



State-of-the-Art Deep Learning for Alzheimer's Disease Classification

Pranav Malik, Dr. Wisam Bukaita

College of Arts and Sciences, Lawrence Technological University



Abstract

Alzheimer's disease is one of the most pressing public health challenges of our time, affecting millions worldwide and placing significant burdens on families and healthcare systems. Early and accurate diagnosis is crucial, as it enables timely interventions that may slow progression and improve quality of life. However, reliable detection remains complex, given that the disease manifests through diverse cognitive decline patterns and structural brain changes. This project applies two state-of-the-art deep learning architectures — a Convolutional Neural Network (CNN) and a Vision Transformer (ViT) — to structural MRI brain scans to classify different stages of Alzheimer's disease. Unlike conventional image analysis methods, these architectures excel at detecting subtle, high-dimensional patterns in neuroimaging data. Both models are trained and evaluated under the same conditions to enable a fair and meaningful comparison. Both models achieved strong classification performance, with DeiT outperforming YOLOv11 with 99% prediction accuracy compared to 87%, due to its ability to capture long-range spatial dependencies in MRI scans that local convolutional filters may miss. These results highlight the advantage of transformer-based architectures for Alzheimer's staging from MRI, suggesting that ViT models represent a promising direction for automated, accurate, and scalable neuroimaging-based diagnosis

Introduction

What is AD? - Alzheimer's Disease, or abbreviated as AD, is a neurodegenerative disease that is the most common cause of dementia in elderly adults. AD occurs when two proteins, Amyloid-beta peptide and Tau, form clumps around brain cells and tangles within them, effectively cutting off communication from outside cells, causing the affected cell to die [4].

How is it diagnosed? - Diagnosis of AD is conducted using both clinical and biological signs. Clinical diagnosis involves searching for "phenotypes", characteristics that can show the effect of a disease (like AD) on cognitive tasks such as language and problem-solving. Biological diagnosis confirms the presence of AD if both Amyloid-beta and Tau are positive in tests [5].

Using Deep Learning in AD Identification - MRI brain scans are an alternative diagnosis method, and so Deep Learning (DL) can be used to recognize the different stages of AD using techniques like object detection, or more commonly, image classification. The goal is for DL models to identify key features in brain scans, such as brain atrophy (the shrinking of brain tissue), a visual sign of AD. So, we can take scans of different levels of brain atrophy, give each stage a label/class, and feed these to the model to train on, which can then be used to recognize and classify any new data given to it.

Statistical Data

- Image Deep Learning models require datasets with vast amounts of images to properly learn features, and are split up into three different categories that align with the three phases of training models on specific image features. These categories are: training (what the model actually uses to learn the required features), validation (used to validate that what is learned was what it was supposed to learn or not), and test (to see if the model can identify the feature, now that we know it has learned the correct feature). However, it is important to note that different deep learning architectures use different techniques to learn the features in the images when the model is being trained, which is explored in the classification model section.
- Data for the deep learning model was pulled from Roboflow Universe [3], where a dataset of 6.4K images was used. The set is split 70-20-10 between train, validate, and test respectively, each containing images of 4 classes representing major stages of Alzheimer's:
- Non-demented (no AD) – no signs of cognitive decline
 - Mild-demented – minor memory lapses, mostly normal function
 - Very mild-demented – noticeable memory loss, difficulty with daily tasks
 - Moderate-demented – significant cognitive decline, requires assistance

The dataset is split into three categories: train (4480 images), validate (1280 images), and test (640 images), and is perfect, as we need the majority of images for training to get the best possible model. As for how much of each class is present in each folder, the most prominent are non-demented and very-mild demented, shown in figure 2. This could indicate that these two stages of AD are the most common, so our model being able to correctly identify them in MRI scans is crucial if it were to be used in diagnosis.

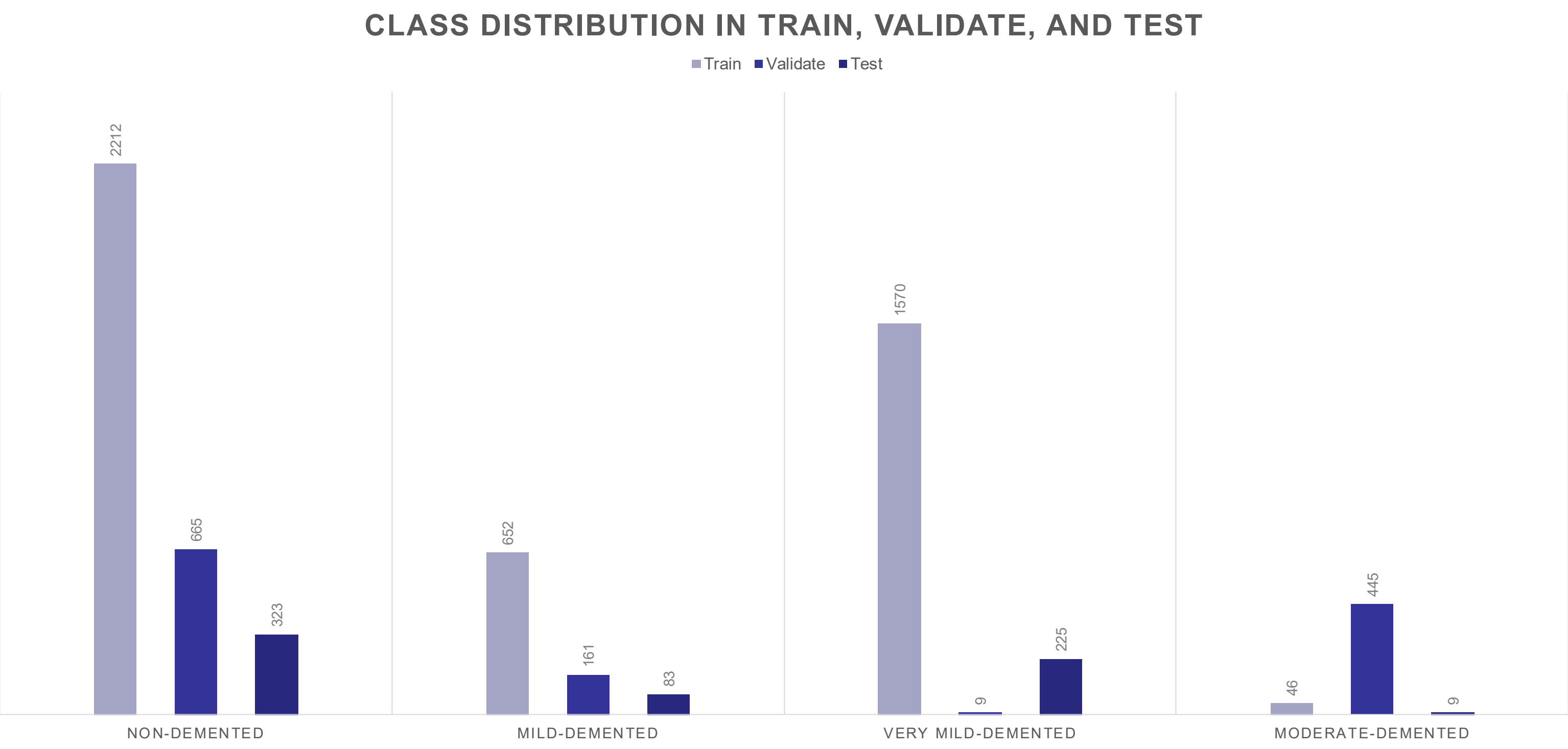


Figure 2: No. of Images for each class in train, validate, and test splits

CNN

Convolutional Neural Networks, or CNN, are a neural network architecture (late 1980s) that in classification, learns to recognize certain patterns, colors, shapes and edges as parts of a feature (e.g. the outline of a person or animal, or the particular shape of an object). CNN works by filtering the image through several layers. The first layer stores the image as a 3D Matrix, or a "tensor", with dimensions 32 x 32 x 3 for an RGB image. Then the convolution layers apply matrices called kernels to parts of the tensors and output a feature map (3b), which contain different features of the image like shapes, edges and textures that make that feature recognizable. The next few layers reduce the complexity of the feature map in order to reduce computation time, and the Fully Connected Layer uses neural networks to produce classification scores using the extracted features. The output shows the presence of the feature in the image using the classification score.

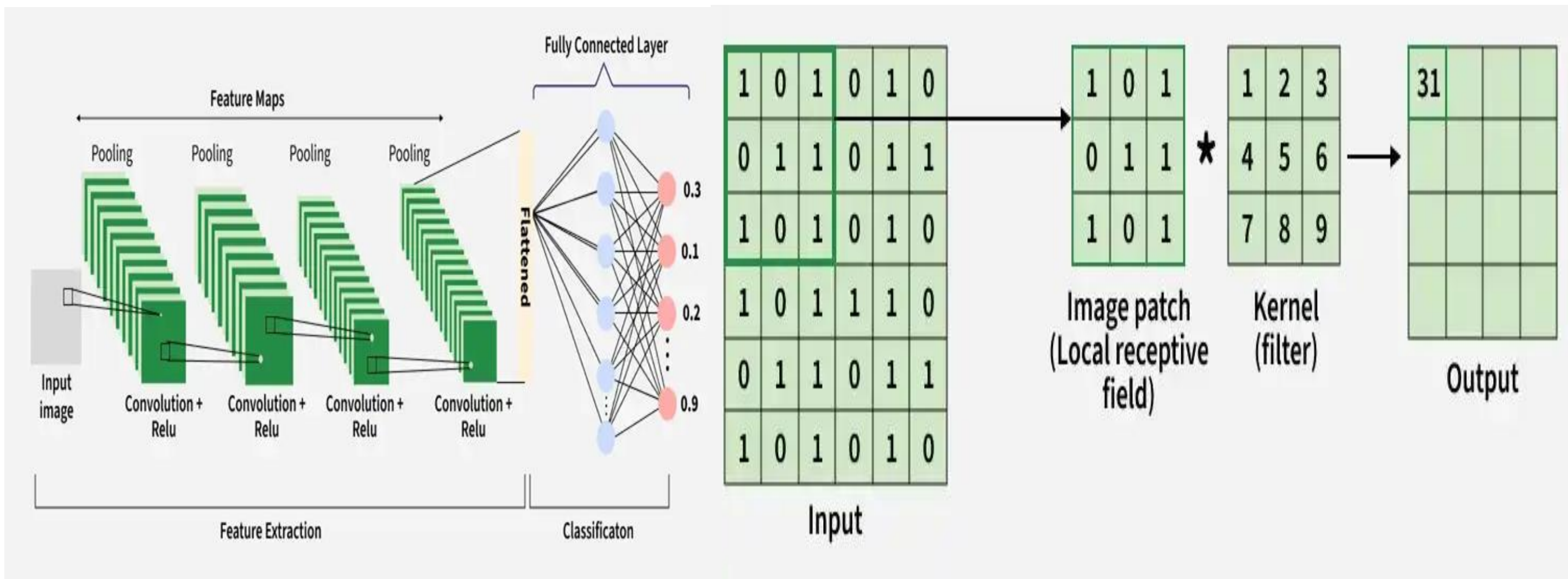


Figure 3a: CNN Diagram [1]

Figure 3b: Convolution [1]

ViT

Vision Transformers are a variant of the Transformer Architecture (2020) that is used to process images for tasks such as classification, and compared to CNNs, which focus on local features with convolution (the process shown in figure 3b), ViTs look at global features to understand what is present in the image, though they require more training than CNNs. The image is split into a certain number of equal-sized patches (3d) or "tokens", flattened into a 1D vector and then maps the vectors into a form the model can use to detect image features. Positional encoding is applied to assign the vectors to their patch location on the original image, and a token called CLS is attached in front of these ordered patch embeddings (patch sequence) that learns the features of each patch to be used for classification later on. The patch sequence is then passed through the transformer encoder, where each patch repeatedly finds relationships with each other across many layers in order to create an overall understanding of the image. The CLS token absorbs all this information and is sent to a small decision-making layer where it then outputs a class.

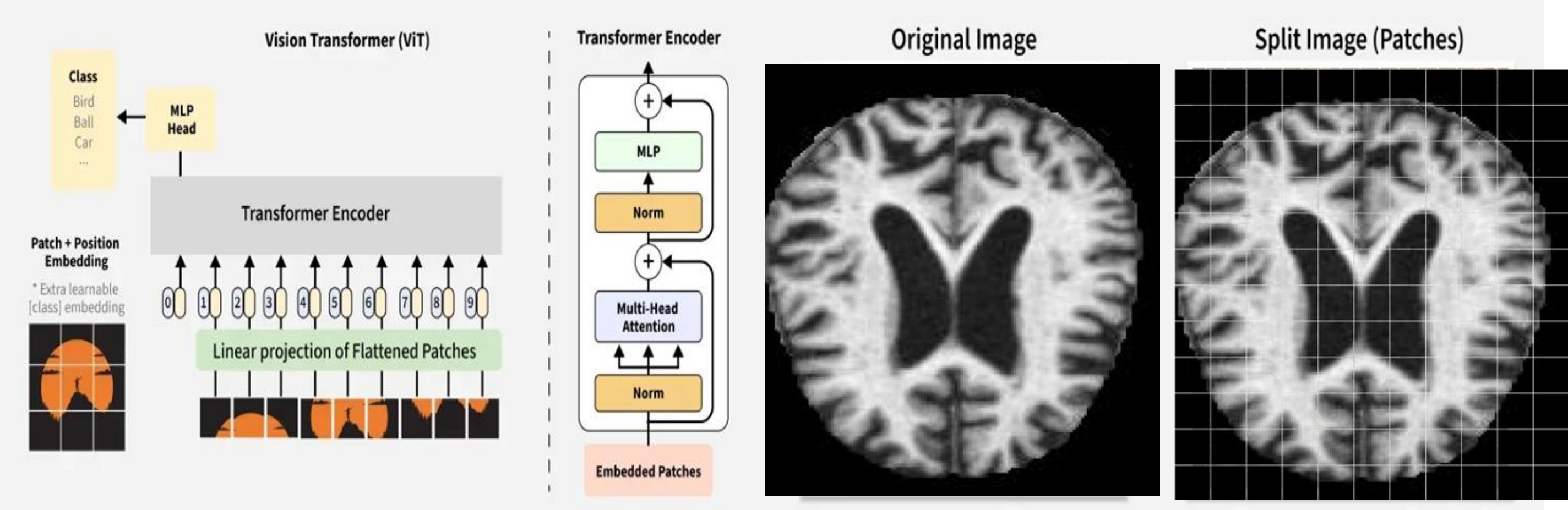


Figure 3c: ViT Diagram [2]

Figure 3d: Image Patches, Dimensions P x P x C [2]

Procedure

Model Selection

CNN: Despite being four decades old, CNN continues to improve as companies release new variants almost every year. Ultralytics in particular has been the leader in CNN Deep Learning models with YOLO (You Only Look Once), releasing around 13 versions across multiple image analysis tasks such as object detection, classification, segmentation, body tracking, and more. For our research we went with the most stable and reliable model v11, specifically the small variant which works on image datasets of around 5-10k files.

ViT: There are many models using ViT architecture, however most are very data hungry meaning they need a vast amount of images, and so will take long periods of time to train. Fortunately, Facebook has developed DeiT (Data-efficient Image Transformers), a classification model that is able to train on small datasets in a relatively small amount of time similar to YOLO.

Training Conditions:

The training was conducted on an LTU Engineering Laptop, which uses the NVIDIA RTX A2000 8GB GPU on Linux Ubuntu 24.04. Both models used the following parameters on the AD Image Dataset to produce a fair comparison and keep within the laptop's computing power. They each took approximately 1-2hr to train & validate as well as test their model on the test split of the dataset (640 images):

- Epochs (Dataset run-throughs) = 50
- Image Resolution = 384 x 384 Pixels
- Image Batch = 4

Results

Looking at the confusion matrices (figures 4a & 4b), we can clearly see that DeiT made significantly fewer errors in its predictions compared to YOLO. For example, DeiT correctly predicted 322/323 images (99.7%) as Non-Demented, while YOLO correctly predicted 306/323 (94.7%). This may not seem like a significant difference, but in reality, that wider gap in error could lead to a patient being incorrectly diagnosed with AD. The training curves (figures 4c & 4d) reflect this further. DeiT converges rapidly, reaching ~99% accuracy by epoch 30 with tightly matched loss curves, indicating strong generalization. YOLO plateaus near 67% by epoch 50 with no clear convergence, suggesting it needed more training.

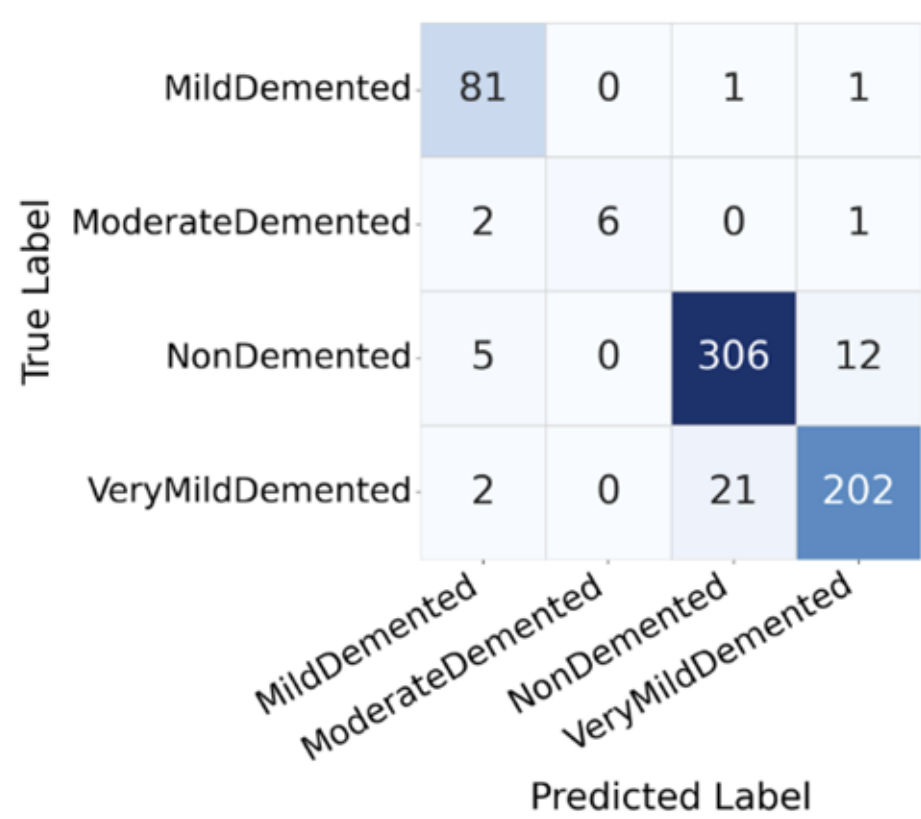


Figure 4a: YOLO Confusion Matrix



Figure 4b: DeiT Confusion Matrix

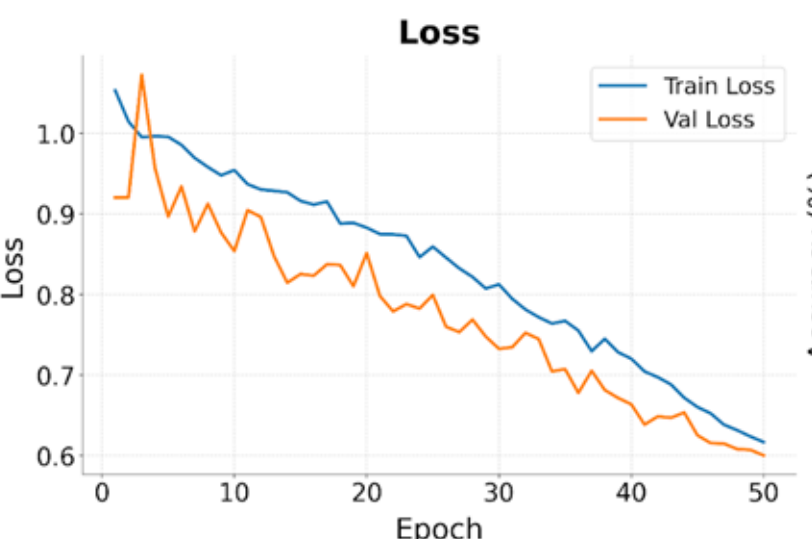


Figure 4c: YOLO Accuracy & Loss Curves

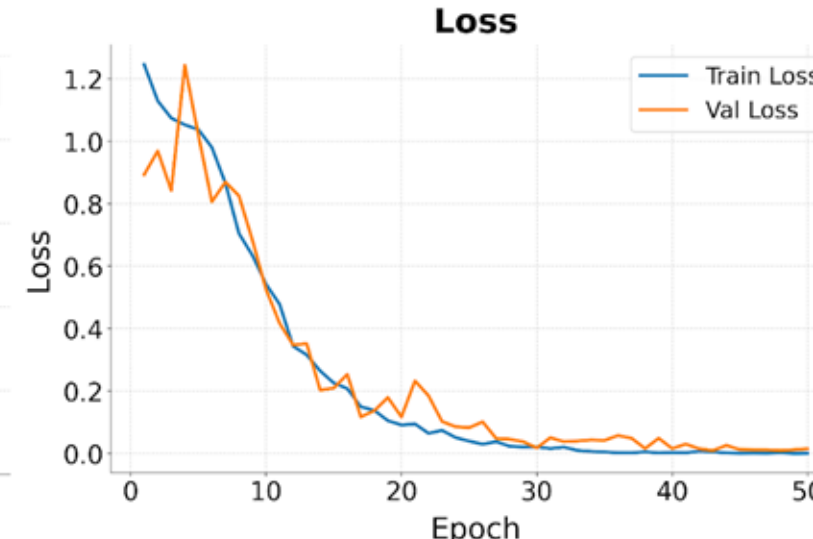


Figure 4d: DeiT Accuracy & Loss Curves

Conclusion

These results confirm our hypothesis that ViT's ability to capture long-range spatial dependencies gives it an advantage over CNNs for MRI-based AD classification. DeiT outperformed YOLO across all four classes of AD, with a mean class accuracy of 99% compared to 87%. A model that performs well can correctly diagnose the stage of AD, preventing incorrect treatment or treatment that arrives too late — both of which have real consequences for the patient. However, these results come with limitations. Both models were trained on a relatively small dataset of 6.4K images, and hardware constraints limited training to 50 epochs on a laptop GPU. These factors may have impacted YOLO's performance in particular, as its training curves showed no clear convergence by epoch 50. These models can always be improved, and so the next steps include training on a larger and more diverse dataset, using a more powerful desktop GPU, and researching newer and better-performing ViTs — with the ultimate goal of developing a reliable, automated tool to assist clinicians in early Alzheimer's diagnosis.

References

- [1] – Intro to CNN: <https://www.geeksforgeeks.org/machine-learning/introduction-convolution-neural-network/>
- [2] – ViTs: <https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/>
- [3] – Dataset: <https://universe.roboflow.com/bhara5h/alzheimers>
- [4] – AD Overview: <https://www.mdpi.com/1420-3049/25/24/5789>
- [5] – AD Diagnosis: [https://www.thelancet.com/journals/laneur/article/PIIS1474-4422\(21\)00066-1/fulltext](https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(21)00066-1/fulltext)
- [6] – AD with Deep Learning: <https://ieeexplore.ieee.org/abstract/document/10613028>