

A 3D Convolutional Neural Network Framework for Alzheimer's Disease Diagnosis Based on Domain-Specific Transfer Learning and Classifier Chains

Mikko Lahtinen^{1*}, Elina Salo¹, Juhani Virtanen²

¹ Department of Neuroimaging and Transfer Learning, Faculty of Medicine, University of Helsinki, Helsinki, Finland.

² Department of Alzheimer's Disease Diagnostics, Faculty of Medicine, University of Eastern Finland, Kuopio, Finland.

*E-mail ✉ mikko.lahtinen@helsinki.fi

Abstract

This work explores diverse architectures for deep convolutional neural networks (CNNs). It analyzes the influence of domain-specific transfer learning on distinguishing between patients with Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and cognitively normal (CN) subjects. Training relied on 1118 3D FDG-PET brain scans acquired through the ADNI initiative. We developed standalone binary classifiers, a multiclass model, and two different chains composed of binary classifiers for sequential multiclass decision-making. Chain A started by separating cognitively normal (CN) individuals from all non-CN cases, then distinguished Alzheimer's disease (AD) from mild cognitive impairment (MCI). Chain B reversed the initial step, first identifying AD against non-AD subjects before classifying MCI versus CN. Every classifier variant was trained from scratch and with domain-specific transfer learning initialized from Med3D weights. All configurations matched or approached current state-of-the-art levels. Chain A stood out, recording 96% accuracy, 95% F1 score, and 99% AUROC. In contrast, domain-specific transfer learning often degraded performance across most models, reducing accuracy by up to 55%. Overall, the study underscores the value of binary classifier chains while prompting reconsideration of domain-specific transfer learning practices.

Keywords: Alzheimer's disease, FDG-PET, 3D CNN, Transfer learning

Introduction

Dementia has emerged as a critical public health challenge, generating substantial financial strain while ranking among the top contributors to death and long-term disability worldwide. Alzheimer's disease (AD) is responsible for the majority of cases, representing 60–80% of dementia diagnoses [1]. Major risk factors include advancing age, hereditary factors, and lifestyle

influences such as diabetes, depression, and nutritional patterns [2].

As a neurodegenerative illness, AD involves progressive neuron degeneration [3], which manifests as serious declines in cognitive abilities. However, research continues to uncover more details about its biological processes [4, 5], while effective curative treatments remain unavailable—existing interventions only moderate symptoms and may slow progression [6]. Timely identification is therefore essential, and recent advances in artificial intelligence (AI) combined with advanced brain imaging techniques offer valuable support for this goal.

Mild cognitive impairment (MCI) functions as a transitional phase between healthy cognition and dementia, often linked to AD. It encompasses early (EMCI) and late (LMCI) forms and features self-reported

Access this article online

<https://smerpub.com/>

Received: 15 August 2024; Accepted: 18 October 2024

Copyright CC BY-NC-SA 4.0

How to cite this article: Lahtinen M, Salo E, Virtanen J. A 3D Convolutional Neural Network Framework for Alzheimer's Disease Diagnosis Based on Domain-Specific Transfer Learning and Classifier Chains. J Med Sci Interdiscip Res. 2024;4(2):147-55. <https://doi.org/10.51847/mlQmYtQCjy>

or clinically observed cognitive issues, frequently alongside amyloid plaques or atrophy patterns resembling those in AD. Progression to AD is common but not inevitable [7]. Ongoing investigations aim to improve MCI categorization, predict progression, and discover biomarkers [8, 9].

Deep learning techniques have been extensively applied to brain scan classification for AD detection [10-12]. Most efforts address binary distinctions or multiclass problems involving AD, MCI, and cognitively normal (CN) groups. Transfer learning, which initializes models with weights from previously trained networks, is frequently employed to improve outcomes [13-15]. This approach adapts knowledge from one task to another; when the source model comes from the same field (medical imaging), it qualifies as domain-specific transfer learning. Additional background is provided in Section 2.

Our study compares the effectiveness of binary and multiclass 3D CNNs, as well as the sequential application of binary models known as classifier chains. We also scrutinize the relevance of transfer learning in medical domains, where appropriate pre-trained resources may be limited, with special attention to domain-specific variants. This leads to the following research questions:

RQ I: Does evidence support superior performance of binary classifiers over multiclass ones?

RQ II: Can chaining binary classifiers capitalize on their advantages for multiclass problems?

RQ III: Does domain-specific transfer learning reliably improve model results?

Similar works

Deep learning applications in medical imaging are expanding rapidly and hold considerable promise for aiding clinical diagnosis of AD and related conditions. Improvements in computational methods now facilitate full 3D brain scan processing [16]. Here we summarize key recent publications on 3D CNN approaches and transfer learning for AD classification.

Turrisi *et al.* [17] examined 3T MRI data using transfer learning from two sources: historical 1.5T MRI models (ADnet, ADnetEX) and standard 2D natural image networks (ResNet 18, 50, 101). Compared to training without pre-training, the medical-domain initialization lifted accuracy from 63% to 99%, while natural-image weights raised it from 71% to 83%.

Yuan *et al.* [18] developed a multimodal Inception-v4 system integrating 3T MRI and MMSE data from ADNI

and external datasets, and later tested it on an independent Chinese cohort. Their approach reached an AUC of 90.8%.

Bloch and Friedrich [19] evaluated DenseNet, EfficientNet, Squeeze-and-Excitation, and other networks on MRI and PET images, contrasting deep learning with classical machine learning while incorporating interpretability techniques. They transferred weights from synthetic MRI training, yielding a top external AUC of 91.74%.

Chakraborty *et al.* [20] extracted discriminative features for AD versus CN differentiation and connected them to genetic markers. Transfer learning attempts yielded only 72% AUC, so the authors preferred a bespoke CNN with expanded and augmented data.

Mora-Rubio *et al.* [21] used natural-image pre-trained models for MRI-based classification, achieving a top binary AD versus CN accuracy of 85.42%.

Soliman *et al.* [22] benchmarked several transfer learning models against a scratch-trained baseline for binary AD/CN tasks (and broader multiclass including Lewy body dementia). The strongest binary validation accuracy was 89% using ResNet50.

The work by Thakur and Snehalatha [10] aligns most closely with ours. They adapted ResNet50V2 to ADNI subsets for multiclass categorization (AD vs. LMCI vs. EMCI vs. CN), achieving an AUROC of 95%—the benchmark we consider state-of-the-art.

Drawing on these studies, binary classifiers typically outperform multiclass classifiers (except in Chakraborty *et al.* [20]). We seek to confirm this pattern (RQ I) and test systematic binary classifier chaining, which has received limited prior attention (RQ II). Transfer learning consistently provided benefits, even from distant domains, which motivates our targeted investigation into domain-specific pre-training (RQ III).

Materials and Methods

Dataset

“Data used in the preparation of this article were obtained 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 magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to

measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD)" [23].

In total, 1118 3D Fluorodeoxyglucose-Positron Emission Tomography (FDG-PET) scans in DICOM format were collected, consisting of 148 AD, 114 LMCI, 356 MCI,

210 EMCI, and 290 CN examinations. **Table 1** details the full breakdown and the assigned training, validation, and test divisions. The images came from 590 distinct patients aged 55–96 years (average 75 years), with males comprising 52% of the group.

Table 1. Overview of sample distribution in the dataset. The number of patients is indicated in brackets behind the number of samples. Percentages are rounded and do not need to sum to 100.00%.

Condition	Training	Testing	Validation	Total
AD	103 (69)	15 (10)	30 (20)	148 (99)
LMCI	80 (54)	11(8)	23 (16)	114 (78)
MCI	249 (121)	36 (17)	71 (34)	356 (172)
EMCI	147 (83)	21 (12)	42 (24)	210 (119)
CN	203 (86)	29 (12)	58 (24)	290 (122)
Total	782 (413)	112 (59)	224 (118)	1118 (590)
Percentage [%]	69.95	10.02	20.04	100

Preprocessing

The scans were initially converted from DICOM to the Neuroimaging Informatics Technology Initiative (NIfTI) [24] format using the dicom2nifti Python library [25]. Next, skull stripping was applied to isolate the brain tissue, removing extraneous elements such as skull, muscle, and other non-brain structures. This step employed HD-BET [26] to focus the data on regions relevant to the classification goal. After extraction, we

performed spatial normalization using the ants package [27], registering all brains to a common template via Symmetric Normalization. We then cropped the volumes to remove unnecessary border padding and resized them to uniform dimensions suitable for neural network input. Finally, we performed intensity normalization to enhance consistency and comparability across the dataset. **Figure 1** illustrates these preprocessing operations as part of the complete study pipeline.

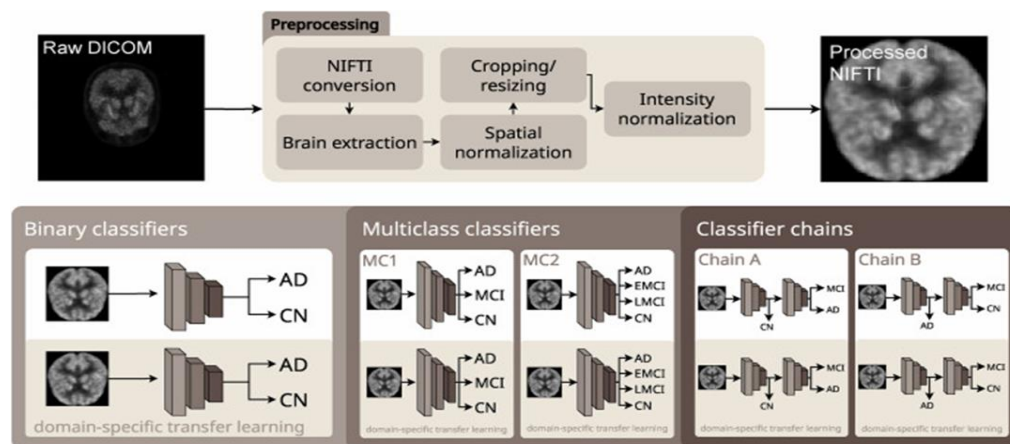


Figure 1. Overview of the study pipeline. Raw DICOM images undergo a series of preprocessing operations, including conversion to NIfTI format, brain extraction, spatial normalization, cropping, resizing, and intensity normalization. The resulting processed NIfTI volumes are then supplied as input to multiple model types: standalone binary classifiers, a multiclass classifier, and chains of binary classifiers. We tested two distinct setups for both the multiclass models and the classifier chains. We assessed all variants with and without domain-specific pre-trained weights.

Input data handling

As noted earlier, the dataset was partitioned into training (70%), testing (10%), and validation (20%) subsets. This

division relied on a custom StratifiedGroupSplit procedure. The approach created 10 randomized folds for cross-validation, ensuring scans from the same patient remained together within a single set (training, testing, or validation) to avoid data leakage. It builds upon scikit-learn's [28] `train_test_split` function, augmented with additional custom code to maintain label stratification and enforce patient-level separation. During training for each fold, a weighted random sampler was applied to promote balanced class exposure, thereby mitigating the inherent class imbalance present in the dataset.

Model implementation

All models in this study were built on the ResNet34 architecture [29] and developed using the PyTorch [30] framework and PyTorch Lightning [31]. ResNet34, introduced by He *et al.* [29], is a well-established deep convolutional neural network (CNN). Thanks to its residual shortcut connections, it achieves greater depth than many contemporary alternatives while maintaining relatively low computational complexity.

ResNet34 was selected because it serves as the foundation of the Med3D [32] project, which provides pre-trained weights tailored for medical imaging applications. These weights were trained on the 3DSeg-8 collection, which comprises eight distinct medical segmentation datasets covering organs such as the brain, heart, and spleen. Although Med3D primarily targets segmentation, the framework has been successfully extended to classification problems, making it directly applicable to the current work.

Experimental design

The experiments utilized the ResNet34-based models outlined above. **Table 2** summarizes the various configurations. For each setup, we trained both with and without loading domain-specific pre-trained weights from Med3D. We trained models for 50–75 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 4.

Table 2. Configurations of ResNet34 for the different experiments.

Experiment	Pre-trained weights	Classes
Binary classification	–	AD vs. CN
Binary classification	Med3D	AD vs. CN
Multiclass classification	–	AD vs. EMCI vs. LMCI vs. CN
	–	AD vs. MCI vs. CN
	Med3D	AD vs. EMCI vs. LMCI vs. CN
	Med3D	AD vs. MCI vs. CN
Classifier chains	–	AD vs. MCI vs. CN
	Med3D	AD vs. MCI vs. CN

Binary classification

The binary classification (BC) setup relied on the ResNet34 backbone. We modified this model to distinguish between AD and CN cases. Training involved 306 images, with 44 allocated for testing and 88 for validation.

Multiclass classification

The foundational architecture was extended to support multiple classes. We explored various setups, but report only the two highest-performing configurations: AD vs. MCI vs. CN (MC1) and AD vs. EMCI vs. LMCI vs. CN (MC2).

Classifier chain

This experiment explored the use of sequential binary classifiers to accomplish multiclass outcomes. To examine the impact of ordering, two separate chains were evaluated (**Figure 1**). Classifier chain A first separated CN images from non-CN cases; the non-CN outputs were subsequently routed to a second classifier that distinguished AD from MCI. Classifier chain B followed the opposite sequence: it first differentiated AD from non-AD images, then classified the non-AD group into CN and MCI.

Evaluation measures

Model performance was quantified using multiple standard metrics, including accuracy, F1 score, and the area under the receiver operating characteristic (ROC) curve (AUROC). To account for class imbalance, we

computed weighted averages for all metrics. When benchmarking against Thakur and Snehalatha [10], we used only accuracy and AUROC for comparison, as the authors did not report F1 score.

Perturbation analysis

We conducted a perturbation analysis to verify model robustness. Each model was tested on its original inputs but with randomly reassigned (pseudo-random) labels. We repeated this procedure 100 times per model. A model was deemed robust if it performed well on unaltered data while producing essentially random predictions under label perturbation.

Results and Discussion

In line with the methodology outlined in Section 4, we developed one binary classifier, one multiclass classifier, and two classifier chains. We assessed all variants with

and without domain-specific transfer learning. This section summarizes the outcomes of these experiments.

RQ I: binary classification vs. multiclass

Several performance indicators were used to compare the different approaches. The binary classifier (BC) delivered the strongest results overall (99% accuracy, 99% F1 score, 100% AUROC), outperforming the remaining models. This outcome aligns with expectations, given the inherent simplicity of binary tasks relative to multiclass problems.

When contrasting MC1 and MC2 with the method presented by Thakur and Snehalatha [10] (**Table 3**), our MC2 configuration achieved comparable results using the same class grouping (MC2: 93% accuracy, 92% F1 score, 99% AUROC; Thakur and Snehalatha [10]: 95 or 98% accuracy, 96% F1 score, 95% AUROC). These findings indicate that distinguishing among the various MCI stages has meaningful value, and that merging them into a single class may be suboptimal.

Table 3. Comparison of binary classification and multiclass classification approaches to Thakur and Snehalatha [10]. Noted in bold is the highest value per metric.

Model	Accuracy	F1 score	AUROC
BC: AD vs. CN	0.99	0.99	1
MC1: AD vs. MCI vs. CN	0.87	0.87	0.98
MC2: AD vs. EMCI vs. LMCI vs. CN	0.93	0.92	0.99
Thakur and Snehalatha [10]: AD vs. EMCI vs. LMCI vs. CN	0.95 or 0.98	0.96	0.95

RQ II: chaining classifiers

This section reports the outcomes from training the two classifier chains outlined previously. The initial step of distinguishing AD from non-AD cases achieved 94% accuracy, marginally surpassing the CN versus non-CN classification at 93%. However, the follow-up differentiation between AD and MCI achieved a notably higher 99% accuracy than the 91% obtained when separating CN from MCI. These patterns imply that MCI subjects show greater similarity to CN individuals in cerebral glucose metabolism.

The classifier chaining strategy was designed to enhance multiclass performance by decomposing the task into

successive binary decisions. The results confirm the effectiveness of this approach (**Table 4**). Both chains surpassed MC1's accuracy. The improvement is particularly pronounced for chain A, which attained a mean accuracy of 96% compared to MC1's 87%—a substantial 9 percentage point gain. Nevertheless, when benchmarked against MC2 or the state of the art, the importance of classifier ordering becomes evident. Chain A produced competitive results (chain a: 96% mean accuracy; MC2: 95% accuracy; Thakur and Snehalatha [10]: 95 or 98% accuracy). In contrast, chain B achieved only 88% accuracy and did not match these levels.

Table 4. Comparison of classifier chains with true multiclass approach. Noted in bold is the highest value for each metric.

Chain	Model	Accuracy	F1 score	AUROC	Mean accuracy	Mean F1 score	Mean AUROC
A	CN vs. non-CN	0.99	0.99	0.99			
	AD vs. MCI	0.92	0.91	0.98	0.96	0.95	0.99
B	AD vs. non-AD	0.95	0.95	0.99			

	CN vs. MCI	0.80	0.7	1	0.88	0.83	0.99
Multiclass (no chain)	MC1: AD vs. MCI vs. CN	0.87	0.87	0.98			

RQ III: The effect of transfer learning

This analysis centers on the impact of incorporating domain-specific pre-trained weights. Because Thakur and Snehalatha [10] did not apply domain-specific transfer learning, their reported results can only be directly compared with our models trained without such pre-trained weights.

Across all tested configurations, performance proved consistently superior when models were trained from scratch rather than initialized with Med3D pre-trained weights (**Table 5**). The smallest decline occurred in the binary classification setup, where AUROC dropped by only 1%. The largest degradation was recorded for MC2, which experienced a 64% reduction in F1 score along with decreases of 55% in accuracy and 34% in AUROC.

Table 5. Comparison of classifiers employing transfer learning and those without for all previously mentioned approaches. For the classifier chains, the table shows the mean value for each metric. Noted in bold is the highest value for each metric.

Model	Accuracy	F1 score	AUROC	Accuracy	F1 score	AUROC
	without domain-specific transfer learning			with domain-specific transfer learning		
BC	0.99	0.99	1	0.99	0.99	0.99
MC1	0.87	0.87	0.98	0.84	0.83	0.86
MC2	0.93	0.92	0.99	0.38	0.28	0.65
Chain A	0.96	0.95	0.99	0.63	0.79	0.81
Chain B	0.88	0.83	0.99	0.97	0.88	0.96
Thakur and Snehalatha [10]	0.98	0.96	0.95	–	–	–

To improve performance with pretrained weights, we tried several strategies.

To improve performance with pre-trained weights, we explored multiple optimization approaches. These included testing a range of hyperparameter settings and combinations, as well as assessing the influence of specific preprocessing steps such as skull stripping and spatial registration. Although these adjustments determined the final configuration employed for all models, further gains in the transfer learning variants proved unattainable.

We observed that classifier chain B not only declined less than the other models but also showed an overall improvement. The average performance gain for chain B across the two scenarios was 0.37. In comparison, the remaining models exhibited mean performance drops ranging from nearly 0 in the binary classifier (BC) to 0.51 in MC2. Chain A recorded a moderate average decline of

0.22. These patterns suggest that a well-ordered classifier chain is more robust than a standard multiclass model.

Robustness of the models

The perturbation analysis confirmed robustness across all models. Performance dropped sharply across all models once labels were randomly reassigned. **Table 6** provides an overview of the distribution of performance metrics over the 100 perturbation runs for each classifier. Only MC2 was included in this analysis, as it outperformed MC1 and would therefore be the preferred choice for subsequent work. For all binary models, accuracy, F1 score, and AUROC values clustered around 50%, consistent with random guessing on the altered labels. In the multiclass models, AUROC values were distributed near 50%, while accuracy and F1 scores centered around 25%, likewise indicating random behavior.

Table 6. Distribution of test performance metric values for the perturbation analysis.

Model	Classification	Accuracy	F1 score	AUROC	Accuracy	F1 score	AUROC
		Without domain-specific transfer learning			With domain-specific transfer learning		
BC	AD vs. CN	0.49	0.49	0.50	0.50	0.49	0.50
MC2	AD vs. LMCI vs. EMCI vs. MCI	0.25	0.24	0.50	0.25	0.24	0.51

Chain A	CN vs. non-CN	0.50	0.50	0.50	0.50	0.50	0.50
	AD vs. MCI	0.50	0.36	0.50	0.50	0.44	0.50
Chain B	AD vs. non-AD	0.50	0.44	0.50	0.50	0.45	0.50
	CN vs. MCI	0.50	0.50	0.50	0.50	0.50	0.50

In this work, we developed and trained several neural network architectures for both binary and multiclass classification of brain scans. We capitalized on the generally stronger performance of binary classifiers by linking them sequentially to address multiclass problems. In addition, we examined whether incorporating domain-specific pre-trained weights could further boost model effectiveness.

With respect to RQ I, the binary classifier demonstrated modestly superior results compared to the multiclass models, though the difference was not dramatic. Interestingly, the four-class configuration (MC1: AD vs. LMCI vs. EMCI vs. CN) yielded higher accuracy than the three-class configuration (MC2: AD vs. MCI vs. CN). We attribute this to dataset characteristics: MC1 used images that clinicians could confidently categorize as either early or late MCI. In contrast, MC2 encompassed a broader and more heterogeneous range of MCI presentations, making differentiation more challenging and resulting in comparatively lower performance.

Regarding RQ II, two classifier chains with differing sequences were trained. Our implementations reached performance levels comparable to, and in some cases exceeding, the current state of the art. Thakur and Snehalatha [10] provide the benchmark, using a similar multiclass network on comparable ADNI data and class definitions. In particular, classifier chain A (CN vs. non-CN followed by AD vs. MCI) outperformed the state-of-the-art model on the equivalent multiclass task. These outcomes support the idea that combining simpler binary models can enhance overall results in complex classification settings. However, the sequence of classifiers in a chain plays a critical role. The notable impact of ordering likely stems from differences in task difficulty, possibly because MCI patterns are more similar to CN than to AD in terms of glucose metabolism. This interpretation is reinforced by the higher accuracy of the AD vs. MCI classifier in chain A compared to the CN vs. MCI step in chain B.

Model robustness was verified through perturbation analysis. When presented with randomly reassigned labels, all models behaved as if they were guessing randomly. This outcome indicates that the networks

learned genuine, meaningful features linked to the true class labels rather than overfitting to the training data.

Finally, addressing RQ III, we re-evaluated all models after initializing them with domain-specific weights from Med3D [32]. Contrary to expectations that domain-relevant knowledge would be beneficial, performance declined substantially in most cases. This may occur because, despite being trained on medical data, the Med3D weights still lacked sufficient similarity to our FDG-PET scans. It remains unclear whether weights derived from other FDG-PET or general brain imaging datasets would prove more advantageous. Our findings suggest that domain alignment alone may not be adequate and that greater dataset similarity could be required for positive transfer effects.

We also observed that classifier chain B (AD vs. non-AD followed by CN vs. MCI) appeared more resilient to using these pre-trained weights. Although beyond the scope of this study, future research should explore whether classifier chains generally offer greater robustness to mismatched pre-training data than individual multiclass models.

Limitations

Several limitations should be acknowledged. First, only domain-specific pre-trained weights were available for testing. Additional experiments with more closely matched weights—ideally from models trained on brain scans or, preferably, FDG-PET brain images—would help clarify the source of the observed performance drop and whether it can be mitigated.

Second, the acquired dataset exhibited class imbalance. This was partially addressed through the weighted random sampler described in Section 4.1, which promoted balanced representation during training. Nonetheless, a perfectly balanced dataset would likely simplify the classification task and reduce potential biases.

Further work

Future efforts will involve expanded experiments on chaining binary classifiers to exploit their simplicity and strong individual performance in more intricate multiclass scenarios. We aim to determine whether this

strategy generalizes well to other problems and delivers similarly substantial gains.

Additionally, we plan to enhance the chained models using ensemble methods, such as boosting or regularization, and to add dropout layers. The latter would also provide further protection against overfitting. Given that domain-specific transfer learning did not yield the anticipated advantages, we intend to explore the underlying causes in greater depth and test whether similar patterns emerge across different medical fields or imaging modalities.

Conclusion

This study examined multiple deep CNN configurations and the influence of domain-specific transfer learning. We implemented and assessed binary and multiclass classifiers based on ResNet34, along with chains of binary classifiers designed to improve multiclass results. Classifier chain A achieved the highest performance, surpassing existing state-of-the-art approaches. Overall, the findings highlight the promise of deep CNN chains for clinical FDG-PET scans and emphasize the need for a more careful examination of transfer learning techniques in this domain.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Alzheimer's Association. 2025 Alzheimer's disease facts and figures. *Alzheimer's Dement.* 2025;21(1):e70235. doi:10.1002/alz.70235
2. Scheltens P, Blennow K, Breteler MMB, De Strooper B, Frisoni GB, Salloway S, et al. Alzheimer's disease. *Lancet.* 2016;388(10043):505-17. doi:10.1016/S0140-6736(15)01124-1
3. Reuter-Lorenz PA, Lustig C. Brain aging: reorganizing discoveries about the aging mind. *Curr Opin Neurobiol.* 2005;15(2):245-51. doi:10.1016/j.conb.2005.03.016
4. Hardy JA, Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. *Science.* 1992;256(5054):184-5. doi:10.1126/science.1566067
5. Guerrero R, Schmidt-Richberg A, Ledig C, Tong T, Wolz R, Rueckert D. Instantiated mixed effects modeling of Alzheimer's disease markers. *Neuroimage.* 2016;142:113-25. doi:10.1016/j.neuroimage.2016.06.049
6. Siemers ER, Sundell KL, Carlson C, Case M, Sethuraman G, Liu-Seifert H, et al. Phase 3 solanezumab trials: secondary outcomes in mild Alzheimer's disease patients. *Alzheimer's Dement.* 2016;12(2):110-20. doi:10.1016/j.jalz.2015.06.1893
7. Roberts R, Knopman DS. Classification and epidemiology of MCI. *Clin Geriatr Med.* 2013;29(4):753-72. doi:10.1016/j.cger.2013.07.003
8. Lee T, Guo L, Chan AS. fNIRS as a biomarker for individuals with subjective memory complaints and MCI. *Alzheimer's Dement.* 2024;20(8):5170-82. doi:10.1002/alz.13897
9. Reas ET, Shadrin A, Frei O, Motazedi E, McEvoy L, Bahrami S, et al. Improved multimodal prediction of progression from MCI to Alzheimer's disease combining genetics with quantitative brain MRI and cognitive measures. *Alzheimer's Dement.* 2023;19(11):5151-8. doi:10.1002/alz.13112
10. Thakur M, Snehalatha U. Multi-stage classification of Alzheimer's disease from 18F-FDG-PET images using deep learning techniques. *Phys Eng Sci Med.* 2022;45(4):1301-15. doi:10.1007/s13246-022-01196-2
11. Farooq A, Anwar S, Awais M, Rehman S. A deep CNN based multi-class classification of Alzheimer's disease using MRI. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST). Beijing: IEEE; 2017. p. 1-6. doi:10.1109/IST.2017.8261460
12. Li F, Cheng D, Liu M. Alzheimer's disease classification based on combination of multi-model convolutional networks. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST). Beijing: IEEE; 2017. p. 1-5. doi:10.1109/IST.2017.8261566
13. Abrol A, Bhattarai M, Fedorov A, Du Y, Plis S, Calhoun V. Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease. *J Neurosci Methods.* 2020;339:108701. doi:10.1016/j.jneumeth.2020.108701

14. Ramzan F, Khan MUG, Rehmat A, Iqbal S, Saba T, Rehman A, et al. A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks. *J Med Syst.* 2020;44(2):37. doi:10.1007/s10916-019-1475-2
15. Wu C, Guo S, Hong Y, Xiao B, Wu Y, Zhang Q, et al. Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks. *Quant Imaging Med Surg.* 2018;8(10):992-1003. doi:10.21037/qims.2018.10.17
16. Rehman Z, Awang MK, Ali G, Faheem M. Deep learning techniques for Alzheimer's disease detection in 3D imaging: a systematic review. *Health Sci Rep.* 2024;7(7):e70025. doi:10.1002/hsr.2.70025
17. Turrisi R, Pati S, Pioggia G, Tartarisco G. Adapting to evolving MRI data: a transfer learning approach for Alzheimer's disease prediction. *Neuroimage.* 2025;307:121016. doi:10.1016/j.neuroimage.2025.121016
18. Yuan Z, Qi N, Zhou Z, Ding J, Chen X, Wu J, et al. Diagnosis of Alzheimer's disease using transfer learning with multi-modal 3D Inception-v4. *Quant Imaging Med Surg.* 2025;15(2):1455-67. doi:10.21037/qims-24-1577
19. Bloch L, Friedrich CM. Systematic comparison of 3D deep learning and classical machine learning explanations for Alzheimer's disease detection. *Comput Biol Med.* 2024;170:108029. doi:10.1016/j.compbimed.2024.108029
20. Chakraborty D, Zhuang Z, Xue H, Fiecas MB, Shen X, Pan W. Deep learning-based feature extraction with MRI data in neuroimaging genetics for Alzheimer's disease. *Genes.* 2023;14(3):626. doi:10.3390/genes14030626
21. Mora-Rubio A, Bravo-Ortíz MA, Quiñones Arredondo S, Saborit Torres JM, Ruz GA, Tabares-Soto R. Classification of Alzheimer's disease stages from magnetic resonance images using deep learning. *PeerJ Comput Sci.* 2023;9:e1490. doi:10.7717/peerj-cs.1490
22. Soliman A, Chang JR, Etminani K, Bytner S, Davidsson A, Martínez-Sánchez B, et al. Adopting transfer learning for neuroimaging: a comparative analysis with a custom 3D convolution neural network model. *BMC Med Inform Decis Mak.* 2022;22(1):318. doi:10.1186/s12911-022-02054-7
23. Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI [Internet]. [cited 2025 Jun 20]. Available from: <https://adni.loni.usc.edu/>
24. National Institute of Mental Health. NIFTI [Internet]. [cited 2025 May 27]. Available from: <https://nifti.nimh.nih.gov/>
25. Brys A. dicom2nifti [software]. Version 2.4.7. 2022 [cited 2025 Jun 20]. Available from: <https://github.com/icometrix/dicom2nifti>
26. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp.* 2019;40(17):4952-64. doi:10.1002/hbm.24750
27. Avants BB. ANTsPy [software]. Version 0.3.8. 2023 [cited 2025 Jun 20]. Available from: <https://github.com/antsx/antspy>
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python [Internet]. 2012. doi:10.48550/ARXIV.1201.0490
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE; 2016. p. 770-8. doi:10.1109/CVPR.2016.90
30. PyTorch Foundation. PyTorch [software]. Version 1.13.1. 2022 [cited 2025 Jun 20]. Available from: <https://pytorch.org/>
31. Lightning AI. PyTorch Lightning [software]. Version 1.9.4. 2023 [cited 2025 Jun 20]. Available from: <https://lightning.ai/>
32. Chen S, Ma K, Zheng Y. Med3D: transfer learning for 3D medical image analysis [Internet]. 2019. doi:10.48550/ARXIV.1904.00625