Predicting conversion from mild cognitive impairment to Alzheimer's disease using a Vision Transformer and hippocampal MRI slices

René Seiger[1, *], Peter Fierlinger[1] for the Alzheimer's Disease Neuroimaging Initiative[**]

[1] Technical University of Munich, TUM School of Natural Sciences,

Physics Department, Garching, Germany

[*] Corresponding author

[**] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at:
https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

Convolutional neural networks (CNNs) have been the standard for computer vision tasks and are frequently applied in medical conditions, such as in Alzheimer's disease (AD). Recently, Vision Transformers (ViTs) have been introduced, which provide a strong alternative to CNNs by discarding the convolution approach in favor of the attention mechanism. This allows modeling global and distant relationships within distinct parts of an image without relying on the strong inductive biases present in CNNs. A common precursor stage of AD is a syndrome called mild cognitive impairment (MCI). However, not all individuals diagnosed with MCI progress to AD. The establishment of reliable classification models that predict converters versus non-converters would be a valuable tool to support clinical decision-making, such as enabling early treatment. Hence, in this investigation a transfer learning approach was used by applying a pretrained ViT model, fine-tuned on the ADNI dataset comprising 575 subjects with MCI. We included baseline T1-weighted structural MRI data from 299 stable MCI and 276 progressive MCI individuals, who developed Alzheimer's disease within 36 months. Inputs to the model were three normalized axial slices covering areas of the hippocampal region, consisting of the combined gray and white matter segmentations. The final model was evaluated over multiple runs to obtain stable performance estimates, yielding an average area under the receiver operating characteristic curve (AUC-ROC) on the test set of 0.74 ± 0.02 (mean ± SD), an accuracy of 0.69 ± 0.03, a sensitivity of 0.65 ± 0.07, a specificity of 0.72 ± 0.06, and a F1-score for the pMCI class of 0.67 ± 0.04. By specifically focusing on axial slices covering the hippocampal region, we aimed to target the brain structure often reported as being the first affected by the disease, while our results indicate that a ViT approach achieves reasonable classification accuracy for predicting the conversion from MCI to AD.

1. Introduction

Convolutional neural networks (CNNs) (Fukushima, 1980; LeCun et al., 1998) long served as the main architecture for visual processing tasks and are widely used for classification and prediction based on magnetic resonance imaging (MRI) data. However, transformer models utilizing the attention mechanism (Bahdanau et al., 2014) and introduced by Vaswani et al. (2017) have recently revolutionized the field of natural language processing (NLP) and have also proven valuable for vision tasks involving image data (Dosovitskiy et al., 2020). The Vision Transformer (ViT) architecture offers several advantages over classical CNN approaches, making it a formidable architecture for the medical domain, where MRI and similar imaging techniques are used to assess disease status and prognosis. While CNNs focus primarily on local or adjacent regions of an image, ViTs are capable of modeling contextualized long-range dependencies on a global scale via the self-attention mechanism. Images are divided into fixed-size patches, which are treated as input tokens, similar to how words are processed in classical Transformers used for natural language processing tasks. When trained on sufficiently large datasets, ViTs are able learn more generalized and flexible representations, while, unlike CNNs, not relying on inductive biases such as locality or translation equivariance (Dosovitskiy et al., 2020; Maurício et al., 2023).

Hence, reliable, accurate, and well-performing deep learning models are needed, as they could serve as valuable tools with the potential to support physicians in the classification and prediction of various diseases, particularly Alzheimer's disease (AD), where early diagnosis is key. AD is one of the most prevalent conditions among elderly individuals, with estimates indicating that over 30 million people worldwide are affected and numbers are steadily increasing (Gustavsson et al., 2023). To date, no cure exists, and its underlying causes remain poorly understood, although research indicates that brain atrophy patterns, which can be detected with MRI, are closely linked to, and likely result from, the pathological accumulation and spread of amyloid-beta and tau proteins (Bloom, 2014). The precursor stage of AD is referred to as mild cognitive impairment (MCI). During this phase, signs of cognitive decline particularly related to memory are already present, although activities of daily life are basically not impaired (Gauthier et al., 2006). Estimates vary, but approximately 15% of individuals with MCI progress to Alzheimer's dementia within two years, and about 33% do so within five years. Conversely, around 26% of individuals with MCI revert to normal cognition (Alzheimer's Association, 2024). It has been shown in people with a rare genetic form of AD that neuropathological changes underlying dementia are already present in the brain several years before any clinical symptoms are evident (Gordon et al., 2018). This relatively long preclinical phase makes early detection using modern deep learning algorithms particularly valuable and promising. Therefore, developing accurate classification

models to project the cognitive trajectory of individuals with MCI presents a critical opportunity to reduce patient suffering and long-term health care costs.

While ViTs have already demonstrated excellent performance in classifying AD patients from non-demented subjects based on MRI data (see Bravo-Ortiz et al., 2024; Mubonanyikuzo et al., 2025) and can be considered a viable alternative to CNNs for this task, studies evaluating their predictive power regarding AD remain scarce. To our knowledge, only one study by Hoang et al. (2023) has explored this question using structural MRI. However, following the methodological procedure reported in their investigation, data leakage likely occurred as previously noted by Valizadeh et al. (2025). Hence, results presented in their work probably represent overly optimistic accuracy metrics rather than an unbiased assessment of a ViT in predicting the conversion from MCI to AD. Further research is needed to determine whether this approach provides state of the art results, as seen in CNNs for example.

To this end, three consecutive axial brain slices covering the hippocampal region were selected as input to our model. We focused on this area located in the temporal cortex, as it is one of the earliest brain structures affected by the disease (Braak & Braak, 1991) and is strongly associated with the processing of memory-related content and memory function (Scoville & Milner, 1957). Thus, by focusing on slices encompassing this region, we aim to capture early, subtle morphological changes that may be indicative of early MCI conversion. One drawback of the ViT architecture, however, is its need for massive training data (Dosovitskiy et al., 2020), which poses a challenge in medical imaging, where available datasets typically include MRI scans from thousands of subjects at best. If not properly addressed, these models are prone to overfitting, memorizing the training data without generalizing to unseen cases. This issue can be mitigated by using pretrained models, which are already trained on millions of images and can be fine-tuned through a transfer learning approach using a specific dataset relevant to the target task. To account for this, we applied a ViT architecture which was pretrained on millions of natural images and fine-tuned it on a dataset of MRI slices for our specific task of MCI-to-AD prediction.

2. Methods

2.1. Magnetic resonance imaging data

Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were used for this investigation. As stated by ADNI, the initiative was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The original goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. The current goals include validating biomarkers for clinical trials, improving the generalizability of ADNI data by increasing diversity in the participant cohort, and to provide data concerning the diagnosis and progression of Alzheimer's disease to the scientific community. For up-to-date information, see adni.loni.usc.edu.

2.2. Demographics and MCI status

The baseline MRI scan from participants with a diagnosis of MCI in ADNI1-3 and ADNIGO was included in the analysis. Subjects were divided according to their assessments into a stable (sMCI) and a progressive (pMCI) group based on the following approach: Individuals in the sMCI group were categorized as stable if they retained this status for a minimum of three years and maintained it throughout all available follow-up assessments provided by ADNI. The pMCI group included participants who progressed from MCI at baseline to Alzheimer's disease (AD) within 36 months. Individuals who transitioned from MCI to AD and subsequently reverted to MCI were excluded. Similarly, participants who converted from MCI to a cognitively normal (CN) status, or who then reverted from CN back to MCI, were also excluded. In total, data from 575 participants were included, with 299 in the sMCI group and 276 in the pMCI group. Details regarding demographics and clinical assessment scores are provided in Table 1.

2.3. MRI data preprocessing and preparation

Structural 3D T1-weighted MRI data provided by ADNI were available either processed (for details see: https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/mri-pre-processing/) or in raw DICOM format. Available DICOM scans were transformed into NifTI format, while all data subsequently underwent the same preprocessing pipeline. The T1-weighted scans were processed using MATLAB (The MathWorks Inc., Natick, Massachusetts, USA) and the SPM12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/, version 7771). Structural images underwent bias field correction to account for intensity inhomogeneities. Gray and white matter probability maps, thresholded at 0.5, were then used as masks to extract brain tissue from the bias-corrected volumes. Normalization to MNI standard space was performed via the forward deformation fields estimated

during segmentation, resulting in one unified volume with dimensions 157 × 189 × 156 and a voxel size of 1 × 1 × 1 mm³. Three consecutive axial slices of the 3D volume of each subject centered around the hippocampal area were extracted (MNI space from z = -17 to z = -19) and assigned to the respective 2D RGB-image channels of the model for further analysis.

2.4. Model architecture – Vision Transformer (ViT)

A ViT architecture 'vit_base_patch16_224' from the TIMM (PyTorch Image Models) library was used as the backbone for our experiments. The model was pretrained on ImageNet and natively processes 224 × 224 RGB images by dividing them into 16 × 16 non-overlapping patches. The original model included a classification head consisting of a single linear layer that maps the 768-dimensional transformer output to 1000 output classes, corresponding to the ImageNet dataset. To adapt this model for our task, we modified the classification head by inserting a dropout layer (p = 0.3) before the final linear layer to reduce overfitting, and adjusted the final output to produce a single value for binary classification. The original model architecture is described in (Dosovitskiy et al., 2020), while the used ViT model with our custom adjustments are depicted in Figure 1.

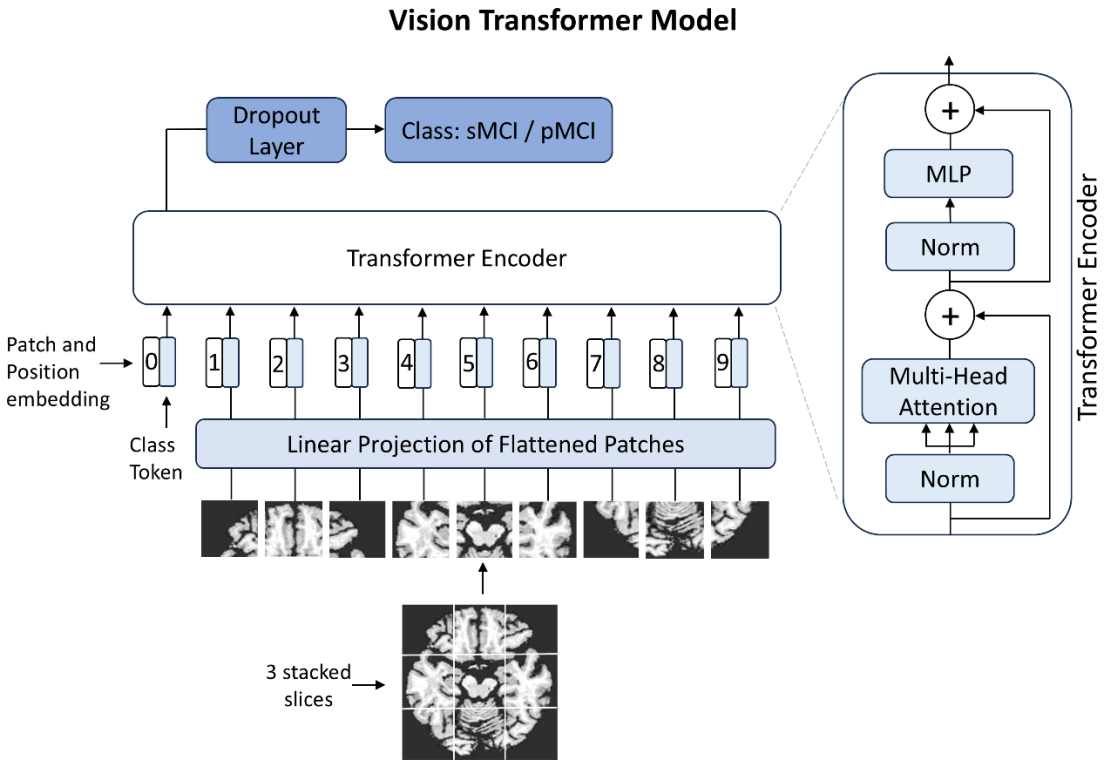

Figure 1: Model architecture featuring a customized MLP (multilayer perceptron) head with a dropout layer, adapted for binary classification. The backbone consists of 12 transformer encoder blocks and 12 multi-head attention modules, as in the pretrained ViT-base model. For illustration purposes, only 9 image patches are shown; the actual model processes 196 input patches. Architecture adapted from (Dosovitskiy et al., 2020).

The three axial MRI slices of each subject were resized to 224 × 224 pixels to match the ViTs input's dimension. To account for heterogeneity in the MRI scans, the volume, comprising the three slices of each subject, was min-max normalized [0, 1] to standardize the intensities. Subsequently the data was linearly transformed to the [-1, 1] range, by subtracting the mean of 0.5 and dividing by the standard deviation of 0.5.

2.5. Model training and fine tuning

The model was trained using PyTorch in a Google Colab environment with an NVIDIA A100 GPU (40 GB memory), 83.5 GB of system RAM. The data were split into training (70%), validation (15%), and test (15%) sets, while maintaining a constant sMCI/pMCI subject ratio across all splits. As data leakage is a serious issue frequently encountered in deep learning studies (Wen et al., 2020), data of each subject was assigned only once to one of the three folds to prevent overly optimistic and biased metrics. Data augmentation was applied during training, including random rotation (±10°) and horizontal flipping. Training was conducted using binary cross-entropy loss with logits, optimized using the AdamW optimizer (learning rate = 1e-5, weight decay = 0.05). The sigmoid function was applied to model outputs at inference time to obtain probability scores for binary classification. The model was trained for 30 epochs with a batch size of 64, and the best model was selected based on the epoch with the lowest validation loss. This process was repeated with different hyperparameter configurations as part of model tuning. The final performance of the model was evaluated on the independent test data set. A classification report was generated using the scikit-learn library, reporting the area under the receiver operating characteristic curve (ROC-AUC), accuracy, sensitivity, specificity, and F1-score.

3. Results

The comparison between the two groups regarding demographics showed no significant difference in gender distribution (Chi-square test, $p = 0.58$). Although the pMCI group was only 1.5 years older than the sMCI group, this age difference reached statistical significance (T-test, $p = 0.013$). Furthermore, both Mini-Mental State Examination (MMSE) and Clinical Dementia Rating Scale – Sum of Boxes (CDR-SOB) scores differed significantly ($p < 0.001$) between groups, with the pMCI cohort demonstrating poorer performance. Detailed results can be found in Table 1.

Table 1: Demographic information and clinical assessment scale scores. MMSE: Mini-Mental State Examination; CDR-SOB: Clinical Dementia Rating – Sum of Boxes.

Characteristic	Stable MCI	Progressive MCI	P-value
N	299	276	
Female (N, %)	120 (40.1%)	118 (42.8%)	0.5806
Male (N, %)	179 (59.9%)	158 (57.2%)	
Age (Mean ± SD)	72.59 ± 7.48	74.09 ± 7.05	0.0132
MMSE (Mean ± SD)	27.99 ± 1.69	26.75 ± 1.80	< 0.001
CDR-SOB (Mean ± SD, Range)	1.20 ± 0.66 (0.5–3.5)	1.99 ± 0.96 (0.5–5.0)	< 0.001

The distribution of conversion times of the pMCI group is depicted in Figure 2.

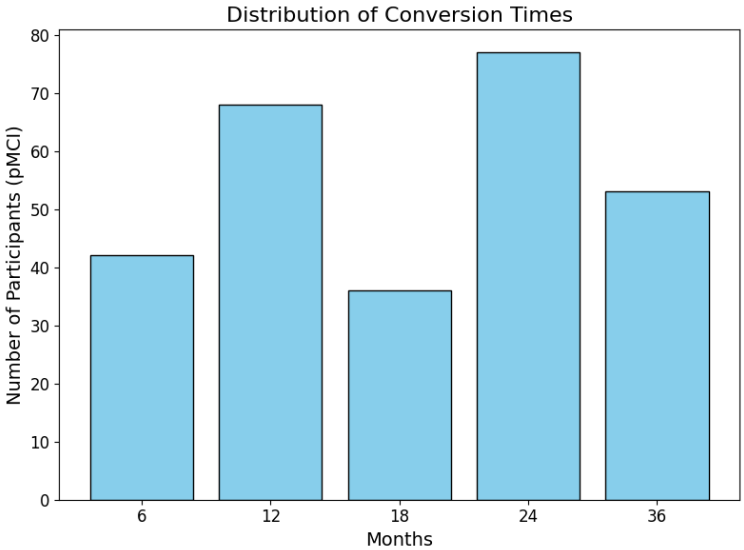

Figure 2: Overview of the distribution of conversion times in the pMCI group. Assessment frequency and intervals varied slightly between participants and across the different ADNI studies, and not all assessment periods were available in every study.

To obtain robust performance estimates and account for stochastic variability introduced by random weight initialization, data shuffling, and dropout, the model was trained and evaluated across 10 independent runs. On average, the model achieved an AUC-ROC of 0.74 with a standard deviation of 0.02 and a range of 0.72 to 0.78, indicating acceptable discriminative ability between the sMCI and pMCI groups. The mean accuracy was 0.69 (SD = 0.03; range = 0.63–0.74), sensitivity was 0.65 (SD = 0.07; range = 0.55–0.74), specificity was 0.72 (SD = 0.06; range = 0.58–0.78), and the F1-score for the pMCI class was 0.67 (SD = 0.04; range = 0.59–0.72).

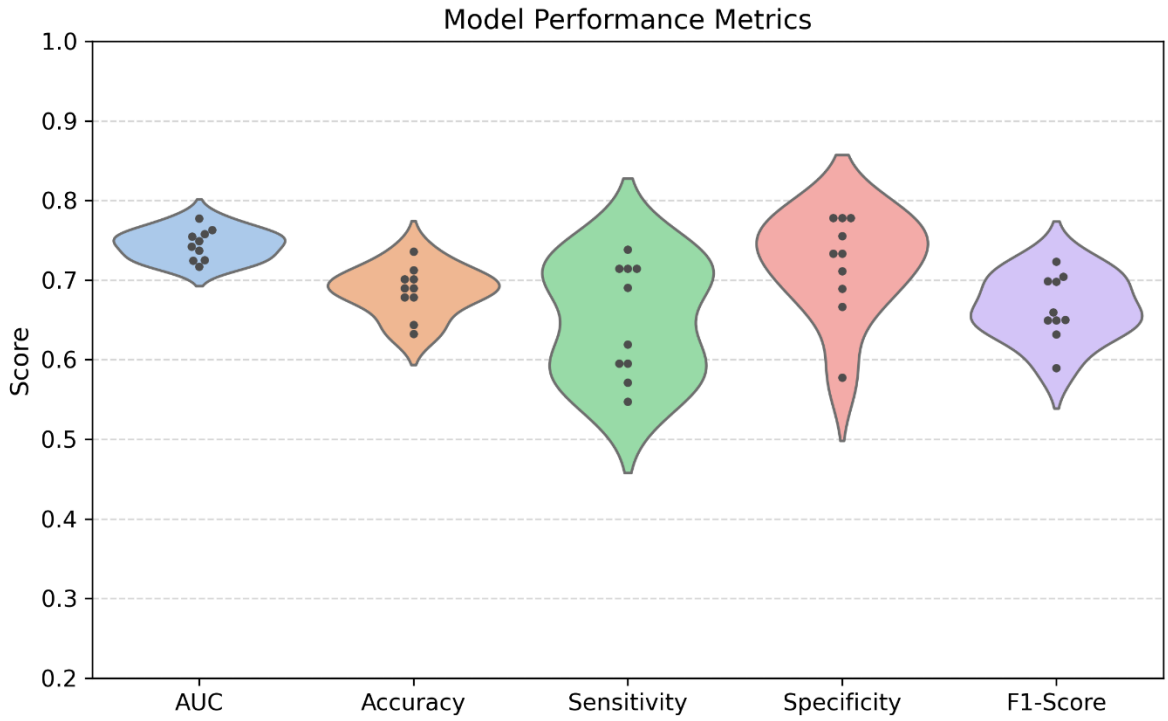

Figure 3: Distribution of model performance metrics across 10 runs of the test set (Area under the Curve (AUC), accuracy, sensitivity, specificity, and F1-score).

Figure 3 shows the distribution of these performance metrics across all runs. The mean ROC and individual ROC curves and corresponding aggregated confusion matrix are displayed in Figure 4.

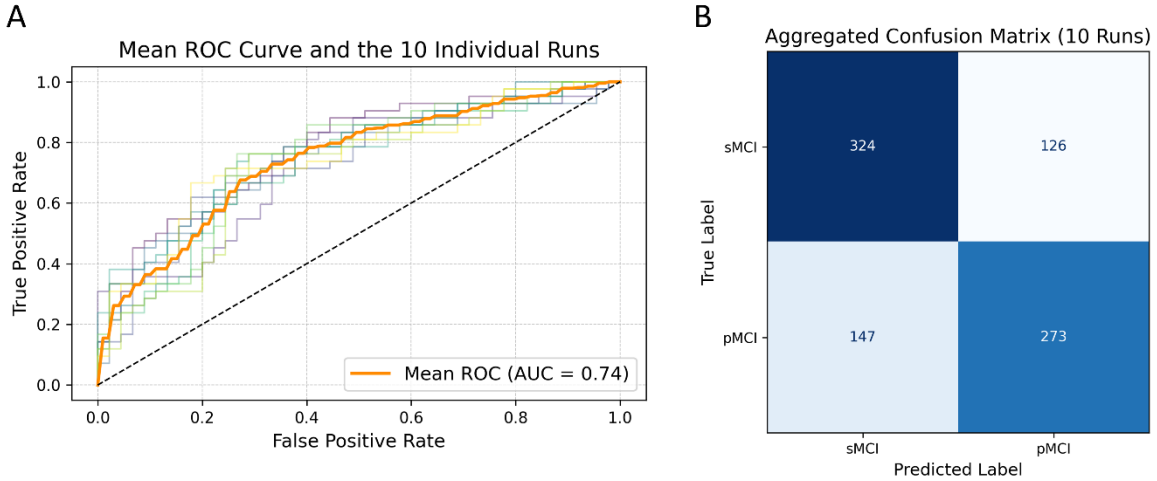

Figure 4: (A) Mean Receiver Operating Characteristic (ROC) curve (orange), including all 10 individual runs on the test dataset. Colored curves represent individual ROC curves from each run, illustrating performance variability. (B) Confusion matrix showing aggregated results from 10 independent model runs on the test dataset. Total counts of true and predicted labels for the sMCI and pMCI classes are summarized.

4. Discussion

In this investigation, we utilized axial T1-weighted MRI brain slices centered on the hippocampal region as input to a pretrained ViT model to predict conversion from MCI to AD. As the hippocampus is one of the first brain regions affected by the disease, we focused our analysis on this area to capture potential early structural changes. Specifically, we mapped three consecutive slices per subject to the three input channels of the 2D ViT model. Currently, no pretrained ViT models for 3D medical imaging trained on large-scale datasets comparable to the scale and diversity of ImageNet exist, which would provide more comprehensive representations of volumetric data and potentially deeper insights into disease progression. However, developing such models poses significant computational and data-related challenges. Hence, establishing large-scale 3D pretrained models that can be fine-tuned for specific clinical tasks remains a promising direction. ViTs represent a relatively recent advancement in computer vision and can complement traditional CNN approaches or even serve as a viable alternative, delivering results on par with CNNs or even surpassing their performance on tasks related to pathology detection (Deininger et al., 2022). Despite their advantages, such as modeling global context and long-range dependencies, ViTs are computationally expensive due to their quadratic complexity with respect to input size and large datasets are needed to reach best performance (Dosovitskiy et al., 2020).

Here, we focused on the implementation of a standard ViT based solely on MRI as input data. Only one prior study has investigated MCI to AD conversion using a pre-trained ViT model with mid-sagittal

slices (Hoang et al., 2023). An accuracy of 83% was reported in distinguishing MCI converters from non-converters. Their accuracy was higher than ours, as we achieved a mean accuracy of 69%. However, the presented steps in their investigation suggest that data from subjects were present in both the training and the validation and test sets, indicating data leakage. Thus, their reported accuracy metrics should be interpreted with caution, as they are likely overly optimistic and may not generalize well. Several studies employing architectures other than ViTs have investigated the prediction of MCI progression to AD based on structural MRI. CNNs have been utilized in works such as Bae et al. (2021), Wen et al. (2020), Oh et al. (2019), Lu et al. (2022), Ashtari-Majlan et al. (2022), Bron et al. (2021), Basaia et al. (2019), and Abrol et al. (2020). CNNs combined with attention mechanisms were explored by Zheng et al. (2023), Ren et al. (2023), Zhang et al. (2021), and W. Zhu et al. (2021). Additionally, hybrid models integrating CNNs with transformer-based architectures have been proposed, including Hu et al. (2023) with a Swin Transformer, Cao et al. (2023) with MobileViT, and Khatri et al. (2024). Some of these studies reported accuracies close to or even exceeding 90%, such as Ren et al. (2023) and Khatri et al. (2024). In particular, the latter employed a ViT as a baseline for evaluating their proposed algorithm. They trained a ViT from scratch, achieving an accuracy of 79% for the MCI-to-AD prediction task. Notably, their comparison with other models showed that a ConVit (d'Ascoli et al., 2022) reached 85% accuracy, while a novel convolution-driven ViT model introduced by the authors achieved 91%. All models in their investigation were trained from scratch rather than leveraging pretrained weights, a choice that fundamentally limits the reliability of the reported results given their small dataset size. Furthermore, insufficient transparency regarding key methodological details complicates the interpretation of the reported accuracies. Given the diversity in methodological approaches, specific techniques, and architectures across studies, a direct comparison of accuracy values does not reliably reflect actual model performance. More importantly, a specific algorithm must generalize well to new and unseen data and uses sound procedures for validation and to prevent any form of data leakage. This was systematically tested by (Wen et al., 2020), who applied various CNN architectures leading to balanced accuracy values ranging from 0.69 to 0.74 for the sMCI vs. pMCI classification task, which is comparable to our results. Future research should emphasize transparent validation protocols and explore hybrid architectures that integrate the strengths of both CNNs and transformers. Compared to other studies in the field, the ViT model proposed here falls within the range of reported metrics; however, it performs not as good as the best models discussed. As already mentioned, data leakage, particularly in studies involving slice-level data, is a well-documented issue in deep learning studies. This occurs when data from the same subject appears in both training and evaluation sets, allowing the model to memorize rather than generalize and thereby inflating performance estimates. While awareness of this problem has grown in recent years, it can still occur—

particularly in studies with limited methodological transparency. In our investigation, we explicitly addressed this issue by ensuring strict subject-level separation during data splitting to prevent any form of data leakage.

In our investigation, we used only baseline MRI scans to train and evaluate our model. This decision reflects realistic clinical scenarios, where prognosis must be based on a single timepoint rather than longitudinal information. Including scans acquired closer to or even after conversion in progressive MCI individuals would have risked biasing the model toward easier cases and inflating performance, thereby limiting generalizability. Furthermore, we specified – as done in almost all related studies – a 36-month period for classifying individuals with MCI as progressive if they changed their status to AD within this period. However, individuals with stable MCI status were only included if they extended beyond that period and remained stable in all available follow-up assessments. This ensured that we did not label individuals as stable who may convert, e.g., shortly after the 36-month period, provided that assessment data were available. A strength of this study is the well-balanced dataset, which supports the reliability of reported metrics such as accuracy and F1-score as meaningful indicators of real-world model performance. While group differences in demographics and clinical scores were observed, these must be acknowledged as potential sources of bias. Specifically, the pMCI group was slightly older and performed worse on cognitive tests, which may have marginally influenced the results. However, the age difference was only 1.5 years on average, and while statistically significant, the effect is likely minimal. Nonetheless, this should be taken into account when interpreting the findings. Furthermore, as the ADNI study spans multiple sites and started already over a decade ago, the dataset includes images acquired using different scanners, field strengths, and acquisition protocols. Although this heterogeneity introduces variability, it also reflects real-world clinical diversity and may, in turn, enhance the generalizability of the model to broader populations. In addition, while MCI and AD statuses in ADNI were clinically defined without requiring biomarker confirmation, it should be noted that recent research is adopting a more biologically driven approach, defining Alzheimer's disease as a continuum based on biomarker evidence (Jack et al., 2018).

Recent advances in model architectures, such as the combination of CNN and transformer architectures, show promising outcomes and represent a viable future direction as indicated above. New developments are already emerging, including revived state space models, in particular Mamba (Gu & Dao, 2023), which have been already applied to imaging tasks with Vision Mamba (L. Zhu et al., 2024). These models are not yet widely utilized in the medical domain but show promising potential also in the area of AD research.

In this work, we solely focused on MRI features, a choice that provides useful information while remaining relatively more accessible compared to other modalities, such as positron emission tomography (PET), where radioactive tracers are used. Furthermore, combining different input features with MRI, such as genetic and clinical data can also be seen as a viable approach (Spasov et al., 2019). Finally, we argue for the establishment of a global, standardized, unique and diverse test set, which could serve as a benchmark for future studies, allowing a fair comparison and realistic and unbiased estimation of each model's performance. This would prevent biased results and limits the occurrence of overly optimistic accuracy metric reports.

5. Conclusion

Early prediction of whether a person with MCI will convert to AD or remain stable is still one of the greatest challenges in neuroscientific research. Given recent developments in the deep learning field with rapidly evolving model architectures, detecting early signs of progression, reflected in altered brain patterns based on structural MRI, is expected to become increasingly precise. Here, a standard ViT model was used, which is thought to capture complex and early patterns of brain atrophy, and was pretrained on large datasets of millions of images. We used three axial MRI slices covering the hippocampal area as input for prediction. With our pretrained and fine-tuned model, we were able to classify sMCI from pMCI individuals with moderate accuracy. However, our results were not as strong as those reported in some other studies where CNNs were primarily applied. To date, due to computational constraints, standard ViT approaches face challenges in processing entire 3D MRI volumes, which can be seen as a drawback compared to recent developments in CNN-based methods, where full 3D inputs are already feasible. However, this may become possible in the near future, potentially enabling even better classification performance, as early brain atrophy patterns are subtle and affect multiple brain regions in a complex manner.

Acknowledgments

R. Seiger gratefully acknowledges the financial support from the Society for Research Promotion Lower Austria through an Excellence Scholarship for Research which was instrumental in conducting this study. Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

We acknowledge the support of large language models (LLMs) for proofreading, stylistic refinements and code assistance.

References

Abrol, A., Bhattarai, M., Fedorov, A., Du, Y., Plis, S., & Calhoun, V. (2020). Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease. *Journal of Neuroscience Methods*, *339*, 108701. https://doi.org/10.1016/j.jneumeth.2020.108701

Alzheimer's Association. (2024). *2024 Alzheimer's disease facts and figures* (pp. 3708–3821). https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13809

Ashtari-Majlan, M., Seifi, A., & Dehshibi, M. M. (2022). A Multi-Stream Convolutional Neural Network for Classification of Progressive MCI in Alzheimer's Disease Using Structural MRI Images. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 3918–3926. https://doi.org/10.1109/JBHI.2022.3155705

Bae, J., Stocks, J., Heywood, A., Jung, Y., Jenkins, L., Hill, V., Katsaggelos, A., Popuri, K., Rosen, H., Beg, M. F., & Wang, L. (2021). Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on a three-dimensional convolutional neural network. *Neurobiology of Aging*, *99*, 53–64. https://doi.org/10.1016/j.neurobiolaging.2020.12.005

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate* (Version 7). arXiv. https://doi.org/10.48550/ARXIV.1409.0473

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, *21*, 101645. https://doi.org/10.1016/j.nicl.2018.101645

Bloom, G. S. (2014). Amyloid-β and Tau: The Trigger and Bullet in Alzheimer Disease Pathogenesis. *JAMA Neurology*, *71*(4), 505. https://doi.org/10.1001/jamaneurol.2013.5847

Braak, H., & Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, *82*(4), 239–259. https://doi.org/10.1007/BF00308809

Bravo-Ortiz, M. A., Holguin-Garcia, S. A., Quiñones-Arredondo, S., Mora-Rubio, A., Guevara-Navarro, E., Arteaga-Arteaga, H. B., Ruz, G. A., & Tabares-Soto, R. (2024). A systematic review of vision transformers and convolutional neural networks for Alzheimer's disease classification using 3D MRI images. *Neural Computing and Applications*, *36*(35), 21985–22012. https://doi.org/10.1007/s00521-024-10420-x

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., Aalten, P., De Deyn, P. P., Biessels, G. J., Claassen, J. A. H. R., Middelkoop, H. A. M., Smits, M., Niessen, W. J., Van Swieten, J. C., Van Der Flier, W. M., Ramakers, I. H. G. B., & Van Der Lugt, A. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage: Clinical*, *31*, 102712. https://doi.org/10.1016/j.nicl.2021.102712

Cao, G., Zhang, M., Wang, Y., Zhang, J., Han, Y., Xu, X., Huang, J., & Kang, G. (2023). End-to-end automatic pathology localization for Alzheimer's disease diagnosis using structural MRI. *Computers in Biology and Medicine*, *163*, 107110. https://doi.org/10.1016/j.compbiomed.2023.107110

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., & Sagun, L. (2022). ConViT: Improving vision transformers with soft convolutional inductive biases. *Journal of Statistical Mechanics: Theory and Experiment*, *2022*(11), 114005. https://doi.org/10.1088/1742-5468/ac9830

Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., Korski, K., & Gaire, F. (2022). *A comparative study between vision transformers and CNNs in digital pathology* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2206.00389

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.11929

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., Cummings, J. L., De Leon, M., Feldman, H., Ganguli, M., Hampel, H., Scheltens, P., Tierney, M. C., Whitehouse, P., & Winblad, B. (2006). Mild cognitive impairment. *The Lancet*, *367*(9518), 1262–1270. https://doi.org/10.1016/S0140-6736(06)68542-5

Gordon, B. A., Blazey, T. M., Su, Y., Hari-Raj, A., Dincer, A., Flores, S., Christensen, J., McDade, E., Wang, G., Xiong, C., Cairns, N. J., Hassenstab, J., Marcus, D. S., Fagan, A. M., Jack, C. R., Hornbeck, R. C., Paumier, K. L., Ances, B. M., Berman, S. B., … Benzinger, T. L. S. (2018). Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: A longitudinal study. *The Lancet Neurology*, *17*(3), 241–250. https://doi.org/10.1016/S1474-4422(18)30028-0

Gu, A., & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2312.00752

Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., Krolak-Salmon, P., Rossini, P. M., Ferretti, M. T., Lanman, L., Chadha, A. S., & Van Der Flier, W. M. (2023). Global estimates on the number of persons across the Alzheimer's disease continuum. *Alzheimer's & Dementia*, *19*(2), 658–670. https://doi.org/10.1002/alz.12694

Hoang, G. M., Kim, U.-H., & Kim, J. G. (2023). Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI. *Frontiers in Aging Neuroscience*, *15*, 1102869. https://doi.org/10.3389/fnagi.2023.1102869

Hu, Z., Wang, Z., Jin, Y., & Hou, W. (2023). VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction. *Computer Methods and Programs in Biomedicine*, *229*, 107291. https://doi.org/10.1016/j.cmpb.2022.107291

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., … Silverberg, N. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

Khatri, U., Shin, S., & Kwon, G.-R. (2024). Convolution Driven Vision Transformer for the Prediction of Mild Cognitive Impairment to Alzheimer's disease Progression. *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 1–7. https://doi.org/10.1109/ICCE59016.2024.10444149

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lu, P., Hu, L., Zhang, N., Liang, H., Tian, T., & Lu, L. (2022). A Two-Stage Model for Predicting Mild Cognitive Impairment to Alzheimer's Disease Conversion. *Frontiers in Aging Neuroscience*, *14*, 826622. https://doi.org/10.3389/fnagi.2022.826622

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, *13*(9), 5521. https://doi.org/10.3390/app13095521

Mubonanyikuzo, V., Yan, H., Komolafe, T. E., Zhou, L., Wu, T., & Wang, N. (2025). Detection of Alzheimer Disease in Neuroimages Using Vision Transformers: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *27*, e62647. https://doi.org/10.2196/62647

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., & Oh, I.-S. (2019). Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning. *Scientific Reports*, *9*(1), 18150. https://doi.org/10.1038/s41598-019-54548-6

Ren, F., Yang, C., & Nanehkaran, Y. A. (2023). MRI-based model for MCI conversion using deep zero-shot transfer learning. *The Journal of Supercomputing*, *79*(2), 1182–1200. https://doi.org/10.1007/s11227-022-04668-0

Scoville, W. B., & Milner, B. (1957). LOSS OF RECENT MEMORY AFTER BILATERAL HIPPOCAMPAL LESIONS. *Journal of Neurology, Neurosurgery & Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

Spasov, S., Passamonti, L., Duggento, A., Liò, P., & Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage*, *189*, 276–287. https://doi.org/10.1016/j.neuroimage.2019.01.031

Valizadeh, G., Elahi, R., Hasankhani, Z., Rad, H. S., & Shalbaf, A. (2025). Deep Learning Approaches for Early Prediction of Conversion from MCI to AD using MRI and Clinical Data: A Systematic Review. *Archives of Computational Methods in Engineering*, *32*(2), 1229–1298. https://doi.org/10.1007/s11831-024-10176-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (Version 7). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., & Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, *63*, 101694. https://doi.org/10.1016/j.media.2020.101694

Zhang, J., Zheng, B., Gao, A., Feng, X., Liang, D., & Long, X. (2021). A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magnetic Resonance Imaging*, *78*, 119–126. https://doi.org/10.1016/j.mri.2021.02.001

Zheng, B., Gao, A., Huang, X., Li, Y., Liang, D., & Long, X. (2023). A modified 3D EfficientNet for the classification of Alzheimer's disease using structural magnetic resonance images. *IET Image Processing*, *17*(1), 77–87. https://doi.org/10.1049/ipr2.12618

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). *Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2401.09417

Zhu, W., Sun, L., Huang, J., Han, L., & Zhang, D. (2021). Dual Attention Multi-Instance Deep Learning for Alzheimer's Disease Diagnosis With Structural MRI. *IEEE Transactions on Medical Imaging*, *40*(9), 2354–2366. https://doi.org/10.1109/TMI.2021.3077079