Comparative Analysis of Deep Learning Architectures for Multiple Sclerosis Lesion Segmentation in Brain MRI: UNet, CDCG-UNet, Swin-UNet

Mert Fidan
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
mertfidan03@gmail.com

Berk Özkan
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
berkozkanjp@gmail.com

M. Shayan Usman
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
mshayanalwaha@gmail.com

Arda Baktır
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
ardabaktir@icloud.com

Emir Kerem Şahin
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
hemirkeremh@gmail.com

Uğur Güdükbay
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
gudukbay@cs.bilkent.edu.tr

***Abstract*—Accurate detection and monitoring of multiple sclerosis (MS) lesions in magnetic resonance imaging (MRI) are critical for diagnosis and treatment planning but remain challenging due to the variability of the lesion and the complexity of volumetric data. Although deep learning has shown promise in automated lesion segmentation, comprehensive comparative evaluations of different architectures in this context are limited. This paper compares standard UNet, CDCG-UNet, and Swin-UNet architectures. These architectures have been used for tumors; however, we apply them to lesions. We evaluate their performance using the F1 score and Intersection over Union (IoU). Experimental results show that Swin-UNet consistently outperforms U-Net and CDCG-UNet across coronal, axial, and sagittal axes. Specifically, Swin-UNet achieved the highest IoU and F1 scores on the combined MSSEG 2016 and Shifts 2.0 dataset in all views, with an IoU of 0.6262 and F1 of 0.6857 in the coronal view, IoU of 0.7015 and F1 of 0.7417 in the axial view, and IoU of 0.5715 and F1 of 0.6368 in the sagittal view.**

I. Introduction

Multiple sclerosis (MS) is a chronic neurological disease characterized by inflammation, demyelination, and neurodegeneration within the central nervous system, significantly impacting patient quality of life [1]. Accurate detection and continuous monitoring of MS lesions through magnetic resonance imaging (MRI) are crucial in diagnosis and treatment planning. However, the interpretation of MRI scans, particularly identifying subtle lesion changes over time, remains challenging due to variability in lesion appearance and the complexity inherent in volumetric data interpretation [2].

Recent advancements in deep learning, specifically in medical image segmentation, have significantly enhanced automated lesion detection capabilities. Various neural network architectures, including UNet and its variants, have demonstrated strong potential for medical segmentation tasks [3]. Nevertheless, comparative assessments of these architectures, particularly in the context of MS lesion detection, remain sparse.

We assess and compare three leading segmentation architectures—standard UNet, CDCG-UNet, and Swin-UNet—based on their effectiveness in accurately identifying MS lesions, using metrics such as the F1 score and Intersection over Union (IoU). We tailor various pre-processing steps to adapt these architectures for lesion segmentation.

The following sections present related work, detailed system architecture, our machine learning methodology, extensive evaluation results, and future directions.

A. Background on Multiple Sclerosis

Multiple sclerosis is an autoimmune disease that predominantly affects adults, with onset typically between 20 and 40 years of age. The disease is characterized by the immune system erroneously attacking the protective myelin sheath surrounding nerve fibers, resulting in inflammation and subsequent neuronal damage. Clinically, MS presents through various symptoms, including motor impairment, cognitive deficits, sensory disturbances, fatigue, and visual problems, significantly impacting daily functioning and quality of life [1].

MS lesions, identifiable via MRI, are focal areas of damage predominantly found in the white matter, though they can also occur in gray matter and spinal cord tissue. The appearance and evolution of these lesions are central to diagnosing MS and monitoring disease progression and treatment efficacy [3].

Typical MRI biomarkers for MS include lesion number, lesion volume, and patterns of lesion distribution.

Given the heterogeneous nature of MS, early and precise identification of lesion changes is essential for personalized patient management. Traditional methods of lesion assessment involve manual delineation by radiologists, a process that is both time-consuming and subject to significant intra- and inter-observer variability [3]. This situation highlights the urgent need for automated, accurate, and consistent segmentation tools to assist lesion tracking.

B. Objectives and Aims

We present a comprehensive comparative analysis of leading segmentation architectures—standard UNet, CDCG-UNet, and Swin-UNet—emphasizing the clinical advantages and performance benefits of Swin-UNet in MS lesion tracking. Our findings underscore the model's potential to enhance precision in clinical decision-making, promote interdisciplinary collaboration, and improve patient outcomes.

We apply Swin-UNet to MS lesion detection for the first time, demonstrating its superior performance through extensive validation using standard medical segmentation metrics such as the F1 score and Intersection over Union (IoU).

II. Related Work

A. Medical Image Segmentation

Medical image segmentation has become a cornerstone of modern diagnostic and treatment planning workflows, enabling precise delineation of anatomical structures and pathological regions within volumetric data. Traditional image processing techniques—such as thresholding, region growing, and edge detection—have seen limited success in handling the variability and complexity of medical images, especially when dealing with subtle or diffuse lesions common in diseases like Multiple Sclerosis (MS) [4].

With the advent of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm for medical image segmentation. Among these, the UNet architecture introduced by Ronneberger et al. [4] has emerged as a foundational model due to its encoder-decoder structure and effective skip connections, allowing for precise localization and semantic context fusion. UNet and its derivatives have demonstrated strong performance across various segmentation tasks, including brain tumors, retinal vessels, and white matter lesions.

Numerous architectural variants have been proposed to enhance segmentation accuracy and generalizability. CDCG-UNet incorporates dense and contextual convolutions to improve feature extraction, particularly in cases of low-contrast boundaries. More recently, transformer-based models such as Swin-UNet have introduced self-attention mechanisms into the segmentation pipeline. This development allows the model to capture long-range dependencies and global contextual information more effectively [5]. These advancements have shown promise in improving lesion detection sensitivity and reducing false positives, particularly in complex clinical scenarios like MS lesion segmentation.

B. MS Lesion Detection Tools

Accurate and reproducible detection of Multiple Sclerosis (MS) lesions from MRI scans is critical for diagnosis, monitoring disease progression, and evaluating treatment efficacy. Traditional approaches for lesion identification have relied heavily on manual annotation by expert radiologists, which is time-consuming, subjective, and prone to high inter-observer variability [1]. Over the past two decades, various automated and semi-automated lesion detection tools have been developed to mitigate these limitations.

Classical approaches include tools such as the Lesion Segmentation Toolbox (LST) for SPM [6] and FSL's lesion growth algorithm (LGA) [7], which utilizes probabilistic models and intensity thresholding techniques. While these methods offer some level of automation, they are often sensitive to scanner variability, require extensive parameter tuning, and lack robustness across different patient populations and lesion types [8].

III. Machine Learning Methodology

A. Dataset and Preprocessing

1) Datasets:

a) Shifts 2.0: The Shifts 2.0 dataset extends the original Shifts benchmark by including data from a high-risk domain: 3D white matter Multiple Sclerosis(MS) lesion segmentation in magnetic resonance imaging(MRI) scans. In MS lesion segmentation, the dataset combines publicly available components, such as ISBI and PubMRI, with unreleased data from the University of Lausanne [9]. Including these diverse sources introduces significant distributional shifts that reflect variations in scanner types, acquisition protocols, and inter-rater annotation guidelines. These shifts make the dataset particularly challenging and well-suited for evaluating predictive performance and uncertainty estimation.

b) MSSEG 2016: The MSSEG 2016 dataset [10], used extensively for multiple sclerosis lesion segmentation challenges, comprises unprocessed and preprocessed MRI data collected during the MICCAI 2016 challenge. The dataset includes imaging data from 15 patients for training and 38 for testing. Patients originate from four distinct MRI scanners at different centers: those with IDs beginning with "01" are acquired using a Siemens 3T Verio scanner in Rennes; those starting with "03" come from a GE Discovery 3T scanner in Bordeaux; patients beginning with "07" are from a Siemens Aera 1.5T scanner in Lyon; and those with "08" come from a Philips Ingenia 3T scanner in Lyon. For each patient, the dataset provides multiple MRI modalities, including 3D FLAIR, 3D T1, T2, DP, and 3D T1 Gd images, along with seven manual segmentations from clinical experts and a consensus segmentation computed using the LOP STAPLE algorithm.

2) Preprocessing Steps:

a) Brain Area Extraction (Skull-Stripping): Brain area extraction is the first step in our pre-processing pipeline, commonly known as skull-stripping. This process removes non-brain tissues such as the skull, scalp, and other extraneous structures from the MRI scans using U-Net and ANTs-based training data with FLAIR volume [11], [7]. By isolating the brain region, the segmentation model is exposed only to the relevant anatomical structures, significantly reducing distractions and minimizing the risk of false positives. In turn, this focused input improves the network's ability to delineate MS lesions accurately.

b) Denoising with Anisotropic Diffusion: To enhance image quality, we apply anisotropic diffusion using MedPy's implementation for denoising, using four iterations with a kappa value of 30 [12]. This technique smooths homogeneous regions while preserving important edge details critical for identifying lesion boundaries. The resulting improvement in the signal-to-noise ratio ensures that subtle features remain intact, allowing the segmentation model to extract more reliable and robust features from the images.

c) Intensity Rescaling with Li Thresholding: Next, we perform intensity rescaling using Li thresholding, an adaptive method that calculates optimal intensity thresholds for the images. This step enhances the contrast between lesions and normal brain tissue by standardizing the dynamic range across scans. Improved contrast is essential for enabling the model to detect and differentiate lesion areas, especially when these lesions present with subtle intensity differences compared to the surrounding tissues.

d) Bias Field Correction Using N4: MRI scans often suffer from intensity inhomogeneities due to scanner imperfections and coil sensitivity variations. To address this issue, we employ N4 bias field correction from SimpleITK [13], which corrects these inconsistencies and produces a uniform intensity distribution throughout each scan. This uniformity is crucial as it reduces variability and prevents the model from being misled by scanner-induced artifacts, thereby enhancing the overall accuracy of the segmentation.

e) Intensity Normalization to a 0–255 Scale: Following bias correction, we normalize the image intensities to a standardized range of 0–255. This normalization ensures that all scans have a consistent dynamic range, which is beneficial for training stability and convergence. Consistent input intensity profiles allow the model to learn discriminative features more effectively, reducing variability that could impact segmentation performance.

f) Standardizing MRI Orientation to LPS: To further enhance data consistency, we standardize the orientation of all MRI scans to the Left-Posterior-Superior (LPS) coordinate system by utilizing the orientation algorithm of SimpleITK's "DICOMOrient" method [14]. This reorientation ensures that anatomical landmarks are consistently positioned across all images, simplifying the segmentation model's task by reducing spatial variability. Uniform orientation is critical when integrating multi-axial views, as it guarantees that corresponding anatomical features are aligned across different slices.

g) Resizing MRI Slices to 512×512 Pixels: Each MRI slice is resized to a resolution of 512×512 pixels to ensure uniform input dimensions. Consistent slice dimensions are critical for efficient batch processing and ensuring that anatomical structures, including lesions, are represented at a similar scale across the dataset. It is fundamental for maintaining the integrity of spatial features during model training. This resizing step, implemented via SciPy's "zoom" method, preserves spatial features while ensuring that scaling is handled correctly across all images.

h) Extraction of Multi-Axial MRI Slices: Finally, the preprocessed 3D MRI volumes are decomposed into 2D slices along the coronal, axial, and sagittal axes. This multi-axial extraction captures complementary brain views, allowing the segmentation model to leverage diverse spatial perspectives. The individual predictions from each plane are later fused using a majority voting scheme, where a voxel is marked as a lesion if at least two of the three axes indicate a lesion. This fusion process enhances the reliability and accuracy of the final 3D lesion mask. This step utilizes NumPy[15] for efficient array manipulation and ensures that slice rotations are standardized for consistent integration of multi-view information. Figure 1 shows sample MRI slices at different stages of our pre-processing pipeline. These visual examples demonstrate how each pre-processing step contributes to enhancing image quality and standardization, which is crucial for consistent model performance across diverse MRI acquisitions.

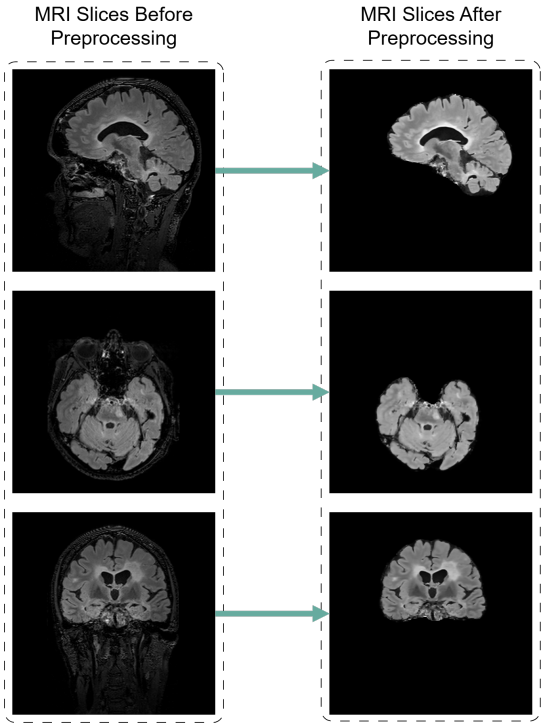

Fig. 1. Example MRI slices during preprocessing

IV. Model Architectures

This study evaluates three deep-learning architectures for MS lesion segmentation: UNet, CDCG-UNet, and Swin-UNet. Each architecture employs different approaches to medical image segmentation with specific mechanisms for feature extraction and contextual understanding.

A. UNet

The U-Net architecture, introduced by Ronneberger et al. [16], is a fully convolutional network designed initially for biomedical image segmentation. Its initial successes were in tasks like segmenting neuronal structures in electron microscopy and cell images, where it achieved superior performance with minimal training data. U-Net's design follows an encoder-decoder paradigm, yielding a characteristic U-shaped architecture that enables context capture and precise localization.

U-Net architecture comprises a multi-layer encoder, the down-sampling path, and a symmetric decoder or up-sampling path, interconnected by skip connections that link corresponding layers. The encoder features repeated blocks consisting of two consecutive 3×3 convolutional operations with ReLU activation, succeeded by a 2×2 max pooling operation. This process progressively reduces the spatial resolution and doubles the number of feature channels at each subsequent stage. This contracting pathway effectively captures hierarchical features, from bare edges to advanced semantic information [16].

Conversely, the decoder executes the inverse operation: each decoder stage commences with an up-sampling step, commonly implemented using a 2×2 transposed convolution that reduces the feature channel count by half. Subsequently, this up-sampled feature map is concatenated with the corresponding encoder-derived feature map of equivalent spatial dimensions via skip connections. Post-concatenation, two additional 3×3 convolutional operations with ReLU activations are employed to integrate the combined features. These skip connections are crucial for retrieving fine-grained spatial details lost during the encoder's down-sampling process, enabling precise structural delineation [16].

Ultimately, a final 1×1 convolutional layer at the decoder's output produces the segmentation mask by converting each feature vector, which consists of 64 channels in the original implementation, into a probability distribution across output classes. The initial U-Net implementation incorporated four down-sampling levels, culminating in 23 convolutional layers, and demonstrated notable efficiency in seamless tile-based predictions and rapid inference for images sized 512×512 pixels [16].

B. CDCG-UNet

CDCG-UNet (Chaotic optimization assisted Dilated Channel Gate U-Net) is a recently introduced variant of the U-Net architecture proposed by Bhagyalaxmi and Dwarakanath (2025) for 3D brain lesion segmentation [17]. Initially developed within the context of brain tumor MRI segmentation. It was explicitly targeting the BraTS challenge. CDCG-UNet enhances the standard U-Net by integrating attention mechanisms and multi-scale context modules. The architecture emphasizes two primary innovations. The first one is the Dilated Channel Gate (DCG) attention block. It leverages dilated convolutions and channel-wise gating mechanisms to refine feature representations. The second one is a chaotic optimization strategy implemented during training to optimize the learning dynamics of the model [17].

Collectively, these innovations aim to improve the accuracy of segmentation results for complex lesion structures while maintaining reasonable model complexity. At its core, CDCG-UNet follows the encoder-decoder structure typical of the original U-Net. It consists of sequential down-sampling convolutional layers (encoder) and corresponding up-sampling layers (decoder), interconnected by skip connections. However, CDCG-UNet distinguishes itself by incorporating Dilated Channel Gate (DCG) attention modules into the standard feature extraction pipeline.

Each DCG attention module enhances feature extraction through two key processes: dilated convolutions and channel-wise gating. The dilated convolutions expand the receptive field of the network, enabling it to capture a broader spatial context around each voxel without reducing spatial resolution through pooling. This property is particularly valuable for detecting lesions and their contextual surroundings at multiple spatial scales [18].

Following dilated convolutions, the channel gating operation applies channel-specific attention to the feature maps. Typically implemented through global pooling or 1×1 convolutions, this mechanism generates attention weights for each channel, emphasizing informative features and suppressing less relevant or noisy channels, similar to squeeze-and-excitation blocks [18]. By selectively highlighting the most important feature channels, the network can better distinguish lesion-related features from confounding structures.

DCG attention blocks are strategically placed within CDCG-UNet, either at the outputs of encoder stages or along the skip connections. This design allows the network to dynamically adapt and prioritize lesion-specific features at multiple hierarchical levels. Furthermore, dilated filters detect subtle lesions by effectively capturing broader contextual information without compromising resolution [18].

Beyond these attention mechanisms, CDCG-UNet maintains the standard U-Net framework, including convolutional layers, pooling, up-convolutions, and concatenation operations, now extended into three dimensions for volumetric imaging data. Thus, CDCG-UNet can be considered an enhanced 3D U-Net architecture with adaptive, channel-wise attention guided by multi-scale contextual cues.

The term "chaotic" in CDCG-UNet refers to the training strategy rather than the architecture itself. Bhagyalaxmi et al. [17] employed a Chaotic Harris Hawk optimization, an evolutionary algorithm to optimize both the segmentation loss function and network parameters during training. This approach resulted in notably high segmentation accuracy, achieving Dice scores in the range of approximately 0.98–0.99

on the BraTS brain tumor segmentation datasets, underscoring the effectiveness of both the architectural enhancements and the chaotic optimization method.

C. Swin-UNet

Swin-UNet, proposed by Cao et al. [19], is a U-Net-like architecture that replaces conventional convolutional layers with Transformer-based modules, making it one of the first pure Transformer segmentation models in medical imaging. The model was introduced to exploit the long-range self-attention capabilities of Transformers for segmentation, addressing the limitation of CNNs, which can miss global context due to their localized receptive fields. Swin-UNet builds on the Swin Transformer, a hierarchical Vision Transformer that uses shifted window attention and adapts it into an encoder-decoder with skip connections analogous to U-Net [20]. In their original work, Cao et al. [19] demonstrated Swin-UNet on multi-organ and cardiac segmentation tasks, achieving accuracy on par or better than CNN-based networks. This showed that a convolution-free approach could successfully perform dense segmentation, which leverages Transformers' strength in modeling global relationships.

The Swin-UNet architecture retains the standard structure of U-Net, consisting of a downsampling encoder, a bottleneck, and an upsampling decoder with lateral skip connections. However, instead of traditional convolutional layers, each component is implemented using Transformer blocks [19].

Initially, the input image is partitioned into small patches, typically of size 4×4 pixels. These patches are flattened and embedded linearly, creating an initial sequence of tokens fed into the Transformer encoder. The encoder is structured hierarchically using Swin Transformer blocks, grouped into multiple stages. At each stage, tokens are organized into local windows for computational efficiency, and self-attention operations are applied. Subsequently, patch merging reduces the resolution of the token representation, similar to pooling layers in convolutional networks, creating increasingly coarse-scale features [19].

At the bottleneck stage, the lowest-resolution level, a Transformer block captures global context across the image, using a shifted-window mechanism that enables interaction between windows at deeper layers [19]. This global attention is valuable in medical imaging applications, where correlations across distant image regions (e.g., lesions appearing in different locations) may exist.

The decoder portion of Swin-UNet reverses this downsampling process. It employs learnable patch-expanding layers to incrementally increase the spatial resolution of the tokens, effectively reversing the encoder's patch merging operation by splitting tokens and reducing the channel dimension. After each upsampling step, decoder tokens are concatenated with corresponding high-resolution encoder tokens through skip connections, preserving and injecting local detail into the decoding stages. Transformer blocks further process these combined representations [19].

The Swin-UNet relies exclusively on self-attention and multi-layer perceptron (MLP) layers rather than convolutions. At lower resolutions, self-attention is constrained to local windows to manage computational complexity effectively. As resolution increases during decoding, the attention mechanism dynamically adapts, allowing the network to recover and refine detailed local structures progressively [19]. Finally, a linear projection layer maps the output token embeddings to a segmentation map. By integrating global contextual understanding and precise local details through Transformer-based attention and skip connections, Swin-UNet effectively addresses dense prediction challenges inherent in medical image segmentation tasks. The architecture's performance has been demonstrated to meet or surpass state-of-the-art convolutional neural network-based models, establishing Transformers as viable and competitive tools in medical imaging segmentation [19].

V. Training & Hyperparameters

A. Data Preparation and Partitioning

The data used in this study were obtained from volumetric MRI scans, primarily focusing on Fluid-Attenuated Inversion Recovery (FLAIR) [7] sequences due to their high sensitivity in detecting demyelinating lesions. Each 3D MRI volume was divided into two-dimensional slices, each measuring 512×512 pixels. Corresponding ground-truth annotations were provided as binary images of identical dimensions (512×512 pixels) aligned precisely with their respective MRI slices.

The dataset was randomly divided into training, validation, and test sets to improve model robustness and reduce the risk of overfitting. An 80:20 split between training and validation data was typically used for experiments conducted using TensorFlow and Keras frameworks. Additionally, an entirely separate test set was reserved to allow unbiased evaluation of model performance, facilitating fair and consistent comparisons across different model architectures and loss functions. Data augmentation techniques were implemented during the training phase, including random horizontal and vertical flipping, slight rotations, and intensity normalization. These augmentations were explicitly selected to mimic realistic variations in lesion appearances, enhancing the model's ability to generalize effectively across various clinical conditions.

B. Model Configuration and Loss Functions

Several neural network architectures were implemented and compared in this study, including enhanced two-dimensional U-Net variants developed using TensorFlow/Keras, an attention-based Channel-Dilated Convolution and Gating (CDCG-UNet), and a transformer-based Swin-UNet implemented in PyTorch. Regardless of the chosen architecture or framework, the input was consistently a single-channel image slice, reshaped and batch-loaded for processing.

Training objectives were specifically chosen to address the inherent class imbalance in lesion segmentation, as lesion areas typically represent a small portion of the image. Loss functions employed were primarily designed to optimize overlap accuracy and included Dice-based metrics [21] and Tversky and

focal Tversky variants [22]. Dice loss was selected to maximize spatial overlap precision. Tversky-based loss variants provide flexibility by assigning differential weights to false positives and false negatives, particularly improving sensitivity to small lesion regions. Additionally, learning rate scheduling techniques were applied to improve training stability: TensorFlow/Keras [23], [24] models generally utilized an exponential decay schedule (reducing the learning rate by approximately 10% after fixed intervals), whereas PyTorch [23] implementations either employed a fixed learning rate or a slight decay schedule, depending on initial testing outcomes.

C. Optimization and Training Procedure

We used an NVIDIA A100 Tensor Core GPU with 80GB of memory during training. All models, namely UNet, CDCG-UNet, and Swin-UNet, were implemented using Python 3.9 and PyTorch 2.2.2, ensuring a reproducible and state-of-the-art deep learning environment. We used a batch size 16 with shuffling to promote diverse mini-batches, facilitating robust parameter updates. Each input image was resized to a fixed dimension of 512×512, and patches were extracted at the same resolution to maintain uniform spatial representation across the dataset. For optimization, we utilized the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-7}$, and Amsgrad option is set to False. The learning schedule followed an exponential decay strategy, with decay steps set to 10,000 and a decay rate of 0.9, starting from an initial learning rate of $\epsilon = 3 \times 10^{-5}$. Training was conducted for 100 epochs, a duration chosen to allow sufficient convergence iterations while avoiding overfitting risks. We employed "Dice Loss" as our loss function because it optimizes the overlap between predicted segmentation and ground truth, a critical consideration in medical image segmentation, especially when dealing with class imbalances.

These training parameters were maintained consistently across both experimental conditions: training with only the MSSEG 2016 [25] dataset and training with a combined dataset of MSSEG 2016 and Shifts 2.0 [9]. This ensured that performance differences could be attributed solely to model architecture and data composition rather than variations in the training regimen.

VI. Evaluation

A. Quantitative Comparison of Models

1) Evaluation Protocol:

a) Train-Test Split for "Only MSSEG 2016": For the "Only MSSEG 2016" condition, the dataset was divided into 15 patients for training and 38 patients for testing, as originally provided by the challenge. The training and testing cohorts were further characterized by scanner origin.

b) Train-Test Split For "MSSEG 2016 and Shifts 2.0": In the "MSSEG 2016 and Shifts 2.0" condition, the MSSEG 2016 data was augmented with the Shifts 2.0 dataset to form a combined training set incorporating a broader range of acquisition protocols and clinical settings. This merged dataset was partitioned into training and testing subsets using an 80-20 split, with 80% of the samples allocated for training and the remaining 20% reserved for testing. Integrating Shifts 2.0 data introduced additional distributional shifts, resulting from introducing a different scanner model for training and more data from other scanners. This process ensured that the evaluation of model robustness and uncertainty quality was not limited to the controlled settings of the MSSEG 2016 challenge alone.

TABLE I
NUMBER OF TRAINING/TESTING SAMPLES FOR "ONLY MSSEG 2016"

Scanner	Training Scans	Testing Scans
Siemens 3T Verio	5	10
GE Discovery 3T	0	8
Siemens Aera 1.5T	5	10
Philips Ingenia 3T	5	10

TABLE II
NUMBER OF TRAINING/TESTING SAMPLES FOR "MSSEG 2016 AND SHIFTS 2.0"

Scanner	Training Scans	Testing Scans
Siemens 3T Verio	11	12
GE Discovery 3T	4	9
Siemens Aera 1.5T	13	12
Philips Ingenia 3T	13	12

c) Validation: For both training conditions, we reserved 20% of the training data for validation. This validation split was used to fine-tune hyperparameters and monitor model performance during training, ensuring our evaluation metrics were robust and indicative of performance on unseen data.

B. Metrics for Assessment

A suite of metrics, which capture both overall accuracy and the quality of lesion delineation, were selected to thoroughly evaluate the performance of segmentation models. Each metric was chosen for its relevance to medical imaging, where the cost of misclassification can be high, and precise localization of lesions is critical [26].

a) Intersection over Union (IoU): Also known as the Jaccard Index, IoU quantifies the overlap between the predicted segmentation and the ground truth by calculating the ratio of their intersection to their union. This metric is particularly significant for medical imaging applications, as it directly reflects how well the model delineates the actual lesion areas [27]. A higher IoU indicates that the model is adept at capturing the precise boundaries of lesions.

b) Dice Score: The Dice Score, often considered the gold standard for evaluating medical image segmentation, measures the similarity between the predicted and ground truth masks [26], [27], [28]. It is susceptible to overlapping regions and is widely used because of its ability to account for the small size of lesion areas relative to the entire image. This metric directly impacts the clinical usability of the segmentation output by ensuring that even small lesions are detected accurately.

c) Precision: Precision is the ratio of true positives to the total number of positive predictions. In a medical setting, high precision is essential to minimize false positives, which could otherwise lead to unnecessary follow-up procedures or undue patient anxiety. It reflects the model's ability to correctly identify lesion areas without erroneously labeling healthy tissue.

d) Recall: Recall, or sensitivity, measures the ratio of true positives to the total number of actual positive cases. For lesion segmentation, achieving high recall is critical because missing a lesion (a false negative) can have severe consequences, including delayed diagnosis or inadequate treatment. Therefore, recall is a key indicator of the model's effectiveness in detecting all actual lesion areas.

e) F1 Score: The F1 Score, as the harmonic mean of precision and recall, provides a single metric that balances both false positives and false negatives. This balance is critical in clinical applications where over-segmentation and under-segmentation can be problematic [29]. The F1 Score offers a comprehensive measure of segmentation quality, reflecting the overall reliability of the model.

VII. Results

This section presents the accuracy measures of U-Net, CDCG-UNet, and, Swin-UNet.

A. Results for Coronal Axis

TABLE III
CORONAL AXIS RESULTS FOR "ONLY MSSEG 2016"

Model	Accuracy	IoU	Dice	Precision	Recall	F1
U-Net	0.9993	0.4216	0.5859	0.4997	0.8408	0.5895
CDCG-UNet	0.9993	0.4103	0.5728	0.4804	0.8640	0.5737
Swin-UNet	0.9993	0.5658	0.6198	0.6229	0.7957	0.6198

TABLE IV
CORONAL AXIS RESULTS FOR "MSSEG 2016 AND SHIFTS 2.0"

Model	Accuracy	IoU	Dice	Precision	Recall	F1
U-Net	0.9994	0.4378	0.6033	0.5044	0.8623	0.6034
CDCG-UNet	0.9994	0.4706	0.6344	0.5447	0.8654	0.6346
Swin-UNet	0.9994	0.6262	0.6857	0.6915	0.8277	0.6857

The performance metrics presented in Tables III and IV show that across the coronal slices, Swin-UNet achieved the highest IoU (0.6262) and F1 score (0.6857), indicating more substantial overlap and segmentation consistency compared to the other models. Figure 2 illustrates these results visually, clearly demonstrating Swin-UNet's advantage in accurately capturing lesion boundaries, especially in views aligned with frontal brain slices.

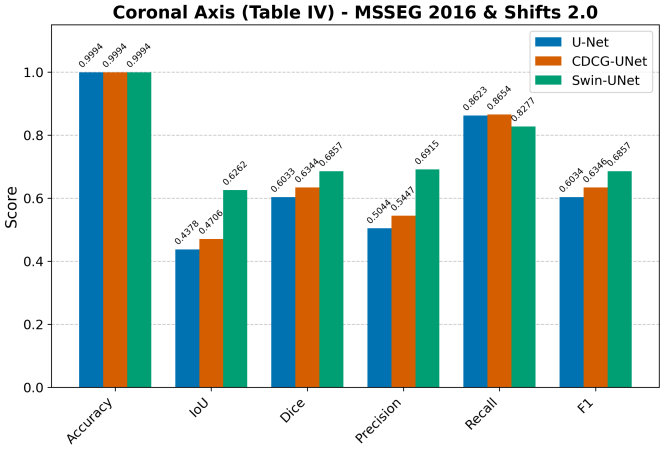

Fig. 2. Coronal axis MSSEG 2016 & Shifts 2.0

B. Results for Axial Axis

TABLE V
AXIAL AXIS RESULTS FOR "ONLY MSSEG 2016"

Model	Accuracy	IoU	Dice	Precision	Recall	F1
U-Net	0.9993	0.3921	0.5459	0.4960	0.8568	0.5481
CDCG-UNet	0.9995	0.4383	0.5897	0.6004	0.7607	0.5899
Swin-UNet	0.9991	0.6072	0.6427	0.6218	0.8808	0.6427

TABLE VI
AXIAL AXIS RESULTS FOR "MSSEG 2016 AND SHIFTS 2.0"

Model	Accuracy	IoU	Dice	Precision	Recall	F1
U-Net	0.9994	0.4191	0.5754	0.5198	0.8734	0.5755
CDCG-UNet	0.9995	0.4509	0.6053	0.5608	0.8716	0.6054
Swin-UNet	0.9994	0.7015	0.7417	0.7427	0.8725	0.7417

The axial view results once again show that Swin-UNet outperforms both U-Net and CDCG-UNet, with IoU (0.7015) and F1 score (0.7417).

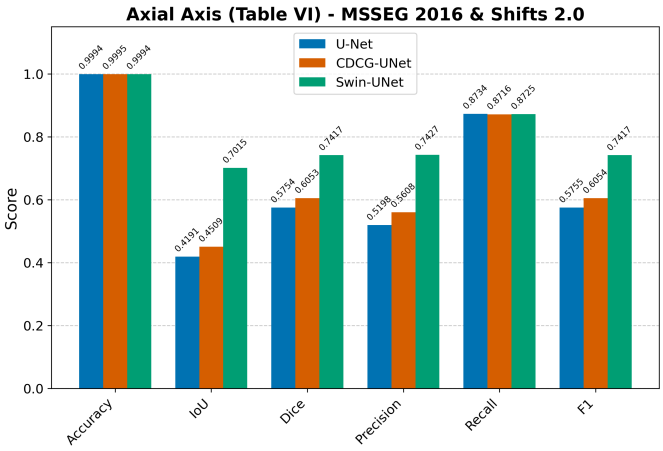

Fig. 3. Axial Axis MSSEG 2016 & Shifts 2.0

TABLE VII
SAGITTAL AXIS RESULTS FOR "ONLY MSSEG 2016"

Model	Accuracy	IoU	Dice	Precision	Recall	F1
U-Net	0.9990	0.4094	0.5698	0.4776	0.8748	0.5699
CDCG-UNet	0.9990	0.3857	0.5452	0.4588	0.8564	0.5473
Swin-UNet	0.9990	0.4932	0.5563	0.5304	0.8154	0.5563

TABLE VIII
SAGITTAL AXIS RESULTS FOR "MSSEG 2016 AND SHIFTS 2.0"

Model	Accuracy	IoU	Dice	Precision	Recall	F1
U-Net	0.9992	0.4436	0.6065	0.5171	0.8732	0.6066
CDCG-UNet	0.9991	0.4426	0.6054	0.5104	0.8858	0.6055
Swin-UNet	0.9993	0.5715	0.6368	0.6491	0.7762	0.6368

C. Results for Sagittal Axis

Although all models performed closely regarding F1 score for the sagittal plane, Swin-UNet is still better (IoU = 0.5715, F1 = 0.6368), indicating its robustness even when lesion shapes are more variable across left-right slices.

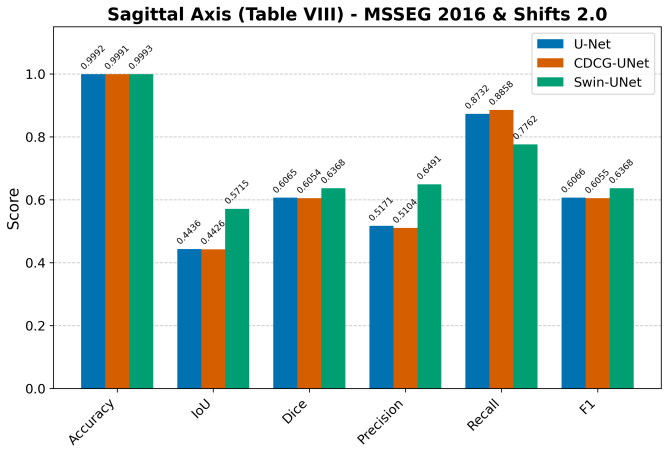

Fig. 4. Sagittal axis MSSEG 2016 & Shifts 2.0

D. Cross-Axis Comparison Using IoU and F1 Scores

To compare the performance of all models across different anatomical views, we analyzed the Intersection over Union (IoU) and F1 scores for each axis—coronal, axial, and sagittal. Figure 5 shows the comparative performance across all axes and models.

The results highlight a consistent trend: Swin-UNet outperforms U-Net and CDCG-UNet across all three axes in both IoU and F1 metrics. This consistency highlights its ability to generalize across anatomical planes and lesion distributions.

- **Coronal axis:** Swin-UNet achieved the highest IoU (0.6262) and F1 score (0.6857), outperforming the next-best model, CDCG-UNet, by over 15% in IoU.
- **Axial axis:** Swin-UNet achieved the strongest performance again, with a remarkable IoU of 0.7015 and an F1 of 0.7417—the highest among all axes and models.
- **Sagittal axis:** Although performance differences were smaller in this view, Swin-UNet led with 0.5715 IoU and 0.6368 F1, showing improved generalization even in noisier contexts.

Overall, Swin-UNet demonstrates the highest segmentation accuracy across all views and excellent stability, validating its strength in capturing both local detail and global context in 3D MRI volumes.

VIII. QUALITATIVE VISUAL RESULTS

We selected representative cases that capture the diversity of lesion appearances, including varying sizes, shapes, and contrasts. These representative cases were chosen based on quantitative performance metrics and visual inspection, ensuring that the selected examples include typical scenarios and challenging cases where lesions are subtle or have irregular boundaries.

For visual comparison, segmentation masks generated by each model were overlaid on the original MRI scans. This overlay technique directly assesses how closely the predicted lesion boundaries match the expert annotations. By presenting these overlays, we can qualitatively compare the strengths and weaknesses of each model.

The overlay visualizations reveal several key insights into the segmentation performance of the models. In the axial view, Swin-UNet accurately delineates lesion boundaries with minimal false positives, outperforming CDCG-UNet and UNet in that order. Although all models capture lesion features in the sagittal plane, Swin-UNet leads in precision, followed by UNet, with CDCG-UNet trailing behind in effectively identifying subtle lesions. Similarly, for the coronal view, Swin-UNet delivers the most accurate and robust segmentation, with CDCG-UNet performing moderately well and UNet showing comparatively lower performance.

In cases with clearly defined lesion boundaries, the Swin-UNet model demonstrates a remarkable dominance in capturing the precise contours of lesions, as evidenced by a prevalence of green overlay areas where the predicted and true labels align. This contrasts with the other models, which in some instances exhibit more red (false negative) regions when compared to Swin-UNet, particularly in areas where lesions are small or exhibit irregular shapes. In challenging cases where lesions are subtle and exhibit low contrast, the predictions of Swin-UNet remain sturdy, indicating fewer omissions and a higher degree of alignment with expert annotations.

IX. DISCUSSION: WHY SWIN-UNET WINS?

The superior performance of Swin-UNet can be attributed to several architectural and methodological innovations. At its core, Swin-UNet adopts a pure Transformer design structured as a U-shaped encoder-decoder with skip connections. Unlike conventional CNNs, the hierarchical Swin Transformer, with its shifted window attention mechanism, enables the model to capture local fine-grained details and long-range global contextual information. This dual capacity is crucial in medical image segmentation, where complex lesion boundaries require

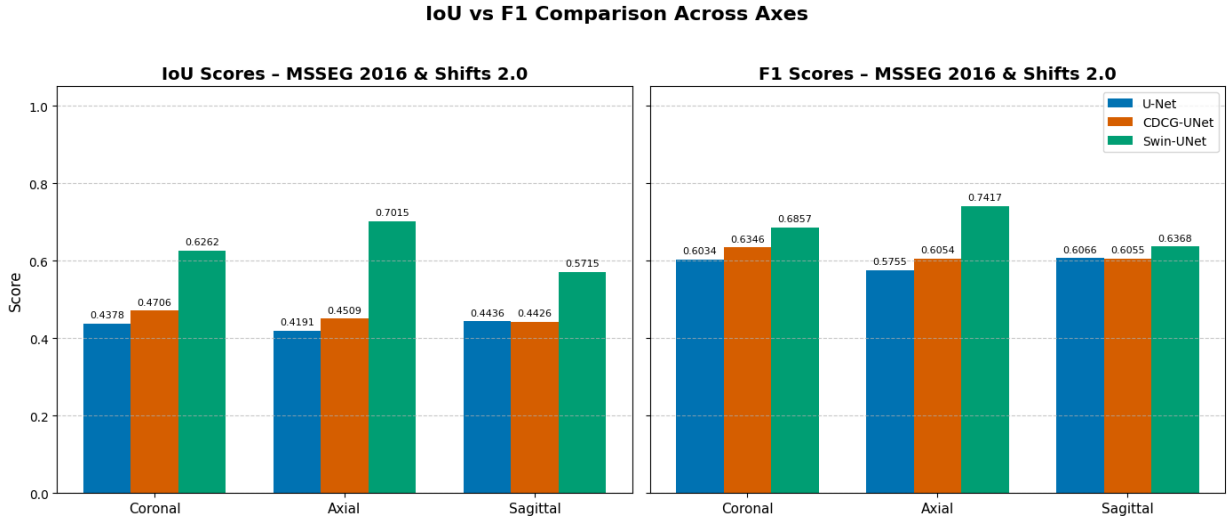

Fig. 5. Bar chart comparing IoU (left) and F1 scores (right) across coronal, axial, and sagittal axes

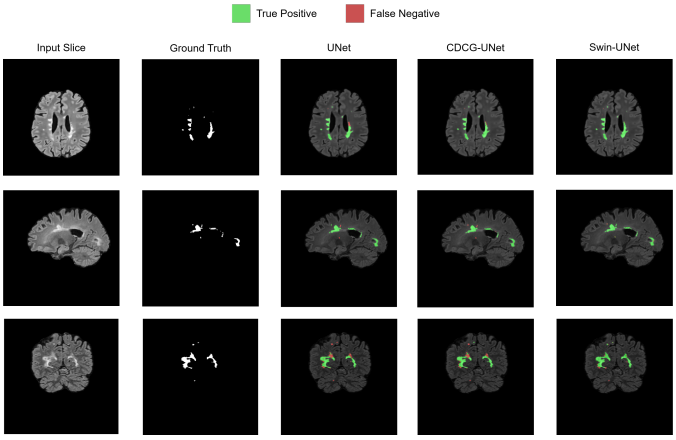

Fig. 6. Example segmentation results

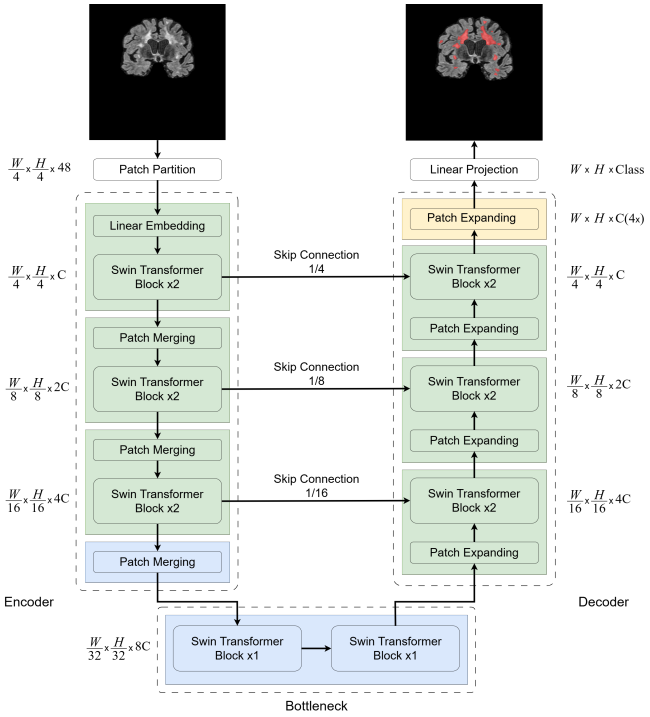

Fig. 7. The architecture of Swin-UNet

precise delineation, and the contextual surroundings help distinguish subtle pathological regions from healthy tissue. The architecture's intrinsic ability to learn feature representations from tokenized image patches underpins its competitive edge over conventional convolution-based approaches [19].

Improved generalization is another integral factor. Swin-UNet benefits from various imaging conditions and patient demographics more efficiently than UNet and CDCG-UNet. This makes the model less sensitive to intensity and structural heterogeneity variations, a common challenge in clinical settings. The attention mechanism further mitigates the effects of these shifts by focusing on the most informative regions of the image, thus enhancing overall generalization and ensuring that the model performs reliably across different clinical settings. For instance, it generalizes better for scans coming from different MRI scanners.

Quantitatively, Swin-UNet consistently achieves higher Dice, IoU, Recall, and F1 scores than UNet and CDCG-UNet. These improvements are directly linked to its advanced ability to model complex and irregular lesion shapes, often poorly handled by traditional convolutional approaches. The pure Transformer approach eliminates the locality constraint inherent in CNNs, allowing for more flexible and effective modeling of global interdependencies within the image [19]. This results in precise lesion detection and reduced false negatives.

From a clinical perspective, the enhanced segmentation quality provided by Swin-UNet translates into more reliable lesion maps, which can directly inform clinical decision-

making processes. Reliable segmentation is vital for monitoring disease progression, planning treatment, and improving patient outcomes [30].

X. Conclusion

This study comprehensively evaluates U-Net, CDCG-UNet, and Swin-UNet for segmenting multiple sclerosis lesions in brain MRI scans across coronal, axial, and sagittal views. Quantitative analyses demonstrate that Swin-UNet consistently achieves superior performance in terms of IoU and F1 score, confirming its capability to generalize across varying anatomical planes and lesion distributions. These results underscore the potential of transformer-based architectures in enhancing the accuracy of medical image segmentation. Future work will extend this analysis to spinal MRI data to assess the generalizability of these models beyond the brain and support broader clinical applications in MS diagnosis and monitoring.

References

[1] W. I. McDonald and M. A. Ron, "Multiple sclerosis: the disease and its manifestations," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 354, pp. 1615–1622, Oct. 1999.

[2] M. P. Wattjes, M. D. Steenwijk, and M. Stangel, "MRI in the diagnosis and monitoring of multiple sclerosis: An update," *Clinical Neuroradiology*, vol. 25, pp. 157–165, Jul. 2015.

[3] M. Hashemi, M. Akhbari, and C. Jutten, "Delve into multiple sclerosis (MS) lesion exploration: A modified attention U-Net for MS lesion segmentation in brain MRI," *Computers in Biology and Medicine*, vol. 145, Jun. 2022. Article no. 105402, 14 pages.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[5] K. Bhagyalaxmi and B. Dwarakanath, "CDCG-UNet: Chaotic optimization assisted brain tumor segmentation based on dilated channel gate attention U-Net model research," *Neuroinformatics*, vol. 23, 2025. Article no. 12, 26 pages.

[6] P. Schmidt and L. Wink, "LST: A lesion segmentation tool for SPM," 2019. Available at https://www.applied-statistics.de/LST_documentation.pdf, Accessed: 2025/04/06.

[7] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, and M. Mühlau, "An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis," *NeuroImage*, vol. 59, no. 4, pp. 3774–3783, 2012.

[8] D. R. van Nederpelt, G. Pontillo, M. Barrantes-Cepas, I. Brouwer, E. M. Strijbis, M. M. Schoonheim, B. Moraal, B. Jasperse, H.-J. M. Mutsaerts, J. Killestein, F. Barkhof, J. P. Kuijer, and H. Vrenken, "Scanner-specific optimisation of automated lesion segmentation in MS," *NeuroImage: Clinical*, vol. 44, 2024. Article no. 103680, 12 pages.

[9] A. Malinin, A. Athanasopoulos, M. Barakovic, M. B. Cuadra, M. J. F. Gales, C. Granziera, M. Graziani, N. Kartashev, K. Kyriakopoulos, P.-J. Lu, N. Molchanova, A. Nikitakis, V. Raina, F. L. Rosa, E. Sivena, V. Tsarsitalidis, E. Tsompopoulou, and E. Volf, "Shifts 2.0: Extending the dataset of real distributional shifts," 2022. Available at https://arxiv.org/abs/2206.15407, Accessed: 2025/04/03.

[10] O. Commowick *et al.*, "Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset," *NeuroImage*, vol. 244, Sep. 2021. Article no. 118589, 8 pages.

[11] N. J. Tustison, P. A. Cook, A. J. Holbrook, H. J. Johnson, J. Muschelli, G. A. Devenyi, J. T. Duda, S. R. Das, N. C. Cullen, D. L. Gillen, M. A. Yassa, J. R. Stone, J. C. Gee, and B. B. Avants, "The ANTsX ecosystem for quantitative biological and medical imaging," *Scientific Reports*, vol. 11, Apr. 2021. Article no. 9068, 13 pages.

[12] O. Maier, "MedPy: Medical image processing in Python, Version 0.5.2." https://loli.github.io/medpy, 2024. [Online].

[13] R. Beare, B. C. Lowekamp, and Z. Yaniv, "Image segmentation, registration and characterization in R with SimpleITK," *Journal of Statistical Software*, vol. 86, no. 8, 2018. DOI: 10.18637/jss.v086.i08, 38 pages.

[14] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research," *Journal of Digital Imaging*, vol. 31, no. 3, pp. 290–303, 2018.

[15] NumPy Developers, "NumPy," 2015. Software available at https://numpy.org, Accessed: 2025/04/06.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI '15, pp. 234–241, 2015.

[17] K. Bhagyalaxmi and D. Bh, "CDCG-UNet: Chaotic optimization assisted brain tumor segmentation based on dilated channel gate attention U-Net model research," *Neuroinformatics*, vol. 23, Jan. 2025. Article no. 12, 26 pages.

[18] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "MISSFormer: An effective transformer for 2D medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1484–1494, 2023.

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 205–218, Feb. 2023.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (Montreal, Canada), pp. 10012–10022, Oct. 2021.

[21] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[22] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[23] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software Available at https://www.tensorflow.org/, Accessed: 2025/04/06.

[24] F. Chollet *et al.*, "Keras," 2015. Software available at https://github.com/fchollet/keras, Accessed: 2025/04/06.

[25] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. Pop, P. Girard, R. Améli, J. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M. Santos, W. Santos, A. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F. Vera-Olmos, N. Malpica, C. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S. Warfield, F. Cotton, and C. Barillot, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific Reports*, vol. 8, pp. 1–17, Sept. 2018. Article no. 13650, 17 pages.

[26] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, Aug. 2015.

[27] N. C. Chung, B. Miasojedow, M. Startek, and A. Gambin, "Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data," *BMC Bioinformatics*, vol. 20, Dec. 2019. Article no. 644, 11 pages.

[28] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, Jul. 1945.

[29] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, pp. 168–192, Aug. 2018.

[30] U. W. Kaunzner and S. A. Gauthier, "MRI in the assessment and monitoring of multiple sclerosis: an update on best practice," *Therapeutic Advances in Neurological Disorders*, vol. 10, pp. 247–261, May 2017.