mild_Moderate)	Final Dense block + Classifier	DenseNet-121	1×10^{-4}
3	3 (Healthy, Very Mild, Mild_Moderate)	2nd last & Final Dense block + Classifier	DenseNet-121	1×10^{-4}
4	2 (Healthy, Dementia)	Classifier only	DenseNet-121	5×10^{-4}
5	2 (Healthy, Dementia)	Classifier only	Inception-v3	5×10^{-4}

Table 1: Experiment Settings

- **Model Training & Assessment Metrics**

Due to the imbalanced dataset, the cross entropy and binary cross entropy loss function (for multi-class and binary classification respectively) is weighed according to the proportion of sample sizes of the different classes.

The key evaluation metric used is AUROC (area under the ROC curve) which is more robust to class imbalance. In addition, the multi-class AUROC is also calculated based on the weighted average to account for class imbalance. Furthermore, "recall" metric (i.e., proportion of predicted "true dementia" cases out of the "dementia" cases) is used to measure the model's effectiveness in predicting a case of Alzheimer's due to the significant implications of an accurate diagnosis.

- **Experiment 1 Results**

After early stopping at Epoch 7, Experiment 1 achieved 0.826 Test AUROC and good recall for "Very Mild" class (61%) and relatively weaker recall for "Mild_Moderate" class (34%). This could be due to the small sample size of "Mild_Moderate" class. Encouragingly, most of the "Mild_Moderate" actual class were predicted as "Very Mild" class (refer to confusion matrix in Appendix 1) – in other words, our model can predict the patient as a dementia case but of different severity. The model could correctly predict 2 out of 3 cases of "Healthy" patients which means 33% of "Healthy" patients will need to undergo "un-necessary" further diagnostic tests which we will discuss further in Section 2.4. Detailed Train/Validate Tensorboard charts and Confusion Matrix can be found in Appendix 1.

	Train	Validate	Test
Loss	0.5679	0.9775	**0.909**
AUROC	0.9326	0.8213	**0.826**
Recall			**61% – Very Mild; 34% – Mild_Moderate; 67% - Healthy**

Table 2 Experiment 1 Results Summary

Based on the tensorboard curves, the training AUROC appeared to plateau at approximately 0.9, suggesting that the pre-trained Densenet convolutional layers may not be able to extract other naunced features from MRI scans. Given the large dataset, this led us to experiment with fine-tuning of the last (experiment 2) and second-last (experiment 3) Dense block of the pre-trained layers in order to learn image features specific or contextualised to MRI scans.

- **Experiment 2 & 3 Results**

A lower learning rate of 1×10^{-4} was applied for fine-tuning, aiming to facilitate a more precise adjustment of the pre-trained model's parameters while mitigating the risk of overfitting.

After early stopping at Epoch 3, Experiment 2 achieved 0.844 Test AUROC (slightly better than Experiment 1). However, there appears to be "Overfitting" because the validation loss is on an increasing trend across epochs and the final validation loss of 1.6977 is higher than Experiment 1. In terms of "recall" metrics, Experiment 2 saw improvement in "Mild_Moderate" and "Healthy" classes, but a significant decline in "Very Mild" class compared to Experiment 1. About 1 out of 3 "Very Mild" patients were mis-predicted as "Healthy" class and this is not in line with our objective of early detection of Alzheimer which will lead to higher subsequent economic burden for the patient. Detailed Train/Validate Tensorboard charts and Confusion Matrix can be found in Appendix 2.

	Train	Validate	Test
Loss	0.0653	1.6977	**0.931**
AUROC	0.9991	0.8187	**0.844**
Recall			**45% – Very Mild; 57% – Mild_Moderate; 75% - Healthy**

Table 3 Experiment 2 Results Summary

Experiment 3 showed more "Overfitting" with a validation loss of 2.8816 which is higher than Experiment 2. Compared to Experiment 1 and 2, Experiment 3 has the best recall of "Healthy" class at 89%, but lowest recall for "Very Mild" and "Mild_Moderate" classes at 29% and 34% respectively. Generally, model prediction is less severe than actual – for example, 64% of "Very Mild" patients are predicted to be "Healthy" and 54% of Mild_Moderate patients are predicted to be "Very Mild". Detailed Train/Validate Tensorboard charts and Confusion Matrix can be found in Appendix 3.

	Train	Validate	Test
Loss	0.0039	2.8816	**Could not complete "Test" run due to running out of memory**
AUROC	1	0.8101	
Recall			

Table 4 Experiment 3 Results Summary

In experiment 2 and 3, performing fine-tuning of the dense convolutional blocks in the pre-trained DenseNet model led to overfitting of the the model and did not improve early detection of Alzheimer. Specifically, there was more overfitting towards the majority (healthy) class. Despite having a fairly large dataset of more than 20,000 images, we noted that these images come from slices from only 347 unique patients. As such, this may have led to insufficient variability in the data and given that Densenet is a highly complex model, led to exacerbation of the model's tendency to memorize patterns from the dominant healthy class, thereby impeding its generalization to less represented classes.

- **Experiment 4 Results**

Given the challenges of misclassification between very mild and mild-moderate stages, as well as class imbalance, we explored a binary classification approach by combining all labels across dementia severities (to "Dementia" vs "Healthy"). Furthermore, this approach enables us to identify all dementia cases, regardless of severity, in order to facilitate early treatment intervention.

Experiment 4, which applied transfer learning for the binary classification task with a learning rate of 5 x 10^{-4}, yielded similar results as Experiment 1 with a Test AUROC of 0.825. In view of the class imbalance between "Healthy" (80% of cases) and "Dementia" (20% of cases), we also explored various decision ("torch.sigmoid(logits)") thresholds to enhance the model's recall for the minority class as

outlined in Table 6, thus prioritizing the accurate detection of dementia cases while minimizing false negatives. Overall, a decision (logits) threshold of 0.15 gives the best recall of "Dementia" cases achieving our objective of early detection of Alzheimer's. However, this is at the expense of a less accurate prediction of the "Healthy" patients where the recall of "Healthy" cases is 54%. Detailed Train/Validate Tensorboard charts and Confusion Matrix can be found in Appendix 4.

	Train	Validate	Test
Loss	1.667	1.968	**1.856**
AUROC	0.875	0.837	**0.825**

Table 5 Experiment 4 Results Summary

Threshold	Recall "Dementia"	Recall "Healthy"	Weighted F1 score
0.5	41%	91%	0.79
0.4	57%	83%	0.78
0.3	75%	73%	0.76
0.25	82%	68%	0.74
0.2	91%	62%	0.71
0.15	95%	54%	0.66

Table 6 Recall & Weighted F1 Metric at Different Decision Threshold Levels

- **Experiment 5 Results**

As alluded to in section 2.2, the Inception-v3 model also demonstrated good performance in medical imaging use-cases. In experiment 5, we applied transfer learning of the pre-trained Inception-v3 model for binary dementia prediction with a training rate of 5×10^{-4}, as well as a weighted binary cross entropy loss in an attempt to mitigate the class imbalance issue. With the same decision threshold, the model yielded a better recall for "Dementia" cases than Experiment 4 (which uses pre-trained DenseNet-121 model). However, the recall for "Healthy" cases is worse than Experiment 4 and is generally lower than 50% - this may be a concern as false positive predicting may lead to unnecessary traumatic distress to these patients. Detailed Train/Validate Tensorboard charts and Confusion Matrix can be found in Appendix 5.

	Train	Validate	Test
Loss	0.7591	0.9377	**0.930**
AUROC	0.8857	0.8271	**0.811**

Table 7 Experiment 5 Results Summary

Threshold	Recall "Dementia"	Recall "Healthy"	Weighted F1 score
0.5	89%	55%	0.65
0.4	96%	43%	0.57
0.3	99%	33%	0.48
0.25	99%	28%	0.43
0.2	100%	22%	0.37
0.15	100%	15%	0.29

Table 8 Recall & Weighted F1 Metric at Different Decision Threshold Levels

2.4 Insights & Implications for Early Detection of Alzheimer's

The **Expected Value Framework (EVF)** is a method typically used to evaluate classification models, particularly in scenarios where the costs associated with misclassification are not equal. It considers not only the accuracy of the model but also the costs and benefits associated with different types of classification errors.

Table 9 and 10 outline the costs (or benefits) associated with each outcome of the classification. For example, in our scenario:

- True positives (i.e., predicting "Dementia" patients as Dementia) will have associated cost savings benefits of about 30% (Homage, 2023). Since the net care cost of dementia per person per year was estimated at S$33,130.81 (Moderate/Mild) and SG$8,370.65 (Very Mild), we estimated the cost savings benefit to be 30% of these costs (Woo et al., 2017).
- False positives (i.e., predicting "Healthy" patients as Dementia) might incur further diagnostics costs such as further psychiatric evaluation; neurological evaluation and laboratory tests (e.g., positron emission tomography) (Homage, 2023)
- False negatives (i.e., predicting "Dementia" patients as Healthy) might lead to lost cost savings due to missed opportunities for early detection and treatment.

		PREDICTED		
		Healthy	**Very Mild**	**Mild_Moderate**
TRUE	**Healthy**	0	-$650	-$650
	Very Mild	-$2,511 per year	+$2,511per year	+$2,511 per year
	Mild_Moderate	-$9,931 per year	+$9,931 per year	+$9,931 per year

Table 9 Cost Benefits for Multi-Class

		PREDICTED	
		Healthy	**Dementia**
TRUE	**Healthy**	0	-$650
	Dementia	-$6,221 per year (average of Very Mild & Mild_Moderate)	+$6,221 per year (average of Very Mild & Mild_Moderate)

Table 10 Cost Benefits for Binary Class

Table 11 outlines the expected value for each multi-class experiment.

Experiment	Expected Value	Implications
1	$629	• Experiment 1, which has the best overall prediction of Dementia (regardless the class) has the highest expected value as it could provide the best early detection.
2	$570	
3	NA; no test run	

Table 11 Expected Values of Multi-Class Experiments

Table 12 outlines the expected value for each threshold level of binary experiments 4 and 5.

Experiment 4: Whilst threshold of 0.15 has the highest expected value of $909 because it could highest recall of "Dementia" cases, we have to be mindful that it also has the highest mis-prediction of "Healthy" cases which will lead to traumatic distress for patients and their families.

Experiment 5: Even though Experiment 5 has better recall of "Dementia" cases than Experiment 4, the expected value for Experiment 5 is lower than $909 (the highest from Experiment 4). This is because Experiment 5 has poorer recall of "Healthy" cases which means the benefits of early detection is off-set by the higher incurred cost of further diagnostic tests. Besides the cost of additional diagnostic

tests, we need to be mindful of the psychological cost due to the un-necessary traumatic distress caused to healthy patients and their families.

Threshold	Expected Value (DenseNet-121)	Expected Value (Inception-v3)
0.5	-$282	$744
0.4	$83	$857
0.3	$494	**$883**
0.25	$652	$872
0.2	$838	$853
0.15	**$909**	$825

Table 12 Expected Values of Binary Class Experiment (DenseNet-121 and Inception-v3)

3. CONCLUSION

3.1 Could we achieve our goal in the proposal?

Overall, the machine learning model based on pre-trained DenseNet-121 and Inception-v3 models could provide early detection of Alzheimer's from MRI scan. We tried different experiments (transfer learning and fine-tuning of the pre-trained models, using multi-class and binary class) to identify the best model type. The Densenet-121 binary classification model with transfer learning yielded the highest expected value of $909 per person per year, with a high recall of 95% while maintaining an acceptable weighted F1 score of 0.66. This allows for early identification of most patients with dementia, thereby ensuring timely follow-up diagnostic action, intervention and improved prognosis.

3.2 What are the difficulties we faced?

When conducting this study, we encountered several difficulties, particularly in the areas of data collection and computational constraints:

Data Collection: The dataset provides less information about metadata, such as patient-specific image identification, the number of images per patient, and brain slice details. We addressed this by extracting necessary information from the image filenames, which enabled proper data organisation and analysis.

Computational Constraints: The dataset included approximately 80,000 large images, requiring significant computational resources. To mitigate this, we created a smaller subset of the dataset, reducing computational demands while maintaining sufficient data variety for robust analysis.

3.3 What are the limitations of our models?

The limitations of our models for image MRI images classification includes:

Generalisability Gap and Overfitting: The models can struggle with generalisability, meaning they might perform well on the specific data they were trained on but perform poorly on unseen data or data from different sources. This is particularly relevant for medical imaging due to variations in imaging protocols and machines.

Interpretability: As with many deep learning methods, this model is highly complex which makes them difficult to understand (often referred to as "black boxes"). In medical applications, it is crucial for practitioners to understand how decisions are made, which is challenging with complex models.

Limitation in image segmentation: Computational constraints limit us from performing image segmentation of the hippocampal region of the brain, a crucial area for determining Alzheimer's disease. Moreover, we do not have the medical expertise necessary to identify and create the training masks.

3.4 What can be done in future?

It is critical to ensure the explainability and trustworthiness of these models for clinicians to understand the rationale behind the predictions and have better discussion with their patients. As next steps, **explainable AI approaches** like LIME (Perturbation-based method) or Saliency Map (Gradient-based method) or Network Dissection could be used to provide interpretability to the model predictions.

Furthermore, we suggest to **perform MRI image segmentation** to highlight the hippocampal region in the images. One of the machine learning methods that could be utilized is the pre-trained Cycle Consistent Generative Adversarial Networks (CycleGAN), especially useful for contrast-enhanced images (T1-weighted) and high-resolution images (T2-weighted). This method was applied in brain tumor segmentation by Azni et al, 2023. Additionally, field experts are needed to accurately define the boundaries for the segmentation masks. With these improvements, we can then apply a similar machine learning process as conducted in this study but focused on the segmented images, which are expected to yield higher performance.

3.5 What ethical considerations to take note of?

We must take note of ethical considerations when deploying our model on early detection of Alzheimer's, so the benefits are balanced with the protection of participant rights, privacy, and fairness.

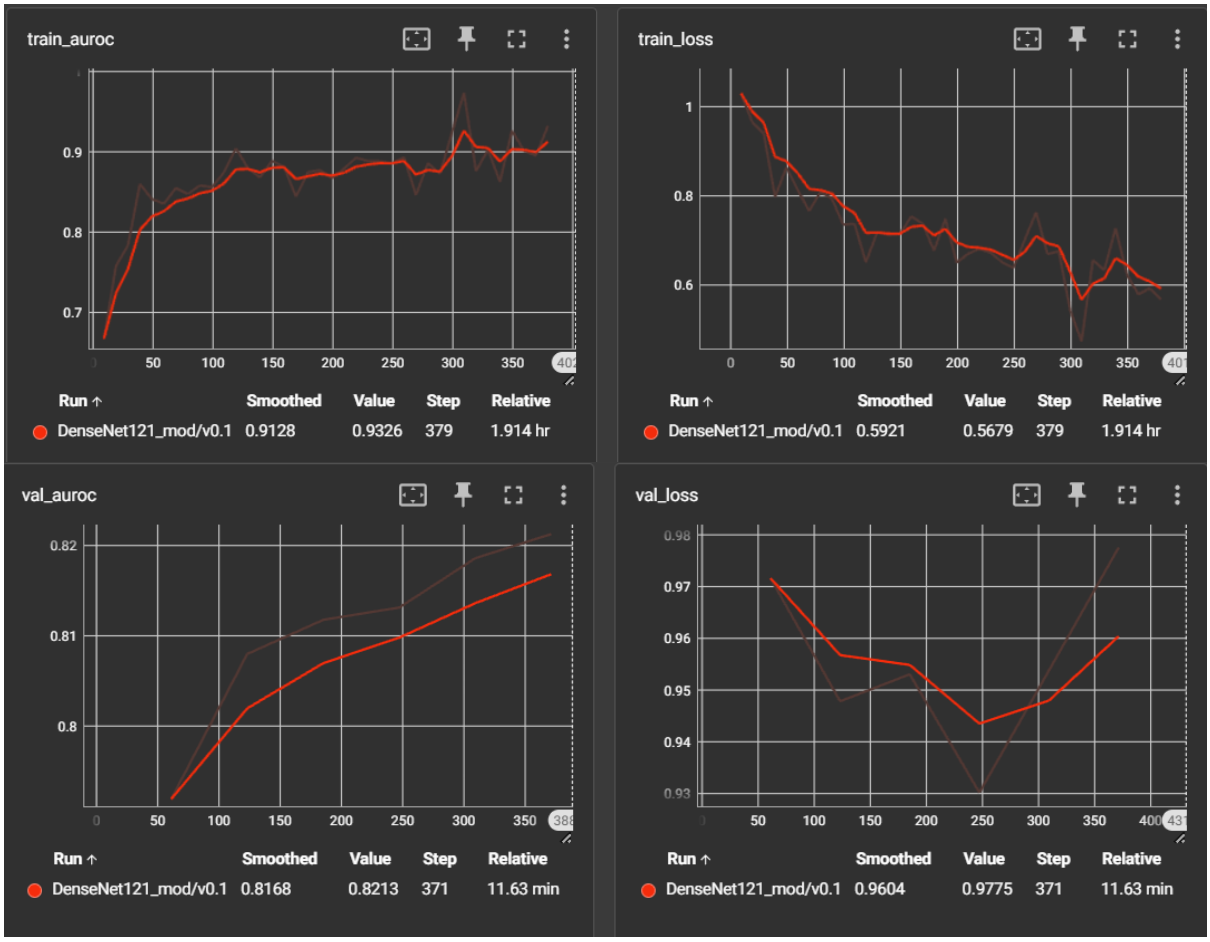
Informed Consent: If this dataset had been obtained by us directly, we would explain how participants were informed about the study's purpose, potential risks, benefits, and their rights to withdraw. Respecting the autonomy and decision-making capacity of participants is essential.

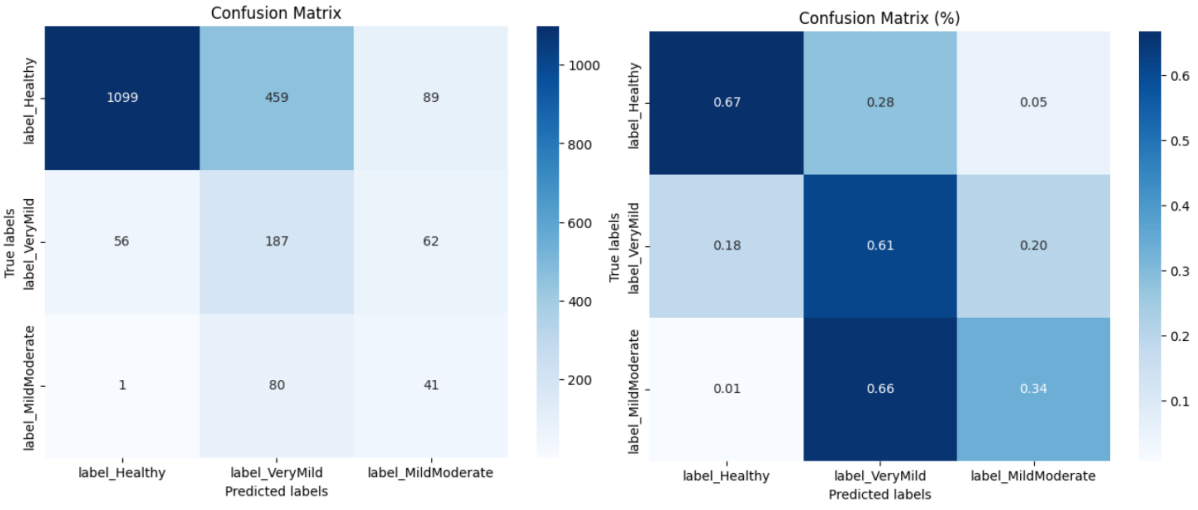
Privacy and Confidentiality: We prioritise the privacy and confidentiality of participant data. Measures were taken to anonymise and securely store the data, protecting it from unauthorised access. Ensuring data security and participant confidentiality is vital to maintain trust and respect in research.

Bias and Fairness: We acknowledge potential biases in the data used for training the machine learning model. We discuss how these biases were addressed and the steps taken to ensure fairness in the model's predictions. We recognise the impact of false positives and false negatives on individuals and the broader implications for healthcare.

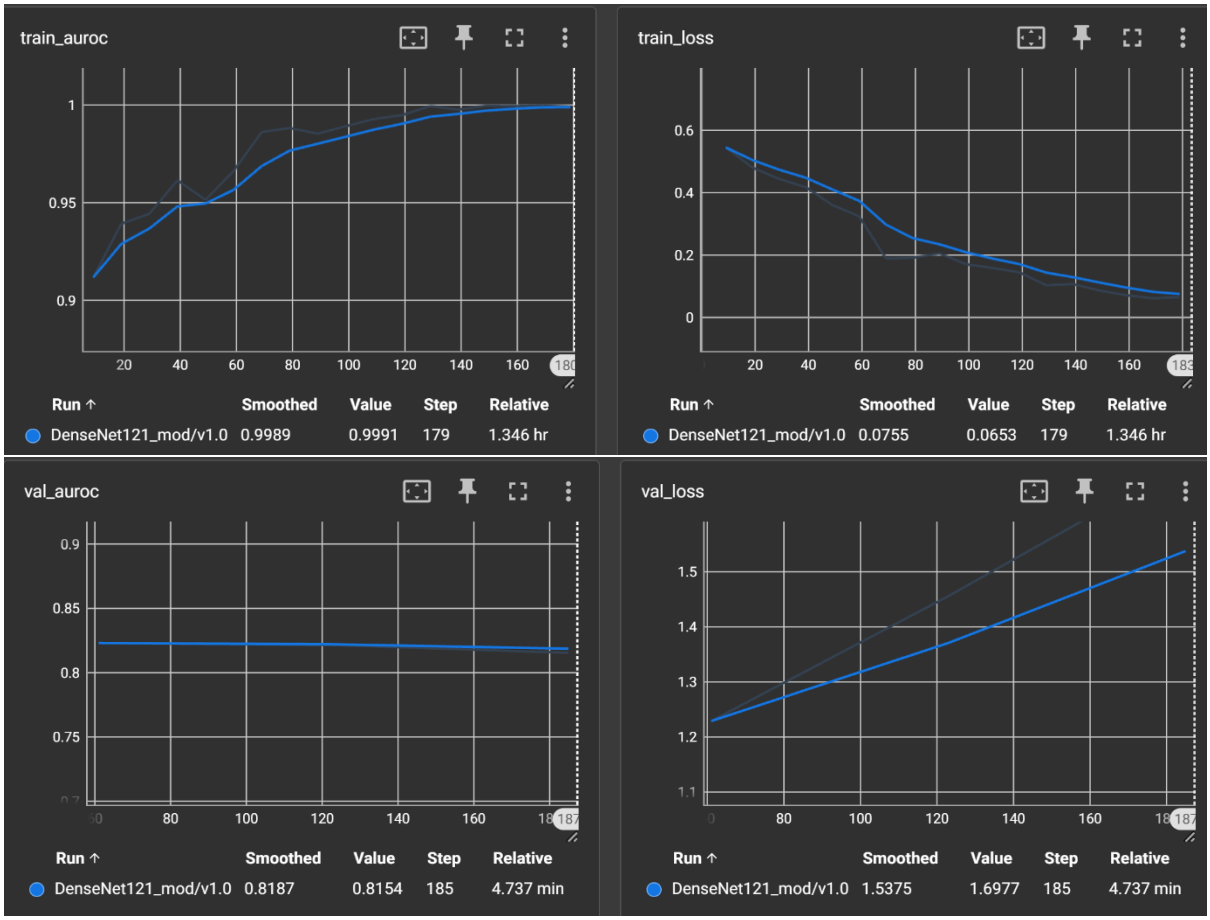
Accountability and Responsibility: We highlight the need for accountability and responsibility in the development and deployment of machine learning models for healthcare. We discuss steps taken to validate the model's performance, assess its limitations, and ensure responsible and ethical use. We acknowledge the potential consequences of relying solely on black-box models without appropriate oversight.

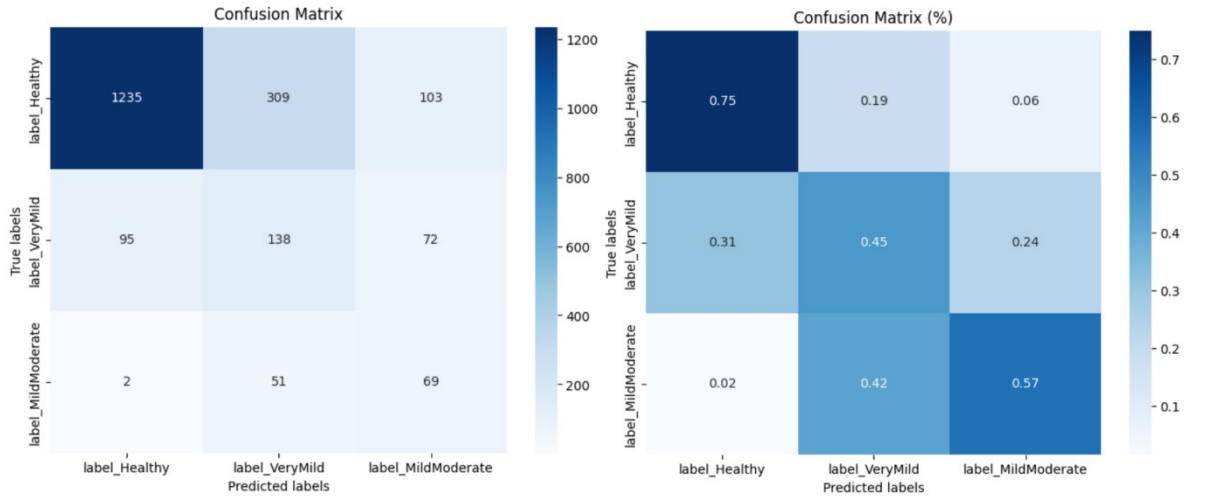
APPENDIX 1 – Experiment 1 Tensorboard Chart & Confusion Matrix



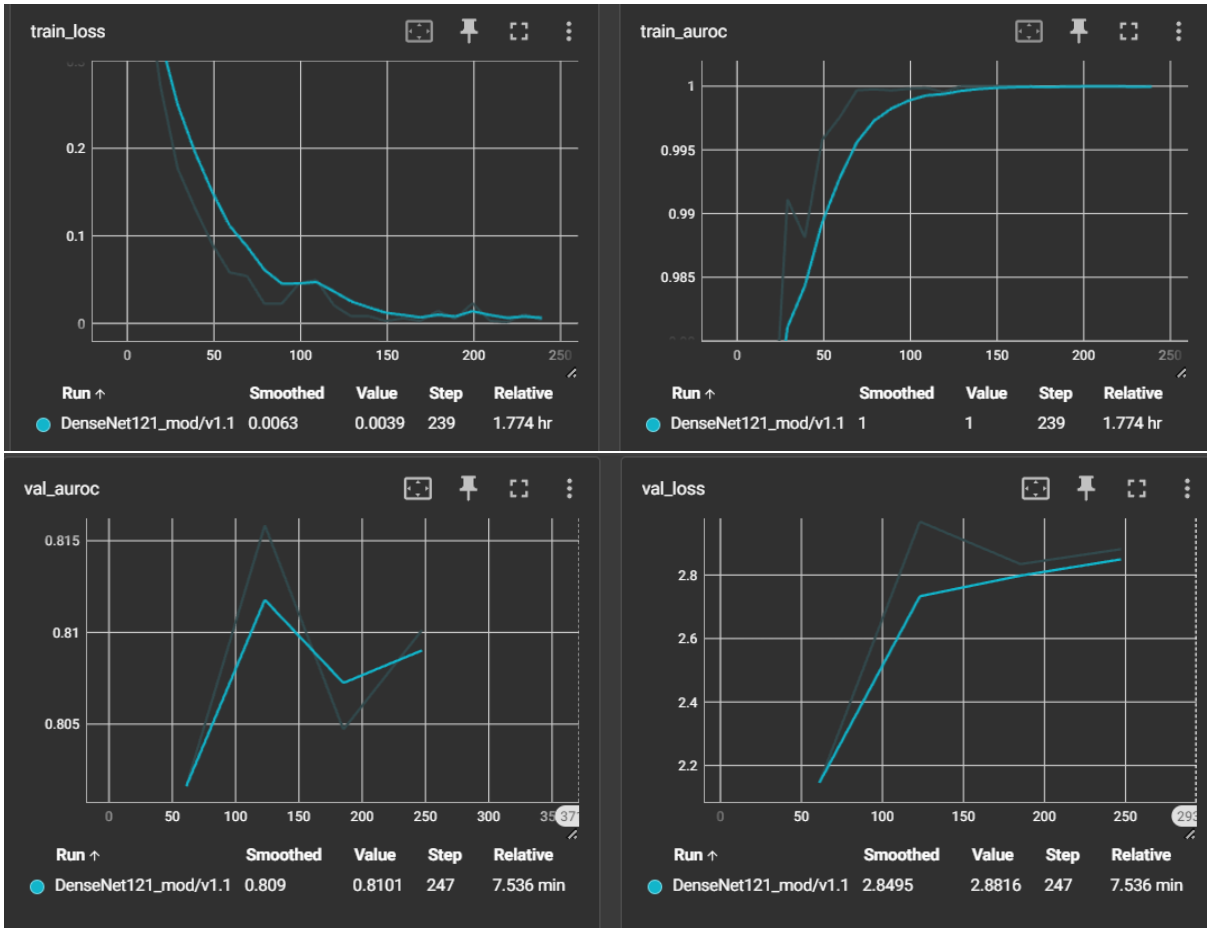


APPENDIX 2 – Experiment 2 Tensorboard Chart & Confusion Matrix

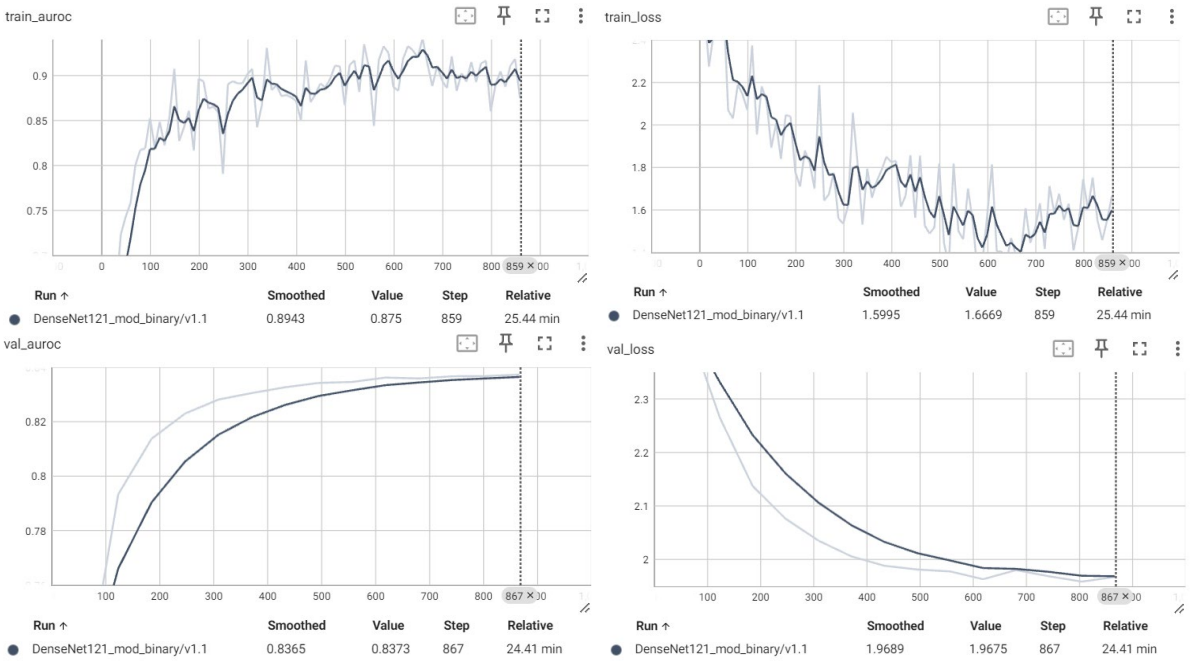




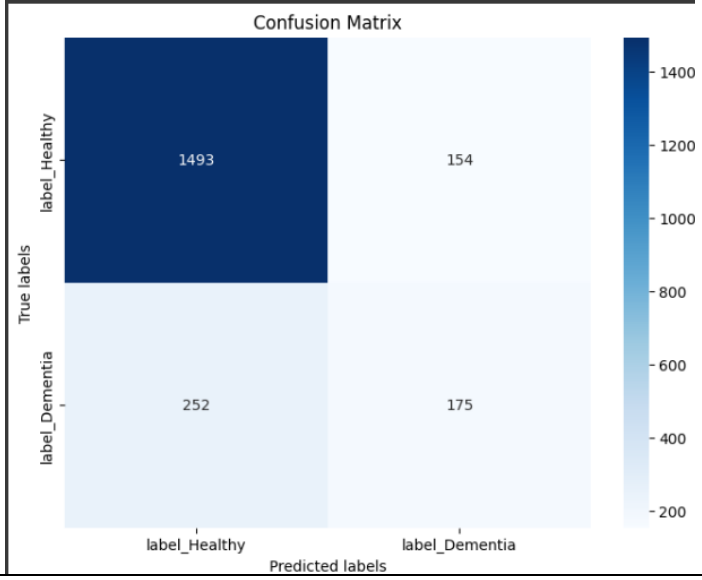
APPENDIX 3 – Experiment 3 Tensorboard Chart & Confusion Matrix



APPENDIX 4 – Experiment 4 Tensorboard Chart & Confusion Matrix



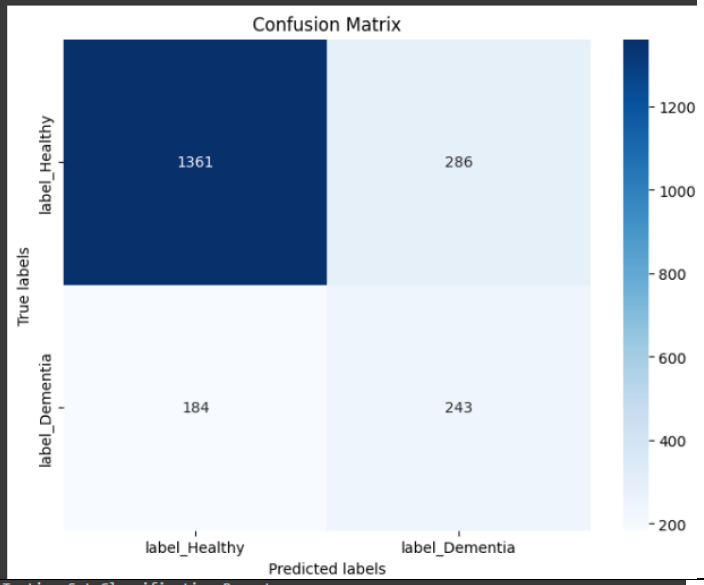
Classification Report & Confusion Matrix at Different Thresholds

Threshold	
0.5	Testing Set Classification Report: precision recall f1-score support 0 0.86 0.91 0.88 1647 1 0.53 0.41 0.46 427 accuracy 0.80 2074 macro avg 0.69 0.66 0.67 2074 weighted avg 0.79 0.80 0.79 2074 

0.4

Testing Set Classification Report:

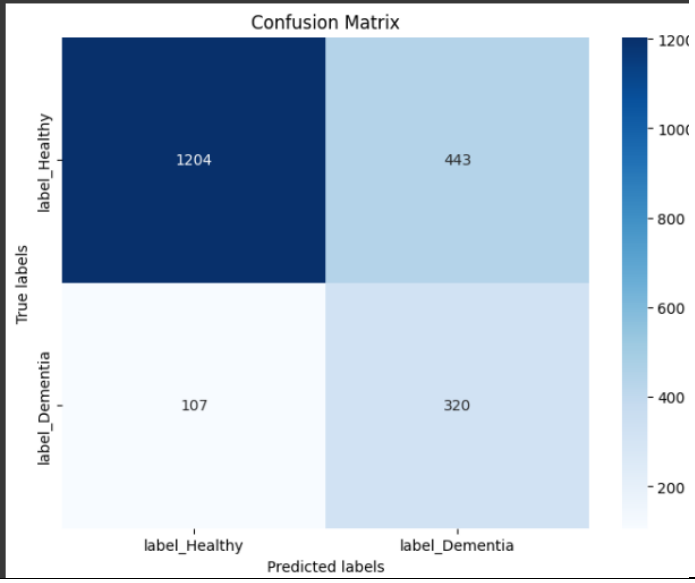
	precision	recall	f1-score	support
0	0.88	0.83	0.85	1647
1	0.46	0.57	0.51	427
accuracy			0.77	2074
macro avg	0.67	0.70	0.68	2074
weighted avg	0.79	0.77	0.78	2074



0.3

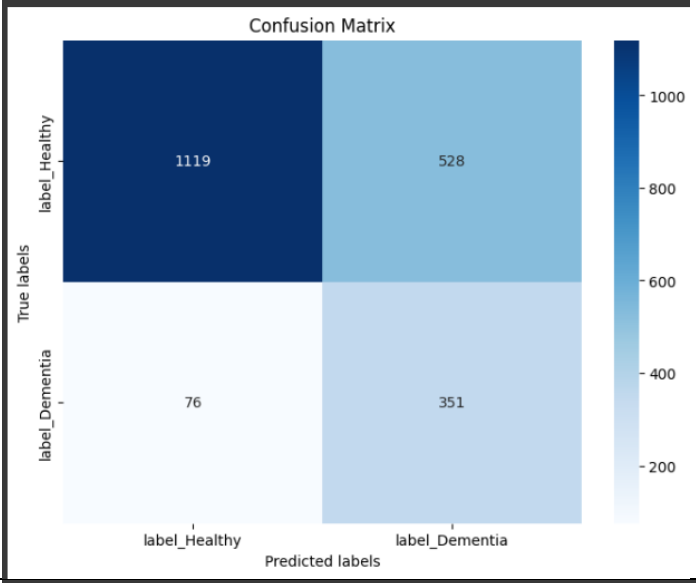
Testing Set Classification Report:

	precision	recall	f1-score	support
0	0.92	0.73	0.81	1647
1	0.42	0.75	0.54	427
accuracy			0.73	2074
macro avg	0.67	0.74	0.68	2074
weighted avg	0.82	0.73	0.76	2074



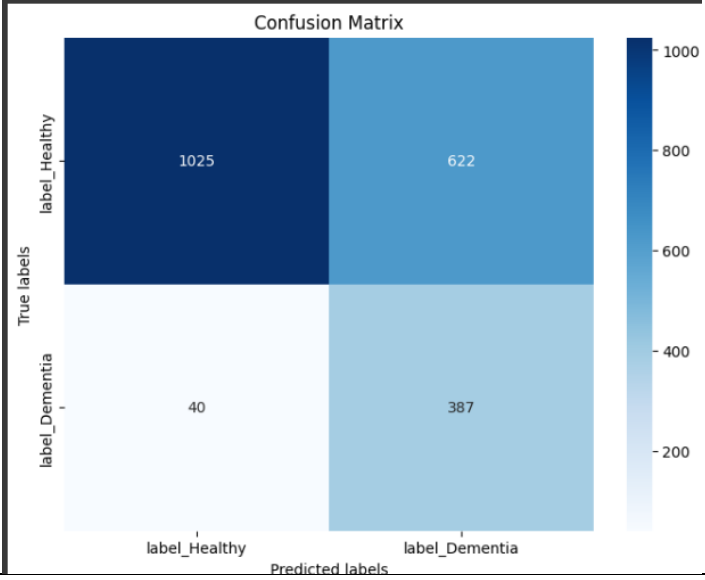
0.25

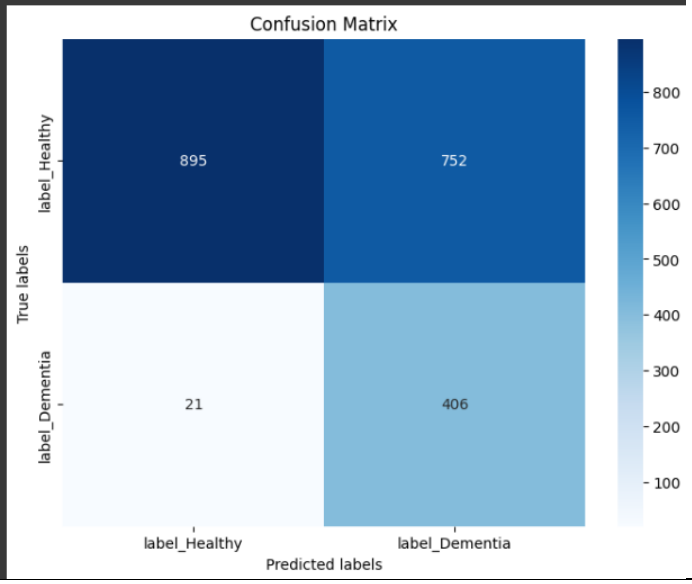
```
Testing Set Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.68      0.79      1647
           1       0.40      0.82      0.54       427

    accuracy                           0.71      2074
   macro avg       0.67      0.75      0.66      2074
weighted avg       0.83      0.71      0.74      2074
```



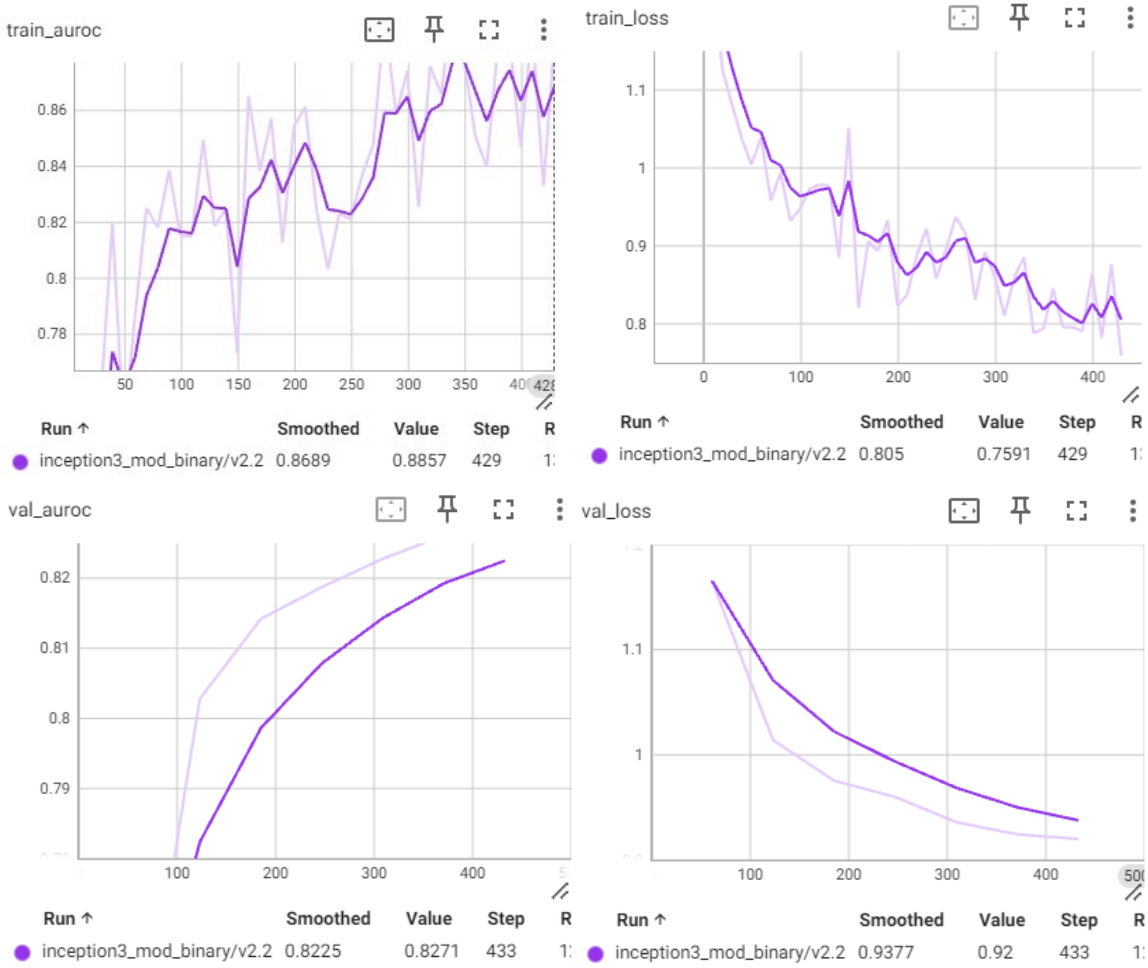
0.20

```
Testing Set Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.62      0.76      1647
           1       0.38      0.91      0.54       427

    accuracy                           0.68      2074
   macro avg       0.67      0.76      0.65      2074
weighted avg       0.84      0.68      0.71      2074
```

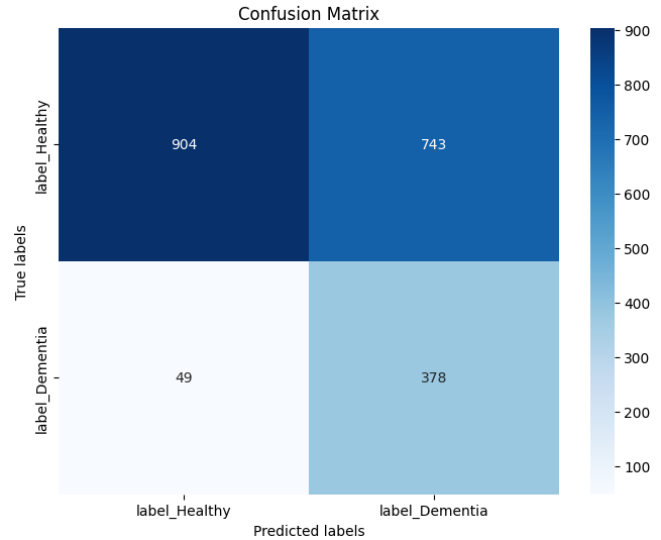


0.15	Testing Set Classification Report: precision recall f1-score support 0 0.98 0.54 0.70 1647 1 0.35 0.95 0.51 427 accuracy 0.63 2074 macro avg 0.66 0.75 0.61 2074 weighted avg 0.85 0.63 0.66 2074 

APPENDIX 5 – Experiment 5 Tensorboard Chart & Confusion Matrix

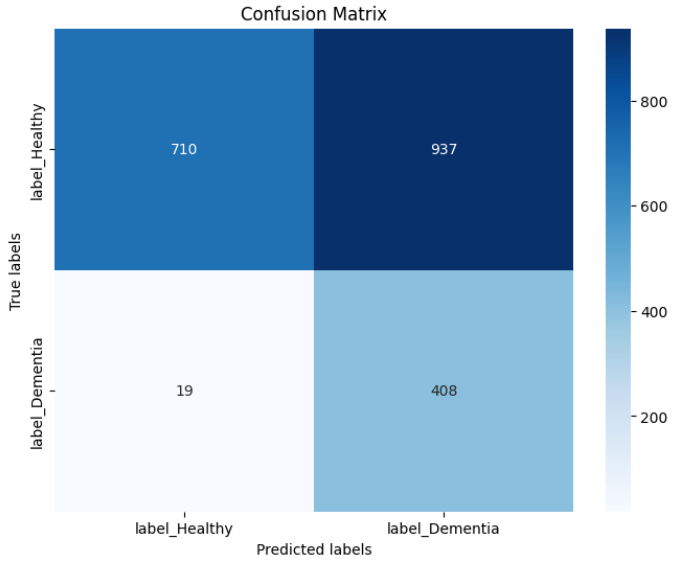


Classification Report & Confusion Matrix at Different Thresholds

Threshold	
0.5	Testing Set Classification Report: precision recall f1-score support 0 0.95 0.55 0.70 1647 1 0.34 0.89 0.49 427 accuracy 0.62 2074 macro avg 0.64 0.72 0.59 2074 weighted avg 0.82 0.62 0.65 2074 

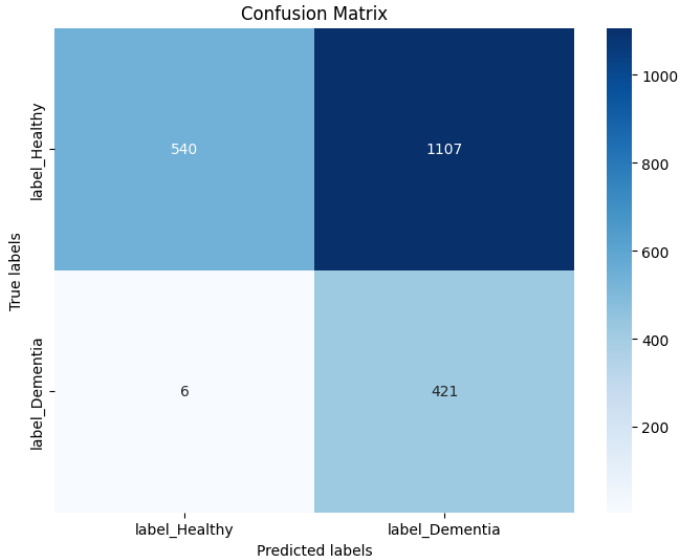
0.4

```
Testing Set Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.43      0.60      1647
           1       0.30      0.96      0.46       427

    accuracy                           0.54      2074
   macro avg       0.64      0.69      0.53      2074
weighted avg       0.84      0.54      0.57      2074
```



0.3

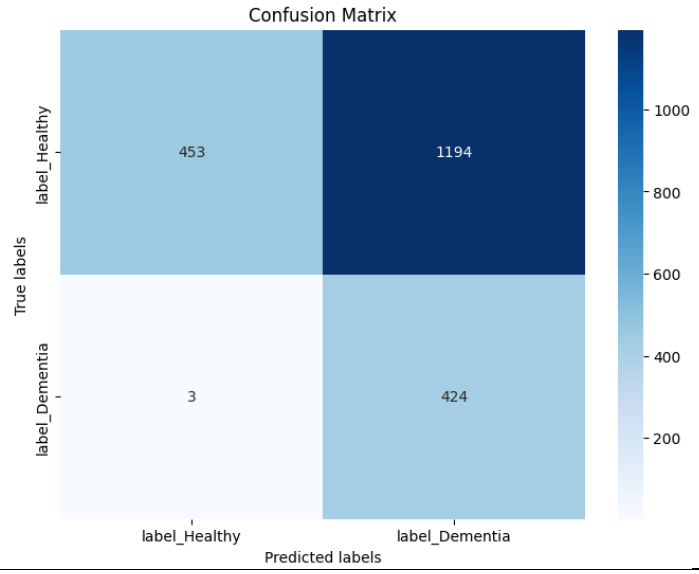
```
Testing Set Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.33      0.49      1647
           1       0.28      0.99      0.43       427

    accuracy                           0.46      2074
   macro avg       0.63      0.66      0.46      2074
weighted avg       0.84      0.46      0.48      2074
```



0.25

Testing Set Classification Report:

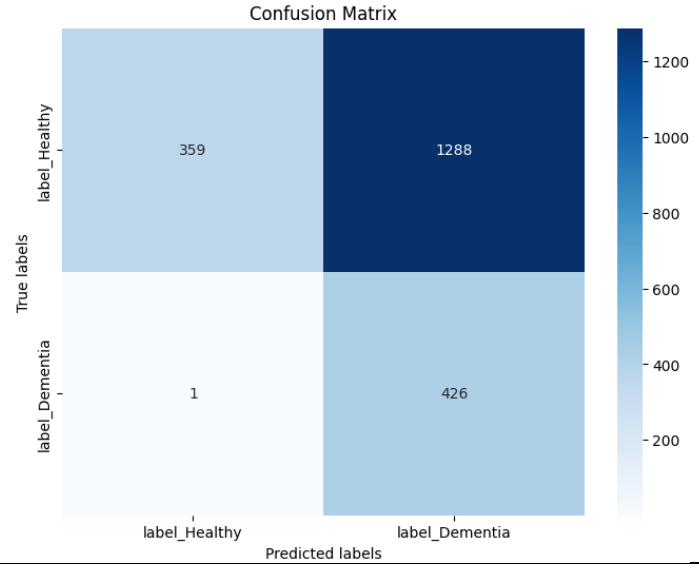
	precision	recall	f1-score	support
0	0.99	0.28	0.43	1647
1	0.26	0.99	0.41	427
accuracy			0.42	2074
macro avg	0.63	0.63	0.42	2074
weighted avg	0.84	0.42	0.43	2074

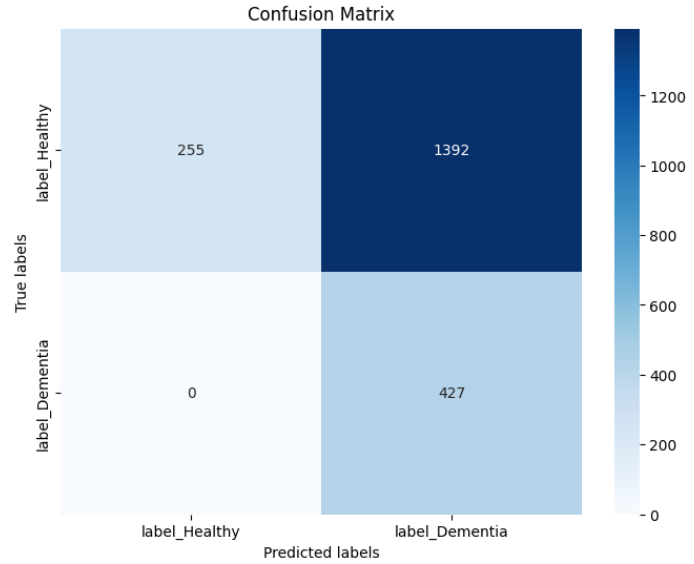


0.20

Testing Set Classification Report:

	precision	recall	f1-score	support
0	1.00	0.22	0.36	1647
1	0.25	1.00	0.40	427
accuracy			0.38	2074
macro avg	0.62	0.61	0.38	2074
weighted avg	0.84	0.38	0.37	2074

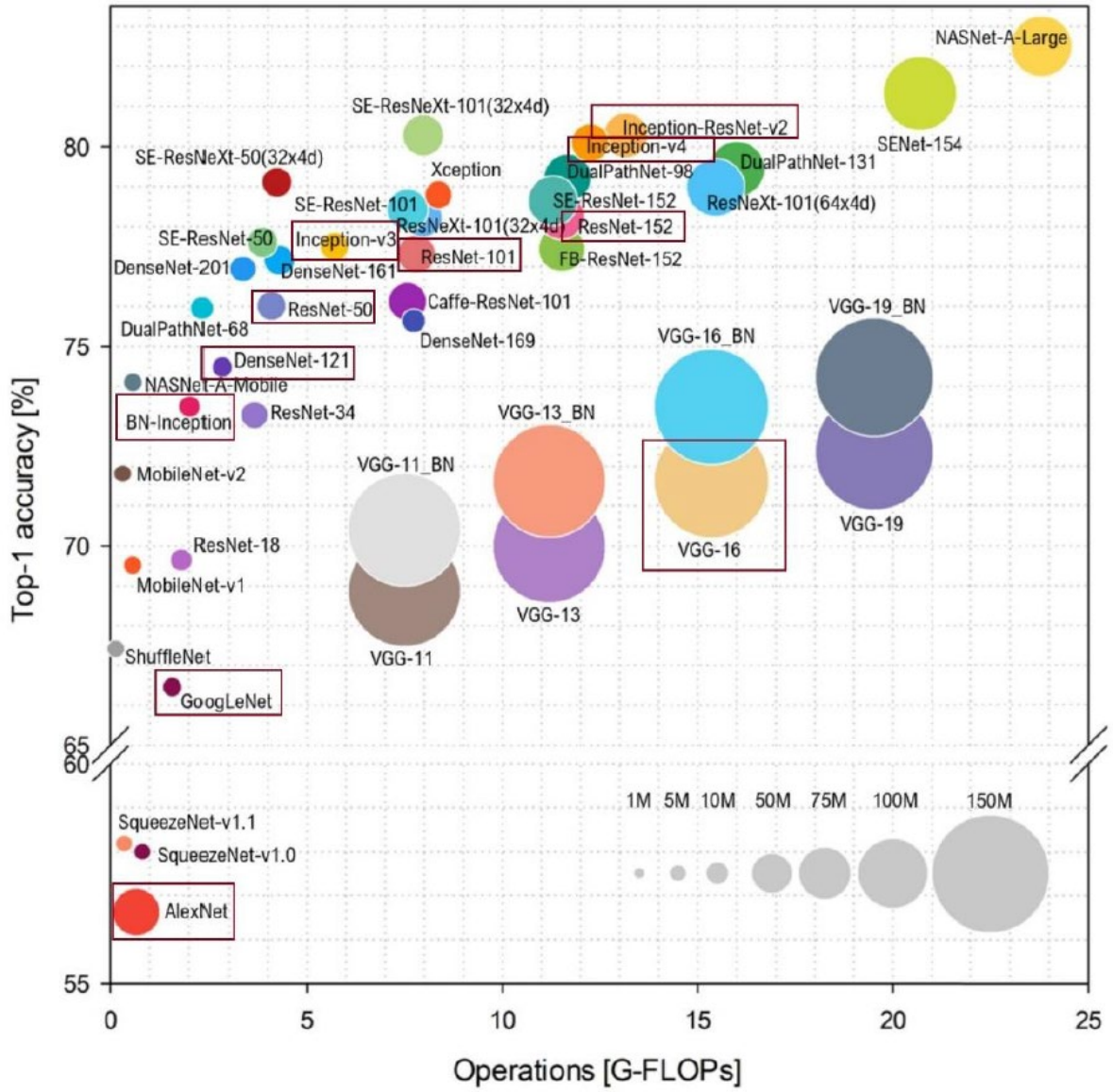


0.15	Testing Set Classification Report: precision recall f1-score support 0 1.00 0.15 0.27 1647 1 0.23 1.00 0.38 427 accuracy 0.33 2074 macro avg 0.62 0.58 0.32 2074 weighted avg 0.84 0.33 0.29 2074 

APPENDIX 6 Model Accuracy vs Computational Complexity for ImageNet dataset

Ball chart below reporting the Top-1 accuracy vs. computational complexity that is called floating-point operations per second (FLOPS) in computing; the size of the balls shows the parameters (Bianco *et al.*, 2018).

DenseNet-121 and Inception-v3 have relatively high accuracy of at least 75% and low computational complexity with manageable number of parameters.



APPENDIX 7 – Contribution of group members

Goh Dai Yong, Adison	• Code writing and programming • Report writing
Rizon Agustan Sinaga	• Code running • Report writing
Gary Wong Yue Whay	• Code running • Report writing
Nur Iryani Binti Halip	• Report writing • Presentation slides

REFERENCES

Advancing early detection. (n.d.). https://www.cdc.gov/aging/healthybrain/issue-maps/early-detection.html

Alzheimer's disease - Symptoms and causes - Mayo Clinic. (2024, February 13). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447

Azni, H. M., Afsharchi, M., & Allahverdi, A. (2023). Improving brain tumor segmentation performance using CycleGAN based feature extraction. *Multimedia Tools and Applications*, 82, 18039-18058. https://doi.org/10.1007/s11042-022-14174-3

Bianco, S., Cadène, R., Celona, L., & Napoletano, P. (2018). Benchmark analysis of Representative deep Neural network architectures. *IEEE Access*, *6*, 64270–64277. https://doi.org/10.1109/access.2018.2877890

Bobinski, M., De Leon, M. J., Węgiel, J., DeSanti, S., Convit, A., Louis, L. S., Rusinek, H., & Wisniewski, H. M. (1999). The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. *Neuroscience*, *95*(3), 721–725. https://doi.org/10.1016/s0306-4522(99)00476-5

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Diogo, V., Ferreira, H., & Prata, D. (2022). Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy*, *14*(1). https://doi.org/10.1186/s13195-022-01047-y

Ghahnavieh, A. E., Luo, S., & Initiative, A. D. N. (2021). Convolutional neural networks for Alzheimer's disease detection on MRI images. *Journal of Medical Imaging*, *8*(02). https://doi.org/10.1117/1.jmi.8.2.024503

Heerema, E. (2014, December 30). *How MRI is used to detect Alzheimer's Disease*. Verywell Health. https://www.verywellhealth.com/can-an-mri-detect-alzheimers-disease-98632

Homage. (2023, September 25). *Living with dementia? Here's how much it might potentially cost - homage*. https://www.homage.sg/resources/dementia-cost/

Huang, G., Liu, Z., Laurens, V. D. M., & Weinberger, K. Q. (2016, August 25). *Densely connected convolutional networks*. arXiv.org. https://arxiv.org/abs/1608.06993

Huang, K. L., Marcora, E., Pimenova, A. A., Di Narzo, A. F., Kapoor, M., Jin, S. C., Harari, O., Bertelsen, S., Fairfax, B. P., Czajkowski, J., Chouraki, V., Grenier-Boley, B., Bellenguez, C., Deming, Y., McKenzie, A., Raj, T., Renton, A. E., Budde, J., Smith, A. V., . . . Goate, A. M. (2017). A common haplotype lowers PU.1 expression in

myeloid cells and delays onset of Alzheimer's disease. *Nature Neuroscience*, *20*(8), 1052–1061. https://doi.org/10.1038/nn.4587

OASIS Alzheimer's detection. (2023, June 18). Kaggle. https://www.kaggle.com/datasets/ninadaithal/imagesoasis

Sajan, C. (2023, June 21). Asian dementia on the rise in Singapore. *The Straits Times*. https://www.straitstimes.com/life/home-design/asian-dementia-on-the-rise-in-singapore

Sato, C., Barthélemy, N. R., Mawuenyega, K. G., Patterson, B. W., Gordon, B. A., Jockel-Balsarotti, J., Sullivan, M., Crisp, M. J., Kasten, T., Kirmess, K. M., Kanaan, N. M., Yarasheski, K. E., Baker-Nigh, A., Benzinger, T. L., Miller, T. M., Karch, C. M., & Bateman, R. J. (2018). Tau Kinetics in Neurons and the Human Central Nervous System. *Neuron*, *97*(6), 1284-1298.e7. https://doi.org/10.1016/j.neuron.2018.02.015

Thyreau, B., Sato, K., Fukuda, H., & Taki, Y. (2018). Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. *Medical Image Analysis*, *43*, 214–228. https://doi.org/10.1016/j.media.2017.11.004

Treatments. (n.d.). Alzheimer's Disease and Dementia. https://www.alz.org/alzheimers-dementia/treatments

Woo, L. L., Thompson, C. L., & Dong, Y. H. (2017). Net informal costs of dementia in Singapore. *Journal of Clinical Gerontology & Geriatrics*, *8*(3), 98–101. https://doi.org/10.24816/jcgg.2017.v8i3.06

Your Guide to Understanding Dementia. (n.d.). https://www.healthhub.sg/live-healthy/yourguidetounderstandingdementia_pdf

Zhang, H. H., & Qie, Y. (2023). Applying Deep Learning to Medical Imaging: A review. *Applied Sciences*, *13*(18), 10521. https://doi.org/10.3390/app131810521