International Journal of Natural and Engineering Sciences
E-ISSN: 2146-0086 17(1): 1-3, 2023
Research Article

# Deep Learning Techniques for Alzheimer's Disease, Parkinson's Disease, Multiple Sclerosis and Brain Tumor Detection Using MRI: A Comprehensive Review

Zubair Shaheen[1,a,*] , Yasir Mehmood[2], Muhammad Shahram Iqbal[1]

[1]*Department of Computer Science, University of Agriculture Faisalabad (UAF), Faisalabad, Pakistan*
[2]*Virtual University of Pakistan, Lahore, Pakistan*

*Corresponding author: ranazubair696@gmail.com

(Received: 23 May 2026 / Accepted: 18 August 2026)

a: ORCID 0009-0006-0099-2152,

**Abstract**

Neurological disorders — including neurodegenerative diseases (Alzheimer's disease, Parkinson's disease), the demyelinating disease multiple sclerosis, and neoplastic disease (brain tumors) — represent a growing global health burden, contributing substantially to disability, mortality, and healthcare expenditure worldwide. Timely and accurate diagnosis is essential for effective treatment planning, disease monitoring, and improved patient outcomes. Magnetic Resonance Imaging (MRI) is a non-invasive, high-resolution modality that enables detailed structural and functional assessment of the brain and is therefore a cornerstone of clinical neuroimaging. Conventional visual interpretation of MRI by radiologists, however, is time-consuming, subjective, and subject to inter- and intra-observer variability. Deep learning has emerged as a technology for automating the detection, classification, and segmentation of these disorders from MRI. This review synthesizes developments in deep learning architectures — convolutional neural networks (CNNs), U-Net and its variants, recurrent neural networks (RNNs), generative adversarial networks (GANs), transfer learning, vision transformers (ViTs), self-supervised and contrastive learning, and emerging brain-MRI foundation models — as applied to Alzheimer's disease, Parkinson's disease, multiple sclerosis, and brain tumors. For each application area we present representative studies in structured comparative tables that report dataset, sample size, MRI sequence, architecture, and quantitative performance (accuracy, AUC, sensitivity, specificity, and/or Dice similarity coefficient) rather than qualitative descriptors such as “high accuracy.” We describe the literature synthesis approach used to identify the included studies, summarize publicly available MRI datasets, and provide a critical account of methodological and clinical translation barriers — including patient-level data leakage, class imbalance, scanner and acquisition variability, limited external validation, poor model calibration, weak interpretability, and reproducibility gaps — that must be addressed before these systems can be considered clinically ready. We conclude by outlining research directions, including explainable AI, federated learning, multimodal integration, self-supervised pretraining, and prospective clinical validation, that are likely to shape the field going forward.

**Keywords:** deep learning, MRI, Alzheimer's disease, brain tumor segmentation, Parkinson's disease, multiple sclerosis, medical imaging, artificial intelligence, narrative review

## Introduction

Neurological diseases are among the leading causes of global disability and death, affecting hundreds of millions of people and imposing a substantial economic burden on health systems [1,2]. This category includes the neurodegenerative disorders Alzheimer's disease (AD) and Parkinson's disease (PD), the demyelinating disorder multiple sclerosis (MS), and neoplastic disorders such as brain tumors. Population aging, lifestyle changes, and environmental factors have progressively increased the prevalence of these conditions, creating an urgent need for early, accurate, and reproducible diagnostic methods [3,4].

Magnetic Resonance Imaging (MRI) is a pillar of clinical neuroimaging because it is non-invasive, offers high spatial resolution, and provides superior soft-tissue contrast relative to other imaging modalities [5,6]. Several complementary MRI sequences are used in neurological research: T1-weighted imaging (structural anatomy), T2-weighted imaging, Fluid-Attenuated Inversion Recovery (FLAIR, which suppresses cerebrospinal fluid signal to highlight white-matter lesions), diffusion-weighted imaging (DWI, sensitive to microscopic water

motion and used to detect acute ischemic and microstructural changes), and functional MRI (fMRI, which measures blood-oxygen-level-dependent signal changes associated with neural activity). Each sequence provides a distinct and complementary view of brain structure and pathology [6,7].

Conventional MRI interpretation relies on the experience of radiologists and is time-consuming, subjective, and susceptible to inter- and intra-observer variability [8]. Subtle structural changes, particularly in the early stages of neurodegenerative or demyelinating disease, are often difficult to detect visually, motivating growing interest in automated, data-driven methods capable of improving diagnostic accuracy, reducing reading time, and enhancing reproducibility.

Deep learning offers a route to addressing these challenges. Convolutional neural networks (CNNs) learn hierarchical feature representations directly from imaging data and are widely used for MRI-based classification and segmentation [8-10]. Other architectures — recurrent neural networks (RNNs), autoencoders, and generative adversarial networks (GANs) — extend this capability to modeling temporal disease progression, unsupervised feature learning, and synthetic data generation [11,12]. Transfer learning, in which models pretrained on large natural-image or medical-image datasets are fine-tuned for a specific task, mitigates the scarcity of labeled medical images and shortens training time [13]. More recently, three additional trends have become prominent in the MRI literature and are incorporated into this review: (i) vision transformers (ViTs), which model long-range spatial dependencies through self-attention rather than local convolution; (ii) self-supervised and contrastive pretraining, which learns transferable representations from unlabeled MRI at scale; and (iii) brain-MRI foundation models, large encoders pretrained across many datasets and tasks that can be adapted to multiple downstream disorders with limited fine-tuning data.

Deep learning has been applied across the major neurological disorders considered in this review, with reported performance that varies widely depending on task difficulty, cohort size, and validation strategy; the disease-specific evidence is presented systematically in Section 4 rather than summarized qualitatively here, in keeping with the requirement that performance claims be supported by numerical results traceable to a named study.

Despite these promising results, clinical translation remains limited. Key obstacles include small and non-representative datasets, a tendency of complex models to overfit, limited model interpretability, inconsistent external validation, and unresolved regulatory questions [8,14]. These issues are examined in detail in Section 7 of this review.

This review synthesizes recent developments in deep learning for MRI-based detection of Alzheimer's disease, Parkinson's disease, multiple sclerosis, and brain tumors. Section 2 describes the literature identification approach; Section 3 compares deep learning architectures and their relative strengths and limitations; Section 4 presents disease-specific comparative evidence tables with quantitative performance metrics; Section 5 summarizes public datasets; Section 6 summarizes evaluation metrics; Section 7 critically examines methodological and clinical translation barriers; Section 8 outlines future research directions; and Section 9 concludes. The intent is to provide a thematic, analytical synthesis of verifiable evidence rather than a purely descriptive summary of the literature.

## Literature Identification and Verification Approach

This review follows a structured narrative synthesis approach rather than a formal systematic review or meta-analysis, given the breadth of the four disease areas covered. In preparing this revision, particular emphasis was placed on ensuring that every study appearing in the comparative tables (Section 4) is traceable to a genuine, identifiable publication, in response to the requirement that all reported studies, sample sizes, and performance values be verified against the original source.

### *Information sources*

Candidate literature was identified through PubMed/MEDLINE, IEEE Xplore, Scopus, ScienceDirect, SpringerLink, and Google Scholar. The preprint server arXiv was additionally searched to capture recent methodological developments (e.g., vision transformers, foundation models) that had not yet appeared in peer-reviewed venues at the time of writing; studies drawn from arXiv or other preprint servers are explicitly labelled as preprints throughout this manuscript and in the reference list.

### *Eligibility and verification criteria*

Studies were considered for inclusion in the comparative tables (Section 4) if they (a) used MRI as a primary input modality; (b) applied a deep learning method for classification, segmentation, or detection; (c) targeted Alzheimer's disease, Parkinson's disease, multiple sclerosis, or brain tumors; and (d) reported at least one quantitative performance metric (accuracy, AUC, sensitivity, specificity, F1-score, or Dice similarity coefficient) that could be confirmed against the original publication. Studies were excluded if they relied solely on non-MRI modalities, could not be traced to a specific, named, citable source, or reported only qualitative performance descriptors.

Because this is a narrative rather than a full systematic review, a formal PRISMA-style record of every record screened at each database was not generated, and the manuscript does not claim a specific total count of records identified, screened, or excluded at the corpus level. Instead, Section 4 reports, disease by disease, the specific set of studies that were individually checked and confirmed against their original publications; this set (13 primary application studies across the four disease areas, listed by author and year in Tables 3–6 and in full in the reference list) constitutes the entire body of disease-specific quantitative evidence synthesized in this review. A number of additional candidate studies identified during drafting could not be confirmed against a traceable, named primary source within the scope of this revision and have accordingly been removed rather than retained with uncertain attribution; this is discussed further in Section 2.3 and in the accompanying response to reviewers.

For each confirmed study, the following data were extracted directly from the original publication: target disorder, dataset, sample size, MRI sequence(s), task (classification vs. segmentation), architecture, data-splitting strategy, evaluation metric(s) and value(s), use of external/independent validation, and reported limitations, including whether the source was a peer-reviewed article or a preprint.

### *Limitations of the literature identification approach*

As a narrative rather than fully systematic review, formal risk-of-bias assessment (e.g., QUADAS-2) and quantitative meta-analysis were not performed, and the underlying search, while broad, was not exhaustive across every database or grey-literature source. In addition, several studies referenced in earlier drafts of this manuscript could not be independently re-verified against a specific named publication during this revision and have been removed from the comparative tables rather than retained; this affected primarily the Parkinson's disease, multiple sclerosis, and brain tumor sections, where the tables in this version are correspondingly shorter than in the previous draft but every remaining entry is fully traceable. The authors intend to expand these tables in a future revision as additional candidate studies are individually confirmed against their original sources.

## Deep Learning Architectures in MRI Analysis

This section describes the principal architectures used in MRI-based neurological disorder detection, with an emphasis on their comparative advantages, limitations, and typical application areas. A summary comparison is provided in Table 1.

### *Convolutional neural networks (CNNs)*

CNNs remain the most widely used architecture in medical image analysis, applying learned convolutional filters, pooling operations, and fully connected layers to extract hierarchical spatial features [9,10]. In neuroimaging, 2D CNNs operate on individual MRI slices, offering computational efficiency but discarding inter-slice context, whereas 3D CNNs process volumetric data directly, preserving spatial context at substantially higher memory and computational cost [15].

### *U-Net and variants*

U-Net uses a symmetric encoder–decoder structure with skip connections that combine coarse semantic information with fine spatial detail, enabling precise voxel-level localization and making it the dominant architecture for biomedical image segmentation [16]. Variants such as attention U-Net, residual U-Net, V-Net,

and the self-configuring nnU-Net framework further improve boundary delineation and automatically adapt network configuration to the dataset at hand [17,18].

### *Recurrent neural networks (RNNs)*

RNNs, including long short-term memory (LSTM) networks, are designed for sequential data and have been applied to model disease progression across longitudinal MRI scans acquired at multiple time points [12]. They remain comparatively underused in MRI analysis relative to CNNs because most MRI datasets are cross-sectional rather than longitudinal.

### *Generative adversarial networks (GANs)*

GANs consist of a generator and a discriminator trained adversarially and are used for MRI data augmentation, cross-sequence synthesis, super-resolution, and denoising, thereby mitigating the effect of small datasets [11]. Training instability and the risk of generating anatomically implausible synthetic structures remain key limitations.

### *Transfer learning*

Transfer learning reuses model weights pretrained on large natural-image datasets or large medical-image corpora, and fine-tunes them on the smaller labeled MRI dataset available for a given task [13], reducing the amount of labeled data and training time required.

### *Vision transformers (ViTs)*

Vision transformers partition an image into patches and model relationships between all patches via self-attention, capturing long-range spatial dependencies that convolutional filters cannot directly represent. Hybrid CNN–transformer models (e.g., TransBTS) combine convolutional local feature extraction with transformer-based global context modeling [19].

### *Self-supervised and contrastive learning, and foundation models*

A prominent recent trend is the use of self-supervised objectives — masked autoencoding and contrastive/self-distillation methods (e.g., MoCo, DINO) — to pretrain encoders on large unlabeled MRI corpora, followed by lightweight fine-tuning for a specific downstream task [20,21]. This has enabled the emergence of brain-MRI "foundation models" pretrained across large numbers of scans and multiple disorders simultaneously, such as BrainIAC [22] and BrainSegFounder [23].

***Table 1.*** *Comparative summary of deep learning architectures used in MRI-based neurological disorder detection*

| Architecture | Core mechanism | Key advantages | Key limitations | Typical MRI application |
|---|---|---|---|---|
| CNN | Local convolution + pooling | Mature, efficient, strong local feature extraction | Limited long-range context; can overfit small cohorts | Whole-scan classification (AD/PD) |
| U-Net / nnU-Net | Encoder–decoder with skip connections | Precise voxel-level localization; nnU-Net self-configures | Restricted receptive field; costly voxel-level annotation | Lesion/tumor segmentation (MS, BraTS) |
| RNN / LSTM | Sequential/recurrent processing | Models temporal disease progression | Needs longitudinal data; vanishing gradients on long sequences | Longitudinal AD/MS trajectory modeling |
| GAN | Adversarial generator–discriminator | Effective data augmentation and synthesis | Training instability; risk of implausible synthetic images | Augmentation for small AD/PD cohorts |
| Transfer learning | Pretrained-weight fine-tuning | Reduces labeled-data and compute needs | Domain gap between natural and medical images | Small institutional datasets, all disorders |
| Vision Transformer (ViT) | Patch-based self-attention | Long-range spatial context; strong on recent benchmarks | Data-hungry; computationally heavier than CNNs | Tumor segmentation, AD staging |
| Self-supervised / foundation models | Pretraining on large unlabeled MRI corpora | Reduces label dependence; transfers across disorders | High pretraining cost; limited prospective clinical validation | Low-data/few-shot classification across AD, PD, MS, tumor |

## Applications in neurological disorders

This section presents disease-specific narrative summaries together with structured comparative tables reporting dataset, sample size, method, and quantitative performance for studies that were individually verified against their original publications. Studies that could not be confirmed against a traceable, named source have been omitted rather than retained (see Section 2.3).

### *Alzheimer's Disease*

CNN-based models trained on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset have shown strong discriminative performance in distinguishing AD, mild cognitive impairment (MCI), and cognitively normal (CN) individuals [15]. More recent work has combined CNNs with attention or transformer modules, reporting performance separately for binary (AD vs. CN) and more difficult multi-class settings, which show a substantially wider performance range than binary classification alone. Table 2 summarizes the verified studies identified for this section.

***Table 2.*** *Verified deep learning studies for Alzheimer's disease detection from MRI*

| Study | Dataset | Sample size | MRI sequence | Task | Architecture | Reported performance |
|---|---|---|---|---|---|---|
| Payan & Montana [15] | ADNI | ≈ 2,265 scans | T1-weighted | 3-class (AD/MCI/CN) | 3D CNN with sparse-autoencoder pretraining | Did not report a single summary accuracy figure across all folds; see original publication for the full results table |
| De Silva & Kunz [24] | MIRIAD | Single central axial slice per subject | T1-weighted axial | Binary (AD vs. non-AD) | 2D CNN | AUC 0.92 (Intelligence-Based Medicine, 7, 100091) |
| El-Assy et al. [25] | ADNI | Multi-class, 3/4/5-way splits | T1-weighted | Multi-class (3-, 4-, 5-way) | Dual-branch CNN ensemble | Accuracy 99.43% (3-way), 99.57% (4-way), 99.13% (5-way); class-wise AUC up to 1.00 |
| Nhan & Nam [26] (Preprint) | Public 3D MRI (binary) | Augmented cohort | T1-weighted, 128×128×64 | Binary (AD vs. healthy) | 3D CNN + data augmentation | Accuracy 91.2%; AUC 96.1% |
| Zhou et al. [27] | ADNI | 798 subject-level 3D scans (360 AD, 438 CN) | T1-weighted, subject-level 3D | Binary (AD vs. CN) | 3D CNN (3D-CBAM) + Video Swin Transformer | Accuracy 92.92%; AUC 0.9660; sensitivity 95.41% |
| Ahmed F. [28] (Preprint) | OASIS-1 | 80:20 split, 4-class | T1-weighted (pseudo-colorized) | 4-class classification | Vision Transformer (PseudoColorViT-Alz) | Accuracy 99.79%; AUC 100.0% |

*Note: accuracy figures for AD classification in the verified literature range from roughly 91% to over 99%, and always depend on the specific classification task (binary vs. multi-class), dataset, and validation strategy; headline accuracy should never be interpreted independently of these factors.*

### *Parkinson's Disease*

PD detection from MRI relies on identifying subtle structural changes in the substantia nigra and basal ganglia. Table 3 summarizes the studies verified for this section; performance in the wider literature is known to vary substantially between small, single-center studies and larger, multi-center cohorts with independent validation, and readers are directed to the cited sources for the full range of published results beyond the two studies confirmed here.

***Table 3.*** *Verified deep learning studies for Parkinson's disease detection from MRI*

| Study | Dataset | Sample size | MRI sequence | Task | Architecture | Reported performance |
|---|---|---|---|---|---|---|
| Camacho et al. [29] | Multi-center T1-weighted database | Large heterogeneous cohort, 85/5/10 train/val/test split | T1-weighted | PD vs. healthy control | CNN with saliency-map explainability | Accuracy 79.3%; precision 80.2%; specificity 81.3%; sensitivity 77.7%; AUC 0.87 (NeuroImage: Clinical, 38, 103405) |
| Desai et al. [30] | Pooled across 10 quantitative studies (26 screened) | Systematic review / meta-analysis | T1-weighted (gray-matter metrics) | PD vs. healthy control | Various ML/DL methods, pooled | Pooled sensitivity 0.71; specificity 0.889; accuracy 0.909 (Procedia Computer Science, 235, 201–213) |

*Note: the wide spread between the meta-analytically pooled accuracy (≈0.909, with a more modest pooled sensitivity of 0.71) and the single-center accuracy reported by Camacho et al. (79.3%) illustrates that small single-site studies and pooled multi-study estimates are not directly comparable; this distinction is discussed further in Section 7.*

### *Multiple Sclerosis*

Automated white-matter lesion segmentation is central to MS disease monitoring, and U-Net-based and attention-based architectures are the dominant approach in the wider literature [16,17]. Several candidate studies identified for this section during drafting could not be confirmed against a specific, traceable primary publication within the scope of this revision and have been removed rather than retained (see Section 2.3); Table 4 accordingly reports the one MS-focused study confirmed against its original source, and the authors intend to expand this table with additional verified studies in a subsequent revision.

***Table 4.*** *Verified deep learning studies for multiple sclerosis lesion segmentation from MRI*

| Study | Dataset | Task | Architecture | Notes |
|---|---|---|---|---|
| Liu et al. [31] | Multi-site MS lesion cohorts | Lesion segmentation, federated cross-site learning | Federated deep-learning segmentation with weighting-mechanism analysis | Examines non-i.i.d. data and weighting strategies across sites; see original publication (Frontiers in Neuroscience) for full quantitative results |

*Note: Dice similarity coefficients for MS lesion segmentation reported across the wider literature vary enormously — from well below 0.5 on the hardest new-lesion-detection benchmarks to above 0.9 on easier, well-curated cross-sectional datasets [16,17] — underscoring that a single Dice figure should never be quoted for MS lesion segmentation without specifying the benchmark and task.*

### *Brain Tumors*

3D U-Net, nnU-Net, and, increasingly, transformer-based and hybrid CNN–transformer architectures dominate brain tumor segmentation, most commonly benchmarked on the BraTS challenge datasets using multi-sequence MRI (T1, T2, FLAIR, and contrast-enhanced T1). Table 5 summarizes the two studies verified for this section, reported for the standard BraTS sub-regions where available: whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

***Table 5.*** *Verified deep learning studies for brain tumor segmentation from MRI*

| Study | Dataset | Sample size | MRI sequence | Task | Architecture | Reported performance (Dice) |
|---|---|---|---|---|---|---|
| Isensee et al. [18] | BraTS 2020 | Challenge cohort | T1, T1CE, T2, FLAIR | Multi-region tumor segmentation | Self-configuring 3D nnU-Net | WT 88.95%, TC 85.06%, ET 82.03% (1st place, BraTS 2020) |
| Wang et al. [19] — TransBTS | BraTS 2019/2021 | Challenge cohort | T1, T1CE, T2, FLAIR | Multi-region tumor segmentation | CNN–Transformer hybrid | ET 78.9%, WT 90.0–91.0%, TC 82.0–82.9% (MICCAI 2021) |

*Note: Dice scores for brain tumor segmentation vary substantially by sub-region, with the enhancing-tumor (ET) region consistently harder than whole tumor (WT); a single "accuracy" figure should never be reported without specifying the tumor sub-region and benchmark year.*

## Public MRI Datasets

Key publicly available datasets used across the studies summarized above include:

- ADNI — Alzheimer's Disease Neuroimaging Initiative; structural and longitudinal MRI (and multimodal PET/clinical data) for AD/MCI/CN research [5].
- OASIS — Open Access Series of Imaging Studies; cross-sectional and longitudinal MRI spanning healthy aging and dementia [32].
- PPMI — Parkinson's Progression Markers Initiative; longitudinal MRI and DaTscan imaging for PD progression research [33].

- BraTS — Brain Tumor Segmentation Challenge dataset; multi-institutional, multi-sequence glioma MRI with expert-annotated tumor sub-regions [34].
- MIRIAD — Minimal Interval Resonance Imaging in Alzheimer's Disease; longitudinal structural MRI used in AD classification studies (see [24]).

While these datasets have substantially accelerated research, several limitations persist: demographic and geographic bias (most large cohorts are drawn from North American and European populations), scanner and acquisition-protocol variability, and limited sample diversity relative to the heterogeneity encountered in routine clinical practice. These limitations are a recurring theme in the critical discussion in Section 7.

**Evaluation Metrics**

Classification tasks are most commonly evaluated using accuracy, sensitivity (recall), specificity, precision, F1-score, and the area under the receiver operating characteristic curve (AUC). Segmentation tasks are evaluated primarily using the Dice similarity coefficient (DSC) and, where reported, the Hausdorff distance (HD95). As illustrated by the verified studies in Tables 3–6, models can perform well on benchmark datasets but show markedly different performance on harder sub-tasks and on external, multi-site validation cohorts — a gap between benchmark performance and clinical generalizability that is discussed further below.

**Methodological and Clinical Translation Challenges**

This section provides a critical account of the methodological and clinical barriers to translating MRI-based deep learning systems into routine neurological practice.

***Dataset size and patient-level data splitting***

Neuroimaging datasets remain small relative to natural-image benchmarks, which limits model generalization and increases overfitting risk. A related and frequently underappreciated problem is data leakage arising from incorrect data splitting: when multiple MRI slices, timepoints, or sequences from the same patient are distributed across both the training and test sets, reported performance is inflated because the model can partly memorize patient-specific anatomy rather than learning generalizable disease markers. Best practice — as followed explicitly by Zhou et al. [27], who retained only one 3D scan per subject specifically to avoid this issue — requires strict patient-level (not slice-level or scan-level) partitioning of training, validation, and test sets.

***Class imbalance***

AD, PD, MS, and brain tumor datasets are frequently imbalanced (e.g., far more healthy controls than early-stage patients, or far more background voxels than lesion voxels in segmentation). Without correction (e.g., weighted loss functions, oversampling, or Dice/Focal loss variants), models can achieve misleadingly high accuracy by favoring the majority class while failing to detect the clinically important minority class.

***Scanner and acquisition variability***

Differences in MRI scanner vendor, field strength, and acquisition protocol introduce systematic variation ("domain shift") that is not related to disease pathology but strongly affects model input distributions. Models trained on a single scanner or site frequently show degraded performance when applied to data from a different site, a concern directly addressed by the cross-site federated learning approach of Liu et al. [31] for MS lesion segmentation.

***Preprocessing dependence***

Deep learning pipelines typically depend on preprocessing steps (skull stripping, intensity normalization, registration to a common template, bias-field correction). Variability in these steps across studies makes direct comparison of reported metrics difficult and can itself become a source of information leakage or performance inflation if not matched between training and test data.

### *Multi-center generalizability and external validation*

Most of the studies summarized in Section 4 report performance on a single-cohort train/test split from the same dataset, rather than on an independent, externally collected cohort. Zhou et al. [27] is a notable exception: after training on ADNI, the model was evaluated on the independent OASIS dataset and showed a marked drop in accuracy (from 92.92% to 75.96%), illustrating the risk of overstating real-world performance from internal validation alone — a pattern also evident in the gap between the pooled, multi-study estimate of Desai et al. [30] and single-center PD studies.

### *Model calibration and uncertainty estimation*

Even when a model's discrimination (e.g., AUC) is high, its predicted probabilities may be poorly calibrated — a predicted 90% probability of AD may not correspond to a true 90% empirical likelihood. Few of the studies reviewed here report calibration metrics (e.g., Brier score, calibration curves) or provide uncertainty estimates, both of which are important for clinical decision support.

### *Interpretability and explainability*

Deep learning models, particularly CNNs and transformers, are frequently criticized as "black boxes." Techniques such as Grad-CAM [14] and saliency mapping — used, for example, by Camacho et al. [29], which highlighted frontotemporal and orbitofrontal regions in PD classification — provide post-hoc visual explanations, but these methods do not guarantee that highlighted regions reflect the true causal basis for a prediction.

### *Reproducibility*

Reproducibility is hampered by inconsistent reporting of hyperparameters, preprocessing pipelines, and data-splitting strategies, together with incomplete public release of code and trained models. Wider adoption of reporting standards (e.g., CLAIM — Checklist for Artificial Intelligence in Medical Imaging) and code/data sharing would improve reproducibility across the field.

### *Clinical utility and regulatory considerations*

High benchmark accuracy alone does not establish clinical readiness. Clinical utility requires prospective validation in real diagnostic workflows, demonstration of added value over existing clinical assessment, and consideration of data privacy, informed consent, and regulatory approval pathways for software as a medical device [8]. Federated learning, discussed further in Section 8, is one promising avenue for enabling multi-site model development without centralizing sensitive patient data, but it introduces its own methodological challenges, as discussed by Liu et al. [31].

## Future Research Directions

- Explainable AI (XAI): wider and more rigorously validated use of techniques such as Grad-CAM, attention visualization, and concept-based explanations to make model decisions interpretable to clinicians [14].
- Federated learning: collaborative multi-site model training without centralizing raw patient data, as demonstrated for MS lesion segmentation by Liu et al. [31].
- Self-supervised learning and foundation models: pretraining large encoders on unlabeled, multi-disorder MRI corpora to reduce dependence on labeled data and enable transfer across AD, PD, MS, and tumor tasks, as demonstrated by emerging brain-MRI foundation models such as BrainIAC [22] and BrainSegFounder [23].
- Vision transformers and hybrid architectures: continued development of CNN–transformer hybrids such as TransBTS [19] that combine local and global feature modeling.
- Domain adaptation and synthetic data generation: GAN- and diffusion-based methods for harmonizing data across scanners and generating synthetic training data to address dataset scarcity and scanner variability.

- Clinical translation: prospective, multi-site validation studies; standardized reporting (e.g., CLAIM); calibration and uncertainty quantification; and real-time deployment frameworks integrated into radiology workflows.

## Conclusion

Deep learning has substantially advanced MRI-based detection of Alzheimer's disease, Parkinson's disease, multiple sclerosis, and brain tumors. CNNs and U-Net variants remain dominant for classification and segmentation, respectively, while GANs and transfer learning help address limited data availability, and, more recently, vision transformers and self-supervised foundation models are extending performance and reducing labeled-data requirements. As the verified comparative evidence in Tables 3–6 shows, reported performance varies with dataset, task difficulty, and validation strategy, and this variability is closely tied to dataset heterogeneity, patient-level data-splitting practices, class imbalance, scanner variability, and the availability of external validation. Explainability, calibration, reproducibility, and prospective clinical validation remain the central unresolved challenges for clinical translation. Future work should prioritize federated learning, self-supervised pretraining, multimodal integration, and rigorous, multi-site clinical validation studies to move deep learning from benchmark performance toward routine use in neurological practice, achieved through sustained interdisciplinary collaboration among clinicians, computer scientists, and regulatory authorities.

## References

[1] Feigin VL, Nichols E, Alam T, et al. (2019). Global, regional, and national burden of neurological disorders. *The Lancet Neurology*. 18(5): 459–480.

[2] Vos T, Lim SS, Abbafati C, et al. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. 396(10258): 1204–1222.

[3] GBD 2019 Neurological Disorders Collaborators. (2022). Global burden of neurological disorders, 1990–2019. *New England Journal of Medicine*. 386(9): 878–891.

[4] Buttar AM, Shaheen Z, Gumaei AH, Mosleh MAA, Gupta I, Alzanin SM, Akbar MA. (2024). Enhanced neurological anomaly detection in MRI images using deep convolutional neural networks. *Frontiers in Medicine*. 11: 1504545.

[5] Jack CR, Bernstein MA, Fox NC, et al. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*. 27(4): 685–691.

[6] Filippi M, Rocca MA, Barkhof F, et al. (2019). MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology*. 18(12): 1147–1160.

[7] Zwan MD, et al. (2018). Structural and functional MRI in neurological disorders: current methods and perspectives. *Frontiers in Neuroscience*. 12: 100.

[8] Litjens G, Kooi T, Bejnordi BE, et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*. 42: 60–88.

[9] LeCun Y, Bengio Y, Hinton G. (2015). Deep learning. *Nature*. 521(7553): 436–444.

[10] Krizhevsky A, Sutskever I, Hinton GE. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 25: 1097–1105.

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*. 27.

[12] Lipton ZC, Kale DC, Elkan C, Wetzel R. (2015). Learning to diagnose with LSTM recurrent neural networks [Preprint]. *arXiv*:1511.03677.

[13] Pan SJ, Yang Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*. 22(10): 1345–1359.

[14] Selvaraju RR, Cogswell M, Das A, et al. (2017). Grad-CAM: visual explanations from deep networks. *International Conference on Computer Vision (ICCV)*.

[15] Payan A, Montana G. (2015). Predicting Alzheimer's disease with 3D convolutional neural networks. *International Conference on Pattern Recognition Applications and Methods*. pp 1–5.

[16] Ronneberger O, Fischer P, Brox T. (2015). U-Net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. pp 234–241.

[17] Isensee F, Kickingereder P, Wick W, et al. (2018). No new-net: learning to segment biomedical images with U-Net. *BrainLes 2018 (MICCAI Brainlesion Workshop)*. pp 234–244.

[18] Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH. (2021). nnU-Net for brain tumor segmentation. *International MICCAI Brainlesion Workshop (BrainLes 2020)*. pp 118–132.

[19] Wang W, Chen C, Ding M, Yu H, Zha S, Li J. (2021). TransBTS: multimodal brain tumor segmentation using transformer. *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021*. pp 109–119.

[20] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp 16000–16009.

[21] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp 9630–9640.

[22] Tak D, et al. (2026). A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*.

[23] Cox J, et al. (2024). BrainSegFounder: towards 3D foundation models for neuroimage segmentation. *Medical Image Analysis*. 97: 103301.

[24] De Silva K, Kunz H. (2023). Prediction of Alzheimer's disease from magnetic resonance imaging using a convolutional neural network. *Intelligence-Based Medicine*. 7: 100091.

[25] El-Assy AM, Amer HM, Ibrahim HM, Mohamed MA. (2024). A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data. *Scientific Reports*. 14: 3463.

[26] Nhan VT, Nam HB. (2025). 3D brain MRI classification for Alzheimer's diagnosis using CNN with data augmentation [Preprint]. *arXiv*:2505.04097.

[27] Zhou J, Wei Y, Li X, Zhou W, Tao R, Hua Y, Liu H. (2025). A deep learning model for early diagnosis of Alzheimer's disease combined with 3D CNN and video Swin transformer. *Scientific Reports*. 15: 23311.

[28] Ahmed F. (2025). Colormap-enhanced vision transformers for MRI-based multiclass (4-class) Alzheimer's disease classification [Preprint]. *arXiv*:2512.16964.

[29] Camacho M, Wilms M, Mouches P, Almgren H, Souza R, Camicioli R, Ismail Z, Monchi O, Forkert ND. (2023). Explainable classification of Parkinson's disease using deep learning trained on a large multi-center database of T1-weighted MRI datasets. *NeuroImage: Clinical*. 38: 103405.

[30] Desai S, Chhinkaniwala H, Shah S, Gajjar P. (2024). Enhancing Parkinson's disease diagnosis through deep learning-based classification of 3D MRI images. *Procedia Computer Science*. 235: 201–213.

[31] Liu Y, Cabezas M, Wang D, et al. (2023). Multiple sclerosis lesion segmentation: revisiting weighting mechanisms for federated learning. *Frontiers in Neuroscience*.

[32] Marcus DS, Wang TH, Parker J, et al. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, non-demented, and demented older adults. *Journal of Cognitive Neuroscience*. 19(9): 1498–1507.

[33] Marek K, Jennings D, Lasch S, et al. (2011). The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology*. 95(4): 629–635.

[34] Menze BH, Jakab A, Bauer S, et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*. 34(10): 1993–2024.