Research Article

A Multi-Seed Comparative Evaluation of Class-Imbalance Mitigation Strategies For Four-Class Alzheimer's MRI Image Classification

N. Sriram[1], Dr. E.K. Girisan[2]

[1]Research Scholar, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Kovaipudur, Coimbatore – 641042, ORDIC id: 0000-0001-7574-2031, Email: sriram.4747@gmail.com
[2]Associate Professor and Head, Dept of Artificial Intelligence and Data Science, SNMV College of Arts and Science, Coimbatore – 641050, ORCID id: 0009-0003-4145-0315, Email: ekgirisan@gmail.com.

Abstract

Even though there is growing evidence that magnetic resonance imaging (MRI) can be used to classify Alzheimer's disease, the classification is still difficult to accomplish due to the unequal distribution of diagnostic classes, especially the lack of the higher stages of the disease in typical image sets. Four class-imbalance mitigation strategies are examined in a controlled fashion for the classification of Alzheimer's MRI images in four classes in this paper. Standard cross entropy loss (CE), weighted cross entropy loss (WCE), weighted random sampling (WRS), and class-balanced focal loss with logit adjustment (CBFL-LA) functions are used to test the ImageNet pretrained ResNet-50 model. In the experiments the MRI images are simply categorized as either Mild Demented, Moderate Demented, or Severely Demented. The MRI images are divided into the following categories: Mild Demented, Moderate Demented and Severely Demented in the experiments. Non-Demented, Very Mild Demented. The image-level training, validation and test splitting approach is used and each approach is tested on four random seeds with accuracy, macro-averaged F1-score and balanced accuracy. The outcomes demonstrate that the WRS has the best mean test accuracy of 96.43% ± 0.33%, and the WCE has the best mean macro-F1-score of 94.11% ± 3.31%. The best mean balanced accuracy in the study is obtained with Standard CE (93.80% ± 3.73%). In the setting evaluated, CBFL-LA's mean performance is lower than that of the other metrics. These results show the effectiveness of imbalance mitigation is related to the chosen evaluation criterion and that the significant increase of the overall accuracy is not necessarily the best measure of class-balanced performance. The study compares in an empirical way commonly used imbalance-handling strategies in a common model and common evaluation protocol. Yet the number of Moderate Demented test images is small and results from splitting the images at the level of the image lead to conclusions that are limited for minority class reliability and patients' independent generalization. The results can guide future studies on Alzheimer's MRI classification, which focus on balancing the methods for handling these imbalances.

Keywords: Magnetic Resonance Imaging (MRI), Alzheimer's disease (AD), Deep learning, ResNet-50, Class imbalance, weighted random sampling, Weighted cross-entropy, Macro-F1 score, Balanced accuracy, multi-seed evaluation.

1. Introduction

Alzheimer's disease (AD) is a progressive condition in which there is a progressive decline of memory, cognitive function and functional independence. Magnetic resonance imaging (MRI) gives a good amount of structural information to determine changes in the brain related to disease. In the literature, MRI images are classified using deep learning methods, such as convolutional neural networks (CNN), which learn discriminative features of the MRI images, and therefore there is hope that a similar approach can be used to classify MRI images. Their performance however, is highly dependent on the characteristics of the dataset, the preprocessing, the choice of the model, and the validation strategy. It is important to note that in an experiment, while good classification performance does not necessarily indicate generalization across datasets and clinical populations [1, 2]. With the assistance of transfer learning, it is possible to adapt a model that has been pre-trained on a large image collection for a new medical image classification task. A deep convolutional network with residual connections called ResNet-50 has been explored for Alzheimer's MRI classification in addition to other methods [3]. To distinguish the different categories of dementia severity, however, can be difficult as the differences between adjacent categories can be small. Hence, it is crucial to assess the classification performance by multiple measures. The other significant challenge is class imbalance, in which underrepresented classes might be underrepresented in the model. This can result in high overall accuracy but low accuracy for minority classes. It has been shown in previous neuroimaging studies that sampling strategies have an impact on the classification performance in imbalanced distributions [4]. The most widely used

mitigation methods are class-weighted loss functions, weighted sampling and focal loss; however, whether they work and which to use will vary according to the characteristics of the set and the criteria used to evaluate the results. Although recent developments, it is still necessary to make transparent comparisons between strategies for imbalance handling under comparable experimental conditions. Comparisons between studies are complicated by differences in data sets, in how data was processed, in how data was split up, and in how the data was evaluated. Additionally, there is a distinction between patient-level and image-level evaluation, as this could inflate generalizability. Four imbalance-handling strategies based on an ImageNet pretrained ResNet-50 network for classification of Alzheimer's MRIs with four classes are compared in this study, which includes standard cross-entropy, weighted cross-entropy, weighted random sampling, and class-balanced focal loss with logit adjustment. The accuracy, macro-averaged F1-score and balanced accuracy are evaluated in experiments for four random seeds. The goal is to systematically evaluate the impact of these strategies in terms of overall classification performance and class recognition with a common experimental framework.

1.1. Objective

The effectiveness of 4-class image classification of MRI images of Alzheimer's disease is examined by applying an ImageNet-pretrained ResNet-50 model, combined with 4 class-imbalance handling strategies.

1.2. Novelty

The study compares four imbalance-handling strategies with a common ResNet-50 architecture while keeping the particles of the stratified dataset fixed for each of the four strategies, and evaluates the four strategies with the aid of multiple evaluation metrics. The emphasis is not on the new architecture of the deep learning network but on comparing the predictive performance and stability.

1.3. Contributions

This study contributes in the following ways:

1. Four class-imbalance handling weighted cross-entropy (WCE), strategies – standard cross-entropy (CE), weighted random sampling (WRS), and class-balanced focal loss with logit adjustment (CBFL-LA) – are compared in terms of a shared ResNet-50 network architecture and identical dataset partitions.
2. Multi-seed assessment: To check how the model performs over repeated training runs, each strategy is assessed with four random seeds.
3. Overall classification performance is evaluated and a trade-off between majority-class and minority-class recognition by multi-metric comparison (accuracy, macro-F1 score and balanced accuracy).
4. Class-wise performance analysis: Rounding to classification metrics and confusion matrix per class are analyzed to find the difference in recognition performance across the four Alzheimer's MRI classes.

1.4. Paper Organization

The related work reviewed in Section 2 includes work on MRI classification for Alzheimer's and on class-imbalance handling techniques. Section 3 introduces the proposed framework, dataset, preprocessing approaches, imbalance handling approaches, setup of the experiment and metrics used for evaluation. The experimental results are presented and discussed in the Section 4. Lastly, Section-5 summarises the study and suggests further research.

2. Related Work

Previously published research work focused on Alzheimer's Disease MRI classifiably, class-imbalance mitigation, evaluation methods and reproducibility are reviewed. It ends with a statement of the research gap that the present study.

2.1 Deep Learning for Alzheimer's MRI Classification

Alzheimer's is an incurable dementia! The extraction of discriminative representations from brain MRI images is an important task that has been tackled using deep learning. The CNNs are used in learning hierarchical features of images and the transfer of pre-trained CNN has been used to adapt a CNN architecture to the medical image classification problem. ResNet-50 is such structure that have been studied for classification task in MRI. Amine and Mourad [3] investigated transfer learning using ResNet-50 and included imbalance-aware training and participant-level validation in their experiments. Their work shows that it is important to take into account the classification performance of the algorithm, not only its accuracy but also the methodology used to validate it.

Recent studies have shown that MRI classification can be extended beyond traditional CNN models. Anbu Devi and Suganthi [5] presented a CNN-based network for structural MRI classification, which used a hybrid class imbalance

approach: SMOTE–Edited Nearest Neighbours and an attribution approach to analyse model decisions. Their work is a demonstration of how the two are combined within a classification pipeline to mitigate imbalance and to be interpretable.

A more recent study has also explored multiclass staging, in which the visual similarity between diagnostic categories from neighbouring stages can be small. Khan, Kim, and Kwon [6] tested a multiscale 3-D CNN for 6-class AD stages on structural MRI. Their class-balanced framework took into account class weighting, focal loss, weighted sampling, and augmenting with less frequent classes. Importantly, they also reported differences in aggregate performance and minority-class performance, which further underscores the importance of looking at class-specific metrics. The authors also noted that it is not always possible to ensure participant independent evaluation in scan-level partitioning.

The results shows that MRI classification results is impacted by the architecture of the structures, class imbalance, and evaluation protocols. When datasets, class definitions, pre-processing, and validation procedures vary, however, performance values obtained in various studies cannot be compared directly. A controlled comparison under a common experimental protocol, would then be valuable to determine the practical impact of the alternative training strategies, while reducing the effects of other factors.

2.2 Class-Imbalance Mitigation Strategies

Multiclass medical image classification is a problem that faces serious methodologic difficulties such as class imbalance. In case the images in the classes varies significantly, the optimization process could be biased to optimize for classes with more training images. However, a model can be very accurate overall but not accurately recognize less common classes.

This is a problem that has been explored in several ways. Weighted cross-entropy allows for different loss components for examples from different classes. Weighted random sampling changes the probabilities of getting a training example from a particular class in order to make minority classes appear more often during the training process. The relative contribution of easily classified examples decreases as a result of focal loss, which causes more attention to be paid to difficult examples. Another advantage of class-balanced focal loss is that it uses class-frequency information as well in the loss function.

In the same way, Anbu Devi and Suganthi [5] investigated a hybrid sampling technique called Edited Nearest Neighbours (ENN) and SMOTE. In an imbalance-aware MRI classification framework, Khan et al. [6] explored the concepts of class weighting, focal loss, and weighted sampling. The above studies indicate that several balancing mechanisms exist, but which ones are effective and efficient varies based on the distribution of the data in the dataset and the evaluation goal.

More recent research confirms that a focus on imbalance handling needs to be considered in addition to class-wise outcomes. To tackle the class imbalance and uncertainty issues, Rahat et al. [7] suggested a multi stream contextual as well as frequency-conscious deep learning framework for classification of MRI image slice. Their study also reveals the restrictions of using image level evaluations when there is no reliable mapping between patients and images.

These methods show that balancing is an ongoing phenomenon and not that a single balancing approach is better, but not that one balancing approach is better. Individually, it is important to look at the effect of ordinary aforementioned factors with logit adjustment under similar conditions, before concluding about which is more effective.

2.3 Evaluation Protocols and Reproducibility

Medical image classification results can also vary significantly due to data partitioning and metrics used to assess model performance. Accuracy is the mostly used to measure the overall effectiveness of predictions, but it can be misleading if the class distribution is not balanced. Macro-F1 computes the F1-score of individual class autonomously and then averages the scores with same weights given to each of the classes. Balanced accuracy also has the advantage of avoiding the problem of overemphasizing the performance of the majority class, by averaging the class-wise recall.

Khan et al. [6] reported both aggregate and class-specific results in their study of imbalance-aware Alzheimer's disease staging. Their analysis shows that it was worthwhile to study the minority-class behaviour separately from classification performance.

Design of the evaluation is also key. Deenadayaal and Shantha rajah [8] presented a multiclass MRI classification framework that has weighted random sampling, interpretability, calibration and participant-level data partitioning.

Their work is a reminder that, when images from the same subject can have similar anatomical features in training and test datasets, leakage is important to prevent. This leakage can result in an over-optimistic generalization performance estimate.

Several trials in which the random seed is changed can help give further evidence of training variability. Repeated-seed testing on one fixed test split is not independent testing on multiple datasets/patient cohorts, however.

In this study, accuracy, macro-F1, and balanced accuracy are used together for each of the four random seeds. This is designed to highlight performance trade-offs between the imbalance-handling strategies assessed. Because the available dataset cannot be said to be patient independent, the results cannot be considered representative of clinical generalization but rather as image level classification performance.

2.4 Recent Developments and Research Gap

More complex architectures, imbalanced objectives, and clinically oriented evaluation of Alzheimer's MRI classification continues to be the subject of recent studies. Khan et al. [6] studied a multiscale three-dimensional CNN which incorporates multiple imbalance handling components. Rahat et al. [7] studied multistream feature learning and uncertainty related aspects, and Deenadayaal and Shantha rajah [8] added weighted sampling, explainability and participant-level partitioning to their work. The studies show that the focus of the research is not limited to improving the accuracy of the aggregates but also extends to the class-wise performance of the aggregates, uncertainty, interpretability, and the reliability of evaluation protocols.

This improvement has not however been enough to ensure that the results reported can be compared directly, as different studies have different data sources, architectures, class definitions or validation protocols. Furthermore, a proposed architecture of combining multiple balancing techniques does not necessarily exhibit the role of each balancing technique. Controlled comparisons can be used to find the variations in the effectiveness due to the model or class imbalance issues.

This study is therefore focused on the need for a consistent empirical evaluation, across repeated training seeds, of four imbalance-handling strategies, all based on a pretrained ResNet-50 architecture, the same data partitions and multiple evaluation measures. The study compares the standard cross-entropy, weighted cross-entropy, weighted random sampling and class-balanced focal loss with logit adjustment. It doesn't suggest a new neural architecture but explores the degree to which these training methods yield different results in an identical experimental configuration.

The study has three main contributions to make:

1. Shared ResNet-50 training/validation protocol: Four imbalance-handling strategies are tested with the same training and validation protocol (ResNet-50).
2. Multi-seed evaluation: Mean performance and standard deviation are reported over the 4 random seeds, to represent variation between training runs, on a fixed split.
3. Class-wise and metric based analysis: Overall metrics are analysed along with class-wise results and confusion matrices to see the trade-offs which are not reflected by just accuracy.

The analysis of the resulting data yields empirical evidence of the relative behaviour of the four common imbalance handling strategies in the selected four-class MRI data set. It does not provide an optimal balance that is universally applicable or a clinically successful performance that is patient independent.

3. Methodology

3.1 Dataset Sketch

It is publicly available Alzheimer's MRI image dataset provided on Kaggle that contains four classes for image classification. The images are partitioned into four classes. The dataset consists of approximately 6400 images with 70% of them used as training, 15% used validation also 15% used testing with a stratified image-level split. The test set consists of 960 images. The approximate class distribution is maintained in the three subsets with stratification. In the present evaluation, the partitioning is image-level because patient identifiers are not present to define patient-independent splits so patient-level generalization is not established.

Table 1. Split of the datasets

Entire dataset consists 6400 images. There are 960 images in test set.

Categories	Training Images	Validation Images	Testing Images	Total
Mild Demented	627	134	135	896
Moderate Demented	45	10	9	64
Non-Demented	2,240	480	480	3,200
Very Mild Demented	1,568	336	336	2,240
Total	4,480	960	960	6,400

Before splitting the data, the total no. of image in each class is shown in figure 1, as output from the notebook when uploaded.

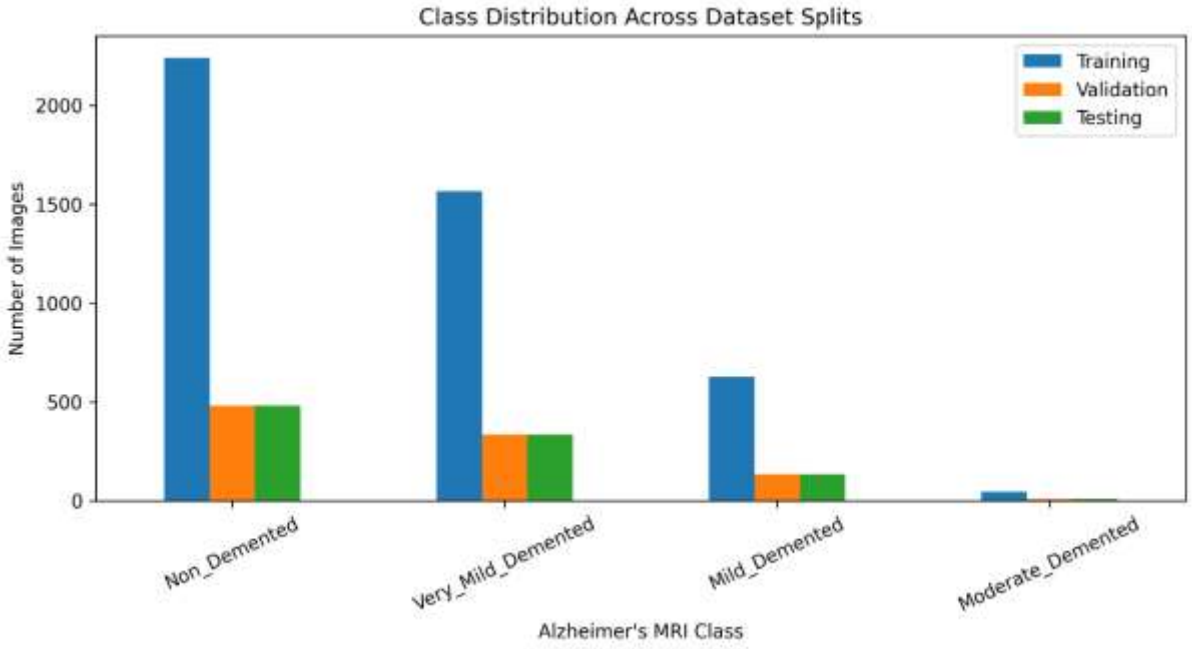


Figure 1. The class distribution of Alzheimer's MRI data

3.2. Image Preprocessing and Data Augmentation

All the images were loaded and converted to RGB format before resizing to 224*224. These images were then transformed to tensors and normalized with the ImageNet mean as well as standard deviations, [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225]. The transformations ensure a uniform representation of the input to ResNet-50. [9].

Image variation was introduced during training by randomly flipping images horizontally with the probability of 0.5, which helped to reduce overfitting. During the validation and testing there was no random flipping applied which

results in consistent evaluation preprocessing. To ensure comparability all four imbalance handling strategies used the same frozen dataset partitions.

Table 2. Image preprocessing configuration

Operation	Parameter	Application
Image conversion	RGB	All subsets
Image resizing	224 X 224 pixels	All subsets
Horizontal-flipping	Probability = 0.5	Training only
Tensor conversion	ToTensor()	All subsets
Normalization	ImageNet mean and standard deviation	All subsets

The steps involved in image preprocessing are shown in Figure 2. First, each MRI image is transformed to RGB also resized to (224 X 224) pixels to ensure that all images have a consistent size as input to ResNet-50. Random horizontal flipping has been applied to the data during the training procedure with a probability of 0.5, meaning 50% of the time. This augmentation is turned off during validation and testing phase to ensure consistent evaluation. Images are then transformed to tensors and normalized with the channel-wise mean and std. of ImageNet. The output of this normalization are the normalized tensors, which are passed to ResNet-50 for classification. To enable a fair comparison, the same preprocessing is kept for all four imbalance-handling strategies.

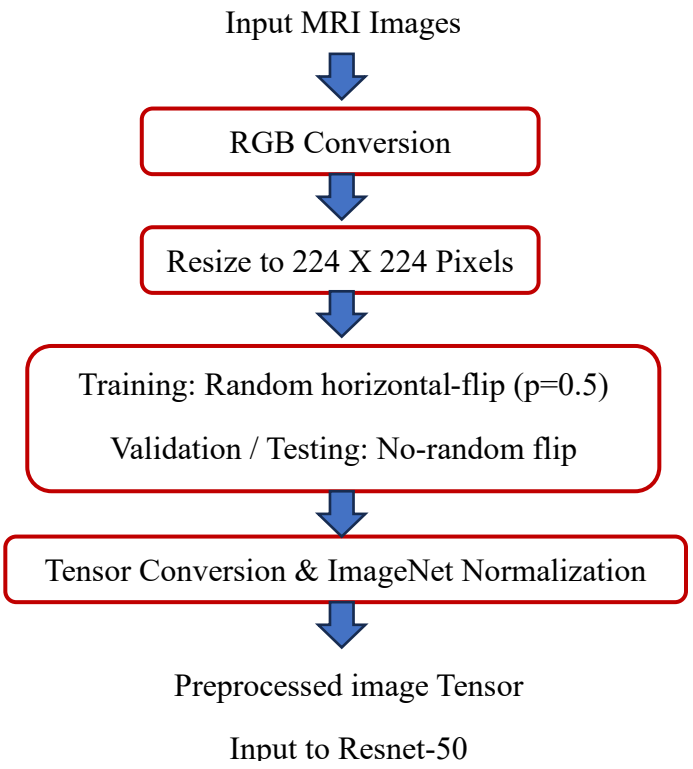

Input MRI Images
RGB Conversion
Resize to 224 X 224 Pixels
Training: Random horizontal-flip (p=0.5)
Validation / Testing: No-random flip
Tensor Conversion & ImageNet Normalization
Preprocessed image Tensor
Input to Resnet-50

Figure 2. The MRI images were preprocessed using the pipeline before classification with ResNet-50

3.3 Proposed ResNet-50-Based Classification Framework

A suggested ResNet-50-Based classification technique is proposed under section 3.3. In 3.3, a proposed classification framework based on ResNet-50 is proposed. The following is the type of material that would work well in your manuscript. It assumes that you are using an image Net-pretrained ResNet-50 model with a modified classification layer for the four MRI classes of Alzheimer's. Double check this with your final program before submitting.

3.3.1 Overall Framework Architecture

The proposed framework can be divided into five main steps: input image acquisition, preprocessing, feature extraction via ResNet-50, four-class probability estimation, and final Alzheimer's MRI image classification. We use the pretrained ResNet-50 network to extract hierarchical image features. Its last classification layer is fine-tuned to make predictions for the four classes it targets.

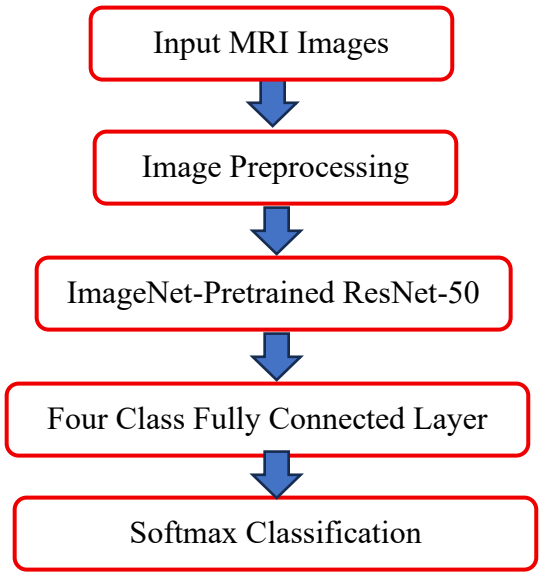


Figure 3. The Proposed Architecture - ResNet-50 based classification system

The backbone feature extraction method and the classification task are shared by the four experimental configurations in the framework. The differences between the configurations are based on the way they deal with class imbalance, which is explained in Section 3.4. This setup allows for a comparison between the models under the same data partition and model structure.

3.3.2. ResNet-50 Feature Extraction

It is a deep CNN which is designed with residual connections to allow training of deeper networks. Residual Block learns a residual function, which is added to the input instead of a direct mapping between all input and output features. This design is useful to propagate information and gradients in the network [10]. The residual mapping is given as:

$$op_feat = \mathcal{F}(x, W) + ip_feat$$

ip_feat – the input feature representation

$\mathcal{F}(x, W)$ – denotes the residual mapping learned using network parameters W

op_feat – denotes the output feature representation

If the size of the input and output are not equal, a projection shortcut can be used to adjust them to match each other. The network gradually filters out low-level and high-level visual characteristics obtained from input and went through a global average pooling layer before the classification.

3.3.3 Four-Class Classification Layer

The four target categories are adapted in the pretrained ResNet-50 classification head. The extracted feature vector is passed to a fully connected layer which produces four output scores (also known as logits).
The classification operation can be represented as:

$$z = W_c h + b_c$$

h - is the feature vector obtained from the backbone
W_c – weight matrix of the classification layer
b_c – bias vector
z – vector belongs four class logits
The logits went through softmax function to give class probabilities.

$$P(y = k\,|x) = \frac{exp(z_k)}{\sum_{j=1}^{4} exp(z_j)}$$

$P(y = k\,|x)$ the predicted probability of the x (input image) is from class k, z_k denotes logit. The highest class suggested is the final predicted class.

$$\hat{y} = \underset{k\ \in \{1,2,3,4\}}{arg\,max\ P(y = k|x)}$$

3.3.4 Integration with Class-Imbalance Strategies
The four training configurations, standard cross-entropy loss (CE), weighted cross-entropy (WCE), weighted random sampling (WRS), and class-balanced focal loss with logit adjustment (CBFL-LA), are used to evaluate the classification framework. Both the classification task and the test set are kept the same for all the configurations.
All the other configurations include some mechanisms for dealing with class imbalance during the training process, while standard cross-entropy is the basic one. The mathematical formulations, the parameter settings and implementation details are given on Section 3.4. Each configuration is tested against of accuracy, macro-F1 score and balanced accuracy, along with class-wise evaluation results.

3.4 Class-Imbalance Handling Strategies
Four approaches were adopted to explore the effect of class imbalance on the classification of Alzheimer's MRI (CE, WCE), weighted random sampling (WRS), class-balanced focal loss with logit adjustment (CBFL-LA)). The same ResNet-50 backbone and partition of the training sets are used for each strategy.

3.4.1 Standard Cross-Entropy (CE)
The classification was based on the standard cross-entropy classification method. It does not explicitly correct for class frequency but penalizes the variation among the class probabilities predicted as well as the ground-truth labels. Given an N samples of training data, the loss is given by:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} log\, p_{i,y_i}$$

p_{i,y_i} probability of prediction that the predicted category for the sample i; y_i is its ground_truth label.

3.4.2 Weighted Cross-Entropy (WCE)
It assigns different loss weight to different classes, to minimize the effect of the majority class. A popular inverse-frequency weighting scheme is:

$$w_c = \frac{N}{Cn_c}$$

N is the total no. of training images, C no. of classes
n_c is the No. of training images in class C.
Weighted cross entropy loss is:

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}\, log\, p_{i,y_i}$$

L_{WCE} – Weighted cross-entropy loss; N – Number of samples used in loss calculation; $\sum_{i=1}^{N}$ – Summation over all training samples from $i = 1\ to\ N$; i - Index of a training image; p_{i,y_i} - Predicted probability assigned to the true class

of image. y_i Ground-truth class label of image i; w_{y_i} - Weight assigned to the true class of image i to address class imbalance;

3.4.3 Weighted-Random Sampling (WRS)

This is the technique of changing the probability of sampling images from underrepresented classes to make it more likely. The sampling weight used for per-image is often:

$$s_i = \frac{1}{n_{y_i}}$$

n_{y_i} is the total no. of images for training in the category associated with sample i

The sampling probability can be expressed as

$$q_i = \frac{s_i}{\sum_{j=1}^{N} s_j}$$

This method is similar to modifying data batches for training, but it does not affect the cross-entropy loss directly. The sampling behaviour will vary according to sampler configuration in the program and may be with or without replacement sampling.

3.4.4 Class-Balanced Focal Loss with Logit Adjustment (CBFL-LA)

The focal loss increases the involvement of the well-categorized examples and decreases the contribution of difficult examples [12]. Class-balanced weighting provides for class-frequency weighting with the help of the effective number of samples [11].

No. of effective sample images in category c is:

Total No. of effective sample images of category c is calculated as:

$$E_c = \frac{1 - \beta^{n_c}}{1 - \beta}$$

The class weight is

$$\widetilde{w_c} = \frac{1 - \beta}{1 - \beta^{n_c}}$$

These weights are standardized so that they have a mean of 1. The value of $\beta = 0.999$

And the focal-loss parameter $\gamma = 2$.

The logit transformed by the log prior to class c is

$$p'_{i,c} = \frac{exp(z'_{i,c})}{\sum_{j=1}^{c} exp(z'_{i,j})}$$

The class-balanced focal loss that has been implemented is

$$L_{CBFL-LA} = -\frac{1}{B}\sum_{i=1}^{B} w_{y_i\left(1-p'_{i,y_i}{}^{\gamma}\right)}\ log(p'_{iy_i})$$

where w_{y_i} is the normalized effective-number weight for the ground-truth class. The implementation uses the adjusted logits to calculate both the focal probability and its logarithm.

3.5 Experimental Setup and Implementation Details

The execution is done with static dataset partition and a ResNet-50 training with the Image_Net-pretraining. Three are 3 groups of datasets named training, testing and validation with the group of 70%, 15% and 15% each. The three groups have images of 4,480, 960 and 960.

To investigate the variability of model training, four random seeds (21, 63, 84 and 105) were used for each of the four imbalance handling strategies. This lead to 16 training runs. The validation group is to develop the model, and test group is to evaluate the final performances. A consistent comparison was ensured by using the same test images in all experiments.

Table 3. Experimental configuration

Parameter	Configuration
Classification model	ResNet-50
Pretrained weights	ImageNet-pretrained (verify successful loading in execution logs)
Number of classes	4
Dataset split	70 -15 -15 (3 Groups)
Size of Input image	224 * 224 pixels
Size of each Batch	32
Maximum epoch	15
Optimization Used	AdamW
Rate of learning	1×10^{-4}
Weight decay	1×10^{-4}

The main experimental parameters that were used to compare four class-imbalance mitigation strategies are summarized in Table 3. The same ResNet-50 classification framework and fixed dataset partitions are used in all techniques. The four randomly selected seeds are used to evaluate the variability between multiple training runs, and the validation macro-F1 is used to direct the selection of checkpoints.

3.6 Evaluation Metrics

The following metrics are especially useful if there are significant disparities in class frequencies [14].

3.6.1 Accuracy

The amount of correctly categorized images out of the whole images of test images is a measure of accuracy.

$$Accuracy\ \ Level = \frac{True\ Positive\ +\ \text{True Negative}}{\text{Treu Positive}\ +\ \text{True Negative}\ +\ \text{False Positive}\ +\ \text{False Negative}}$$

An equivalent formula is for multiclass classification is

$$\text{Accuracy} = \frac{\sum_{c=1}^{c} \text{TP}_c}{\text{N}}$$

Here TP_c is the total no. of correctly classified images belonging to category c , c total classes N total test images.

3.6.2 Precision - Recall - F1 Score

First is the measurement of purity of the model in terms of positive predictions.

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}$$

Second is the model's ability to find all positive instances of the class.

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}$$

The class-wise F1-score performance metric combination of recall as well as precision.

$$\mathrm{F1}_c = 2\frac{\mathrm{Precision}_c . \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}$$

Here TP_c, FP_c and FN_c - true positive, false positives , false negatives for call c, respectively.

3.6.3 Macro-F1 Score

Macro-F1 treats the F1 scores from each class equally by computing the class wise average of F1-scores.

$$\mathrm{Macro} - \mathrm{F1} = \frac{1}{c}\sum_{c=1}^{c} \mathrm{F1}_c$$

It is not the same as accuracy, which can be misleading if the largest class makes up a large percentage of the test images. This makes it convenient to study performance on the imbalanced Alzheimer's MRI classes.

3.6.4 Balanced Accuracy

This is the mean value of all the classes.

$$\mathrm{Balanced\ Accuracy} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{Recall}_c$$

This metric gives each class equal weight, irrespective of its frequency in the test set [14].

3.6.5 Mean and Standard Deviation Across Seeds

The above of the four random-seed runs were computed for each method.

$$s = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(x_r - \bar{x})^2}$$

where x_r is the metric obtained from run r, $R = 4$ is the number of runs, $\bar{x}$ denotes mean and s sample standard deviation. Results are reported as mean ± standard deviation.

These statistics report on the variation from one run to another on a fixed data partition. These should not be considered as variability between independent datasets or patients.

3.6.6 Confusion Matrix

A confusion matrix was created for each classification approach in order to compare the true category groups against the predicted group category. Rounding errors are indicated by the numbers that lie off diagonally, and the numbers that lie on the diagonal are the correct predictions. The matrices were analyzed to identify the classes from the Alzheimer's MRI that are more likely to be misclassified.

4. Results and Discussion

4.1 Experimental Results

A ResNet-50 based deep model for four-class Alzheimer's MRI image classification was used to assess the performance of four class-imbalance handling strategies. The strategies that were evaluated were the Weighted Cross-Entropy (WCE), standard Cross Entropy Loss (CE), Weighted Random Sampling (WRS), and Class-Balanced Focal Loss with Logit Adjustment (CBFL-LA). The evaluation criteria of accuracy, macro-F1 score and balanced accuracy were evaluated in order to compare the categorization effectiveness as well as the classification performance among the groups.

The experiments were conducted four times with the seeds prepared in random manner and the mean and standard deviations are presented. This evaluation is used to get an idea of how well the different models perform when trained again and again. The repeated runs should not be interpreted as independent dataset level runs, but the same test partition was used for the comparisons.

4.2 Comparative Performance Analysis

Table 4. Class-imbalance handling strategies performance comparison

Method	Acc (%)	Macro-F1 (%)	Balanced Accuracy (%)
Cross Entropy (CE)	96.04 ± 0.37	94.08 ± 1.56	93.80 ± 3.73
Weighted Cross-Entropy (WCE)	94.95 ± 1.43	94.11 ± 3.31	92.59 ± 4.98
Weighted Random Sampling (WRS)	96.43 ± 0.33	94.05 ± 1.47	91.59 ± 2.20
CBFL-LA	92.45 ± 2.01	92.47 ± 1.83	90.83 ± 2.67

Values represent the mean ± standard deviation and were obtained from four different random-seed runs. The highest mean of each metric is marked as bold.

Results show different compromises between the analysed strategies. The overall accuracy of WRS was the first with the mean accuracy of 96.43% and a stddev of 0.33%, showing relatively consistent overall accuracy in the four runs. Although the difference was not large, WCE had the best mean macro-F1 score (94.11%). The overall mean balanced accuracy of CE was the highest (93.80%), indicating good overall recall in all four classes.

The accuracy of CBFL-LA was 92.45%, macro-F1 score was 92.47%, and balanced accuracy was 90.83%. With the setup used in this experiment it was not superior on these combined measures. The results suggest that the proposed class-balanced weighting, focal loss and logit adjustment may not necessarily yield better results in every dataset and training setup.

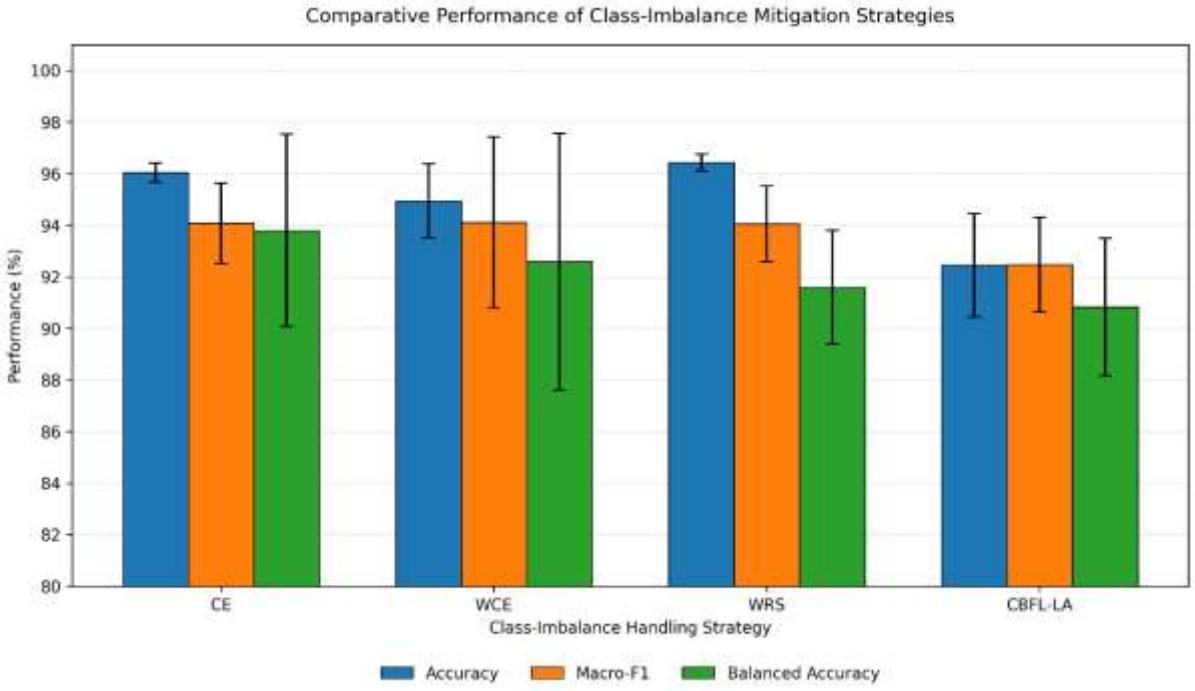


Figure 5. Comparative Performance of the Four Strategies

To visually compare the results of accuracy, macro-F1, and balanced accuracy as shown in Table 5, a comparison was made by plotting the results on the graphs of Figure 5. The standard-deviation error bars show the variation between the four runs of the random seed. The figure illustrates the overall classification performance and the class-balanced performance.

4.3 Discussion of Class-Imbalance Handling Strategies

Analysis of the result reveals that the selection of the imbalance-handling technique has an impact on achieving optimal balance between the overall accuracy and class-wise accuracy. All in all, the results of standard CE were good, indicating that it continues to perform well even if the class distribution of the training data is imbalanced.

WCE gives higher weight to classes based on the inverse of their frequency of training samples. It has the highest mean macro-F1 score, indicating that it is a viable option if performance across classes is crucial. The standard deviation on balanced accuracy is relatively large, however, which suggests that there are run-to-run variations to consider when interpreting the results.

WRS samples minority-class examples more often in order to change the composition of training batches. In these experiments, it shows to be consistent with respect to mean accuracy and low accuracy standard deviation. However, its overall accuracy was not always superior to CE, as there is no direct relationship between the overall accuracy and the average recall obtained per class.

CBFL-LA is a combination of effective-number based class weighting, focal modulation and logit adjustment. The results indicate that the combination of these mechanisms did not produce the maximum aggregate performance in this experiment and that optimizing one mechanism is not necessarily optimal for the other. Before conclusions can be drawn for this method further investigation of the hyper-parameters and the behaviour of the minorities is required.

Table 5. Class-Wise Classification Performance Across Four Random Seeds

Method	Class	Precision (%)	Recall (%)	F1-score (%)
CE	Mild Demented	94.72 ± 6.11	93.52 ± 4.25	93.91 ± 1.55
CE	Moderate Demented	92.95 ± 8.79	88.89 ± 15.71	89.71 ± 6.81
CE	Non-Demented	96.93 ± 1.90	97.19 ± 1.37	97.04 ± 0.50
CE	Very Mild Demented	95.79 ± 2.81	95.61 ± 2.27	95.65 ± 0.84
WCE	Mild Demented	95.67 ± 0.73	94.07 ± 2.34	94.85 ± 1.35
WCE	Moderate Demented	100.00 ± 0.00	86.11 ± 16.67	91.88 ± 9.87
WCE	Non-Demented	95.66 ± 2.09	95.62 ± 1.45	95.63 ± 1.28
WCE	Very Mild Demented	93.63 ± 2.20	94.57 ± 2.41	94.08 ± 1.59
WRS	Mild Demented	97.02 ± 2.34	95.00 ± 0.71	95.99 ± 0.98
WRS	Moderate Demented	100.00 ± 0.00	77.78 ± 9.07	87.28 ± 5.77
WRS	Non-Demented	97.19 ± 0.58	97.29 ± 0.17	97.24 ± 0.25
WRS	Very Mild Demented	95.08 ± 0.64	96.28 ± 1.02	95.67 ± 0.73

Method	Class	Precision (%)	Recall (%)	F1-score (%)
CBFL-LA	Mild Demented	94.22 ± 4.08	88.33 ± 0.93	91.15 ± 1.86
CBFL-LA	Moderate Demented	100.00 ± 0.00	88.89 ± 9.07	93.93 ± 5.11
CBFL-LA	Non-Demented	94.00 ± 1.26	93.75 ± 3.72	93.85 ± 2.07
CBFL-LA	Very Mild Demented	89.77 ± 5.02	92.34 ± 1.62	90.96 ± 2.38

All the evaluation criteria for each class are shown in Table-6 for each approach to solving the imbalance problem. The results demonstrate that there are class differences in performance. The highest mean F1 scores were for the Non-Demented class for WRS, and for the Very Mild Demented class for CE and WRS. There was, however, a wide range of recall for some of the methods for the Moderate Demented class. These results should be viewed with caution, because there are only nine images of Moderate Demented patients in the test set.

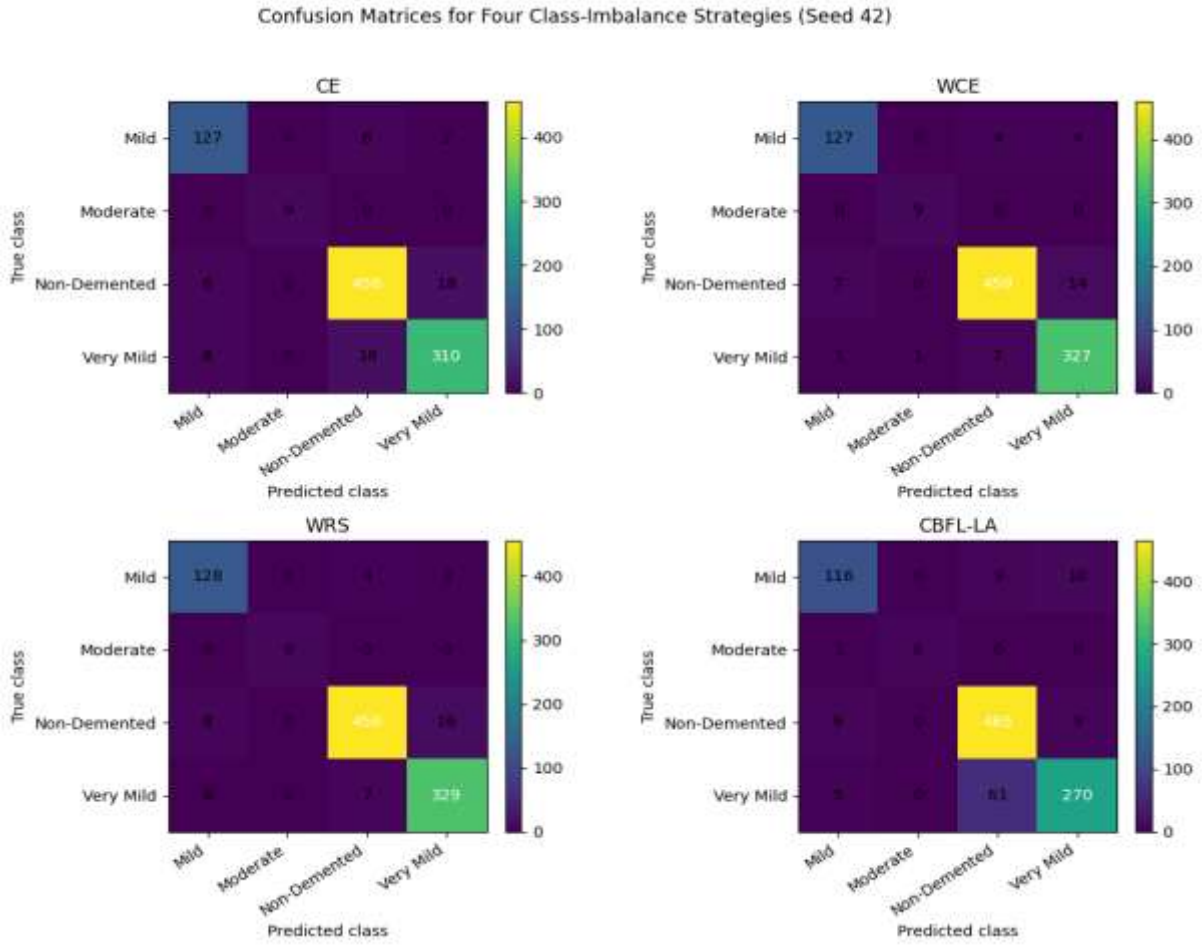


Figure 4. Confusion Matrices of the Four Class-Imbalance Mitigation Strategies (Seed 42).

Confusion Matrices of the four class-imbalance mitigation strategies (Seed 42). Table 6 presents an overview of the evaluation criteria for the Four Class-Imbalance mitigation strategies, class-wise. In the confusion matrices for each of the four strategies displayed in figure 4 to further explore the distribution of correct predictions and misclassifications among individual classes of Alzheimer's MRI. The figure shows the prediction patterns of the classes CE, WCE, WRS, and CBFL-LA – differences in errors that are not apparent from the overall evaluation metrics.

4.4 Implications of the Results

The comparative evaluation has shown that there was no single method that had the top mean score for all three metrics. The best accuracy was obtained by WRS, the best macro F1 was obtained by WCE, while CE obtained the best balanced accuracy. Therefore, the method selected should be based on the evaluation objective and not just on accuracy.

WRS is a good choice of the considered configurations for an application where overall correct classification is primary. If both class-wise precision and recall are crucial, WCE and CE considerations should be taken into account in addition to WRS considerations. The results suggest that using multiple evaluation measures are appropriate for comparing models trained on imbalanced medical image datasets.

Table 6. Test Performance Across Individual Random Seeds

Method	Seed	Accuracy (%)	Macro-F1 (%)	Balanced Accuracy (%)
CE	21	96.04	95.46	93.56
CE	63	95.52	94.94	95.45
CE	84	96.25	93.99	97.46
CE	105	96.35	91.93	88.73
WCE	21	95	96.01	95.63
WCE	63	96.46	97.27	97.38
WCE	84	93.02	89.79	86.29
WCE	105	95.31	93.37	91.07
WRS	21	96.15	93.71	91.33
WRS	63	96.67	95.97	94.54
WRS	84	96.15	94.08	91.24
WRS	105	96.77	92.41	89.23
CBFL-LA	21	93.54	91.88	88.79
CBFL-LA	63	92.08	92.41	90.52
CBFL-LA	84	94.38	94.97	94.68
CBFL-LA	105	89.79	90.62	89.33

Above table, Table 7 presents the effectiveness of each method on the test-set for random seeds 21, 63, 84, and 105. The results show inter-run variability on a fixed data partition, with the same data being divided into the equal no. of partitions each iteration. While the accuracy of WRS remained fairly stable across the four runs, while the accuracy belongs WCE fluctuated more. WCE's macro-F1 and balanced accuracy also had a higher variation across the 4 runs, whereas WRS remained fairly stable. These results are not independent tests conducted on different data sets or patient populations and are in support of the mean and standard deviation analysis presented in Table 5.

4.5 Limitations of the Experimental Results

Limitations

It is vital to consider some limitations of these datasets in the findings. There is only nine images in the test set from the Moderate Demented class, which makes the results sensitive to predictions for this minority class. In addition, patient identifiers were not available, so data was partitioned at the image level and patient independent separation was not possible. The four random-seed runs were made on the same test partition, which means that they show the variability in training, and not the differences between independent patient cohorts. Evaluating the dataset used was done on one public dataset with limited evidence of generalizability. Validation with larger samples of minority class, patient level splits and external datasets is needed before evaluating clinical applicability. The results do not constitute a clinical diagnostic effectiveness.

5. Future Work

The limitations of the current study will be addressed in future studies. More comprehensive and representative datasets will be explored for better representation of minority classes, especially the Moderate Demented. When reliable patient identifiers exist, patient-level data partitioning will be used to assess the generalizability to new patients. External validation with independent datasets will also be performed to evaluate the robustness of the models by analyzing them on different datasets. Additional experiments will examine class-wise consistency of performance and the efficacy of the imbalance-mitigation strategies in different experiment setups. These activities seek to increase the validity and generalizability of Alzheimer's MRI classification and increase the basis for the evaluation of its potential clinical utility.

6. Conclusion

In this work, four class-imbalance handling methods for four-class Alzheimer's MRI image classification with ResNet-50 were compared. Four random-seed runs were used to compare CE, WCE, WRS, and CBFL-LA by their accuracy, macro-F1 score, and balanced accuracy. The outcomes showed that the effectiveness of each of the strategies varied with the criterion being used for evaluation. The highest mean accuracy was obtained by WRS and was 96.43% ± 0.33%, and WCE is the winner with the mean macro-F1 score of value 94.11% ± 3.31%, while CE was the highest mean balanced accuracy of 93.80% ± 3.73%. The results suggest that there is no single best strategy for handling imbalances, which is why it is important to take into account multiple performance measures when evaluating imbalanced medical image classification. These findings, however, must be considered with regard to the limitations of the data set, such as the fact that only nine Moderate Demented images were included in the test set, and that it was impossible to include patient identifiers so that the images could be partitioned at the patient level. Thus, the results reflect comparative performance with the selected data set, and not generalization and/or clinical diagnosis effectiveness for patients independently.

References

[1] M. Menagadevi, S. Devaraj, N. Madian, and D. Thiyagarajan, “Machine and deep learning approaches for Alzheimer disease detection using magnetic resonance images: An updated review,” Measurement, vol. 226, article 114100, 2024. https://doi.org/10.1016/j.measurement.2023.114100

[2] R. D. Resen, L. S. Alzubaidi, H. A. Alwzwazy, and M. I. Capel, “Review of deep learning models for Alzheimer’s disease detection: MRI-centric approaches and multimodal extensions,” Frontiers in Artificial Intelligence, vol. 9, article 1707043, 2026, doi: 10.3389/frai.2026.1707043.

[3] J. M. Amine and M. Mourad, “Toward accurate Alzheimer’s detection: Transfer learning with ResNet50 for MRI-based diagnosis,” Frontiers in Neuroscience, vol. 19, article 1664418, 2025, doi: 10.3389/fnins.2025.1664418.

[4] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, and J. Ye, “Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study,” NeuroImage, vol. 87, pp. 220–241, 2014. https://doi.org/10.1016/j.neuroimage.2013.10.005

[5] M. Krithika Alias Anbu Devi and K. Suganthi, “A Convolutional Mixer-Based Deep Learning Network for Alzheimer's Disease Classification from Structural Magnetic Resonance Imaging,” Diagnostics, vol. 15, no. 11, article 1318, 2025. https://doi.org/10.3390/diagnostics15111318

[6] F. F. Khan, J.-H. Kim, and G.-R. Kwon, “Multi-scale 3D CNN with imbalance-aware training for structural MRI-based six-class Alzheimer disease staging,” Computers in Biology and Medicine, vol. 216, article 111949, 2026. https://doi.org/10.1016/j.compbiomed.2026.111949

[7] S. K. R. U. I. Rahat et al., “Multi-stream contextual and frequency-aware deep learning with evidential uncertainty for Alzheimer’s MRI slice classification,” Discover Artificial Intelligence, vol. 6, article 1375, 2026, doi: 10.1007/s44163-026-02337-2.
[8] T. Deenadayalan and S. P. Shantharajah, “PredatorNet: a clinically trustworthy deep learning framework for multi-class Alzheimer's disease classification using MRI,” Frontiers in Aging Neuroscience, vol. 18, article 1835340, 2026. https://doi.org/10.3389/fnagi.2026.1835340
[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
[10] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-Balanced Loss Based on Effective Number of Samples,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9268–9277. https://doi.org/10.1109/CVPR.2019.00949
[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324
[12] PyTorch Contributors, “CrossEntropyLoss,” PyTorch Documentation. Available: https://docs.pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html. Accessed October 26, 2026.
[13] Scikit-learn Developers, “Model Evaluation: Quantifying the Quality of Predictions,” Scikit-learn Documentation. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed October 26, 2026.
[14] PyTorch Contributors, “Data Loading and Processing: WeightedRandomSampler,” PyTorch Documentation. Available: https://docs.pytorch.org/docs/stable/data.html. Accessed October 26, 2026.
[15] PyTorch Contributors, “Normalize,” Torchvision Documentation. Available: https://docs.pytorch.org/vision/main/generated/torchvision.transforms.Normalize.html. Accessed October 27, 2026.