2026 IEEE International Conference on Signal Processing, Information, Communication and Systems (SPICSCON)
13-14 August 2026, Faculty of ECE, Bangladesh Army University of Engineering & Technology (BAUET), Qadirabad, Natore-6431, Bangladesh

TriFusion-ResViTNet: A Hybrid Feature-Level Fusion Framework for MRI-Based Rare Neurological Disease Classification

1st Safeyeatur Rahman
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
srsaman16653@gmail.com

2nd Syeda Jannatul Ferdous
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
syedajannatulferdous96@gmail.com

3rd Md. Abid Hasan Apurbo
Department of Computer Science
BRAC University
Dhaka-1212, Bangladesh
abidhasanapurbo@gmail.com

4th Monaum Islam Rifat
MBBS
Eastern Medical College
Cumilla, Bangladesh
monayemrifat180@gmail.com

5th Sharupya Barua
Department of Computer Science and Engineering
Premier University
Chittagong, Bangladesh
engineer.sharupyabarua@gmail.com

6th Masiat Mohammad Momshad
Department of Computer Science
BRAC University
Dhaka-1212, Bangladesh
masiatmohammad007@gmail.com

7th Tauhidur Rahman Sakib*
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
tauhidurrahmansakib@gmail.com

***Abstract*—MRI-based classification of rare neurological disorders remains difficult because disease patterns overlap and public data are limited. This paper presents TriFusion-ResViTNet, a feature-level fusion framework combining ResNet50 spatial features, Vision Transformer contextual features, and radiomics-inspired texture-frequency descriptors. The branch embeddings are concatenated, jointly processed, and adaptively reweighted before classification. Experiments used a balanced, quality-controlled Kaggle subset of 1,400 MRI images across five disease classes with an 80:10:10 stratified split. The model achieved 97.9% accuracy, a weighted F1-score of 0.977, and a ROC-AUC of 0.992, outperforming baseline and ablation models under the same protocol. However, the improvement over ResNet50 was modest and required approximately twice the computational cost. Interpretability analyses were qualitative only. Since patient identifiers and acquisition metadata were unavailable and external validation included only 60 images, the findings demonstrate feasibility rather than confirmed clinical generalization.**

***Index Terms*—MRI classification, neurological disease classification, Vision Transformer, ResNet50, radiomics, feature-level fusion, gated attention, deep learning.**

I. Introduction

Neurological MRI classification remains challenging because rare diseases often exhibit subtle and overlapping anatomical patterns, while available datasets are limited and imaging characteristics vary across scanners. Deep learning has improved automated neurological disease recognition through convolutional networks, transfer learning, and computer-aided neuroimaging analysis [1], [2]. Earlier methods relied on structural MRI and handcrafted features for white-matter disorders and epilepsy classification [3], [4], whereas later CNN-based approaches addressed multi-class brain MRI, ischemic stroke, hydrocephalus, and demyelinating diseases [5]–[8]. Recent studies have explored MRI biomarkers, transfer learning, transformer-assisted augmentation, radiomics, and multimodal fusion [9]–[13]. Autoencoder ensembles and CNN-based multimodal frameworks have also demonstrated the value of combining complementary representations in biomedical classification [14], [15]. However, many hybrid methods integrate model decisions rather than allowing different feature representations to interact directly. Convolutional networks provide strong local spatial representations but limited global context [17], whereas vision transformers capture long-range dependencies but generally require more training data [18]. Texture-frequency descriptors offer compact and stable information but lack adaptive learning capability

979-8-3195-1270-3/26/$31.00 ©2026 IEEE

[25]. To exploit their complementary strengths, this paper proposes TriFusion-ResViTNet, which combines ResNet50, Vision Transformer, and radiomics-inspired representations through feature concatenation, cross-branch interaction, and gated adaptive weighting before a single joint classifier.

The contribution is primarily engineering and empirical rather than foundational. Established components are integrated into a feature-level fusion framework and systematically evaluated for rare neurological disease classification through baseline comparisons, ablation analysis, external validation, interpretability assessment, and statistical testing. The main contributions are as follows:

- **Feature-Level TriFusion** – CNN, transformer, and radiomics-inspired embeddings are concatenated, jointly interacted, and adaptively reweighted before a single classifier.
- **Complementary Representations** – A five-variant ablation examines the non-redundant contributions of local spatial, global contextual, and texture-frequency features under limited data.
- **Transparent Evaluation** – Six matched-protocol baselines, computational trade-off analysis, per-case errors, and statistical validation are reported.
- **Bounded Interpretability** – Grad-CAM, edge, contour, and intensity analyses are presented as qualitative evidence, with their unquantified limitations stated explicitly.

The remainder of this paper is organized as follows: Section II describes the proposed methodology, Section III presents the experimental results and analysis, and Section IV concludes the study and discusses future work.

II. Methodology

A. Dataset Overview

Experiments used 1,400 quality-controlled 2D grayscale images from the public Kaggle dataset "Rare Neurological Diseases MRI–Curated Edition" [16], covering FMD, HSD, MD, PCH, and WWS. Patient identifiers, counts, demographics, institutions, scanner details, acquisition parameters, and raw DICOM/NIfTI data were unavailable. Readable, correctly labelled, anatomically informative images were retained, while corrupted, noisy, low-quality, duplicate or near-duplicate, incomplete, mislabelled, and ambiguous images were excluded with medical-reviewer support. The dataset was balanced at 280 images per class, which reduces class bias but does not reflect clinical prevalence. An 80:10:10 stratified split yielded 1,120 training, 140 validation, and 140 test images. External validation used 60 unseen images, with the distribution reported in Table I, and therefore provides only preliminary cross-domain evidence. Because patient or study identifiers were unavailable, patient-level separation could not be confirmed; duplicates were removed before splitting and augmentation was limited to training, although residual multi-slice overlap cannot be excluded.

TABLE I
Class Distribution Across Partitions

Class	Train	Val.	Test	Primary Total	Ext.
FMD	224	28	28	280	11
HSD	224	28	28	280	11
MD	224	28	28	280	12
PCH	224	28	28	280	11
WWS	224	28	28	280	15
Total	**1120**	**140**	**140**	**1400**	**60**

B. Data Preprocessing and Augmentation

All images were resized to 224×224 pixels. Since the released data are pre-exported 8-bit grayscale images, intensities in $[0, 255]$ were scaled to $[0, 1]$ as

$$I_{\text{norm}}(x, y, c) = \frac{I(x, y, c)}{255}. \tag{1}$$

This applies only to the pre-quantized image files, not raw DICOM or NIfTI data, which may require percentile clipping, brain-mask z-score normalization, or bias-field correction. Each image was replicated into three identical channels to match the ImageNet-pretrained ResNet50 and ViT inputs. This adds no image information but increases input and first-layer computation; future work will adopt a single-channel stem. Random rotation, horizontal flipping, zooming, and width and height shifts were applied only to training images. Validation and test images received deterministic resizing and normalization. Splitting preceded augmentation to prevent leakage, and the test set was evaluated once after validation-based model selection.

C. Proposed TriFusion-ResViTNet Architecture

The architecture in Fig. 1 uses three parallel feature-extraction branches followed by shared feature-level fusion and a single classifier. The branches learn complementary representations but do not make independent predictions or use voting.

Spatial branch. An ImageNet-pretrained ResNet50 [17] extracts hierarchical anatomical, boundary, and local disease features. Its strong locality prior suits limited data, yielding 97.1% standalone accuracy in Table II, although long-range relationships are learned only through deeper layers. Global average pooling produces f_R.

Contextual branch. A Vision Transformer [18] divides the image into patches and applies positional encoding and multi-head self-attention to model global dependencies. Despite being more data-dependent and achieving 94.8% standalone accuracy, it captures context not explicitly represented by the CNN. Its classification token forms f_V.

Radiomics-inspired branch. GLCM descriptors [25] encode contrast, correlation, energy, homogeneity, and dissimilarity, while wavelet features [27] capture multi-scale frequency variation. These deterministic features remain stable with limited data and explicitly represent texture and intensity changes. A dense layer projects them to f_X.

The branches provide partly non-redundant information. In Table II, ResNet50 with radiomics reaches 97.3%, compared with 96.5% for ResNet50 with ViT, while the complete model combines both texture stability and global context.

Feature-level fusion. The embeddings are concatenated as

$$f_c = [f_R \| f_V \| f_X], \tag{2}$$

and transformed through a cross-branch interaction layer:

$$h = \sigma_r(W_1 f_c + b_1), \tag{3}$$

where $\sigma_r(\cdot)$ denotes ReLU. A gated adaptive module estimates a sample-dependent importance vector,

$$g = \sigma(W_2 h + b_2), \qquad g \in (0,1)^d, \tag{4}$$

where $\sigma(\cdot)$ is the logistic sigmoid, and reweights the shared representation as

$$z = g \odot h. \tag{5}$$

Because g depends on the input, less informative dimensions can be suppressed for each scan rather than contributing a fixed share, as in decision-level fusion. The full model achieves 97.9% accuracy and outperforms all two-branch variants in Table II, supporting the value of complementary representations rather than additional capacity alone. Finally, z passes through fully connected layers with ReLU, batch normalization, and dropout, followed by a softmax classifier for FMD, HSD, MD, PCH, and WWS.

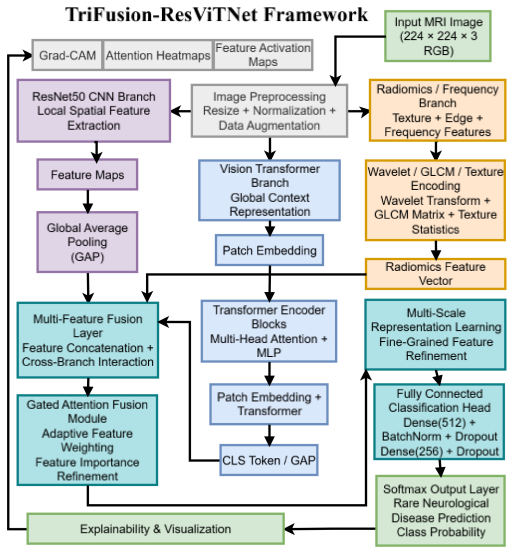


Fig. 1. Proposed TriFusion-ResViTNet framework for MRI-based rare neurological disease classification.

D. Training, Evaluation, and Validation

All models were implemented in TensorFlow 2.19 and Python 3.10 with fixed seeds and trained on an NVIDIA RTX 3060 GPU. TriFusion-ResViTNet used Adam and categorical cross-entropy:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{6}$$

where $C = 5$. Training used 20 epochs, batch size 32, dropout, and training-only augmentation. Because the data were class-balanced, no weighting or resampling was applied. The best validation checkpoint was tested once, while all baseline and ablation models used identical splits, preprocessing, and training settings.

Performance was measured using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices, with 60 external images used for cross-domain evaluation. Grad-CAM [26] and related visualizations provided qualitative interpretation, while Cohen's Kappa, McNemar's test, and confidence intervals assessed reliability.

III. Results and Discussion

A. Classification Performance and Confusion Matrix Analysis

TriFusion-ResViTNet achieved balanced performance across all five classes. The precision, recall, and F1-score were 0.972, 0.996, and 0.984 for FMD; 0.991, 0.957, and 0.974 for HSD; and 0.963, 0.948, and 0.955 for MD. PCH achieved the highest recall of 1.000 with an F1-score of 0.989, while WWS reached 0.994 precision and 0.978 F1-score. Overall, the model obtained 97.9% accuracy, with weighted precision, recall, and F1-score of 0.980, 0.976, and 0.977, respectively. On the other hand, The confusion matrix in Fig. 2 shows a strong diagonal pattern, indicating that most test images were correctly classified. Only limited off-diagonal errors occurred, mainly between diseases with similar structural MRI characteristics.

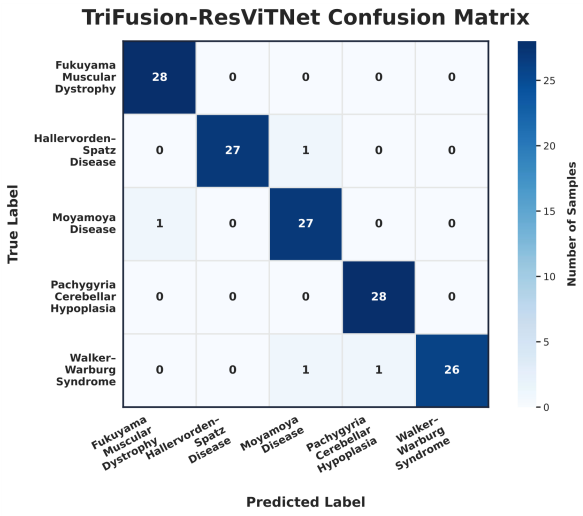


Fig. 2. Confusion matrix of TriFusion-ResViTNet on the held-out test set.

B. Misclassification and ROC Analysis

In the primary test set, four errors occurred: FMD→HSD, HSD→PCH, MD→FMD, and WWS→PCH. The external set

contained nine errors, mainly involving MD, PCH, and WWS. MD was the most difficult class, whereas FMD and PCH were easiest on the primary set. Confusion may reflect overlapping structural patterns among FMD, PCH, and WWS, while MD may require vascular or multi-sequence information unavailable in isolated 2D slices. External acquisition differences may also have contributed, but these explanations remain tentative because the errors were not radiologically reassessed. Image quality was not quantitatively linked to prediction error; therefore, the higher external error rate may reflect image quality, domain shift, or both.

The one-vs-rest ROC curves in Fig. 3 show strong separation across all five classes, with an overall ROC-AUC of 0.992 and high true-positive rates at low false-positive rates.

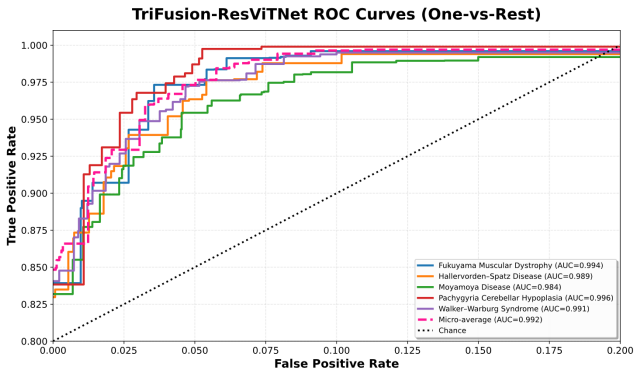


Fig. 3. One-vs-rest ROC curves of TriFusion-ResViTNet.

C. Training Performance Analysis

The curves in Fig. 4 show stable convergence. Training accuracy increased steadily and validation accuracy followed a similar trend, while both loss curves decreased gradually, indicating a reliable optimization process. The limited gap between training and validation performance suggests that dropout regularization and training-only augmentation improved generalization under limited data. These curves are reported for model selection only and are not used for performance claims.

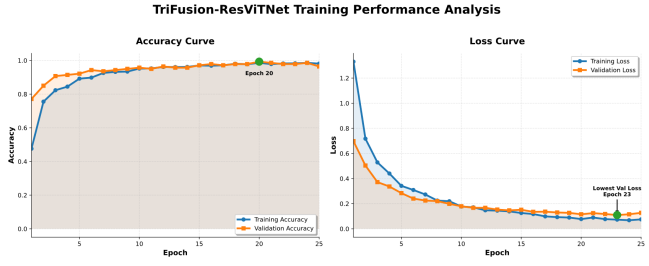


Fig. 4. Training and validation performance curves of TriFusion-ResViTNet.

D. Ablation Study

Table II reports five reduced variants trained under the identical protocol. All configurations were implemented in this work using the same split, preprocessing, optimizer, and evaluation script, so the differences are attributable to architecture. The complete framework achieved the best result on every metric. The ordering of the variants is discussed in Section II-D: the transformer is the weakest single branch at this data scale, radiomics complements the CNN more than the transformer does, and the three-branch model exceeds every two-branch variant.

TABLE II
ABLATION STUDY OF DIFFERENT MODEL CONFIGURATIONS

Configuration	Acc. (%)	Prec.	Rec.	F1	AUC
ResNet50 only [17]	97.1	0.972	0.969	0.970	0.986
ViT only [18]	94.8	0.949	0.944	0.946	0.974
ResNet50 + ViT	96.5	0.966	0.962	0.964	0.982
ResNet50 + radiomics	97.3	0.974	0.971	0.972	0.988
ViT + radiomics	95.7	0.958	0.952	0.955	0.978
TriFusion-ResViTNet	**97.9**	**0.980**	**0.976**	**0.977**	**0.992**

E. Comparative Performance Analysis

Six ImageNet-pretrained baselines [19]–[24] were retrained under the same protocol for the matched comparison in Table III. TriFusion-ResViTNet performed best, achieving 97.9% accuracy and 0.992 ROC-AUC, although its gain over ResNet50 was modest.

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS

Model	Acc. (%)	Prec.	Rec.	F1	AUC
VGG16 [19]	81.6	0.822	0.807	0.814	0.881
MobileNetV2 [20]	83.4	0.839	0.826	0.832	0.894
InceptionV3 [21]	85.1	0.856	0.842	0.848	0.907
Xception [22]	86.3	0.868	0.854	0.860	0.918
DenseNet121 [23]	87.5	0.879	0.866	0.872	0.931
EfficientNetB0 [24]	89.2	0.895	0.883	0.889	0.947
TriFusion-ResViTNet	**97.9**	**0.980**	**0.976**	**0.977**	**0.992**

F. Cross-Dataset Comparison

The locked model was evaluated without fine-tuning on 60 unseen external MRI images. As shown in Fig. 5, performance decreased slightly, suggesting domain shift across imaging conditions. Given the small sample, this provides preliminary transportability evidence rather than confirmed generalization.

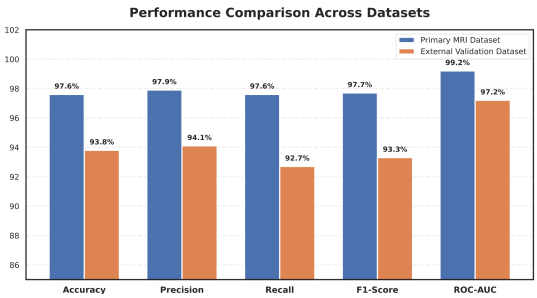


Fig. 5. Comparative performance analysis on the primary and secondary MRI datasets.

G. External Validation Analysis

The external confusion matrix in Fig. 6 shows generally stable performance, with most errors involving the visually similar MD, PCH, and WWS classes. Given the small 60-image external set, the results provide preliminary cross-domain evidence rather than confirmed clinical generalization.

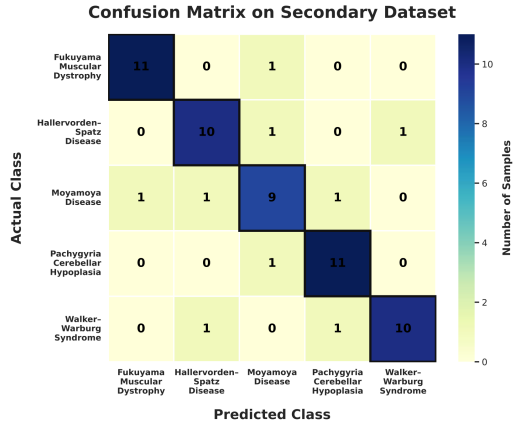


Fig. 6. External validation confusion matrix on the secondary MRI dataset.

H. Visual Interpretation and Feature Map Analysis

Fig. 7 presents Grad-CAM, edge, contour, and intensity analyses, showing activations over anatomical regions and highlighting structural and intensity variations. The visual evidence suggests that the model primarily responds to image content rather than background areas. However, consistency across samples, agreement with expert annotations, and differences between correct and incorrect predictions were not quantified. Thus, the findings provide qualitative interpretability only, not validated lesion localization or clinical evidence.

The feature maps in Fig. 8 illustrate hierarchical learning: early layers capture edges and local intensity changes, while deeper layers produce sparser, more organized activations over discriminative regions. This progression indicates increasing abstraction of MRI features across network depth. Nevertheless, these maps are descriptive rather than quantitative evidence of clinically meaningful representation learning.

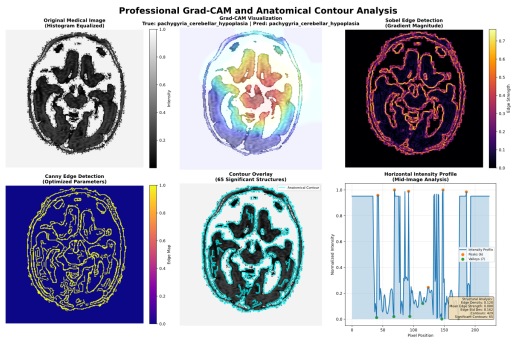


Fig. 7. Visual interpretation using Grad-CAM, edge detection, contour analysis, and intensity profiling.

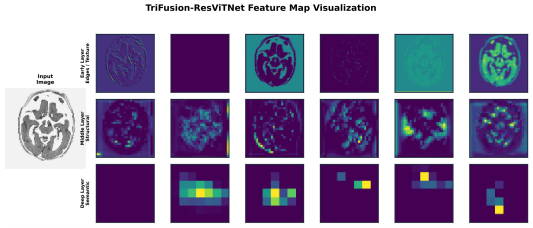


Fig. 8. Feature map visualization of TriFusion-ResViTNet.

I. Computational Complexity and Accuracy–Cost Trade-off

Complexity was measured under identical RTX 3060 conditions using parameters, FLOPs, and per-image inference time (Table IV); the ResNet50 and ViT rows follow [17], [18]. Compared with ResNet50, TriFusion-ResViTNet improves accuracy from 97.1% to 97.9% and ROC-AUC from 0.986 to 0.992, but roughly doubles parameters, FLOPs, and inference time. It is therefore better suited to accuracy-focused offline analysis than constrained or real-time deployment. ResNet50 + radiomics offers a more efficient alternative at 97.3% accuracy, 27.1,M parameters, and 22.3,ms. Since a 0.8-point gain on 140 test images is approximately one image, close rankings remain indicative and require repeated runs and larger test sets.

TABLE IV
COMPUTATIONAL COMPLEXITY ANALYSIS

Model	Params	FLOPs	Time (ms)
ResNet50 baseline [17]	25.6 M	4.1 G	18.4
ViT branch [18]	21.8 M	3.6 G	21.7
ResNet50 + ViT	48.2 M	7.9 G	35.6
ResNet50 + radiomics	27.1 M	4.4 G	22.3
TriFusion-ResViTNet	**50.4 M**	**8.3 G**	**39.8**

J. Statistical Validation

Table V shows strong prediction agreement, significant improvement over the baseline, and a 95% confidence interval of 95.8–98.4%. However, this interval indicates that small differences among the best models cannot be resolved reliably with only 140 test images.

TABLE V
STATISTICAL VALIDATION RESULTS

Analysis	Result	Interpretation
Cohen's Kappa	0.96	Strong agreement
McNemar's test	$p = 0.017$	Significant improvement
95% confidence interval	95.8–98.4%	Stable performance

IV. CONCLUSION

This paper presented TriFusion-ResViTNet, a feature-level fusion framework combining ResNet50 spatial features, Vision Transformer contextual features, and radiomics-inspired texture-frequency descriptors through cross-branch interaction and gated adaptive weighting. It achieved 97.9% accuracy, an F1-score of 0.977, and a ROC-AUC of 0.992, outperforming

six baselines and five ablation variants under a matched protocol. The contribution is a practical integration of established components rather than a new learning mechanism, and the modest gain over ResNet50 comes at approximately twice the computational cost, favouring accuracy-focused offline use.

The findings remain limited by unavailable patient identifiers, acquisition metadata, scanner information, and demographics; thus, patient-level separation could not be verified and residual multi-slice leakage cannot be excluded. The manually balanced, quality-filtered dataset does not reflect clinical prevalence or real-world image variability, and image quality was not quantitatively linked to error. The small test and external sets of 140 and 60 images limit resolution of close model differences. Moreover, only exported 2D slices were used without volumetric or multi-sequence context, particularly limiting vascular diagnoses such as MD, while interpretability remained qualitative and unvalidated by experts. Future work will use larger unfiltered multi-centre datasets with patient-level partitioning, volumetric and multi-sequence inputs, expert-annotated saliency analysis of correct and incorrect cases, single-channel lightweight models, and prospective clinical evaluation.

ACKNOWLEDGMENT

The authors thank the medical professional who assisted in MRI scan verification and acknowledge the use of AI-assisted tools for language and grammar refinement during manuscript preparation.

REFERENCES

[1] A. A. Lima, M. F. Mridha, S. C. Das, M. M. Kabir, Md. R. Islam, and Y. Watanobe, "A comprehensive survey on the detection, classification, and challenges of neurological disorders," *Biology*, vol. 11, no. 3, p. 469, Mar. 2022, doi: 10.3390/biology11030469.

[2] S. S. A. Ali, K. Memon, N. Yahya, and S. Khan, "Deep learning frameworks for MRI-based diagnosis of neurological disorders: a systematic review and meta-analysis," *Artificial Intelligence Review*, vol. 58, no. 6, Mar. 2025, doi: 10.1007/s10462-025-11146-5.

[3] R. Schiffmann and M. S. van der Knaap, "Invited article: An MRI-based approach to the diagnosis of white matter disorders," *Neurology*, vol. 72, no. 8, pp. 750–759, Feb. 2009, doi: 10.1212/01.wnl.0000343049.00540.c8.

[4] S. Duchesne, N. Bernasconi, A. Bernasconi, and D. L. Collins, "MR-based neurological disease classification methodology: Application to lateralization of seizure focus in temporal lobe epilepsy," *NeuroImage*, vol. 29, no. 2, pp. 557–566, Sep. 2005, doi: 10.1016/j.neuroimage.2005.07.052.

[5] S. A. Yazdan, R. Ahmad, N. Iqbal, A. Rizwan, A. N. Khan, and D.-H. Kim, "An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD," *Tomography*, vol. 8, no. 4, pp. 1905–1927, 2022, doi: 10.3390/tomography8040161.

[6] J. Jose and J. James, "An MRI based ischemic stroke classification – a mechanism oriented approach," *Annals of Indian Academy of Neurology*, vol. 25, no. 6, pp. 1019–1028, Nov. 2022, doi: 10.4103/aian.aian_365_22.

[7] M. Hashimoto, M. Ishikawa, E. Mori, and N. Kuwana, "Diagnosis of idiopathic normal pressure hydrocephalus is supported by MRI-based scheme: a prospective cohort study," *Cerebrospinal Fluid Research*, vol. 7, no. 1, p. 18, Oct. 2010, doi: 10.1186/1743-8454-7-18.

[8] D. Zhou *et al.*, "M-DDC: MRI based demyelinative diseases classification with U-Net segmentation and convolutional network," *Neural Networks*, vol. 169, pp. 108–119, Oct. 2023, doi: 10.1016/j.neunet.2023.10.010.

[9] H. Yousef, B. M. Tortei, and F. Castiglione, "Predicting multiple sclerosis disease progression and outcomes with machine learning and MRI-based biomarkers: a review," *Journal of Neurology*, vol. 271, no. 10, pp. 6543–6572, Sep. 2024, doi: 10.1007/s00415-024-12651-3.

[10] L. Saba *et al.*, "Brain MRI-based Wilson disease tissue classification: an optimised deep transfer learning approach," *Electronics Letters*, vol. 56, no. 25, pp. 1395–1398, Nov. 2020, doi: 10.1049/el.2020.2102.

[11] M. M. E. Yurtsever *et al.*, "Evaluating augmented deep networks for MRI-based brain anomaly classification," *Cluster Computing*, vol. 29, no. 3, Feb. 2026, doi: 10.1007/s10586-025-05893-x.

[12] M. B. Goeldlin *et al.*, "CADMUS," *Neurology*, vol. 102, no. 1, p. e207977, Dec. 2023, doi: 10.1212/wnl.0000000000207977.

[13] X. Chen *et al.*, "Noninvasive molecular diagnosis of craniopharyngioma with MRI-based radiomics approach," *BMC Neurology*, vol. 19, no. 1, p. 6, Jan. 2019, doi: 10.1186/s12883-018-1216-z.

[14] T. R. Sakib *et al.*, "Predicting super-class drug mechanisms from large-scale IC50 cell-line sensitivity profiles," in *Proc. IEEE TCSET*, 2026, pp. 1–6, doi: 10.1109/TCSET65181.2026.11461133.

[15] T. R. Sakib *et al.*, "CDFNet based fusion network for interpretable Alzheimer's disease prediction using hybrid imaging and clinical meta-ensemble learning," in *Proc. IEEE TCSET*, 2026, pp. 1–6, doi: 10.1109/TCSET65181.2026.11461211.

[16] AhsanNeural, "Rare Neurological Diseases MRI – Curated Edition," *Kaggle*, 2026. [Online]. Available: https://www.kaggle.com/datasets/ahsanneural/rare-neurological-diseases-mri-curated-edition. Accessed: May 5, 2026.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[18] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[23] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[24] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[25] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[27] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, Jul. 1989, doi: 10.1109/34.192463.