Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33443

A Comparative Study of Parameter Settings for Lightweight MRI-Based Three-Class Alzheimer's Disease Classification

Yuhan Gu

Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo, China
ovitch20051208@gmail.com

Abstract. Alzheimer's disease (AD) three-class classification based on MRI remains a difficult task because the boundaries among cognitively normal (CN), mild cognitive impairment (MCI), and AD are often unclear. This study examined how different parameter settings influence lightweight MRI classification performance under limited computational resources. A public MRI dataset was used, and the original four classes were remapped into three categories. Transfer learning experiments were then carried out with several MobileNet variants. The results showed that MobileNetV2 generally performed better than MobileNetV1. Increasing the input size from 96×96 to 160×160 improved the best testing performance. And data augmentation helped generalization. The best result was obtained with MobileNetV2 160×160 0.35 using RGB input and data augmentation. It achieved a test accuracy of 60.60% and a weighted F1 score of 0.61. However, the overall performance remained moderate, and stantial confusion involving MCI was still observed. Overall, the study shows that parameter selection has a clear effect on lightweight MRI classification performance. It provides a practical reference for MRI-based Alzheimer's disease classification under limited computational resources.

Keywords: Alzheimer's disease, MRI classification, three-class classification, transfer learning

1. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the most common cause of cognitive decline, memory loss, and behavioral changes [1]. Mild Cognitive Impairment (MCI) is commonly regarded as an intermediate clinical stage between normal cognitive aging and dementia, and individuals with MCI are at elevated risk of progressing to AD [2]. Therefore, distinguishing Cognitively Normal (CN) individuals, MCI patients, and AD patients is clinically meaningful for early identification, disease monitoring, and treatment planning. Among current neuroimaging, Magnetic Resonance Imaging (MRI) is one of the most widely used modalities due to its non-invasive nature and excellent soft tissue contrast [3].

MRI-based Alzheimer's disease analysis has been studied using both conventional machine learning and deep learning methods. Traditional machine learning methods often involve extracting features from neuroimaging data, such as grey matter density or cortical thickness, followed by classifiers such as Support Vector Machine (SVM) or random forest [4]. By contrast, recent research

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33443

typically use available MRI datasets and deep learning models for automatic classification. It is worth noting that many existing studies have focused on binary classification tasks, such as AD versus CN. Compared with this method, three-class classification of CN, MCI, and AD is more clinically meaningful because this setting can better reflect the cognitive continuum observed in real-world clinical practice.

However, Alzheimer's Disease three-class classification in MRI remains highly challenging. Adjacent groups often exhibit substantial overlap in neuroanatomical patterns. In particular, the boundaries between CN and MCI, and between MCI and AD, are often blurred [5]. Because cognitive decline in Alzheimer's disease develops gradually, and MCI represents a heterogeneous intermediate stage [6]. Moreover, classification is always highly influenced by the choice of model and the settings of the parameter. Therefore, it is important to identify suitable settings under limited computational resources.

In this study, a public MRI dataset is used to investigate three-class classification of CN, MCI, and AD. This task is chosen because the CN/MCI/AD setting better reflects the clinical cognitive spectrum than simpler binary classification. The main purpose of this work is not to design a new complex model. Instead, it aims to examine how different parameter setting affect the results. The experiment mainly compares model backbone, input resolution, colour mode, and data augmentation to find a more suitable setting for MRI-based classification under limited computing resources.

2. Method

2.1. Dataset preparation

A publicly available brain MRI dataset was downloaded from Kaggle and used in this study [7]. In MRI classification, data leakage happens when images from the same subject appear in both training and testing [8]. To reduce the risk of leakage, the dataset used in this study ensures slices from the same subject are not split across both the training and testing sets. However, the dataset has a major limitation: the original class distribution was highly imbalanced.

The original dataset shown in Table 1 was organised into four categories: MildDemented, ModerateDemented, NonDemented, and VeryMildDemented. The number of samples in each category differed substantially. Especially, lacking the sample of ModerateDemented.

Table 1. Original four-class data distribution in the training and testing sets

Class	Training	Testing
MildDemented	13575	659
ModerateDemented	1039	168
NonDemented	22705	40852
VeryMildDemented	18124	4505

In this study, the original four classes were remapped into three classes. The mapping rule was defined as follows: NonDemented was mapped to CN, VeryMildDemented was mapped to MCI, and MildDemented together with ModerateDemented were merged into AD. This mapping was used to create a more practical three-class task and to reduce the class imbalance in the original dataset. It followed the severity ordering of the original labels in the dataset [9]. In the training set, CN was formed from 2000 NonDemented images, MCI from 2000 VeryMildDemented images, and AD from 1700 MildDemented plus 300 ModerateDemented images. In the testing set, CN was formed from

DOI: 10.54254/2755-2721/2026.GU33443

500 NonDemented images, MCI from 500 VeryMildDemented images, and AD from 400 MildDemented plus 100 ModerateDemented images. The selection is totally random through using a random seed.

2.2. Model architectures

This study used lightweight CNN backbones for transfer learning. Transfer learning was adopted because it provides a practical way to train image classification models under limited computational resources. In addition, it allows a more stable comparison across different model settings. MobileNetV1 and MobileNetV2 were used in the experiments [10, 11]. Different variants shown in Table 2 were compared in order to examine the effect of model version, input size, and width multiplier on classification performance. These models require fewer parameters and lower computational cost, making them suitable for experiments under limited computing resources.

Table 2. MobileNet variants used in this study

Model	Input size	Width multiplier
MobileNetV1	96 × 96	0.25
MobileNetV2	96 × 96	0.35
MobileNetV2	160 × 160	0.35
MobileNetV2	160 × 160	0.5

2.3. Experimental configuration

All experiments shared a common set of training parameters shown in Table 3.

Table 3. Common experimental settings

Setting	Value
Training cycles	20
Learning rate	0.0005
Training processor	CPU
Validation set size	20%
Batch size	32
Auto-weight classes	Off
Profile int8 model	On

At the same time, several settings were changed across experiments shown in Table 4 to compare their influence on classification performance.

Table 4. Variable settings across experiments

Exp	Size	Color	Model	Augmentation	Threshold
1	96×96	Gray	MobileNetV1 0.25	On	0.2
2	96×96	Gray	MobileNetV2 0.35	On	0.2
3	96×96	RGB	MobileNetV2 0.35	On	0.2
4	160×160	RGB	MobileNetV2 0.35	On	0.2

DOI: 10.54254/2755-2721/2026.GU33443

Table 4. (continued)

5	160×160	RGB	MobileNetV2 0.35	Off	0.2
6	160×160	RGB	MobileNetV2 0.5	On	0.2

2.4. Evaluation metrics

Model performance was evaluated using accuracy, weighted precision, weighted recall, weighted F1 score, confusion matrix, and area under the ROC curve (AUC). Accuracy was reported for validation and model testing stages. And test accuracy was used as the main indicator of final classification performance. Weighted precision and weighted recall were used to evaluate classification performance across all classes while taking class proportions into account. Weighted F1 score was considered particularly important in this study because it combines both precision and recall into a single measure. In MRI classification, MCI is always difficult to identify and distinguish, but the situation for CN is exactly the opposite. Thus, weighted F1 can better reflect the overall classification quality than accuracy. Additionally, confusion matrices were used to examine the misclassification patterns among AD, CN, and MCI while AUC was reported as an auxiliary indicator of classification performance.

3. Results and discussion

3.1. Overall results

In this study, different configurations are designed to make the comparison.

Table 5. Overall experimental results

Exp	Size	Color	Backbone	Aug	Val Acc	Test Acc	Weighted F1
1	96×96	Gray	V1-0.25	On	57.3%	54.80%	0.55
2	96×96	Gray	V2-0.35	On	63.8%	57.47%	0.58
3	96×96	RGB	V2-0.35	On	42.7%	59.87%	0.60
4	160×160	RGB	V2-0.35	On	68.4%	60.60%	0.61
5	160×160	RGB	V2-0.35	Off	68.4%	57.80%	0.58
6	160×160	RGB	V2-0.5	On	72.8%	59.20%	0.60

Table 5 summarizes the results of all six experiments. Experiment 1 was used as the baseline. It employed MobileNetV1 96*96 0.25 with grayscale images. Compared with this baseline, the modification of the configuration helps achieve better performance.

Among all experiments, the best overall testing result was obtained in Experiment 4, which used MobileNetV2 160*160 0.35 with RGB images and data augmentation on. This configuration achieved the highest test accuracy of 60.60% and the highest weighted F1 score of 0.61. By contrast, although Experiment 6 produced the highest validation accuracy of 72.8%, its test accuracy was lower (59.20%). It suggests that a larger model width did not further improve generalization. In addition, the comparison between Experiments 4 and 5 shows that removing data augmentation reduced test performance.

Overall, the results indicate three main trends. First, MobileNetV2 generally performed better than MobileNetV1. Second, increasing the input size from 96×96 to 160×160 improved the best

DOI: 10.54254/2755-2721/2026.GU33443

testing performance. Third, data augmentation contributed positively to generalization. These findings suggest that model backbone, input resolution, and augmentation strategy were all important factors in this study.

3.2. Comparison of different parameter settings

Overall, replacing MobileNetV1 with MobileNetV2 led to improvements in validation accuracy, test accuracy and weighted F1 score. However, the magnitude of this improvement was relatively limited. Backbone selection alone was not the main factor determining the final performance.

The effect of color mode and input size was more complex. For the 96×96 setting, changing the input from grayscale to RGB caused a noticeable drop in validation accuracy, while the test accuracy increased slightly. Overall, this change did not produce a clear or stable improvement in performance. A possible explanation is that MRI slices are closer to intensity-based medical images rather than natural color images. RGB input often only duplicates similar information across three channels, without introducing substantial new features. As a result, grayscale input may sometimes reduce redundancy and noise.

A different trend was observed when the input size was increased. Under the RGB setting, increasing the image size from 96×96 to 160×160 led to a clear improvement in validation accuracy, but only a limited improvement in test accuracy. This suggests that the main difficulty of the three-class classification task was not simply caused by insufficient image resolution. Instead, the difficulty appears to come more from the intrinsic overlap among the classes themselves. In other words, a higher resolution mainly improved the model's ability to fit the validation data, but did not lead to a similarly large gain in generalization.

It showed that the benefit of data augmentation was mainly reflected in test performance. When augmentation was enabled, the weighted F1 score was slightly higher which means a better balance between precision and recall. This suggests that augmentation helped improve generalization rather than only improving fitting on the validation set.

3.3. Confusion matrix analysis

The confusion matrices show that CN was usually the easiest class to classify. It had the most stable class-level performance across different experiments shown in Table 6 and Table 7. In contrast, MCI was more difficult to distinguish. It is understandable because it represents an intermediate stage between normal cognition and dementia [12].

Table 6. Test confusion matrices of the baseline and best model

Model	True class	AD	CN	MCI
Baseline (Exp 1)	AD	47.4%	12.6%	40.0%
Baseline (Exp 1)	CN	7.2%	62.6%	30.2%
Baseline (Exp 1)	MCI	29.2%	16.4%	54.4%
Best model (Exp 4)	AD	49.0%	12.4%	38.6%
Best model (Exp 4)	CN	6.8%	70.4%	22.8%
Best model (Exp 4)	MCI	19.4%	18.2%	62.4%

DOI: 10.54254/2755-2721/2026.GU33443

Table 7. Test per-class F1 scores across experiments

Exp	AD F1	CN F1	MCI F1
1	0.52	0.65	0.48
2	0.50	0.73	0.50
3	0.56	0.72	0.54
4	0.56	0.70	0.56
5	0.49	0.69	0.55
6	0.52	0.73	0.54

Another common pattern in the confusion matrices is that AD was often misclassified as MCI and some CN samples were also predicted as MCI. This shows that most confusion was concentrated around MCI. Therefore, overall accuracy alone was not enough to describe model performance, and weighted F1 score together with confusion matrices were also used in this study.

It should be noted that the testing results reported here were all under the confidence threshold of 0.2. When the threshold was increased above 0.3, the proportion of uncertain predictions rose quickly. Therefore, a lower threshold was used in the final testing stage to reduce uncertain outputs and make the comparison between experiments clearer.

Experiment 4 was selected as the final configuration in this study. Although Experiment 6 achieved higher validation accuracy, Experiment 4 produced the best test accuracy and weighted F1 score. This suggests that a moderate model size was more effective than simply increasing model capacity. Despite these improvements, the overall performance remained moderate, and confusion involving MCI was still evident. This reflects the essential difficulty of three-class MRI classification.

4. Conclusion

This study compared different parameter settings with lightweight model architectures in a three-class MRI classification. Using the public MRI dataset, this study shows that MobileNetV2 generally performs better than MobileNetV1. Higher input resolution was beneficial. Data augmentation improved generalisation. This indicates that parameter selection has a clear influence on classification performance, even when lightweight models are used. Among all experiments, the best result is Experiment 4 whose setting is 160×160, RGB, MobileNetV2 0.35, augmentation On. Although the overall performance remained moderate, there is still significant confusion between MCI and adjacent categories. Future work may further explore more diverse datasets, stricter clinical labels, and stronger model architectures.

References

[1] Jedidi, H., Jedidi, Z., & Salmon, E. (2010). La maladie d'Alzheimer. Revue Médicale de Liège, 65(3), 163–168.
[2] Lee, J. (2023). Mild cognitive impairment in relation to Alzheimer's disease: An investigation of principles, classifications, ethics, and problems. Neuroethics, 16, 16.
[3] Rashid, H. U., Utrera-Martines, M. A., Jorge, J., Martin de Moraes, P., Umar, M. N., Khan, K., & Ur Rehman, H. (2016). Cyclen-based Gd^{3+} complexes as MRI contrast agents: Relaxivity enhancement and ligand design. Bioorganic & Medicinal Chemistry, 24, 5663–5684.
[4] Thakur, A., Agarwal, S., Kumar, K., Nazir, N., Kaur, K., & Dahiya, O. (2023). A comparative study of machine learning techniques for the detection of Alzheimer's disease. In 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE.

DOI: 10.54254/2755-2721/2026.GU33443

[5] Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups. Alzheimer's & Dementia, 7(3), 270–279.
[6] Morris, J. C. (2012). Revised criteria for mild cognitive impairment may compromise the diagnosis of Alzheimer disease dementia. Archives of Neurology, 69(6), 700–708.
[7] Mohanty, S. (2026). OASIS Alzheimer's detection multi-class dataset. Kaggle. https://www.kaggle.com/datasets/shreyanmohanty/oasis-alzheimers-detection-multi-class-dataset
[8] Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., & Diciotti, S. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Scientific Reports, 11, 22544.
[9] Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies (OASIS): Longitudinal MRI data in nondemented and demented older adults. Journal of Cognitive Neuroscience, 22(12), 2677–2684.
[10] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications.
[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[12] Petersen, R. C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., & Fratiglioni, L. (2014). Mild cognitive impairment: A concept in evolution. Journal of Internal Medicine, 275(3), 214–228.