Efficient Net- B0 baseline for three class Alzheimer MRI with imbalance remedies

Subash Luitel, Goo-Rak Kwon

Abstract

Class imbalance in structural MRI can bias Alzheimer disease classifiers toward majority categories and raise the risk of clinically costly false negatives. We compare two simple and widely used remedies within the same EfficientNet-B0 transfer learning pipeline for three class slice level classification of AD, CN, and MCI. Method A uses balanced sampling each epoch with standard cross entropy. Method B uses natural sampling with class weighted cross entropy. With identical preprocessing, augmentations, label smoothing 0.05, the AdamW optimizer, and a two-phase schedule with 8 epochs of warm up followed by 10 epochs of fine tuning, we evaluate on a stratified 15% validation split that reflects natural prevalence with AD 18,440 slices, CN 26,892, and MCI 42,744. Method B attains accuracy 0.9938, macro F1 0.9938, and AD recall 0.9946, exceeding Method A with 0.9839, 0.9829, and 0.9646. AD false negatives decreased from 98% of method A to 15% of method B. The gain arises from exposure to all unique training samples each epoch and from an objective that matches the unbalanced validation distribution by reweighting errors without altering batch priors. These results support class weighted loss as a strong and architecture agnostic baseline for imbalanced AD MRI and agree with prior evidence that weighting can outperform synthetic oversampling on related Alzheimer datasets.

Keywords: Alzheimer's disease | Balanced sampling | Class imbalance | class weighted cross entropy | EfficientNet- B0 | MRI | Transfer learning

1. INTRODUCTION

Alzheimer disease (AD) is a progressive neurodegenerative disorder with major societal and clinical burden. Early and reliable detection remains challenging because structural changes can be subtle in prodromal stages, and clinical deployment also demands transparent models that clinicians can trust[1]. Recent work highlights both the promise of machine learning for Alzheimer disease and the necessity of explainibility for clinical acceptance[2]. In parallel, MRI based deep learning pipelines commonly rely on transfer learning to overcome data scarcity and computational limits, achieving strong multi class performance on dementia staging tasks[3], [4]. A persistent obstacle is class imbalance, where cognitively normal (CN) and mild cognitive impairment (MCI) typically outnumber Alzheimer disease. This imbalance biases optimization toward majority classes and increases the risk of clinically costly Alzheimer false negatives[5]. Prior studies in AD MRI have addressed the skew with data level resampling and with loss reweighting. Random oversampling and synthetic methods such as Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) are widely used to build class balanced training batches and often improve minority class sensitivity in AD datasets[6], [7]. Broader surveys in medical imaging and imbalanced learning recommend class balanced mini batches as a simple baseline and a fair comparator to loss weighting. At the same time, several

* This study was supported by research funds from Chosun University, 2025.

Manuscript : 2025.09.27
Revised : 2025.10.22
Confirmation of Publication: 2025.11.11
Corresponding Author : Goo-Rak Kwon e-mail : grkwon@chosun.ac.kr

comparisons report that weight balancing can be competitive or superior across accuracy, precision, recall, and macro F1 when evaluation remains unbalanced, which motivates a controlled head to head test under a fixed recipe[8]. Our focus is a controlled comparison of two simple and widely used imbalance remedies within the same EfficientNet-B0 transfer learning pipeline for three class MRI slice classification that covers Alzheimer disease, cognitively normal, and mild cognitive impairment. Method A uses balanced sampling each epoch with standard cross entropy, following the common practice of constructing class balanced batches in AD MRI. Method B uses natural sampling with class weighted cross entropy. We keep the validation split unbalanced to reflect natural prevalence, we quantify effects on Alzheimer recall as a clinical priority, and we analyze why loss weighting can outperform sampler based balancing by aligning the training objective with the evaluation distribution while exposing all unique samples each epoch. Under a carefully matched EfficientNet-B0 fine tuning recipe, class weighted cross entropy enhances Alzheimer sensitivity and improves macro F1 compared with balanced sampling, particularly when the validation set preserves real world class ratios. We conclude with practical guidance on when to prefer loss weighting over per epoch balanced sampling in Alzheimer MRI classification, noting the tradeoffs between stability, sensitivity, and generalizability to real world class distributions. We select EfficientNet-B0 with ImageNet pretraining because it offers a strong balance between accuracy and computational efficiency, and we keep the model, the data processing, and the training schedule identical so that the comparison of imbalance remedies is clear and fair.

2. METHODOLOGY

In this study we use EfficientNet- B0 as the backbone for Alzheimer detection. Figure 1 provides an overall methodology of our study, and the subsections present detailed description of each component, describing each stage in detail. We keep the model, data processing, and training schedule consistent so that the comparison of imbalance remedies is clear and fair. We select EfficientNet-B0, pre trained on ImageNet, because it offers a

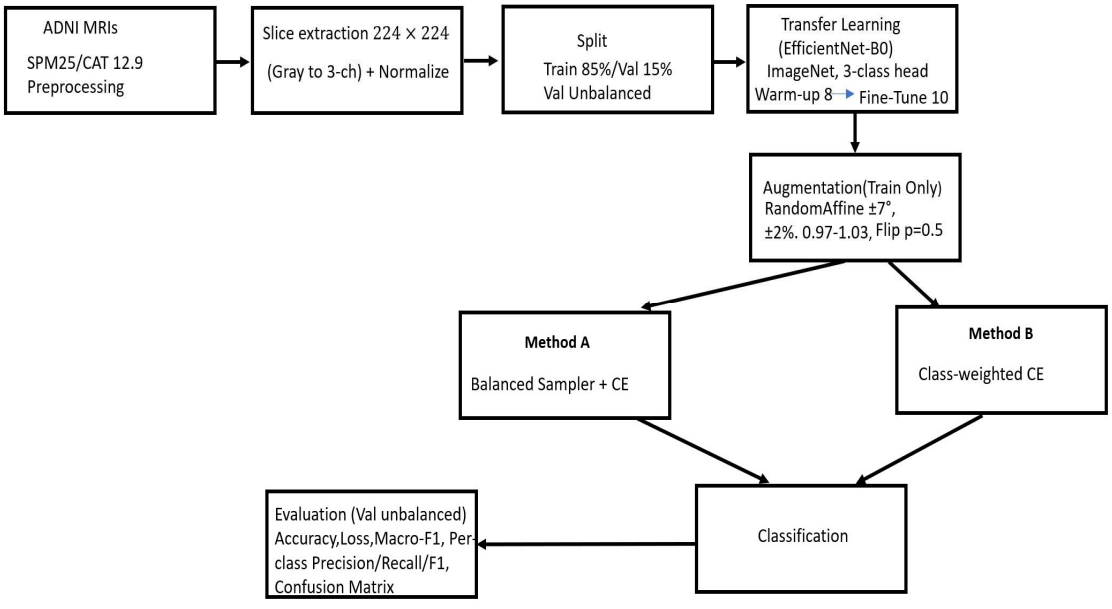

Figure 1. Overall methodology of our study

strong balance between accuracy and computational efficiency.

2.1. Data & Preprocessing

We use T1-weighted MRIs from the Alzheimer's Disease Neuroimaging Initiative cohort[9]-[11]. All volumes are preprocessed offline with Statistical Parametric Mapping (SPM25) and computational anatomy toolbox (CAT12.9). The SPM25 is available at https://github.com/spm/spm/releases and CAT12.9 at https://github.com/ChristianGaser/cat12/releases. The pipeline includes bias field correction, tissue segmentation, spatial normalization to MNI space, and skull stripping. The same upstream steps are applied to every experiment. After preprocessing, we generate a metadata table in CSV format with subject identifiers and diagnostic labels for downstream slicing. The dataset includes 18,440 slices labeled Alzheimer's disease, 26,892 labeled cognitively normal, and 42,744 labeled mild cognitive impairment.

2.2. Slice preparation and split

From every preprocessed volume we extract axial slices that pass through the hippocampal region and save each slice as a (224×224) PNG. The slices are single-channel images; to meet the network input we replicate the gray channel to form three channels. Intensities are standardized with the ImageNet mean and standard deviation. The validation and test pipelines do not resize or crop; they only convert the image to a tensor and apply the same normalization.

During training only, we apply light geometric jitter to improve robustness to small misregistration while preserving anatomy. We use RandomAffine with rotation ±7°, translation ≤2%, and scale 0.97–1.03, and we include horizontal flip with probability 0.5.

2.3. Training and transfer learning

We fine-tune EfficientNet-B0 initialized from ImageNet with a three class linear head for AD, CN, and MCI. Training uses cross entropy with label smoothing set to 0.05, the Adam with decoupled weight decay (AdamW) optimizer with weight decay 1×10^{-4}, and automatic mixed precision. First is a warm up of 8 epochs with the backbone frozen so only the classifier head is trained. Second is 10 epochs of end-to-end fine tuning with a reduced learning rate. We use Reduce on Plateau on the validation loss with factor 0.5 and patience 2 and early stopping with patience 6. The checkpoint with the minimum validation loss is restored for reporting.

2.4. Imbalance strategies compared

2.4.1. Method A (balanced sampler + cross entropy)

In this method, each training epoch draws class homogeneous batches so that mini-batches contain equal counts from AD, CN, and MCI. The loss remains standard cross entropy. Here, this data level rebalance reduces gradient bias toward majority classes and can improve minority class sensitivity, as reported in AD MRI studies using oversampling and balanced batches. The tradeoff is that minority images repeat more often, which can add duplication noise if augmentation is light or if evaluation remains unbalanced.

2.4.2. Method B (class weighted cross entropy)

In this method, natural sampling is preserved and the loss assigns a larger penalty to errors on underrepresented

classes. The weight for each class is proportional to the inverse of its frequency in the training split and the weights are normalized to sum to one. This keeps batch priors unchanged while aligning the optimization objective with the unbalanced evaluation distribution. It reduces the need to duplicate minority images and can improve Alzheimer sensitivity when validation reflects real-world prevalence. The class-weighted cross-entropy for a single sample is,

$$L_{\mathrm{WCE}}(x,y) = -\sum_{c\in\{\mathrm{AD,CN,MCI}\}} \widetilde{w_c}\; y_c \log p_c \quad (1)$$

where, w_c is the Normalized class weight for class c and given by,

$$\widetilde{w_c} = \frac{1/n_c}{(1/n_{\mathrm{AD}})+(1/n_{\mathrm{CN}})+(1/n_{\mathrm{MCI}})} \quad (2)$$

Here, x is a slice, y is one hot target, p_c is the predicted probability for class c, and n_c is the number of training samples in class c, Weights are computed once on the training split and reused for all epochs. The optimizer, schedule, and augmentations are the same as in Method A. Moreover, the data pipeline and the composition of mini-batches are unchanged and training uses natural sampling. The validation set remains unbalanced.

2.5. Classification and evaluation

Inference uses a single forward pass for each slice and the label is assigned by argmax over the class probabilities. We do not use test time augmentation and we do not apply decision thresholds. Metrics are computed on the unbalanced validation split and include accuracy, macro F1, and per class precision recall and F1. Confusion matrices are reported for error analysis.

3. Experiment result and discussion

We evaluate two imbalance remedies within the same EfficientNet-B0 pipeline. Method A uses a balanced sampler with cross entropy while Method B uses class weighted cross entropy with natural sampling. All metrics are reported on the unbalanced 15% validation split using argmax predictions and no test time augmentation.

3.1. Training dynamics

Figures 2 and 3 present loss and accuracy, respectively, across the 10 epoch fine-tuning phase for both imbalance strategies. The vertical dashed line marks the epoch with the minimum validation loss and those weights are restored for all reported metrics. Method B descends faster and reaches a lower validation loss than Method A, and maintains a slightly higher validation accuracy in late epochs. Because Reduce on Plateau lowers the learning rate when validation loss stalls and early stopping halts when improvements cease, pilot runs with more epochs did not yield a better validation checkpoint.

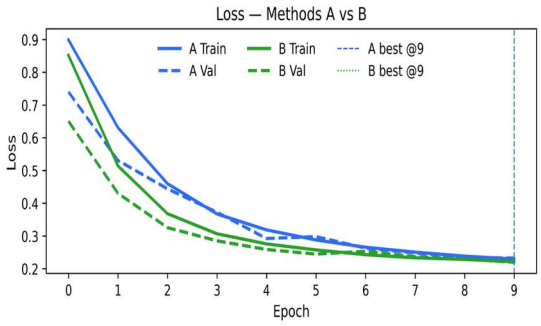

Figure 2. Loss for Methods A and B

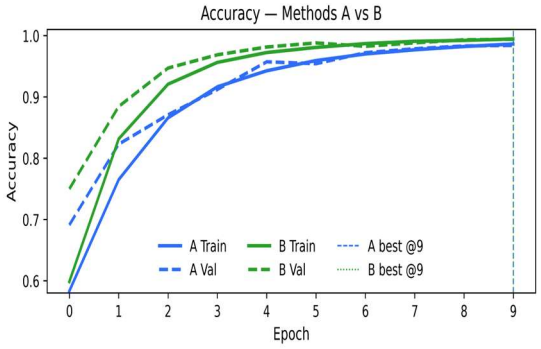

Figure 3. Accuracy for Methods A and B

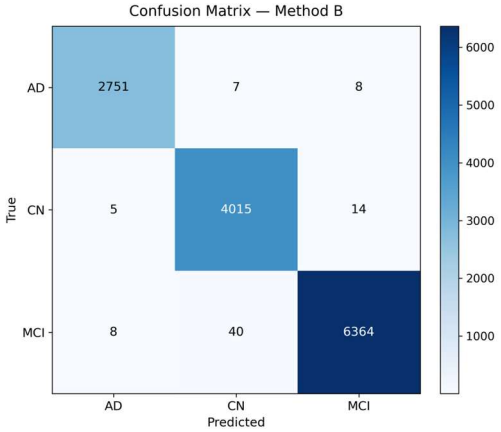

Figure 5. Confusion matrix for Method B

3.2. Error Patterns

Figures 4 and 5 show confusion matrices on the unbalanced validation set. It can be seen that under Method A, most Alzheimer errors are Alzheimer to MCI confusions, yielding 98 Alzheimer false negatives. Under Method B, Alzheimer false negatives are reduced to 15 while performance for cognitively normal and mild cognitive impairment remains high. The diagonal entries for Alzheimer strengthen visibly under Method B, indicating a more favorable operating point for clinical sensitivity.

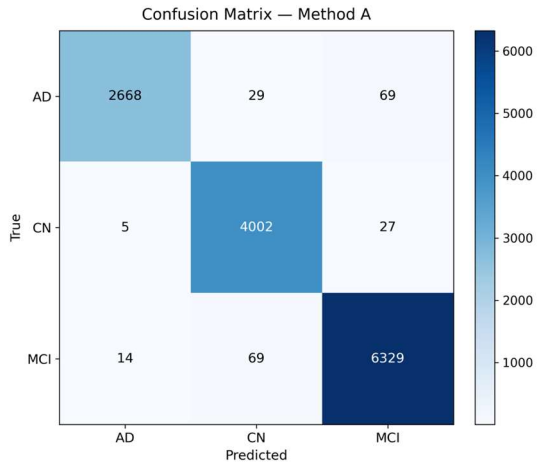

Figure 4. Confusion matrix for Method A

3.3. Comparison

Table 1 presents the comparison of the two imbalance strategies on the unbalanced validation split with argmax predictions and no test time augmentation. Method B increases accuracy from 0.9839 to 0.9938 and macro F1 from 0.9829 to 0.9938. The clinically critical Alzheimer recall rises from 0.9646 to 0.9946, reducing Alzheimer false negatives from 98 to 15. These gains mirror the lower validation loss for Method B and the improved Alzheimer diagonal in the confusion matrices. Because evaluation preserves the natural prevalence, reweighting by inverse class frequency aligns optimization with the evaluation distribution and directly improves minority class sensitivity without changing architecture or inference.

Table 1. Comparison of the two imbalance strategies (Method A vs Method B)

Metric	Method A	Method B
Accuracy	0.9839	0.9938
Loss	0.2323	0.2197
Macro-F1	0.9829	0.9938
AD Precision	0.9929	0.9953
AD Recall	0.9646	0.9946
AD F1	0.9785	0.9949

AD False Negatives	98	15
CN Precision	0.9761	0.9884
CN Recall	0.9921	0.9953
CN F1	0.9840	0.9918
CN False Negatives	32	19
MCI Precision	0.9851	0.9966
MCI Recall	0.9871	0.9925
MCI F1	0.9861	0.9945
MCI False Negatives	83	48

3.4. Experiment Environment

Experiments were run on Windows 11 with Python 3.12.3 through Anaconda. The models used PyTorch 2.7.0 with CUDA 11.8 and torchvision 0.22.0, executed on two NVIDIA TITAN RTX GPUs. Core libraries included OpenCV 4.11.0, nibabel 5.3.2, pandas 2.2.2, and matplotlib 3.9.2. Training and evaluation were carried out in Jupyter Notebook. MRI preprocessing was performed in MATLAB R2024b using SPM25 and CAT12.9.

4. Conclusion

We present a strong transfer learning baseline for 3-class AD, CN, and MCI slice classification using EfficientNet- B0, comparing two imbalance remedies under an identical recipe with 8 epochs of warm up and 10 epochs of fine tuning, light geometric augmentation, label smoothing 0.05, AdamW, and automatic mixed precision AMP. On a stratified 85/15 split with an unbalanced validation set and argmax inference without test time augmentation, class weighted cross entropy with natural sampling outperforms balanced sampling with standard cross entropy across headline metrics. Accuracy is 0.9938 versus 0.9839 and macro F1 is 0.9938 versus 0.9829. The largest gain is in AD recall, 0.9946 versus 0.9646, reducing AD false negatives from 98 to 15. Because it changes only the loss weighting and adds no architectural or inference overhead, the method is easy to adopt.

This study operates at the slice level and uses a single dataset and a single preprocessing pipeline based on SPM25 and CAT12.9. Next steps include patient level evaluation and external validation, aggregation of slices to subject level predictions, robustness checks across scanners and sites and preprocessing variants, probability calibration with analysis of operating points for clinical sensitivity and specificity, and exploration of richer encoders such as 2.5D and 3D or transformer hybrids together with broader imbalance baselines including effective number, focal loss, LDAM, and logit adjustment.

ACKNOWLEDGMENT

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database Available at: https://adni.loni.usc.edu

REFRENCES

[1] I. Malik, A. Iqbal, Y. H. Gu, and M. A. Al-antari, "Deep Learning for Alzheimer's Disease Prediction: A Comprehensive Review," Jun. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI).

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." [Online]. Available: http://gradcam.cloudcv.org

[3] M. Raghu, C. Zhang, G. Brain, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," Dec, 2019.

[4] M. Tan and Q. V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019.

[5] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif Intell Rev*, vol. 57, no. 6, Jun. 2024.

[6] E. Öter, A. Hastalığı, S. Için, V. Dengeleme, Y. Karşılaştırmalı, and B. Çalışması, "A Comparative Study on Data Balancing Methods for Alzheimer's Disease Classification," 2024.

[7] T. Perumal, N. Mustapha, R. Mohamed, and F. M. Shiri, "A Comprehensive Overview and Comparative Analysis on Deep Learning Models," *Journal on Artificial Intelligence*, vol. 6, no. 1, pp. 301–360, 2024.

[8] M. Ryspayeva and O. Salykova, "Effect of Data Balancing Methods on MRI Alzheimer's Classification," *in 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST), IEEE,* May, pp. 1–7, 2025.

[9] C. R. Jack et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," Apr. 2008.

[10] Faizaan Fazal Khan, and G.-R. Kwon, "Toward Clinically Trustworthy Alzheimer's Diagnosis: Combining EfficientNetV2B0 and XAI Techniques for MRI Analysis," *Korean Inst. Smart Media,* vol. 14, no. 6, pp. 60–66, June 2025.

[11] Vyshnavi Ramineni, and G.-R. Kwon, " A Stacked CNN Approach for Accurate Classification of AD Severity from T1-Weighted MRI Slices," *Korean Inst. Smart Media*, vol. 14, no. 10, pp. 90–97, Oct. 2025.

Authors

Subash Luitel

He has completed his Master Degree from the Department of Computer Engineering, Chosun University, in 2018. He is pursuing his Ph.D. in the Department of Information and Communication Engineering at Chosun University, Gwangju, Republic of Korea. His research interests involve Machine Learning and Medical Imaging.

Goo-Rak Kwon

He got a Ph.D. from the Department of Mechatronic Engineering, Korea University, in 2007. He has been a professor at Chosun University, since 2017. His research focus includes medical image analysis, A/V signal processing, video communication, and applications. He is a senior member of the IEEE.