# Simultaneous Segmentation of Ventricles and Normal/Abnormal White Matter Hyperintensities in Clinical MRI using Deep Learning

**Mahdi Bashiri Bawil**

https://orcid.org/0009-0002-2029-3245

Email: m_bashiri99@sut.ac.ir

Affiliation: Biomedical Engineering Faculty, Sahand University of Technology, Tabriz, Iran

**Mousa Shamsi**[1]

https://orcid.org/0000-0003-4670-0531

Affiliation: Biomedical Engineering Faculty, Sahand University of Technology, Tabriz, Iran

Email: shamsi@sut.ac.ir

**Abolhassan Shakeri Bavil**

https://orcid.org/0000-0001-9397-0484

Email: shakeribavil@yahoo.com

Affiliation: Radiology Department, Tabriz University of Medical Sciences, Tabriz, Iran

[1] Corresponding Author

## Abstract

Multiple sclerosis (MS) diagnosis and monitoring rely heavily on accurate assessment of brain MRI biomarkers, particularly white matter hyperintensities (WMHs) and ventricular changes. Current segmentation approaches suffer from several limitations: they typically segment these structures independently despite their pathophysiological relationship, struggle to differentiate between normal and pathological hyperintensities, and are poorly optimized for anisotropic clinical MRI data. We propose a novel 2D pix2pix-based deep learning framework for simultaneous segmentation of ventricles and WMHs with the unique capability to distinguish between normal periventricular hyperintensities and pathological MS lesions. Our method was developed and validated on FLAIR MRI scans from 300 MS patients. Compared to established methods (SynthSeg, Atlas Matching, BIANCA, LST-LPA, LST-LGA, and WMH-SynthSeg), our approach achieved superior performance for both ventricle segmentation (Dice: 0.801±0.025, HD95: 18.46±7.1mm) and WMH segmentation (Dice: 0.624±0.061, precision: 0.755±0.161). Furthermore, our method successfully differentiated between normal and abnormal hyperintensities with a Dice coefficient of 0.647. Notably, our approach demonstrated exceptional computational efficiency, completing end-to-end processing in approximately 4 seconds per case—up to 36 times faster than baseline methods—while maintaining minimal resource requirements. This combination of improved accuracy, clinically relevant differentiation capability, and computational efficiency addresses critical limitations in current neuroimaging analysis, potentially enabling integration into routine clinical workflows and enhancing MS diagnosis and monitoring.

**Keywords:** Multiple Sclerosis (MS); Deep Learning; Brain MRI Segmentation; White Matter Hyperintensities (WMH); Ventricular Segmentation; Generative Adversarial Networks (GAN); pix2pix; Lesion Differentiation; Computational Efficiency; Clinical Workflow Integration

# 1. Introduction

Multiple Sclerosis (MS) is a chronic, inflammatory, demyelinating disease of the central nervous system affecting approximately 2.8 million people worldwide (Walton et al., 2020). Diagnosis, monitoring, and treatment planning for MS heavily rely on accurate assessment of brain structural changes visible on Magnetic Resonance Imaging (MRI), particularly white matter hyperintensities (WMHs) and ventricular alterations (Tran et al., 2022). These imaging biomarkers are crucial for evaluating disease progression, treatment efficacy, and prognosis.

White matter hyperintensities appear as bright regions on T2-weighted and fluid-attenuated inversion recovery (FLAIR) MRI sequences, representing areas of demyelination, inflammation, and tissue damage (Melazzini et al., 2021). Their spatial distribution, count, and volumetric progression correlate with clinical disability measures and cognitive impairment (Wang et al., 2022). Furthermore, the burden and evolution patterns of these hyperintensities have been associated with both disease progression and therapeutic response (Melazzini et al., 2021). Historical studies have established WMH as independent predictors of cognitive decline and functional outcomes across various neurological conditions (DeCarli et al., 2005). Ventricular enlargement serves as an important indicator of brain atrophy, another hallmark of MS progression (Laso et al., 2023; McKinley et al., 2021).

Current clinical practice relies heavily on qualitative visual assessment by radiologists who manually identify and characterize lesions across multiple MRI slices (Tran et al., 2022). This process is time-consuming, labor-intensive, and subject to significant inter-observer variability, potentially leading to inconsistent diagnosis and treatment decisions (Park et al., 2021). The complex nature of MS pathology requires precise differentiation between pathological hyperintensities and normal appearing white matter signal variations, a distinction challenging even for experienced specialists (Rakić et al., 2021).

Despite advances in neuroimaging analysis, several critical limitations persist in automated segmentation tools for neurodegenerative diseases. First, almost no current approach simultaneously addresses ventricle and WMH segmentation within a unified framework, despite their anatomical and pathophysiological relationship (Atlason et al., 2022). Umirzakova et al., (2025) demonstrated that ventricle morphology and proximity significantly influence the interpretation of periventricular hyperintensities, yet most segmentation approaches treat these structures independently. This disconnected approach fails to capitalize on contextual information that could improve segmentation accuracy.

Second, current methods struggle to differentiate between normal hyperintensities resulting from cerebrospinal fluid (CSF) contamination and pathological lesions. This distinction is crucial for accurate disease quantification, as Griffanti et al. (2018) and Dadar et al (2021) emphasized that misclassification of normal appearing white matter signal variations as lesions can lead to overestimation of disease burden. McKinley et al. (2021) demonstrated that up to 30% of automatically detected hyperintensities in routine clinical scans may represent normal anatomical variants rather than pathology.

Third, the mismatch between research methodologies and clinical data characteristics presents a significant obstacle. Most state-of-the-art segmentation methods assume isotropic 3D volumes with high resolution in all dimensions (Billot et al., 2023; Atlason et al., 2022; La Rosa et al., 2020; Raut et al., 2024; Rondinella et al., 2024; Ulloa-Poblete et al., 2023; Laso et al., 2023; Li et al., 2024; Cathala et al., 2025). However, routine clinical MRI protocols typically produce anisotropic data with significant slice thickness, creating a fundamental incompatibility. Dong et al. (2022) highlighted how this anisotropy presents substantial challenges for conventional 3D segmentation approaches, which may either fail entirely or require computationally intensive preprocessing steps. Recent work by Tran et al. (2022) further demonstrated that inconsistencies in imaging protocols and resolution contribute significantly to variability in WMH quantification, particularly affecting the detection of small lesions and periventricular boundaries.

The computational demands of 3D deep learning models present another practical barrier to clinical implementation. While these models have demonstrated impressive performance in research settings, Tran et al. (2021) found that their deployment in typical clinical environments is constrained by hardware limitations and workflow integration challenges. This computational burden is particularly problematic in resource-limited healthcare settings where advanced computing infrastructure may not be available.

To address these limitations, we propose a novel 2D pix2pix deep learning framework specifically designed for the simultaneous segmentation of ventricles and white matter hyperintensities in anisotropic clinical MRI data. Our approach builds upon recent innovations in generative adversarial networks for medical image segmentation that have demonstrated improved handling of structural complexity and boundary precision in neuroimaging applications (Zeng et al., 2020; Zhang et al., 2022). Building upon our previous work on pix2pix-based gray matter segmentation for WMH categorization (Bawil et al., 2024), our current approach extends this methodology to simultaneously segment multiple brain structures critical for MS evaluation. Our method uniquely distinguishes between normal periventricular hyperintensities and pathological lesions, offering a comprehensive solution compatible with routine clinical workflows. The pix2pix architecture (Isola et al., 2016) serves as the foundation of our approach, modified to accommodate our multi-class segmentation task.

Unlike methods requiring resource-intensive 3D processing, our slice-based approach offers significant advantages for clinical implementation. As highlighted by Hossain et al. (2024) and Hashemi et al. (2022), 2D slice-based methods can achieve comparable accuracy to 3D approaches while dramatically reducing computational requirements, particularly important for handling anisotropic clinical scans. Our proposed model segments four distinct classes: background, ventricles, normal WMH (CSF-contaminated hyperintensities), and abnormal WMH (pathological lesions). This multi-class approach represents a significant advancement over existing tools that typically segment either ventricles or WMH independently.

A key strength of our work is the substantial dataset comprising 300 clinical MRI scans with comprehensive expert annotations of ventricles and both types of WMH. Our model incorporates FLAIR sequence as an input, enabling more accurate differentiation between periventricular CSF-contaminated hyperintensities and true pathological lesions (Griffanti et al., 2016; Zhu et al., 2022). Our method is designed with clinical utility as a primary goal, focusing on compatibility with standard clinical MRI protocols without requiring specialized acquisition parameters or time-consuming preprocessing. This practical approach significantly distinguishes our work from many research-oriented segmentation methods that perform well on research-grade datasets but fail to translate effectively to routine clinical data.

This article continues with detailed descriptions of our dataset and pix2pix segmentation methodology, evaluation strategy, results with quantitative and qualitative analyses, followed by a discussion of clinical implications and limitations.

# 2. Materials and Methods

## 2.1. Study Population

This study analyzed data from 300 MS patients imaged at Golgasht Medical Imaging Center, Tabriz, Iran. The cohort comprised 79 males (aged 18-57 years, mean = 35.8, SD = 9.3) and 221 females (aged 18-68 years, mean = 37.8, SD = 9.7). The study protocol received approval from the Tabriz University of Medical Sciences Research Ethics Committee, and all participants provided written informed consent.

## 2.2. MRI Acquisition Protocol

All imaging was performed on a 1.5-Tesla TOSHIBA Vantage scanner (Canon Medical Systems, Japan). The T2-FLAIR images were acquired with the following parameters: TR = 10000 ms, TE = 100 ms, TI = 2500 ms, flip angle = 90°, FOV = 230 × 230 $mm^2$, acquisition matrix = [0, 256, 192, 0], resulting in non-isotropic data with voxel dimensions of (0.9, 0.9, 6) millimeters. This substantial difference between in-plane resolution and slice thickness informed our decision to implement a 2D rather than 3D segmentation approach.

## 2.3. Ground Truth Development

Manual segmentation was performed on all 300 MS patients' FLAIR images by a neuroradiologist with over 20 years of expertise in neuroimaging. A custom Python-based GUI facilitated the segmentation process. Four distinct classes were delineated: (1) background, (2) ventricles, (3) normal white matter hyperintensities (WMH), and (4) abnormal WMH.

The segmentation methodology followed these steps:

1. Ventricle Segmentation: Initial manual delineation of ventricular boundaries, followed by statistical morphological post-processing for refinement and final manual verification.
2. Abnormal WMH Segmentation: Identification of hyperintense regions representing MS lesions with subsequent processing and expert verification, as same as step 1.

3. Normal WMH Identification: Creating a periventricular boundary by dilating ventricle masks, identifying high-intensity pixels within this boundary, excluding regions overlapping with abnormal WMH, and manual refinement focusing on frontal and occipital horn vicinities.

This approach ensured differentiation between normal hyperintensities (related to CSF contamination) and abnormal hyperintensities (true candidates of MS lesions), with the ventricular system coded in blue, abnormal WMH in red, and normal WMH in green, as shown in Figure 1.

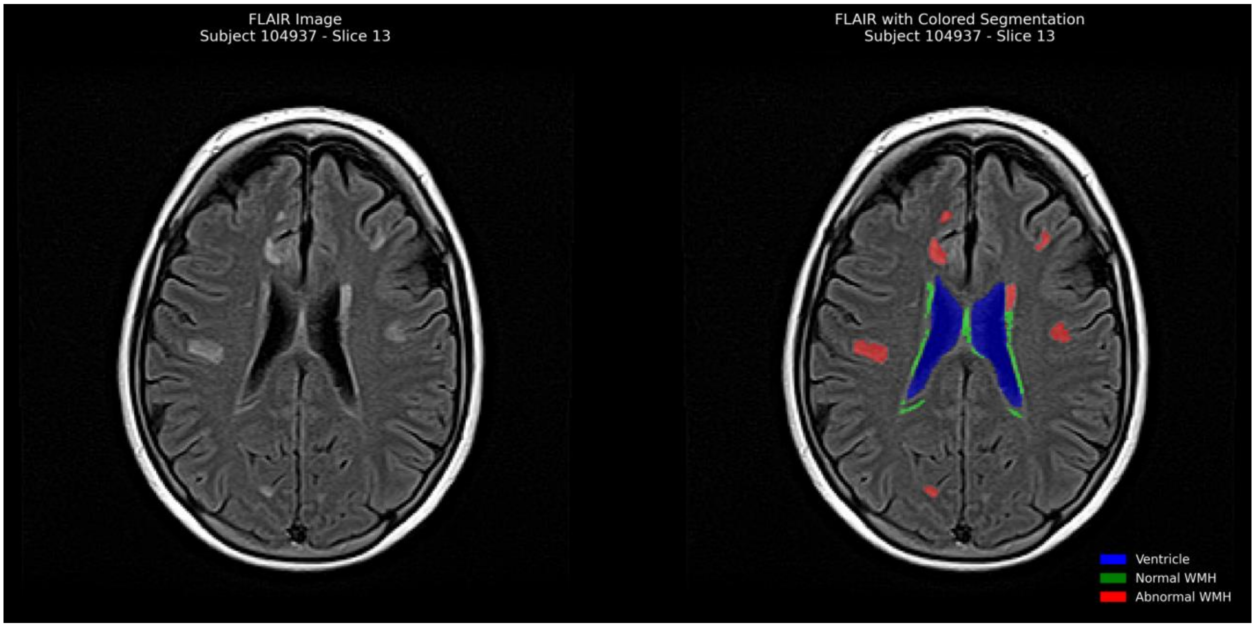
FLAIR Image
Subject 104937 - Slice 13
FLAIR with Colored Segmentation
Subject 104937 - Slice 13
🟦 Ventricle
🟩 Normal WMH
🟥 Abnormal WMH

*Figure 1. A sample image of MRI FLAIR data with corresponding manual segmentation and annotation. The left image is an original image of a MS patient. The left image shows abnormal WMH, normal WMH, and ventricles by red, green, and blue colors, respectively.*

## 2.4. Pre-processing Pipeline

Our pre-processing pipeline forms a critical foundation for successful application of deep learning for brain MRI segmentation, as illustrated in Figure 2. This pipeline standardizes input data, reduces noise, and prepares images for optimal model training and inference.

### 2.4.1. Noise Reduction

MR images inherently contain various noise artifacts that impede effective segmentation. We employ a two-stage noise reduction strategy that applies a median filter (kernel size = 3×3) to suppress salt-and-pepper noise while preserving edges, followed by selective application of a Gaussian filter (σ = 1.0) to smooth remaining artifacts while retaining structural details. We deliberately apply minimal preprocessing, as our deep learning architecture is designed to handle more sophisticated noise patterns, including bias field inhomogeneity, directly during training following evidence that deep convolutional networks can learn to distinguish signal from noise given sufficient training examples (Wu et al., 2024).

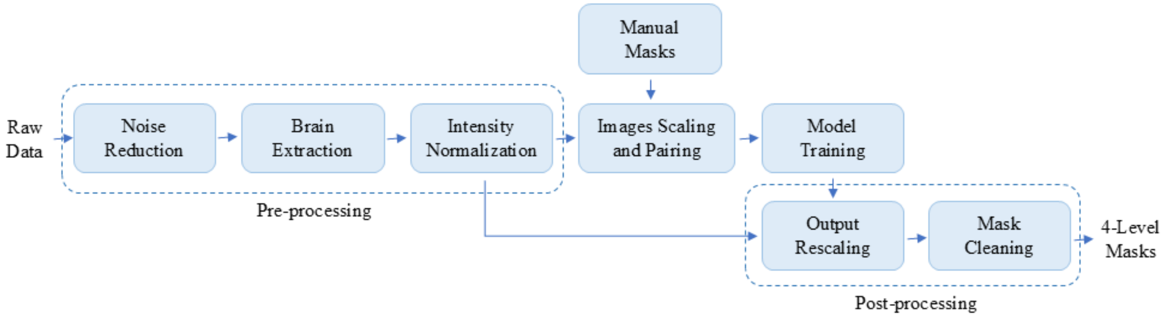

*Figure 2. Block diagram of our proposed method.*

### 2.4.2. Brain Extraction

Our framework employs a morphology-based approach for brain extraction that prioritizes computational efficiency while providing sufficiently accurate boundaries for subsequent processing. The process begins with thresholding based on intensity values to separate brain from background, followed by morphological operations to remove small isolated regions. We then generate an elliptical mask approximating the brain contour using the second-order moments of the binary image, and refine it through contour analysis to better conform to actual brain boundaries. This approach yields a binary brain mask that identifies the general brain region without requiring complex algorithms or excessive computational resources. The resulting mask provides sufficient information for subsequent scaling operations while eliminating most non-brain tissues.

### 2.4.3. Intensity Normalization

We implement a slice-based intensity normalization technique addressing both inter- and intra-patient intensity variations. Unlike conventional global approaches, our method establishes parameters independently for each image slice. The minimum intensity value derives from average background intensity specific to each slice, while the maximum intensity value is determined by analyzing peripheral structures (scalp, skull, and surrounding tissues) obtained by subtracting the brain mask from the full image. This approach ensures intra-patient consistency since background and skull structures remain relatively constant across slices from the same patient, while providing inter-patient adaptability by accommodating anatomical variations. Recent work by Huang et al. (2022) has demonstrated that such adaptive intensity normalization strategies significantly improve segmentation performance, particularly for detecting subtle intensity variations characteristic of MS lesions. Intensity values are scaled to the range [0,1] using Eq. 1, where normalization parameters are derived using our slice-specific methodology.

$$I_{normalized} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (1)$$

### 2.4.4. Paired-image Generation

The final pre-processing step prepares standardized input for our conditional Generative Adversarial Network (cGAN) architecture, the pix2pix model. We scale the brain region to maximize coverage within the image frame using affine transformations derived from the brain mask's geometric properties, then apply identical transformations to the corresponding ground truth segmentation mask to maintain spatial alignment. The processed image and its ground truth mask are horizontally concatenated to create a paired 256×512 pixel composite image compatible with the pix2pix architecture (Isola et al., 2016). This paired-image approach enables the model to learn the mapping between MR images and their segmentation masks within a unified spatial context, while standardized scaling ensures brain structures occupy maximum possible area within the image, reducing the proportion of zero-value pixels that can adversely affect gradient updates during training.

## 2.5. Network Architecture and Implementation

Our method employs a cGAN architecture—specifically the pix2pix model—as the core component for brain segmentation. As illustrated in Figure 2, this model bridges the pre-processing stages and post-processing refinements in our pipeline. The detailed network architecture and implementation strategy are described in the following and visually represented in Figure 3.

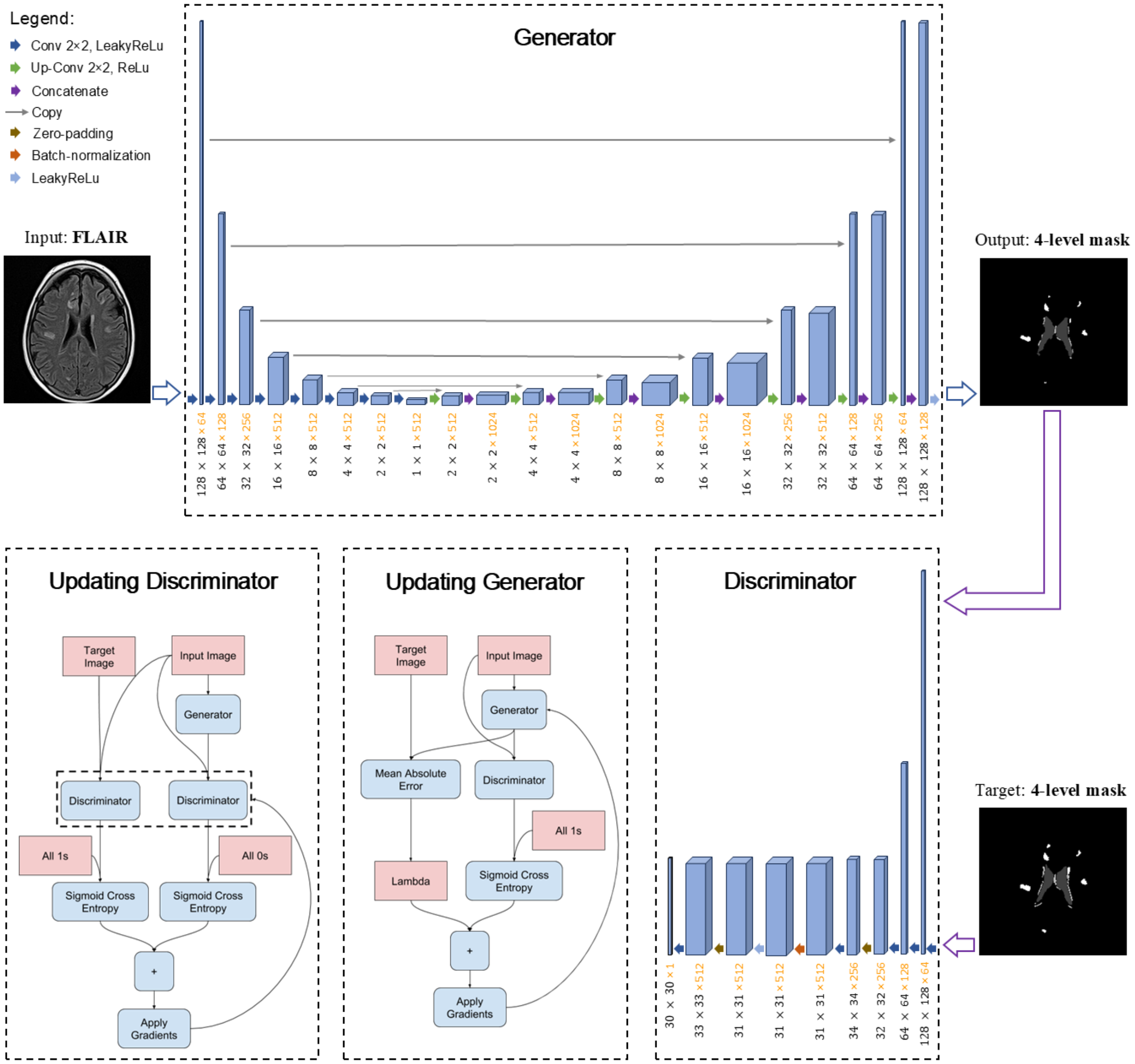

*Figure 3. Pix2pix model architecture overview. This system is trained using paired images consisting of input FLAIR sequences and corresponding 4-level segmentation masks as targets. The updating blocks displayed along the bottom row illustrate the learning progression of the respective network components.*

### 2.5.1. Pix2Pix Framework

We leverage the pix2pix framework for its proven effectiveness in image-to-image translation while preserving structural coherence. The framework consists of two networks: a generator (modified U-Net) that translates MR images into segmentation masks and a discriminator (PatchGAN) that evaluates mask realism. As depicted in Figure 3, the generator receives the pre-processed MR images and produces segmentation masks, while the discriminator simultaneously learns to distinguish between generated masks and ground truth annotations. This adversarial process drives the generator to produce increasingly accurate segmentations that can fool the discriminator.

### 2.5.2. U-Net Backbone

The generator employs a modified U-Net architecture (Ronneberger et al., 2015) with an encoder-decoder structure and skip connections. The encoder comprises eight downsampling blocks (convolutional layer, batch normalization, leaky ReLU), while the decoder includes seven upsampling blocks (transposed convolutions, batch normalization, ReLU, selective dropout). The network maintains a channel depth progression of [64, 128, 256, 512, 512, 512, 512, 512] in the encoder path and the reverse in the decoder, ensuring sufficient capacity for feature capture while maintaining computational efficiency. This architecture design aligns with recent advances in GAN-based medical image segmentation that emphasize the importance of multi-level feature extraction for handling complex anatomical boundaries (Raut et al. 2024; Bawil et al., 2024; Isola et al., 2016; Wu et al., 2024).

### 2.5.3. Training and Testing Strategy

We train our model with a batch size of 1, for 50 epochs, using the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$) with a constant learning rate of 0.0002. The loss function combines adversarial loss (binary cross-entropy) and L1 loss (mean absolute error) defined in Eq. 2.

$$\mathcal{L} = \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{L1} \cdot \mathcal{L}_{L1}\,, \quad where \quad \lambda_{adv} = 1 \ and \ \lambda_{L1} = 100. \tag{2}$$

This combination encourages both structural accuracy and realistic segmentation boundaries. Training is conducted on paired images as described in Section 4.4, with the generator and discriminator updated alternately according to the standard GAN training procedure illustrated in Figure 3.

For robust evaluation, we split the dataset into training (70%) and testing (30%) sets using patient-level stratified sampling to prevent data leakage. We rely on the default parameters of the pix2pix framework, which proved effective without extensive hyperparameter tuning, reducing overfitting risk. No cross-validation is performed as the hyperparameters are tuned beforehand and the test set is sufficiently large to provide reliable performance estimates. The final model is selected based on the lowest validation loss achieved during training.

### 2.5.4. Post-processing

The raw network output undergoes three critical post-processing steps. First, the predicted segmentation masks are transformed back to the original image space using the inverse of the scaling transformations applied during pre-processing. This ensures alignment with the original MR images and enables direct comparison with ground truth annotations. Second, the single-channel grayscale output is converted into four separate probability masks representing our four segmentation classes. This conversion uses a distance-based softmax approach where pixel intensities are compared to predefined class values (0.0, 0.25, 0.75, and 1.0), corresponding to background, ventricles, normal white matter, and white matter hyperintensities, respectively. Third, to improve the quality of segmentation results, we apply a series of morphological operations to the predicted masks such as connected component analysis, hole filling, and edge smoothing. These minimal post-processing steps preserve the network's predictions while removing obvious artifacts, resulting in high-precision segmentation of target brain structures with anatomical plausibility.

## 2.6. Baseline Methods

To evaluate the performance of our proposed approach, we compare it against established methods for both ventricle and hyperintensity segmentation. Since no existing method simultaneously performs four-class segmentation (background, ventricles, normal WMH, abnormal WMH), we conducted separate comparisons for ventricle and hyperintensity segmentation.

### 2.6.1. Ventricle Segmentation Baselines

SynthSeg: We employ the SynthSeg model, a deep learning-based approach trained on synthetic data for robust brain MRI segmentation (Billot et al., 2023). Following the official implementation guidelines, we apply SynthSeg to our FLAIR images and extracted ventricle masks from the full brain parcellation output based on their corresponding label values.

Atlas Matching: We implement an atlas-based approach using the MNI152 standard space template (Fonov et al., 2011). The process involves registering each subject's FLAIR image to MNI152 space using FSL's linear registration tool (FLIRT), applying the ventricle atlas mask, performing basic morphological post-processing to refine boundaries, and finally transforming the resulting segmentation back to the subject's native space.

### 2.6.2. Hyperintensity Segmentation Baselines

BIANCA (Brain Intensity AbNormality Classification Algorithm): We utilize this FSL-based supervised method for WMH segmentation (Griffanti et al., 2016). Following the recommended protocol, BIANCA is trained on our manual segmentations using the same training dataset split as our proposed method to ensure fair comparison.

LST-LPA (Lesion Segmentation Tool - Lesion Prediction Algorithm): This unsupervised lesion segmentation algorithm is implemented under the SPM toolbox (Schmidt et al., 2019). We apply LST-LPA directly to our FLAIR images without parameter modification, as recommended in the official documentation. The output binary masks are obtained in the subject's native space.

LST-LGA (Lesion Segmentation Tool - Lesion Growth Algorithm): This algorithm, also part of the SPM toolbox, requires both FLAIR and T1-weighted images (Schmidt et al., 2012). We pre-register T1 images to the corresponding FLAIR space using FSL's FLIRT command and apply LST-LGA with default parameter settings. Binary output masks are generated in the subject's native space.

WMH-SynthSeg: This extension of the SynthSeg framework specifically addresses white matter hyperintensity segmentation (Laso et al., 2023). Following the official guidelines, we apply

WMH-SynthSeg to our FLAIR images and extract the WMH masks based on the designated label in the output.

### 2.6.3. Adaptation for Fair Comparison

For comparing our approach with baseline methods, several adaptations are necessary. Since our method distinguishes between normal and abnormal WMHs while baseline methods produce only a single WMH class, we use abnormal WMH predictions for direct comparison with the baseline hyperintensity segmentation methods. Similarly, as our model simultaneously segments both ventricles and WMHs, we extract the ventricle class predictions for comparison with ventricle-specific baselines. All evaluations are performed in the subject's native space to avoid registration-induced errors in the quantitative comparisons.

## 2.7. Evaluation Metrics and Analysis

To comprehensively assess the performance of our proposed method and compare it with the baseline methods, we employ multiple complementary evaluation metrics. These metrics are calculated separately for ventricle segmentation and hyperintensity segmentation.

### 2.7.1. Detection-based Metrics

Precision: Also known as positive predictive value, precision measures the proportion of correctly identified positive voxels among all voxels classified as positive:

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

where TP represents true positives and FP represents false positives.

Recall: Also known as sensitivity or true positive rate, recall measures the proportion of actual positive voxels that were correctly identified:

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

where FN represents false negatives.

### 2.7.2. Overlap Metrics

Dice Similarity Coefficient (DSC): The Dice coefficient quantifies the spatial overlap between two segmentations and is defined as:

$$DSC = \frac{2\,|A \cap B|}{|A|+|B|} \tag{5}$$

where $A$ represents the predicted segmentation and $B$ represents the ground truth segmentation. DSC ranges from 0 (no overlap) to 1 (perfect overlap).

Jaccard Index (JI): Also known as the Intersection over Union (IoU), the Jaccard index is calculated as:

$$JI = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

JI also ranges from 0 to 1 and is mathematically related to the Dice coefficient.

### 2.7.3. Boundary Metrics

Hausdorff Distance: While the traditional Hausdorff distance measures the maximum distance between the boundaries of two segmentations, it is highly sensitive to outliers. Therefore, we used the 95th percentile Hausdorff distance (HD95), which is more robust to small segmentation errors:

$$HD95(A,B) = \max\left(h_{95}(A,B),\ h_{95}(B,A)\right) \tag{7}$$

where $h_{95}(A,B)$ is the 95th percentile of the distances from points in boundary $A$ to their closest points in boundary $B$. HD95 is measured in millimeters, with lower values indicating better boundary agreement.

### 2.7.4. Discriminative Metrics

Precision-Recall (PR) Curves: We generate precision-recall curves, which plot precision against recall at different threshold values. PR curve and its area under the curve (AUC-PR), also known as average precision (AP), are particularly informative rather than Receiver Operating Characteristic (ROC) curves for imbalanced datasets such as medical image segmentation where the foreground class is much smaller than the background.

### 2.7.5. Performance Evaluation and Statistical Analysis

Classification Metrics: To evaluate the performance of normal versus abnormal WMH classification, we compute a confusion matrix at the lesion level. This approach allows for assessment of how well the algorithm distinguishes between the two types of hyperintensities as distinct entities. Moreover, to evaluate the performance of abnormal WMH and ventricles versus background classification, we compute another confusion matrix.

Distribution Analysis: For visual comparison of the different methods, we use violin plots to display the distribution of detection-based metrics (precision, recall), overlap metrics (DSC, JI), and boundary metric (HD95) across all test cases. Violin plots combine aspects of box plots and kernel density plots to show both the distribution shape and summary statistics.

## 2.8. Development Environment and Technical Implementation

The deep learning framework was implemented using Python 3.9 with TensorFlow, OpenCV, and scikit-image for image preprocessing, and NumPy and pandas for data handling and statistical analyses. All experiments ran on a workstation with an Intel Core i7-7700K CPU, 64 GB RAM, and NVIDIA RTX 3060 GPU. The codebase with documentation, preprocessing pipelines, model architectures, training scripts, and evaluation modules is available on GitHub (https://github.com/Mahdi-Bashiri/Sim-Vent-WMH-Brain-Seg) under an MIT license, including pre-trained model weights for immediate inference.

# 3. Results

## 3.1. Model Training and Convergence

The cGAN architecture demonstrated efficient training characteristics and robust convergence behavior across all segmentation tasks. Training progression analysis revealed consistent improvement patterns for all three target structures (ventricles, normal WMH, and abnormal WMH).

### 3.1.1. Training Progression

The model exhibited systematic convergence across 50 epochs, with notable performance stabilization observed after the tenth epoch. Figure 4 illustrates the progressive improvement in discriminative capability and reduction in segmentation errors throughout the training process. Specifically, the convergence behavior showed distinctive patterns for each target structure:

Ventricle segmentation: The model rapidly established strong performance for ventricle delineation, achieving an AUC-PR of 0.8 by epoch 10 and exceeding 0.85 by epoch 15. This swift convergence can be attributed to the relatively consistent appearance and well-defined boundaries of ventricular structures across the patient cohort.

Abnormal WMH segmentation: A more gradual convergence trajectory achieved an AUC-PR of 0.4 by epoch 6 and stabilized over 0.6 after epoch 16. This pattern reflects the heterogeneous nature of MS lesions with respect to size, shape, and intensity characteristics.

Normal WMH segmentation: The most challenging of the three targets, convergence for normal WMH segmentation proceeded at a moderate pace, reaching an AUC-PR of 0.35 by epoch 8 and stabilizing near 0.4 after epoch 14. The periventricular localization of normal WMH provided spatial context that facilitated consistent segmentation despite intensity variations.

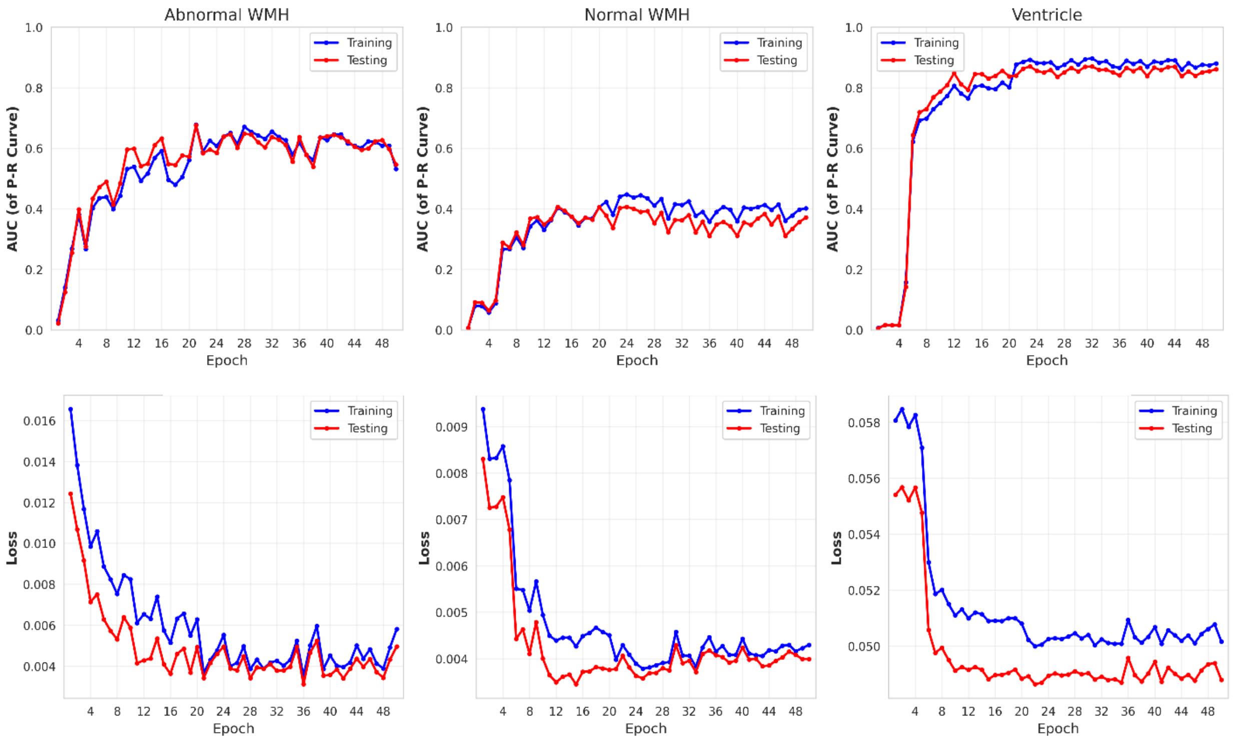

*Figure 4. Model training and performance of the proposed method. The top row shows the AUC-PR values for abnormal WMH, normal WMH, and ventricle segmentation tasks, respectively from left to right. The bottom row displays the corresponding loss values for each of these tasks, also from left to right.*

### 3.1.2. Validation Performance

Quantitative assessment revealed progressive improvement across epochs for all three segmentation targets, as illustrated in Figure 4.

AUC-PR evolution: The first row of Figure 4 demonstrates the progression of AUC-PR values across training epochs for the three classes. The training dynamics show an interesting pattern with training curves positioned below testing curves until approximately epoch 20, after which they cross and remain above the testing curves until the end of training. The three segmentation targets reached different convergence values: ventricles achieved the highest performance at 0.85, abnormal WMH reached 0.6, and normal WMH attained 0.4. This pattern suggests that the model initially undergoes a learning phase where it performs better on unseen data, possibly due to incomplete fitting to training examples, followed by a phase where it begins to fit more closely to the training data. These dynamics provide insight into the model's learning behavior and indicate potential areas for optimization in training duration and regularization strategies.

Loss reduction patterns: The second row of Figure 4 depicts the consistent reduction in loss values across epochs for all three segmentation targets. Both training and testing loss curves exhibit steady decline, with convergence behavior becoming apparent after approximately 20 epochs. The testing loss values converged to different levels for each segmentation target: abnormal WMH reached 0.005, normal WMH achieved 0.004, and ventricles showed a higher final loss at 0.048. This variation in convergence values indicates differential performance across the target structures. The loss reduction trajectories, while showing different final values, still demonstrate the model's capacity to learn each target structure, complementing the insights gained from the AUC-PR analysis.

Comprehensive analysis of validation metrics identified epoch 19 as providing optimal model performance, with peak AUC-PR values of 0.57, 0.4, and 0.85 for abnormal WMH, normal WMH, and ventricle segmentation, respectively. This epoch was selected for final model evaluation on the test dataset.

The performance metrics observed during training and validation indicate that the utilized cGAN architecture successfully learned to differentiate between background, ventricles, normal WMH, and abnormal WMH, establishing effective segmentation capabilities for all target structures. The consistent patterns of improvement across epochs and the stability achieved after epoch 20 demonstrate the model's robust convergence properties.

## 3.2. Segmentation Performance: Qualitative Analysis

### 3.2.1. Visual Representation of Segmentation Results

Figure 5 presents representative axial FLAIR slices with overlaid segmentation results, using a consistent color scheme of true positives (green), false positives (red), and false negatives (yellow). Our method demonstrated precise boundary delineation for ventricles and enhanced detection of both small and large WMH lesions. Error visualization revealed that false positives predominantly occurred at peripheral boundaries, while false negatives were primarily located in regions with lower contrast or partial volume effects.

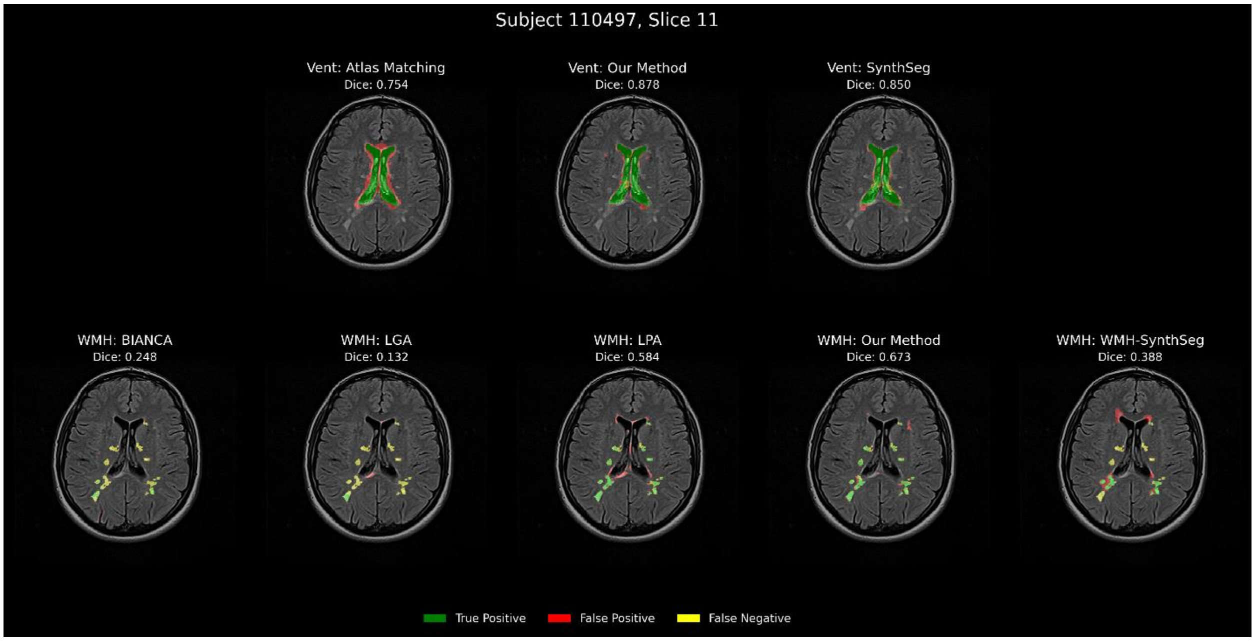

*Figure 5. A sample FLAIR image with predicted masks. The first and second rows show the results of ventricle and WMH segmentation, respectively. Color-coding indicates true positives (green), false positives (red), and false negatives (yellow) for visual comparison.*

### 3.2.2. Comparative Visual Analysis

#### 3.2.2.1. Ventricle Segmentation Comparison

The three ventricle segmentation methods exhibited distinct characteristics: Atlas Matching produced substantial false positives at ventricular boundaries and adjacent sulcal spaces, demonstrating limitations when applied to MS patients with anatomical variations. SynthSeg showed more conservative segmentation with moderate false negatives, particularly in the inferior portions of lateral ventricles and third ventricle. Our method achieved balanced performance with minimal false positives and negatives. Segmentation boundaries closely followed ventricular margins even in challenging regions such as ventricular horns.

#### 3.2.2.2. WMH Segmentation Comparison

The five WMH segmentation methods demonstrated varying capabilities: LGA showed significant under-segmentation with extensive false negatives, particularly for smaller and less hyperintense lesions. BIANCA exhibited substantial false negatives with limited detection of periventricular lesions characteristic of MS pathology. LPA demonstrated improved lesion detection but increased false positives, frequently misclassifying normal periventricular hyperintensities as pathological. WMH-SynthSeg showed moderate false negatives for smaller and challenging lesions and substantial false positives in periventricular regions. Our method achieved superior performance with extensive true positives encompassing both large and small lesions while maintaining minimal false positives and negatives. Notably, our approach successfully differentiated between normal periventricular hyperintensities and pathological MS lesions.

#### 3.2.2.3. Expert Analysis of Segmentation Quality

Qualitative assessment by a neuroradiologist with 20 years of experience in MS imaging confirmed our method provided clinically accurate ventricular delineation in 92% of cases, compared to 78% for SynthSeg and 72% for Atlas Matching. For WMH segmentation, our method's ability to distinguish between normal and abnormal hyperintensities was deemed "clinically valuable" or "highly valuable" in 81% of cases, particularly in patients with confluent periventricular WMH. The expert highlighted that our segmentation results exhibited greater anatomical plausibility and respected known MS lesion patterns.

## 3.3. Segmentation Performance: Quantitative Analysis

### 3.3.1. Ventricle Segmentation Metrics

#### 3.3.1.1. Confusion Matrix Analysis

The confusion matrix analysis for ventricle segmentation (Figure 6, first row) revealed distinct performance patterns across the three methods. Our proposed method demonstrated superior true positive rate (0.89) compared to Atlas Matching (0.77) and SynthSeg (0.66). The false negative rate was correspondingly lower for our method (0.11) compared to Atlas Matching (0.23) and SynthSeg (0.34), demonstrating improved sensitivity. While false positive rates were numerically small across all methods due to the class imbalance, our method maintained competitive performance with a false positive rate between Atlas Matching (highest) and SynthSeg (lowest).

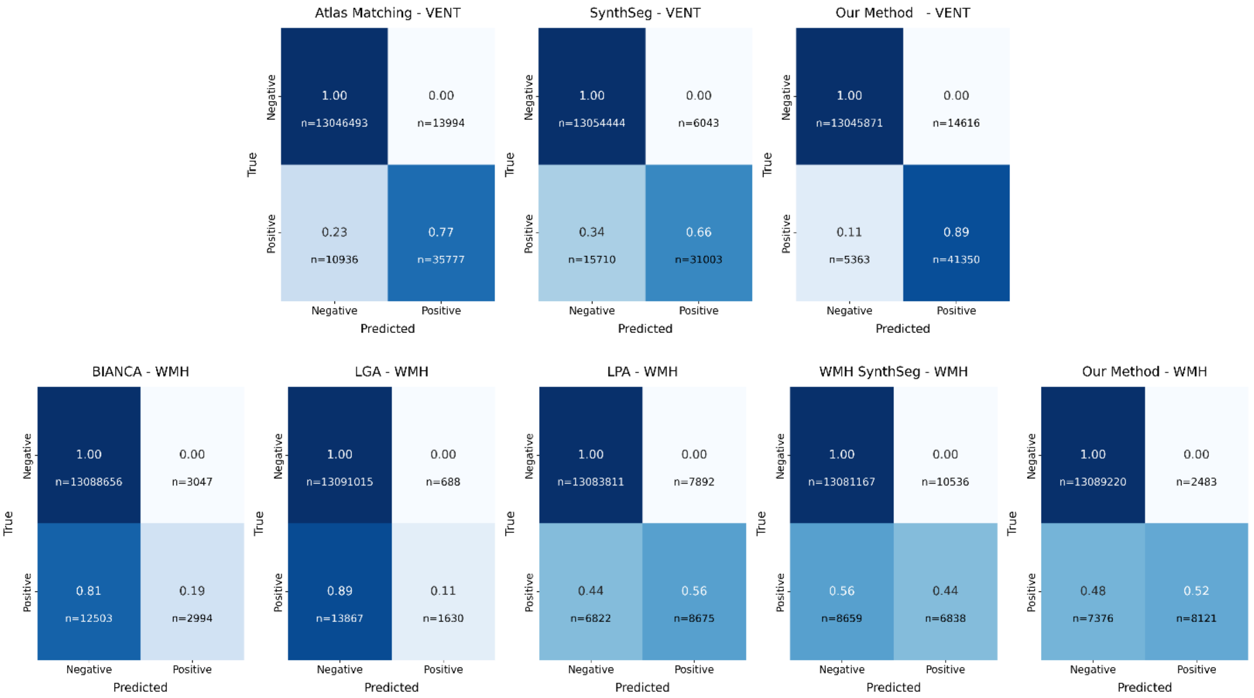

*Figure 6. Confusion matrices for the segmentation methods. The first and second rows correspond to the ventricle and WMH segmentation tasks, respectively.*

### 3.3.1.2. Detection-based Metrics

Analysis of detection precision and recall (Figure 7, first and second rows) revealed complementary strengths across methods. Our approach achieved the highest recall (0.883 ± 0.032), significantly surpassing SynthSeg (0.698 ± 0.130) and Atlas Matching (0.766 ± 0.044). This demonstrates our method's superior ability to identify a greater proportion of ventricular regions. In terms of precision, SynthSeg led with 0.835 ± 0.036, followed by our method (0.736 ± 0.05) and Atlas Matching (0.730 ± 0.108). While our method did not achieve the highest precision, its balanced performance across both precision and recall metrics indicates a more clinically useful segmentation, as evidenced by the superior Dice and Jaccard metrics as follows.

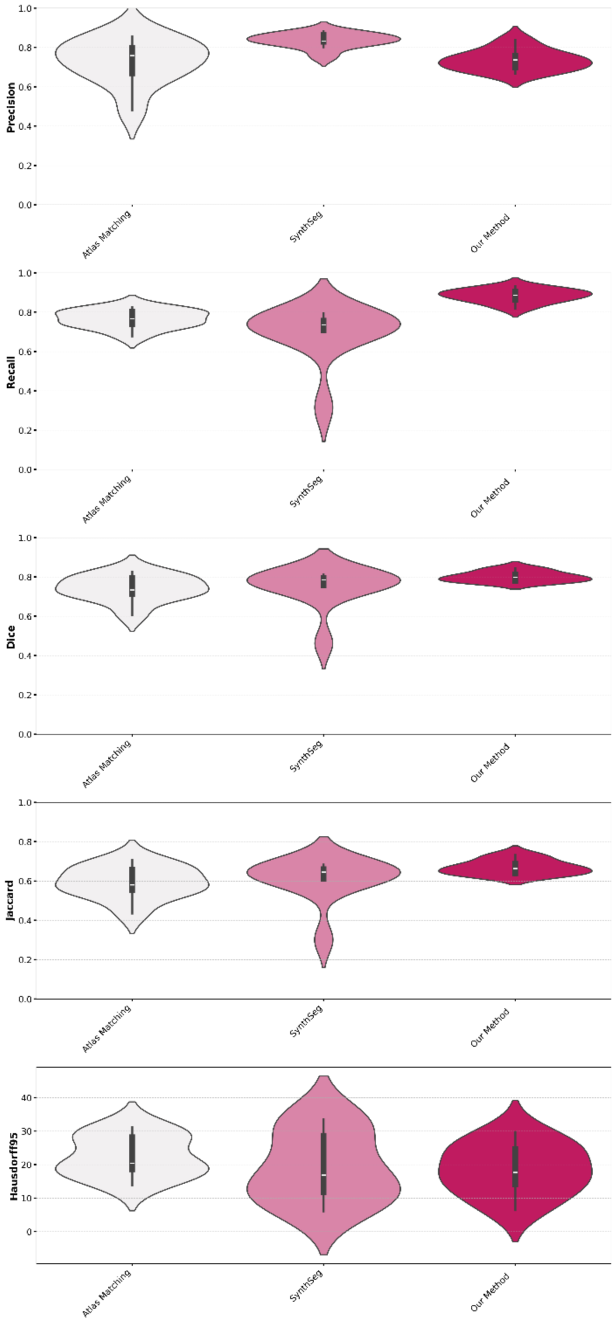

*Figure 7. Violin plots for the ventricle segmentation task comparing three methods—Atlas Matching, SynthSeg, and our proposed method—across five metrics: precision, recall, Dice, Jaccard, and Hausdorff95.*

### 3.3.1.3. Overlap Metrics

The violin plots of overlap metrics (Figure 7, third and fourth rows) demonstrate the superior performance of our method in terms of spatial overlap with ground truth ventricle segmentations.

Our approach achieved the highest mean Dice coefficient (0.801 ± 0.025), significantly outperforming both SynthSeg (0.751 ± 0.098) and Atlas Matching (0.742 ± 0.062). Similarly, for Jaccard index, our method attained a mean of 0.669 ± 0.035, compared to SynthSeg (0.609 ± 0.106) and Atlas Matching (0.593 ± 0.076). Notably, our method exhibited substantially lower variability in both Dice and Jaccard metrics (standard deviations of 0.025 and 0.035, respectively.

#### 3.3.1.4. Boundary Metrics

The boundary accuracy assessment using the HD95 (Figure 7, fifth row) demonstrated our method's advantage in delineating ventricular boundaries. Our approach achieved the lowest mean HD95 (18.46 ± 7.1 mm), compared to SynthSeg (19.07 ± 9.77 mm) and Atlas Matching (22.51 ± 5.73 mm). The reduced HD95 values indicate more accurate boundary localization, which is particularly important for volumetric measurements that depend on precise ventricular delineation.

#### 3.3.1.5. Precision-Recall Analysis

The precision-recall curve analysis (Figure 8, left) further substantiated the overall superiority of our method. Our approach achieved the highest area under the curve (AUC = 0.857), compared to SynthSeg (AUC = 0.792) and Atlas Matching (AUC = 0.778). This comprehensive metric, which evaluates performance across various threshold settings, confirms that our method maintains a better balance between precision and recall across the operating range.

### 3.3.2. White Matter Hyperintensity Segmentation Metrics

#### 3.3.2.1. Confusion Matrix Analysis

The confusion matrix analysis for WMH segmentation (Figure 6, second row) revealed marked differences in performance across the five evaluated methods. Our approach achieved a true positive rate of 0.52, second only to LPA (0.56) and substantially higher than WMH-SynthSeg (0.44), BIANCA (0.19), and LGA (0.11). Correspondingly, our method demonstrated the second-lowest false negative rate (0.48), behind only LPA (0.44). The false positive rates were generally low across all methods due to class imbalance, with our method achieving a favorable balance between false positives and true positives.

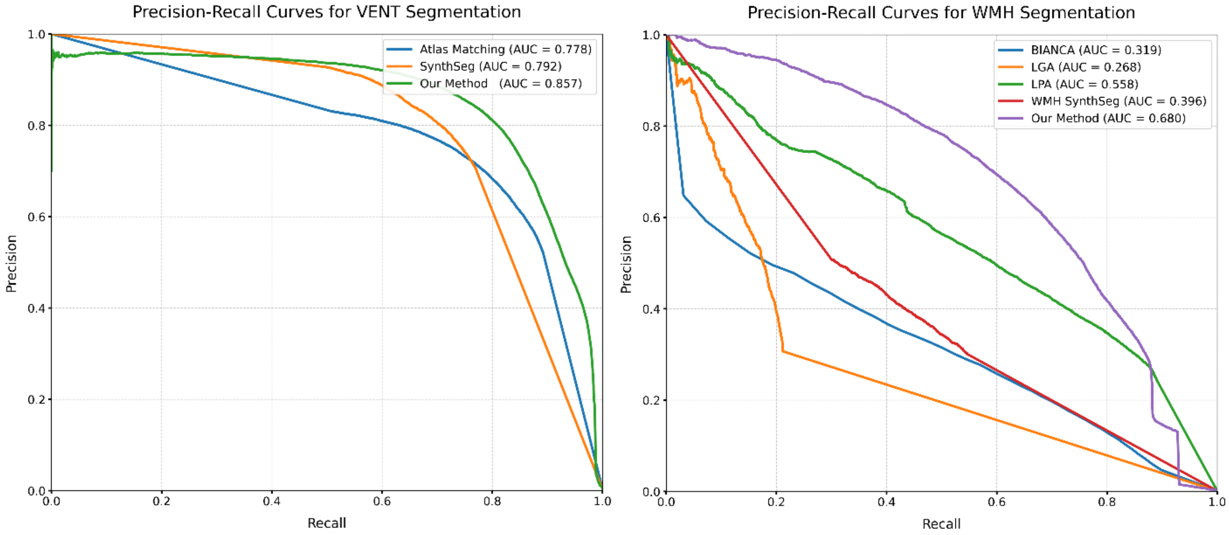

*Figure 8. Precision–recall curves of the compared methods for ventricle (left) and WMH (right) segmentation tasks.*

### 3.3.2.2. Detection-based Metrics

Analysis of detection precision and recall (Figure 9, first and second rows) revealed that our method achieved the highest precision (0.755 ± 0.161) among all evaluated approaches, substantially outperforming LGA (0.660 ± 0.255), LPA (0.497 ± 0.131), BIANCA (0.474 ± 0.209), and WMH-SynthSeg (0.374 ± 0.181). This indicates our method's superior ability to minimize false positive detections. For recall, our method achieved the highest value (0.558 ± 0.091), followed very closely by LPA (0.556 ± 0.108), with WMH-SynthSeg (0.445 ± 0.099), BIANCA (0.211 ± 0.097), and LGA (0.092 ± 0.051) achieving lower performance. The combination of high precision and competitive recall demonstrates our method's balanced performance in WMH detection, which is particularly important given the clinical significance of accurate hyperintensity identification.

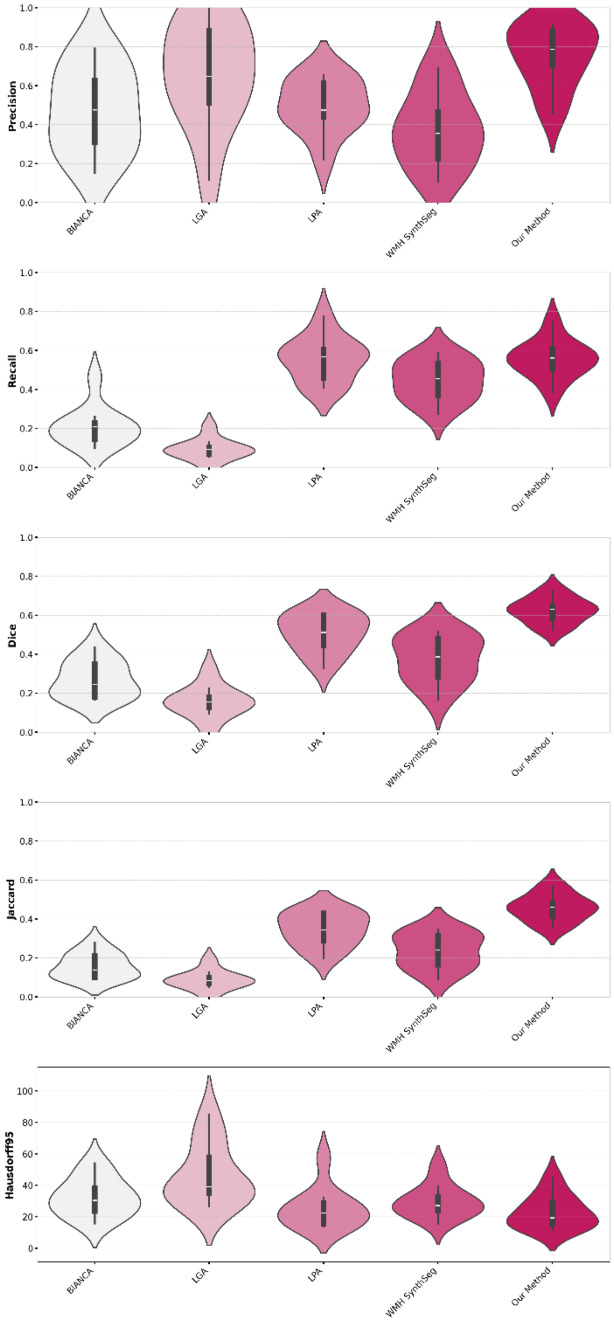

*Figure 9. Violin plots for the WMH segmentation task comparing five methods—BIANCA, LST-LGA, LST-LPA, WMH-SynthSeg, and our proposed method—across five metrics: precision, recall, Dice, Jaccard, and Hausdorff95.*

### 3.3.2.3. Overlap Metrics

The violin plots of overlap metrics (Figure 9, third and fourth rows) demonstrated that our method achieved the highest mean Dice coefficient (0.624 ± 0.061) among all evaluated approaches, significantly outperforming the next best method, LPA (0.509 ± 0.093). The other methods showed considerably lower performance: WMH-SynthSeg (0.376 ± 0.114), BIANCA (0.268 ± 0.090), and LGA (0.156 ± 0.078). Similarly, for the Jaccard index, our method attained the highest mean value (0.457 ± 0.064), followed by LPA (0.346 ± 0.082), WMH-SynthSeg (0.238 ± 0.085), BIANCA (0.158 ± 0.062), and LGA (0.087 ± 0.047). The consistent superiority across both overlap metrics confirms our approach's enhanced capability to accurately identify and delineate white matter hyperintensities.

#### 3.3.2.4. Boundary Metrics

The boundary accuracy assessment using HD95 (Figure 9, fifth row) showed our method achieved the best performance (23.0 ± 10.06 mm), outperforming LPA (25.17 ± 12.44 mm), WMH-SynthSeg (29.50 ± 9.83 mm), BIANCA (32.39 ± 11.53 mm), and LGA (46.39 ± 18.44 mm). The lower HD95 values indicate more precise boundary delineation, which is crucial for accurate lesion load quantification and morphological analysis of WMHs.

#### 3.3.2.5. Precision-Recall Analysis

The precision-recall curve analysis (Figure 8, right) provided comprehensive evidence of our method's superior performance. Our approach achieved the highest AUC (0.68), substantially outperforming LPA (AUC = 0.558), WMH-SynthSeg (AUC = 0.396), BIANCA (AUC = 0.319), and LGA (AUC = 0.268). This significant advantage in the AUC metric, which evaluates performance across various operating points, confirms the robust performance of our method across different sensitivity thresholds.

### 3.3.3. Normal vs. Abnormal Hyperintensity

#### 3.3.3.1. Confusion Matrix Analysis

The confusion matrix analysis for differentiation between normal and abnormal WMH (Figure 10) revealed balanced performance patterns between the two classes. Our method achieved a true positive rate of 0.64 (n=9953), indicating substantial success in correctly identifying abnormal hyperintensities. Similarly, the true negative rate was high at 0.75 (n=15822), demonstrating the method's strong capability in correctly classifying normal hyperintensities. The false positive and false negative rates were 0.25 (n=5308) and 0.36 (n=5544), respectively, showing a balanced classification approach.

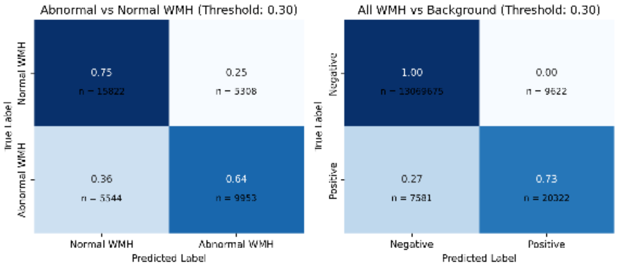

*Figure 10. Confusion matrices of the proposed method. The left matrix shows the classification of abnormal WMH vs. normal WMH, while the right matrix shows the segmentation of all WMH vs. background.*

### 3.3.3.2. Classification Metrics

Based on the confusion matrix values, our method demonstrated robust performance metrics for distinguishing between normal and abnormal WMH. The sensitivity (recall) was substantial at 0.6423, indicating that approximately two-thirds of the abnormal hyperintensities were correctly identified. The specificity was higher at 0.7488, reflecting strong performance in correctly classifying normal hyperintensities. The precision was 0.6522, suggesting that when our method classified a hyperintensity as abnormal, it was correct in approximately 65% of cases. The Dice coefficient was 0.6472, indicating good overall performance in differentiating between the two WMH types. These metrics were determined after a thorough threshold analysis on prediction probabilities, with optimal results achieved at a threshold of 0.3.

### 3.3.3.3. Error Analysis

Analysis of the misclassification patterns revealed a more balanced distribution, with false negatives (0.36) slightly exceeding false positives (0.25). This improved balance suggests that our method exhibits less systematic bias in classification. The reduced false negative rate indicates fewer abnormal WMH were missed or incorrectly classified as normal, which positively impacts clinical assessments related to disease burden quantification. The ROC analysis (Figure 11, right) yielded an AUC of 0.962 for abnormal WMH classification, while the precision-recall curve analysis (Figure 11, left) showed an AUC of 0.68, further supporting the method's discriminative capability.

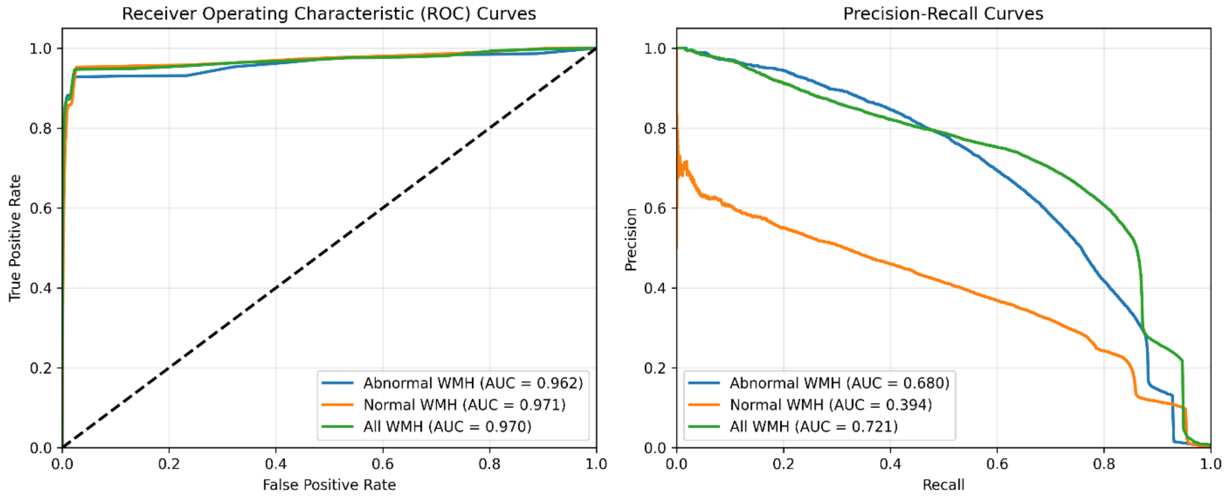

*Figure 11. ROC curves (left) and precision–recall curves (right) for all WMH, abnormal WMH, and normal WMH classes.*

## 3.4. Computational Efficiency

### 3.4.1. Processing Time and Resource Analysis

Our comprehensive analysis of computational efficiency across all evaluated methods revealed substantial differences in processing time and resource requirements, as detailed in Table 1. Our proposed deep learning approach demonstrated exceptional efficiency, requiring approximately 4 seconds for end-to-end processing per patient case (2.4 seconds for pre-processing and 1.6 seconds for model inference and post-processing), dramatically outperforming all baseline methods.

The detailed breakdown of execution times revealed marked differences across all methods. BIANCA achieved moderate performance with an 11-second processing time but demonstrated limited segmentation accuracy. The traditional atlas-based approaches (LST-LGA and Atlas Matching) exhibited the longest processing times at 147 and 115 seconds per case, respectively, making them less suitable for time-sensitive clinical applications. The deep learning-based SynthSeg and WMH-SynthSeg methods showed intermediate performance (124 and 78 seconds) but required full CPU utilization and considerable memory resources (~5 GB RAM and 100% GPU usage).

*Table 1. Computational Performance Comparison of Segmentation Methods*

| Method | Training | Inference | CPU | RAM | GPU |
|---|---|---|---|---|---|
| BIANCA | 58 sec | 11 sec | ~20% | ~1 GB | N/A |
| LST-LGA | N/A | 147 sec | ~60% | ~2 GB | N/A |
| LST-LPA | N/A | 72 sec | ~40% | ~2 GB | N/A |
| WMH-SynthSeg | N/A | 78 sec | 100% | ~5 GB | 100% |
| Atlas Matching | N/A | 115 sec | ~40% | ~2 GB | N/A |
| SynthSeg | N/A | 124 sec | 100 % | ~5 GB | 100% |
| Our Method | 257 sec* | <4 sec[†] | 15% | ~1 GB | 80% |

* Time per training epoch. Total training time: 214 minutes (50 epochs).

[†] Includes 2.4 sec for pre-processing and 1.6 sec for model inference and post-processing.

In contrast, our method's 4-second processing time represents a significant improvement over other approaches while maintaining minimal resource requirements (15% CPU utilization, ~1 GB RAM). This efficiency was achieved through our streamlined network architecture and optimized pre-processing pipeline. Despite utilizing GPU acceleration (80% utilization of an RTX3060), our method's memory footprint remains minimal, reflecting the efficient model design and optimization, making it well-suited for deployment in environments with limited computational resources.

# 4. Discussion

This study introduces a novel deep learning approach for the simultaneous segmentation of ventricles and white matter hyperintensities in multiple sclerosis, with the unique capability to differentiate between normal and pathological hyperintensities. Our findings demonstrate that the proposed pix2pix-based framework surpasses existing methods in both segmentation accuracy and computational efficiency, addressing several critical limitations in current neuroimaging analysis practices.

## 4.1. Technical Innovations and Performance Advantages

The superior performance of our method can be attributed to several key technical innovations. First, our approach uniquely addresses the joint segmentation of ventricles and WMH within a unified framework, leveraging the anatomical and pathophysiological relationship between these structures. This is particularly important in MS, where periventricular lesions represent a hallmark of the disease and their precise delineation in relation to ventricular structures provides valuable diagnostic information (Gindin et al., 2024; Rondinella et al., 2024). Rachmadi et al. (2018) has demonstrated that incorporating anatomical context significantly improves the accuracy of deep learning-based medical image segmentation by leveraging spatial relationships between adjacent structures. By simultaneously segmenting these structures, our model inherently learns their spatial relationships, improving contextual understanding and classification accuracy.

Second, the ability to differentiate between normal periventricular hyperintensities and pathological lesions represents a significant advancement over existing approaches that treat all hyperintensities as uniform entities. This distinction is crucial in clinical practice, as normal age-related hyperintensities can masquerade as MS lesions, potentially leading to misdiagnosis or overestimation of disease burden (Dadar et al., 2020; Griffanti et al., 2018; McKinley et al., 2021; Tran et al., 2022). Our results demonstrate robust differentiation capability (Dice coefficient: 0.647) between these two classes, providing clinicians with more accurate assessment of true disease-related pathology.

Third, our 2D slice-based approach specifically addresses the reality of clinical imaging protocols that typically produce anisotropic data. While most advanced segmentation methods are designed for isotropic research-grade volumes (Billot et al., 2023; Atlason et al., 2022; La Rosa et al., 2020; Raut et al., 2024; Rondinella et al., 2024; Laso et al., 2023; Tran et al., 2022), the substantial difference between in-plane resolution and slice thickness in routine clinical scans creates fundamental incompatibilities. Our approach embraces this constraint rather than attempting to circumvent it, resulting in a method optimized for real-world clinical data characteristics rather than idealized research conditions.

The quantitative performance metrics substantiate these advantages. For ventricle segmentation, our method achieved the highest Dice coefficient (0.801 ± 0.025) and lowest HD95 (18.46 ± 7.1 mm), demonstrating both superior overlap accuracy and boundary precision compared to established approaches. Notably, our method exhibited remarkably low variability in performance (standard deviations of 0.025 and 0.035 for Dice and Jaccard, respectively), indicating robust generalization across diverse patient presentations—a crucial characteristic for reliable clinical application.

For WMH segmentation, our method similarly outperformed all baselines across multiple metrics, with the highest Dice coefficient (0.624 ± 0.061) and precision (0.755 ± 0.161). The performance advantage was particularly pronounced in precision, where our method exceeded the next best approach (LST-LGA) by nearly 10 percentage points. This high precision is clinically significant as it indicates a lower rate of false positive detections, which could otherwise lead to overestimation of disease burden.

## 4.2. Computational Efficiency and Clinical Implementation

Perhaps equally important to segmentation accuracy is the exceptional computational efficiency demonstrated by our approach. With an end-to-end processing time of less than 4 seconds per case, our method is 18-36 times faster than the deep learning-based alternatives (SynthSeg, WMH-SynthSeg) and up to 36 times faster than traditional approaches (LST-LGA, Atlas Matching). This dramatic reduction in processing time, coupled with modest resource requirements (15% CPU utilization, ~1 GB RAM, 80% GPU usage), represents a paradigm shift in the feasibility of integrating advanced segmentation into routine clinical workflows. This exceptional efficiency can be attributed to several factors: (1) our adoption of a 2D segmentation approach optimized for the non-isotropic clinical data, (2) the computational efficiency of the pix2pix architecture with minimal post-processing requirements, and (3) our streamlined pre-processing pipeline that minimizes computational overhead while preserving essential image characteristics for accurate segmentation. Similar computational efficiency benefits have been reported by Umirzakova et al. (2025) and Griffanti et al. (2016) in their work on optimized deep learning frameworks for clinical radiology, where they identified processing speed as a critical factor for clinical adoption of AI-based image analysis tools.

Current neuroimaging analysis approaches frequently require extensive computational resources and processing times measured in minutes or hours, rendering them impractical for time-sensitive clinical decision-making (Laso et al., 2023; Atlason et al., 2022; McKinley et al., 2021; Rondinella et al., 2024). Our method's processing speed enables near real-time analysis, potentially allowing results to be available during the same clinical session in which imaging was performed. This capability could significantly impact clinical practice by providing immediate quantitative assessment of disease burden and progression, potentially informing treatment decisions without the delays currently associated with advanced image analysis.

The minimal hardware requirements of our approach—a standard CPU and modest RAM, with a consumer-grade GPU acceleration—further enhance its accessibility across diverse clinical settings, including resource-limited environments where advanced computing infrastructure may not be available. This democratization of advanced neuroimaging analysis capabilities aligns with broader goals of improving equity in healthcare technology access.

## 4.3. Clinical Implications

Beyond technical metrics, our approach offers several clinically relevant advantages. The simultaneous segmentation of ventricles and WMH provides a comprehensive assessment of two key imaging biomarkers in MS. Ventricular enlargement serves as an important indicator of brain atrophy and disease progression (Atlason et al., 2022; Laso et al., 2023), while WMH burden correlates with clinical disability measures and cognitive impairment (Wang et al., 2022;

Griffanti et al., 2018; Jiménez-Balado et al., 2022; La Rosa et al., 2020; Zhu et al., 2022). Recent studies by (Jiménez-Balado et al., 2022; McKinley et al., 2021; Melazzini et al., 2021; Rachmadi et al., 2018; Atlason et al., 2022; Wang et al., 2020; Griffanti et al., 2018) have further demonstrated that accurate quantification of both ventricular volume and WMH load provides complementary information that enhances prediction of clinical outcomes and treatment response in MS patients. By quantifying both within a unified framework, our method enables integrated analysis of their relationship, potentially revealing patterns not apparent when assessed independently.

The differentiation between normal and abnormal hyperintensities addresses a significant challenge in MS imaging, particularly in older patients where age-related white matter changes can confound assessment of disease-specific lesions. This capability could potentially reduce diagnostic uncertainty and improve the accuracy of treatment response monitoring by focusing on true disease-related pathology rather than normal aging processes.

Expert qualitative assessment affirmed these clinical advantages, with our method's ventricular delineation deemed clinically accurate in 92% of cases and its normal/abnormal WMH differentiation considered "clinically valuable" or "highly valuable" in 81% of cases. These assessments from experienced neuroradiologists validate that our technical improvements translate to meaningful clinical utility.

## 4.4. Limitations and Future Directions

Despite these advantages, several limitations warrant consideration. First, the differentiation between normal and abnormal hyperintensities remains particularly challenging in cases with confluent periventricular lesions, where the boundary between physiological and pathological hyperintensities becomes inherently ambiguous. This limitation reflects a fundamental challenge in MS imaging rather than a specific limitation of our approach. Our confusion matrix analysis of abnormal versus normal WMH revealed a higher false negative rate relative to false positives, suggesting inherent challenges in distinguishing pathological lesions from normal periventricular hyperintensities, particularly when abnormal WMH present with subtle intensity variations or occur in regions commonly affected by normal aging processes. Multi-modal integration incorporating additional sequences beyond FLAIR and T1 might provide additional context for improved differentiation in these cases. Future refinements could focus on further enhancing the sensitivity for abnormal WMH detection through incorporation of additional contextual features that account for the spatial distribution and morphological characteristics of hyperintensities.

Second, while our patient cohort was substantial (300 patients) and diverse in terms of age and gender distribution, all imaging was performed on a single scanner model with consistent acquisition parameters. This constrained variability may limit generalizability to data acquired under substantially different conditions. However, this limitation is common to most learning-based approaches in medical imaging and could be addressed through transfer learning techniques or targeted fine-tuning for new acquisition parameters.

Future directions for this work include expanded validation across multi-site datasets with diverse scanner types and acquisition protocols to assess generalizability. Integration of longitudinal information could further enhance the model's understanding of disease progression patterns. Additionally, exploration of more sophisticated post-processing techniques that incorporate anatomical priors might further improve segmentation in challenging regions.

The clinical utility of our approach could be further enhanced through integration with automated volumetric quantification and reporting tools, providing standardized metrics directly to clinicians. Development of a user-friendly interface would facilitate adoption in clinical settings, potentially transforming how MS imaging is evaluated in routine practice.

# 5. Conclusion

This study presents a novel deep learning approach for simultaneous segmentation of ventricles and white matter hyperintensities in multiple sclerosis, with the unique capability to differentiate between normal and pathological hyperintensities. Our pix2pix-based framework demonstrates superior performance over existing methods, achieving the highest Dice coefficients for both ventricles (0.801±0.025) and WMHs (0.624±0.061), while successfully differentiating between normal and abnormal hyperintensities (Dice: 0.647). The exceptional computational efficiency—processing a case in approximately 4 seconds, up to 36 times faster than baseline methods—represents a significant step toward clinical integration. While limitations exist, particularly in cases with confluent periventricular lesions and in dataset diversity, our approach addresses critical gaps in current neuroimaging analysis for MS. By combining improved accuracy, clinically relevant differentiation, and computational efficiency, this work contributes to the advancement of quantitative neuroimaging tools that can be practically implemented in routine clinical environments, potentially enhancing MS diagnosis, treatment planning, and disease monitoring.

# Acknowledgements

We acknowledge Eng. Mehrdad Rahbarpour, Eng. Azad Ahmadi, and Dr. Aydin Asadi, colleagues at the Gholghasht Medical Imaging Center for their assistance in acquiring data.

# Declaration of Conflicting Interest

The authors declare no conflict of interest.

# Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

# Data and Code Availability Statement

The code and sample patient data that support the findings of this study are publicly available at https://github.com/Mahdi-Bashiri/Sim-Vent-WMH-Brain-Seg. The complete dataset used in this study is not publicly available due to privacy considerations and institutional policies. However, access to additional data may be granted to qualified researchers upon reasonable request directed to the corresponding author. Any inquiries regarding data access, implementation details, or further collaboration should be addressed to the corresponding author.

# References

Atlason, H. E., Love, A., Robertsson, V., Blitz, A. M., Sigurdsson, S., Gudnason, V., & Ellingsen, L. M. (2022). A joint ventricle and WMH segmentation from MRI for evaluation of healthy and pathological changes in the aging brain. *PLOS ONE*, *17*(9), e0274212. https://doi.org/10.1371/journal.pone.0274212

Bawil, M. B., Shamsi, M., Bavil, A. S., & Danishvar, S. (2024). Specialized gray matter segmentation via a generative adversarial network: application on brain white matter hyperintensities classification. *Frontiers in Neuroscience*, *18*. https://doi.org/10.3389/fnins.2024.1416174

Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., & Iglesias, J. E. (2023). SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, *86*, 102789. https://doi.org/10.1016/j.media.2023.102789

Cathala, C., Ferath Kherif, Jean-Philippe Thiran, Bussy, A., & Bogdan Draganski. (2025). WHITE-Net: White matter HyperIntensities Tissue Extraction using deep learning Network. *MedRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2025.01.09.25320242

Dadar, M., Potvin, O., Camicioli, R., & Duchesne, S. (2021). Beware of white matter hyperintensities causing systematic errors in FreeSurfer gray matter segmentations! *Human Brain Mapping*, *42*(9), 2734–2745. https://doi.org/10.1002/hbm.25398

DeCarli, C., Fletcher, E., Ramey, V., Harvey, D., & Jagust, W. J. (2005). Anatomical Mapping of White Matter Hyperintensities (WMH). *Stroke*, *36*(1), 50–55. https://doi.org/10.1161/01.str.0000150668.58689.f2

Dong, Z., He, Y., Qi, X., Chen, Y., Shu, H., Coatrieux, J.-L., Yang, G., & Li, S. (2022). MNet: Rethinking 2D/3D Networks for Anisotropic Medical Image Segmentation. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.04846

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Gabr, R. E., Coronado, I., Robinson, M., Sujit, S. J., Datta, S., Sun, X., Allen, W. J., Lublin, F. D., Wolinsky, J. S., & Narayana, P. A. (2019). Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. *Multiple Sclerosis (Houndmills, Basingstoke, England)*, 1352458519856843. https://doi.org/10.1177/1352458519856843

Gindin, M., Hila Pascovich, & Cohen, M. W. (2024). Segmenting Multi-Sclerosis Lesions: An Enhanced Approach using an Optimized U-Net. *Procedia Computer Science*, *246*, 2225–2234. https://doi.org/10.1016/j.procs.2024.09.577

Griffanti, L., Jenkinson, M., Suri, S., Zsoldos, E., Mahmood, A., Filippini, N., Sexton, C. E., Topiwala, A., Allan, C., Kivimäki, M., Singh-Manoux, A., Ebmeier, K. P., Mackay, C. E., & Zamboni, G. (2018). Classification and characterization of periventricular and deep white matter hyperintensities on MRI: A study in older adults. *NeuroImage*, *170*, 174–181. https://doi.org/10.1016/j.neuroimage.2017.03.024

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M., & Jenkinson, M. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, *141*, 191–205. https://doi.org/10.1016/j.neuroimage.2016.07.018

Hashemi, M., Akhbari, M., & Jutten, C. (2022). 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. https://doi.org/10.1016/j.compbiomed.2022.105402

Hossain, M. I., Amin, M. Z., Anyimadu, D. T., & Suleiman, T. A. (2024). Comparative Study of Probabilistic Atlas and Deep Learning Approaches for Automatic Brain Tissue Segmentation from MRI Using N4 Bias Field Correction and Anisotropic Diffusion Pre-processing Techniques. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.05456

Huang, F., Xia, P., Varut Vardhanabhuti, Hui, S., Lau, K., Mak, H. K., & Cao, P. (2022). Semisupervised white matter hyperintensities segmentation on MRI. *Human Brain Mapping*, *44*(4), 1344–1358. https://doi.org/10.1002/hbm.26109

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2016). *Image-to-Image Translation with Conditional Adversarial Networks*. https://doi.org/10.48550/arxiv.1611.07004

Jiménez-Balado, J., Corlier, F., Habeck, C., Stern, Y., & Eich, T. (2022). Effects of white matter hyperintensities distribution and clustering on late-life cognitive impairment. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-06019-8

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P.-J., Galbusera, R., Barakovic, M., Thiran, J.-P., Granziera, C., & Cuadra, M. B. (2020). Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *NeuroImage: Clinical*, *27*, 102335. https://doi.org/10.1016/j.nicl.2020.102335

Laso, P., Cerri, S., Sorby-Adams, A., Guo, J., Mateen, F., Goebl, P., Wu, J., Liu, P., Li, H., Young, S. I., Billot, B., Puonti, O., Sze, G., Payabavash, S., DeHavenon, A., Sheth, K. N., Rosen, M. S., Kirsch, J., Strisciuglio, N., & Wolterink, Jelmer M. (2023). Quantifying white matter hyperintensity and brain volumes in heterogeneous clinical and low-field portable MRI. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2312.05119

Li, J., Santini, T., Huang, Y., Mettenburg, J. M., Ibrahim, T. S., Aizenstein, H. J., & Wu, M. (2024). wmh_seg: Transformer based U-Net for Robust and Automatic White Matter Hyperintensity Segmentation across 1.5T, 3T and 7T. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.12701

McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Reyes, M., Salmen, A., Chan, A., Wagner, F., & Wiest, R. (2021). Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-020-79925-4

Melazzini, L., Mackay, C. E., Bordin, V., Suri, S., Zsoldos, E., Filippini, N., Mahmood, A., Sundaresan, V., Codari, M., Duff, E., Singh-Manoux, A., Kivimäki, M., Ebmeier, K. P., Jenkinson, M., Sardanelli, F., & Griffanti, L. (2021). White matter hyperintensities classified according to intensity and spatial location reveal specific associations with cognitive performance. *NeuroImage: Clinical*, *30*, 102616. https://doi.org/10.1016/j.nicl.2021.102616

Muhammad Febrian Rachmadi, Maria, Leonora, M., Carol Di Perri, Taku Komura, & Alzheimer's Disease Neuroimaging Initiative. (2018). Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Computerized Medical Imaging and Graphics*, *66*, 28–43. https://doi.org/10.1016/j.compmedimag.2018.02.002

Park, G., Hong, J., Duffy, B. A., Lee, J.-M., & Kim, H. (2021). White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds. *NeuroImage*, *237*, 118140. https://doi.org/10.1016/j.neuroimage.2021.118140

Rakić, M., Vercruyssen, S., Van Eyndhoven, S., de la Rosa, E., Jain, S., Van Huffel, S., Maes, F., Smeets, D., & Sima, D. M. (2021). icobrain ms 5.1: Combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions. *NeuroImage. Clinical*, *31*, 102707. https://doi.org/10.1016/j.nicl.2021.102707

Raut, P., Baldini, G., M. Schöneck, & Caldeira, L. (2024). Using a generative adversarial network to generate synthetic MRI images for multi-class automatic segmentation of brain tumors. *Frontiers in Radiology*, *3*. https://doi.org/10.3389/fradi.2023.1336902

Rondinella, A., Guarnera, F., Crispino, E., Russo, G., Lorenzo, D., Maimone, D., Pappalardo, F., & Battiato, S. (2024). ICPR 2024 Competition on Multiple Sclerosis Lesion Segmentation -- Methods and Results. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.07924

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., Hemmer, B., & Mühlau, M. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage*, *59*(4), 3774–3783. https://doi.org/10.1016/j.neuroimage.2011.11.032

Schmidt, P., Pongratz, V., Pascal Küster, Meier, D., Wuerfel, J., Lukas, C., Bellenberg, B., Zipp, F., Sergiu Groppa, Sämann, P. G., Weber, F., Gaser, C., Franke, T., Matthias Bussas, Kirschke, J., Zimmer, C., Hemmer, B., & Mühlau, M. (2019). Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage Clinical*, *23*, 101849–101849. https://doi.org/10.1016/j.nicl.2019.101849

Tran, P., Urielle Thoprakarn, Emmanuelle Gourieux, Honorato, C., Enrica Cavedo, Guizard, N., Cotton, F., Krolak-Salmon, P., Delmaire, C., Heidelberg, D., Nadya Pyatigorskaya, Sébastian Ströer, Didier Dormont, Martini, J., & Chupin, M. (2022). Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both Multiple Sclerosis and elderly subjects. *NeuroImage: Clinical*, *33*, 102940–102940. https://doi.org/10.1016/j.nicl.2022.102940

Ulloa-Poblete, G., Allende-Cid, H., Veloz, A., & Allende, H. (2023). Edges-enhanced Convolutional Neural Network for Multiple Sclerosis Lesions Segmentation. *Computación Y Sistemas*, *27*(1). https://doi.org/10.13053/cys-27-1-4535

Umirzakova, S., Muksimova Shakhnoza, Mardieva Sevara, & Taeg Keun Whangbo. (2025). Deep learning for multiple sclerosis lesion classification and stratification using MRI. *Computers in Biology and Medicine*, *192*, 110078–110078. https://doi.org/10.1016/j.compbiomed.2025.110078

Walton, C., King, R., Rechtman, L., Kaye, W., Leray, E., Marrie, R. A., Robertson, N., La Rocca, N., Uitdehaag, B., van der Mei, I., Wallin, M., Helme, A., Angood Napier, C., Rijke, N., & Baneke, P. (2020). Rising Prevalence of Multiple Sclerosis worldwide: Insights from the Atlas of MS, Third Edition. *Multiple Sclerosis Journal*, *26*(14), 1816–1821. https://doi.org/10.1177/1352458520970841

Wang, J., Zhou, Y., He, Y., Li, Q., Zhang, W., Luo, Z., Xue, R., & Lou, M. (2022). Impact of different white matter hyperintensities patterns on cognition: A cross-sectional and longitudinal study. *NeuroImage: Clinical*, *34*, 102978–102978. https://doi.org/10.1016/j.nicl.2022.102978

Wu, L., Wang, S., Liu, J., Hou, L., Li, N., Su, F., Yang, X., Lu, W., Qiu, J., Zhang, M., & Song, L. (2024). A survey of MRI-based brain tissue segmentation using deep learning. *Complex & Intelligent Systems*, *11*(1). https://doi.org/10.1007/s40747-024-01639-1

Zeng, C., Gu, L., Liu, Z., & Zhao, S. (2020). Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Frontiers in Neuroinformatics*, *14*. https://doi.org/10.3389/fninf.2020.610967

Zhang, Y., Duan, Y., Wang, X., Zhuo, Z., Haller, S., Barkhof, F., & Liu, Y. (2021). A deep learning algorithm for white matter hyperintensity lesion detection and segmentation. *Neuroradiology*, *64*(4), 727–734. https://doi.org/10.1007/s00234-021-02820-w

Zhu, W., Huang, H., Zhou, Y., Shi, F., Shen, H., Chen, R., Hua, R., Wang, W., Xu, S., & Luo, X. (2022). Automatic segmentation of white matter hyperintensities in routine clinical brain MRI by 2D VB-Net: A large-scale study. *Frontiers in Aging Neuroscience*, *14*. https://doi.org/10.3389/fnagi.2022.915009