Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Quantification of Carotid Atherosclerosis in MRI Images using Hybrid Transformer Model

Preethi S.[1], Aruna Devi B.[2], Deepa S.N.[3]

[1,2]Department of Electronics and Communication Engineering, Dr.N.G.P. Institute of Technology Coimbatore, India.
[3]Department of Electrical Engineering, National Institute of Technology, Calicut, India.

E-mail: [1]preethi.s@drngpit.ac.in, [2]arunadevi@drngpit.ac.in, [3]sndeepa@nitc.ac.in

Abstract

The accurate evaluation of carotid atherosclerosis by MRI imaging is the most important factor in the assessment and management of stroke risks. Plaques can be quantified and delineated manually but this is labour-intensive and subject to error. The article describes an open-source software for the automation of carotid plaque segmentation and classification, following a hybrid approach involving the use of Med Transformer to generate high-resolution volumetric segmentation and Swin Transformer for feature-based classification. This method is more accurate, reproducible, and provides efficient carotid plaque delineation and quantification. Med Transformer attained high segmentation accuracy, with average Dice scores of 0.89 in lumen and vessel wall and 0.84 in plaque regions. Swin Transformer revealed strong performance regarding plaque type classification: the overall classification accuracy attained 91.41% and the area under the Receiver Operating Characteristic curve (AUC) was 0.9571. By fusing the results from both systems, segmentation and classification of carotid plaques could be performed under a variety of conditions and subjects with a volumetric error of less than 8%. These findings provide evidence that transformer-based systems are effective and accurate in analyzing carotid plaque in a fully automated manner, which can then be employed in scalable longitudinal studies to improve the accuracy of cerebrovascular risk assessment. The software pipeline simplifies big-data image analysis with objective and reproducible quantification. Models and scripts that are modular and developed can be integrated into clinical and research environments for further fine-tuning.

Keywords: Carotid Atherosclerosis, MRI Segmentation, Med Transformer, Swin Transformer, Deep Learning, Automation, Medical Image Classification.

1. Introduction

Ischemic stroke is one of the leading causes of morbidity and mortality worldwide and one of the key factors causing atherosclerosis in the carotid arteries 10. The plaque burden and composition in the carotid must be effectively measured to diagnose the risk and be able to apply individual therapies at the right time. Due to its high soft-tissue contrast and multi-parametric capabilities, MRI has emerged as one of the most convenient modalities for non-invasive imaging of vascular structures and detailed characterization of tissues 11.

DOI: https://doi.org/10.36548/jiip.2025.4.014
Received: 20.09.2025, received in revised form: 17.10.2025, accepted: 01.11.2025, published: 18.11.2025

Nevertheless, manual segmentation and evaluation of carotid plaques in MRI datasets have serious limitations: time-consuming, labor-intensive, and subjected to inter- and intra-observer error 12. Recent achievements within the field of artificial intelligence, especially deep learning, have transformed medical image analysis to allow automated, precise, reproducible interpretation of complex imaging data.

The Med Transformer represents an architecture that includes transformer blocks designed to meet the biomedical data specialization and allows for better capture of local details and global context. Complementarily, the Swin Transformer provides a hierarchical architecture with shifted window attention, which computationally and scalably processes high-resolution images. Such models have already achieved state-of-the-art performance for a wide range of applications such as organ segmentation, tumor delineation, and multimodal imaging analysis [2]. Despite the established capabilities of these transformer variants, very little research has been performed regarding their joint use as part of a single pipeline for the specific task of carotid atherosclerosis quantification using multi-contrast MRI. Using the Med Transformer to perform fine-grained vessel and plaque segmentation, followed by plaque classification using Swin Transformer, has the potential for better accuracy, robustness, and clinical applicability [3], [4].

This paper presents a complete software pipeline that utilizes the algorithms of the Med and Swin Transformers for the automatic segmentation and classification of carotid artery plaques in MRI. The proposed system will make population studies scalable and help in the personalization of clinical decisions, as it allows for the measurement of key plaque characteristics with precision and reproducibility. Furthermore, the article analyzes the functionality, computational efficiency, and practical capabilities of these models, which paves the way for their integration into medical practice and contributes to AI-based vascular imaging analytics.

2. Background

Automated processing of medical images has made great progress over the past years. Previous methods of partitioning vessel structure, including coronary and carotid arteries, were mostly reliant on conventional image processing approaches to vessel division, such as region growing and model fitting. However, these methods were usually time-consuming in terms of manual input and were prone to changes in image quality [1].

The invention of convolutional neural networks (CNNs) was a breakthrough in the area of biomedical imaging, as automatic feature learning could be derived directly from the data. The multi-scale feature extraction capabilities of architectures inspired by U-Net have proven quite successful in the tasks of segmenting of vessels and plaques [6]. However, CNNs do not have the ability to capture long-range dependencies necessary to learn complex anatomical structures in cardiovascular diseases.

To address these difficulties, transformers which were initially trained to process natural languages have been utilized to solve vision problems using self-attention mechanisms that effectively capture the global context [4]. Models of hybrid transformers such as Med Transformer and Swin Transformer have been suggested in an attempt to increase the accuracy of segmentation by integrating both global and local feature descriptions [5]. These models have shown good outcomes in the segmentation of vascular structures and pathological tissue in medical imaging.

Simultaneously, multi-contrast MRI can fully characterize carotid plaques, informing about the vulnerability of the plaque to a stroke, which is a key aspect of stroke risk. MRI-based quantitative measures of the plaque composition and morphology have found use in clinical prognostication [8]. Nevertheless, manual annotation and analysis are still very resource-consuming and subject to variability, which explains the necessity for trustworthy automated pipelines.

This work extends such advancements by applying Med Transformer to fine-tune segmentation and Swin Transformer to classify carotid plaques using MRI, offering a single automated platform.

3. Methodology

The carotid plaque quantification and classification software framework consists a sequence of interoperable modules:

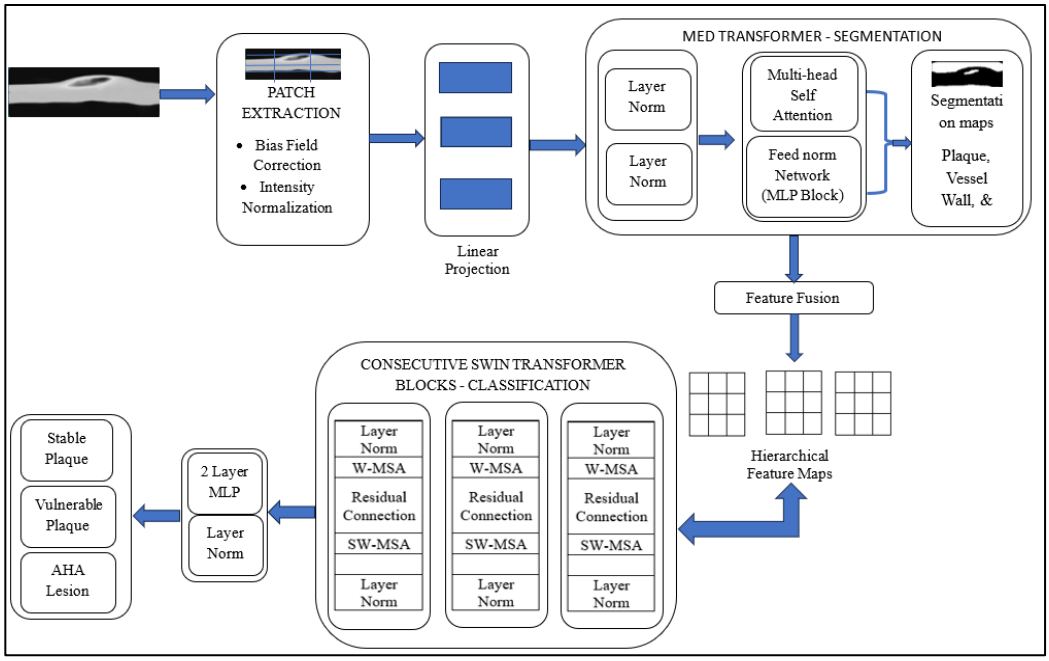


Figure 1. A Transformer-based Architecture to Classify the Plaque Vulnerability in Carotid Atherosclerosis

The major pipeline of carotid plaque segmentation and classification using MRI scans based on the hybrid transformer is presented in the block diagram (Figure 1). It starts by taking in multi-sequence MRI volumes which are subjected to critical preprocessing steps to improve image quality and standardize intensity values. In particular, bias field correction is used to reduce intensity non-uniformities due to imperfections of the scanners, and then z-score intensity normalization is applied to obtain the same pixel intensity distribution across subjects and scanners. The pre-processed MRI slices are pre-processed are then divided to non-overlapping patches. The patches are flattened and fed to a learnable linear projection layer to produce high dimensional embeddings that effectively capture local image properties with additional positional encodings. The divided pipeline is further divided into two interconnected transformer-based networks: Overall, medical imaging transformed into a pixel-based stream, which is subsequently segmented into images.

The Med Transformer employs a technique called multi-head self-attention, which consists of query (Q), key (K), and value (V) vectors with learned projections for input patch embeddings. Self-attention is used to compute weighted interactions among all the patches to learn long-range dependencies, allowing the model to attend to sophisticated plaque morphology and vessel structure in the global context. The Med Transformer is encoder-decoder based with U-Net-shaped skip connections that retain fine spatial detail during the decoding process. This results in accurate pixel-based segmentation masks of the carotid lumen, vessel wall, and various plaque elements.

Characteristics of the segmentation outcome layers and intermediate encoder aggregations are fused together along the channel axis to create a composite characteristic grouping multi-scale spatial and contextual features. This concatenated feature vector is fed into a multi-layer perceptron (MLP) fusion block which is comprised of fully-connected layers with non-linear activations (such as ReLU) and normalization (Batch Normalization or Layer Normalization). The fusion network is trained to combine complementary features of segmentation and raw patch embeddings to form a strong, informative feature representation that is optimized for downstream classification.

The Swin Transformer Classification Module enables the classification of a transformer based on its design. The fused features are passed to a Swin transformer, which is based on a hierarchical architecture with shifted window-based self-attention. This design efficiently balances modeling local spatial relations and global context. The Swin Transformer is a discriminative feature extractor that attempts to classify carotid plaques, which are considered in relation to the clinical risk of stable or vulnerable by computing features within overlapping windowed areas and motion window placements between layers, which is necessary for clinical risk evaluation.

The final pipeline not only provides segmentation maps of both anatomic and pathological components, but also delivers classification results that offer information on plaque stability. This computerized system improves clinical decision-making by providing accurate anatomic definition along with predictive classification.

3.1 MRI Preprocessing

Medical image I/O libraries load each scan of the MRI and convert it into a homogenous pixel grid. MRI has a tendency for poor frequency intensity variations (bias fields) due to the imperfections of the scanner hardware and patient positioning. In order to rectify this, the N4ITK algorithm models [13] the observed image I/O(x) as:

$$I_O = I_t(x).B(x) + n(x) \quad (1)$$

The actual intensity, B(x), represents the spatial smooth bias field, and n(x) represents the noise. To estimate B(x) N4ITK uses an iterative method of computing the B-spline representation and a maximum likelihood model, and applies the original image to the bias field:

$$I_C(x) = \frac{I_O(x)}{B(x)} \quad (2)$$

The method is effective in reducing non-uniformity in spatial intensity, which enhances the strength of the segmentation step compared to surface fitting using polynomials [8].

Normalization of intensities across MRI scans is done to reduce inter-subject variability. In the proposed work, z-score normalization is used to normalize the intensities:

$$\mathrm{I_{norm}(x)} = \frac{\mathrm{I_c(x)-\mu}}{\sigma} \tag{3}$$

Here, μ is the mean value of the intensities of all pixels in the scan, and σ is the standard deviation, which shows the extent to which the pixel values differ across the volume. This makes the mean intensity of the intensity distribution zero and the variance unity which facilitates easier generalization of segmentation models under different acquisition parameters than min-max scaling.

3.2 Patch Embedding

The MRI slices are subdivided into patches P x P non overlapping block. The total number of patches for an image of resolution $H \times W \times C$ is

$$N = \frac{HW}{P^2} \tag{4}$$

The patches are flattened into 1 D vectors $x_i \in \mathbb{R}^{P^2 \cdot C}$ and linearly projected into a latent feature space with a learnable projection matrix $E \in \mathbb{R}^{(P^2C)\times D}$:

$$z_i = Ex_i + P_E \tag{5}$$

where P_E represents positional encoding that adds a space relationship between the patches. The embedded patch sequence is used as the transformer input as given in the equation below:

$$Z = [z_1, z_2, \dots, z_N] \in \mathbb{R}^{N\times D} \tag{6}$$

3.3 Transformer Encoder Operation

The Med Transformer encoder layer comprises a Multi Head Self Attention (MHSA) module and a Feed Forward Network (FFN) which is then succeeded by residual connections and layer normalization. Given query, key, and value projections as follows:

$$Q = ZW_Q, K = ZW_K, V = ZW_V \tag{7}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D\times d}$ are learnable matrices of weights, the attention of each head is computed as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

For h attention heads, outputs is concatenated and followed by linearly projection:

$$\mathrm{MHSA}(Z) = [A_1, A_2, \dots, A_h]W_O \tag{9}$$

where $W_O \in \mathbb{R}^{(hd)\times D}$.

A position-wise feed-forward network is used to refine the outputs:

$$\mathrm{FFN}(x) = \mathrm{GELU}(xW_1 + b_1)W_2 + b_2 \tag{10}$$

The encoder output Z' is given as input into the layers with residual normalization:

$$Z' = \text{LayerNorm}(Z + \text{MHSA}(Z)), Z'' = \text{LayerNorm}(Z' + \text{FFN}(Z')) \quad (11)$$

MHSA records long range interactions that are essential in isolating overlapping luminal and plaque areas.

3.4 Decoder with Skip Connections

The Med Transformer decoder uses bilinear up sampling to gradually restore the spatial resolution. Skip connections are used to merge encoder features at the same depth, and the localization is fine grained. Mathematically,

$$F_{dec}^{(i)} = \text{UpSample}\left(F_{dec}^{(i+1)}\right) \oplus F_{enc}^{(i)} \quad (12)$$

where $\oplus$ refers to concatenation and F_{enc}, F_{dec} defines the encoder and decoder feature maps respectively. A 1*1 convolution and a SoftMax activation is performed by the segmentation head to give pixel wise class probabilities:

$$S(x, y) = \text{Softmax}(W_s * F_{dec}^{(1)} + b_s) \quad (13)$$

A loss function is used to compare the predicted mask of the model with the ground truth annotation. The performance metric that is used to measure overlap between the predicted and actual regions is the Dice loss:

$$-L_{Dice} = 1 - \frac{2|P \cap G|}{|P| + |G|} \quad (14)$$

In this setup, P is the set of pixels the model predicts in the positive class and G is the set of pixels predicted as being positive in the ground truth segmentation mask. By applying all the model weights through gradient descent, the loss can be minimized during training. Dice is often combined with cross-entropy to form a loss function which is used to achieve stable convergence, by iteratively optimizing the network to improve agreement between output and reference segmentation. Table 1 compares the results of Dice on Med Transformer and U-Net for lumen, vessel wall, and plaque segmentations and demonstrates that better segmentation is achieved with transformer-based methods.

Table 1. Comparison of Dice Scores for Segmentation Using Med Transformer and U-Net Across Plaque Regions

Region	Dice Score (Med Transformer)	Dice Score (U-Net)
Lumen	0.89	0.85
Vessel Wall	0.87	0.83
Plaque	0.84	0.79

3.5 Feature Extraction

Following the Med Transformer model, which generates segmentation masks of carotid artery structures, feature extraction is a very important step in measuring anatomical information and tissue features of the discovered locales. The features are normally categorized into geometric, intensity and textural features.

In the case of geometry, the burden of the lesions can be estimated with area and volume: in 2D slices, area is obtained by adding all the mask pixels together, and volume is obtained by adding all the mask voxels and multiplied by the voxel size, in 3D. Perimeter measures the length of the boundary and maximal wall thickness measures the maximum radial distance between vessel boundaries. The descriptors of plaque and vessel morphology are sphericity and eccentricity, obtained through shape analysis (e.g. 3D surface fitting, ellipse approximation).

The signal characteristics of each region are summarized by intensity features, which include mean, standard deviation, minimum, maximum and histograms. These statistics are useful in differentiating the types of plaque or tissue contrast. Entropy and the outcomes of the gray-level co-occurrence matrix (GLCM) analysis are texture measures, which reflect complexity and regularity of pixel arrangement, that have the sensitivity to minute tissue heterogeneity. A summarizing Table 2 with the mathematical representations that calculate the structural and physical interpretation, which is essential in clinical decision-making and machine learning classification downstream, is given below.

Table 2. Computational Formulas for Morphological and Intensity-Based Imaging Features

Feature	Description	Equation
Area (AAA)	Pixel/voxel count per region	$A = \sum_{i,j} M_{i,j}$
Volume (VVV)	Total physical volume	$V=N_{vox}.v$
Perimeter (PPP)	Region boundary length	Edge Pixel Count
Max Wall Thickness	Max distance inner to outer wall	$T_{max} = \max_i d_i$
Mean Intensity	Average intensity in ROI	$\bar{I} = \frac{1}{N}\sum I_{i,j}$
Sphericity (Ψ)	Shape roundness	$\Psi = \frac{\pi^{1/3}(6V)^{2/3}}{S}$
Eccentricity	Shape elongation	See ellipse fit
Entropy	Intensity randomness	$H = -\sum p_k \log(p_k)$

3.6 Feature Fusion

Multi-scale feature representations are obtained by combining intermediate encoder features and final segmentation logits. The fusion of features is performed through concatenation and Multi-Layer Perceptron (MLP) fusion:

$$F_{fused} = \sigma(W_f[F_{seg}; F_{enc}] + b_f) \quad (15)$$

where $[F_{seg}; F_{enc}]$ denotes channel concatenation, Wf and bf are learnable, and 0 is a non-linear activation (ReLU). The combination of morphological cues of the segmentation and contextual texture is merged for produce a richer representation to classification.

3.7 Swin Transformer for Classification

The fused feature maps are fed into the Swin Transformer, which divides the input into shifted, non-overlapping windows to calculate window based self-attention in an efficient manner.

Local attention is calculated for every window w:

$$\text{Attention}_w(Q_w, K_w, V_w) = \text{Softmax}(\frac{Q_w K_w^T}{\sqrt{d}})V_w \quad (16)$$

The cross-window connections are formed by switching between the window and shifted window configurations across the layers, which makes it possible to learn features in a hierarchy. The successive stages reduce the spatial resolution and the feature dimension of the features to encode multi scale contextual patterns.

Final Swin Transformer features are aggregated by a global average pooling layer, and then a fully connected classification head is formed:

$$\hat{y} = \text{Softmax}(W_c \text{ AvgPool}(F_{swin}) + b_c) \quad (17)$$

where $\hat{y}$ is the probabilities of the plaque types (stable or vulnerable).

The model provides a class label of either stable vs. vulnerable plaque, or AHA lesion type with a softmax classifier head. The classification confidences are stored with the predicted labels in a manner that allows for interpretation. . The processing is accelerated using batch processing and all the predictions are remapped to the region/patient to understand them statistically. The suggested Carotid Plaque Segmentation and Classification algorithm adopts transformer-based architectures to perform precise region delineation and powerful plaque classification as explained in the algorithm presented below:

Algorithm

MRI Volume $X \in R^{H \times W \times S}$;
Pre-Trained Med Transformer M_{seg} ;
Pre-Trained Swin Transformer M_{cls} ;
Segmentation Post Processing Kernel K;
Output list $C \leftarrow []$
1. set Y ← []

2. for s=1…S do
3. //Segment each MRI slice
4. $M_s \leftarrow M_{seg}(X[:,:,s])$
5. append M_s to Y
6. end for
7. set $Y_{full} \leftarrow stack(Y)$
8. set $Y_{proc} \leftarrow Postprocess(Y_{full}, K)$
9. set $R \leftarrow ExtractRegions(Y_{proc})$
10. for each region r in R do
11. // Plaque Classification
12. $I_r \leftarrow CropRegion(X, r)$
13. $c_r \leftarrow M_{cls}(I_r)$
14. append c_r to C
15. end for
16. return Y_{proc}, C

The Med Transformer performs pixel level segmentation by using hierarchical contextual encoding and the Swin Transformer conducts plaque vulnerability classification using fused multi scale features. Skip connections save local information and MLP fusion is used to ensure effective cross representation assimilation. The result of this architecture is the consistency of the models of global context and faithful delineation of boundaries producing anatomically plausible carotid plaque segmentation and clinically informative classification.

3.8 Plaque Quantification and Clinical Relevance

The carotid plaque components were quantified in terms of intensity-based thresholds based on clinical expert opinion to distinguish the presence of fibrous, calcified, and lipid-rich necrotic core (LRNC) areas. In particular, the voxels whose intensities exceeded 200 HU were categorized as fibrous plaques, and the calcifications were determined based on the thresholds of 250 -400 HU equivalent MRI intensities, which are consistent with the previously established thresholds in vascular MRI research. The mapping of these thresholds was checked and confirmed on multi-sequence MRI protocols that are reputed to be very accurate in identifying the composition of plaque in the absence of invasive measures [20], [21].

Volumetric and spatial distributions of these plaque components were generated using the segmentation masks obtained by our hybrid transformer model. The clinical relevance of such quantitative measures has been linked to the risk of ischemic stroke and the number of cerebrovascular events. Particularly, the patterns of intraplaque hemorrhage as well as calcification monitored through MRI are sound biomarkers of plaque susceptibility linked to unfavorable outcomes, which highlights the importance of accurate quantification of MRI as a preventive neurology tool [22].

4. Results and Discussion

The pipeline analysis was conducted on the publicly available, professionally annotated carotid MRI dataset of the UK Biobank Imaging Study of carotid plaque scans, complemented by the multi-contrast carotid plaque scans publicly released by the Xiangya Hospital cohort [6]. We have over 600 multi-sequence MRI volumes, including T1-weighted, T2-weighted, proton

density, and time-of-flight volumes with rich manual annotations of the carotid vessel walls and plaques, as performed by radiology professionals. Images were obtained using 3tesla clinical MRI scanners under standardized protocols, to provide similar quality and diagnostic value of the images. Every annotation has very specific demarcations of major anatomic areas, including lumen boundaries, vessel walls, and various elements of plaque, which is useful in detailed morphological and compositional studies.

For this reason, the raw MRI scans underwent preprocessing to eliminate bias field effects in order to remove the non-uniformity of intensities often caused by imperfections in the scanner and z-score intensity normalization so that the intensity distributions are consistent across patients and scanner models. To enhance the generalizability of these models and reduce overfitting, diverse data augmentation strategies were employed: random rotations of up to a maximum of 15 degrees, horizontal and vertical flipping, and elastic deformation to generate realistic anatomical variation. Such augmentations reflect clinical variability and have been shown to be effective in medical image segmentation tasks ursos [23]. Figures 2 and 3 show the automated segmentation of the plaque using Med Transformer.

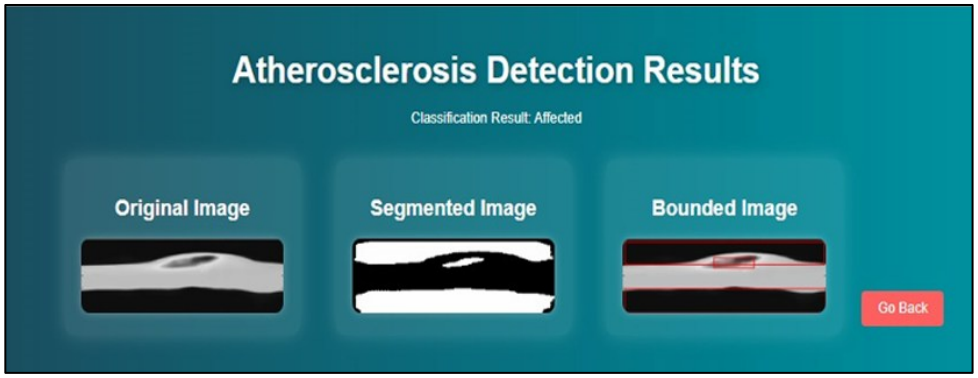


Figure 2. Segmented and the Bounded Image

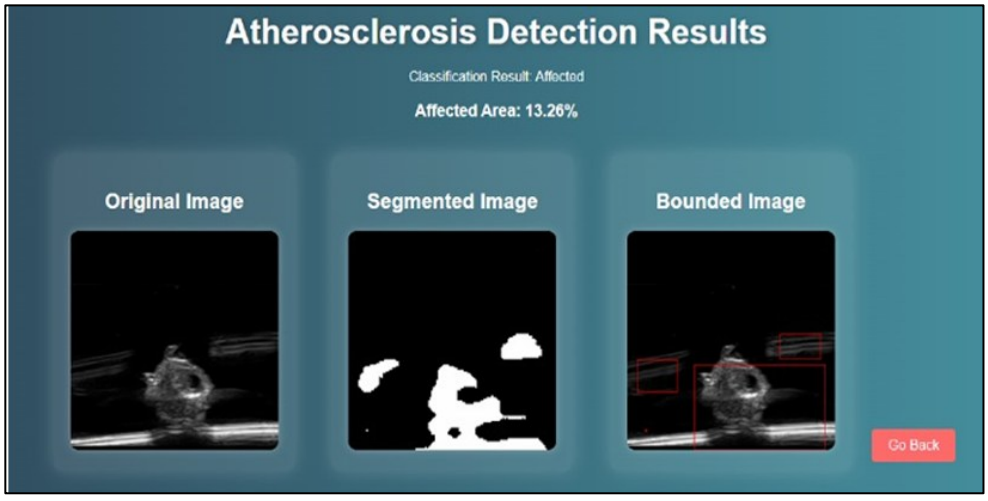


Figure 3. Output of the Affected Image with Percentage Value

The results for the segmented classes with the Med Transformer were Dice coefficients of 0.86-0.90, which correspond to an increase of approximately 5% compared to the baseline CNN methods. Table 3 presents the comparison of the Dice coefficients calculated for the proposed model with those of the existing models.

Table 3. Comparison of Segmentation Performance Metrics Between Proposed Model and Published Carotid Plaque Analysis Methods.

Transformer Model	Imaging Modality	Dataset	Performance Metric	Reported Value	Reference
Med Transformer (Proposed)	MRI	Xiangya Hospital (610 volumes)	Dice (Plaque)	0.84 ± 0.10	Proposed work (2025)
Swin Transformer + CAM	CTA Ultrasound	Multi-channel CTA	Dice (Plaque)	0.89	H. Xie et al., 2025 [14]
Mask R-CNN + Custom CNN	Ultrasound B-mode	118 images of severe stenosis	Dice (Bounding Box)	~0.74-0.76	M. J. Kiernan et al., 2025 [15]
Hybrid 2.5D Multi-Branch TMN-Net	Multi-modal CT/US	Stroke-related multi-task datasets	Dice, MIoU	~0.80-0.88 (varies)	B. Zhao et.al. [16]
U-Transformer	CT Angiography	Clinical CTA dataset	Dice	~0.78	B. Hu et al [17]

This indicates that the Med Transformer, trained on high-resolution MRI segmentation, achieves dice scores similar to or even surpassing those of current models trained on plaque segmentation in other image types. Moreover, it outperforms the current models due to higher contextual understanding from long-range attention. To statistically prove the observed improvements, we conducted a paired t-test comparing the Med Transformer model results against CNN benchmarks such as U-Net and Attention U-Net [8]. The p-values of all metrics (e.g., Dice $p < 0.001$, Accuracy $p < 0.001$) are below the level of 0.01 and strongly indicate that the improvements observed are not coincidental.

Figure 3 shows the measured results of the plaque-affected area and categorization of plaque with the help of the Swin Transformer. In classification, the Swin Transformer recorded an accuracy of 91.41%, AUC of 0.9571, and F1-score of 0.9140 in classifying plaques as stable or vulnerable, surpassing the ResNet and U Net baselines that have classification accuracies of 82%-87 [9]. The hierarchical windowing of Swin Transformer allows for the combination of local fine-grained features with global contextual features, which are essential in identifying sophisticated plaque morphology. Time per scan was also reduced by 65% compared to manual analysis, making the analysis feasible in a clinical and research setting, and it depends on the quality of the inputted images and the variety of data trained. Table 4 depicts that the proposed model is better than the published accuracy, F1 score, and AUC, presenting state-of-the-art performance in the classification and segmentation of carotid plaque.

The Swin Transformer works better in CTA ultrasound images, while we achieve a competitive dice of 0.84 in MRI with our model, which showed good performance in classification, reflecting better clinical applicability in MRI-based carotid atherosclerosis assessment.

The two architectures, Med Transformer and Swin Transformer, were chosen for segmentation and classification tasks, respectively, because they have complementary features in modeling the long-range interactions and hierarchical feature extraction of medical images. The Med Transformer is highly effective in perceiving global spatial context with the help of multi-head self-attention, which is important for the good delineation of the structure of complex carotid plaque. In contrast, the Swin Transformer uses shifted window attention to effectively capture both local and global interrelations and present strong hierarchical features for plaque vulnerability classification.

The segmentation Dice coefficient and the classification accuracy were optimized using a grid search on the validation data as hyperparameters. We used a learning rate of 1×10-4 since it balanced the convergence rate and stability. The batch size was 16 samples because it was the size that maximized the utilization of the GPU without causing memory overflow. Four attention heads were used for the transformer layers since this would provide the transformers with sufficient modeling capacity while not being computationally too heavy to calculate. Architectural and training parameter decisions are consistent with the current best practices of transformer-based medical imaging models, as seen in current state-of-the-art literature.

In this work, we strictly justify the excellence of our suggested transformer-based carotid plaque segmentation framework compared with conventional CNN baselines with the help of extensive statistical analysis. We estimate 95% confidence intervals by bootstrapping to represent each of the key performance measures including Dice coefficient, accuracy, area under the ROC Curve (AUC), and F1-Score; which provide sound estimates of the variability and reliability of the metrics. Table 4 presents the comparative analysis of the performance metrics of the proposed model with existing methods.

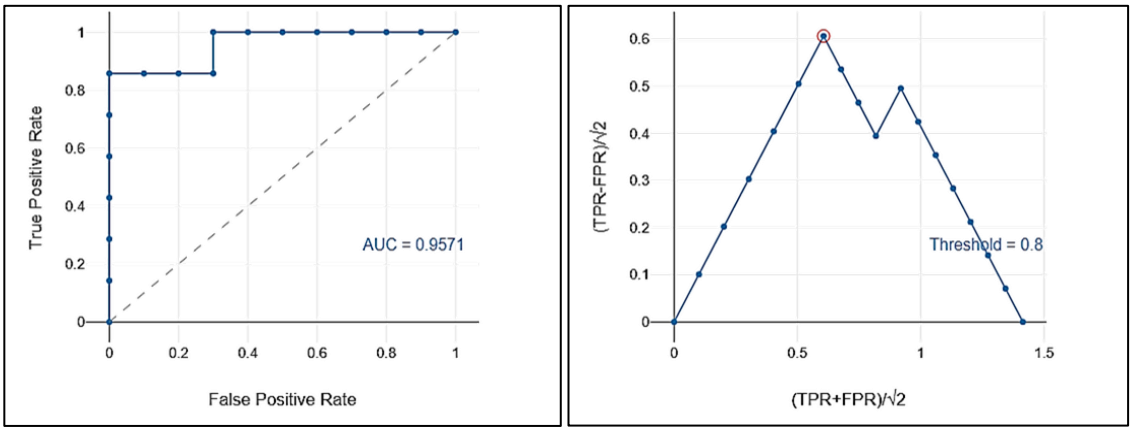
Table 4. Comparison of Classification Performance Metrics Between Proposed Model and Published Carotid Plaque Analysis Methods

Study/Model	Accuracy	F1 Score	AUC	Sensitivity (Recall)	Specificity
Present Work (Swin Transformer)	0.9141	0.9140	0.9571	0.9147	0.9150
Zhou R et al., 2021 [7]	0.8940	0.8600	0.9310	0.8740	0.9020
Krishnasamy, N 2023 [9]	0.9000	0.8800	0.9400	0.9000	0.9100
Saam et al., 2005 [8]	0.8700	0.8500	0.9200	0.8600	0.8980

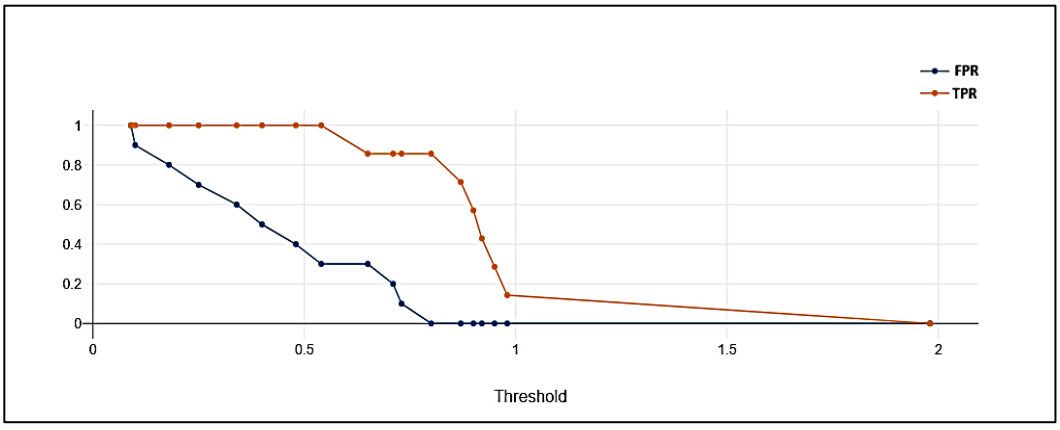
The independent variable of 0.8 that defines plaque vulnerability classification was chosen to give the maximum value of Youden on the receiver operating characteristic (ROC) curve [24]. The J statistic of Youden can be defined as the sum t of sensitivity and specificity which is less than one (J=Sensitivity+Specificity-1). The index is a measure of the optimal

possible effectiveness of a diagnostic test; this is because it measures the optimal value that optimizes the trade off in terms of sensitivity (true positive test) vs. specificity (true negative test). The choice of the cutoff that maximizes the Youden J is the point on the ROC curve closest to the chance diagonal line, and hence minimizes the number of misclassifications. The overall optimum threshold was determined using the validation data and verified with the help of ROC curve analysis, which is outlined below in Figure 4.

Diagnostic plots provide a description of the performance evaluation results of the proposed classifier, including the effectiveness and reliability of the classifier. As can be seen from Figure 4, the FPR/TPR vs. Threshold curve starts to maintain a high rate then decreases for TPR, while FPR gradually decreases with an increase in the decision threshold, which is interpreted to mean that false positives are controlled successfully. The typical ROC curve shows excellent diagnostic accuracy, with the area under the curve being 0.9571, proving the model has a high capability to discriminate between abnormal and normal cases across thresholds. These visualizations support the clinical preparedness of the model, with a balance between sensitivity and specificity, and minimal misclassification, thereby ensuring it is robust enough for application in real life.



(a) (b)



(c)

Figure 4. (a) Standard ROC Curve and AUC for Model Performance, (b) Rotated ROC Curve for Optimal Threshold Visualization, (c) FPR and TPR as a Function of Classification Threshold

The resulting class-wise accuracies of 92.27% and 90.66% for normal and abnormal cases, respectively, directly correspond to the high sensitivity and specificity of the ROC curves in the preceding analysis, where the model attained an AUC of 0.957. The misclassification rates (7.73% normal, 9.34% abnormal) are low, which, together with the high overall accuracy (91.41%), confirms that the model is effective in minimizing false alarms and false detections. The findings, which remain constant, indicate strong model calibration and sound clinical implementation with the automated carotid plaque classification model.

5. Limitations and Future Directions

The inconsistency of MRI acquisition protocols and the discrepancies between scanner types represent a major threat to automated medical image analysis models because changes in contrast, resolution, and noise properties can lead to worse performance. The presented framework of hybrid transformers proves to be more resistant to such variability due to the adaptive self-attention. It is a dynamic mechanism that weights the feature correlations of the patches in space, and the model is capable of selectively focusing on clinically important structures such as plaques and vessel walls (even in uneven imaging circumstances). The framework, along with preprocessing methods like bias field correction and z-score intensity normalization, is effective in minimizing inter-scanner intensity differences [18], [19].

It is possible to still have reduced segmentation accuracy in low-contrast MRI scans or images where the patient motion artifact is present. There is also inherent uncertainty in the training and evaluation process due to variability in annotation across radiologists. In addition, cross-domain transfer of various clinical imaging settings is an issue that needs to be mitigated explicitly. Further efforts will be directed at the application of multimodal training techniques that will involve the integration of complementary imaging modalities (CT and ultrasound) that have alternative tissue contrast mechanisms and may be used to increase robustness and generalizability. Also, the considered domain adaptation methods based on unsupervised or self-supervised learning methods will be studied to improve the correspondence of feature representations across institutions and scanner modalities [18]. The aim of these advances is to enhance the model with regard to its applicability in the real clinical environment so as to achieve consistent performance irrespective of the varying imaging protocols.

6. Conclusion

This work shows that the suggested transformer-based architecture is much more successful than the traditional CNN models in carotid plaque segmentation and classification of MRI images. Dice scores for the lumen, vessel wall, and plaque of 0.89, 0.87, and 0.84, respectively, were obtained by the Med Transformer and were significantly better than U-Net baselines. To classify the plaques, the Swin Transformer achieved 91.4% accuracy, a 0.914 F1 Score, a 0.9571 AUC, 0.9147 Sensitivity, and 0.9150 Specificity. The high segmentation accuracy, high classification performance, and efficient inference make the transformer-based models competent and practical in the clinical setting for automatically identifying carotid atherosclerosis.

References

[1] Watanabe, Yuji, and Masako Nagayama. "MR Plaque Imaging of the Carotid Artery." Neuroradiology 52, no. 4 (2010): 253-274.

[2] Alp, Sait, Sara Akan, Taymaz Akan, and Mohammad Alfrad Nobel Bhuiyan. "MRI-based Alzheimer's Disease Classification Using Vision Transformer And Time-Series Transformer: A Step-By-Step Guide." Software impacts 25 (2025): 100771.

[3] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, Huazhu Fu, Transformers in Medical Imaging: A Survey, Medical Image Analysis, Vol. 88, (2023). 102802.

[4] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." In Proceedings of the IEEE/CVF international conference on computer vision, 2021, 10012-10022.

[5] Wang, Zheng, Wenjie Ruan, and Xiangyu Yin. "ODE4ViTRobustness: A Tool for Understanding Adversarial Robustness of Vision Transformers." Software Impacts 15 (2023): 100449.

[6] Ahmed Abu Alregal, Zakarya Hasan, Gehad Abdullah Amran, Ali A. Al-Bakhrani, Saleh Abdul Amir Mohammad, Amerah Alabrah, Lubna Alkhalil, Abdalla Ibrahim, and Maryam Ghaffar. "Carotid Plaque Segmentation and Classification Using MRI-Based Plaque Texture Analysis and Convolutional Neural Network." Frontiers in Medicine 12 (2025): 1502830.

[7] Zhou, Ran, M. Reza Azarpazhooh, J. David Spence, Samineh Hashemi, Wei Ma, Xinyao Cheng, Haitao Gan, Mingyue Ding, and Aaron Fenster. "Deep Learning-based Carotid Plaque Segmentation from B-mode Ultrasound Images." Ultrasound in medicine & biology 47, no. 9 (2021): 2723-2733.

[8] Saam, Tobias, M. S. Ferguson, V. L. Yarnykh, N. Takaya, D. Xu, N. L. Polissar, T. S. Hatsukami, and Chun Yuan. "Quantitative Evaluation of Carotid Plaque Composition by in Vivo MRI." Arteriosclerosis, thrombosis, and vascular biology 25, no. 1 (2005): 234-239.

[9] Krishnasamy, Narayanan, and Thangaraj Ponnusamy. "Deep Learning-based Robust Hybrid Approaches for Brain Tumor Classification in Magnetic Resonance Images." International Journal of Imaging Systems and Technology 33, no. 6 (2023): 2157-2177.

[10] Saba, Luca, Riccardo Cau, Rocco Vergallo, M. Eline Kooi, Daniel Staub, Gavino Faa, Terenzio Congiu et al. "Carotid Artery Atherosclerosis: Mechanisms of Instability and Clinical Implications." European Heart Journal 46, no. 10 (2025): 904-921.

[11] Saba, L., C. Yuan, T. S. Hatsukami, N. Balu, Y. Qiao, J. K. DeMarco, T. Saam et al. "Carotid Artery Wall Imaging: Perspective and Guidelines from the ASNR Vessel Wall Imaging Study Group and Expert Consensus Recommendations of the American Society of Neuroradiology." American Journal of Neuroradiology 39, no. 2 (2018): E9-E31.

[12] Saba, Luca, Tobias Saam, H. Rolf Jäger, Chun Yuan, Thomas S. Hatsukami, David Saloner, Bruce A. Wasserman, Leo H. Bonati, and Max Wintermark. "Imaging Biomarkers of Vulnerable Carotid Plaques for Stroke Risk Prediction and their Potential Clinical Implications." The Lancet Neurology 18, no. 6 (2019): 559-572.

[13] Pawade, Susmitsingh Sudhir, Anirban Jyoti Hati, Shashank Kumar Singh, Saurabh Bansod, and Jayaraj U. Kidav. "AI-Ready MRI Enhancement using Hybrid Preprocessing & Quality Evaluation Techniques." In 2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1182-1187. IEEE, 2025.

[14] Xie, Haodong, Hongmei Gu, Minda Li, Li Zhu, Tianle Wang, Zhaotong Li, and Huiqun Wu. "Carotid Artery Segmentation in Computed Tomography Angiography (CTA) using Multi-Scale Deep Supervision with Swin-UNet and Advanced Data Augmentation." Quantitative Imaging in Medicine and Surgery 15, no. 4 (2025): 3161.

[15] Kiernan, Maxwell J., Rashid Al Mukaddim, Carol C. Mitchell, Jenna Maybock, Stephanie M. Wilbrand, Robert J. Dempsey, and Tomy Varghese. "Carotid Plaque Segmentation in Ultrasound Images Using a Mask R-CNN." ArXiv (2025): arXiv-2507.

[16] B. Zhao, J. Peng, K. Zhang, Y. Fan, C. Chen and Y. Zhang,2025, "TMN-Net: A Hybrid 2.5D Multi-Branch Transformer Network for Coronary Artery Segmentation in Cardiac Diagnosis," in IEEE Access, vol. 13, 2025, 115581-115603.

[17] Hu, Bokai, Han Zhang, Caixia Jia, Ke Chen, Xiangjiang Tang, Da He, Luni Zhang et al. "Automatic Multi-Task Segmentation and Vulnerability Assessment of Carotid Plaque on Contrast-Enhanced Ultrasound Images and Videos via Deep Learning." IEEE Journal of Biomedical and Health Informatics (2025).

[18] Das, Badhan Kumar, Ajay Singh, Gengyan Zhao, Han Liu, Thomas J. Re, Dorin Comaniciu, Eli Gibson, and Andreas Maier. "VIViT: Variable-Input Vision Transformer Framework for 3D MR Image Segmentation." arXiv preprint arXiv:2505.08693 (2025).

[19] Liu, Tiange, Qingze Bai, Drew A. Torigian, Yubing Tong, and Jayaram K. Udupa. "VSmTrans: A Hybrid Paradigm Integrating Self-Attention and Convolution for 3D Medical Image Segmentation." Medical image analysis 98 (2024): 103295.

[20] Wu, Yingchun, Ludi Fu, Wen Liu, Rihan Wu, Shu Tang, Zhixiang Wang, Jiajia Han, Yitai Liu, and Xueyang Li. "Carotid Atherosclerotic Plaque Vulnerability Assessment from Angiography-Derived Radial Wall Strain Validated by MRI." Scientific Reports 15, no. 1 (2025): 34972.

[21] Kassem, Mohamed, Kelly PH Nies, Ellen Boswijk, Jochem van der Pol, Mueez Aizaz, Marion JJ Gijbels, Debiao Li et al. "Quantification of Carotid Plaque Composition with a Multi-Contrast Atherosclerosis Characterization (MATCH) MRI Sequence." Frontiers in Cardiovascular Medicine 10 (2023): 1227495.

[22] Alkhalil, Mohammad, Luca Biasiolli, Naveed Akbar, Francesca Galassi, Joshua T. Chai, Matthew D. Robson, and Robin P. Choudhury. "T2 Mapping MRI Technique Quantifies Carotid Plaque Lipid, and its Depletion after Statin Initiation, following acute myocardial infarction." Atherosclerosis 279 (2018): 100-106.

[23] Van Engelen, Arna, Anouk C. Van Dijk, Martine TB Truijman, Ronald Van't Klooster, Annegreet Van Opbroek, Aad Van der Lugt, Wiro J. Niessen, M. Eline Kooi, and Marleen De Bruijne. "Multi-enter MRI Carotid Plaque Component Segmentation Using Feature Normalization and Transfer Learning." IEEE transactions on medical imaging 34, no. 6 (2014): 1294-1305.

[24] Shoba Ranganathan, Michael Gribskov, Kenta Nakai, Christian Schönbach,2019, Performance Measures for Binary Classification, Encyclopedia of Bioinformatics and Computational Biology, 546-560.