Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 07 Aug 2026
Revised 11 Sep 2026
Accepted 29 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (23s): 01–26

KEYWORDS:
Alzheimer's disease; Magnetic resonance imaging; Transfer learning; Data leakage; Explainable artificial intelligence; Ensemble learning

Explainable deep transfer learning for multi-stage Alzheimer's disease classification from brain MRI

Harsimran Bains[1], Khushboo Bansal[2]

[1]Department of Computer Science and Application, Faculty of Engineering Technology and Computing, Desh Bhagat University, Mandi Gobindgarh, Punjab, India1
[2]Department of Computer Science and Engineering, Faculty of Engineering Technology and Computing, Desh Bhagat University, Mandi Gobindgarh, Punjab, India2
bainsharsimran91@gmail.com, deputy.directorengg@deshbhagatuniversity.in

ABSTRACT: Deep learning models trained on public Kaggle MRI collections often report near-perfect accuracy for staging Alzheimer's disease, yet duplicated images, untraceable augmented copies and untested explanations can make such results unreliable. This study presents a leakage-aware benchmark of six transfer-learning convolutional networks for four-stage classification of axial brain MRI slices. Exact duplicates were removed by MD5 hashing, leaving 3,741 of 5,099 unaugmented images, which were split 70:15:15 with stratification. VGG16, InceptionV3, Xception, InceptionResNetV2, NASNetMobile and ConvNeXtTiny were trained with an identical classification head, class-weighted loss and two-stage fine-tuning, and the three best models on the validation set were fused by soft voting. Models were compared with agreement measures, bootstrap confidence intervals and McNemar tests, and Grad-CAM heatmaps were assessed for plausibility and faithfulness. VGG16 achieved the highest accuracy of 85.41% (95% CI 82.56–88.26%), with a macro-F1 of 0.890, a Matthews correlation coefficient of 0.781 and a macro ROC-AUC of 0.971; it was significantly better than the other five backbones ($p < 0.001$) and the fastest at inference (7.05 ms per image). The weighted ensemble (84.16%) did not improve on VGG16 ($p = 0.360$), and most errors occurred between the NonDemented and VeryMildDemented stages. All models focused on brain tissue, but for three backbones deleting the highlighted regions changed predictions no more than random deletion, whereas VGG16 produced the most faithful heatmaps. Careful data curation, statistical testing and faithfulness checks are essential for trustworthy MRI-based staging, and a well fine-tuned VGG16 provides an accurate, efficient and interpretable baseline.

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia. Its course runs from subtle cognitive change through very mild and mild dementia to a stage in which daily independence is lost, and the underlying neurodegeneration begins years before the clinical diagnosis is made. Structural magnetic resonance imaging (MRI) captures the macroscopic footprint of this process, including cortical thinning, hippocampal and medial temporal atrophy and enlargement of the ventricles, and it is therefore one of the most widely used imaging modalities for the computer-aided detection and staging of AD (Gao and Lima, 2022; Menagadevi et al., 2024; Hassen et al., 2024). Distinguishing the earliest stages from normal ageing remains difficult, however, because the structural changes of very mild dementia are small and overlap with the variability of healthy brains. Deep learning has transformed automated medical image analysis. Convolutional neural networks (CNNs) and transfer learning from large natural-image collections now support diagnosis across a wide range of conditions, including gastric cancer (Bhardwaj et al., 2022), chest diseases on radiographs (Kakkar et al., 2022), congenital disorders in newborns and fetuses (Kaur et al., 2023), paediatric brain disorders on MRI (Kumar et al., 2025a), knee arthritis

(Kumar et al., 2025b), orthopaedic conditions (Shah et al., 2025) and diabetic foot ulcers (Koul et al., 2026), and related learning methods have been applied to disease forecasting (Kumar et al., 2023). For AD, recent systematic reviews document a rapid growth of deep-learning and transfer-learning approaches for MRI-based detection and stage-wise classification (Sisodia et al., 2023; Mohammed et al., 2024; Chamakuri and Janapana, 2025; Easwaramoorthy et al., 2026). Alongside MRI, other data sources such as retinal photographs (Cheung et al., 2022), electroencephalography (Jain and Srivastava, 2025), handwriting (Pace et al., 2026) and longitudinal multimodal records (El-Sappagh et al., 2020) are being explored, but structural MRI remains the most established input for stage classification. A large share of recent MRI studies is built on openly available Kaggle collections of axial brain slices labelled in four stages: NonDemented, VeryMildDemented, MildDemented and ModerateDemented. These collections have made the problem accessible and have supported increasingly sophisticated architectures. They also carry three risks that are rarely examined. First, some releases contain repeated files and pre-augmented copies that cannot be traced to their source images, so that a random image-level split can place copies of test images in the training set. Analyses of brain MRI classification have shown that slice-level splitting and other forms of data leakage can inflate reported accuracy substantially (Wen et al., 2020; Yagis et al., 2021). Second, new architectures are often compared with baselines trained under different settings, and differences between models are seldom tested for statistical significance. Third, heatmaps from methods such as Grad-CAM are frequently presented as evidence of clinical plausibility, although saliency maps can look convincing without reflecting the model's actual decision process (Adebayo et al., 2018; Bloch and Friedrich, 2024). This study addresses these gaps with a controlled, leakage-aware benchmark of six transfer-learning CNNs for four-stage AD classification from structural MRI slices. The six backbones were chosen to represent distinct design families: a plain deep convolutional stack (VGG16), multi-branch inception modules (InceptionV3), depthwise-separable convolutions (Xception), residual inception modules (InceptionResNetV2), a searched architecture (NASNetMobile) and a modernised convolutional network (ConvNeXtTiny). The main contributions are as follows:

(i) A curation protocol that removes exact duplicates by MD5 hashing, excludes untraceable augmented copies, and quantifies residual near-duplicates across subsets with perceptual hashing before any model is trained.

(ii) A controlled comparison of six architecture families under one preprocessing pipeline, classification head, class-weighted loss and two-stage fine-tuning schedule, so that differences can be attributed to the backbone.

(iii) A statistically supported evaluation that reports agreement measures, bootstrap 95% confidence intervals and paired McNemar tests, together with the computational cost of each model.

(iv) A validation-selected ensemble of the strongest backbones, tested against the strongest individual network.

(v) A quantitative explainability analysis that measures not only whether Grad-CAM and Grad-CAM++ heatmaps fall on brain tissue (plausibility) but also whether removing the highlighted regions changes the prediction.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset, the curation protocol, the models and the evaluation methods. Section 4 presents and discusses the results, including the limitations of the study, and Section 5 concludes the paper.

2. RELATED WORK

2.1 Deep Learning for MRI-Based Alzheimer's Disease Classification

CNN-based transfer learning is the dominant approach to MRI-based AD classification. Sharma et al. (2022) used a CNN with a VGG16 feature extractor to detect AD from MRI scans, illustrating the continued relevance of classical ImageNet backbones for this task, and EL-Geneedy et al. (2023) proposed an MRI-based deep learning approach aimed at accurate AD detection. Later work has moved towards purpose-built architectures. Prasath and Sumathi (2024) designed a pipelined deep learning architecture for AD detection, Lashkary et al. (2025) applied a ConvMixer-based model to detect disease progression from structural MRI, and Ullah et al. (2026) introduced LADNET, an MRI-based deep network for AD detection. Rudraraju and Lakshmi (2026) addressed severity detection and progression analysis with efficient deep learning techniques, which is the four-stage task considered here. Attention mechanisms and hybrid designs are a second line of development. Mohan et al. (2027) combined gather-excite attention with an adaptive residual-DenseNet strategy, and Veluchamy et al. (2025) proposed HY-Deepnet, an optimised deep transfer-learning framework for autonomous AD diagnosis. Several studies add an optimisation or feature-engineering stage to the deep network: Mahendran and P M (2022) embedded feature selection into a deep learning framework for early detection, Ravikanti and Santhanam (2026) integrated features with an improved PEBO

algorithm before an ensemble deep network, and T.S. and N. (2025) used a Taylor-dingo-optimised RP-net to segment MRI before classification. Lakhan et al. (2023) addressed data sharing between institutions with a federated, evolutionary deep CNN. Recent reviews summarise this progress and note that high accuracies are commonly reported on public datasets, while heterogeneity in data handling and evaluation makes the results difficult to compare (Menagadevi et al., 2024; Hassen et al., 2024; Mohammed et al., 2024; Chamakuri and Janapana, 2025).

2.2 Modalities and Multimodal Approaches

AD detection has also been pursued with data other than structural MRI. Kandasamy et al. (2026) modelled tissue probabilistically on fluid-attenuated inversion recovery (FLAIR) MRI, Cheung et al. (2022) developed a deep learning model that detects AD from retinal photographs in a multicentre case-control study, Jain and Srivastava (2025) combined the synchrosqueezing transform of electroencephalography signals with deep transfer learning, and Pace et al. (2026) analysed multimodal handwriting with a dual-backbone network. El-Sappagh et al. (2020) used multimodal time-series data in a multitask model of disease progression, and Mohammed et al. (2026) proposed IFMXCN, an explainable model that fuses multimodal inputs. These studies show that complementary sources can strengthen AD assessment, but they rely on data that are not routinely available, and structural MRI remains the most common basis for stage classification (Easwaramoorthy et al., 2026). Explainability is increasingly expected of clinical AI models. Bloch and Friedrich (2024) systematically compared the explanations produced by 3D deep learning and classical machine-learning models for AD detection, highlighting that explanations themselves require evaluation. Adebayo et al. (2018) showed that some saliency methods produce maps that are largely insensitive to the trained model, so visual plausibility alone is not sufficient evidence of faithful explanation. On the evaluation side, Wen et al. (2020) reviewed CNN-based AD classification and identified data leakage in a substantial share of published studies, and Yagis et al. (2021) demonstrated that slice-level splitting of 2D brain MRI inflates classification accuracy compared with subject-level splitting. Taken together, these studies show that the contribution of the backbone, the effect of data curation and the faithfulness of explanations are rarely examined within one controlled framework, which is the gap addressed in this study.

3. METHODOLOGY

This study develops a leakage-aware deep-learning framework for four-class staging of Alzheimer's disease from axial brain magnetic resonance imaging (MRI) slices. The framework comprises dataset acquisition and deduplication, data-integrity screening and exploratory image analysis, stratified partitioning with class-imbalance handling, image preprocessing and training-time augmentation, transfer learning of six architecturally distinct convolutional neural networks (CNNs), a validation-selected ensemble, multidimensional statistical evaluation and quantitative explainability. Each backbone is trained and evaluated on identical partitions with the same augmentation policy, classification head, optimiser, callbacks and model-selection rule, so that differences in performance can be attributed to the architecture rather than to the experimental conditions. Beyond classification accuracy, each model is examined for agreement beyond chance, statistical uncertainty, computational cost and the faithfulness of its visual explanations. The complete workflow is illustrated in **Figure 1** and described in the following subsections.

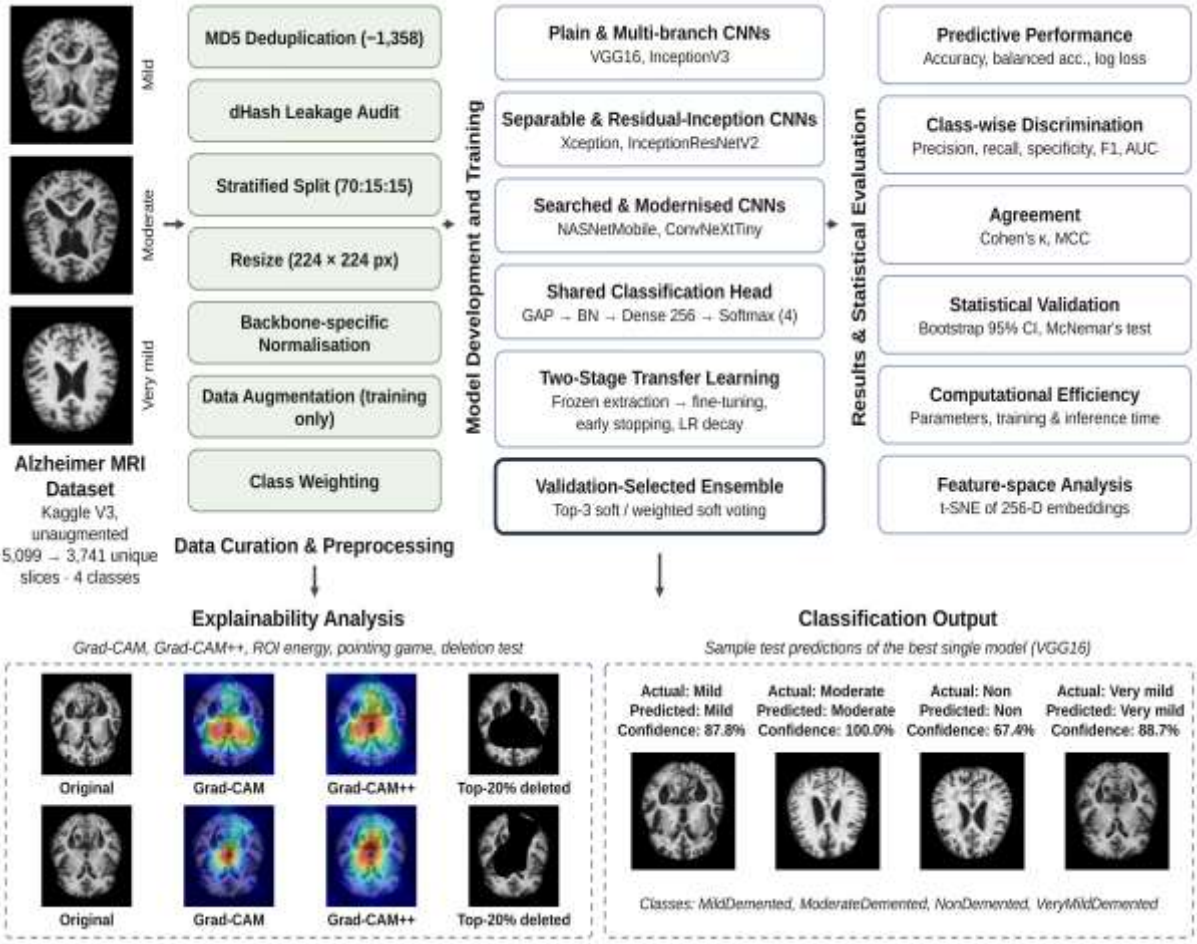


Figure 1: Proposed leakage-aware methodology

3.1 Dataset Description

The publicly available Alzheimer Disease Dataset (Alzheimer_Dataset_V3) hosted on Kaggle was used (Kaggle, 2026). The release contains an *Unaugmented* folder of 5,099 JPEG axial brain MRI slices labelled in four dementia stages, MildDemented, ModerateDemented, NonDemented and VeryMildDemented (**Figure 2**), and an *Augmented* folder of 33,984 synthetic variants. Only the unaugmented images were used. The augmented files carry random identifiers and cannot be traced back to their source images, so including them would place geometrically altered copies of validation and test images in the training set and inflate the measured performance. The dataset provides image-level labels only; it contains no subject identifiers, acquisition parameters, demographic information or pixel-level annotations. Its class-wise composition and partitioning are reported in Section 3.3.

Natural Resources for Human Health
eISSN: 2583-1194

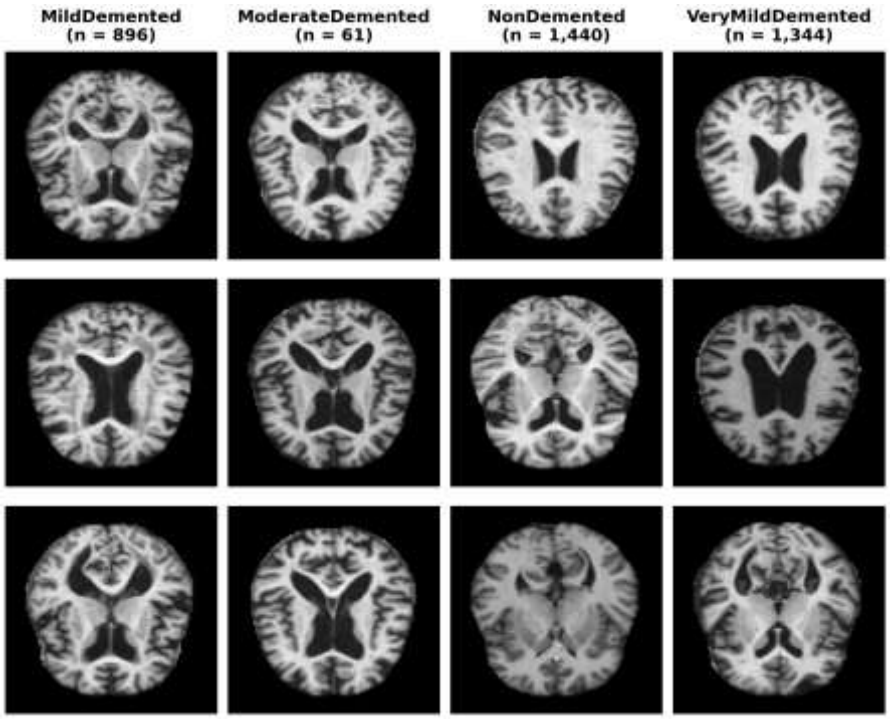


Figure 2: Representative MRI slices of the four classes

3.2 Data Integrity Screening and Exploratory Analysis

Before model development, every file was hashed with the MD5 algorithm so that byte-identical images could be detected regardless of their file names. Repeated files within a class were reduced to a single copy, and identical files carrying different labels were to be removed entirely; no such cross-label conflicts were found. Deduplication removed 1,358 files (26.6%) and left 3,741 unique images. The duplication was strongly class-dependent (Table 2): ModerateDemented fell from 971 to 61 images (−93.7%) and MildDemented from 1,344 to 896 (−33.3%), whereas NonDemented and VeryMildDemented contained no exact duplicates. Without this step, copies of the same slice would have been distributed across the training and test subsets.

Deduplication cannot detect visually near-identical images, such as adjacent slices from the same subject. To quantify this residual risk, a 64-bit difference hash (dHash) was computed for every image. The image is converted to greyscale, resized to 9 × 8 pixels by bilinear interpolation, and each bit records the sign of the horizontal intensity gradient:

$$b_{r,c} = \mathbf{1}[I(r, c+1) > I(r, c)], \quad r = 1, \ldots, 8, \;\; c = 1, \ldots, 8 \tag{1}$$

The number of hash values shared by two subsets D_i and D_j was then counted:

$$S_{ij} = \left| \mathcal{H}(D_i) \cap \mathcal{H}(D_j) \right| \tag{2}$$

The training and test subsets shared 192 hash values, the training and validation subsets 188, and the validation and test subsets 94. Because the dataset provides no subject identifiers, a subject-level split was not possible; these counts are therefore reported as a limitation of the benchmark rather than removed silently.

Class-level image properties were characterised on up to 300 randomly drawn training images per class (all 43 for ModerateDemented). For every class, the pixel-wise mean and standard-deviation images, the intensity distribution of brain pixels (grey level > 10) and the segmented brain area were computed (**Figure 3**, **Table 1**). The brain region of interest (ROI) was obtained by Gaussian smoothing with a 5 × 5 kernel followed by Otsu thresholding (Otsu, 1979), which selects the grey level that maximises the between-class variance of the intensity histogram:

$$\tau^* = \underset{\tau}{\operatorname{argmax}}\, \omega_0(\tau)\, \omega_1(\tau) [\mu_0(\tau) - \mu_1(\tau)]^2 \tag{3}$$

where ω_0, ω_1 are the probabilities and μ_0, μ_1 the mean intensities of the two classes separated by threshold τ. Objects smaller than 200 pixels were removed, holes smaller than 300 pixels were filled, and only the largest connected component (LCC) was retained:

$$M = \mathrm{LCC}\left(\mathrm{Fill}\left(\mathrm{Open}(G_{5\times5} * I > \tau^*)\right)\right) \quad (4)$$

The mean images show progressive ventricular enlargement and cortical thinning from NonDemented to ModerateDemented, and the segmented brain area decreased from 41.7% of the image in NonDemented to 36.9% in ModerateDemented (**Table 1**). These differences are consistent with the expected atrophy pattern, although they are measured on single slices and are therefore also influenced by slice position. To inspect anatomical detail, contrast-limited adaptive histogram equalisation (CLAHE; Zuiderveld, 1994), global histogram equalisation, k-means intensity clustering (k = 3), simple linear iterative clustering superpixels (SLIC; Achanta et al., 2012), Canny edges and Sobel gradient magnitude were also computed (**Figure 4**). These operations were used for visual characterisation only and were never applied to the classifier inputs.

Table 1. Class-level image statistics of the training subset

Class	Images analysed	Brain area, mean ± SD (% of image)	Brain-pixel intensity, mean ± SD (0–255)
MildDemented	300	37.79 ± 2.87	132.96 ± 62.94
ModerateDemented	43	36.93 ± 3.20	135.43 ± 64.37
NonDemented	300	41.74 ± 4.11	147.06 ± 60.59
VeryMildDemented	300	39.75 ± 3.70	139.00 ± 62.09

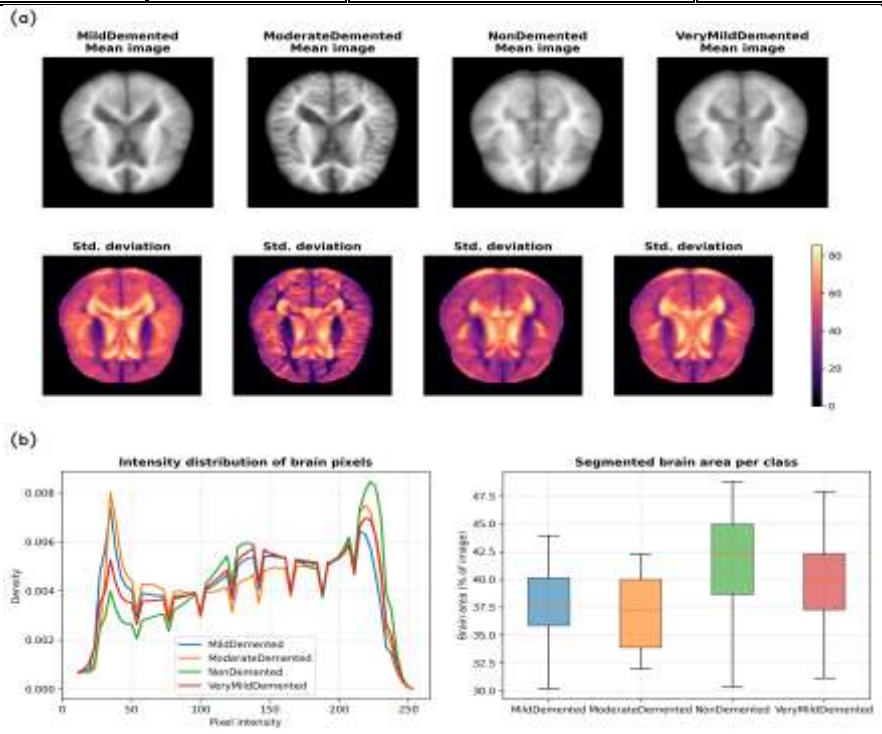


Figure 3: Class-wise mean, variability, intensity and brain-area analysis

Natural Resources for Human Health
eISSN: 2583-1194

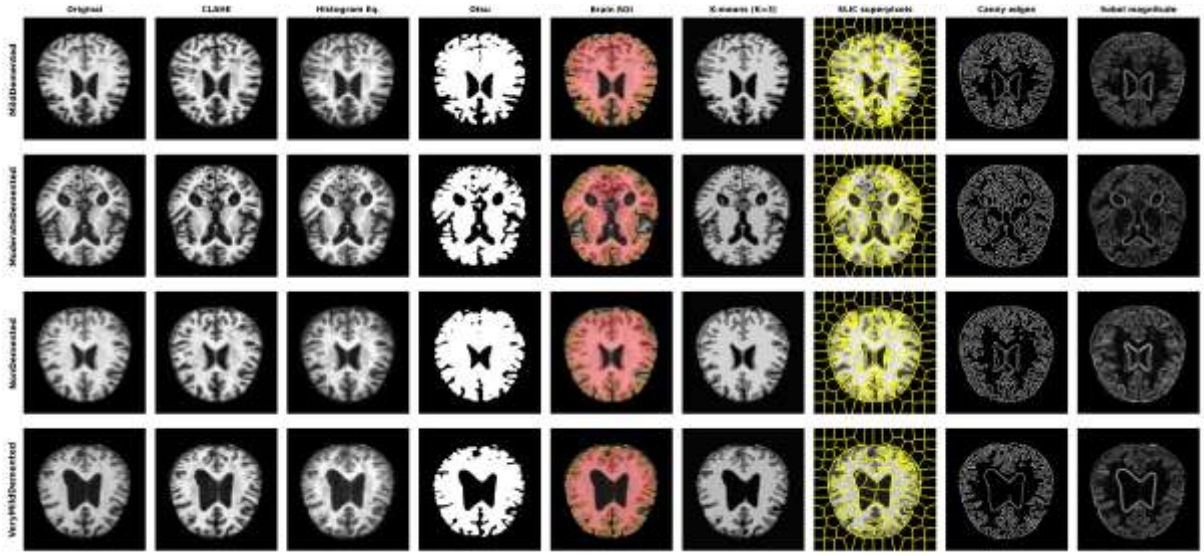


Figure 4: Enhancement and segmentation of representative MRI slices

3.3 Stratified Data Partitioning and Class-Imbalance Handling

A two-step stratified hold-out protocol with a fixed random seed of 42 produced three disjoint subsets from the 3,741 unique images. First, 15% of the images were reserved as an independent test set; the remaining 85% was divided again so that the validation subset formed 15% of the complete dataset, leaving 70% for training. Stratification preserved the class proportions in every subset:

$$D = D_{train} \cup D_{val} \cup D_{test}, \qquad D_i \cap D_j = \varnothing \ \ (i \neq j) \tag{5}$$

The split yielded 2,618 training, 561 validation and 562 test images (Table 2). The validation subset was used for early stopping, learning-rate scheduling and ensemble selection, and the test subset was used only once, for the final evaluation.

Class imbalance was severe after deduplication: the training subset contained 1,008 NonDemented but only 43 ModerateDemented images, a ratio of 23.4:1. Balanced class weights were therefore computed from the training labels alone, so that no information from the validation or test subsets influenced optimisation. For class k,

$$w_k = \frac{N}{K \cdot n_k} \tag{6}$$

where N denotes the training-set size, $K = 4$ is the class count and n_k is the number of training images in class k. The weights enter the categorical cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log p(y_i \mid x_i) \tag{7}$$

where y_i is the true class of image x_i and $p(y_i \mid x_i)$ is the softmax probability assigned to it. The resulting weights ranged from 0.649 (NonDemented) to 15.221 (ModerateDemented).

Table 2. Class composition, data partitions and training class weights

Class	Original	Unique	Share (%)	Train	Val.	Test	Weight
MildDemented	1,344	896	23.95	627	134	135	1.0439
ModerateDemented	971	61	1.63	43	9	9	15.2209
NonDemented	1,440	1,440	38.49	1,008	216	216	0.6493
VeryMildDemented	1,344	1,344	35.93	940	202	202	0.6963

Class	Original	Unique	Share (%)	Train	Val.	Test	Weight
Total	**5,099**	**3,741**	**100.00**	**2,618**	**561**	**562**	**–**

3.4 Image Preprocessing and Data Augmentation

A standardised TensorFlow pipeline processed all images. Each file was read from disk, decoded as a three-channel RGB tensor and resized to 224 × 224 pixels by bilinear interpolation, the input resolution shared by every backbone. Decoded images were cached as 8-bit integers to reduce memory use and cast to 32-bit floating point at batch time. Training data were reshuffled at every epoch, whereas validation and test data were kept in a fixed order; all subsets were batched in groups of 32 and prefetched. Augmentation was implemented as layers inside the model graph, so that it was active only in training mode and was bypassed automatically during validation, testing, explanation and inference. Three geometric transformations were applied in sequence (**Table 3**, **Figure 5**). Photometric transformations were deliberately excluded, because MRI intensity carries tissue information and altering it could change the apparent degree of atrophy.

Table 3. Training-time augmentation policy

Transformation	Setting	Variability simulated
Random flip	Horizontal	Left–right orientation of the axial slice
Random rotation	Factor 0.05 (±18°)	Head tilt and patient positioning
Random zoom	±10% in height and width	Differences in head size and field of view

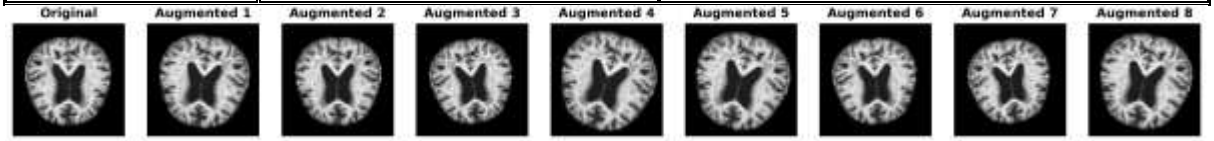


Figure 5: Training-time augmentation examples

After augmentation, each tensor passed through the normalisation expected by its ImageNet-pretrained backbone (Deng et al., 2009), implemented as a model layer so that training and explanation used identical inputs:

$$\tilde{x} = \begin{cases} x_{BGR} - \mu_{IN}, & \text{VGG16} \\ x/127.5 - 1, & \text{InceptionV3, Xception, InceptionResNetV2, NASNetMobile} \\ x, & \text{ConvNeXtTiny} \end{cases} \quad (8)$$

where μ_{IN} = (103.939, 116.779, 123.68) are the ImageNet channel means in BGR order, and ConvNeXtTiny applies its own built-in normalisation layer. The processing order was therefore identical across architectures: RGB decoding → resizing → augmentation (training only) → backbone-specific normalisation → feature extraction.

3.5 Applied Algorithms

Six transfer-learning models were evaluated under an identical protocol. They were chosen to represent six distinct design families rather than variants of one family: a plain deep convolutional stack (VGG16), multi-branch inception modules (InceptionV3), depthwise-separable convolutions (Xception), inception modules with residual connections (InceptionResNetV2), a cell found by neural architecture search (NASNetMobile) and a modernised CNN that borrows design choices from vision transformers (ConvNeXtTiny). All six share one design pattern (**Figure 6**): an ImageNet-pretrained backbone without its original classifier, followed by an identical classification head. Keeping the head constant isolates the contribution of the feature extractor. Let $F \in \mathbb{R}^{H\times W\times C}$ denote the final backbone feature map. Global average pooling (GAP) compresses each channel into a single descriptor:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F(i,j,c) \quad (9)$$

The pooled vector z is batch-normalised, regularised by dropout and passed through a 256-unit fully connected layer with ReLU activation:

Natural Resources for Human Health
eISSN: 2583-1194

$$h = \mathrm{Drop}_{0.30}\left(\mathrm{ReLU}\left(W_1\,\mathrm{Drop}_{0.40}(\mathrm{BN}(z)) + b_1\right)\right) \tag{10}$$

A softmax output layer with $K = 4$ units converts the logits o_c into class probabilities, and the predicted class is the one with the highest probability:

$$p_c = \frac{e^{o_c}}{\sum_{k=1}^{K} e^{o_k}}, \qquad \hat{y} = \underset{c}{\mathrm{argmax}}\; p_c \tag{11}$$

The 256-dimensional activation h also served as the feature embedding for the t-SNE analysis (Section 3.8).

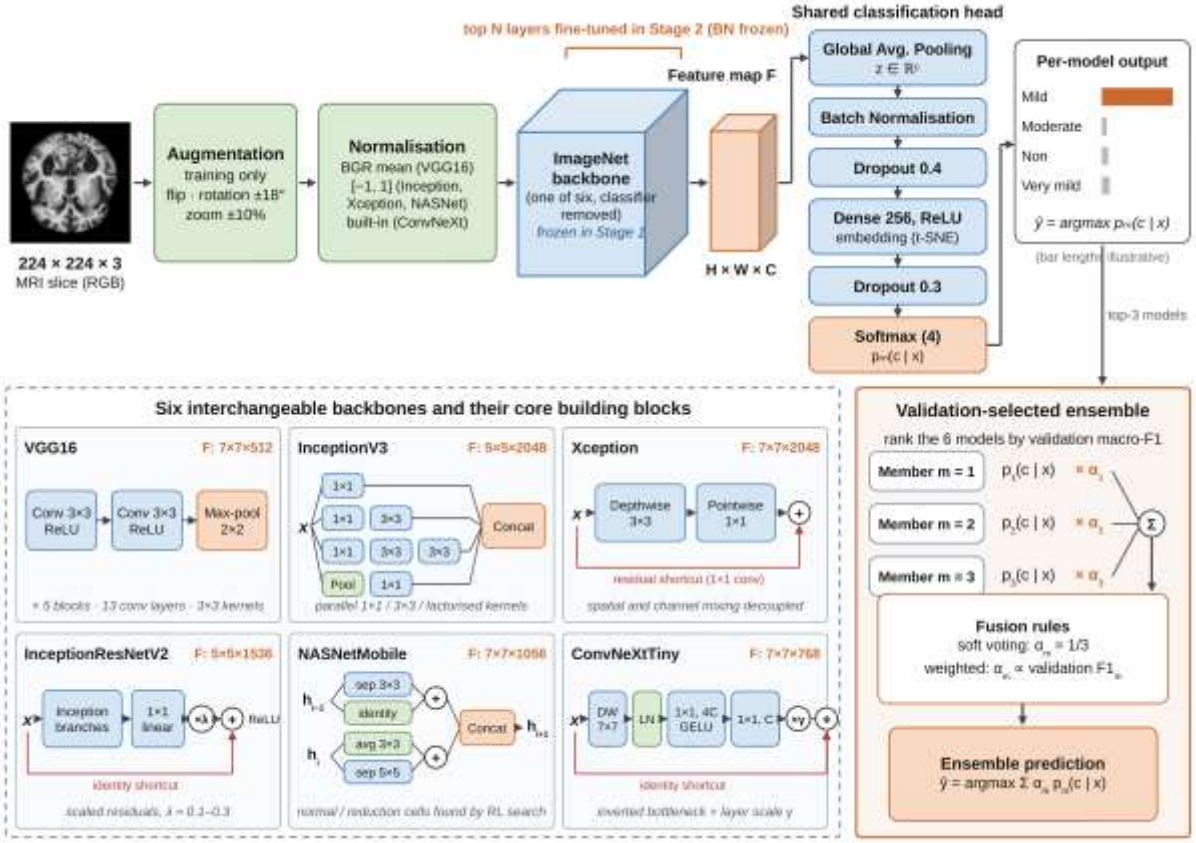


Figure 6: Architecture of the transfer-learning framework

VGG16. VGG16 (Simonyan and Zisserman, 2015) stacks 13 convolutional layers with 3 × 3 kernels in five blocks separated by 2 × 2 max-pooling. Each layer applies

$$x_l = \mathrm{ReLU}(W_l * x_{l-1} + b_l) \tag{12}$$

where * denotes convolution. Stacking small kernels enlarges the receptive field while keeping the number of parameters per layer low. Its final feature map is 7 × 7 × 512, and the last eight layers (blocks 4 and 5) were fine-tuned. InceptionV3 (Szegedy et al., 2016) processes the input through parallel branches with different receptive fields and concatenates their outputs along the channel axis:

$$y = \left[f_{1\times1}(x);\; f_{3\times3}(f_{1\times1}(x));\; f_{3\times3}\left(f_{3\times3}(f_{1\times1}(x))\right);\; f_{1\times1}(P(x))\right] \tag{13}$$

where P is pooling and $[\cdot;\cdot]$ denotes concatenation. Large kernels are factorised into smaller or asymmetric ones, so that fine and coarse structures, such as narrow sulci and enlarged ventricles, are captured at the same depth. At 224 × 224 input, the final map is 5 × 5 × 2,048. Xception (Chollet, 2017) replaces standard convolutions with depthwise-separable convolutions, which first filter each channel spatially and then mix channels with a 1 × 1 convolution, and adds residual shortcuts:

$$y = \mathrm{PW}_{1\times1}\left(\mathrm{DW}_{3\times3}(x)\right) + W_s\,x, \qquad \frac{\mathrm{cost}_{sep}}{\mathrm{cost}_{std}} = \frac{1}{C_{out}} + \frac{1}{k^2} \tag{14}$$

where W_s is an identity or 1 × 1 projection shortcut and k the kernel size. Decoupling spatial and cross-channel mixing reduces computation by almost an order of magnitude for 3 × 3 kernels. The final map is 7 × 7 × 2,048. InceptionResNetV2 (Szegedy et al., 2017) wraps inception modules in residual connections, and scales each residual branch by a small factor λ to stabilise training of the very deep network:

$$y = \mathrm{ReLU}\big(x + \lambda \cdot W_{1\times1}\,\mathcal{F}_{inc}(x)\big), \qquad \lambda \in [0.1, 0.3] \tag{15}$$

It is the largest backbone in this study (54.74 M parameters), and its final map is 5 × 5 × 1,536. NASNetMobile (Zoph et al., 2018) uses normal and reduction cells whose internal operations were discovered by reinforcement-learning-based neural architecture search. Each cell combines pairs of hidden states with searched operations o (separable convolutions, pooling or identity) and concatenates the results:

$$h_{i+1} = \mathrm{Concat}_{b=1}^{B}\big[\, o_{b,1}(I_{b,1}) + o_{b,2}(I_{b,2})\big], \qquad I_{b,\cdot} \in \{h_i,\ h_{i-1},\ \ldots\} \tag{16}$$

where B = 5 blocks per cell and I are the inputs selected by the search. It is the lightest backbone (4.55 M parameters), with a final map of 7 × 7 × 1,056. ConvNeXtTiny (Liu et al., 2022) modernises a residual CNN with design choices taken from vision transformers: large 7 × 7 depthwise kernels, layer normalisation (LN), an inverted bottleneck with four-fold expansion, GELU activation and a learnable per-channel layer scale γ:

$$y = x + \gamma \odot W_2\,\mathrm{GELU}\Big(W_1\,\mathrm{LN}\big(\mathrm{DWConv}_{7\times7}(x)\big)\Big) \tag{17}$$

Its final map is 7 × 7 × 768. The per-model configuration is summarised in **Table 4**.

Table 4. Configuration of the six transfer-learning models

Model	Core building block	Final feature map	Params (M)	Fine-tuned layers	Trainable (M)
VGG16	Plain 3 × 3 convolution stack	7 × 7 × 512	14.85	Last 8	13.11
InceptionV3	Multi-branch inception module	5 × 5 × 2,048	22.34	Last 60	11.06
Xception	Depthwise-separable convolution + residual	7 × 7 × 2,048	21.40	Last 30	9.45
InceptionResNetV2	Inception module + scaled residual	5 × 5 × 1,536	54.74	Last 80	13.36
NASNetMobile	Searched normal/reduction cells	7 × 7 × 1,056	4.55	Last 80	1.32
ConvNeXtTiny	Depthwise 7 × 7 + inverted bottleneck + LN	7 × 7 × 768	28.02	Last 40	18.67

3.6 Two-Stage Training Protocol and Model Selection

Every network followed an identical two-stage fine-tuning schedule, so that the randomly initialised head was adapted before any pretrained weight was altered. Let $\theta = \{\theta_B, \theta_H\}$ denote the backbone and head parameters. In Stage 1 (frozen feature extraction), only the head was trained; in Stage 2 (selective fine-tuning), the top N layers of the backbone, $\theta_B^{(N)}$, were unfrozen and trained together with the head:

$$\text{Stage 1: } \min_{\theta_H}\mathcal{L}\big(\theta_B^{fixed}, \theta_H\big), \qquad \text{Stage 2: } \min_{\theta_H,\,\theta_B^{(N)}}\mathcal{L}(\theta_B, \theta_H) \tag{18}$$

Stage 1 ran for up to 20 epochs with the Adam optimiser (Kingma and Ba, 2015) at a learning rate of 10^{-3}, and Stage 2 for up to 15 epochs at 10^{-5}. The depth of unfreezing was chosen per backbone to cover roughly its last one or two stages (**Table 4**).

Batch-normalisation layers inside the backbone were kept frozen in both stages, because statistics re-estimated from batches of 32 images are noisy and would overwrite the stable ImageNet statistics. Both stages minimised the class-weighted cross-entropy of Eq. (7). Early stopping monitored validation accuracy with a patience of five epochs and restored the weights of the best epoch, and the learning rate was reduced on a validation-loss plateau:

$$\eta_{t+1} = \begin{cases} \max(0.5\,\eta_t,\ 10^{-7}), & \text{if } \mathcal{L}_{val} \text{ has not improved for 2 epochs} \\ \eta_t, & \text{otherwise} \end{cases} \tag{19}$$

Per-epoch training and validation metrics were logged, and wall-clock training time was recorded from the start of Stage 1 to the end of Stage 2. The complete procedure is summarised in Algorithm 1.

Algorithm 1. Leakage-aware training, ensembling and explanation of the six transfer-learning models

Input: unaugmented MRI set D; backbones $\mathcal{B}$ = {VGG16, InceptionV3, Xception, InceptionResNetV2, NASNetMobile, ConvNeXtTiny}; seed 42

Output: trained models f_b, ensemble prediction $\hat{y}$, test metrics with 95% CIs, McNemar p-values and explanation scores

1. Remove exact duplicates by MD5 hash; audit near-duplicates across subsets with dHash (Eqs. 1–2).
2. Split D into D_{train}, D_{val}, D_{test} (70:15:15, stratified); compute class weights w_k from D_{train} (Eq. 6).
3. **for** each backbone $b \in \mathcal{B}$ **do**
    4. Build: input → augmentation (training only) → normalisation (Eq. 8) → b (ImageNet weights, no top) → shared head (Eqs. 9–11).
    5. **Stage 1:** freeze b; train the head with Adam (10^{-3}) and the weighted loss (Eq. 7) for ≤ 20 epochs; early stopping on validation accuracy.
    6. **Stage 2:** unfreeze the last N_b layers (BN frozen); fine-tune with Adam (10^{-5}) for ≤ 15 epochs; restore the best-epoch weights.
    7. Store validation and test probabilities p_b; compute test metrics (Eqs. 21–25) and bootstrap CIs (Eq. 26).
    8. Compute Grad-CAM and Grad-CAM++ on the fixed explanation set; compute E_{ROI}, PG, Δ_{CAM} and Δ_{rand} (Eqs. 28–32).
9. **end for**
10. Rank the six models by validation macro-F1; fuse the top three by soft and weighted soft voting (Eq. 20).
11. Compare the best ensemble and the best single model with every other model using McNemar's test (Eq. 27).

3.7 Ensemble Construction, Evaluation Metrics and Statistical Validation

To test whether complementary architectures outperform the top individual backbone, the six models were ranked by macro-averaged F1 on the validation subset, and the three highest-ranked models formed the ensemble; the test subset played no part in this choice. Two fusion rules were applied to the members' softmax outputs. Soft voting averaged the class probabilities with equal weights, and weighted soft voting scaled each member's probabilities by its validation macro-F1, normalised to sum to one:

$$\hat{p}(c \mid x) = \sum_{m=1}^{M} \alpha_m\ p_m(c \mid x), \qquad \alpha_m = \frac{F1_m^{val}}{\sum_{j=1}^{M} F1_j^{val}}, \qquad M = 3 \tag{20}$$

with a_m = 1/3 for soft voting. Equal-weight voting across the complete pool of six networks served as a reference. The predicted class was the one with the highest fused probability.

Performance was measured on the held-out test subset. For each class k, treated one-versus-rest, the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) give:

$$P_k = \frac{TP_k}{TP_k+FP_k}, \quad R_k = \frac{TP_k}{TP_k+FN_k}, \quad Sp_k = \frac{TN_k}{TN_k+FP_k}, \quad F1_k = \frac{2P_kR_k}{P_k+R_k} \tag{21}$$

where P, R and Sp denote precision, recall (sensitivity) and specificity. These were macro-averaged so that each class counts equally regardless of its size; F1 was additionally reported weighted by class support. Overall accuracy and balanced accuracy were computed as

$$Acc = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\,[\hat{y}_i = y_i], \qquad BAcc = \frac{1}{K}\sum_{k=1}^{K} R_k \tag{22}$$

Agreement beyond chance was summarised by the multiclass Matthews correlation coefficient (Gorodkin, 2004) and Cohen's kappa (Cohen, 1960):

$$MCC = \frac{c\cdot s-\sum_k p_k t_k}{\sqrt{(s^2-\sum_k p_k^2)(s^2-\sum_k t_k^2)}} \tag{23}$$

$$\kappa = \frac{p_o-p_e}{1-p_e} \tag{24}$$

where c counts the correct predictions, s is the sample size, p_k and t_k the numbers of samples predicted as and truly belonging to class k, and p_o and p_e the observed and chance agreement. Threshold-free discrimination was assessed with the macro-averaged one-versus-rest area under the receiver operating characteristic curve (ROC-AUC) and under the precision–recall curve (PR-AUC), and probabilistic quality with the log loss:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\ \log p_{ik} \tag{25}$$

Computational cost was described by the total and trainable parameter counts, the size of the saved weights, the training time and the mean inference time per image, measured over the full test subset in batches of 32 after one warm-up batch.

Because a single test-set estimate carries sampling uncertainty, non-parametric bootstrap confidence intervals (Efron and Tibshirani, 1993) were computed for accuracy and macro-F1. In each of B = 1,000 iterations, 562 test indices were drawn with replacement (seed 42), the metric g was recomputed, and the 95% percentile interval was obtained as

$$CI_{95\%}(g) = [Q_{0.025}(\{g_b\}_{b=1}^{B}),\ Q_{0.975}(\{g_b\}_{b=1}^{B})] \tag{26}$$

Since all models are evaluated on the same images, overlapping intervals are not a formal test of difference. Paired differences were therefore tested with McNemar's test (McNemar, 1947) on the discordant predictions n_{10} (only the first model correct) and n_{01} (only the second correct):

$$\chi^2 = \frac{(|n_{10}-n_{01}|-1)^2}{n_{10}+n_{01}} \tag{27}$$

with one degree of freedom; when $n_{10} + n_{01} < 25$, the exact binomial test was used instead. The top ensemble and the strongest individual network were each contrasted with all remaining models, and $p < 0.05$ was considered significant.

3.8 Explainability and Quantitative Faithfulness Assessment

Class-discriminative heatmaps were produced with Grad-CAM (Selvaraju et al., 2017) and Grad-CAM++ (Chattopadhay et al., 2018) on the final feature map A of each backbone. Gradients were taken of the pre-softmax score y^c of the predicted class rather than of its probability, because softmax saturation at high confidence suppresses the gradient signal. Grad-CAM weights each channel A^k by its spatially averaged gradient:

$$\alpha_k^c = \frac{1}{HW}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \qquad L^c = \text{ReLU}(\sum_k \alpha_k^c\, A^k) \tag{28}$$

Grad-CAM++ replaces the uniform average with pixel-wise coefficients derived from higher-order gradients, which better separates several instances of a discriminative pattern:

$$a_{ij}^{kc} = \frac{\left(\partial y^c / \partial A_{ij}^k\right)^2}{2\left(\partial y^c / \partial A_{ij}^k\right)^2 + \sum_{a,b} A_{ab}^k \left(\partial y^c / \partial A_{ij}^k\right)^3} \tag{29}$$

$$w_k^c = \sum_{i,j} a_{ij}^{kc}\ \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right), \qquad L_{++}^c = \mathrm{ReLU}\left(\sum_k w_k^c\, A^k\right) \tag{30}$$

Each map was upsampled to 224 × 224 by bicubic interpolation and min–max normalised to [0, 1]. Heatmaps that look anatomically plausible are not necessarily faithful to the model's decision, so both properties were measured on a fixed, class-balanced set of 84 test images (25 per class, and all 9 ModerateDemented images), explained in the same way for each network (**Figure 7**). Plausibility was assessed with the ROI energy, the share of heatmap mass inside the brain mask M of Eq. (4), and with the pointing game (Zhang et al., 2018), which records whether the heatmap maximum falls inside M:

$$E_{ROI} = \frac{\sum_{i,j} L_{ij} M_{ij}}{\sum_{i,j} L_{ij}}, \qquad PG = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left[\operatorname*{argmax}_{(i,j)} L_{ij}^{(t)} \in M^{(t)}\right] \tag{31}$$

Because the brain occupies about 37–42% of each image (**Table 1**), E_{ROI} values well above this level indicate genuine concentration on brain tissue. Faithfulness was measured with a deletion test (Petsiuk et al., 2018): the 20% of pixels with the highest Grad-CAM values were set to zero (mask T_{20}), and the resulting drop in the predicted-class probability was compared with that obtained by zeroing the same number of randomly chosen pixels (mask R_{20}):

$$\Delta_{CAM} = p_c(x) - p_c\left(x \odot (1 - T_{20})\right), \qquad \Delta_{rand} = p_c(x) - p_c\left(x \odot (1 - R_{20})\right) \tag{32}$$

A faithful heatmap satisfies $\Delta_{\mathrm{CAM}} > \Delta_{\mathrm{rand}}$. For qualitative inspection, Grad-CAM and Grad-CAM++ maps were displayed for two images per class, for one image per class across the six architectures, and for the eight most confident misclassifications of each model. Finally, the 256-dimensional embeddings h of Eq. (10) for all 562 test images were projected to two dimensions with t-distributed stochastic neighbour embedding (t-SNE; van der Maaten and Hinton, 2008; perplexity 30, PCA initialisation), which minimises the Kullback–Leibler divergence between neighbourhood distributions in the embedding space (P) and the map (Q):

$$\mathrm{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{33}$$

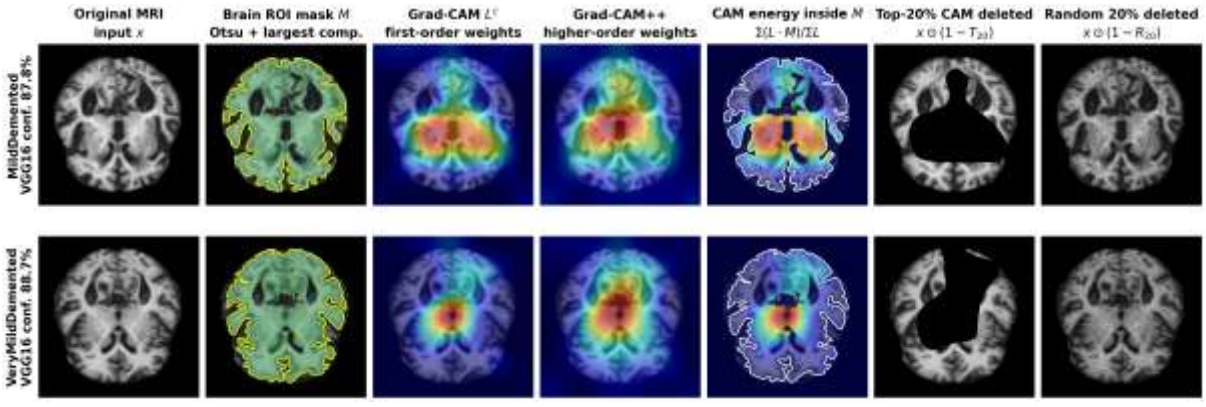


Figure 7: Explainability assessment pipeline

4. RESULTS AND DISCUSSION

4.1 Comparative Classification Performance

The six backbones differed markedly on the held-out test set of 562 images (Table 5). VGG16 was the strongest individual network, with an accuracy of 85.41%, a macro-F1 of 0.890 and the highest specificity (0.944). ConvNeXtTiny ranked second, about nine percentage points behind. Xception, InceptionV3 and InceptionResNetV2 formed a middle group within 1.0 percentage point of each other, with overlapping confidence intervals, and NASNetMobile was clearly the weakest. Balanced

accuracy exceeded overall accuracy for five of the six models because all backbones apart from NASNetMobile recognised most of the nine ModerateDemented test images, which weigh equally with the large classes in the macro average.

Table 5. Test-set classification performance (n = 562)

Model	Acc. (%)	BAcc (%)	Prec.	Rec.	Spec.	F1 (M)	F1 (W)
VGG16	85.41	89.06	0.890	0.891	0.944	0.890	0.854
Weighted ensemble (top-3)	84.16	88.18	0.878	0.882	0.940	0.880	0.841
Soft-voting ensemble (top-3)	83.27	87.53	0.871	0.875	0.937	0.873	0.833
Soft-voting ensemble (all 6)	80.78	85.27	0.852	0.853	0.927	0.852	0.807
ConvNeXtTiny	76.16	82.51	0.817	0.825	0.910	0.820	0.762
Xception	70.11	71.48	0.700	0.715	0.886	0.706	0.699
InceptionV3	69.93	77.17	0.670	0.772	0.887	0.706	0.700
InceptionResNetV2	69.22	71.24	0.708	0.712	0.884	0.709	0.691
NASNetMobile	60.32	59.21	0.651	0.592	0.851	0.613	0.600

Acc.: accuracy; BAcc: balanced accuracy; Prec., Rec., Spec.: macro-averaged precision, recall and specificity; F1 (M), F1 (W): macro-averaged and support-weighted F1.

Model capacity did not predict accuracy. InceptionResNetV2, the largest network (54.74 M parameters), ranked fifth, whereas VGG16, the oldest and architecturally simplest backbone, ranked first. A likely explanation is the depth of adaptation during fine-tuning. Unfreezing the last eight VGG16 layers made 88% of its parameters trainable, so most of the network could move from natural-image features towards MRI-specific ones. The corresponding share was only 24% for InceptionResNetV2 and 29% for NASNetMobile. Because only 2,618 training images were available after deduplication, the backbones that could re-learn the most of their mid-level filters at a low learning rate benefited most. VGG16 also keeps a 7 × 7 final map, whereas InceptionV3 and InceptionResNetV2 reduce the 224-pixel input to 5 × 5, which may discard part of the fine spatial detail of sulcal widening and ventricular shape.

4.2 Training Behaviour and Convergence

The validation curves (**Figure 8**) show that the two-stage protocol behaved as intended. In Stage 1, all heads converged within roughly ten epochs to validation accuracies between 0.65 and 0.75, which reflects the limited transferability of frozen ImageNet features to MRI. Fine-tuning then separated the models. VGG16 gained the most, rising from 0.754 at the end of Stage 1 to a best validation accuracy of 0.845, followed by ConvNeXtTiny (0.747 to 0.784). InceptionResNetV2 did not improve in Stage 2, and its best epoch remained in Stage 1. Early stopping ended Stage 1 of Xception after 12 epochs and Stage 2 of InceptionResNetV2 and NASNetMobile after 8 and 7 epochs, respectively. Training and validation accuracies stayed within a few percentage points of each other throughout (for VGG16, 0.852 and 0.816 in the final epoch), so overfitting was not the limiting factor. For VGG16, ConvNeXtTiny and InceptionV3, validation accuracy was still rising when the epoch budget ended, which suggests that a longer fine-tuning stage could have improved these models further.

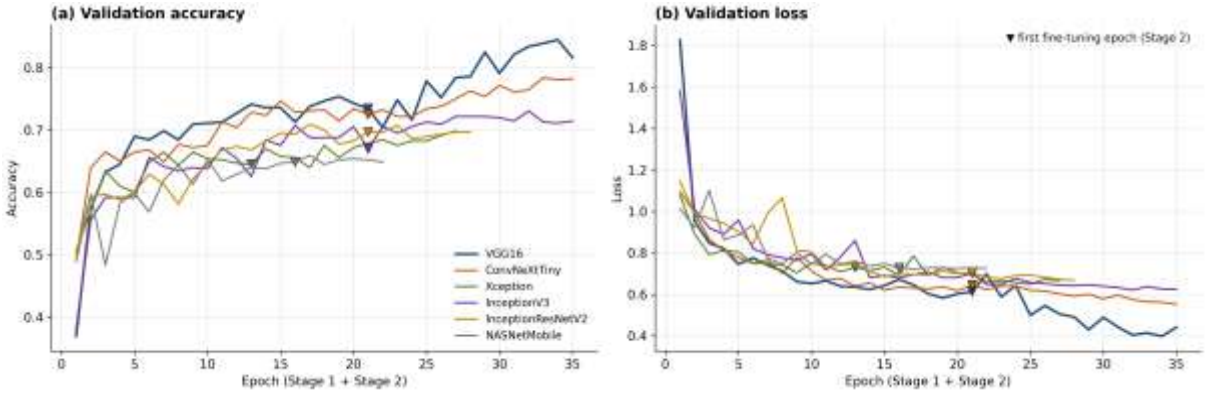


Figure 8: Validation accuracy and loss of the six backbones

4.3 Agreement, Discrimination and Statistical Uncertainty

Agreement beyond chance followed the same ranking (**Table 6**). VGG16 reached an MCC of 0.781 and a Cohen's $\varkappa$ of 0.781, which corresponds to substantial agreement with the reference labels, while NASNetMobile reached only 0.407. The bootstrap intervals separate VGG16 from the remaining single models: its lower accuracy bound (82.56%) lies above the upper bound of ConvNeXtTiny (79.54%) and of every other backbone. Because all models were evaluated on the same images, this separation was confirmed with paired McNemar tests (Section 4.6), which showed VGG16 to be significantly more accurate than each of the other five backbones (all $p < 0.001$).

Table 6. Agreement, discrimination and bootstrap 95% confidence intervals

Model	MCC	$\varkappa$	ROC-AUC	PR-AUC	Log loss	Acc. 95% CI (%)	F1 (M) 95% CI
VGG16	0.781	0.781	0.971	0.951	0.380	82.56–88.26	0.868–0.911
Weighted ensemble (top-3)	0.763	0.762	0.967	0.944	0.465	81.14–87.19	0.856–0.903
Soft-voting ensemble (top-3)	0.749	0.749	0.965	0.942	0.471	80.25–86.48	0.850–0.896
Soft-voting ensemble (all 6)	0.712	0.711	0.959	0.930	0.538	77.40–83.99	0.826–0.877
ConvNeXtTiny	0.645	0.644	0.934	0.885	0.558	72.94–79.54	0.795–0.845
Xception	0.550	0.549	0.900	0.797	0.680	66.19–73.84	0.624–0.766
InceptionV3	0.551	0.550	0.908	0.824	0.654	65.84–73.67	0.638–0.762
InceptionResNetV2	0.540	0.539	0.904	0.816	0.681	65.48–72.95	0.640–0.766
NASNetMobile	0.407	0.406	0.853	0.694	0.818	56.40–64.41	0.506–0.685

ROC-AUC and PR-AUC are macro-averaged one-versus-rest values; $\varkappa$: Cohen's kappa; CI: bootstrap 95% confidence interval (1,000 resamples).

The threshold-independent measures were higher than the accuracy-based ones across all architectures (**Figure 9**). Even the weakest backbone ranked a positive image above a negative one in 85% of pairs, and the micro-averaged ROC curve of VGG16 dominated those of all other models across the full range of false-positive rates. The gap between a high AUC and a moderate accuracy for the middle group suggests that these networks often assigned substantial probability to the correct class without ranking it first, so that many of their errors arise from small margins between neighbouring stages rather than from a complete failure to recognise the class. VGG16 also produced the lowest log loss (0.380), showing that its probabilities

were the best calibrated as well as the most accurate. For VGG16, the per-class AUC ranged from 0.946 (VeryMildDemented) to 1.000 (ModerateDemented).

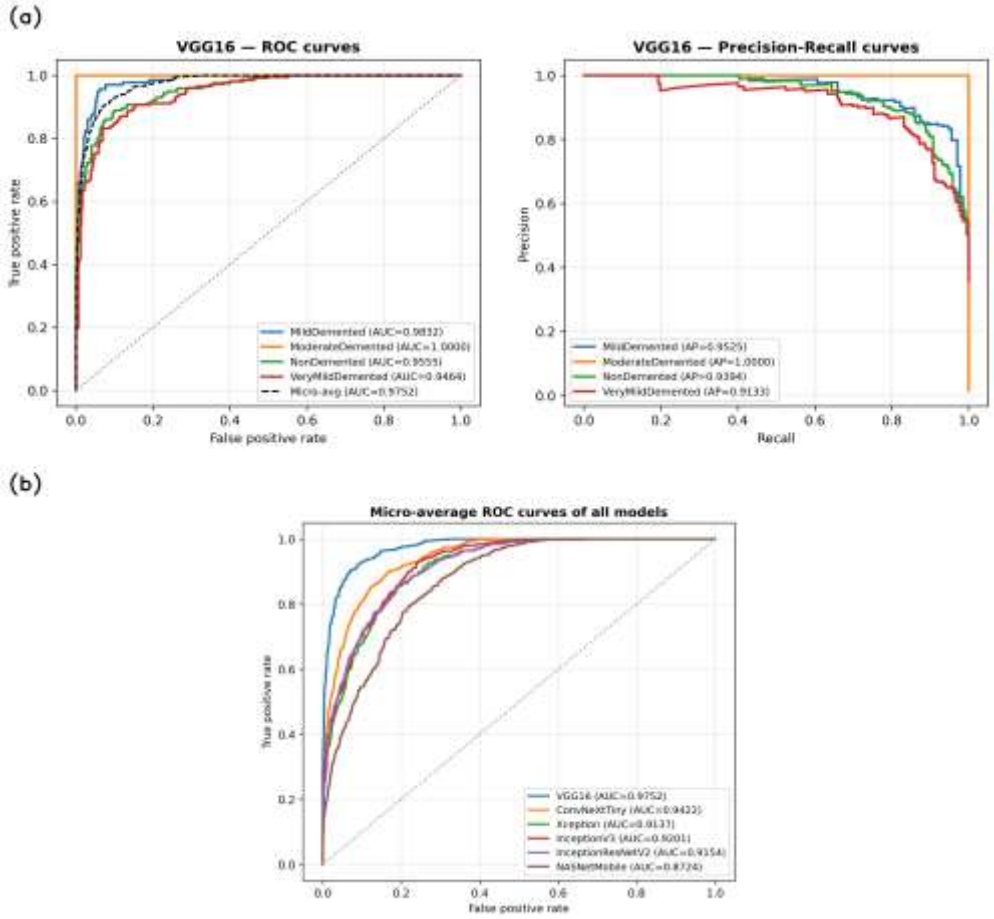


Figure 9: ROC and precision–recall curves

4.4 Class-wise Diagnostic Performance

The class-wise results (**Table 7**, **Figure 10**) show a consistent order of difficulty across all architectures. VeryMildDemented was the most difficult class for all networks, with F1 between 0.509 (NASNetMobile) and 0.823 (VGG16). MildDemented and NonDemented were recognised more reliably, and VGG16 reached F1 values of 0.872 and 0.865 for them. ModerateDemented was classified perfectly by VGG16, ConvNeXtTiny and the weighted ensemble. This result must be read with caution, because the test subset contained only nine ModerateDemented images after deduplication; a single error would lower the recall by 11 percentage points. InceptionV3 detected all nine ModerateDemented images but assigned seven other images to this class, giving a precision of only 0.563, which shows how strongly the weight of 15.2 on this class can push a weaker model towards over-predicting it.

Table 7. Class-wise diagnostic performance of VGG16 on the test set

Class	Precision	Recall	Specificity	F1	AUC	Support
MildDemented	0.862	0.881	0.956	0.872	0.983	135
ModerateDemented	1.000	1.000	1.000	1.000	1.000	9
NonDemented	0.842	0.889	0.896	0.865	0.956	216
VeryMildDemented	0.856	0.792	0.925	0.823	0.946	202

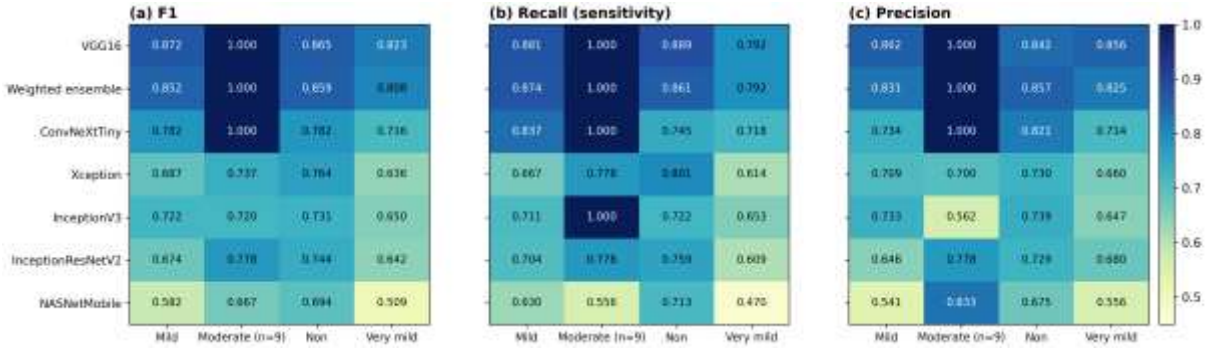


Figure 10: Class-wise F1, recall and precision

4.5 Confusion and Error Analysis

The normalised confusion matrices of the two strongest single networks (Figure 11) show that errors were concentrated between adjacent disease stages. For VGG16, the largest off-diagonal entry was VeryMildDemented predicted as NonDemented (14.9% of that class), followed by the reverse confusion (7.9%), and these two directions together accounted for more than half of its misclassifications. ConvNeXtTiny showed the same pattern with a weaker diagonal: 19.4% of NonDemented images were labelled VeryMildDemented and 14.4% moved in the opposite direction, while confusion between MildDemented and VeryMildDemented also rose (11.9% and 13.9%). Neither network wrongly assigned any image to ModerateDemented, and errors that skipped a stage, such as MildDemented predicted as NonDemented, remained below 7% for both models. This agrees with the clinical picture, in which very mild dementia involves subtle and inconsistent structural change that overlaps with normal ageing on a single axial slice.

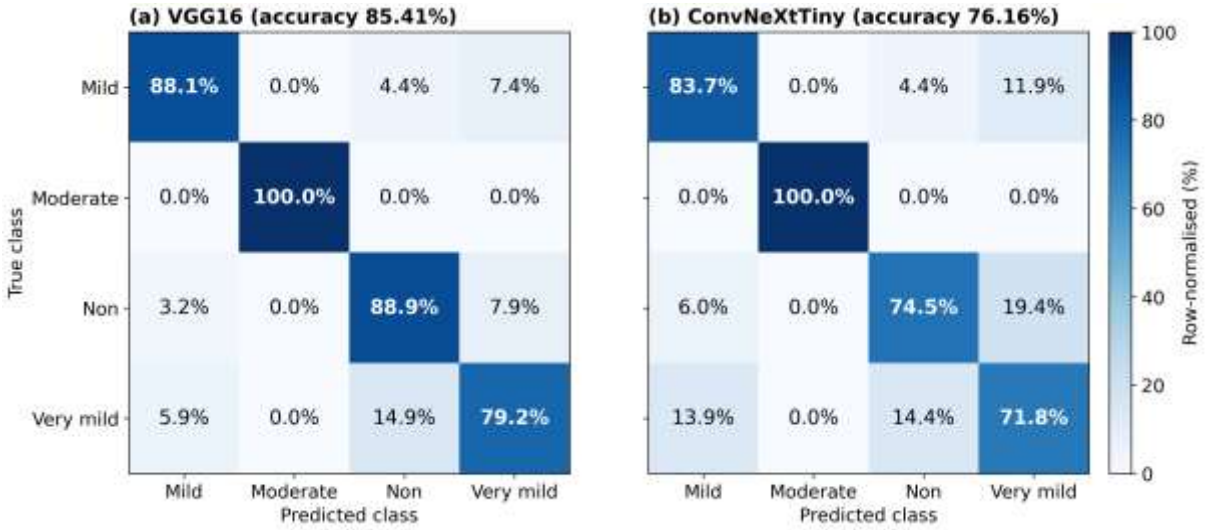


Figure 11: Normalised confusion matrices of the two best single models

The eight most confident misclassifications of VGG16 (**Figure 12**) illustrate the same boundary. All eight were confusions between adjacent stages, with confidences between 0.81 and 0.96, and six of them involved a true VeryMildDemented image, four of which were predicted as NonDemented. Unlike the correctly classified examples, the Grad-CAM maps of these errors were often displaced from the ventricles towards the posterior cortex, the frontal convexity or the skull margin, suggesting that the decision was not anchored on the main atrophy markers in these cases. High-confidence errors of this kind are the most important failure mode for screening, because they are the least likely to be questioned by a reader.

Natural Resources for Human Health
eISSN: 2583-1194

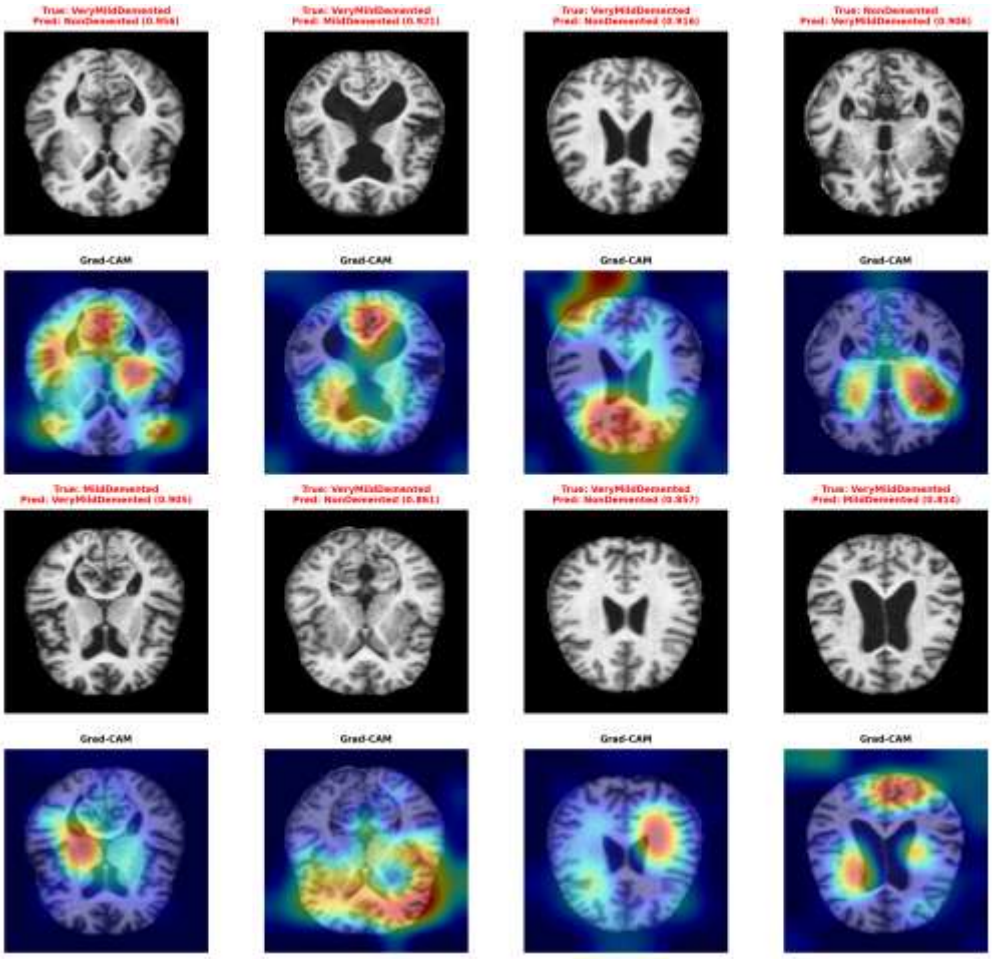


Figure 12: Most confident misclassifications of VGG16 with Grad-CAM

4.6 Validation-Selected Ensemble

Ranking by validation macro-F1 selected VGG16, ConvNeXtTiny and InceptionV3 as ensemble members. Weighted soft voting performed best among the fusion rules, reaching 84.16% accuracy and a macro-F1 of 0.880, followed by equal-weight soft voting and the six-model ensemble (**Table 5**). All three ensembles were significantly better than every single backbone except VGG16. However, none of them exceeded VGG16. The weighted ensemble corrected 25 images that VGG16 misclassified, but introduced 18 new errors on images VGG16 had classified correctly; the difference was not significant (McNemar $p = 0.360$; **Table 8**). The ensemble also had a higher log loss than VGG16 (0.465 versus 0.380).

Table 8. Paired McNemar tests on the test set

Comparison	Only first correct	Only second correct	*p-value*
Weighted ensemble vs VGG16	18	25	0.360
Weighted ensemble vs ConvNeXtTiny	65	20	< 0.001
Weighted ensemble vs Xception	116	37	< 0.001
Weighted ensemble vs InceptionV3	99	19	< 0.001
Weighted ensemble vs InceptionResNetV2	113	29	< 0.001
Weighted ensemble vs NASNetMobile	162	28	< 0.001

Comparison	Only first correct	Only second correct	*p-value*
VGG16 vs ConvNeXtTiny	84	32	< 0.001
VGG16 vs Xception	120	34	< 0.001
VGG16 vs InceptionV3	117	30	< 0.001
VGG16 vs InceptionResNetV2	119	28	< 0.001
VGG16 vs NASNetMobile	172	31	< 0.001

Two factors explain why fusion did not help. First, the gap between VGG16 and the next-best member was large, so averaging pulled the stronger model's probabilities towards weaker, less calibrated ones; weighting by validation F1 reduced but could not remove this effect. Second, the members made correlated errors at the same NonDemented–VeryMildDemented boundary, so their mistakes did not cancel. Ensembling is most useful when members are similarly accurate and err on different images, and this condition was not met here. The single VGG16 model is therefore the more parsimonious choice, as it is as accurate as the ensemble at one-third of the inference cost.

4.7 Computational Efficiency

VGG16 was also the most efficient model to deploy (Table 9). It had the shortest inference time (7.05 ms per image) and trained in under 12 minutes. NASNetMobile had the fewest parameters (4.55 M) and the smallest weight file, but its many small branching operations made it more than three times slower than VGG16 at inference, so parameter count was a poor proxy for speed. InceptionResNetV2 was the largest, the slowest at inference and the most expensive to store, without any gain in accuracy. ConvNeXtTiny offered the second-best accuracy at roughly three times the inference time of VGG16. All models processed an image in about 30 ms or less, which is well within the requirements of offline clinical decision support.

Table 9. Computational cost of the six backbones (S1, S2: training stages)

Model	Params (M)	Trainable (M)	Weights (MB)	Epochs (S1 + S2)	Training (min)	Inference (ms)
VGG16	14.85	13.11	164.4	20 + 15	11.7	7.05
Xception	21.40	9.45	161.6	12 + 15	10.9	9.67
InceptionV3	22.34	11.06	178.7	20 + 15	7.5	11.53
ConvNeXtTiny	28.02	18.67	261.9	20 + 15	14.2	22.91
NASNetMobile	4.55	1.32	30.8	15 + 7	2.9	23.64
InceptionResNetV2	54.74	13.36	327.9	20 + 8	14.0	30.08

4.8 Explainability and Faithfulness of the Heatmaps

Grad-CAM maps produced for the same four test images (**Figure 13**) differed considerably, even when the models agreed on the label. VGG16 maps remained largely inside the brain and were often centred on the lateral ventricles and the posterior periventricular region, where ventricular enlargement is one of the most visible structural markers of Alzheimer's disease. InceptionV3 and Xception produced broader maps that, on the ModerateDemented example, extended beyond the skull, and InceptionResNetV2 spread its activation over large parts of the cortex. NASNetMobile placed strong activation outside the brain on the ModerateDemented example, and ConvNeXtTiny often highlighted small peripheral cortical regions or the skull margin.

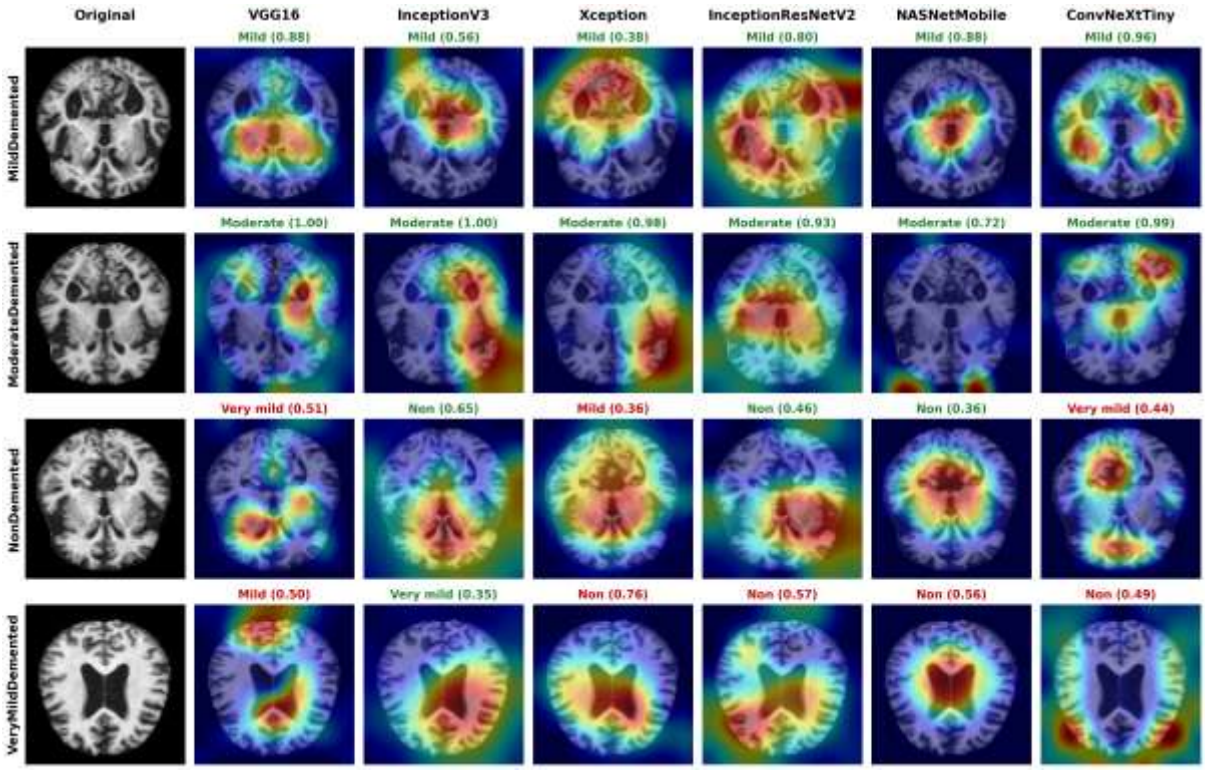


Figure 13: Grad-CAM heatmaps of the six backbones

The quantitative assessment (Table 10) shows why visual impressions alone are insufficient. Each network placed more heatmap energy inside the brain ROI (47.0–72.2%) than the brain's share of the image (36.9–41.7%), so none relied mainly on background. However, plausibility and faithfulness diverged. NASNetMobile had the highest ROI energy but the lowest pointing-game score, and deleting its top-20% Grad-CAM pixels reduced its confidence less than deleting random pixels. ConvNeXtTiny showed the same reversal, and for InceptionV3 the CAM-guided and random deletions were equivalent. For these three models the heatmaps therefore do not identify the evidence the network actually uses, despite looking anatomically reasonable. In contrast, deleting the regions highlighted for VGG16 reduced its confidence by 50.9 percentage points, 11.9 points more than random deletion, and Xception showed a similar margin (11.0 points). VGG16 thus combined the highest accuracy with the most faithful explanations, which strengthens confidence that its decisions rest on the anatomy its heatmaps display.

Table 10. Quantitative assessment of the Grad-CAM heatmaps

Model	ROI energy (%)	Pointing game (%)	Δ_{CAM} (%)	Δ_{rand} (%)	$\Delta_{CAM} - \Delta_{rand}$ (points)
VGG16	60.7	70.2	50.9	39.0	+11.9
Xception	61.8	75.0	41.5	30.6	+11.0
InceptionResNetV2	54.8	78.6	38.8	34.6	+4.2
InceptionV3	47.0	63.1	27.3	27.3	0.0
NASNetMobile	72.2	56.0	25.9	29.6	−3.7
ConvNeXtTiny	57.2	70.2	43.8	48.4	−4.6

Δ_{CAM} and Δ_{rand}: mean drop in predicted-class probability after zeroing the top-20% Grad-CAM pixels or 20% random pixels (Eq. 32).

This divergence is consistent with earlier findings that saliency methods can produce convincing maps that are insensitive to the model's parameters or decision (Adebayo et al., 2018). It shows that Grad-CAM figures should be accompanied by a faithfulness test before they are used to argue that a model is clinically trustworthy.

4.9 Feature-Space Analysis

The t-SNE projections of the 256-dimensional embeddings (**Figure 14**) revealed a structure common to every architecture: the test images formed two large lobes, each containing all four classes. Because this split appears regardless of architecture and is not aligned with the labels, it most likely reflects a non-diagnostic factor, such as slice level or acquisition source, that dominates the learned representation. Class structure was clearest for VGG16, whose MildDemented embeddings concentrated in the central region linking the two lobes, while NonDemented images occupied the outer parts of both lobes. ModerateDemented images formed small, tight clusters across networks, which is consistent with the very small number of unique images in this class. NonDemented and VeryMildDemented were intermixed for all models, explaining why this pair dominated the errors.

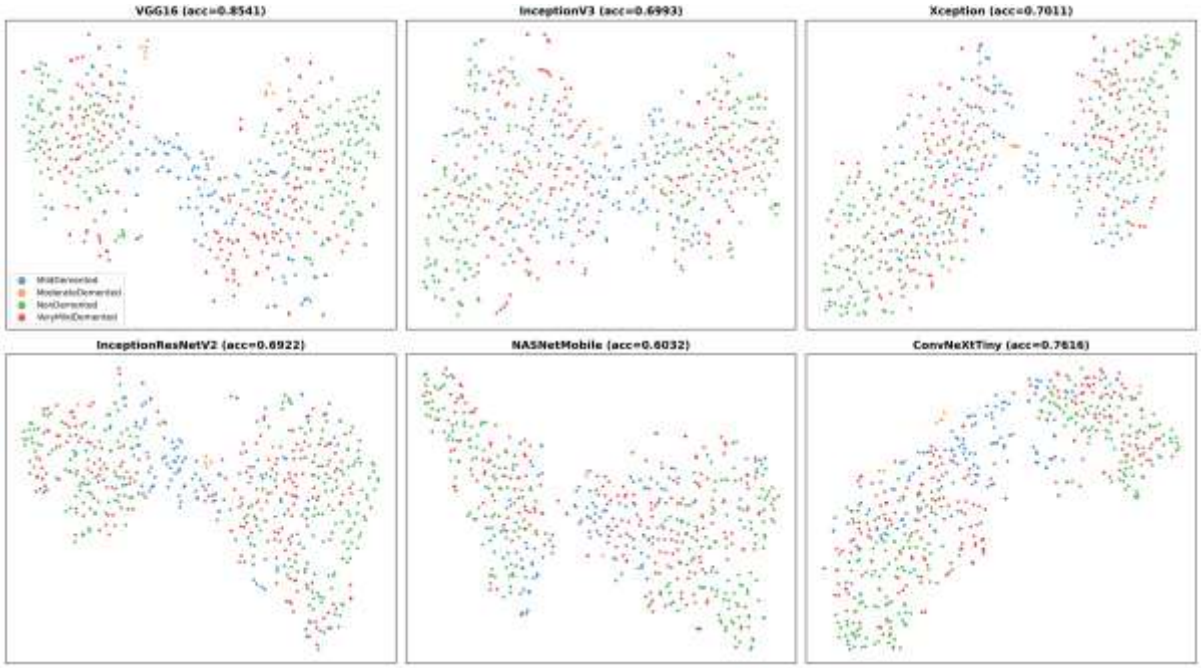


Figure 14: t-SNE projections of test-set embeddings

4.10 Comparison with Previous Studies and Effect of Data Curation

Many studies that use Kaggle Alzheimer MRI collections report accuracies above 95% for four-stage classification. The present results are lower, and the difference is largely explained by the curation steps of this study. More than a quarter of the files in the unaugmented release were exact duplicates, including 93.7% of the ModerateDemented images, and the release also offers an augmented folder whose files cannot be traced to their originals. When such collections are split at the image level without deduplication, identical or augmented copies of test images can enter the training set, and accuracy then measures recall of seen images rather than generalisation. Earlier analyses of brain MRI classification have shown that slice-level splits and data leakage can inflate accuracy substantially (Wen et al., 2020; Yagis et al., 2021). By removing exact duplicates, excluding untraceable augmented copies, using the test set only once and reporting confidence intervals and paired tests, the present benchmark trades headline accuracy for estimates that are more likely to hold on new data.

4.11 Limitations

Several limitations should be acknowledged. First, the dataset lacks subject identifiers, so the split was made at the image level and residual near-duplicate slices of the same subjects remain across subsets. Second, deduplication left only 61 ModerateDemented images, nine of them in the test subset, so the perfect scores for this class have wide uncertainty and should not be generalised. Third, all experiments used a single dataset, a single stratified split and a single random seed;

repeated or cross-validated splits and an external test cohort, such as data from an independent imaging study, would give firmer estimates. Fourth, the images are 2D JPEG slices of unknown acquisition protocol, so volumetric information and intensity calibration are unavailable. Fifth, the training budget was identical for all models, and the validation curves of the three strongest backbones were still rising when it ended. Finally, the explainability analysis measures plausibility against an automatically segmented brain mask and faithfulness by pixel deletion; neither substitutes for evaluation against expert-annotated regions.

5. CONCLUSION

This study benchmarked six transfer-learning CNNs from distinct design families for four-stage Alzheimer's disease classification from axial MRI slices under a leakage-aware protocol. Deduplication removed 26.6% of the unaugmented images, including 93.7% of the ModerateDemented class, which shows that data curation must precede any comparison of models on public slice-level collections. Under identical preprocessing, classification head and training schedule, VGG16 was the most accurate model, with 85.41% accuracy (95% CI 82.56–88.26%), a macro-F1 of 0.890, an MCC of 0.781 and a macro ROC-AUC of 0.971, and it was significantly better than each of the other five backbones. It was also the fastest at inference. Model size did not predict performance, and the depth of fine-tuning appeared more important than capacity on a dataset of this size. The validation-selected weighted ensemble reached 84.16% accuracy but did not improve on VGG16, because its members made correlated errors at the boundary between NonDemented and VeryMildDemented, which accounted for most misclassifications of all models. The explainability analysis showed that anatomically plausible heatmaps are not necessarily faithful. Every model concentrated its Grad-CAM energy on brain tissue, but for half of the evaluated networks removing the highlighted regions affected the prediction no more than removing random pixels. VGG16 combined the highest accuracy with the most faithful heatmaps. Visual explanations should therefore be accompanied by a quantitative faithfulness test before they are used to support clinical trust. Future work should evaluate the models with subject-level splits on cohorts that provide subject identifiers, validate them on independent multi-site data, extend the analysis to 3D volumes, examine longer and deeper fine-tuning, and compare the heatmaps with expert-annotated regions. Probability calibration and uncertainty estimation for the VeryMildDemented class, where clinical decisions are most difficult, are also important directions before such models can support screening in practice.

REFERENCES

[1] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2274-2282. https://doi.org/10.1109/TPAMI.2012.120

[2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps, In: Advances in Neural Information Processing Systems 31 (NeurIPS 2018), pp. 9505-9515. https://doi.org/10.48550/arXiv.1810.03292

[3] Bhardwaj, P., Bhandari, G., Kumar, Y., Gupta, S., 2022. An investigational approach for the prediction of gastric cancer using artificial intelligence techniques: a systematic review. Archives of Computational Methods in Engineering 29, 4379-4400.

[4] Bloch, L., Friedrich, C.M., 2024. Systematic comparison of 3D deep learning and classical machine learning explanations for Alzheimer's disease detection. Computers in Biology and Medicine 170, 108029. https://doi.org/10.1016/j.compbiomed.2024.108029

[5] Chamakuri, R., Janapana, H., 2025. A systematic review on recent methods on deep learning for automatic detection of Alzheimer's disease. Medicine in Novel Technology and Devices 25, 100343. https://doi.org/10.1016/j.medntd.2024.100343

[6] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 839-847. https://doi.org/10.1109/WACV.2018.00097

[7] Cheung, C.Y., Ran, A.R., Wang, S., Chan, V.T.T., Sham, K., Hilal, S., Venketasubramanian, N., Cheng, C.-Y., Sabanayagam, C., Tham, Y.C., Schmetterer, L., McKay, G.J., Williams, M.A., Wong, A., Au, L.W.C., Lu, Z., Yam, J.C.,

Tham, C.C., Chen, J.J., et al., 2022. A deep learning model for detection of Alzheimer's disease based on retinal photographs: a retrospective, multicentre case-control study. The Lancet Digital Health 4, e806-e815. https://doi.org/10.1016/S2589-7500(22)00169-8

[8] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1800-1807. https://doi.org/10.1109/CVPR.2017.195

[9] Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46. https://doi.org/10.1177/001316446002000104

[10] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[11] Easwaramoorthy, S.V., T., I., Venkatachalam, K., S., H., Elangovan, B., 2026. Deep learning for Alzheimer's disease and mild cognitive impairment detection, In: Ramamurthy, K. (Ed.), Deep Learning Applications in Neuroinformatics. Academic Press, pp. 81-103. https://doi.org/10.1016/B978-0-443-41459-6.00008-0

[12] Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, New York. https://doi.org/10.1201/9780429246593

[13] EL-Geneedy, M., Moustafa, H.E.-D., Khalifa, F., Khater, H., AbdElhalim, E., 2023. An MRI-based deep learning approach for accurate detection of Alzheimer's disease. Alexandria Engineering Journal 63, 211-221. https://doi.org/10.1016/j.aej.2022.07.062

[14] El-Sappagh, S., Abuhmed, T., Riazul Islam, S.M., Kwak, K.S., 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing 412, 197-215. https://doi.org/10.1016/j.neucom.2020.05.087

[15] Gao, S., Lima, D., 2022. A review of the application of deep learning in the detection of Alzheimer's disease. International Journal of Cognitive Computing in Engineering 3, 1-8. https://doi.org/10.1016/j.ijcce.2021.12.002

[16] Gorodkin, J., 2004. Comparing two K-category assignments by a K-category correlation coefficient. Computational Biology and Chemistry 28, 367-374. https://doi.org/10.1016/j.compbiolchem.2004.09.006

[17] Hassen, S.B., Neji, M., Hussain, Z., Hussain, A., Alimi, A.M., Frikha, M., 2024. Deep learning methods for early detection of Alzheimer's disease using structural MR images: a survey. Neurocomputing 576, 127325. https://doi.org/10.1016/j.neucom.2024.127325

[18] Jain, S., Srivastava, R., 2025. Enhanced EEG-based Alzheimer's disease detection using synchrosqueezing transform and deep transfer learning. Neuroscience 576, 105-117. https://doi.org/10.1016/j.neuroscience.2025.04.041

[19] Kaggle, 2026. Alzheimer Disease Dataset (Alzheimer_Dataset_V3), contributed by ashrafulhossenakash. Kaggle. https://www.kaggle.com/datasets/ashrafulhossenakash/alzheimer-disease-dataset. Date accessed: 29th October 2026.

[20] Kakkar, B., Johri, P., Kumar, Y., Park, H., Son, Y., Shafi, J., 2022. An IoMT-based federated and deep transfer learning approach to the detection of diverse chest diseases using chest X-rays. Human-centric Computing and Information Sciences 12.

[21] Kandasamy, V., Simic, V., Pamucar, D., Bacanin, N., 2026. Hierarchical probabilistic tissue modelling with deep learning for Alzheimer's disease detection from fluid-attenuated inversion recovery magnetic resonance imaging. Engineering Applications of Artificial Intelligence 181, 115508. https://doi.org/10.1016/j.engappai.2026.115508

[22] Kaur, K., Singh, C., Kumar, Y., 2023. Diagnosis and detection of congenital diseases in new-borns or fetuses using artificial intelligence techniques: a systematic review. Archives of Computational Methods in Engineering 30, 3031-3058.

[23] Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, In: 3rd International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1412.6980

[24] Koul, A., Kumar, Y., Modi, N., 2026. Deep learning-based detection of diabetic foot ulcers using convolutional neural networks and transfer learning. Procedia Computer Science 283, 4149-4158. https://doi.org/10.1016/j.procs.2026.06.475

[25] Kumar, Y., Bhardwaj, P., Shrivastav, S., Mehta, K., 2025a. Predicting paediatric brain disorders from MRI images using advanced deep learning techniques. Neuroinformatics 23, 9.

[26] Kumar, Y., Garg, P., Singh, R., Singh, C., 2025b. Optimized deep learning techniques for prediction and prognosis of severe and moderate knee arthritis: a comprehensive analysis.

[27] Kumar, Y., Koul, A., Kaur, S., Hu, Y.-C., 2023. Machine learning and deep learning based time series prediction and forecasting of ten nations' COVID-19 pandemic. SN Computer Science 4, 1-27.

[28] Lakhan, A., Grønli, T.-M., Muhammad, G., Tiwari, P., 2023. EDCNNS: Federated learning enabled evolutionary deep convolutional neural network for Alzheimer disease detection. Applied Soft Computing 147, 110804. https://doi.org/10.1016/j.asoc.2023.110804

[29] Lashkary, M., Bansal, K., Jain, S., Shukla, P.K., Tekchandani, H., Dhaka, V.S., 2025. Detection of Alzheimer's disease progression from structural MRIs using ConvMixer based deep learning model. Procedia Computer Science 258, 420-429. https://doi.org/10.1016/j.procs.2025.04.278

[30] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s, In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 11966-11976. https://doi.org/10.1109/CVPR52688.2022.01167

[31] Mahendran, N., P M, D.R.V., 2022. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. Computers in Biology and Medicine 141, 105056. https://doi.org/10.1016/j.compbiomed.2021.105056

[32] McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153-157. https://doi.org/10.1007/BF02295996

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

[34] Mohammed, E.M., Fakhrudeen, A.M., Alani, O.Y., 2024. Detection of Alzheimer's disease using deep learning models: a systematic literature review. Informatics in Medicine Unlocked 50, 101551. https://doi.org/10.1016/j.imu.2024.101551

[35] Mohammed, N., Mohammed, Z.A., Mohammed, S., Mohammed, N.A., Rajyalakshmi, P., 2026. IFMXCN: Intellectual flow map-based explainable deep learning model for Alzheimer's disease detection using multimodal input. Neurocomputing 664, 132171. https://doi.org/10.1016/j.neucom.2025.132171

[36] Mohan, R., Mali, S., Ilango, P., Jainulabdeen, J., 2027. Gather-excite attention-driven deep learning model for Alzheimer disease detection from MRIs using adaptive residual-DenseNet strategy. Psychiatry Research: Neuroimaging 364, 112344. https://doi.org/10.1016/j.pscychresns.2026.112344

[37] Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62-66. https://doi.org/10.1109/TSMC.1979.4310076

[38] Pace, C.D., Nardone, E., D'Alessandro, T., 2026. A dual backbone deep learning framework for early Alzheimer's disease detection via multimodal handwriting analysis. Pattern Recognition Letters 210, 117-123. https://doi.org/10.1016/j.patrec.2026.09.040

[39] Petsiuk, V., Das, A., Saenko, K., 2018. RISE: Randomized input sampling for explanation of black-box models, In: British Machine Vision Conference (BMVC). https://doi.org/10.48550/arXiv.1806.07421

[40] Prasath, T., Sumathi, V., 2024. Pipelined deep learning architecture for the detection of Alzheimer's disease. Biomedical Signal Processing and Control 87, 105442. https://doi.org/10.1016/j.bspc.2023.105442

[41] Ravikanti, D.K., Santhanam, S., 2026. An efficient framework for Alzheimer disease detection using improved PEBO-based feature integration and ensemble deep learning network. Biomedical Signal Processing and Control 120, 110102. https://doi.org/10.1016/j.bspc.2026.110102

[42] Rudraraju, A., Lakshmi, S.V., 2026. Alzheimer's disease severity detection and progression analysis using efficient deep learning techniques. Biomedical Signal Processing and Control 120, 110069. https://doi.org/10.1016/j.bspc.2026.110069

[43] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization, In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74

[44] Shah, P., Kumar, Y., Bhowmik, T., 2025. Deep learning techniques for the diagnosis and detection of orthopedic conditions: a systematic review of recent advances and challenges.

[45] Sharma, S., Guleria, K., Tiwari, S., Kumar, S., 2022. A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer disease using MRI scans. Measurement: Sensors 24, 100506. https://doi.org/10.1016/j.measen.2022.100506

[46] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, In: 3rd International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1409.1556

[47] Sisodia, P.S., Ameta, G.K., Kumar, Y., Chaplot, N., 2023. A review of deep transfer learning approaches for class-wise prediction of Alzheimer's disease using MRI images. Archives of Computational Methods in Engineering 30, 2409-2429. https://doi.org/10.1007/s11831-022-09870-0

[48] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning, In: Proceedings of the AAAI Conference on Artificial Intelligence 31, 4278-4284. https://doi.org/10.1609/aaai.v31i1.11231

[49] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception architecture for computer vision, In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[50] T.S., S., N., K., 2025. Taylor-dingo optimized RP-net for segmentation toward Alzheimer's disease detection and classification using deep learning. Computational Biology and Chemistry 116, 108359. https://doi.org/10.1016/j.compbiolchem.2025.108359

[51] Ullah, U., Habib, S., Islam, M., 2026. LADNET: An MRI-based deep learning approach for Alzheimer's disease detection. Computers in Biology and Medicine 208, 111647. https://doi.org/10.1016/j.compbiomed.2026.111647

[52] van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579-2605.

[53] Veluchamy, S., Bhuvaneswari, R., Ashwini, K., Alshathri, S., El-Shafai, W., 2025. HY-Deepnet: A new optimal deep transfer learning empowered framework for an autonomous Alzheimer's disease detection and diagnosis system. Engineering Science and Technology, an International Journal 68, 102058. https://doi.org/10.1016/j.jestch.2025.102058

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

[55] 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. https://doi.org/10.1038/s41598-021-01681-w

[56] Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S., 2018. Top-down neural attention by excitation backprop. International Journal of Computer Vision 126, 1084-1102. https://doi.org/10.1007/s11263-017-1059-x

[57] Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable image recognition, In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 8697-8710. https://doi.org/10.1109/CVPR.2018.00907

[58] Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization, In: Heckbert, P.S. (Ed.), Graphics Gems IV. Academic Press, San Diego, pp. 474-485. https://doi.org/10.1016/B978-0-12-336156-1.50061-6