

SpineFusion for plane aware local and global MRI classification of multiple sclerosis and myelitis

Salih Taha Alperen Özçelik

Bingöl University

Elif İrem Kılıç

Fırat University

Hasan Güler

hasanguler@firat.edu.tr

Fırat University

Article

Keywords: multiple sclerosis, myelitis, magnetic resonance imaging, cross-attention, ConvNeXt, Swin Transformer

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11162695/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SpineFusion for plane aware local and global MRI classification of multiple sclerosis and myelitis

Salih Taha Alperen Özçelik¹, Elif İrem Kılıç² & Hasan Güler^{3,*}

¹ Department of Electrical and Electronics Engineering, Faculty of Engineering, Bingöl University, Bingöl 12000, Türkiye; sozcelik@bingol.edu.tr; ORCID: 0000-0002-7929-7542

² Department of Electrical-Electronics Engineering, Faculty of Engineering, Fırat University, 23119 Elazığ, Türkiye; elifiremkillic@hotmail.com; ORCID: 0009-0003-9963-2429

³ Department of Electrical-Electronics Engineering, Faculty of Engineering, Fırat University, 23119 Elazığ, Türkiye; hasanguler@firat.edu.tr; ORCID: 0000-0002-9917-3619

* Correspondence: hasanguler@firat.edu.tr

Abstract

Deep learning models for MRI classification must combine local lesion appearance with broader anatomical context and acquisition orientation. We introduce SpineFusion, an end-to-end dual encoder that links ConvNeXt-Tiny and Swin-T representations through bidirectional cross-attention and an axial/sagittal plane embedding. We evaluated the architecture independently on three public datasets: a 2,746-image spinal multiple sclerosis (MS)-myelitis dataset (MSM), a 3,427-image brain MS/healthy dataset (BT-MS), and an audited 11,955-image, eight-class neurological dataset (MCND). Stratified image-level 10-fold cross-validation with inner validation yielded accuracies of $95.34 \pm 1.07\%$ on MSM, $97.78 \pm 0.85\%$ on BT-MS, and $99.18 \pm 0.35\%$ on MCND. On MSM, SpineFusion exceeded Swin-T by 1.56 percentage points but improved on ConvNeXt-Tiny by only 0.11 points, indicating that the convolutional branch provided most discriminative signal. Residual errors primarily involved MS-myelitis confusion. Grad-CAM showed that local activations often overlapped anatomy, but it did not explain the transformer branch or validate lesion localization. Because reliable subject identifiers were unavailable, neighboring slices may have crossed folds and the estimates may be optimistic. These results support the feasibility of plane-aware local-global fusion across independently trained MRI tasks; patient-grouped external validation remains necessary before clinical interpretation.

Keywords: *multiple sclerosis; myelitis; magnetic resonance imaging; cross-attention; ConvNeXt; Swin Transformer*

Introduction

Multiple sclerosis (MS) is an immune-mediated disorder of the central nervous system in which inflammatory demyelination and neuroaxonal injury produce lesions distributed across the brain and spinal cord. Magnetic resonance imaging (MRI) is central to demonstrating dissemination in space and time and to excluding alternative diagnoses [1, 2]. In the spinal cord, MS lesions are often short-segment and eccentric, whereas longitudinally extensive inflammatory myelitis may span several vertebral segments. These tendencies are clinically informative but are not absolute; lesion burden, acquisition plane, pulse sequence, image quality, and coexisting inflammation can make slice-level visual differentiation difficult [3]. Recent MS lesion research has likewise emphasized that limited annotated cohorts, noise, and incomplete imaging data can materially affect automated MRI analysis [16]. Computational decision support is therefore attractive, but evaluation protocols must prevent optimistic leakage and model outputs must remain research evidence rather than autonomous diagnoses.

The public MS-Myelitis spinal MRI collection introduced by Tatli et al. contains 2746 images in Healthy, MS, and Myelitis folders. Their transfer-transfer MSNet combined frozen DenseNet201 and ResNet50 features, three feature selectors, shallow classifiers, and iterative voting, reaching 97.63% accuracy and 97.23% F1 [4]. KhayyamNet subsequently combined Xception, a convolutional branch, and a Vision Transformer with minimum-redundancy maximum-relevance selection and a random forest, reporting $98.15 \pm 0.80\%$ accuracy on the same collection [5]. These systems demonstrate the discriminative value of complementary feature families, but both emphasize extracted features followed by a separate classifier rather than a single end-to-end local-global fusion network.

Transformer-based work also shows that contribution cannot be judged by accuracy alone. MS-Trust integrates a compact convolutional tokenizer, causal and global attention, channel recalibration, and MixUp/CutMix regularization; it was assessed on separate MS/healthy and MS/myelitis/healthy datasets with ablation and benchmark comparisons [6]. Its scientific contribution is architectural and methodological even where its reported accuracy is below feature-engineering systems. This distinction is important because results from different splits, patient-grouping rules, augmentations, and label taxonomies are not directly interchangeable.

Recent studies show parallel developments in local-global MRI modelling. A hybrid CNN-Transformer combined convolutional detail extraction with transformer context and paired the model with saliency analysis for multimodal glioma MRI [17]. A joint transformer architecture modeled inter-slice dependencies for three-dimensional brain MRI classification [18]. More recently, cross-attention was used to couple CNN-derived local features with transformer-derived global representations for multi-level neurological MRI classification [19]. Together, these studies support complementary local and global representations, while also showing that explanation scope and validation design remain part of the scientific claim.

Three gaps motivate the present study. First, local plaque texture and longer-range anatomical context should interact during optimization rather than be concatenated only after independent feature extraction. Second, axial and sagittal images encode different spatial evidence, yet a conventional image classifier receives no explicit plane identity. Recent externally validated work has shown that MRI orientation can itself be classified reliably, supporting its treatment as structured acquisition metadata rather than a nuisance variable [20]. Third, a model proposed for one small benchmark should be tested for architectural portability on additional MRI taxonomies, while keeping the datasets and their learned weights strictly separate.

The principal contributions are:

SpineFusion, an end-to-end dual-encoder network that couples ConvNeXt-Tiny local tokens and Swin-T global tokens through bidirectional cross-attention and conditions the classifier on a learned axial/sagittal plane embedding;

three independent, stratified 10-fold evaluations on a spinal three-class task, a brain two-class task, and an eight-class neurological task, with inner-validation early stopping and no sharing of weights or images across datasets; and

transparent ablation and qualitative analysis using encoder baselines, removal of the plane token, Grad-CAM examples, and representative residual error pairs.

A key limitation is established at the outset: the available image exports and experiment manifests did not provide reliable subject-grouping identifiers for all datasets. Cross-validation was therefore performed at image level. Neighboring slices from the same source examination may occur in different folds, and the results may overestimate patient-level generalization. All claims and comparisons in this article are restricted accordingly.

Classical MS MRI classification commonly separates representation learning from decision making. In MSNet, multiple pretrained-network layers generated feature vectors that were filtered by ReliefF, chi-square, and neighborhood component analysis before k-nearest-neighbor and support-vector-machine predictions were fused [4]. The approach attained high accuracy on the Firat spinal dataset but required a multi-stage feature-engineering pipeline. KhayyamNet broadened the representation set with Xception, CNN, and Vision Transformer features, selected variables by MRMR, and classified them with a random forest [5]. It currently provides the strongest reported accuracy on that public three-class benchmark, although its nested image-level protocol and shallow final classifier differ from the present end-to-end design.

For brain MRI, Macin et al. introduced the BT-MS collection and the computationally lightweight ExMPLPQ method. Their 3427 axial and sagittal images came from 72 individuals with MS and 59 healthy individuals; handcrafted local-phase-quantization features and Fine kNN reached 98.37%, 97.75%, and 98.22% accuracy for axial, sagittal, and combined subsets, respectively [7]. This dataset provides a second MS classification setting with different anatomy and a binary label space.

MS-Trust represents a complementary deep-learning direction. Its causal-global dual attention, squeeze-excitation recalibration, and regularization experiments emphasize model design, ablation, and subject grouping [6]. The study illustrates why raw accuracy should not be treated as a universal ranking: the number of subjects, the definition of the task, split granularity, and evaluation design govern the difficulty of the reported number.

SpineFusion builds on two modern visual backbones. ConvNeXt translates hierarchical convolutional design into a training recipe competitive with vision transformers and retains strong locality [9]. Swin Transformer forms hierarchical representations with shifted-window self-attention, allowing efficient aggregation beyond a single local neighborhood [10]. Multi-head attention provides a natural operator for exchanging information between token sets [11]. Instead of merging final logits, the proposed block lets local queries attend to global keys and values and simultaneously lets global queries attend to local keys and values. The plane embedding then supplies acquisition-orientation context before classification.

Results

Primary spinal benchmark

Table 2 summarizes the ten held-out test folds on MSM. SpineFusion attained $95.34 \pm 1.07\%$ accuracy, $94.60 \pm 1.34\%$ macro-F1, $94.59 \pm 1.36\%$ macro-recall, and 0.986 ± 0.005 macro-AUC. It exceeded Swin-T and was close to ConvNeXt-Tiny. The absolute accuracy gain over ConvNeXt-Tiny was 0.11 percentage points, which is small relative to fold-to-fold variability. The plane embedding increased accuracy from $94.90 \pm 1.50\%$ to $95.34 \pm 1.07\%$; this is reported as an ablation difference, not evidence of statistical superiority.

Table 2. Performance on MSM: reported literature results and ten-fold test mean $\pm$ standard deviation for the present experiments.

Method	Accuracy (%)	Macro-F1 (%)	Macro-recall (%)	Macro-AUC
Tatli et al. (reported) [4]	97.63	97.23	97.22	-
KhayyamNet (reported) [5]	98.15 ± 0.80	-	-	-
DenseNet201	95.05 ± 1.41	94.02 ± 1.68	94.08 ± 1.79	0.990 ± 0.004
ConvNeXt-Tiny	95.23 ± 1.19	94.42 ± 1.38	94.38 ± 1.37	0.986 ± 0.007
Swin-T	93.78 ± 2.70	92.62 ± 3.28	92.47 ± 3.33	0.986 ± 0.012
SpineFusion, no plane	94.90 ± 1.50	94.09 ± 1.86	94.04 ± 1.78	0.986 ± 0.008
SpineFusion	95.34 ± 1.07	94.60 ± 1.34	94.59 ± 1.36	0.986 ± 0.005

Literature rows are quoted from the cited publications and were not reproduced in the present pipeline. Their
feature extraction, nested validation, and classifier procedures are not identical to the end-to-end protocol
used here.

Class-wise results show that Healthy was the easiest category, with F1 $97.45 \pm 1.04\%$. MS was the most difficult, with F1 $91.30 \pm 2.33\%$, while Myelitis reached $95.04 \pm 2.19\%$. The asymmetry indicates that the remaining errors are concentrated in the clinically related MS-versus-Myelitis distinction rather than in separating abnormal from Healthy slices.

Independent evaluations on BT-MS and MCND

With separately initialized and trained weights, SpineFusion achieved $97.78 \pm 0.85\%$ accuracy on BT-MS and $99.18 \pm 0.35\%$ on MCND (Table 3). Macro-F1 was $97.70 \pm 0.89\%$ and $99.20 \pm 0.35\%$, respectively. These experiments demonstrate that the same architectural recipe can be optimized for different anatomies and label spaces; they are not external tests of a single frozen model and therefore do not establish cross-dataset domain generalization.

Table 3. SpineFusion performance on the independently trained secondary datasets, ten-fold test mean $\pm$ standard deviation.

Dataset	C	Accuracy (%)	Macro-F1 (%)	Macro-recall (%)	Macro-AUC
BT-MS (Dataset-2)	2	97.78 ± 0.85	97.70 ± 0.89	97.54 ± 1.00	0.994 ± 0.004
MCND (Dataset-3)	8	99.18 ± 0.35	99.20 ± 0.35	99.20 ± 0.34	0.999 ± 0.001

All folds are image-level. The unusually high MCND result may reflect repository-specific visual signatures and slice dependence; it should not be interpreted as patient-level or clinical performance.

Within MCND, the weakest class F1 values were observed for AD-MildDemented (0.980) and AD-VeryMildDemented (0.979), whereas Normal reached 1.000 and MS reached 0.999. This pattern is internally consistent with the visual similarity of adjacent Alzheimer severity labels, but the magnitude of the overall score also raises concern about image-source cues and correlated slices.

Qualitative review

Figure 5 shows correct held-out examples. In MSM, activation often overlapped the cord or spinal canal, although some maps extended into adjacent tissue. BT-MS maps emphasized portions of cerebral white matter and image boundaries. In MCND, the glioma example showed strong activation over the tumor region, while the MS and Normal examples exhibited more distributed responses. The examples support spatial plausibility in selected cases but do not constitute lesion localization validation.

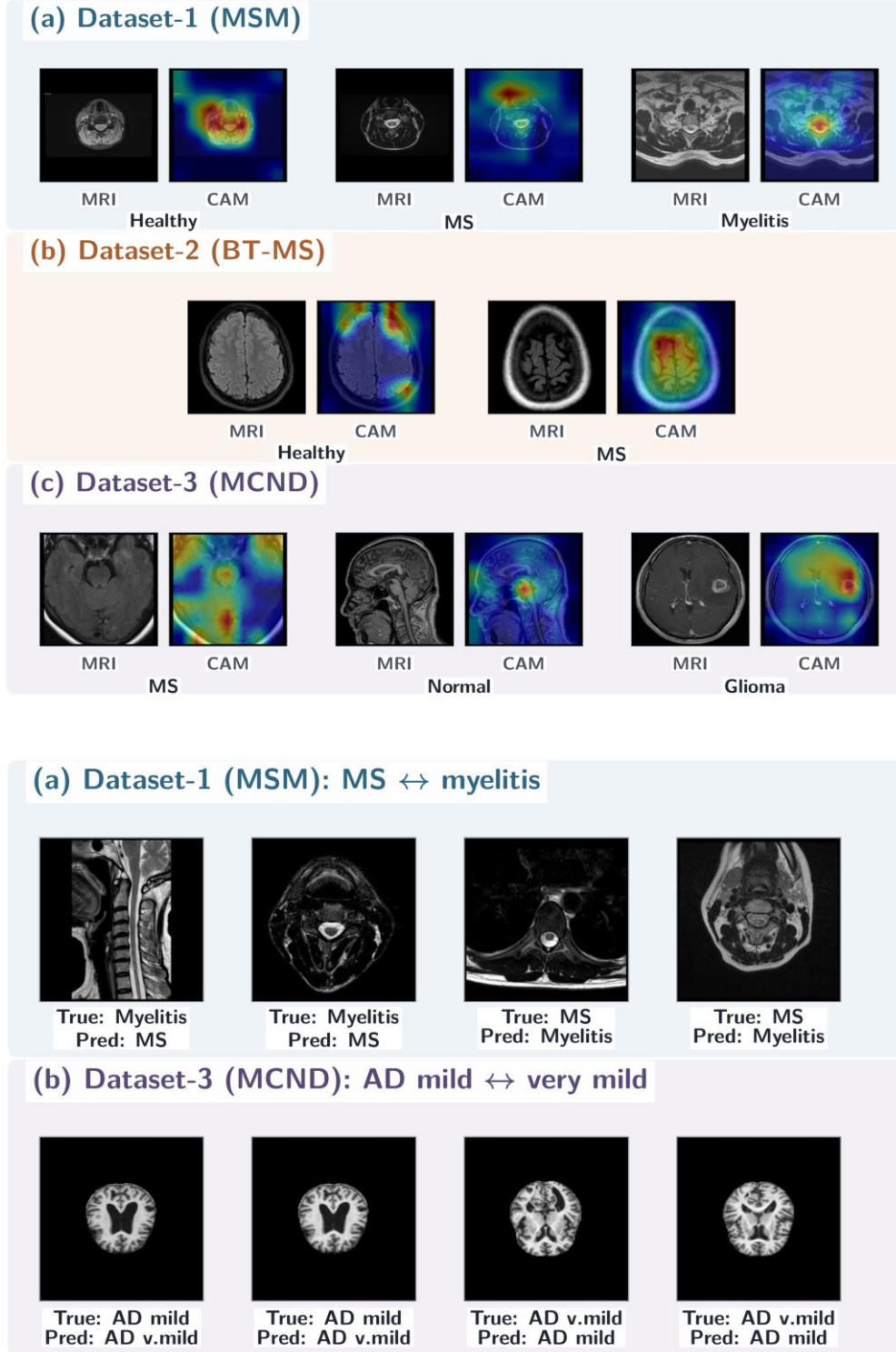


Figure 5. Qualitative sensitivity maps and representative residual errors. The upper montage presents Grad-CAM from the local ConvNeXt branch for correctly classified MSM, BT-MS, and MCND slices. The lower montage shows residual MSM confusion between MS and myelitis and MCND confusion between mild and very mild Alzheimer disease. Maps are illustrative sensitivity visualizations and not lesion-localization metrics.

Residual hard cases

Representative errors in Figure 5 align with the class-wise results. MSM mistakes occurred in both directions between MS and Myelitis and included axial as well as sagittal appearances. MCND mistakes commonly exchanged AD-MildDemented and AD-VeryMildDemented. BT-MS was omitted from the montage because its relatively few errors did not form one comparably dominant clinical pair.

Discussion

SpineFusion did not exceed the reported 97.63% MSNet or $98.15 \pm 0.80\%$ KhayyamNet accuracy on MSM [4, 5]. This result should be interpreted in light of method class and protocol rather than framed as a failed state-of-the-art claim. MSNet and KhayyamNet construct multiple frozen feature spaces, select variables, and apply shallow classifiers or voting. SpineFusion is optimized end to end and selects its checkpoint with an inner validation set that is separate from the outer test fold. The comparison rows are therefore informative context, not a controlled head-to-head experiment.

The ablation results temper the architectural claim. ConvNeXt-Tiny alone was already a strong local encoder, and fusion improved its mean accuracy by only 0.11 percentage points. In contrast, the fused system exceeded Swin-T by 1.56 points, suggesting that the convolutional branch contributes the dominant discriminative representation on this slice-level spinal benchmark. The plane token added 0.44 points over the otherwise identical no-plane variant. Because these differences are small compared with cross-fold standard deviations and were not supported by a prespecified paired test, the evidence supports architectural feasibility, not confirmed superiority.

The two secondary experiments demonstrate portability of the training recipe across brain and spinal images. BT-MS adds a binary brain MRI task with both axial and sagittal views, while MCND adds eight labels spanning neurodegeneration, tumors, MS, and Normal. However, each dataset was retrained independently, so the results do not show that learned features transfer between institutions or diseases. A stronger generalization experiment would train on one source and evaluate, without retraining, on an independently collected source with harmonized labels and patient-level partitions.

The 99.18% MCND accuracy warrants particular caution. The dataset aggregates images from multiple repositories, and class labels may correlate with source-specific acquisition, preprocessing, resolution, or background signatures. In addition, absent subject identifiers permit adjacent slices from the same volume to enter different folds. Both effects can make image-level recognition substantially easier than diagnosis in a new patient. The Grad-CAM and hard-case figures provide a useful qualitative audit, but they cannot rule out reliance on confounding features.

Further limitations include the aspect-ratio plane heuristic for MSM, lack of MRI sequence metadata, single-slice processing without three-dimensional context, and Grad-CAM computed only from the local branch. No external clinical cohort, uncertainty calibration, reader study, or lesion-level annotation was available. Future work should recover or construct subject identifiers, use grouped cross-validation, evaluate across scanners and centers, compare plane metadata with sequence-aware embeddings, and validate attention with segmentation or point annotations. Reporting parameter counts, FLOPs, calibration, and decision-curve utility would further clarify the practical trade-offs.

In conclusion, this study introduced SpineFusion, an end-to-end MRI classifier that exchanges local ConvNeXt-Tiny and global Swin-T token information through bidirectional cross-attention and adds a learned axial/sagittal plane embedding before classification. Across three independently trained image-level 10-fold experiments, the method reached $95.34 \pm 1.07\%$ accuracy on the primary MSM spinal task, $97.78 \pm 0.85\%$ on BT-MS, and $99.18 \pm 0.35\%$ on MCND. On MSM, performance was close to the stronger ConvNeXt baseline and residual errors centered on MS versus Myelitis. The work establishes a reproducible fusion framework and an honest baseline for further investigation; patient-wise, multi-center validation is required before clinical interpretation.

Methods

Datasets

Three public MRI collections were analyzed. They differ in anatomy, label taxonomy, and source organization and were never concatenated into one classifier. A fresh SpineFusion model was trained for each dataset (Figure 1). Dataset names were standardized as MSM (Dataset-1), BT-MS (Dataset-2), and MCND (Dataset-3).

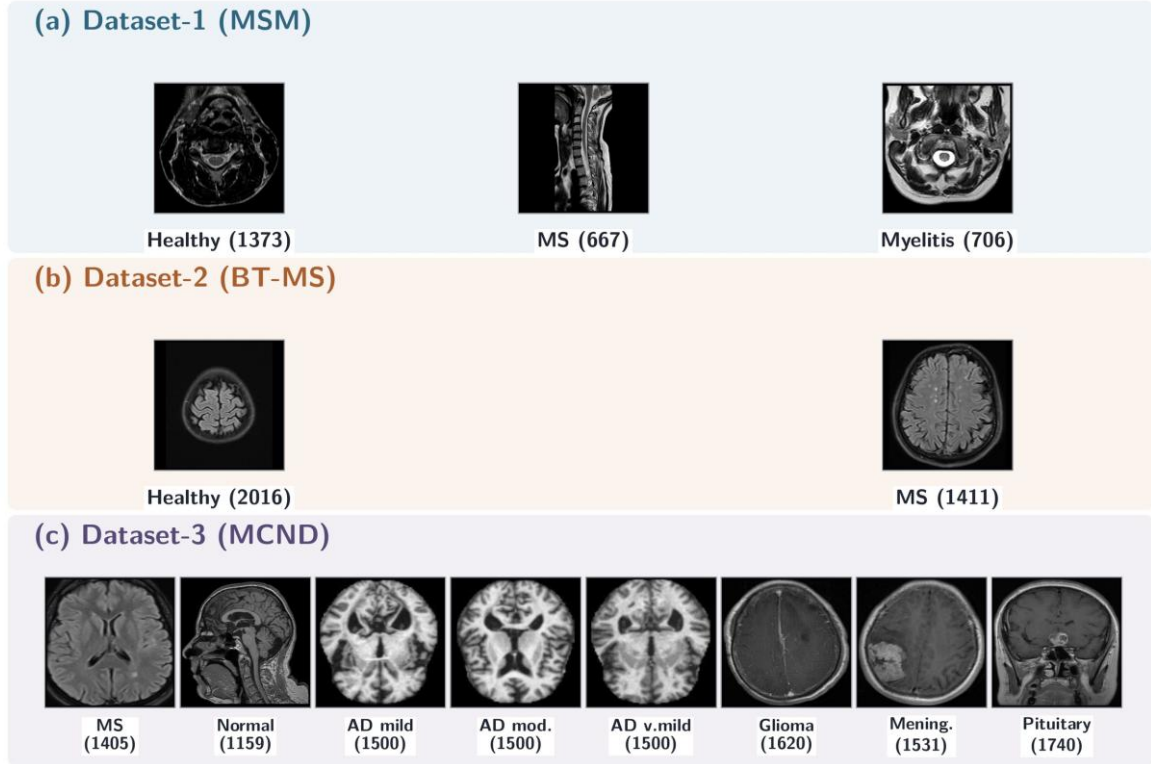


Figure 1. Representative slices from the three datasets, trained and evaluated independently. (a) Dataset-1 (MSM): MS-Myelitis Spinal MRI, Healthy (1373), MS (667), and Myelitis (706). (b) Dataset-2 (BT-MS): Burak Tasci MS Brain MRI, Healthy (2016) and MS (1411). (c) Dataset-3 (MCND): eight neurological classes; one image per class is shown, with audited class counts in parentheses.

Table 1. Audited dataset composition used in the experiments.

Dataset	Anatomy	Classes (C)	Images (N)	Plane source
MSM (Dataset-1)	Spinal	Healthy, MS, Myelitis (3)	2746	Aspect-ratio heuristic
BT-MS (Dataset-2)	Brain	Healthy, MS (2)	3427	Filename: axial/sagittal
MCND (Dataset-3)	Brain	3 AD stages, 3 tumors, MS, Normal (8)	11,955	Heuristic; all axial in audit

MSM folder counts were Healthy 1373, MS 667, and Myelitis 706. The MCND data card advertises 16,400 images, whereas the downloaded and programmatically audited eight-folder copy used here contained 11,955 readable images; all reported results use the audited copy.

MSM originated from Firat University and was released with the MSNet study [4]; the public copy is hosted on Kaggle and is also cited by KhayyamNet [5]. The local manifest contained 1373 Healthy, 667 MS, and 706 Myelitis images. Some secondary descriptions exchange the MS and Myelitis counts; this work follows the class-folder labels in the audited download. The filenames follow generic patterns such as '1 (n).jpg' and do not provide dependable subject identifiers. Plane labels were inferred from aspect ratio, with wide images treated as sagittal:

$$plane(I) = \text{sagittal if } W/H \geq 1.35; \text{ axial otherwise}$$

(1)

The resulting MSM plane counts were 577 sagittal and 2169 axial images. The heuristic is an engineering approximation and is treated as a limitation rather than a radiological annotation.

BT-MS is the public brain MRI dataset associated with ExMPLPQ [7]. Its source study reports 72 participants with MS and 59 healthy participants, with 1652 axial and 1775 sagittal images. The audited class folders contained 2016 Healthy and 1411 MS images. Plane labels were parsed from the axial/sagittal filename structure. The experiment manifest did not include a validated subject-to-slice grouping map, so the present folds are image-level despite the subject counts reported by the source study.

MCND is a Kaggle compilation of Alzheimer disease, brain-tumor, MS, and healthy MRI sources [8]. The audited working copy contained AD-MildDemented (1500), AD-ModerateDemented (1500), AD-VeryMildDemented (1500), BT-glioma (1620), BT-meningioma (1531), BT-pituitary (1740), MS (1405), and Normal (1159), for a total of 11,955 images. All were labeled axial by the pipeline's scan. Because MCND aggregates 2D images from several pre-existing repositories without a common subject identifier, its high slice-level scores should be interpreted cautiously.

Preprocessing and augmentation

Every image was converted to three channels and letterboxed to 224×224 pixels. Letterboxing preserves the native aspect ratio and adds zero-valued padding rather than cropping the anatomy. Pixel values were scaled to $[0, 1]$ and normalized with ImageNet channel statistics [15]:

$$\hat{x} = (x/255 - \mu)/\sigma, \quad \mu = (0.485, 0.456, 0.406), \quad \sigma = (0.229, 0.224, 0.225) \quad (2)$$

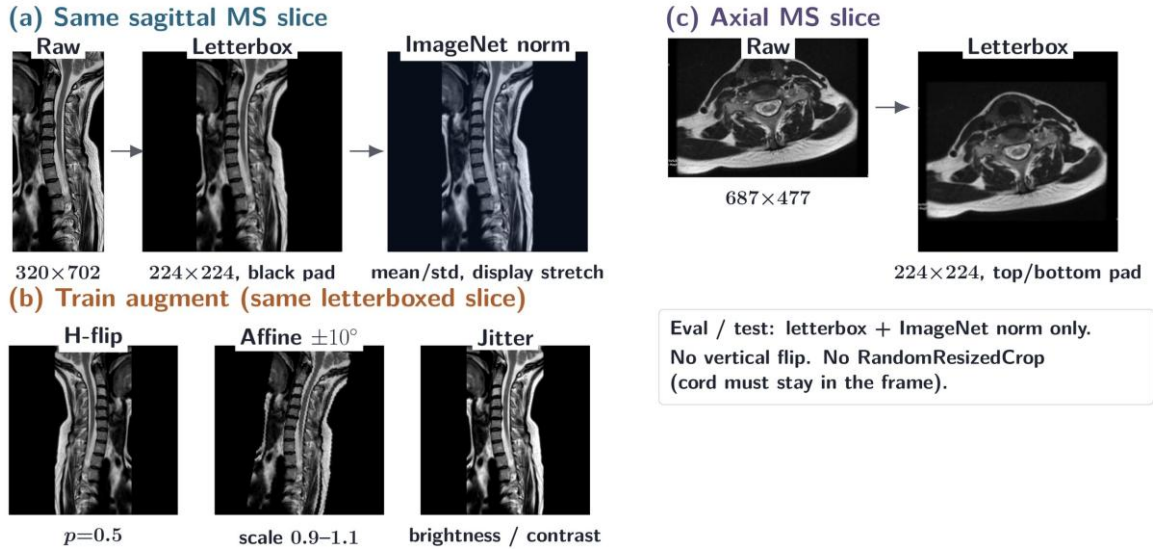


Figure 2. Preprocessing on Dataset-1 (MSM) slices. (a) The same sagittal MS slice at native resolution, after 224×224 letterboxing, and after ImageNet normalization (display-stretched). (b) Training augmentations of the letterboxed slice: horizontal flip, affine rotation up to ± 10 degrees with 5% translation and 0.9-1.1 scaling, and brightness/contrast jitter. (c) An axial MS slice, for which letterboxing pads the top and bottom. Evaluation uses letterboxing and normalization only. Vertical flip and RandomResizedCrop are disabled so that the cord remains in frame and plane identity is preserved.

Training-only augmentation comprised horizontal flipping with probability 0.5, random affine transformation (rotation ± 10 degrees, translation up to 5%, and scale 0.9-1.1), and ColorJitter with magnitude 0.15 for brightness and contrast. Vertical flipping and RandomResizedCrop were disabled. Mixup was configured with $\alpha = 0$ in the reported runs. Validation and test images underwent only letterboxing and normalization.

Evaluation protocol

Each dataset was evaluated independently with stratified 10-fold cross-validation using seed 42 (Figure 3). In each outer fold, approximately 10% of images formed the held-out test set. The remaining images were divided again, with 10% used as an inner validation set and the balance used for training. Augmentation was applied only to training images. Validation macro-F1 controlled early stopping (patience 7; maximum 20 epochs), and the checkpoint with the best validation macro-F1 was loaded before the test fold was evaluated once.

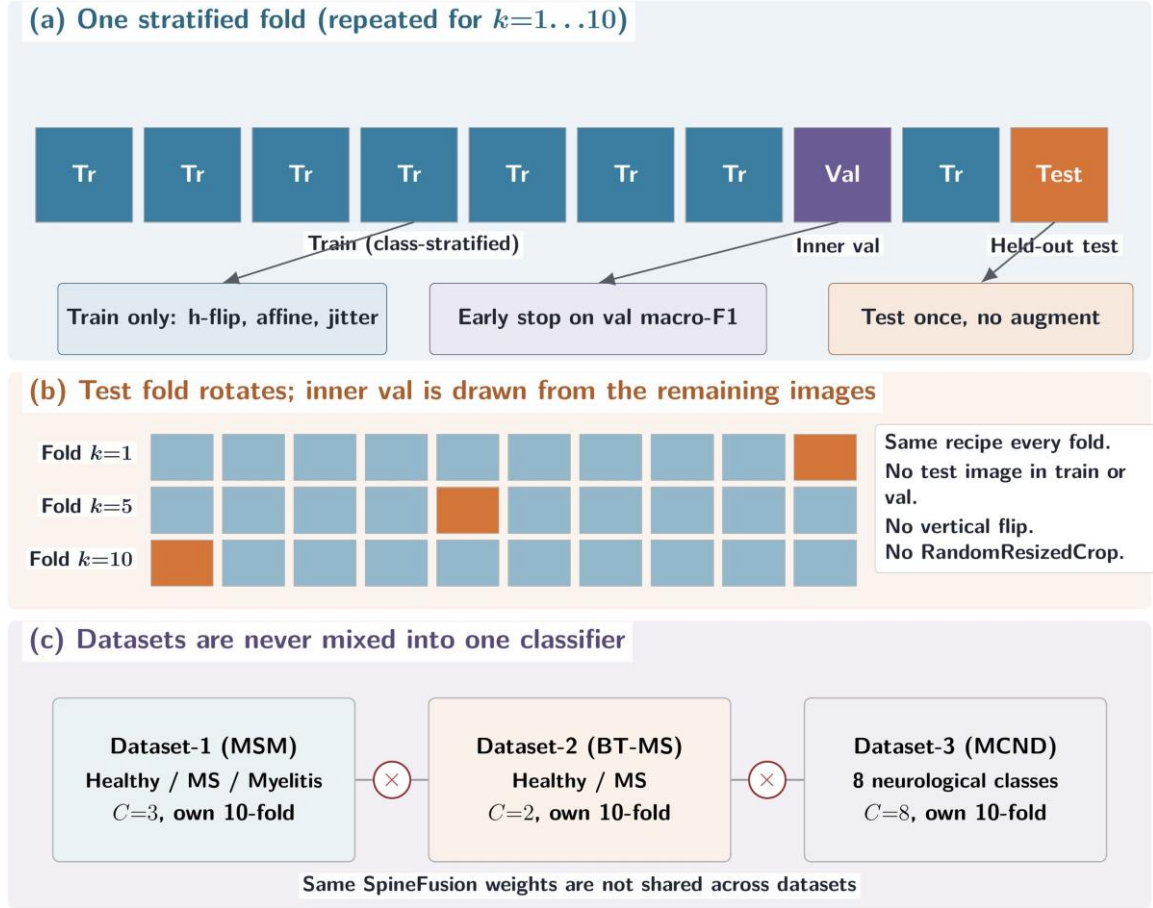


Figure 3. Independent evaluation protocol. (a) One stratified fold contains a class-stratified training set, an inner validation set, and a held-out test set. Augmentation is applied only to training; early stopping monitors validation macro-F1; and the test fold is evaluated once. (b) The test block rotates over $k = 1, \dots, 10$ while inner validation is drawn from the remaining images. (c) MSM, BT-MS, and MCND are never mixed, and model weights are not shared across datasets. Splitting is image-level, not patient-level.

We report the mean and sample standard deviation across the ten held-out test folds for accuracy, macro-precision, macro-recall, macro-F1, weighted F1, and macro-area under the receiver-operating-characteristic curve (macro-AUC). For multiclass tasks, AUC was calculated one-versus-rest and macro-averaged. Per-class metrics were computed from each fold and then averaged. Cross-fold standard deviations describe protocol variability; they are not independent-patient confidence intervals. No hypothesis test was prespecified for model superiority, so small differences are described relative to fold variability without claims of statistical significance.

SpineFusion architecture

SpineFusion combines a convolutional local encoder, a shifted-window transformer global encoder, bidirectional cross-attention, and a learned plane embedding (Figure 4). The output dimension C is set to 3 for MSM, 2 for BT-MS, and 8 for MCND. The architecture is the same across tasks, but all weights are trained separately.

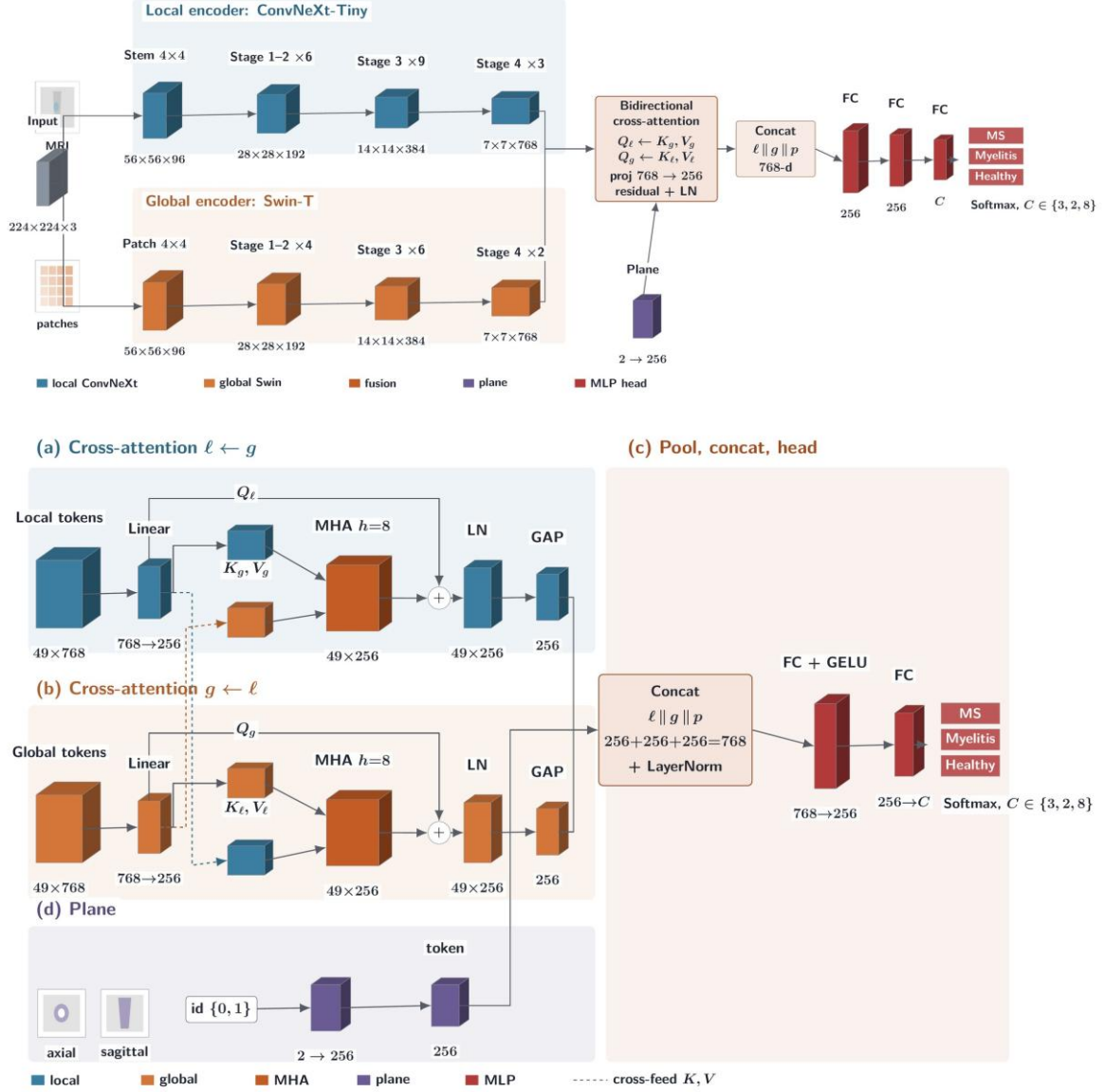


Figure 4. SpineFusion architecture and fusion design. The upper diagram shows the local ConvNeXt-Tiny and global Swin-T encoders, bidirectional cross-attention, learned plane embedding, and classification head. The lower diagram details the two cross-attention paths, residual normalization, pooled local and global vectors, and concatenation with the plane vector.

Encoders

For input x in $R^{(B \times 3 \times 224 \times 224)}$, the ConvNeXt-Tiny branch [9] yields a $7 \times 7 \times 768$ terminal feature volume. After channel normalization, spatial positions are flattened into T_l in $R^{(B \times 49 \times 768)}$. Its convolutional inductive bias is intended to preserve local plaque texture. The Swin-T branch [10] uses a 4×4 patch embedding and hierarchical shifted-window attention, also producing T_g in $R^{(B \times 49 \times 768)}$. It supplies broader cord or brain context. Separate linear maps project both token sets to the fusion width $d = 256$:

$$Z_l = T_l W_l, Z_g = T_g W_g, W_l, W_g \in R^{(768 \times 256)} \quad (3)$$

Bidirectional cross-attention

The fusion block uses eight attention heads (head dimension 32). Scaled dot-product attention follows the standard transformer operator [11]:

$$\text{Attn}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_h})V \quad (4)$$

Local queries attend to global keys and values, while global queries attend to local keys and values. Each path then applies residual addition followed by layer normalization:

$$H_l = \text{LN}[Z_l + \text{MHA}(Z_l, Z_g, Z_g)], H_g = \text{LN}[Z_g + \text{MHA}(Z_g, Z_l, Z_l)] \quad (5)$$

Global average pooling over the 49 positions produces h_l and h_g in R^{256} . This symmetric exchange differs from simple feature concatenation because local and global tokens condition one another before spatial pooling.

Plane embedding

A trainable lookup E in $R^{(2 \times 256)}$ maps plane identity (0 = axial, 1 = sagittal) to p in R^{256} . For the no-plane ablation, p is replaced by the zero vector while the remainder of the architecture is unchanged. Plane identity is contextual metadata rather than a target and is not predicted by the model.

$$p = E[\text{plane}] \quad (6)$$

Classification head

The pooled vectors and plane embedding are concatenated and normalized. A two-layer multi-layer perceptron with GELU activation maps 768 features to 256 hidden units and then to C class logits. Dropout 0.3 is applied before each linear layer, and softmax converts logits to class probabilities:

$$u = \text{LN}([h_l; h_g; p]), z = W_2 \text{GELU}(W_1 u), \hat{y} = \text{softmax}(z) \quad (7)$$

Loss and optimization

Class weights were computed separately from the training partition of each fold using inverse frequency, $w_c = N/(Cn_c)$. The implementation applies weighted, label-smoothed cross-entropy (epsilon = 0.05) and then a focal modulation with gamma = 1.5 [12]. For sample i , the operational loss can be summarized as:

$$L = (1/B)\sum_i (1 - p_{(i,y_i)})^\gamma w_{(y_i)} CE_i \quad (8)$$

Optimization used AdamW [13] with learning rate 3×10^{-5} and weight decay 0.05, cosine learning-rate decay, and two warm-up epochs. During epoch 1 the pretrained encoders were frozen and only the fusion/head parameters were optimized; full fine-tuning followed. Batch size was 8 with automatic mixed precision. Training stopped at 20 epochs or after seven epochs without improvement in validation macro-F1.

Baselines and ablation

On MSM, four comparisons used the same preprocessing and fold protocol: DenseNet201, ConvNeXt-Tiny, Swin-T, and SpineFusion without the plane embedding. DenseNet201 provides an end-to-end CNN reference related to one of the feature networks used by MSNet; ConvNeXt-Tiny and Swin-T isolate the local and global encoders; and the no-plane variant tests the incremental effect of orientation metadata. Reported MSNet and KhayyamNet results are included as literature values, not reruns, because their code paths and validation procedures differ from ours.

Explainability and error review

Grad-CAM [14] was computed from the final ConvNeXt feature maps by differentiating the relevant class score, weighting activation channels by global-average-pooled gradients, applying a rectifier, and bilinearly upsampling the map to 224×224 . Maps were overlaid on letterboxed input images. They are used only as qualitative indications of spatial sensitivity; no lesion masks were available, and therefore no localization accuracy is claimed. Representative correct predictions and recurring error pairs were selected from held-out folds for Figure 5.

Ethics declarations

Ethics approval. Ethical review and approval were waived because this study used only publicly available, de-identified secondary image datasets and involved no new recruitment, intervention, or access to identifiable clinical records.

Informed consent. Not applicable to this secondary analysis of public, de-identified datasets.

Use of generative artificial intelligence

During manuscript preparation, the authors used OpenAI Codex (GPT-5 family) to assist with English-language drafting, document formatting, and consistency checks. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Data availability

Publicly available datasets were analyzed in this study. The MS–Myelitis Spinal MRI dataset is available at <https://www.kaggle.com/datasets/turkertuncer/ms-myelitis>; the Multiple Sclerosis Brain MRI dataset is available at <https://www.kaggle.com/datasets/buraktaci/multiple-sclerosis>; and the Multi-Class Neurological Disorder (MCND) dataset is available at <https://www.kaggle.com/datasets/alifatahi/multi-class-neurological-disorder-mcnd-dataset>. No new patient-level clinical data were generated in this study.

References

- [1] Thompson, A. J. et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162-173 (2018). [https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2).
- [2] Filippi, M. et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol.* **15**, 292-303 (2016). [https://doi.org/10.1016/S1474-4422\(15\)00393-2](https://doi.org/10.1016/S1474-4422(15)00393-2).
- [3] Beh, S. C., Greenberg, B. M., Frohman, T. & Frohman, E. M. Transverse myelitis. *Neurol. Clin.* **31**, 79-138 (2013). <https://doi.org/10.1016/j.ncl.2012.09.008>.
- [4] Tatli, S. et al. Transfer-transfer model with MSNet: an automated accurate multiple sclerosis and myelitis detection system. *Expert Syst. Appl.* **236**, 121314 (2024). <https://doi.org/10.1016/j.eswa.2023.121314>.
- [5] Dehghanpour, M., Fateh, M., Mohammadpoory, Z. & Ferdowsi, S. KhayyamNet: a parallel multiscale feature fusion framework for accurate diagnosis of multiple sclerosis and myelitis. *Mach. Learn. Knowl. Extr.* **8**, 62 (2026). <https://doi.org/10.3390/make8030062>.
- [6] Alzahrani, S. M. MS-Trust: a transformer model with causal-global dual attention for enhanced MRI-based multiple sclerosis and myelitis detection. *Complex Intell. Syst.* **11** (2025). <https://doi.org/10.1007/s40747-025-01945-2>.
- [7] Macin, G. et al. An accurate multiple sclerosis detection model based on exemplar multiple parameters local phase quantization: ExMPLPQ. *Appl. Sci.* **12**, 4920 (2022). <https://doi.org/10.3390/app12104920>.
- [8] Fatahi, A. et al. MCND dataset-multi-class brain MRI scans. *Kaggle*. <https://www.kaggle.com/datasets/alifatahi/multi-class-neurological-disorder-mcnd-dataset> (accessed 15 August 2026).
- [9] Liu, Z. et al. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11976-11986 (2022). <https://doi.org/10.1109/CVPR52688.2022.01167>.

- [10] Liu, Z. et al. Swin Transformer: hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012-10022 (2021). <https://doi.org/10.1109/ICCV48922.2021.00986>.
- [11] Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998-6008 (2017).
- [12] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, 2980-2988 (2017). <https://doi.org/10.1109/ICCV.2017.324>.
- [13] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (2019).
- [14] Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 618-626 (2017). <https://doi.org/10.1109/ICCV.2017.74>.
- [15] Deng, J. et al. ImageNet: a large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248-255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>.
- [16] Alsayat, A. et al. RAUM-GANs: a multi-layer GAN-enhanced framework for accurate multiple sclerosis lesion segmentation in MRI. *Sci. Rep.* **15**, 43933 (2025). <https://doi.org/10.1038/s41598-025-27787-z>.
- [17] Zeineldin, R. A. et al. Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Sci. Rep.* **14**, 3713 (2024). <https://doi.org/10.1038/s41598-024-54186-7>.
- [18] Alp, S. et al. Joint transformer architecture in brain 3D MRI classification: its application in Alzheimer's disease classification. *Sci. Rep.* **14**, 8996 (2024). <https://doi.org/10.1038/s41598-024-59578-3>.
- [19] Veerlapalli, P. et al. Cross-attention guided explainable deep transformer model for multi-level classification of rare neurological disorders using MRI images. *Sci. Rep.* **16**, 19357 (2026). <https://doi.org/10.1038/s41598-026-58491-1>.
- [20] Kim, J. et al. Automated characterization of abdominal MRI exams using deep learning. *Sci. Rep.* **15**, 27044 (2025). <https://doi.org/10.1038/s41598-025-11985-w>.

Author contributions

Conceptualization, E.I.K.; methodology, E.I.K. and S.T.A.Ö.; software, S.T.A.Ö. and E.I.K.; formal analysis, H.G.; validation, H.G. and S.T.A.Ö.; investigation, S.T.A.Ö. and E.I.K.; writing—original draft preparation, S.T.A.Ö.; writing—review and editing, S.T.A.Ö., E.I.K. and H.G.; supervision, H.G. All authors reviewed and approved the final manuscript.

Funding

434

This research was funded by the Scientific Research Projects Unit of Firat University.

435

Competing interests

436

The authors declare no competing interests.

437