# Comparative Analysis of Convolutional and Transformer-Based Architectures for 4-Class Alzheimer's Disease Classification using MRI Data

Github

**Umar Zubair   Shahmir Qazi**

## Abstract

Automatic classification of Alzheimer's Disease (AD) stages using Magnetic Resonance Imaging (MRI) is crucial for early diagnosis and intervention. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have shown promise in this area. This paper presents a comparative study evaluating the performance of various architectures, including standard CNNs (ResNet50, MobileNetV2), a Hybrid Vision Transformer baseline, and hybrid models combining CNN backbones with Data-efficient image Transformer (DeiT) components for 4-class AD classification (Cognitively Normal, Early Mild Cognitive Impairment, Late Mild Cognitive Impairment, Alzheimer's Disease) primarily using the combination of Oasis and ADNI datasets. Our findings indicate significant challenges with using deeper CNNs like ResNet50, which exhibited overfitting and lower accuracy (approx. 75% when combined with DeiT) compared to lighter architectures. A baseline Hybrid ViT model, combining MobileNetV2 features with a Transformer, achieved high performance (87-95% accuracy on ADNI/OASIS datasets). Hybrid DeiT models showed improvements over their ResNet50 counterpart (MobileNetV2+DeiT reaching 85%), suggesting the potential of Transformer-based approaches, although they did not surpass the high-performing baseline in our experiments. This work highlights the importance of architecture selection for AD MRI classification, favoring lighter or hybrid models over standard deep CNNs like ResNet50 for the datasets used.

## 1. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that poses a significant global health challenge, characterized by cognitive decline and memory loss (Gauthier et al., 2021). Early and accurate diagnosis is paramount for timely intervention and management strategies that can potentially slow disease progression and improve patient quality of life. Neuroimaging techniques, particularly Magnetic Resonance Imaging (MRI), play a vital role by providing detailed structural information about the brain, revealing patterns of atrophy associated with different stages of AD, from Cognitively Normal (CN) through Mild Cognitive Impairment (MCI) – often subdivided into Early (EMCI) and Late (LMCI) stages – to full-blown AD (Jack Jr et al., 2010).

Automating the classification of these AD stages from MRI scans using machine learning, especially deep learning, has become an active area of research (Jo et al., 2019). Convolutional Neural Networks (CNNs) have been widely applied due to their success in general computer vision tasks. Architectures like ResNet (He et al., 2016) are often employed as standard benchmarks. However, the high dimensionality of MRI data and often limited dataset sizes in medical imaging can pose challenges, potentially leading to issues like overfitting, especially with very deep networks like ResNet50.

Recently, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) and their variants, such as the Data-efficient image Transformer (DeiT) (Touvron et al., 2021), have emerged as powerful alternatives for image classification. ViTs process images by dividing them into patches and using self-attention mechanisms to capture global dependencies, contrasting with the local receptive fields of CNNs. DeiT specifically introduced techniques, including knowledge distillation using a distillation token, to train transformers effectively even without massive pre-training datasets, making them potentially suitable for medical imaging tasks (Yuan et al., 2021). Hybrid approaches, combining the local feature extraction strengths of CNNs with the global context modeling of Transformers, have also been explored for AD classification (Li et al.; gao).

. **AUTHORERR: Missing \icmlcorrespondingauthor.**

*Proceedings of the* $38^{th}$ *International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

Despite the potential of various architectures, selecting the optimal model remains crucial. Preliminary investigations suggested that while standard deep CNNs like ResNet50 are powerful, they might struggle with the nuances of AD MRI data or require extensive datasets and regularization to avoid overfitting. This prompted a comparative study focusing on the efficacy of different models.

This paper investigates the performance of several deep learning architectures for the 4-class classification task (CN, EMCI, LMCI, AD) using the widely recognized Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. We compare a standard ResNet50, evaluate lighter CNNs like MobileNetV2 (as part of hybrid models), explore hybrid DeiT models (ResNet50+DeiT, MobileNetV2+DeiT), and benchmark against a high-performing Hybrid Vision Transformer baseline. Our primary objective is to evaluate the suitability of these models and specifically investigate the reported underperformance and overfitting issues associated with ResNet50 in this context, in order to identify more effective architectural choices for AD classification from MRI scans.

The remainder of this paper is structured as follows: Section 2 details the dataset, preprocessing steps, data augmentation techniques, and the architectures of the models evaluated. Section 3 presents the quantitative results, including accuracy, loss metrics, and confusion matrices. Section 4 interprets these results, comparing model performances and discussing potential reasons for the observed outcomes. Section 5 summarizes the findings and suggests directions for future research. Section 6 outlines the contributions of this work.

## 2. Methodology

This section describes the dataset, data preparation procedures, model architectures, and training setup used in our comparative analysis.

### 2.1. Dataset

The primary dataset used for training and evaluating most models in this study was derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see `www.adni-info.org`.

We focused on the 4-class classification task, distinguishing between Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Alzheimer's Disease (AD) subjects based on their structural MRI scans.

Additionally, our baseline ”Hybrid ViT” model, which achieved the highest performance, was trained on a larger, combined dataset comprising images from both ADNI and the Open Access Series of Imaging Studies (OASIS) dataset (`www.oasis-brains.org`), totaling approximately 30,000 images. This larger dataset likely contributed to its robust performance. For the direct comparison experiments (ResNet50+DeiT vs MobileNetV2+DeiT), models were trained and evaluated primarily on the processed ADNI subset.

### 2.2. Data Preprocessing and Augmentation

ADNI+OASIS datasets were used obtained from kaggle , a custom dataloader was made that combined images from both of them randomly upto 30,000 images. and then the models were trained , and validated on each of them sepereately. The process was also done with only a ADNI dataset as well.

Prior to model training, the MRI images underwent several preprocessing steps. Images were loaded and decoded into 3 channels (potentially by duplicating grayscale channels if necessary for compatibility with pre-trained models). All images were resized to a uniform input size of $160 \times 160$ pixels, (224,224) for DeiT. Pixel intensity values were normalized to the range [0, 1] by dividing by 255. Labels were converted to one-hot encoded vectors corresponding to the four classes (CN=0, EMCI=1, LMCI=2, AD=3).

The dataset was split into training and validation sets using an 80/20 ratio. Stratification based on the class labels was employed during the split using `sklearn.model_selection.train_test_split` to ensure that the class distribution was preserved in both sets, which is crucial for handling potentially imbalanced medical datasets.

To increase the diversity of the training data and mitigate overfitting, especially given the often limited size of medical imaging datasets, we applied online data augmentation using tf.keras.layers. The following augmentations were randomly applied to the training images during training:

- **Random Horizontal Flip:** Flips the image horizontally with a 50% probability.
- **Random Rotation:** Rotates the image by a random angle within a factor of 0.1 (corresponding to approx. ±36 degrees).
- **Random Zoom:** Zooms the image in or out by a random factor within the range [0.9, 1.1].

- **Random Contrast:** Adjusts the image contrast by a random factor within the range [0.9, 1.1].

These augmentations introduce variations in orientation, scale, and contrast, helping the model learn more robust and generalizable features. Augmentation was applied only to the training set, not the validation set. Data loading and augmentation were managed using tf.data.Dataset pipelines for efficiency, including shuffling, batching (with a batch size of 16), caching, and prefetching.

## 2.3. Model Architectures

We implemented and compared several architectures:

### 2.3.1. Baseline: Hybrid Vision Transformer

Our primary baseline was a Hybrid Vision Transformer model reported to achieve high accuracy (87-95% depending on evaluation set) on a combined ADNI+OASIS dataset. While the exact implementation details were part of earlier work, hybrid models typically leverage a CNN backbone for initial feature extraction followed by a Transformer encoder to capture global dependencies (Li et al.; gao). We hypothesize this baseline utilized MobileNetV2 as the CNN backbone due to its strong performance observed independently. Its success on the larger dataset serves as a benchmark for models trained primarily on ADNI.

### 2.3.2. ResNet50

We initially considered ResNet50 (He et al., 2016), a widely used deep CNN architecture with 50 layers featuring residual connections to facilitate training. However, initial experiments (not detailed here) indicated issues with overfitting and suboptimal performance on our dataset, aligning with our central hypothesis.

### 2.3.3. MobileNetV2

As a lighter alternative to ResNet50, MobileNetV2 (Sandler et al., 2018) uses depthwise separable convolutions, making it significantly more parameter-efficient. We used it both as a potential standalone model (leading to the high-performing baseline) and as a backbone for hybrid models.

### 2.3.4. Vision Transformer

In our initial attempt, as stated, we used a hybrid model, combining a pretrained CNN, MobileNetV2 for feature extraction and then using a ViT(Vision Transformer for our multiclass classification. This yielded an accuracy of 89-95%, providing a very strong base for the rest of our project.

### 2.3.5. Hybrid DeiT Models

To leverage the potential of transformers while mitigating data requirements, we implemented hybrid models incorporating components inspired by the Data-efficient image Transformer (DeiT) (Touvron et al., 2021). DeiT builds upon the ViT architecture but introduces a distillation token trained via knowledge distillation from a teacher model, enabling better performance with smaller datasets. Our implementation included the core ViT structure (patch embedding, positional embedding, transformer encoder blocks) and added the distillation token and head, although the complex distillation training process itself (requiring a separate teacher model during training) was simplified in the experiments reported here, focusing primarily on the architectural combination. We tested two variants:

- **ResNet50+DeiT:** Using ResNet50 (potentially truncated) as the initial feature extractor feeding into a DeiT-like transformer module.
- **MobileNetV2+DeiT:** Using MobileNetV2 as the feature extractor backbone feeding into a DeiT-like transformer module. This represents the ”improved version” whose results are presented. The architecture used a patch size of 16x16, an embedding dimension of 192, 3 attention heads, 4 transformer layers, and MLP dimension of 768, approximating a DeiT-Tiny configuration.

The DeiT model components included PatchEmbedding, TransformerBlock (with Multi-Head Attention and MLP sub-blocks), learnable class and distillation tokens, and positional embeddings, consistent with the ViT/DeiT structure.

## 2.4. Training Details

All models were implemented using TensorFlow (Abadi et al., 2016) and Keras (Chollet et al., 2015). Training utilized mixed precision (`mixed_float16`) to potentially accelerate training and reduce memory usage. The Adam optimizer (Kingma & Ba, 2014) was used with an initial learning rate (e.g., $3 \times 10^{-4}$ for the DeiT hybrids). Models were trained using Categorical Crossentropy as the primary classification loss (calculated on the CLS token output for DeiT). For the DeiT model with a distillation token, a combined loss involving both the standard cross-entropy and a distillation loss (Kullback-Leibler divergence between student's distillation token output and teacher's output, as described in (Touvron et al., 2021)) was implemented within the custom train$_{s}tep$.

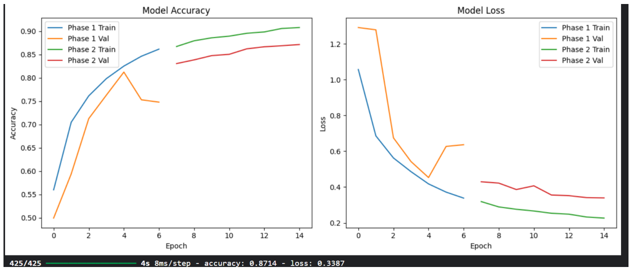
Training was performed for a set number of epochs (e.g., 7+6 epochs for the DeiT hybrids) with a batch size of 16. Callbacks were used to save the best model based on validation classification accuracy (ModelCheckpoint) and to stop training early if validation accuracy did not improve for 3 consecutive epochs (EarlyStopping), restoring the best weights found.

## 3. Results

This section presents the performance of the different models evaluated on the 4-class AD classification task using the ADNI dataset (unless otherwise specified). Key metrics include training/validation accuracy and loss curves, and confusion matrices for the best performing models.

### 3.1. Baseline Hybrid ViT Performance

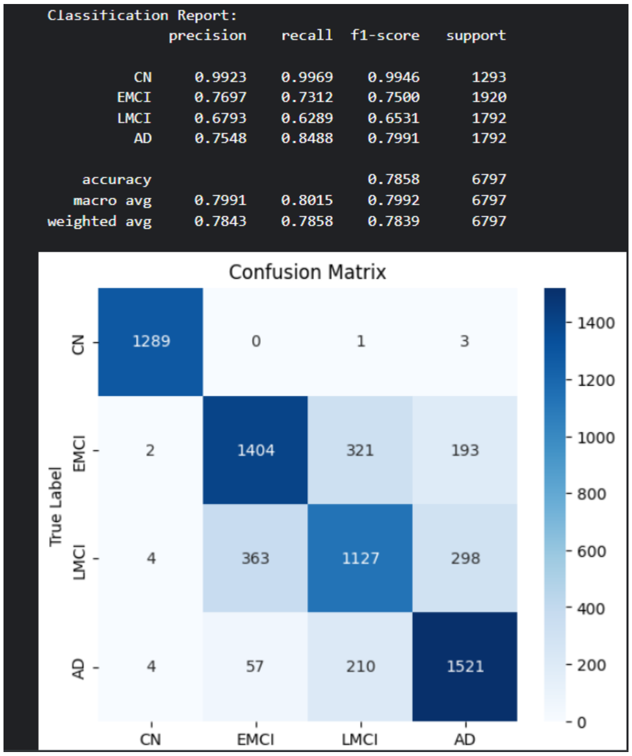
The baseline Hybrid ViT model, trained on the combined ADNI+OASIS dataset (approx. 30,000 images), demonstrated strong performance. On the ADNI validation set, it achieved an accuracy of 87.1% with a loss of 0.3387. When evaluated on the OASIS dataset partition, it achieved 95.0% accuracy with a loss of 0.1732. The classification report showed strong performance across classes, particularly for CN and AD, with a weighted average F1-score of approximately 0.878.



```
425/425 ━━━━━━━━━━━━━━━━━━━━ 4s 8ms/step - accuracy: 0.8714 - loss: 0.3387
```

### 3.2. ResNet50 Performance

As mentioned in the introduction, our initial experiments and subsequent hybrid testing involving ResNet50 indicated suboptimal performance. When combined with a DeiT head (ResNet50+DeiT architecture), the model achieved a peak validation accuracy of only around 77%. Furthermore, training curves often showed signs of overfitting, where training accuracy significantly diverged from validation accuracy, or validation performance plateaued early. This supports the premise that the complexity of ResNet50 might not be well-suited for this specific task and dataset size without significant regularization or pre-training strategies beyond the scope of this comparative study.

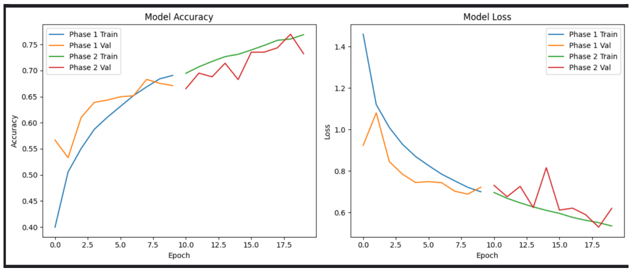


```
    Classification Report:
                  precision    recall  f1-score   support

              CN     0.9923    0.9969    0.9946      1293
            EMCI     0.7697    0.7312    0.7500      1920
            LMCI     0.6793    0.6289    0.6531      1792
              AD     0.7548    0.8488    0.7991      1792

        accuracy                         0.7858      6797
       macro avg     0.7991    0.8015    0.7992      6797
    weighted avg     0.7843    0.7858    0.7839      6797
```

Confusion Matrix

| True Label | CN | EMCI | LMCI | AD |
|---|---|---|---|---|
| CN | 1289 | 0 | 1 | 3 |
| EMCI | 2 | 1404 | 321 | 193 |
| LMCI | 4 | 363 | 1127 | 298 |
| AD | 4 | 57 | 210 | 1521 |

*Figure 1.* Performance Metrics for the Baseline Hybrid ViT Model on ADNI Validation Set and OASIS Dataset partition. Includes Classification Report and Confusion Matrix (ADNI).



Performance Metrics for the ResNet Hybrid ViT Model on ADNI Validation Set and OASIS Dataset partition.

### 3.3. Swin Transformer Performance

Direct application of Swin Transformer architecture trained from scratch resulted in poor convergence, achieving only around 20% accuracy. This highlights the known challenge of training transformers on smaller datasets without tailored strategies like those employed in DeiT or extensive domain-specific pre-training.

### 3.4. Hybrid DeiT Performance

We evaluated two hybrid DeiT models trained primarily on the ADNI dataset subset.

### 3.4.1. ResNet50 + DeiT

This configuration yielded a maximum validation accuracy of approximately 75%, showing an improvement over pure Transformer approaches but still significantly underperforming compared to the baseline and the MobileNetV2-based hybrid.

### 3.4.2. MobileNetV2 + DeiT

This hybrid model, leveraging the lighter MobileNetV2 backbone, demonstrated considerably better performance, reaching a validation accuracy of approximately 85%. The training progression for this model is shown in Figures 2 and 3. While the training accuracy continued to increase, the validation accuracy plateaued around 83-85%, indicating some degree of overfitting or limitations compared to the baseline trained on more data. Figure 4 shows the confusion matrix for this model, illustrating its performance across the four classes. Misclassifications primarily occurred between the adjacent MCI stages (EMCI and LMCI) and between LMCI and AD.

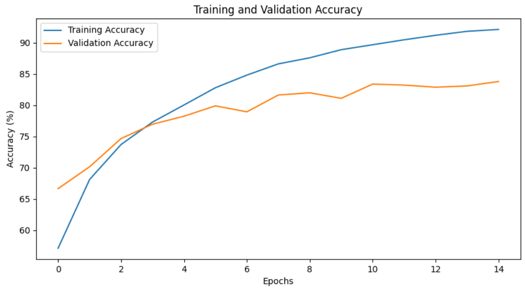


*Figure 2.* Training and Validation Accuracy curves for the MobileNetV2+DeiT Hybrid Model.

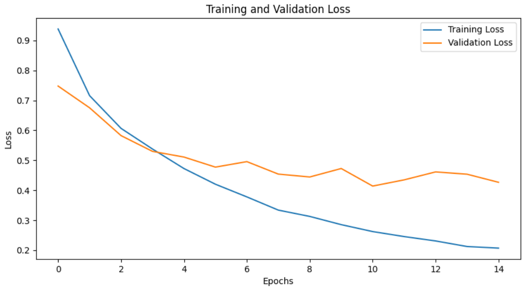


*Figure 3.* Training and Validation Loss curves for the MobileNetV2+DeiT Hybrid Model.

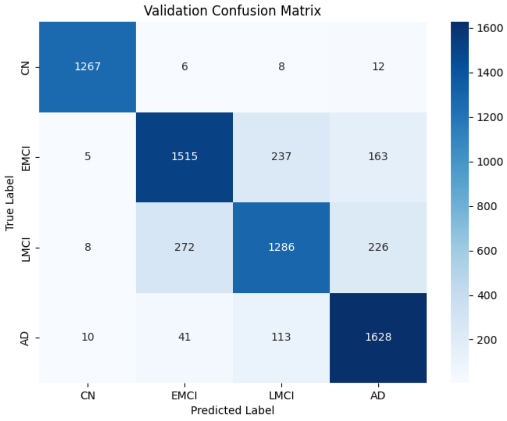


*Figure 4.* Validation Confusion Matrix for the MobileNetV2+DeiT Hybrid Model.

## 4. Discussion

The results highlight several important points regarding the application of different deep learning architectures to the 4-class classification of Alzheimer's Disease stages using ADNI MRI data.

Our primary premise – that deeper standard CNNs like ResNet50 may not be the optimal choice for this task compared to alternatives – appears to be supported by the results. The ResNet50+DeiT hybrid achieved only 75% accuracy, and initial standalone ResNet50 experiments (not fully detailed) suggested significant overfitting. This could be attributed to the high complexity and parameter count of ResNet50 relative to the size and potential subtleties of the ADNI dataset subset used for these comparisons. While ResNet architectures excel on large-scale natural image datasets like ImageNet, medical imaging tasks often present different challenges, including smaller dataset sizes and less distinct inter-class variations, where overly complex models can fail to generalize well.

In contrast, the Hybrid ViT baseline model, reportedly trained on a larger combined ADNI+OASIS dataset, achieved significantly higher accuracy (87-95%). This suggests that both the hybrid architecture (likely leveraging MobileNetV2's efficient feature extraction combined with Transformer's global context modeling, similar to (Li et al.; gao)) and the larger, more diverse training dataset contributed to its success.

The MobileNetV2+DeiT hybrid further supports the effectiveness of using a lighter CNN backbone, achieving 85% accuracy, outperforming the ResNet50+DeiT variant by 10 percentage points. This indicates that MobileNetV2's architecture is better suited for extracting relevant features from the MRI data without succumbing to the same degree of overfitting observed with ResNet50 in this context.

The inclusion of DeiT components appears beneficial compared to training standard ViTs or Swin Transformers from scratch, which failed catastrophically (2% accuracy). This aligns with the principles outlined in the DeiT paper (Touvron et al., 2021), suggesting that techniques like the distillation token (even if full distillation training wasn't the focus here) or other data-efficient adaptations can make transformers more viable for moderately sized datasets. However, the 85% accuracy of the best DeiT hybrid still fell short of the baseline Hybrid ViT. This could be due to several factors: the baseline potentially having a more optimized architecture, the significant advantage conferred by the larger combined training dataset used for the baseline, or the DeiT hybrids requiring more extensive hyperparameter tuning or longer training convergence time than allocated in our experiments (approx. 13 epochs shown).

The confusion matrix for the MobileNetV2+DeiT model (Figure 4) shows that while CN and AD classes are classified relatively well, there is considerable confusion between the intermediate stages, particularly EMCI and LMCI, and between LMCI and AD. This is expected, as these represent transitional phases of the disease continuum where neuroimaging biomarkers may overlap significantly.

Limitations of this study include the reliance on a specific preprocessing pipeline and augmentation strategy, the exploration of a limited set of architectures and hyperparameters, and the discrepancy in training data size between the baseline and the other comparative models. The 1.5-month project duration also constrained the extent of experimentation possible. Future work could involve more rigorous hyperparameter optimization, exploring different CNN backbones and Transformer variants, utilizing transfer learning more extensively, and validating results on external datasets. Investigating the full knowledge distillation training paradigm for the DeiT hybrids could also potentially bridge the performance gap with the baseline.

## 5. Conclusion

This study compared the performance of ResNet50, MobileNetV2, and hybrid DeiT-based models against a high-performing Hybrid Vision Transformer baseline for 4-class Alzheimer's Disease stage classification using ADNI MRI data. Our results indicate that the deep ResNet50 architecture, either standalone or as a backbone for a hybrid DeiT model, underperformed compared to lighter architectures, exhibiting lower accuracy (75% for ResNet50+DeiT) and potential overfitting. In contrast, the MobileNetV2+DeiT hybrid achieved a more promising accuracy of 85%. However, neither DeiT hybrid matched the performance of the Hybrid ViT baseline (87-95%), which benefited from training on a larger combined ADNI+OASIS dataset. The failure of standard ViT/Swin models trained from scratch underscores the data requirements for large transformers. We conclude that for this specific task and dataset configuration, lighter CNN architectures like MobileNetV2, especially within hybrid frameworks, offer a more effective approach than the complex ResNet50. While DeiT shows potential for improving transformer data efficiency, further optimization or larger datasets may be needed to reach state-of-the-art performance in this domain. Future work should focus on refining hybrid architectures, exploring advanced pre-training and distillation techniques, and validating on diverse datasets.

## 6. Contributions

The main contributions of this work are:

- A comparative evaluation of ResNet50, MobileNetV2, and hybrid DeiT architectures for 4-class AD classification on ADNI MRI data.
- Empirical evidence supporting the hypothesis that ResNet50 faces challenges (lower accuracy, overfitting) compared to lighter CNNs (MobileNetV2) and hybrid models in this specific application context.
- Implementation and evaluation of hybrid models combining CNN backbones (ResNet50, MobileNetV2) with DeiT-inspired transformer components.
- Benchmarking against a high-performing Hybrid ViT baseline model.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

Chollet, F. et al. Keras. `https://keras.io`, 2015.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

Gauthier, S., Rosa-Neto, P., Morais, J. A., and Webster, C. World Alzheimer report 2021: Journey through the diagnosis of dementia. Technical report, Alzheimer's Disease International (ADI), London, UK, September 2021. URL https://www.alzint.org/resource/world-alzheimer-report-2021/.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Jo, T., Nho, K., and Saykin, A. J. Deep learning for neuroimaging: a validation study. *PloS one*, 14(3):e0212724, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014.

Li, Y., Wang, K., Hou, Y., Li, F., and Yang, J.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers distillation through attention, 2021.

Yuan, L., Chen, Y., Wang, T., Wang, W., Shi, Y., Jiang, Z., Tay, F. E. H., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021.