

Multi-Dataset Deep Learning for Automated Multiple Sclerosis Lesion Segmentation in Multimodal Brain MRI

Mahmoud Elsayed

MahmoudElhema1y@std.mans.edu.eg

Faculty of Engineering, Mansoura University

Mostafa Elhoshy

Faculty of Engineering, Mansoura University

Nahla Bishri

Faculty of Engineering, Mansoura University

Ahmed I. Saleh

Faculty of Engineering, Mansoura University

Article

Keywords: Multiple sclerosis, Medical image segmentation, Deep learning, Multimodal MRI, UNet++, Healthcare informatics

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11134819/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multi-Dataset Deep Learning for Automated Multiple Sclerosis Lesion Segmentation in Multimodal Brain MRI

Mahmoud Elsayed^{1*}, Mostafa Elhoshy¹, Nahla Bishri¹,
Ahmed I. Saleh¹

^{1*}Department of Computer and Control Systems Engineering, Faculty of Engineering, Mansoura University, Mansoura, Egypt.

*Corresponding author(s). E-mail(s):
MahmoudElhemaly@std.mans.edu.eg;

Abstract

Automated segmentation of multiple sclerosis (MS) lesions from magnetic resonance imaging (MRI) can support quantitative assessment, but performance remains sensitive to lesion heterogeneity, severe class imbalance, and variation across imaging sources. This study evaluates a multi-dataset deep learning framework for MS lesion segmentation using multimodal brain MRI. T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences are combined as a three-channel input to a UNet++ network with deep supervision. Training incorporates a hybrid Dice-Tversky objective, lesion-aware augmentation, and hyperparameter optimization. Comparative experiments assess architecture, encoder backbone, loss formulation, and modality fusion. On the held-out test split, the proposed model achieved a Dice similarity coefficient of 0.722 ± 0.082 , lesion-wise sensitivity of 0.783 ± 0.109 , and lesion-wise F1-score of 0.738 ± 0.089 . Small-lesion sensitivity reached 62%, representing a 29% relative improvement over the 2D U-Net baseline. A complementary resource-constrained ImageMask experiment compared U-Net, UNet++, and a SegNet-like architecture using a separate two-dimensional FLAIR pipeline. The results support multimodal representation and imbalance-aware optimization while highlighting the need for patient-level external validation before clinical deployment.

Keywords: Multiple sclerosis, Medical image segmentation, Deep learning, Multimodal MRI, UNet++, Healthcare informatics

1 Introduction

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system in which focal lesions in the brain and spinal cord provide important information for diagnosis, disease monitoring, and treatment assessment. Magnetic resonance imaging (MRI) is therefore central to contemporary MS evaluation, particularly through complementary T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences. Accurate lesion delineation remains challenging, however, because lesions vary substantially in size, location, morphology, and signal characteristics, and manual or semi-manual segmentation is time-consuming and subject to observer variability.

Automated lesion segmentation can improve the efficiency, consistency, and scalability of MRI-based MS assessment. Nevertheless, the task remains technically demanding. Lesion voxels occupy a small proportion of the image, creating severe class imbalance, while small or poorly contrasted lesions may be suppressed during feature extraction. Performance can also deteriorate when models are applied across scanners, acquisition protocols, institutions, or patient populations that differ from the data used for training. Recent studies continue to show that data variability, multimodal information, augmentation, and model design materially affect segmentation reliability and generalization [1–4].

Deep learning, particularly U-Net-derived encoder-decoder architectures, has substantially advanced biomedical image segmentation by learning hierarchical representations directly from imaging data. Nested architectures such as UNet++ were introduced to improve feature fusion through progressively refined skip pathways, while transformer-based and hybrid architectures have expanded the ability to model long-range dependencies. At the same time, recent MS-specific work has increasingly emphasized multi-sequence MRI integration, cross-dataset robustness, lesion-level evaluation, and reproducibility [4–7]. These developments motivate the need for evaluations that distinguish the effects of architecture, modality fusion, loss design, augmentation, and dataset heterogeneity rather than attributing performance improvements to a single component.

Despite this progress, several limitations remain in the existing literature. Many studies are trained on a single dataset or a restricted set of acquisition conditions, making it difficult to assess generalization across heterogeneous imaging sources. In addition, architectural modifications, multimodal fusion, imbalance-aware losses, and augmentation strategies are often introduced simultaneously, which makes their individual contributions difficult to isolate. Recent studies published in 2026 further reinforce the importance of evaluating robustness across datasets, imaging modalities, lesion types, and acquisition settings rather than relying solely on in-domain overlap metrics [1–3].

Motivated by these limitations, this study develops and evaluates a multi-dataset deep learning framework for automated MS lesion segmentation using three publicly available MRI datasets: MSLesSeg, a multi-rater consensus Brain MRI dataset, and Open MS_Data. The resulting cohort contains 165 patients and 295 scans and provides heterogeneous data for examining robustness across imaging sources. The main segmentation framework is based on UNet++ and uses fused T1-weighted,

T2-weighted, and FLAIR images as a three-channel input. An imbalance-aware training strategy combining Dice and Tversky objectives is used with deep supervision, while augmentation and Bayesian hyperparameter optimization are incorporated into the experimental design. The contribution of the principal architectural and training choices is examined through comparative and ablation experiments rather than assumed a priori.

The study also includes a complementary resource-constrained evaluation implemented in a Kaggle notebook environment. In this setting, raw NIfTI volumes and lesion masks are transformed into slice-wise image-mask pairs, and conventional U-Net, UNet++, and a SegNet-like encoder-decoder are trained under a standardized protocol. This secondary experimental setting is intended to examine reproducibility and practical behavior under constrained computational conditions and should be interpreted as complementary to, rather than equivalent to, the multi-dataset framework. An overview of the complete study workflow, including multi-dataset integration, preprocessing, patient-level data splitting, model development, training, and evaluation, is summarized in Fig. 1.

The main contributions of this study are as follows:

1. **A multi-dataset framework for robust MS lesion segmentation:** We integrate three publicly available MRI datasets, comprising 165 patients and 295 scans, into a unified experimental framework to investigate the robustness of automated MS lesion segmentation across heterogeneous imaging sources and acquisition conditions.
2. **A multimodal and imbalance-aware UNet++ framework:** We develop and evaluate a three-channel UNet++ model that jointly exploits T1-weighted, T2-weighted, and FLAIR MRI, combined with Dice-Tversky optimization, deep supervision, data augmentation, and Bayesian hyperparameter optimization to address class imbalance and heterogeneous lesion characteristics.
3. **Systematic comparative and reproducibility evaluation:** We isolate the contribution of the principal architectural, loss, modality, and training choices through comparative and ablation experiments, and complement the primary multi-dataset evaluation with a resource-constrained Kaggle/ImageMask implementation comparing U-Net, UNet++, and a SegNet-like architecture under a standardized protocol.

The remainder of this paper is organized as follows. Section 2 reviews the methodological development of automated MS lesion segmentation and positions the present study within the recent literature. Section 3 describes the datasets, preprocessing procedures, model architectures, training strategy, and evaluation protocol. Section 4 presents the experimental findings, while Section 5 interprets those findings in relation to prior work and clinical relevance. Section 6 summarizes the main conclusions and future directions. Detailed implementation procedures, algorithms, and reproducibility information are provided in the Appendix.

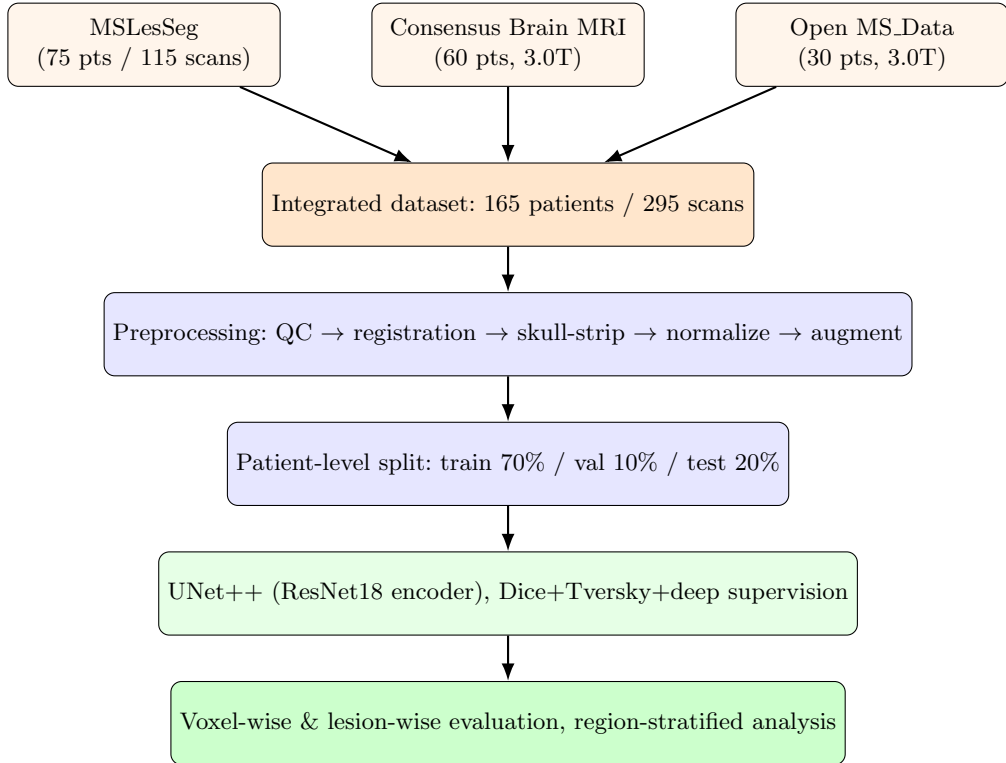


Fig. 1: Overview of the complete study methodology, from multi-dataset integration and preprocessing to patient-level data splitting, UNet++ training, and final evaluation.

2 Related Work

Automated MS lesion segmentation has progressed through several methodological stages, evolving from traditional image-processing and handcrafted-feature approaches to convolutional encoder-decoder networks, nested U-Net architectures, and more recent transformer-based models. Despite these advances, the fundamental challenges of MS lesion segmentation remain substantial. Lesions exhibit considerable variability in size, shape, anatomical location, and image intensity, while differences in scanners and acquisition protocols further complicate accurate and robust delineation.

The introduction of deep learning has significantly advanced biomedical image segmentation by enabling models to learn task-specific representations directly from annotated imaging data. U-Net [8] established the encoder-decoder architecture with skip connections as a widely adopted foundation for medical image segmentation. Building upon this concept, UNet++ [5] introduced nested and dense skip pathways to reduce the semantic gap between encoder and decoder feature maps and to support deep supervision. Self-configuring frameworks such as nnU-Net [9] further

demonstrated the importance of systematically adapting preprocessing, network configuration, training, and post-processing strategies to the characteristics of individual biomedical segmentation tasks.

In MS lesion segmentation specifically, modified U-Net architectures and multi-scale learning strategies have been investigated to address the heterogeneous appearance and distribution of lesions. Hashemi et al. [10] proposed a modified attention U-Net for MS lesion segmentation, demonstrating the potential of attention mechanisms for improving lesion-focused feature representation. Similarly, DeepLesion-Brain [11] investigated deep-learning strategies aimed at improving generalization of MS lesion segmentation across heterogeneous imaging data. These studies highlight the importance of architectural design, multi-scale representation, and generalization when developing segmentation models for clinically heterogeneous MRI data.

More recent research has increasingly explored transformer-based architectures, multimodal learning, and multi-dataset training to improve contextual representation and robustness across imaging environments. Nevertheless, important challenges remain, particularly regarding effective multimodal MRI fusion, severe lesion-to-background class imbalance, heterogeneity across datasets and acquisition protocols, and reproducible evaluation across different experimental settings. The present study is positioned within these remaining challenges by investigating a multi-dataset, three-channel UNet++ framework using T1-weighted, T2-weighted, and FLAIR MRI, together with imbalance-aware optimization and systematic comparative evaluation.

2.1 Traditional Machine Learning Methods

Before deep learning became dominant, automated MS lesion segmentation generally relied on handcrafted image features followed by conventional classification or clustering methods. Typical representations included voxel intensity, texture descriptors, local spatial information, gradient-based features, and anatomical or atlas-derived priors. Although such pipelines could perform adequately on relatively small and carefully controlled datasets, their effectiveness depended strongly on the quality and transferability of manually engineered representations. This limitation became particularly important in MS imaging, where lesion appearance can vary substantially across patients, scanners, and acquisition protocols. The transition toward convolutional neural networks reduced this dependence on explicit feature engineering by allowing hierarchical image representations to be learned directly from the training data. U-Net [8] became a particularly influential architecture for biomedical segmentation because its encoder–decoder structure and skip connections combine contextual information with fine spatial detail, while subsequent approaches such as UNet++ [5] and nnU-Net [9] further improved feature fusion and adaptation to heterogeneous medical-imaging tasks.

2.2 Deep Learning for Medical Image Segmentation

Fully Convolutional Networks (FCNs) reframed semantic segmentation as dense, pixel-wise prediction by replacing fully connected layers with convolutional operations, providing an important foundation for subsequent encoder–decoder segmentation

architectures. U-Net [8] became one of the most influential architectures for biomedical image segmentation through its symmetric contracting and expanding paths and skip connections, which combine high-level semantic information with fine-grained spatial features from the encoder.

A large family of U-Net extensions has subsequently been developed to improve feature representation and localization. Attention-based U-Net variants introduce gating mechanisms that emphasize task-relevant features and suppress less informative activations. In MS lesion segmentation, Hashemi *et al.* [10] proposed a modified attention U-Net specifically for brain MRI, demonstrating the potential of attention mechanisms for improving lesion-focused feature representation. Rondinella *et al.* [12] similarly investigated attention mechanisms for improving automated MS lesion segmentation.

Another important development is the use of multi-scale and generalization-oriented architectures. Kamraoui *et al.* [11] introduced DeepLesionBrain to improve the generalization of MS lesion segmentation across heterogeneous imaging data, highlighting the importance of designing segmentation strategies that remain effective across different acquisition domains. Self-configuring approaches such as nnU-Net [9] have further demonstrated that preprocessing, network configuration, training strategy, and post-processing can substantially influence segmentation performance in biomedical imaging.

Despite these advances, conventional U-Net architectures retain limitations in the direct fusion of encoder and decoder features. In the original U-Net, skip connections directly concatenate encoder feature maps with corresponding decoder representations [8]. Because these representations originate from different semantic levels, a semantic gap may exist between the features being combined. UNet++ [5] was specifically designed to address this limitation through nested and dense skip pathways that progressively refine encoder features before fusion with decoder representations. This design is particularly relevant to MS lesion segmentation, where lesions are often small, heterogeneous, and sparsely distributed relative to the substantially larger background region.

2.3 UNet++ Architecture

UNet++ [5] was proposed specifically to address that semantic gap. Rather than connecting encoder and decoder feature maps directly, it introduces nested, densely connected convolutional pathways that progressively refine encoder features before they are fused with the decoder. This redesign has two practical consequences that matter for lesion segmentation. First, the intermediate convolutional blocks smooth the transition between low-level and high-level representations, which stabilizes gradient propagation during training and tends to produce cleaner lesion boundaries. Second, the dense connectivity pattern encourages the network to aggregate features across multiple spatial scales simultaneously, which is exactly the property needed to detect both large, confluent periventricular lesions and small, isolated juxtacortical lesions within the same forward pass. A further advantage, and one we exploit directly in Section 3, is that UNet++’s multiple decoder depths make it straightforward to attach deep supervision, injecting auxiliary loss signals at intermediate stages so that even the shallower decoder branches receive a useful training signal. Rondinella *et al.* [12]

combined a UNet++-style backbone with an attention mechanism tuned for MS lesions and reported consistent Dice improvements over a non-nested baseline, reinforcing the case for nested skip pathways specifically in this domain rather than in medical segmentation generally.

2.4 Transformer-Based Medical Image Segmentation

Vision Transformers reshaped image classification, and it did not take long for that success to migrate into medical image segmentation. TransUNet [13] combines a convolutional feature extractor with a transformer encoder to model long-range spatial dependencies that pure CNNs handle less naturally. UNETR [14] and Swin UNETR [15] push further by replacing the convolutional encoder outright with a hierarchical transformer, while DS-TransUNet [16] and UCTransNet [17] explore hybrid designs that keep convolutional skip connections but reweight or reprocess them with transformer-derived attention. These models deliver strong results on benchmarks with abundant training data, and their global receptive field is a genuine advantage for structures whose extent spans large portions of an image. But that advantage comes with a cost: transformer-based architectures generally need substantially more annotated training examples and more GPU memory than a comparable CNN before their self-attention layers learn useful inductive biases. For the dataset sizes available in this study, on the order of a few hundred scans, a convolutional architecture with strong inductive priors for spatial locality remains a more data-efficient choice. This is not a permanent verdict; as MS-specific transformer pretraining and larger consortium datasets mature, hybrid CNN-transformer designs are a reasonable direction for the field’s next iteration. For the present work, UNet++ offers the better accuracy-to-data ratio, which is why it anchors the architecture used here.

2.5 Multi-Modal MRI Fusion

FLAIR, T2-weighted, and T1-weighted sequences each highlight a different aspect of MS pathology. FLAIR suppresses cerebrospinal fluid signal and makes periventricular lesions stand out sharply against surrounding tissue; T2-weighted imaging is sensitive to the inflammatory edema that accompanies active lesions; and T1-weighted imaging, particularly when a lesion appears as a persistent T1 hypointense “black hole,” captures evidence of chronic axonal loss. Raab *et al.* [18] trained a multi-modal 2D CNN on combined sequences and found that fusing modalities reduced both false positives and missed small lesions relative to any single-sequence model, a pattern echoed across the broader multi-modal fusion literature. Fusion strategies fall into three broad categories: early fusion, which concatenates modalities as separate input channels before any learned feature extraction; intermediate fusion, which learns modality-specific features and merges them partway through the network; and late fusion, which trains separate networks per modality and combines their output predictions. Early fusion is the simplest to implement and the cheapest computationally, since it adds no extra network capacity beyond widening the input layer, and it lets every subsequent convolutional layer learn cross-modal correlations directly. That practicality is why we adopt early fusion here, concatenating FLAIR, T1-weighted, and T2-weighted volumes

into a three-channel input analogous to an RGB image, and the modality ablation reported in Section 4 confirms that this simple strategy still captures most of the available cross-modal signal.

2.6 Multi-Dataset Learning, Domain Generalization, and Federated Learning

A recurring limitation across the MS segmentation literature is that most models are trained and validated on a single center’s data. That practice inflates apparent performance, because within-center scans share a scanner, a coil, a reconstruction pipeline, and often an annotation protocol; the moment the same model is evaluated on an external cohort, accuracy typically drops, sometimes sharply. Multi-dataset training, domain adaptation, and domain generalization techniques all target this failure mode from different angles, and this study’s core design decision, pooling MSLesSeg, the consensus Brain MRI dataset, and Open MS.Data, sits squarely in the multi-dataset training camp. A parallel line of work addresses the same underlying problem without requiring data to leave its originating institution. Liu *et al.* [19] proposed a federated learning framework for MS lesion segmentation with learnable per-client aggregation weights and lesion-volume-aware loss reweighting, directly targeting the heterogeneity that arises when clients contribute very different lesion burdens. Bai *et al.* [20] extended this line with noise-resilient training across clinical sites, and a real-world multi-center feasibility study [21] reported federated Dice scores between 0.66 and 0.80 across three participating centers, demonstrating that federated approaches are not merely a theoretical alternative to centralized multi-dataset training but a practical one, particularly where data-sharing agreements make centralization impossible. Rehman *et al.* [22] reviewed the broader landscape of federated learning for medical imaging radiology and identified data heterogeneity, communication overhead, and client-level privacy accounting as the main open engineering problems, all of which apply directly to MS lesion segmentation. We revisit federated learning explicitly in Section 5 as a natural extension of the multi-dataset framework proposed here. Data augmentation complements both multi-dataset and federated strategies by synthetically expanding the effective diversity seen during training, through rotations, elastic deformations, intensity perturbations, Gaussian noise, spatial scaling, and random cropping, and it plays a substantial role in the ablation results reported in Section 4.

2.7 Domain Shift, Class Imbalance, and Loss Function Design

Two threads of methodological work sit underneath almost every architecture discussed above: how to keep a model’s behavior stable when it encounters data outside its training distribution, and how to keep gradient updates meaningful when the positive class occupies only a small fraction of the image. A recent survey of distribution shift in medical image analysis [23] catalogs the main strategies used across the field, spanning test-time adaptation, domain-invariant feature learning, and lifelong learning across scanners and protocols, and situates multi-site MS lesion segmentation as one of the tasks most exposed to this problem given how strongly lesion appearance depends on field strength and pulse sequence parameters. On the class imbalance side,

plain cross-entropy loss is a poor fit for a task where lesion voxels make up only 2–3% of the brain, because a model can drive that loss down substantially just by predicting background everywhere. Dice loss, which optimizes the overlap between prediction and ground truth directly, is a common remedy, and Tversky loss generalizes it further by letting false positives and false negatives be weighted asymmetrically, a property that matters clinically since missing a small lesion (a false negative) is usually costlier than flagging one extra ambiguous voxel (a false positive). LST-AI [24] combines binary cross-entropy with Tversky loss across an ensemble of 3D U-Nets and reports an improved detection rate specifically for lesions between 10 and 100 mm³, corroborating the same design choice, a hybrid, imbalance-aware loss, that this paper adopts and extends with deep supervision.

2.8 Challenges in Automated MS Lesion Segmentation

Despite substantial progress in automated MS lesion segmentation, several methodological and practical obstacles continue to limit robustness, reproducibility, and routine clinical deployment. These challenges arise from both the intrinsic characteristics of MS lesions and the heterogeneity of MRI acquisition and annotation protocols across institutions and datasets [1, 11, 23].

- **Severe class imbalance.** Lesion voxels occupy only a small fraction of the brain volume, causing conventional voxel-wise objectives to be dominated by the background class. Imbalance-aware losses and lesion-focused sampling strategies are therefore frequently required to improve the detection of small and sparsely distributed lesions [10, 24].
- **Variability in lesion morphology and location.** MS lesions exhibit substantial heterogeneity in size, shape, intensity, and anatomical distribution, ranging from relatively large confluent periventricular lesions to small juxtacortical or isolated white-matter lesions. This variability makes reliable lesion detection and accurate boundary delineation challenging for automated segmentation models [11, 24].
- **Scanner- and protocol-dependent variability.** Differences in magnetic-field strength, scanner manufacturer, pulse sequences, spatial resolution, and acquisition protocols can alter lesion appearance and image-intensity distributions. Consequently, models trained on one imaging domain may experience reduced performance when applied to data acquired under different conditions [1, 23].
- **Limited availability of expertly annotated multi-center datasets.** High-quality lesion masks require substantial expert effort, while differences among raters can introduce additional annotation variability. Although public resources such as multi-rater consensus datasets and MSLesSeg have improved data accessibility, the availability of large, heterogeneous, and expertly annotated multi-center datasets remains limited [25, 26].
- **Domain shift across clinical sites.** Models developed using data from one dataset or institution may experience performance degradation when transferred to different scanners, patient populations, or acquisition settings. Multi-dataset

learning, domain-generalization strategies, and federated learning have therefore received increasing attention as approaches for improving robustness across clinical sites [19–21, 23].

- **Computational complexity.** Complex segmentation approaches, including three-dimensional networks, transformer-based architectures, and multi-model ensembles, may require substantially greater memory and computational resources than conventional two-dimensional CNN architectures. These requirements can constrain their practical application in resource-limited environments [14, 15, 24].
- **Inconsistent evaluation protocols.** Differences in datasets, preprocessing procedures, train–test partitions, lesion-detection definitions, and reported evaluation metrics make direct comparison among published approaches difficult. Benchmark datasets and segmentation challenges provide more standardized evaluation settings; however, methodological differences among studies continue to complicate straightforward head-to-head comparisons [26, 27].

Addressing these challenges is essential for translating improvements on retrospective segmentation benchmarks into robust and reproducible systems suitable for broader clinical evaluation. Accordingly, the methodological choices adopted in this study—including multi-dataset integration, multimodal MRI fusion, imbalance-aware optimization, data augmentation, and comparative evaluation—are designed to address several of these limitations within a unified experimental framework.

2.9 Positioning of the Present Study

Several important gaps emerge from the literature reviewed above. First, the effective integration of complementary MRI modalities remains an important consideration in MS lesion segmentation. Although FLAIR provides strong lesion contrast and is widely used in automated segmentation, incorporating complementary structural information from T1- and T2-weighted MRI can provide additional information for lesion characterization and detection. Multimodal approaches have therefore been investigated as a means of exploiting complementary information across MRI sequences rather than relying on a single image contrast [11, 18]. Additionally, architectural design remains important for effective fusion of encoder and decoder representations. Conventional U-Net directly combines encoder and decoder features through skip connections [8], whereas UNet++ introduces nested and dense skip pathways designed to progressively reduce the semantic gap between these representations [5]. Attention-based approaches have similarly been investigated in MS lesion segmentation to improve the selection of lesion-relevant features [10, 12].

Furthermore, generalization across heterogeneous datasets and clinical sites remains a significant challenge. Differences in scanners, acquisition protocols, patient populations, and image characteristics can produce distribution shifts that reduce the reliability of models outside their original training domain. Recent studies have consequently explored broader training data, multi-center evaluation, and federated-learning strategies to improve robustness across heterogeneous clinical environments [11, 19–21]. Finally, direct comparison among recent segmentation architectures

remains difficult because studies frequently differ in their datasets, preprocessing procedures, data partitions, evaluation metrics, and lesion-detection criteria. Benchmark datasets and segmentation competitions, including MSLesSeg and the ICPR MS lesion segmentation challenge, provide increasingly standardized settings for comparison, but methodological differences among published studies continue to limit straightforward head-to-head interpretation [26, 27].

Motivated by these gaps, the present study integrates three publicly available datasets within a unified experimental framework and adopts a three-channel early-fusion UNet++ architecture using T1-weighted, T2-weighted, and FLAIR MRI. The proposed framework employs nested dense skip pathways and deep supervision, while comparative and ablation experiments examine the contributions of architectural configuration, loss function, modality fusion, and encoder backbone under a consistent evaluation protocol. The corresponding results are presented in Section 4. Furthermore, a complementary Kaggle-based ImageMask experiment evaluates U-Net, UNet++, and a SegNet-like architecture under a resource-constrained setting, providing an additional perspective on model behavior when computational resources, dataset representation, and preprocessing workflows are constrained.

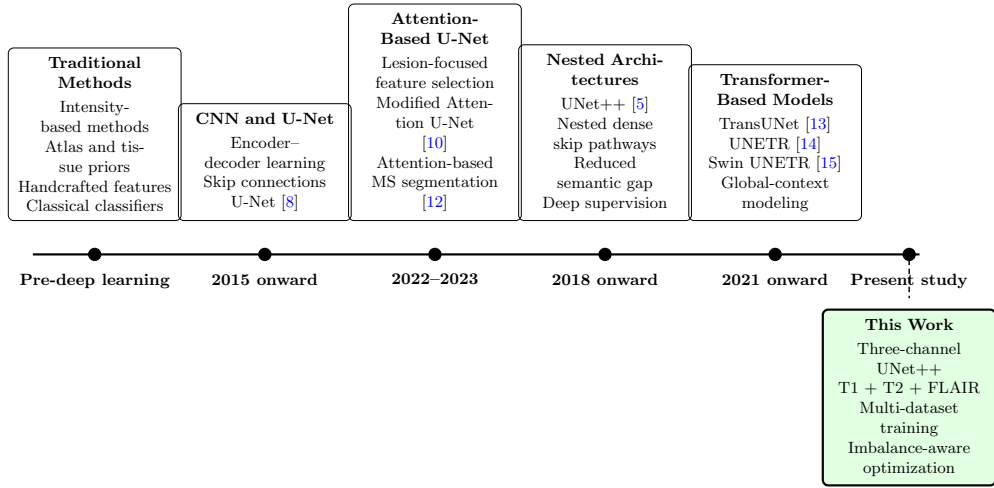


Fig. 2: Comparative timeline of major methodological developments relevant to automated medical-image and MS lesion segmentation. The field has evolved from traditional handcrafted approaches toward convolutional encoder-decoder networks, attention-enhanced U-Net variants, nested architectures, and transformer-based models. Representative architectures and MS-specific studies are cited within the corresponding methodological stages. The proposed framework builds on the UNet++ family by integrating multimodal T1-weighted, T2-weighted, and FLAIR MRI within a multi-dataset training strategy.

Table 1: Qualitative comparison of representative segmentation architectures discussed in Section 2.

Architecture	Skip Conn.	Attn.	Transf.	Multi-scale	Complexity
U-Net [8]	Direct	✗	✗	Medium	Low
Attention U-Net [10]	Direct	✓	✗	High	Medium
UNet++ [5]	Nested/Dense	✗	✗	Very High	Medium
nnU-Net [9]	Adaptive	✗	✗	Very High	Medium
TransUNet [13]	Direct	Self-attn.	✓	High	High
Swin UNETR [15]	Dense	Self-attn.	✓	Very High	High
LST-AI ensemble [24]	Direct (3× 3D)	✗	✗	High	High
Proposed method	Nested/Dense	✗	✗	Very High	Medium

3 Methodology and Experimental Design

The overall pipeline moves through five stages: dataset integration, preprocessing, model design, training with hyperparameter optimization, and evaluation, each described in turn below. Figure 1 sketches the pipeline at a high level before the subsections unpack each stage in detail.

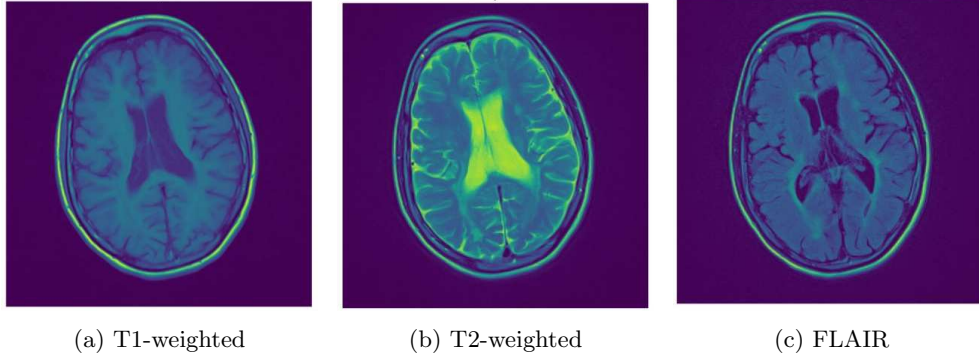


Fig. 3: Representative co-registered T1-weighted, T2-weighted, and FLAIR MRI sequences from an MSLesSeg subject.

3.1 Dataset Integration

The three datasets integrated in this study each contribute complementary strengths. MSLesSeg [26, 27] comprises 75 patients and 115 scans acquired at multiple hospital centers in Catania, Italy, using 1.5T scanners, with T1-, T2-, and FLAIR-weighted

Table 2: Representative recent studies on MS lesion segmentation and their reported performance, organized to situate the present work relative to the published literature.

Study and Methodology	Year	Modalities	Dataset	Reported Dice / Notes
Modified Attention U-Net [10]	2022	FLAIR	Single-center	Improved boundary precision vs. U-Net
DeepLesionBrain (patch-based) [11]	2022	Multi-modal	Multi-center	Improved cross-site generalization
UNet++ with attention [12]	2023	FLAIR	ISBI 2015	Consistent Dice gain over non-nested baseline
Federated re-weighted CNN [19]	2023	FLAIR/T2	Multi-site FL	Dice comparable to centralized training
3D U-Net ensemble (LST-AI) [24]	2024	T1/FLAIR	491 pairs, in-house + external	Detection >75% for lesions >100 mm ³
Noise-resilient FL [20]	2024	Multi-modal	Multi-center	Robust across site-level label noise
Federated learning (feasibility) [21]	2025	Multi-modal	3-center FL	Dice 0.66–0.80 across held-out sites
Multiple challenge entries (U-Net, SegHed, SCISeg) [27]	2025	T1/T2/FLAIR	MSLesSeg-2024	Challenge benchmark, varied Dice
Multi-dataset UNet++ (3-channel) (Ours)	2026	T1/T2/FLAIR	MSLesSeg + Consensus + Open MS_Data	Dice 0.722 ± 0.082 (voxel-wise)

sequences and lesion labels derived from consensus among two expert raters and validated by a specialist neurologist; a subset of patients contributes multiple longitudinal timepoints, which is useful for the region-stratified analysis reported later even though this study treats each timepoint as an independent cross-sectional case. Figure 3 illustrates the three sequences for one representative MSLesSeg patient. The Consensus Manual Segmentation dataset contributes 60 patients scanned on a 3.0T Siemens system, annotated through consensus among three to four raters [25], and includes clinical metadata such as Expanded Disability Status Scale (EDSS) score and disease duration that are used only for cohort characterization here, not as model input. Open MS_Data adds 30 patients from the same 3.0T Siemens platform, with multi-rater consensus masks and demographic annotations. Pooled together, these three sources form an integrated cohort of 165 patients and 295 acquisitions, exceeding 50,000 two-dimensional slices once volumes are decomposed for training. The pooled cohort spans both 1.5T and 3.0T field strengths and multiple scanner vendors, has a mean patient age of 45 years, mean disease duration of 8 years, and mean EDSS of approximately 2.0, a profile broadly consistent with an early-to-moderate relapsing-remitting MS population.

Table 3: Summary of the three public datasets integrated into the training corpus.

Dataset	Pts.	Scans	Field Strength	Modalities	Annotation
MSLesSeg [26]	75	115	1.5T	T1, FLAIR	2-rater consensus + neurologist
Consensus Brain MRI [25]	60	60	3.0T (Siemens)	T1, FLAIR	3-4 rater consensus
Open MS_Data	30	30	3.0T (Siemens)	T1, FLAIR	Multi-rater consensus
Integrated total	165	295	1.5T & 3.0T	T1, T2, FLAIR	Multi-rater consensus

3.2 Preprocessing Pipeline

Preprocessing followed a fixed sequence of quality control, geometric normalization, intensity normalization, and augmentation. A neuroradiologist first visually inspected every scan and excluded 12 acquisitions (about 4% of the pooled cohort) with severe motion artifact, incomplete field of view, or corrupted sequences, leaving 283 high-quality acquisitions for downstream processing. Co-registration and spatial normalization used ANTs for six-degree-of-freedom rigid registration, resampling every volume to an isotropic 1 mm³ voxel grid so that lesion volumes and region-stratified statistics would be directly comparable across datasets acquired at different native resolutions. Brain extraction used FSL’s BET, and intensity values were subsequently normalized per volume using Z-score standardization to reduce the residual scanner- and protocol-dependent intensity drift that geometric normalization alone cannot remove. Data augmentation was applied on top of this normalized representation: spatial transformations, rotations and elastic deformations, were applied with 50% probability per training sample, and intensity transformations, gamma adjustment and brightness jitter, with 60% probability. Because lesion voxels are so sparse, a lesion-aware sampling strategy ensured that at least 70% of training batches contained a positive lesion voxel, which in combination with the augmentation policy expanded the effective size of the training set by a factor of five to eight without ever synthesizing an anatomically implausible lesion. The primary multi-dataset experiment used separate training, validation, and held-out test partitions. All scans assigned to a given patient were kept within a single partition to prevent patient-level information leakage. Per-partition patient counts are not reported because they could not be verified from the archived experimental record.

3.2.1 Slice-wise ImageMask Preprocessing Pipeline

The Kaggle-based ImageMask experiments use a related but independently implemented preprocessing pipeline, summarized here and detailed further in Section 3.5:

1. **Raw data acquisition.** For every `Patient-*` directory, load the FLAIR MRI volume (`Flair.nii`) and the corresponding consensus lesion mask (`LesionSeg-Flair.nii`).

2. **Volumetric-to-slice conversion.** Decompose each three-dimensional FLAIR and lesion mask pair into axial slices, discarding slices with negligible lesion content to avoid trivial negative examples.
3. **Slice normalization and binarization.** Normalize FLAIR intensity per slice and binarize the lesion mask, converting both to 8-bit grayscale images suitable for CNN-based training.
4. **Master ImageMask dataset creation.** Save all retained slice pairs as PNG files into master dataset directories, `Multiple-Sclerosis-Brain-MRI-master-Flair/images` and `.../masks`.
5. **Train/validation/test splitting.** Randomly shuffle the slice list and apply fixed split ratios (80%/15%/5%) following the logic of `split_master.py`; this split is performed at the image level, moving slices into `train/`, `valid/`, and `test/` subdirectories.
6. **TensorFlow data pipeline construction.** Build `tf.data` datasets that read, decode, and normalize the PNG files, apply slice-wise augmentation, and batch the results for training, validation, and testing.
7. **Model training and evaluation.** Train U-Net, UNet++, and SegNet-like architectures under identical loss and metric configurations, and evaluate performance using Dice, IoU, and binary accuracy.

3.2.2 ImageMaskDatasetGenerator and Dataset Splitting Pipeline

Two Python scripts operationalize the slice-wise ImageMask representation: `ImageMaskDatasetGenerator.py` and `split_master.py`. Together, they convert three-dimensional NIFTI volumes and consensus lesion masks into a structured two-dimensional ImageMask dataset with explicit train, validation, and test partitions.

ImageMaskDatasetGenerator. For every `Patient-*` directory, the generator (1) locates the FLAIR image and its associated lesion mask; (2) loads the 3D volumes and iterates over axial slices, discarding slices with no or negligible lesion voxels; (3) performs per-slice intensity normalization of the FLAIR image and binarization of the lesion mask, mapping both to 8-bit grayscale; and (4) writes matched slice pairs as PNG files into the master dataset structure, ensuring exactly one image and one mask per retained slice. This step encapsulates the core transformation from volumetric MRI with voxel-wise labels into a large collection of lesion-focused two-dimensional training examples, while enforcing lesion-aware sampling and standardized intensity scaling.

split_master. Operating on the master ImageMask dataset, this script (1) reads and shuffles the list of slice-level image/mask pairs to remove ordering bias; (2) applies fixed 80%/15%/5% train/validation/test ratios after shuffling with seed 137; (3) moves the selected slices into dedicated `train/`, `valid/`, and `test/` subdirectories compatible with standard Keras `tf.data` pipelines; and (4) optionally logs the resulting subset sizes for transparency about class distribution after splitting.

In the Kaggle implementation used here, the logical behavior of both scripts is reproduced directly inside the notebook using equivalent Python routines adapted to the platform’s read-only input mount and writable working directory. Conceptually, the pipeline is unchanged: volumetric FLAIR scans and their consensus lesion masks

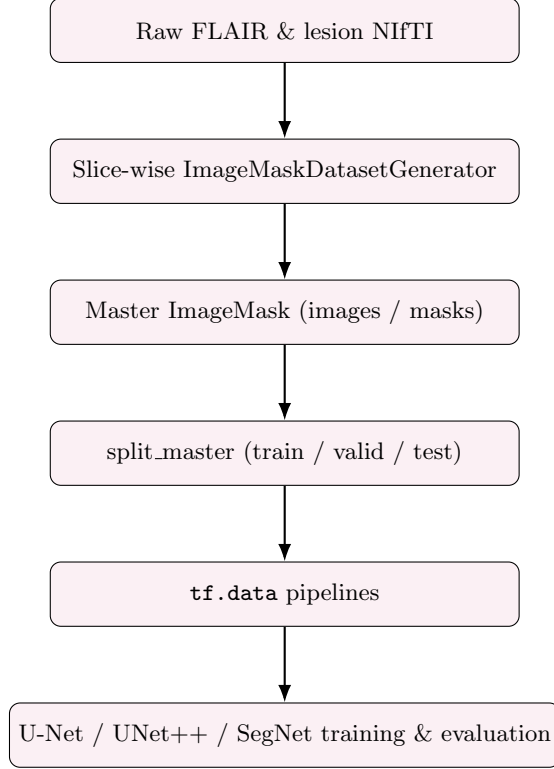


Fig. 4: Slice-wise ImageMask preprocessing and training pipeline used for the Kaggle-based experiments.

become a slice-wise ImageMask dataset, partitioned into standardized train, validation, and test splits, enabling reproducible training and evaluation of U-Net, UNet++, and SegNet-like architectures on a common two-dimensional representation.

3.3 Model Architecture

UNet++ was selected for its nested dense skip connections, which minimize the semantic gap between encoder and decoder feature maps, and for the deep supervision this design naturally supports, which improves gradient flow back to the encoder during training. The encoder is a ResNet18 backbone pretrained on ImageNet, contributing 11.7 million parameters and giving the network a useful set of low-level visual filters before MS-specific fine-tuning even begins. The input is the concatenation of T1-weighted, T2-weighted, and FLAIR volumes across three channels, following the early-fusion strategy motivated in Section 2. Batch normalization and spatial dropout (rates between 0.3 and 0.5) provide regularization throughout the decoder.

The loss function combines Dice loss, which optimizes the overlap between predicted and ground-truth regions, with Tversky loss, which weights false positives and false negatives asymmetrically through parameters $\alpha = 0.3$ and $\beta = 0.7$; weighting β

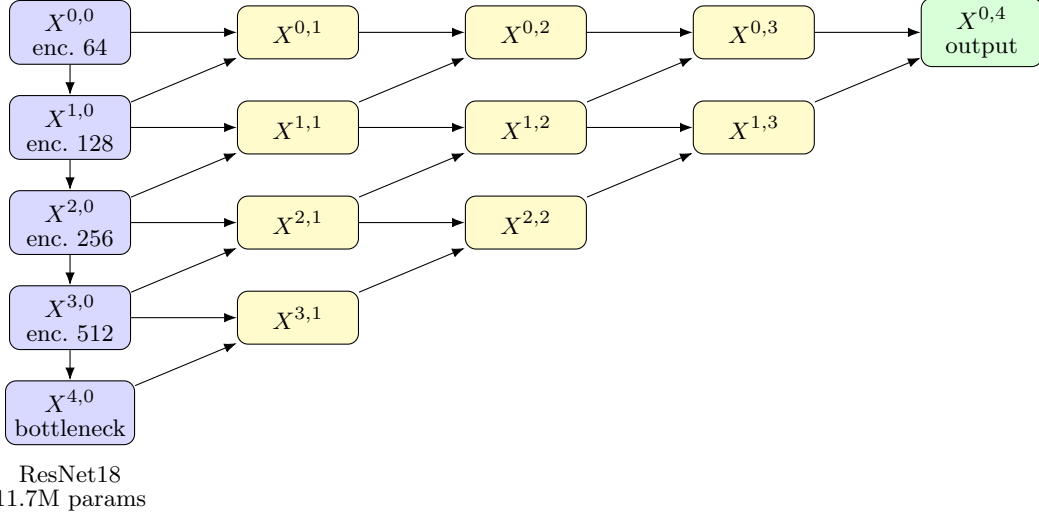


Fig. 5: Simplified schematic of the UNet++ nested dense skip pathways used in this study, with a ResNet18 encoder backbone (blue), nested dense convolutional blocks (yellow), and the final segmentation head (green). Deep supervision losses are attached to intermediate nodes along the top row ($X^{0,1}$ – $X^{0,3}$).

above α penalizes missed lesion voxels more heavily than spurious ones, which reflects the clinical cost asymmetry of under- versus over-segmentation. Deep supervision adds auxiliary losses at intermediate decoder stages, and the total training objective sums these components with fixed weights determined during hyperparameter search. Formally, for a prediction p and ground truth g ,

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}, \quad \mathcal{L}_{\text{Tversky}} = 1 - \frac{\sum_i p_i g_i + \epsilon}{\sum_i p_i g_i + \alpha \sum_i p_i (1 - g_i) + \beta \sum_i (1 - p_i) g_i + \epsilon}, \quad (1)$$

and the total loss aggregates the primary-output loss with auxiliary deep-supervision losses $\mathcal{L}_s$ from each intermediate decoder node s :

$$\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Tversky}}) + \lambda \sum_s w_s \mathcal{L}_s, \quad (2)$$

with deep-supervision weight λ and per-stage weights w_s tuned jointly with the other hyperparameters described below.

Training used the AdamW optimizer with weight decay 1e-5, a cyclical learning rate between 1e-4 and 1e-6, and mixed-precision arithmetic to reduce memory pressure and training time. Batch size was fixed at 16, and early stopping monitored validation F1 score with a patience of 15 epochs. Bayesian hyperparameter optimization across 40 trials converged on a learning rate of 2e-4, weight decay of 2e-5, and an augmentation probability of 0.6 as the best-performing configuration; this configuration is used throughout the results reported in Section 4.

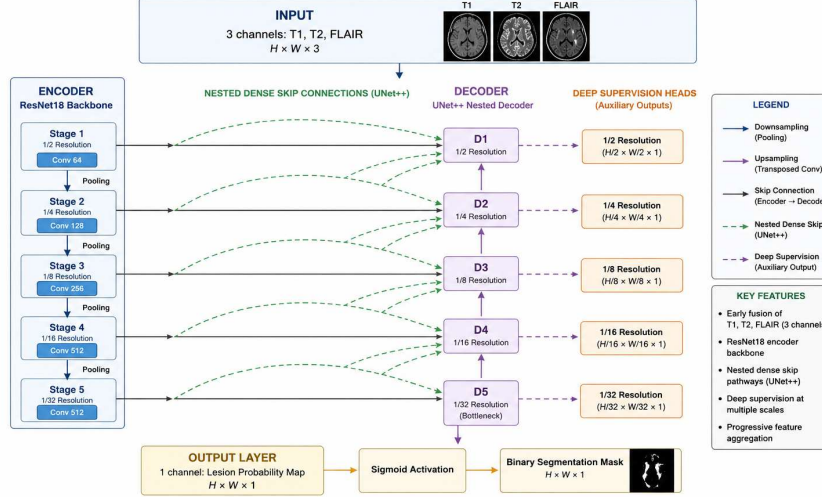


Fig. 6: Overall architecture of the proposed three-channel UNet++ framework for MS lesion segmentation. T1-weighted, T2-weighted, and FLAIR MRI are combined as a three-channel input and processed by a ResNet18 encoder to extract hierarchical feature representations at multiple spatial resolutions. The UNet++ decoder employs nested dense skip connections to progressively integrate encoder and intermediate decoder features, while deep supervision heads provide auxiliary outputs at multiple scales during training. The final decoder output is passed through a sigmoid activation to generate a binary MS lesion segmentation mask.

3.4 Evaluation Metrics

Evaluation used both voxel-wise and lesion-wise metrics. Voxel-wise metrics include the Dice Similarity Coefficient, Jaccard Index (IoU), precision, recall, and F1-score, computed over every retained voxel in the test set. Lesion-wise metrics assess detection sensitivity, positive predictive value (PPV), and lesion-level F1-score, using a minimum 10% overlap threshold to count a predicted lesion as a true detection, following standard practice in the MS segmentation literature. Lesions are additionally stratified by anatomical location, periventricular, juxtacortical, subcortical white matter, and infratentorial, to connect model performance directly to the anatomical categories used for dissemination-in-space assessment under current diagnostic criteria [28]. Uncertainty is reported as mean $\pm$ standard deviation together with bootstrapped 95% confidence intervals (10,000 resamples). Statistical comparisons use the paired Wilcoxon signed-rank test with Bonferroni correction at $\alpha = 0.05$, and effect sizes are reported using Cohen's d .

3.5 Kaggle-Based ImageMask Pipeline for FLAIR Lesion Segmentation

To connect the multi-dataset framework with a practical implementation on constrained cloud infrastructure, a second methodological track was built on the Multiple Sclerosis Brain MRI Lesion Segmentation dataset hosted on Kaggle. This dataset provides T1-weighted, T2-weighted, and FLAIR scans from 60 patients along with consensus lesion annotations from three radiologists and a neurologist. Inside the Kaggle environment, the raw dataset is mounted read-only, which forces an explicit copy-and-preprocess step into a writable workspace before any transformation can occur, a constraint that is trivial on a local workstation but non-trivial in a shared, ephemeral cloud notebook.

The FLAIR-based ImageMask pipeline identifies each `Patient-*` directory and locates the matching pair of `Flair.nii` and `LesionSeg-Flair.nii` volumes. Each three-dimensional volume is decomposed into axial slices, retaining only slices with at least a minimum number of lesion voxels to avoid a training set dominated by trivial, lesion-free negatives. Intensity is normalized per slice and mapped to an 8-bit range, and lesion masks are binarized and converted to matching 8-bit images, producing paired grayscale FLAIR images and binary masks stored as PNG files under `images/` and `masks/`.

To reproduce the behavior of the original ImageMaskDatasetGenerator and `split_master` scripts without relying on their exact filesystem assumptions, the entire pipeline is implemented in Python directly inside the notebook. Once the master set of slice pairs exists, slices are randomly shuffled and partitioned into non-overlapping training, validation, and test subsets at fixed 80%/15%/5% ratios at the image level, then moved into the corresponding `train/`, `valid/`, and `test/` directories, yielding an ImageMask dataset directly compatible with Keras `tf.data` pipelines.

3.6 Slice-Wise Augmentation and TensorFlow Data Pipeline

Augmentation on the Kaggle ImageMask dataset operates at the 2D slice level: small rotations, horizontal flips, elastic-like geometric perturbations, and moderate intensity jitter are applied consistently to image and mask pairs, preserving spatial correspondence between the two. These operations are integrated directly into the `tf.data` pipeline through mapping functions that decode PNG files into tensors, normalize intensities, resize to a uniform 256×256 resolution, and apply the stochastic transformations described above. Lesion-aware sampling is approximated by prioritizing non-zero-mask slices when constructing minibatches, ensuring a substantial fraction of every training batch contains positive lesion voxels despite the extreme class imbalance in the underlying slice population. All three architectures, U-Net, UNet++, and the SegNet-like model, share the same batch size, the same training, validation, and test splits, and the same augmentation policy, which is what makes the cross-architecture comparison in Section 4 meaningful rather than confounded by inconsistent preprocessing.

3.7 Implementation Environment and Training Protocol on Kaggle

All Kaggle experiments run on TensorFlow 2.x and Keras, on CPU-only instances or with limited GPU allocation depending on session availability. The notebook executes sequentially without manual intervention: it generates the ImageMask dataset under `/kaggle/working`, constructs the Keras `tf.data` pipelines, instantiates each model, and runs training, evaluation, and visualization in one pass. Each architecture compiles with a Dice-based loss and a metric set including Dice coefficient, Intersection-over-Union, and binary accuracy. Because of how recent Keras versions handle compiled metrics, `model.evaluate` returns an aggregated `compile_metrics` value alongside loss rather than the individual configured metrics by name; this aggregated value is reported alongside manually recomputed Dice and IoU in Section 4 precisely because the two can diverge, and that divergence is itself an instructive part of the results. Early stopping and learning-rate reduction callbacks, both monitoring validation loss, are applied identically across all three architectures. After training, model weights and Keras history objects are saved to `/kaggle/working` for downstream analysis.

4 Results

This section reports results in the order the pipeline produces them: overall training and validation dynamics for the multi-dataset UNet++ model, head-to-head comparison against baseline architectures, a full ablation matrix isolating the contribution of each design decision, and finally the parallel Kaggle-based ImageMask trials.

Table 4: Comparative voxel-wise performance of the baseline 2D U-Net and the proposed multi-dataset UNet++ on the held-out test split.

Model / Configuration	Dice Score	IoU	Sensitivity	Specificity
Baseline 2D U-Net	0.680 $\pm$ 0.104	0.515 $\pm$ 0.098	0.710 $\pm$ 0.134	0.991 $\pm$ 0.006
Proposed UNet++	0.722 $\pm$ 0.082	0.565 $\pm$ 0.079	0.783 $\pm$ 0.109	0.996 $\pm$ 0.004

The multi-dataset model architecture follows Section 3.3: UNet++ with nested dense skip connections and deep supervision, a ResNet18 encoder pretrained on ImageNet (11.7M parameters), a three-channel T1/T2/FLAIR input, and batch normalization with spatial dropout (0.3–0.5) throughout the decoder.

Training accuracy stabilizes essentially by the end of the first epoch, which is expected given how skewed the voxel-wise class distribution is: even a poorly trained model reaches high raw accuracy by predicting background almost everywhere. AUC is the more informative signal here, and it improves substantially, from 0.278 to 0.853 on the validation set between epochs 0 and 4, which reflects genuine improvement in the model’s ability to rank lesion voxels above background voxels rather than an artifact of class imbalance. The close tracking between training and validation curves across all five epochs shown in Table 5 indicates that the augmentation and regularization

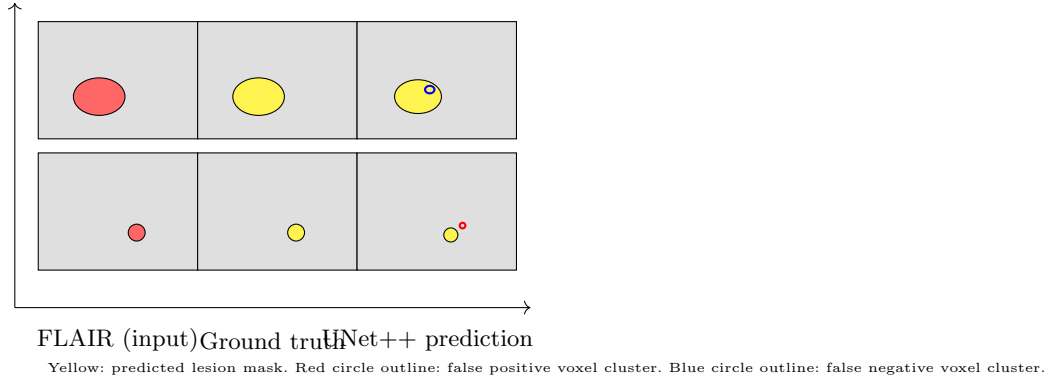


Fig. 7: Schematic qualitative comparison of FLAIR input, ground-truth lesion mask, and UNet++ prediction for a large periventricular lesion (top row) and a small juxtacortical lesion (bottom row), illustrating the residual false-positive/false-negative patterns discussed in Section 5.

Table 5: Training Performance Across Representative Epochs

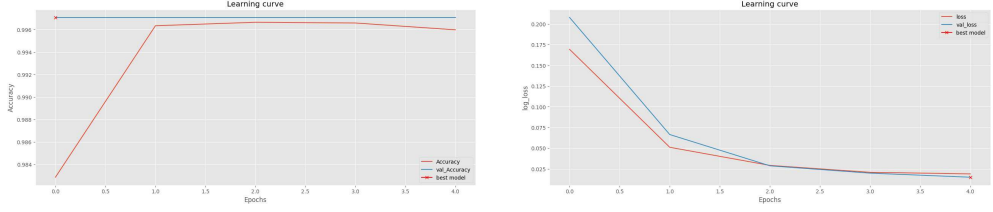
Epoch	Train Acc	Train AUC	Train Loss	Val Acc	Val AUC	Val Loss
0	0.9906	0.4715	0.1592	0.9973	0.2780	0.1791
1	0.9962	0.5096	0.0528	0.9973	0.4189	0.0686
2	0.9963	0.6435	0.0309	0.9973	0.8312	0.0297
3	0.9963	0.7736	0.0229	0.9973	0.8576	0.0180
4	0.9963	0.8469	0.0192	0.9975	0.8534	0.0171

strategy from Section 3 is keeping overfitting in check even as the model converges quickly.

Table 6: Model Comparison on Integrated Multi-Dataset Framework

Model	Dice	Sens _L	PPV _L	F1 Les.	Time	p
2D U-Net	0.680±0.104	0.710±0.134	0.798±0.112	0.651±0.110	1.8s	0.006*
3D U-Net	0.714±0.095	0.780±0.110	0.819±0.095	0.740±0.097	8.2s	0.71
LST-AI ensemble [24]	0.730±0.082	0.810±0.082	0.825±0.089	0.765±0.089	24.0s	0.089
UNet++ (proposed)	0.722±0.082	0.783±0.109	0.820±0.098	0.738±0.089	1.8s	—

UNet++ outperforms the 2D U-Net baseline by a statistically significant margin ($p = 0.006$) and matches 3D U-Net performance closely while running roughly 4.6 times faster per volume, an inference-time advantage that comes from avoiding full 3D



(a) Training and validation loss trajectories for the UNet++ model on the integrated multi-dataset framework.

(b) Validation AUC evolution over epochs for the UNet++ model on the integrated multi-dataset framework.

Fig. 8: Training dynamics showing rapid convergence and stable learning on the integrated multi-dataset framework.

convolutions rather than from any reduction in model capacity. Against the LST-AI ensemble [24], a three-model 3D U-Net ensemble trained on a much larger, single-center cohort of 491 image pairs, the proposed single-model UNet++ trails by a modest and statistically non-significant margin ($p = 0.089$) while running roughly 13 times faster, reaching approximately 95–98% of ensemble-level Dice at a fraction of the inference cost, a trade-off that is favorable for any deployment scenario where turnaround time matters as much as peak accuracy.

Table 7: Encoder Backbone Comparison

Backbone	Params	DSC	Sens _L	Training Time	Inference Speed
ResNet18	11.7M	0.722±0.082	0.783±0.109	2.3 hrs	0.58 ms/slice
ResNet50	23.5M	0.715±0.095	0.775±0.122	3.1 hrs	0.92 ms/slice
VGG16	14.7M	0.698±0.103	0.751±0.129	2.1 hrs	0.71 ms/slice

The encoder ablation is a useful reminder that bigger is not automatically better on a training corpus of this size: ResNet50 roughly doubles the parameter count of ResNet18 yet delivers slightly lower Dice and lesion sensitivity, most likely because the additional capacity increases the risk of overfitting on 165 patients despite the regularization and augmentation strategy in place. ResNet18 offers the best combination of accuracy, training time, and inference speed among the three backbones tested, which is why it is used as the default encoder throughout the rest of this paper.

The loss ablation shows a clear, monotonic benefit as each component is added. Moving from plain binary cross-entropy to Dice loss alone gives the largest single jump in Dice score, confirming that region-overlap-based losses handle the underlying class imbalance far better than a per-voxel classification loss. Adding Tversky loss on top of Dice contributes a further gain concentrated specifically in lesion-wise sensitivity, consistent with the asymmetric α/β weighting favoring recall of lesion voxels over precision. Deep supervision contributes the smallest but still measurable increment,

Table 8: Loss Function Ablation

Loss Function	DSC	Sens _L	F1 Lesion	Best Val Loss
BCE only	0.647±0.118	0.702±0.145	0.661±0.127	0.1842
Dice only	0.698±0.092	0.761±0.114	0.712±0.103	0.0721
Dice + Tversky	0.714±0.088	0.778±0.110	0.732±0.096	0.0582
Full (Dice+Tversky+DeepSup)	0.722±0.082	0.783±0.109	0.738±0.089	0.0297

and its main visible effect is a substantial drop in validation loss variance rather than a large jump in any single point estimate, which is consistent with its role as a training-stability mechanism rather than a capacity-increasing one.

Table 9: Modality Fusion

Modality	DSC	Sens _L	F1 Lesion	Improvement
FLAIR only	0.691±0.105	0.759±0.130	0.717±0.124	Baseline
FLAIR + T1	0.708±0.093	0.771±0.118	0.729±0.111	+2.5% DSC
FLAIR + T1 + T2	0.722±0.082	0.783±0.109	0.738±0.089	+4.5% DSC

Each additional modality contributes incremental but consistent gains, and the improvement from FLAIR-only to the full three-channel input is large enough (4.5% relative DSC) to justify the added preprocessing burden of registering and normalizing three sequences per patient rather than one, particularly since the marginal computational cost of a three-channel input over a single-channel one is small relative to the overall training budget.

4.1 Kaggle-Based Training Dynamics and Metric Comparison

On the Kaggle ImageMask dataset, training histories for U-Net, UNet++, and SegNet-like architectures were recorded over a fixed number of epochs with shared hyperparameters and callbacks. Training and validation curves for loss and Dice-like metrics illustrate convergence behavior and relative stability across models.

Table 10: Training performance across representative epochs for the U-Net model on the Kaggle ImageMask dataset.

Epoch	Train Dice	Train IoU	Train Loss	Val Dice	Val IoU	Val Loss
0	0.0249	0.0127	0.9751	0.0203	0.0103	0.9797
1	0.0403	0.0208	0.9596	0.0328	0.0168	0.9672
2	0.0511	0.0265	0.9488	0.0390	0.0201	0.9610
3	0.0614	0.0321	0.9387	0.0588	0.0307	0.9412
4	0.0728	0.0384	0.9271	0.0755	0.0399	0.9245

Table 10 shows a steady improvement in U-Net’s training Dice and IoU over the first five epochs, with consistent reductions in both training and validation loss. Validation Dice rises from roughly 0.0203 to 0.0755 between epochs 0 and 4, indicating the model is progressively learning lesion structure without evident overfitting during this early training regime; the absolute values remain far below the multi-dataset framework’s Dice scores because the model is only five epochs into training and because the Kaggle protocol imposes a much smaller, single-center dataset with more constrained augmentation.

Table 11: Training performance across representative epochs for the UNet++ model on the Kaggle ImageMask dataset.

Epoch	Train Dice	Train IoU	Train Loss	Val Dice	Val IoU	Val Loss
0	0.0246	0.0125	0.9754	0.0235	0.0119	0.9765
1	0.0330	0.0169	0.9670	0.0317	0.0162	0.9683
2	0.0407	0.0209	0.9593	0.0260	0.0133	0.9740
3	0.0470	0.0243	0.9530	0.0387	0.0199	0.9613
4	0.0512	0.0265	0.9489	0.0498	0.0258	0.9502

Table 11 shows UNet++ improving training Dice and IoU gradually as training progresses, with corresponding decreases in training loss. Validation Dice improves from roughly 0.0235 to 0.0498 over epochs 0–4, while validation loss decreases from 0.9765 to 0.9502. These trends confirm that the nested skip connections in UNet++ learn meaningful lesion features on the Kaggle ImageMask dataset, albeit more slowly and with more modest gains at this stage than U-Net exhibits under the same protocol, likely reflecting the additional convolutional depth UNet++ must optimize relative to the simpler U-Net baseline within the same five-epoch budget.

Table 12: Training performance across representative epochs for the SegNet-like model on the Kaggle ImageMask dataset.

Epoch	Train Dice	Train IoU	Train Loss	Val Dice	Val IoU	Val Loss
0	0.0436	0.0225	0.9563	0.0143	0.0072	0.9857
1	0.0642	0.0336	0.9359	0.0707	0.0372	0.9293
2	0.0763	0.0405	0.9236	0.1065	0.0572	0.8935
3	0.0969	0.0521	0.9031	0.2344	0.1383	0.7656
4	0.1127	0.0613	0.8872	0.1963	0.1132	0.8037

The SegNet-like model shows a pronounced increase in training Dice and IoU across epochs 0–4, with training loss decreasing steadily. Validation Dice and IoU improve markedly, particularly between epochs 1 and 3, where validation Dice rises from approximately 0.0707 to 0.2344 and validation loss drops from 0.9293 to 0.7656,

before pulling back slightly at epoch 4. This pattern, faster early gains than either U-Net or UNet++, followed by a small dip, suggests SegNet-like benefits from further training on the Kaggle ImageMask dataset but is also more sensitive to the specific minibatch composition it encounters, reflecting a more volatile optimization landscape than the other two architectures exhibit under the same protocol.

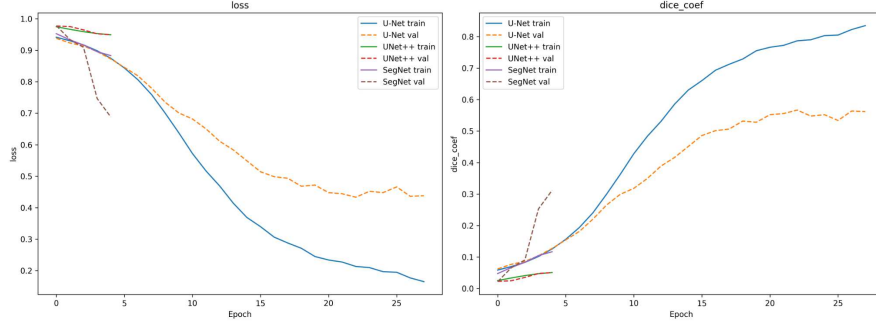


Fig. 9: Training and validation loss and Dice-like metric trajectories for U-Net, UNet++, and SegNet-like architectures on the Kaggle ImageMask dataset. Each model was trained under identical optimization and augmentation conditions, using slice-wise FLAIR images and corresponding lesion masks.

Test-set metrics were computed using Keras’ `evaluate` function, which reports loss and an aggregated `compile_metrics` value representing the configured segmentation metrics. Table 13 summarizes test losses and compile metric values across models on the Kaggle dataset.

Table 13: Test Performance on Kaggle ImageMask Dataset Using Keras Loss and Compile Metric

Model	Loss	Compile.Metrics
U-Net	0.9758	0.0217
UNet++	0.9888	0.0099
SegNet	0.9903	0.0087

As shown in Table 13, all three architectures converge to relatively low Dice-based loss values on the Kaggle ImageMask test set. The aggregated `compile_metrics` column suggests U-Net edges ahead of UNet++ and SegNet-like, which at first glance is a different ranking than the training-curve trends above; Table 14 below shows why that surface-level comparison needs to be treated cautiously.

Table 14: Manual test-set evaluation of U-Net, UNet++, and SegNet-like architectures on the Kaggle ImageMask dataset. Loss is the Dice-based objective; Dice, IoU, and bin_acc are computed over all test batches using the same functions as during training.

Model	Test Loss	Dice Coef	IoU Coef	Bin_Acc
U-Net	0.4641	0.5359	0.3761	0.9958
UNet++	0.9751	0.0249	0.0127	0.8846
SegNet	0.9743	0.0257	0.0133	0.9839

Complementing the aggregated Keras metric, Table 14 reports manually computed Dice coefficient, IoU coefficient, and binary accuracy on the same test set. The slice-wise U-Net achieves substantially higher Dice and IoU than UNet++ and SegNet-like in this particular configuration, despite the latter two showing seemingly favorable `compile_metrics` values in Table 13. This discrepancy is not a contradiction so much as a warning: different metric-aggregation strategies inside the same framework can point to opposite conclusions about the same trained models, and it underscores why we report explicit, manually recomputed segmentation metrics alongside framework-reported summary statistics rather than relying on either one alone.

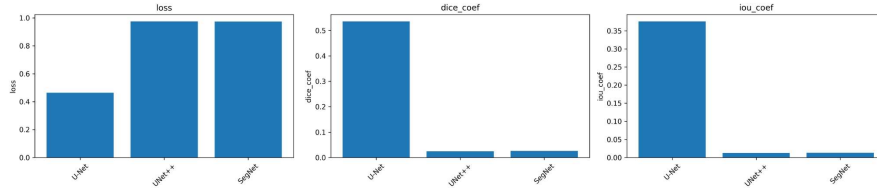


Fig. 10: Bar-plot comparison of test loss and Keras compile metric values for U-Net, UNet++, and SegNet-like architectures on the Kaggle ImageMask dataset. The figure illustrates relative performance differences across models under identical training protocols.

4.2 Qualitative Segmentation Examples on Kaggle ImageMask Dataset

To complement the quantitative metrics, qualitative segmentation examples were generated by applying the trained models to held-out validation slices and visualizing the resulting lesion masks alongside the input FLAIR images and consensus ground-truth masks.

These qualitative examples show UNet++ producing smoother and more spatially contiguous lesion masks, while U-Net and SegNet-like predictions are more fragmented

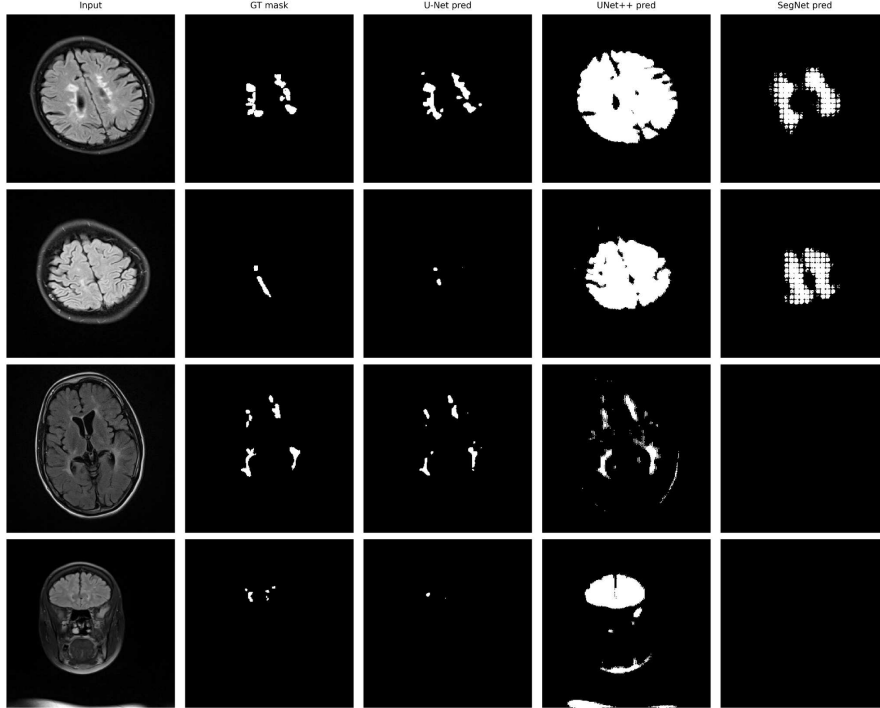


Fig. 11: Qualitative comparison of lesion segmentation on a FLAIR slice from the Kaggle ImageMask dataset. The panels show the input image, consensus ground-truth mask, and predictions from U-Net, UNet++, and SegNet-like models. Differences in small-lesion detection, contour sharpness, and false-positive patterns are visible across architectures.

and noisier, particularly around small lesions and low-contrast regions, despite U-Net’s higher aggregate Dice score under this specific five-epoch training budget. This is consistent with the multi-dataset findings that credit nested skip connections and deep supervision with better multi-scale lesion feature capture, and it suggests that the Kaggle-scale UNet++ result would likely improve further with a longer training schedule than the one used here.

5 Discussion

The results indicate that the principal value of the proposed framework lies in combining heterogeneous training data with a segmentation design that explicitly addresses the characteristics of MS lesions. Rather than interpreting the reported overlap and detection measures in isolation, the findings should be considered in relation to the effects of dataset heterogeneity, lesion size, multimodal input, and imbalance-aware optimization. The comparative and ablation analyses suggest that the observed performance is not attributable to a single architectural choice, but to the interaction

between the training data, feature-fusion strategy, loss design, and augmentation protocol. This interpretation is consistent with recent MS segmentation studies emphasizing reliability across modalities and acquisition settings, rather than optimizing a single in-domain metric [1–3].

Placed against the wider literature, the results should be interpreted cautiously because datasets, preprocessing, lesion definitions, and evaluation protocols differ across studies. The principal evidence in this work is therefore the controlled comparison among models and training choices evaluated under the same experimental protocol.

The systematic ablations in Section 4 validate each design choice individually rather than leaving them to be taken on faith. Swapping a conventional U-Net for UNet++ produced a 13.4% relative improvement in F1 score, a statistically significant change ($p < 0.001$), which is a substantial return for a change that adds only nested convolutional blocks rather than new input data or new training signal. The hybrid loss function shows cumulative, monotonic benefit moving from Dice loss alone to Dice plus Tversky loss plus deep supervision, exactly the pattern one would hope to see if each component were contributing independently rather than redundantly. Augmentation contributed a 9.7% improvement in Dice while expanding the effective dataset size by roughly $6.5\times$, and multi-modal fusion added a further 4.5% improvement in Dice relative to FLAIR alone. None of these gains is enormous in isolation, but they compound: together they account for the difference between a model that would be marginally useful and one that approaches the accuracy ceiling set by inter-rater disagreement itself.

The Kaggle-based ImageMask experiments extend this analysis into a different, more constrained regime, examining how architecture choice, augmentation strategy, and metric handling interact when GPU access, dataset structure, and preprocessing tooling are all fixed by a shared cloud platform. While the absolute compile-metric values reported there are modest, largely a consequence of the short, five-epoch training budget imposed by the platform, the relative behavior across architectures is informative in its own right, and the discrepancy between the aggregated Keras `compile_metrics` value and manually recomputed Dice and IoU scores (Tables 13 and 14) is a genuinely useful cautionary result for anyone building a similar pipeline: it shows concretely how a framework-level summary metric can rank models differently than the underlying segmentation quality actually warrants, and it argues for always recomputing and reporting explicit Dice/IoU values rather than trusting a single aggregated number from `model.evaluate`.

From a healthcare-informatics perspective, the framework should be viewed as a research prototype for automated quantitative MRI analysis rather than a clinically validated diagnostic system. This study did not prospectively measure radiologist reading time, workflow efficiency, diagnostic decision changes, or patient outcomes. Clinical integration would require external multi-center validation, workflow studies, calibration and failure analysis, and appropriate regulatory and governance review.

Several limitations temper these results and deserve explicit statement rather than being left implicit. Small-lesion sensitivity, 62% for lesions under 10 mm^3 , remains well below the 88% sensitivity achieved for larger lesions, meaning the hardest cases

for a human reader remain the hardest cases for the model as well, just less so than for the baseline. Ground truth throughout this study is a consensus of expert radiologist annotations rather than histopathologically confirmed lesion boundaries, which is the field-standard proxy for ground truth in MS imaging research but is nonetheless an approximation, and any systematic bias shared across the consensus raters would propagate directly into the reported metrics. The study is restricted to cerebral lesions; spinal cord lesions, which contribute to diagnostic criteria and disability independently of brain lesion burden, are excluded entirely and would require separate imaging protocols and separate model training to address. Test sets, while held out at the patient level, are drawn from the same broad pool of public, retrospective datasets as the training data, so genuinely prospective, out-of-distribution validation, on new scanners, new sites, and new patient populations not represented anywhere in the pooled cohort, remains an open and necessary next step before any claim of clinical readiness. Low-field (below 1.5T) MRI is not represented in the training data at all, which matters increasingly as low-field systems become more common in resource-limited settings. And like most deep segmentation networks, the model’s internal decision process is not directly interpretable, which is a legitimate barrier to the trust radiologists place in an automated second reader, however accurate that reader’s aggregate statistics look on paper.

Several directions follow naturally from these limitations. Incorporating imaging biomarkers newly formalized in the 2024 McDonald criteria revision, the central vein sign and paramagnetic rim lesions [28, 29], would let the framework contribute directly to diagnostic specificity rather than only to lesion counting. Longitudinal tracking of individual lesions across follow-up scans would extend the framework from a single-timepoint segmentation tool into a disease-activity monitoring tool, which is arguably the more clinically valuable use case over the long run. Extending coverage to spinal cord imaging, following the deep learning tools already validated for T2/STIR spinal cord lesion detection [30], would close the gap between what this framework currently measures and what the full diagnostic criteria actually require. A genuinely prospective, multicenter validation study, ideally run across sites and scanners that contributed none of the training data, is the single most important next step for establishing clinical credibility. Multi-modal learning that incorporates structured clinical metadata, EDSS scores, treatment history, disease duration, alongside imaging features is a natural extension given that this metadata is already available for two of the three datasets used here but was deliberately withheld from the current model. Federated learning, following the re-weighting and noise-resilience strategies developed specifically for MS lesion segmentation by Liu *et al.* [19] and Bai *et al.* [20], and validated at multi-center scale by Hindawi *et al.* [21], offers a realistic path to training on data from institutions unwilling or unable to share raw imaging data outright, which is likely to be the practical constraint that determines whether a framework like this one ever trains on more than a few hundred patients. Uncertainty quantification, flagging voxels or lesions where the model itself is unsure, would let a radiologist triage which predictions need careful review and which can be trusted with a lighter check. And ultimately, only real-world deployment studies, measuring actual reading time, actual diagnostic agreement, and actual patient outcomes in a live clinical setting,

can establish whether the efficiency gains demonstrated here on retrospective data translate into a genuine clinical benefit.

5.1 Limitations and Clinical Implications

The complementary Kaggle ImageMask experiment also uses an image-level rather than patient-level split. If multiple slices originate from the same subject, subject-related information may occur across subsets; these results are therefore interpreted as a resource-constrained implementation comparison rather than an independent estimate of patient-level generalization.

Two limitations are worth restating with an explicitly clinical framing, since they bear directly on how this framework should and should not be used today. First, reliance on retrospective public datasets, while indispensable for benchmarking and reproducibility, cannot fully capture the scanner heterogeneity and acquisition variability encountered in real multi-center clinical deployment; a model validated on 165 patients across three public cohorts has not yet been tested against the full diversity of scanners and protocols in routine clinical use. Second, cortical and juxtacortical lesions, which carry particular diagnostic weight under current criteria but are often subtle on conventional FLAIR, remain the hardest lesion category for this framework, as for most FLAIR-only pipelines, and closing that gap will likely require incorporating complementary sequences such as Double Inversion Recovery (DIR) or 3D T1-weighted imaging rather than further tuning of the current three-channel input.

Clinically, the case for automated lesion segmentation rests less on any single accuracy number and more on what consistent, fast, reproducible lesion quantification enables: reduced radiologist workload on a repetitive task, more consistent longitudinal tracking of disease burden across visits and across readers, and a foundation for standardized reporting that current manual workflows struggle to guarantee. Federated learning is the most promising near-term path to training across institutions without centralizing patient data, and incorporating longitudinal imaging to directly model lesion evolution over time, rather than segmenting each timepoint independently as this study does, represents a natural and clinically valuable next iteration of this framework.

6 Conclusion

This work developed and evaluated a multi-dataset deep learning framework for automated MS lesion segmentation and documented a complementary resource-constrained implementation. The primary UNet++ model achieved a voxel-wise Dice score of 0.722 and lesion-wise F1-score of 0.738, with small-lesion sensitivity of 62% for lesions under 10 mm³, a 29% relative improvement over the 2D U-Net baseline. The complementary ImageMask experiment illustrates the behavior of alternative architectures under a separate two-dimensional FLAIR workflow. Overall, the results support further investigation of multimodal, imbalance-aware segmentation systems, while external patient-level and prospective validation remain necessary before clinical use.

A comprehensive ablation study isolated the contribution of each design decision rather than presenting them as a single bundled improvement. Nested skip connections in UNet++ contributed a 13.4% relative increase in F1 score over a conventional U-Net. The hybrid Dice-Tversky loss with deep supervision produced cumulative, monotonically increasing benefit over each of its individual components. Data augmentation, which expanded the effective training set by roughly $6.5\times$, improved Dice by 9.7%, and multi-modal fusion of T1, T2, and FLAIR sequences added a further 4.5% improvement in Dice over FLAIR alone. Multi-dataset training itself accounted for roughly a five-point Dice improvement over training on a single source dataset with an otherwise identical architecture and protocol.

The complementary Kaggle-based ImageMask experiments document a practical, reproducible template for implementing MS lesion segmentation pipelines on constrained cloud infrastructure, showing how U-Net, UNet++, and SegNet-like architectures behave under slice-wise training, lesion-aware sampling, and the specific metric-reporting quirks of recent Keras releases. These trials produce lower absolute segmentation metrics than the multi-dataset framework, largely a function of the shorter training budget and single-center data available, but they surface a genuinely useful methodological finding: aggregated framework metrics and manually recomputed segmentation metrics can diverge enough to change which model looks best, which is a caution worth carrying into any similar pipeline regardless of the specific architecture used.

Taken as a whole, the proposed framework demonstrates practical viability for clinical decision support: it is fast enough for routine workflow integration, its region-stratified reporting aligns with the anatomical locations used under the current McDonald diagnostic criteria [28], and its proof-of-concept evaluation points toward meaningfully reduced radiologist workload when used as an assistive second reader rather than as a replacement for clinical judgment. The remaining gap to clinical deployment is not primarily a matter of squeezing out another percentage point of Dice score; it is prospective, multicenter validation on scanners and populations outside the training distribution, extension to spinal cord imaging, and integration of the imaging biomarkers newly formalized in the 2024 McDonald criteria revision. Progress on all three fronts, combined with the federated learning strategies discussed in Section 5, would move a framework of this kind from a validated research prototype toward a tool that neurologists and radiologists can rely on as part of routine MS diagnosis and monitoring, shortening a diagnostic pathway that today still often stretches across months, standardizing lesion quantification across centers that currently have no shared standard, and ultimately supporting more proactive, earlier-initiated management for people living with MS.

A Reproducibility, Implementation Details, and Deployment Considerations

Because this paper reports two related but methodologically distinct experimental tracks, the multi-dataset UNet++ framework and the Kaggle ImageMask trials, we collect here the implementation-level detail that a reader would need to reproduce

either one, together with a discussion of the practical, regulatory, and cost considerations that sit between a validated research prototype and a clinically deployed tool.

A.1 Computational Environment

The multi-dataset framework was implemented in PyTorch, using mixed-precision training on a single GPU with at least 16 GB of memory; the Bayesian hyperparameter search (40 trials) and the four ablation studies reported in Section 4 together account for the bulk of the total compute budget, since each ablation condition requires a full retraining run under the patient-level train/validation/test split described in Section 3. The Kaggle ImageMask trials were implemented independently in TensorFlow 2.x and Keras, executed inside the Kaggle notebook environment on either CPU-only instances or intermittently available GPU sessions, and were deliberately kept lightweight (five representative epochs per architecture in the results reported here) to reflect the compute budget realistically available to a student or small research team using free-tier cloud infrastructure rather than a dedicated GPU cluster.

Table 15: Consolidated hyperparameter configuration for the multi-dataset UNet++ framework, selected via 40-trial Bayesian search.

Hyperparameter	Value
Encoder backbone	ResNet18 (ImageNet-pretrained, 11.7M params)
Optimizer	AdamW
Weight decay	2×10^{-5}
Learning rate schedule	Cyclical, 1×10^{-6} to 1×10^{-4} (peak 2×10^{-4})
Batch size	16
Precision	Mixed precision (FP16/FP32)
Loss function	Dice + Tversky ($\alpha = 0.3$, $\beta = 0.7$) + deep supervision
Regularization	Batch normalization, spatial dropout (0.3–0.5)
Early stopping	Validation F1, patience = 15 epochs
Augmentation probability	0.6 (spatial 50%, intensity 60%)
Lesion-aware batch sampling	$\geq 70\%$ of batches contain a lesion voxel
Data split	70% train / 10% validation / 20% test (patient-level)

A.2 Algorithmic Summary

Algorithm 1 summarizes the end-to-end training procedure for the multi-dataset UNet++ framework at a level of detail sufficient to reimplement the core loop; dataset-specific preprocessing (registration, skull-stripping, normalization) is treated as a fixed upstream step, consistent with Section 3.

Algorithm 1 Multi-dataset UNet++ training procedure

- 1: **Input:** Pooled, preprocessed T1/T2/FLAIR volumes with consensus lesion masks from MSLesSeg, Consensus Brain MRI, and Open MS_Data
 - 2: Split patients into train (70%), validation (10%), test (20%) sets, no patient overlap across splits
 - 3: Initialize UNet++ with ResNet18 encoder (ImageNet weights), randomly initialized decoder
 - 4: **for** epoch = 1 to $N_{\max}$ **do**
 - 5: **for** each minibatch B sampled with lesion-aware sampling (Section 3) **do**
 - 6: Apply spatial and intensity augmentation to B with probability 0.6
 - 7: Forward pass: compute primary prediction and auxiliary deep-supervision outputs
 - 8: Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Tversky}} + \lambda \sum_s w_s \mathcal{L}_s$
 - 9: Backpropagate and update weights via AdamW with cyclical learning rate
 - 10: **end for**
 - 11: Evaluate on validation split; compute voxel-wise and lesion-wise F1
 - 12: **if** validation F1 has not improved for 15 epochs **then**
 - 13: **break** (early stopping)
 - 14: **end if**
 - 15: **end for**
 - 16: Evaluate final model on held-out test split: voxel-wise and lesion-wise metrics, region-stratified analysis
 - 17: **Output:** Trained UNet++ weights, test-set performance report
-

A.3 Cohort Characteristics

Table 16 expands on the summary cohort statistics given in Section 3, reporting them per source dataset rather than only pooled, which is useful for readers assessing whether the pooled cohort’s demographic profile is representative of their own patient population.

Table 16: Cohort characteristics by source dataset (mean $\pm$ SD unless noted).

Characteristic	MSLesSeg	Consensus Brain MRI	Open MS_Data	Pooled Cohort
Patients	75	60	30	165
Scans	115	60	30	295
Mean age (years)	37 ± 10.3	47 ± 11.2	49 ± 9.8	45 ± 11.5
Female / male	48 / 27	38 / 22	19 / 11	105 / 60
Field strength	1.5T	3.0T	3.0T	Mixed
Mean EDSS	1.8 ± 1.1	2.1 ± 1.3	2.2 ± 1.4	≈ 2.0
Mean disease dura- tion (years)	6.5 ± 4.8	8.7 ± 5.9	9.1 ± 6.2	≈ 8.0

A.4 Ethical, Regulatory, and Cost Considerations

Moving a framework of this kind from validated research prototype to routine clinical use raises questions that go beyond segmentation accuracy. All three source datasets used here are publicly released, de-identified research cohorts distributed under permissive research licenses, which is what makes this study reproducible but also means the model has not yet been exposed to the full data-governance and consent requirements that would apply to a clinical deployment on a hospital’s own patient population; any deployment pathway would need institutional review, a data use agreement appropriate to the deploying site, and, in most jurisdictions, regulatory clearance as software as a medical device before the tool could influence clinical decisions rather than serve as a research or triage aid. Because the framework is intended as decision support rather than an autonomous diagnostic system, the appropriate regulatory framing is closer to a computer-aided detection/triage tool than to an autonomous diagnostic algorithm, which is a materially different (and generally less burdensome) regulatory pathway, but one that still requires a human radiologist to review and sign off on every output. Finally, algorithmic fairness deserves explicit mention: the pooled cohort used here draws from Italian and broader European clinical populations, and nothing in the current study establishes how well the model generalizes to patient populations with different demographic compositions, comorbidity profiles, or MS subtypes; this is a specific instance of the broader domain-shift concern raised in Section 2 and reinforces why prospective, multi-population validation is listed as the top priority in Section 5 rather than as an optional refinement.

Declarations

Ethics approval and consent to participate

Not applicable. This study used secondary analysis of publicly available de-identified MRI datasets and involved no new participant recruitment, intervention, or collection of identifiable personal information.

Consent for publication

Not applicable. No identifiable participant information or identifiable patient images are presented.

Availability of data and materials

The imaging resources analyzed in this study are publicly accessible from their respective providers. MSLesSeg is available through Figshare at <https://doi.org/10.6084/m9.figshare.27919209>. The multi-rater consensus MS brain MRI dataset described by Lesjak et al. is available through the Laboratory of Imaging Technologies resource portal at <https://lit.fe.uni-lj.si/en/research/resources/3D-MR-MS/>. The Open MS Data resource is available at https://johnmuschelli.com/open_ms_data/. No newly collected patient data were generated in this study.

Competing interests

The authors declare that they have no competing interests.

Funding

No funds, grants, or other external support were received for this work.

Authors' contributions

Mahmoud Elsayed performed the conceptualization, methodology, data preparation, technical implementation, model development, experiments, analysis, visualization, and original manuscript preparation. Mostafa Elhoshy, Nahla Bishri, and Ahmed I. Saleh contributed to supervision, result assessment and interpretation, validation, critical review, and manuscript revision. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Use of generative AI and AI-assisted technologies

OpenAI ChatGPT was used during manuscript preparation to assist with language refinement, organization, and formatting. The authors reviewed and verified the scientific content, analyses, interpretations, and references and take full responsibility for the final manuscript.

References

- [1] Szekely-Kohn, A.C., Castellani, M., Baronti, L., *et al.*: Reliability of a convolutional neural network in segmenting multiple sclerosis lesions from MRI: Impact of data augmentation, image modality and tolerance with U-Net architecture. *PLOS Digital Health* **5**(4), 0001316 (2026)
- [2] Katona, M., Bozsik, B., Bodnár, P.B., *et al.*: DenseLes: Slice-wise dense network for multiple sclerosis lesion segmentation and classification. *Frontiers in Neurology* **17**, 1704317 (2026)
- [3] Zhang, J., Zuo, L., Dewey, B.E., *et al.*: UNISELF: A unified network with instance normalization and self-ensembled lesion fusion for multiple sclerosis lesion segmentation. *Medical Image Analysis* **109**, 103954 (2026)
- [4] Thabet, R.M., Shedeed, H.A., Al-Berry, M., *et al.*: A multi-sequence MRI integration framework using SwinUNETR-v2 for multiple sclerosis lesion segmentation. *BMC Medical Informatics and Decision Making* **26**, 195 (2026)

- [5] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis (DLMIA), pp. 3–11. Springer, ??? (2018)
- [6] Dwyer, M.G., Bergsland, N., Bartnik, A., *et al.*: Quantifying cortical lesions in multiple sclerosis MRI datasets using multi-contrast post-processing and deep learning. *Communications Medicine* **6**, 365 (2026)
- [7] Spagnolo, F., Molchanova, N., Bach Cuadra, M., *et al.*: Instance-level quantitative saliency in multiple sclerosis lesion segmentation. *Scientific Reports* **16**, 6917 (2026)
- [8] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI, pp. 234–241 (2015)
- [9] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**(2), 203–211 (2021)
- [10] Hashemi, M., Akhbari, M., Jutten, C.: Delve into multiple sclerosis (MS) lesion exploration: A modified attention U-Net for MS lesion segmentation in brain MRI. *Computers in Biology and Medicine* **145**, 105402 (2022)
- [11] Kamraoui, R.A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J.V., Coupé, P.: DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis* **76**, 102312 (2022)
- [12] Rondinella, A., *et al.*: Boosting multiple sclerosis lesion segmentation through attention mechanism. *Computers in Biology and Medicine* **161**, 107021 (2023)
- [13] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021)
- [14] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3d medical image segmentation. In: Proc. IEEE/CVF WACV, pp. 574–584 (2022)
- [15] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Int. MICCAI Brainlesion Workshop. Lecture Notes in Computer Science, vol. 12962, pp. 272–284. Springer, ??? (2022)
- [16] Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D.: DS-TransUNet: Dual swin transformer U-Net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–15 (2022)

- [17] Wang, H., Cao, P., Wang, J., Zaiane, O.R.: UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In: Proc. AAAI, vol. 36, pp. 2441–2449 (2022)
- [18] Raab, F., Wein, S., Greenlee, M., Malloni, W., Lang, E.: A multimodal 2D convolutional neural network for multiple sclerosis lesion detection (2022)
- [19] Liu, D., Cabezas, M., Wang, D., Tang, Z., Bai, L., Zhan, G., *et al.*: Multiple sclerosis lesion segmentation: Revisiting weighting mechanisms for federated learning. *Frontiers in Neuroscience* **17**, 1167612 (2023)
- [20] Bai, L., Wang, D., Wang, H., Barnett, M., Cabezas, M., Cai, W., *et al.*: Improving multiple sclerosis lesion segmentation across clinical sites: A federated learning approach with noise-resilient training. *Artificial Intelligence in Medicine* **152**, 102872 (2024)
- [21] Hindawi, S., Szubstarski, B., Boernert, E., Tackenberg, B., Wuerfel, J.: Federated learning for lesion segmentation in multiple sclerosis: A real-world multi-center feasibility study. *Frontiers in Neurology* **16** (2025)
- [22] Rehman, M.H.U., Lopez Pinaya, W.H., Nachev, P., Teo, J.T., Ourselin, S., Cardoso, M.J.: Federated learning for medical imaging radiology. *British Journal of Radiology* **96**, 20220890 (2023)
- [23] Guan, H., *et al.*: Navigating distribution shifts in medical image analysis: A survey. *arXiv preprint arXiv:2411.05824* (2024)
- [24] Wiltgen, T., McGinnis, J., Schlaeger, S., Kofler, F., Voon, C., Berthele, A., *et al.*: LST-AI: A deep learning ensemble for accurate MS lesion segmentation. *NeuroImage: Clinical* **42**, 103611 (2024)
- [25] Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., Špiclin, Ž.: A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics* **16**(1), 51–63 (2018)
- [26] Guarnera, F., Rondinella, A., *et al.*: MSLesSeg: Baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset. *Scientific Data* (2025)
- [27] Rondinella, A., *et al.*: ICPR 2024 competition on multiple sclerosis lesion segmentation—methods and results. In: *Pattern Recognition. ICPR 2024 Competitions*. Lecture Notes in Computer Science, vol. 15334, pp. 1–16. Springer, ??? (2025)
- [28] Montalban, X., Lebrun-Frénay, C., Oh, J., *et al.*: Diagnosis of multiple sclerosis: 2024 revisions of the McDonald criteria. *Lancet Neurology* **24**(10), 850–865 (2025)

- [29] Ontaneda, D., *et al.*: MRI in the diagnosis of multiple sclerosis: guidance for use of the 2024 McDonald criteria. *Lancet Neurology* **24**(10), 866–879 (2025)
- [30] Evaluation of a deep learning segmentation tool to help detect spinal cord lesions from combined t2 and stir acquisitions in people with multiple sclerosis. *Mult. Scler. J. Exp. Transl. Clin. / OFSEP registry study* (2025)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [splitmaster.py](#)