IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2026, accepted 30 April 2026, date of publication 6 May 2026, date of current version 14 May 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3690923

APPLIED RESEARCH

Performance Assessment of AI-Assisted Thrombus and Vessels Segmentation in Acute Ischemic Stroke Patients in Emergency Settings

GIADA MINGHINI[1], LUIGI BELLINI[2], ARMANDO UGO CAVALLO[3], MARIAFRANCESCA TRULLI[2], SIMONE DI CANDIA[2], FRANCESCO GIUSEPPE GARACI[2], ANDREA MIOLA[1], ENRICO CALORE[4], SEBASTIANO FABIO SCHIFANO[1,4], CRISTIAN ZAMBELLI[1], (Member, IEEE), AND VALERIO DA ROS[2]

[1]Università degli Studi di Ferrara, 44121 Ferrara, Italy
[2]Dipartimento di Biomedicina e Prevenzione, Università degli Studi di Roma "Tor Vergata", 00133 Rome, Italy
[3]Istituto Dermopatico dell'Immacolata (IDI) IRCCS, 00167 Rome, Italy
[4]Istituto Nazionale di Fisica Nucleare (INFN), 44121 Ferrara, Italy

Corresponding author: Giada Minghini (giada.minghini@unife.it)

This work was supported in part by Italian National Institute for Nuclear Physics (INFN) for granting access to the COKA computing systems used for training the models considered in this study; and in part by the PRIN 2022 Project Finanziato dall'Unione europea-Next Generation EU, Missione 4 Componente 1, under Grant CUP F53D23002370006.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the "Comitato Etico Lazio 1" and performed in line with the Declaration of Helsinki.

ABSTRACT Large Vessel Occlusions (LVOs) in the M1 segment are among the most frequent and severe causes of Acute Ischemic Stroke (AIS). In this study, we investigate the use of Artificial Intelligence (AI) to detect these occlusions through Deep Neural Networks (DNN). Although AI-assisted segmentation shows considerable promise in medical imaging, the trade-offs between functional and non-functional performance metrics, particularly in clinical and emergency settings, remain insufficiently explored. To address this gap, we adopt an interdisciplinary approach that integrates medical and engineering perspectives to detect and characterize LVOs in the M1 segment of the middle cerebral artery (MCA). We train a state-of-the-art nnU-Net model on a recent NVIDIA GH200 system using late-phase CT angiography (CTA) images from a retrospective cohort of 198 patients. Our results indicate that segmentation performance is influenced by anatomical factors such as collateral circulation and thrombus location, and predictive quality is not strictly dependent on the specific DNN architecture. The model achieves a sensitivity of 0.93 in identifying the thrombus–vessel interface. Furthermore, analysis of non-functional metrics demonstrates that inference costs can be limited to approximately €0.08 per 1000 patients. These findings support the deployment of GPU-based systems in on-premise hospital environments, providing neuroradiologists with an effective assistive and decision-support tool for the detection of LVOs and the planning of surgical interventions in AIS.

INDEX TERMS Artificial intelligence, image segmentation, cerebrovascular disorders, computed tomography angiography, deep neural networks, ischemic stroke.

I. INTRODUCTION

The adoption of Artificial Intelligence (AI) in medicine has grown rapidly in recent years, opening new opportunities to improve healthcare delivery by supporting physicians in diagnosis and treatment planning through advanced data

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

analysis techniques applied to Electronic Medical Records (EMR) [1]. Medical imaging represents one of the most informative components of EMR data, with techniques such as Computed Tomography (CT) and Magnetic Resonance (MR) playing a central role in modern diagnostic workflows [2]. Among the many image analysis tasks, image segmentation is fundamental for identifying anatomical structures or pathological regions using mathematical models such as edge detection or thresholding techniques [3]. Traditionally, segmentation is performed manually or with semi-automatic tools, making it highly dependent on operator expertise [4] and difficult to scale for large retrospective studies or high-throughput clinical environments [5].

To address these limitations, AI-assisted approaches based on Deep Neural Networks (DNN) have been widely adopted for medical image segmentation. Convolutional neural networks (CNNs), and more recently transformer-based architectures, have demonstrated strong performance in various segmentation tasks thanks to their ability to learn hierarchical representations directly from imaging data [6]. Furthermore, advances in High-Performance Computing (HPC) systems and specialized accelerators for deep learning [7] have significantly improved the feasibility of deploying these models in automated analysis pipelines [8].

Despite these technological advances, a notable gap remains between the rapid development of deep learning architectures and their effective integration into clinical practice. In the context of medical image segmentation, most studies primarily focus on functional performance metrics such as Dice coefficient, sensitivity, or accuracy, evaluated from a purely technical perspective. However, several critical aspects are often overlooked. First, segmentation performance may be fundamentally constrained by anatomical and physiological factors that are not captured by standard evaluation metrics. Second, the clinical interpretation and practical relevance of segmentation results are rarely discussed. Third, deployment-related aspects—including inference latency, computational requirements, and energy consumption—are seldom considered, even though they are essential for real-world adoption of AI systems in hospital environments.

This disconnect is particularly evident in the case of Large Vessel Occlusion (LVO) detection in acute ischemic stroke (AIS), a time-critical clinical scenario where rapid and accurate diagnosis is essential for effective treatment [9]. Computed Tomography (CT) is the most widely used imaging modality for AIS diagnosis due to its availability and fast acquisition time [10], [11]. In particular, CT Angiography (CTA) enables visualization of intracranial arterial structures through contrast enhancement, allowing clinicians to identify thrombi as interruptions of arterial blood flow. In some cases, thrombus visibility can be improved by exploiting collateral circulation (CC), which allows blood reperfusion distal to the occlusion via retrograde filling from leptomeningeal arteries [12], [13]. This compensatory vascular mechanism plays a key role in determining patient outcomes and guiding treatment strategies [12], [14], as therapeutic decisions may vary depending on the thrombus location and whether the occlusion involves the horizontal M1 segment alone or extends to its bifurcation [15].

In current clinical practice, AIS management typically begins with the manual localization of the thrombus responsible for the LVO and the assessment of its characteristics (e.g., location, length, and extension) from CT images [16]. However, this manual workflow conflicts with the highly time-sensitive nature of stroke treatment, where early intervention is crucial to improve patient outcomes [17]. Consequently, AI-assisted segmentation tools have the potential to significantly accelerate thrombus detection and characterization.

Several recent studies have explored DNN-based approaches for automated thrombus segmentation in the middle cerebral artery (MCA) [18], [19]. While these approaches report promising segmentation performance, most existing research focuses exclusively on functional metrics and does not investigate the underlying anatomical factors that may limit segmentation accuracy. Moreover, the literature rarely considers the non-functional characteristics of deep learning models, such as training cost, inference latency, and energy consumption. These aspects are essential for evaluating the Total Cost of Ownership (TCO) of AI-assisted medical systems and determining their practical feasibility in real-world clinical settings.

In this context, the aim of this work is to provide a comprehensive assessment of AI-assisted thrombus segmentation for AIS by jointly analyzing functional segmentation performance and non-functional deployment metrics. Specifically, the study investigates whether segmentation performance is primarily constrained by neural network architecture or by anatomical characteristics of the thrombus and collateral circulation, while also evaluating the computational cost and energy efficiency of different deep learning models.

The key contributions of this work are summarized as follows:

- Training and evaluation of multiple state-of-the-art nnU-Net (no-new-UNet) models [20] for AI-assisted processing of late-phase CTA images for thrombus identification and anatomical characterization, using a retrospective cohort of 198 patients.
- Comprehensive statistical analysis of functional segmentation metrics (including overlap- and spatial-based measures) to assess the limits of AI-assisted thrombus characterization. Our framework demonstrated measurable performance outcomes, achieving a 93% sensitivity in identifying the thrombus-vessel interface and proving that spatial-localization metrics like Surface Dice provide a 30% higher and more clinically relevant performance score for thrombus localization than traditional volumetric overlap (Dice scores).

TABLE 1. Summary of demographic and clinical characteristics of the cohort of 198 patients. The age is reported as the mean ± the standard deviation.

	Total	Percentage
Demographic		
Male	80	40.4%
Age	75 ± 13	
Risk factors		
Hypertension	142	71.7%
Atrial Fibrillation	95	47.9%
Hyperlipidemia	57	28.8%
Cardiac Disease	51	25.7%
Diabetes	40	20.2%
Smoking	32	12.6%
Neoplasia	15	7.6%
Recent TIA/Attacks	26	3.1%
None	16	8.1%
Carotid Stenosis	25	12.6%
Etiology		
Atherosclerosis	93	47.0%
Cardiac	80	40.4%
Unknown	18	9.1%
Other	7	3.5%
Occlusion laterality		
Right	99	50.0%
Left	99	50.0%

- Comparative evaluation of seven nnU-Net architectural variants [21] and transformer-based UNet models to determine whether segmentation performance is constrained by anatomical characteristics (e.g., collateral circulation status and thrombus location) rather than architectural design.
- First comprehensive characterization, to the best of our knowledge, of these DNN models on the NVIDIA Grace Hopper GH200 platform [22], measuring training and inference performance together with energy consumption to estimate the Total Cost of Ownership and assess the feasibility of AI-assisted segmentation in routine AIS emergency workflows. Our findings demonstrate that inference costs can be reduced to approximately €0.08 per 1000 patients, establishing a clear baseline for deploying these architectures in on-premises emergency hospital environments.

The remainder of this article is organized as follows: Section II reviews literature on LVOs and AI-assisted segmentation methods. Section III presents the features of the patient cohort and imaging acquisition protocol. Section IV describes the DNN models tested for image segmentation and the GH200 system used in the analysis. Section V presents the functional metrics of the DNN models and their anatomical limitations. Section VI discusses the challenges and main results of AI-assisted thrombus segmentation. Section VII merges the non-functional metrics of the DNN models running on the GH200 system with the functional ones. Lastly, Section VIII concludes this study.

TABLE 2. Summary of stroke outcome scores on the cohort of 198 patients. mRS means modified Rankin Scale, mTICI is modified thrombolysis in cerebral infarction, ASPECTS is alberta stroke program early CT score.

Stroke Outcome Scores	Total	Percentage
Three-months mRS		
mRS ≤ 2	57	28.8%
mRS > 2	105	53.0%
Unknown	36	18.2%
mTICI		
0–2a	35	17.8%
2b–3	163	82.2%
Pre-event ASPECTS		
0–6	19	12.3%
7–9	42	27.1%
9–10	92	59.4%
Not assessable	2	1.2%

II. RELATED WORKS

The application of AI in AIS encompasses a broad spectrum of tasks, including infarct area segmentation, hemorrhage detection, and clinical outcome prediction. Among these, LVO represents a common cause of AIS. Despite its clinical importance, fully automated AI-assisted thrombus detection methods for LVO remain limited in the current literature, highlighting inherent technical challenges.

Kobold et al. [23] used a U-Net for thrombus segmentation on MRI images, demonstrating the adaptability of encoder–decoder architectures across imaging modalities. Despite a 77% thrombi detection rate, the Dice score (0.42) indicated only moderate overlap with annotations, emphasizing the difficulty of the thrombus segmentation task.

Tolhuisen et al. [24] developed a dual Convolutional Neural Network (CNN) approach for non-contrast CT (NCCT) analysis: one network for brain hemisphere asymmetry detection and another for HAS identification. Researchers tested a neural network inspired by AlexNet, which featured 3D convolutional layers and ReLU activations, for detecting thrombi in 103 images. The model achieved a sensitivity of 0.86 and a specificity of 0.65.

In [25], Mojtahedi et al. developed an automated technique to detect thrombi by combining data from both NCCT and CTA scans. They created a two-encoder U-Net that uses a weighted-sum fusion technique to merge information from the two scan types. When tested on a dataset of 328 patients, their method achieved high accuracy (Surface Dice score of 0.78 and a Dice score of 0.62), and a low error rate (only 4% of the cases missed). This work demonstrates a successful dual-modality approach for accurate thrombus segmentation.

Zoetmulder et al. [26] introduced BL-UNet and Polar-UNet to address thrombus localization and segmentation, combining CTA and NCCT data from a multicenter cohort. The Polar-UNet is better suited to model the elongated geometry of thrombi, while the BL-UNet focuses on boundaries. However, the moderate Dice and localization precision

indicate that accurate delineation remains challenging even with heterogeneous multicenter data.

Liang et al. [27] proposed a Context Pooling Module to enhance segmentation by utilizing local and global context information on CTA scans. A further refinement network corrected the inaccuracies of the initial segmentation results, particularly at the level of the thrombus-tissue interfaces, obtaining a Dice score of 0.43 in 29 patients.

Soh and Rajapakse [28] proposed a U-Net Transformer architecture that combines convolutional layers for local feature extraction with transformer blocks to capture long-range spatial dependencies. The reported HD95 (12.63mm) suggests that boundary precision can still vary, potentially impacting volumetric measurements in clinical scenarios.

The literature exposes the challenge of achieving a high Dice score for small object detection. In McConnell et al. [21], Standard nnU-Net achieved a comparable Dice score for hepatic vessel (0.64 versus our 0.58) and a Dice score in colon cancer tumor segmentation (0.46 versus our 0.45). While nnU-Net is efficient for segmenting large, distinct organs such as the liver and pancreas, it is not effective for smaller and more complex anatomical structures within a vast and similar background space, as it happens in the brain.

In contrast to the main works in the literature, we deepen the insights into the relationships between functional and non-functional metrics of the systems running the AI models, and not only on the medical standpoints. We believe this is relevant to set a computing environment that can be installed in hospital premises, whose structural characteristics are very different from those of a computing center.

Berzoini et al. [29] focused on mapping U-Net and its variants to edge hardware. The work achieved a Dice score of 93.04%, while improving energy efficiency by 12.7× compared to standard GPU implementations. It explicitly charts the ''Energy-Accuracy'' Pareto front for different models.

The same research group investigated a framework for optimizing medical segmentation models for power-constrained environments in [30]. They demonstrate how 8-bit quantization maintains high Dice scores (>0.90) while reducing power adsorbed by over 60%, compared to full-precision models.

Dörrich et al. [31] relate the Dice score to energy consumption during the training process of a generic deep neural network for image segmentation. The outcome is that using reduced precision can reduce energy consumption during training by up to $40 - 50\%$, with a negligible drop ($< 0.5\%$) in Dice score.

Fornt et al. [32] explored a hardware-software co-design flow for segmentation tasks. By using binary segmentation and mixed-precision, they achieved an 11× increase in energy efficiency while retaining almost the same model accuracy compared to standard full-precision benchmarks.

Finally, in our work in [33] we showed that a U-Net model can run effectively on field programmable gate array (FPGA) devices using 8-bit integer computation, achieving a prediction accuracy of approximately 95% in Dice and 91% in IoU scores. These results are comparable to those measured when running the floating-point models on GPUs and CPUs. We also found that, considering the energy efficiency, the improvements are approximately a factor of 6.7× compared to the CPU, and 1.6× compared to the P100 GPU manufactured with the same technological process.

Despite these advances, several limitations remain in the current literature on AI-assisted thrombus segmentation. First, most studies primarily focus on improving segmentation architectures or introducing task-specific network modules, while evaluating performance exclusively through functional metrics such as Dice score, sensitivity, or surface distance. Second, the relationship between segmentation performance and anatomical characteristics of the thrombus, such as its location or the presence of collateral circulation, is rarely investigated. Third, existing works seldom consider the practical deployment of AI models in clinical environments, where computational efficiency, inference latency, and energy consumption are critical factors.

In this work, we address these limitations by proposing a comprehensive evaluation framework that jointly analyzes functional segmentation metrics and non-functional system-level characteristics. Specifically, we evaluate multiple state-of-the-art segmentation architectures under varying anatomical conditions and characterize their computational and energy requirements on a modern high-performance computing platform. This approach enables a deeper understanding of the practical and anatomical constraints affecting AI-assisted thrombus segmentation in AIS.

Moreover, we emphasize the need to consider the AI not only from the pure functional metrics standpoint, but rather as a system-wide exploration.

III. DATASET FEATURES

In this section, we highlight the clinical and demographic features of the dataset considered in this work, along with the strategies and protocols devised in imaging to build the cohort investigated with AI-assisted segmentation methods.

A. PATIENT COHORT

From January 2015 until December 2024, we analyzed patients admitted to the stroke emergency unit of Tor Vergata Hospital in Rome (Italy) with an AIS diagnosis. The study was conducted in accordance with the Declaration of Helsinki and received approval from the institutional review board (IRB). In many cases, informed consent could not be obtained, given the acute emergency setting of CTA image retrieval.

Out of 1480 patients with LVO accessing the stroke unit, we have selected 421 with occlusion on the M1 segment. The choice for this is twofold: i) M1 occlusion is the main cause of AIS, as reported in several studies [34], [35], and ii) they show more robust clinical evidence regarding the benefits of revascularization therapy [36]. A further selection has been done, removing those CTAs with corrupted images,

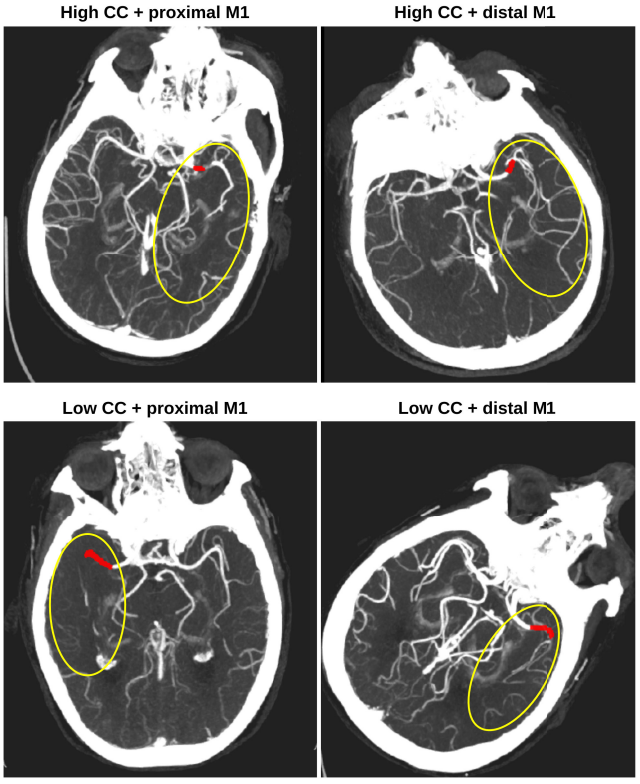


FIGURE 1. Stacked reconstructions of axial CTA slices illustrating representative cases stratified according to collateral circulation (CC), status (high or low), and occlusion location along the M1 segment of the MCA (proximal or distal M1). In the top-row the cases for high-CC with proximal M1 and distal M1 occlusions, while in the bottom-row those for low-CC with proximal M1 and distal M1 occlusions. The CC status is enclosed within the yellow oval, while the occlusion site is highlighted in red.

incomplete acquisition, and with large motion artifacts due to the emergency settings in which they have been acquired. Normal anatomical variations and minor imaging artifacts were retained to ensure that model performance reflects real-world clinical conditions. The final dataset is a cohort of 198 patients, statistically representative of the initial (421) population; 118 are women and 80 men, the mean age is 75 years ±13, and the right-sided occlusion occurs in 99 patients. The demographic and clinical characteristics of the cohort are summarized in Table 1, while Table 2 reports the stroke outcome scores.

The dataset was stratified based on two criteria: thrombus location and cerebral collateral status. Thrombus location encompassed both the proximal M1 segment and thrombi involving the MCA bifurcation (referred to as distal M1 for brevity). Collateral status was categorized using the Miteff scoring system, dividing patients into High CC (good Miteff score with visible major MCA branches distal to occlusion) and Low CC (moderate or poor Miteff scores) [37]. This approach resulted in four patient groups: High CC + proximal M1 ($n = 65$), High CC + distal M1 ($n = 62$), Low CC + proximal M1 ($n = 36$), and Low CC + distal M1 ($n = 35$). Details are shown in Table 3. We did not perform the Synthetic Over-Sampling Minority Technique

TABLE 3. Summary of the dataset partitioning strategy, detailing the distribution of the 198 patient volumes according to baseline collateral circulation (High vs. Low CC) and occlusion site (Proximal vs. Distal M1).

Cerebral Collateral (CC) Status	Proximal M1	Distal M1	Total Patients
High CC	65	62	**127**
Low CC	36	35	**71**
Total Patients	**101**	**97**	**198**

(SMOTE) to perfectly balance the four strata to avoid the generation of anatomical artifacts and biologically invalid synthetic geometries. This divides our dataset into four splits, whose examples are illustrated in Figure 1, displaying the CT scans stacked along the axial direction.

B. IMAGE ACQUISITION PROTOCOL

We acquired CTA images with a standard clinical protocol. After excluding intracranial hemorrhage with preliminary non-enhanced axial brain CT, we performed triphasic post-contrast acquisition. The arterial phase began $5 - 10$ s after automatically injecting $50 - 70$ mL of iodine contrast media (ICM) bolus at $4 - 5$ mL/s into an antecubital vein, capturing from the aortic arch to the brain apex. We acquired the venous phase 15 s after the arterial phase, and the late venous phase 30 s after the arterial phase. Both acquisitions covered the region from the skull base to the apex of the brain.

For optimal visualization of CC, we have built our dataset using only late-phase CTAs from a triphasic acquisition procedure. Other techniques, such for example, the time-resolved Maximum Intensity Projections (t-MIP) can not be used, because for a large fraction of patients (approximately 37%), one or more phases are missing or are qualitatively insufficient due to the emergency setting in which they have been acquired, making it impossible to build a large number of samples to efficiently train the AI-models.

Manual segmentations were performed using the Medical Imaging Interaction Toolkit (MITK) [38] with a freehand drawing tool to delineate thrombi, occluded vessels, and contralateral healthy vessels. An example of an annotated image is shown in Figure 2. All segmentations were made by two radiologists with over five years of specialized neuroimaging expertise. Inter-observer reliability between the two raters was assessed using the intraclass correlation coefficient (ICC) with a two-way mixed-effects model, absolute agreement, and single measures (ICC(3,1)). The analysis has been performed using the R software (version 4.5.1) on a subset of 30 patients, achieving a score of 0.84 [95% CI: 0.70 - 0.92] ($p < 0.001$), indicating good inter-observers agreement and supporting the robustness of the reference annotations. While we did not perform an intra-observer variability analysis due to the prohibitive time constraints of manual 3D volumetric segmentation across thousands of CTA slices, we rigorously quantified inter-observer variability. By demonstrating strong agreement between two senior radiologists, we ensured that our ground-truth annotations are robust, reproducible, and reflective of a consensus clinical standard rather than a single individual's interpretation.

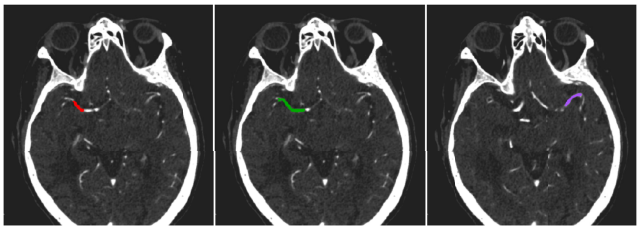


FIGURE 2. Axial CTA slices with ground truths. The thrombus is highlighted in red (left), the occluded vessel segment is shown in green (center), and the contralateral healthy vessel is shown in purple (right).

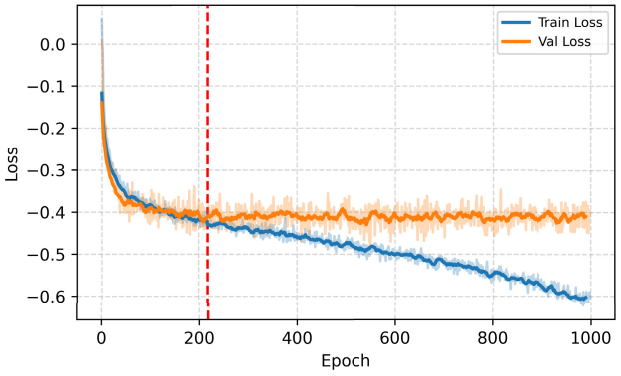


FIGURE 3. Training and validation curves for Standard nnU-Net. The plot displays the training loss (blue) and validation loss (red) on the y-axis over 1000 epochs. The epoch 218, highlighted by the vertical bar, is the best model employed to avoid overfitting.

To ensure dataset consistency while preserving clinically relevant image features, we implemented a preprocessing pipeline using Python 3.10.3 with NumPy, NiBabel, and SimpleITK libraries. The pipeline included: (i) z-axis cropping to isolate skull-specific regions, (ii) voxel resampling for uniform spatial resolution, and (iii) intensity standardization by clipping values between 0 and 500 Hounsfield Units (HU), followed by intensity normalization to the [0, 1] range. The final preprocessed contrast-enhanced CTA images of each patient in the dataset have a standardized dimension of $512 \times 512 \times 200$ pixels and a spatial resolution of $0.4 \times 0.4 \times 0.625$ mm^3, accounting for a total of 39600 slices.

IV. AI SEGMENTATION MODELS

This section introduces the nnU-Net models we have used for detecting vessels and thrombi, together with the computing setup for running training and evaluating our DNN. We underline that leveraging HPC resources is necessary to overcome the computational intensity of 3D medical image segmentation, allowing for practical training times and comprehensive model evaluation.

A. NNU-NET ARCHITECTURES, TRAINING, AND TEST

We selected the nnU-Net model as our primary architecture based on several key considerations. First, nnU-Net is a deep learning-based semantic segmentation model specifically designed for 3D volumetric medical imaging [20], fitting well our CT angiography data characteristics. Second, this architecture automatically adapts to different datasets, minimizing the need for manual hyperparameter tuning. Furthermore, the base architecture is highly extensible; it can readily incorporate advanced components such as residual connections, dense connectivity, and attention mechanisms, facilitating the exploration of various architectural variants.

The Standard (base) nnU-Net architecture used in our work consists of downsampling and upsampling branches with 3D convolutional layers. The downsampling path employs stride convolutions that reduce all spatial dimensions, while the upsampling path progressively reconstructs spatial resolution using transposed convolution layers. Skip connections are critical to the architecture, bridging corresponding depth levels between the downsampling and upsampling blocks to preserve fine-grained contextual details.

Beyond the Standard (STD) version of the nnU-Net [20], we also explored the six variants proposed by [21], namely Residual (RES), Dense (DNS), Inception (INC), Spatial-Single-Attention (SSA), Spatial-Multi-Attention (SMA), and Channel-Spatial-Attention (CSA). The RES integrates residual connections to facilitate gradient flow and improve convergence during training, while the DNS employs dense connectivity patterns, where spatial dimensions remain constant, but the number of channels increases with each layer. The INC introduces specialized parallel building blocks designed for improving computational performance in DNN. The SSA and SMA variants incorporate specialized spatial attention modules that are differentiable neural network components designed to dynamically emphasize important features (e.g., specific pixels, channels, or tokens) while suppressing irrelevant information. Lastly, the CSA combines both channel-wise and spatial attention, aiming to further improve feature representation by jointly modeling inter-channel and spatial dependencies.

We also evaluated more recent transformer-based UNet variants, specifically TransUNet [39] and Swin-UNet [40], optimized under comparable training conditions to ensure fair evaluation. TransUNet integrates Vision Transformers (ViT) into the U-Net decoder branches to capture long-range dependencies, while Swin-UNet employs hierarchical Swin Transformers with shifted windowing mechanisms for efficient multi-scale feature extraction. We have tested both on multiple configurations, with different image resolutions (512×512, 224×224, 112×112 pixels) and backbone architectures (ResNet50-ViT-B and ViT-L).

We trained each DNN model selected for the multiclass segmentation of MCA and thrombi for 1000 epochs. In Figure 3, we provide an example of loss curves during the training process for the STD model without lack of generality. To avoid model overfitting, we considered the best epoch during training for which the values for training and validation losses differ by less than 10^{-3}. In particular, the STD was trained with a non-smoothed loss function to prevent distortion of the Dice score calculation. Without smoothing, we ensured the model maintained high

sensitivity to minute variations, preserving the integrity of our segmentation performance and enabling more nuanced detection of vascular structures and thrombi [20]. We implemented a K-fold cross-validation strategy with $K = 5$ to enhance model robustness and validate the results [41]. The final segmentation accuracy is the average of the scores obtained in the folds. We did not perform any additional post-processing steps after training to preserve the outputs of the nnU-Net models as raw images, for transparent and unbiased performance assessment.

B. COMPUTING SETUP

Training, inferencing, and benchmarking of the DNN were performed on a single-node NVIDIA GH200 Grace Hopper Superchip [22] hosted by the joint HPC facility of the University of Ferrara and the National Institute of Nuclear Physics (INFN). This system features one NVIDIA H100 GPU with 96 GB of memory, paired with a 72-core Grace ARM CPU and 580 GB of RAM. The H100 is the first GPU platform released by Nvidia based on the Hopper architecture, featuring several optimizations for running AI applications [42]. Within the GH200 system, the CPU and the GPU are interconnected through a high-speed, low-latency chip-to-chip link named NVLink-C2C [43]. The reason for using this system lies in the fact that GPUs are today a *de facto* standard for running AI applications since they can boost the computing performance of classical CPUs by two or more orders of magnitude [7]. Moreover, in our case, they are crucial since nnU-Net is based on the computation of large 3D convolutional floating-point matrices, which significantly slows down the computation when run on classical CPUs. Also, we aim to investigate the possibility of performing quasi-real-time online processing of CT scans to support the routine activities of radiologists.

V. FUNCTIONAL METRICS ANALYSIS

To evaluate the segmentation accuracy of our models, we compared predictions with GTs using both overlap-based metrics, such as the Dice score, Surface Dice, and Recall-Optimized Location (ROLOC), and spatial-based metrics such as the 95th percentile Hausdorff distance (HD95), and Center of Mass (CoM) error. The combination of both is essential for the comprehensive evaluation of thrombus segmentation, as relying on volumetric overlap only can lead to incorrect assessment for small anatomical structures where small spatial misalignments are heavily penalized.

The Dice score [44] measures the volumetric overlap between the GT and the inference outcome, returning a score between 0 and 1, where 0 indicates no spatial overlap, and 1 represents a perfect match. It is defined with this formula:

$$Dice = \frac{2\,|V_A \cap V_B|}{|V_A| + |V_B|} \tag{1}$$

where V_A and V_B are the number of voxels in the GT and segmentation prediction, respectively.

Unlike the standard Dice coefficient, the Surface Dice [45] evaluates the agreement between the ground truth and predicted boundaries based on a spatial tolerance τ. It considers surface regions to be overlapping only if the distance between them is below this threshold (set to $\tau = 1$mm), yielding a normalized score between 0 and 1. For this reason, the Surface Dice allows for better evaluation of subtle spatial differences between surfaces that would be otherwise penalized by the Dice score. The Surface Dice is defined as:

$$Surface\ Dice = \frac{2\,|S_A \cap S_B|}{|S_A| + |S_B|} \tag{2}$$

where S_A and S_B are the number of pixels on the surface of the GT and the surface of the segmentation prediction, respectively.

The ROLOC combines Precision and Recall metrics commonly used in Machine Learning applications [46] using a weighted formula to provide a balanced assessment of segmentation performance. The ROLOC definition is:

$$ROLOC = \alpha \cdot P + (1 - \alpha) \cdot R \tag{3}$$

where the parameter α is the weighting factor ($\alpha = 0.5$ in our study, giving equal weight to both metrics), P is the Precision, and R the Recall.

The HD95 is an improved version of the standard Hausdorff Distance (HD), defined as

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{4}$$

$$HD(A, B) = max(h(A, B), h(B, A)) \tag{5}$$

$$HD95(A, B) = P_{95} HD(A, B) \tag{6}$$

It measures the distance between the GT (A) and prediction points (B) in both directions, and uses the 95th percentile of these distances – instead of the maximum – to be more robust to outliers [47].

The CoM Error measures the Euclidean distance between the centers of mass (centroids) of the predicted and GT volumes, quantifying how well the volume inferred is spatially aligned with the actual GT.

All the metrics are computed with Python packages scipy 1.17.0 and scikit-learn 1.8.0.

A. SEGMENTATION RESULTS

In Table 4 we show the vessel and thrombus detection performance of the DNN architectures tested in terms of Dice score (without considering stratifications), reporting median values with first and third quartile ranges in parentheses. The nnU-Net models exhibit consistent vessel segmentation performance, with median Dice scores in the narrow range [0.58 − 0.59] and no statistically significant differences ($p > 0.05$). Thrombus detection performance shows similar homogeneity across all nnU-Net variants, with SSA reaching a slightly higher median Dice of 0.40 [IQR: 0.23 − 0.52], while scores for the others range from 0.15 to 0.38, highlighting negligible performance differences.

TABLE 4. Median Dice scores with interquartile ranges [Q1; Q3] for Vessels and Thrombi segmentation across different the nnU-Net variants, TransUnet and Swin-Unet.

DNN Model	Dice score	
	Vessels	Thrombi
STD	0.58 [0.53;0.65]	0.36 [0.19;0.48]
RES	0.58 [0.53;0.64]	0.35 [0.21;0.50]
INC	0.58 [0.52;0.65]	0.38 [0.21;0.50]
DNS	**0.59 [0.52;0.64]**	0.38 [0.23;0.50]
SSA	0.59 [0.51;0.65]	**0.40 [0.23;0.52]**
SMA	0.58 [0.53;0.65]	0.37 [0.23;0.51]
CSA	0.59 [0.52;0.65]	0.38 [0.21;0.52]
TransUnet	0.51 [0.47;0.58]	0.18 [0.08;0.26]
Swin-Unet	0.50 [0.42;0.56]	0.15 [0.06;0.20]

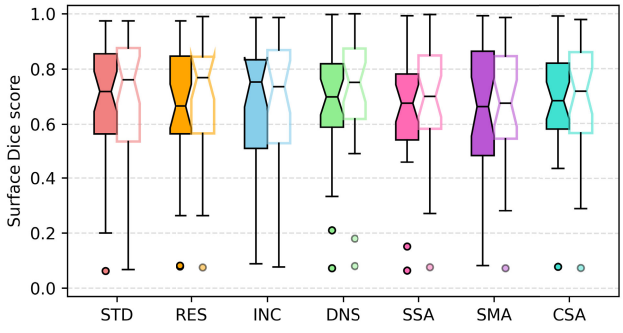


FIGURE 4. Surface Dice score of nnU-Net variants for intracranial thrombus detection, with (empty boxes) and without TTA (solid boxes).

In contrast, transformer-based architectures show overall lower performance on thrombi segmentation, even though on vessels they exhibit comparable performance. The best TransUNet setup ((R50-ViT-B at 224 × 224 resolution) achieved a Dice of 0.51 for vessels and 0.18 for thrombi, whereas the top-performing Swin-UNet configuration (SwU-Swin-Tiny-P4W7 at 224 × 224 resolution) reached 0.50 and 0.15, respectively. For this reason, in the following analyses, we focus only on the nnU-Net variants we have selected.

To further enhance the nnU-Net performance, we optimized two computationally significant inference parameters: Test Time Augmentation (TTA) and the step-size [20], [21]. The TTA is a technique that applies several transformations (i.e., rotations, flips, scaling) to input images during inference, averaging the predictions to improve robustness and potentially enhance segmentation accuracy. The step-size controls the overlap for the sliding-window inference, leading to more precise boundary detection.

We opted for an empirical tuning process that selected the optimal configuration as the best trade-off between segmentation accuracy and computational efficiency for potential clinical deployment.

Figure 4 reports the surface dice achieved by all nn-Unet variants, and Figure 5 the trend of the median Surface Dice

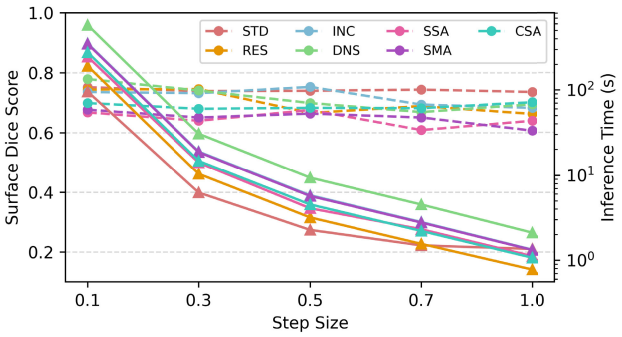


FIGURE 5. Median Surface Dice score (dashed line-points measured on the left-scale) and inference times (solid line-points measured on the right log-scale) as a function of the step size for the nnU-Net variants for intracranial thrombus detection.

and the inference times in seconds for five different step sizes, namely 0.1, 0.3, 0.5, 0.7, and 1.0. Running without TTA and keeping the default value of 0.5 for the step size, the segmentation accuracy does not degrade for all the nnU-Nets variants, achieving a median Surface Dice ranging between 0.65 and 0.75, while the computation time is reduced by approximately 86%. In the same way, reducing the step size slightly improves segmentation accuracy, for example, for step sizes of 0.1 and 0.3, the Surface Dice increases by up to 11%. However, this gain significantly increased inference times by approximately 40× and 2.8×, respectively. Conversely, larger step sizes reduce inference latency by 34% (0.7) and 40% (1.0), but decrease the Surface Dice by up to 9%. Given the limited accuracy gains at smaller step sizes and the non-negligible performance degradation at larger ones, the default value of 0.5 was kept as it provides the best trade-off between segmentation accuracy and computational efficiency. For these reasons, we run all the inferences with TTA disabled since the computation is faster.

In Figure 6, we illustrate four representative segmentation outcome scenarios. Panel (a) shows a successful segmentation of both vessels and a proximal thrombus, with all structures correctly identified. Panel (b) presents a successful segmentation of vessels and a distal thrombus, with no extension of the affected vessel beyond the thrombus location. In panel (c), the vessel-thrombus interface of a proximal thrombus is correctly detected, but the vessel is not fully reconstructed beyond the occlusion, resulting in a gap. Finally, panel (d) depicts a failure case in which the thrombus is not predicted. Across all scenarios, the contralateral healthy vessel is consistently identified due to contrast agent perfusion, which makes it radiodense on CTA images. Importantly, when thrombi are detected, their localization is consistently correct at the vessel–thrombus interface, even when their full spatial extent is not completely segmented. This behavior is reflected in the quantitative results: across the 198 patients in the dataset, thrombus presence and location were correctly identified in 184 cases, corresponding to a sensitivity of 93%.

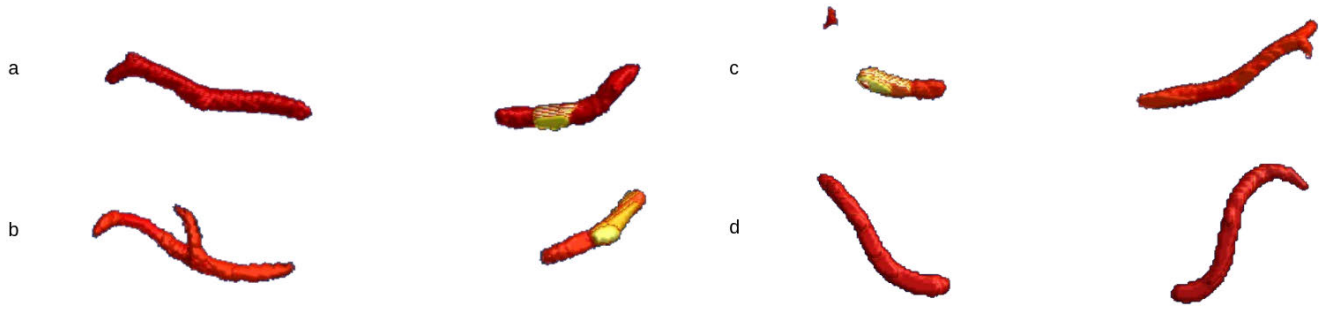


FIGURE 6. Visualization of MCA predictions by a standard nnU-Net. The rows display three-dimensional predictions of vessels (shown in red) with occlusions (shown in yellow) within. In (a), the thrombus is located in the proximal M1 segment; in (b), the thrombus is located in the distal M1 segment. In both cases, the occluded vessel is not fully extended. In (c), the proximal thrombus is not extended, and in (d), no thrombus is predicted.

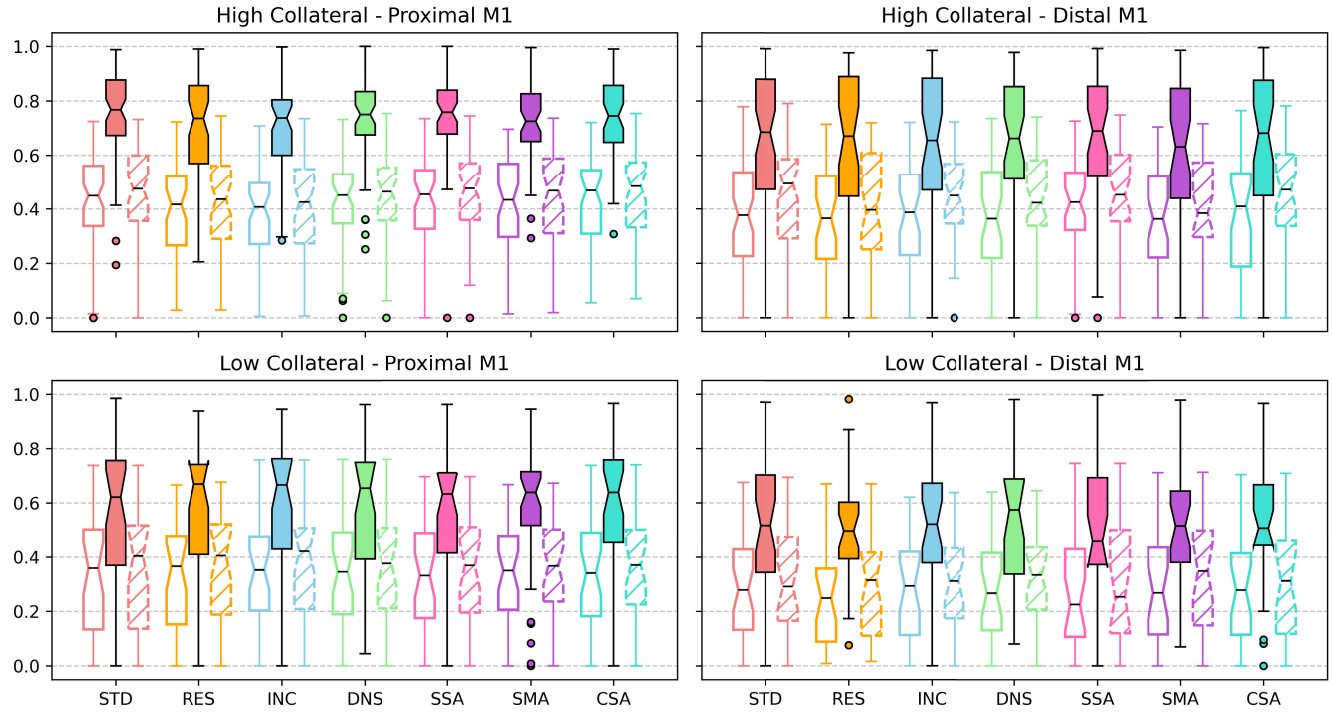


FIGURE 7. Performance comparison of nnUNet variants for thrombus detection across different clinical subgroups, stratified by collateral circulation status (High/Low collateral) and thrombus location (proximal/distal M1), in terms of overlap-based metrics: Dice score (empty boxes), Surface Dice (filled boxes), and ROLOC (hatched boxes). Box plots show median, interquartile ranges, and outliers for each architecture-condition combination.

B. SEGMENTATION PERFORMANCE ACROSS STRATA

Figure 7 shows the comparison of the performance of each nnU-Net variant in terms of overlap-based metrics. Each model was assessed across four distinct dataset subgroups based on collateral status and occlusion location, namely High CC-Proximal M1, High CC-Distal M1, Low CC-Proximal M1, and Low CC-Distal M1. The analysis reveals distinct performance patterns across these clinical scenarios. The High CC-Proximal M1 condition (upper left panel) demonstrated the most favorable segmentation performance, with all architectures achieving median Dice scores in the $[0.4, 0.5]$ range and Surface Dice consistently above 0.7. Conversely, the Low CC-Distal M1 scenario (bottom right panel) presented the most challenging condition, exhibiting significantly reduced performance with a median Dice between the $[0.2, 0.4]$ range and Surface Dice ranging in the $[0.4, 0.6]$ range across architectures. ROLOC consistently followed a similar trend to the Dice but with generally slightly higher values across all four scenarios. To validate the results, we analyzed Dice scores across patient groups using the non-parametric Kruskal-Wallis test. This method was selected because the Anderson-Darling test confirmed that the data were not normally distributed. With statistical significance set at $p < 0.05$, the analysis—conducted in MATLAB R2019b—revealed a significant difference between the Low CC-Distal group and all other groups.

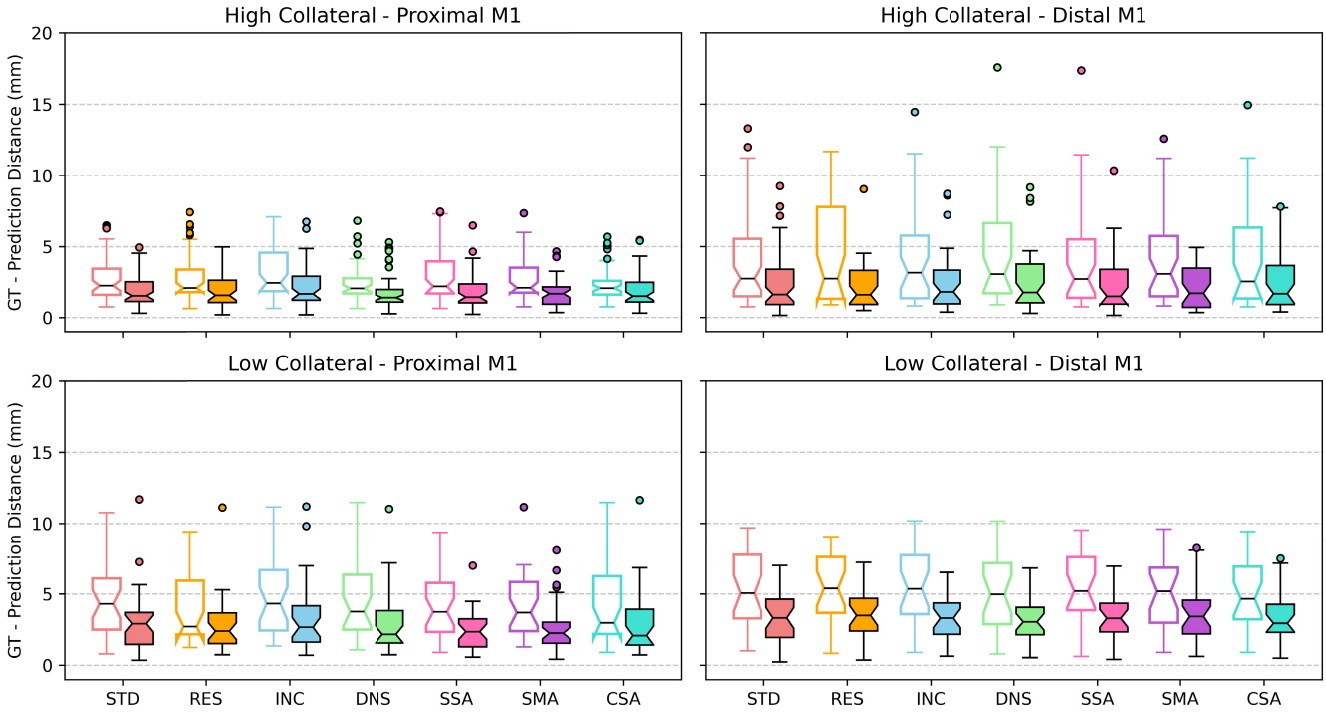


FIGURE 8. Performance of nnUNet variants for thrombus detection across different clinical scenarios, stratified by collateral circulation status (High/Low collateral) and thrombus location (proximal/distal M1), in terms of spatial-based metrics: HD95 (empty boxes) and CoM Error (filled boxes). Box plots show median, interquartile ranges, and outliers for each architecture-condition combination.

Figure 8 shows the results in terms of spatial-based metrics. In the High CC-Proximal M1 case, most of the architectures achieve an HD95 value below 3 mm and a CoM Error consistently under 2 mm. Both Low CC conditions (bottom panels) demonstrated larger spatial values compared to their High CC counterparts, indicating lower segmentation confidence, with HD95 values in the [4, 5] mm range and CoM Error in the [2, 3] mm range. The Low CC-Distal M1 scenario consistently exhibits the highest spatial difference between predictions and GTs across all the models, being a challenging task due to smaller thrombus size and increased vascular complexity at the MCA bifurcation, and indicating that segmentation accuracy is mainly constrained by anatomical and imaging factors rather than model architecture.

VI. DISCUSSION ON THROMBUS SEGMENTATION

The results we have reported in Section V-A highlight that all the nnU-Net variants tested achieve approximately the same level of performance in terms of segmentation quality. This is more related to intrinsic issues in delineating thrombi, and less to the architecture of the DNN model used. In fact, unlike large and homogeneous organs such as the liver and kidney—where nnU-Net typically excels—thrombi are small and possess irregular morphologies [48]. Moreover, on contrast-enhanced CT scans, thrombi are often isointense to the background brain parenchyma. This lack of tissue contrast, combined with poorly defined boundaries, makes them challenging to distinguish. Also, image noises, motion artifacts, and other distortions in CT scans can significantly compromise thrombus visibility, requiring manual annotation by neuro-radiologists [49]. An additional critical factor is the timing between the CTA acquisitions and the contrast agent injection [50]. An early scan, performed before the contrast agent is concentrated enough in the blood, causes incomplete vessel visibility, leading to false negatives because the occlusion sites are not detected. Conversely, delayed acquisitions may obscure arterial structures or simulate false occlusions [51], [52], [53]. Furthermore, in patients with compromised collateral circulation, attenuated vascular signals impede accurate occlusion identification, as reduced perfusion pressure limits contrast delivery to the affected vascular area [54], [55]. For these reasons, our results, consistent with other studies in the literature, confirm that nnU-Net performs better on the vessel segmentation task, where vascular structures exhibit clear radiodensity differences relative to adjacent tissues [56], [57]. Looking at collateral status and the location of occlusions, the performance is approximately 50% better for proximal M1 occlusions with good collateral circulation than for those with low collateral circulation.

Figure 9 shows two representative cases with contrasting performance: a challenging scenario where a zero volumetric overlap (Dice: 0.00) corresponds to a successful thrombus detection (Surface Dice: 0.28), and a successful case with large spatial agreement between Dice (0.71) and Surface Dice (0.92). As we see, the volumetric Dice score is not the

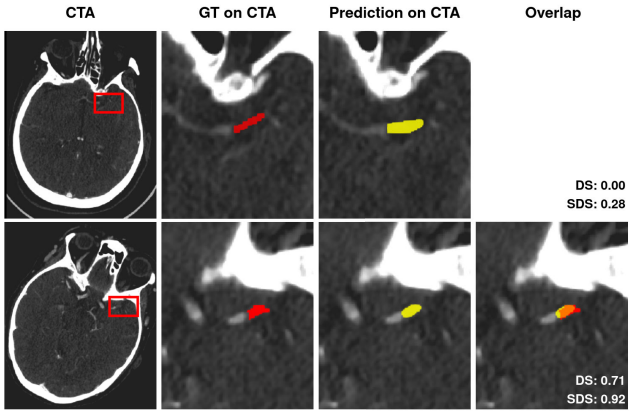


FIGURE 9. Examples of thrombus segmentation performance. DS is the Dice score, and SDS is the Surface Dice score. Top row: the case with no volumetric overlap (DS=0.00) with successful thrombus detection (SDS=0.28). Bottom row: successful segmentation with high agreement between DS (0.71) and SDS (0.92). For both cases, we show in sequence the CTA with the occlusion site in a red bounding box, the GT in red, the prediction in yellow, and the overlap of the two in orange.

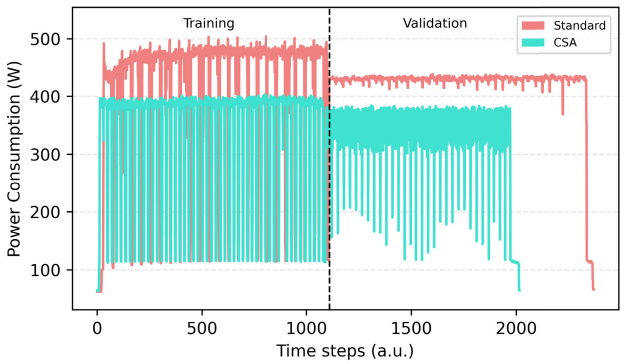


FIGURE 10. Power draw of STD and CSA variants during training and validation on the GH200 system. Dashed vertical lines indicate the transition from the training phase to the validation. Time is reported in the horizontal axes in arbitrary units (a.u.).

best figure of merit, while the Surface Dice better reflects the nnU-Net ability for thrombi localization, providing approximately 30% higher score. As a general result, the nnU-Net model successfully detected thrombi by identifying Hounsfield Unit (HU) changes corresponding to contrast flow interruption, achieving a remarkable result of 93% of true positives, and failing in 14 out of 198 cases (7% false negative rate).

VII. NON-FUNCTIONAL METRICS ANALYSIS

Given that all nnU-Net variants demonstrated comparable segmentation accuracies, we investigated whether significant differences existed instead in non-functional metrics. Beyond segmentation performance, the clinical deployment feasibility of AI-assisted systems depends critically on practical constraints including training costs, inference latency, energy consumption, and operational expenses – aspects typically overlooked in existing thrombus segmentation literature but

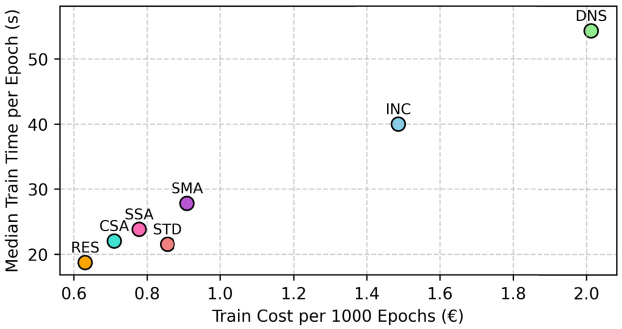


FIGURE 11. Trade-off between training cost (in €) per 1000 epochs and median training time per epoch (in seconds) for different nnU-Net architectural variants running on the GH200 system.

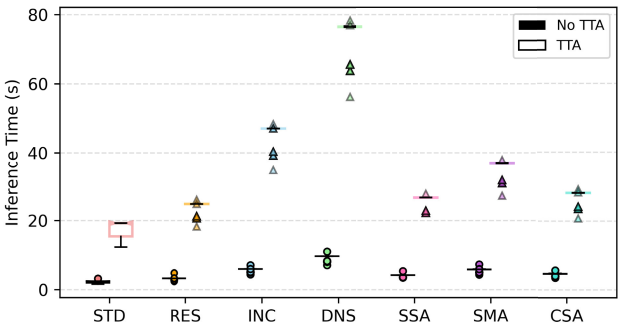


FIGURE 12. Comparison of inference latency (in seconds) of nnU-Net variants for intracranial thrombus detection, evaluated with (triangle fliers) and without the TTA (round fliers) approach.

essential for real-world hospital implementation. Modern GPU systems provide exceptional computational power for deep learning applications. However, their operation comes with a considerable energy cost. This high energy burden is particularly notable in clinical settings, where these systems process vast amounts of medical imaging data from large patient populations.

Achieving energy-efficient AI models in clinical applications requires an optimization between model accuracy and computational overhead. This balance becomes especially critical in medical imaging scenarios where real-time diagnostic capabilities must be maintained without compromising diagnostic precision, particularly in time-critical acute stroke workflows where inference latency directly impacts treatment decisions. We assessed power profiles and their associated energy costs per training and inference in our GH200 system. All the results in this section consider an average electricity cost of 0.30 EUR (€)/kWh [58] and the power draw of the GH200 is extracted with the *nvidia-smi* tool [59].

A. TRAINING

In this section, we discuss the energy consumption and the cost of the training and validation phases running on the GH200 system. In particular, in this analysis, we focus

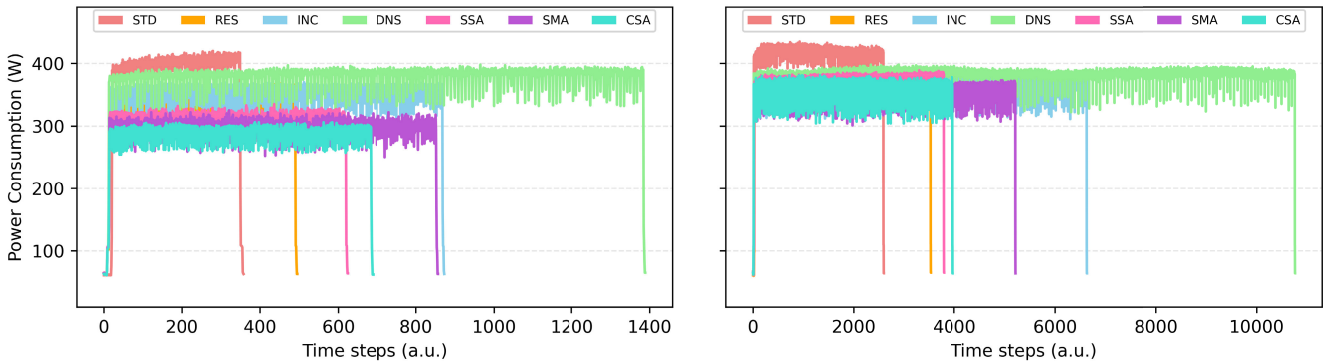


FIGURE 13. Power draw over time in arbitrary units (a.u.) for the seven nnU-Net variants during inference, evaluated both without test-time augmentation (No TTA, left) and with test-time augmentation (TTA, right).

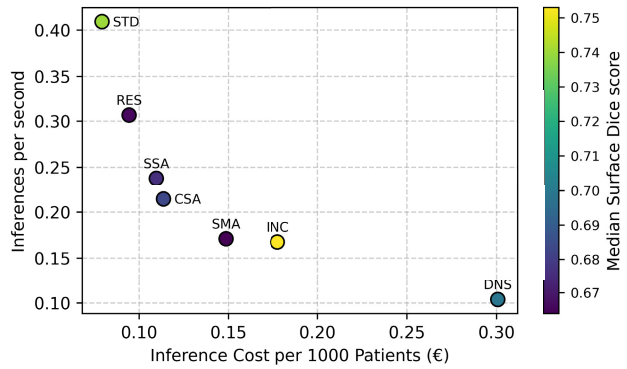


FIGURE 14. Inference-time versus inference-cost. Each point corresponds to a specific architecture, with colors encoding the median Surface Dice score.

on the temporal evolution of energy demand, providing crucial insights for understanding the practical implications of architectural choices in clinical implementation scenarios.

Figure 10 presents a detailed comparison of power draw profiles between STD and CSA variants, measured over 50 epochs. The visualization only shows these two architectures because they represent the highest and lowest power draw profiles among all the variants. The STD demonstrates significantly higher power consumption, averaging approximately 480 W, while the CSA maintains a lower range of 380 − 400 W. The model validation phase reveals markedly different power profiles for both architectures. The STD shows reduced power draw (approximately 430 W) compared to the training phase, while CSA maintains a similar baseline consumption (350−380 W) but with notably increased variability. Both architectures exhibit regular power fluctuations during training accuracy metrics assessment, which occur at the end of each epoch.

Our analysis of training costs revealed that different models have distinct computational profiles, which is shown in Figure 11. The training cost per 1000 epochs ranged from €0.60 for the RES to €2 for the DNS, with corresponding median training times spanning from 17 to 53 seconds per epoch. Attention-based mechanisms (CSA, SSA, SMA) demonstrated favorable cost-efficiency profiles, clustering around €0.70 − 0.90 per 1000 epochs with training times of 22 − 28 seconds per epoch. Lastly, the STD architecture achieves low training costs (€0.85 per 1000 epochs) with 21 seconds per epoch.

B. INFERENCE

This section evaluates the inference latency of all nnU-Net variants to determine their suitability for real-time clinical applications.

Understanding inference latency is critical for time-sensitive stroke diagnosis, where rapid image processing can directly impact patient outcomes and treatment decision timelines [17]. Figure 12 shows the inference latency, in seconds, for all the nnU-Net architectures, with and without TTA. Without TTA, all variants demonstrated consistently fast inference times (approximately 5 − 10 seconds per patient), with minimal variations. Applying TTA significantly increased the computational load, causing a steep jump in inference times. Of all the variants, the DNS is the slowest with TTA enabled, taking up to 80 s per patient, while the STD remains relatively fast at around 20 s per patient.

Having demonstrated that TTA significantly increases inference time, we extended our analysis to examine its contribution to the power consumption characteristics of each model and their associated energy costs per inference, as shown in Figure 13. In the No-TTA scenario, power consumption exhibits considerable variations across the architectures, ranging from around 260 W for the CSA (cyan) to approximately 380 W for the DNS (green). Conversely, under TTA conditions, power consumption remains relatively uniform across all DNN, stabilizing within a narrower range of 300−400 W. Notably, across both experimental conditions, the STD demonstrates the highest energy consumption, requiring approximately 400 W in the No-TTA condition and 430 W with TTA enabled. A critical observation is the substantial difference in inference time between conditions:

TABLE 5. Summary of main literature studies for AI-assisted thrombus segmentation and detection. The first column reports the authors, the second lists the image acquisition modalities, the third the size of the dataset, the forth the AI method used by the study, the fifth and sixth report if the data comes from different centers and if an external dataset has been used to validate the results, the seventh describes the main purpose of the study, and the last column reports the metrics used to measure the results achieved by each work and the corresponding values.

Author	Imaging modality	Dataset size	AI Methodology	Multi center	External validation	Clinical application	Evaluation metrics
Kobold et al. [24]	MRI	61	U-Net	No	No	Thrombus segmentation	Dice: 0.42 Detection rate: 0.77
Tolhuisen et al. [25]	NCCT	103	2x AlexNet	Yes	No	Thrombus detection and segmentation	Sensitivity: 0.86 Specificity: 0.65
Mojtahedi et al. [26]	CTA + NCCT	328	U-Net + Weighted-Sum	Yes	No	Thrombus segmentation	Dice: 0.62 Surface Dice : 0.78
Zoetmulder at al. [27]	CTA + NCCT	187	BL-UNet, Polar UNet	Yes	No	Thrombus localization and segmentation	Dice: 0.44 Localization Precision: 0.56
Brugnara et al. [61]	CTA	1179	RetinaNet	Yes	Yes	Thrombus detection	Sensitivity: 0.94 Specificity: 0.83 overall FROC score: 0.79
Liang et al. [28]	CTA	29	Attention Unet + CPM	No	No	Thrombus segmentation	Dice: 0.43
Soh et al. [29]	MRI, CT	239	U-Net Transformer	No	No	Ischemic stroke small lesion segmentation	Dice: 0.47 HD95: 12.63 mm
Amukotuwa et al. [62]	CTA	477	Automated tool (RAPID CTA)	No	No	LVO detection	Sensitivity: 0.94 Specificity: 0.76
Zhu et al. [19]	CTA + NCCT	499	U-Net-like	No	Yes	Thrombus segmentation	Dice: 0.71 HD95: 1.31 mm
Our work	**CTA**	**198**	**Standard nnU-Net**	**No**	**No**	**LVO detection**	**Dice: 0.45** **Surface Dice: 0.77** **HD95: 2.5 mm** **Sensitivity: 0.93**

the TTA configuration requires approximately $6 - 7.5\times$ longer runtime than the No-TTA setting. Additionally, we can notice the best trade-off in the STD performance profile. While maintaining consistently high power consumption across both conditions, the STD variant completes the inference phase in a significantly shorter time duration compared to other models.

The scatter plot in Figure 14 reports the throughput against the accuracy. STD demonstrates the highest throughput (0.41 inferences per second) at the lowest cost (€0.08 per 1000 patients) while maintaining a good segmentation accuracy (median Surface Dice score: 0.74), highlighting its favorable balance between computational efficiency and performance. The DNS is the least efficient model tested, with the lowest throughput (0.10 inferences per second) and the highest cost (€0.30 per 1000 patients), while not providing a significant increase in accuracy compared to other models (median Surface Dice score of 0.70). All the other architectures are positioned in the middle, with throughputs in the range of 0.16 to 0.31 inferences per second and costs between €0.10 to €0.18 per 1000 patients. At the same time, they maintain comparable accuracy, with median Surface Dice scores ranging from 0.67 to 0.75.

VIII. CONCLUSION

In this work, we have explored the use of state-of-the-art UNet architectures and more recent Transformer concepts for LVO detection in the MCA-M1 segment and assessed the best trade-off between segmentation accuracy and non-functional parameters. We have trained seven variants of the nnU-Net model on 198 late-phase CTA scans, divided into four strata based on thrombus location and CC status. A key finding of our study is that while thrombus detection remains challenging, the models successfully identified and located occlusions in 93% of cases, mainly delineating vessel-thrombus boundaries accurately. This makes the model a reliable tool to help radiologists localize thrombi and assess patient conditions quickly in time-critical emergency settings. On transformer-based models, the results achieved are in line with those of nnU-Net for vessel segmentation, but significantly lower for the thrombi.

In Table 5 we list some of the major works published in the literature. In terms of Surface Dice score, our results are in line with those achieved by Mojtahedi et al., and for both studies, it significantly improves the results compared to the volumetric Dice score. Considering the HD95 metric, our results are comparable with those reported by Zhu et al. and significantly better than those of Soh et al., despite a similar Dice score value measured. For this metric, the errors range between 3 and 5 mm; considering that the thickness of a CT slice is 0.625 mm, this translates into an error of a few slices (between 5 and 8), which does not hamper the localization of thrombi. Finally, looking at the sensitivity, we can compare with the results of Zhu et al. and Brugnara et al. However, we underline that, even if, in first approximation, our study achieves similar results with those in the literature, a direct comparison is not easy to do, since different image acquisition modalities (MRI, NCCT, CTA, or multimodal

combinations), dataset sizes, and evaluation metrics are used. We remark that, in general, the full segmentation of the thrombi remains a challenging task since the value of the volumetric Dice score achieved is low. We believe that this is not dependent on the architecture model used, but it is fundamentally constrained by anatomical complexities of the thrombi, with the Low CC-Distal M1 subgroup presenting the most challenging scenario for both overlap and spatial accuracy metrics. Also, we highlight that the majority of the studies have no external validation, confirming the difficulties we have faced in accessing clinical datasets under the severe regulation protocols applied in medicine.

In contrast to the other works, beyond accuracy metrics, we have also considered the non-functional metrics (i.e., power consumption, energy cost) relevant to assessing the deployment of our AI model in hospital premises, where the structure and organization are quite different from those of a computing center. Among the models we have evaluated, the STD nnU-Net achieves the best balance, combining strong segmentation accuracy (median Surface Dice 0.74) with low inference time latency ($\sim 5-10$ s per patient), and low operational costs (€0.08 per 1000 patients). These results suggest that the model we have proposed can be sustained and installed in hospital premises for making automatic quasi-real-time CT scans processing, helping the radiologists in the patient conditions assessment admitted to the stroke unit.

In future works, we plan to proceed along three main directions: (i) we aim to expand the dataset to include cases with varying image quality, artifacts, and suboptimal contrast timing to comprehensively assess system robustness in routine clinical practice where such variations are common; (ii) to include multicenter validation studies to assess model generalizability across different acquisition protocols and institutional practices; and (iii) to make prospective clinical validation within a workflow integration framework at our institutions to evaluate real-world advantages in routine stroke management.

REFERENCES

[1] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," *Minimally Invasive Therapy Allied Technol.*, vol. 28, no. 2, pp. 73–81, 2019.

[2] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in U.S. images and pulmonary nodules in CT scans," *Sci. Rep.*, vol. 6, no. 1, p. 24454, Apr. 2016.

[3] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[4] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From CNN to transformer: A review of medical image segmentation models," *J. Imag. Informat. Med.*, vol. 37, no. 4, pp. 1529–1547, Mar. 2024.

[5] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study," *Lancet*, vol. 392, no. 10162, pp. 2388–2396, Dec. 2018.

[6] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[7] C. Silvano et al., "A survey on deep learning hardware accelerators for heterogeneous HPC platforms," *ACM Comput. Surveys*, vol. 57, no. 11, pp. 1–39, Nov. 2025.

[8] G. Minghini, A. U. Cavallo, A. Miola, V. Sisini, E. Calore, F. Fortini, R. Micheloni, P. Rizzo, S. F. Schifano, F. V. D. Sega, and C. Zambelli, "An HPC pipeline for calcium quantification of aortic root from contrast-enhanced CCT scans," *IEEE Access*, vol. 11, pp. 101309–101319, 2023.

[9] J. K. Nicholls, J. Ince, J. S. Minhas, and E. M. L. Chung, "Emerging detection techniques for large vessel occlusion stroke: A scoping review," *Frontiers Neurol.*, vol. 12, Jan. 2022, Art. no. 780324.

[10] C. A. Potter, A. S. Vagal, M. Goyal, D. B. Nunez, T. M. Leslie-Mazwi, and M. H. Lev, "CT for treatment selection in acute ischemic stroke: A code stroke primer," *RadioGraphics*, vol. 39, no. 6, pp. 1717–1738, Oct. 2019.

[11] M. Abdalkader, J. E. Siegler, J. S. Lee, S. Yaghi, Z. Qiu, X. Huo, Z. Miao, B. C. V. Campbell, and T. N. Nguyen, "Neuroimaging of acute ischemic stroke: Multimodal imaging approach for acute endovascular therapy," *J. Stroke*, vol. 25, no. 1, pp. 55–71, Jan. 2023.

[12] G. Maguida and A. Shuaib, "Collateral circulation in ischemic stroke: An updated review," *J. Stroke*, vol. 25, no. 2, pp. 179–198, May 2023.

[13] M. Sperti, F. Arba, A. Acerbi, G. Busto, E. Fainardi, and C. Sarti, "Determinants of cerebral collateral circulation in acute ischemic stroke due to large vessel occlusion," *Frontiers Neurol.*, vol. 14, May 2023.

[14] M. Mangiardi, A. Bonura, G. Iaccarino, M. Alessiani, M. C. Bravi, D. Crupi, F. R. Pezzella, S. Fabiano, E. Pampana, F. Stilo, G. Alfano, and S. Anticoli, "The pathophysiology of collateral circulation in acute ischemic stroke," *Diagnostics*, vol. 13, no. 14, p. 2425, Jul. 2023.

[15] C. Jiang, Y. Li, F. Hao, J. Yang, B. Wang, and Y. Fan, "Y-configuration double-stent-retriever thrombectomy for refractory thrombus in middle cerebral artery bifurcation: A case report," *Medicine*, vol. 100, no. 11, Mar. 2021, Art. no. e24993.

[16] E. M. M. Santos, H. A. Marquering, O. A. Berkhemer, W. H. van Zwam, A. van der Lugt, C. B. Majoie, and W. J. Niessen, "Development and validation of intracranial thrombus segmentation on CT angiography in patients with acute ischemic stroke," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e101985.

[17] W. J. Powers, A. A. Rabinstein, T. Ackerson, O. M. Adeoye, N. C. Bambakidis, K. Becker, J. Biller, M. Brown, B. M. Demaerschalk, B. Hoh, E. C. Jauch, C. S. Kidwell, T. M. Leslie-Mazwi, B. Ovbiagele, P. A. Scott, K. N. Sheth, A. M. Southerland, D. V. Summers, and D. L. Tirschwell, "Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: A guideline for healthcare professionals from the American heart association/American stroke association," *Stroke*, vol. 50, no. 12, pp. e344–e418, Dec. 2019.

[18] K. Zhu, F. Bala, J. Zhang, F. Benali, P. Cimflova, B. J. Kim, R. McDonough, N. Singh, M. D. Hill, M. Goyal, A. Demchuk, B. K. Menon, and W. Qiu, "Automated segmentation of intracranial thrombus on NCCT and CTA in patients with acute ischemic stroke using a coarse-to-fine deep learning model," *Amer. J. Neuroradiology*, vol. 44, no. 6, pp. 641–648, Jun. 2023.

[19] L. Cui, Z. Fan, Y. Yang, R. Liu, D. Wang, Y. Feng, J. Lu, and Y. Fan, "Deep learning in ischemic stroke imaging analysis: A comprehensive review," *BioMed Res. Int.*, vol. 2022, no. 1, Jan. 2022, Art. no. 2456550.

[20] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[21] N. McConnell, N. Ndipenoch, Y. Cao, A. Miron, and Y. Li, "Exploring advanced architectural variations of nnUNet," *Neurocomputing*, vol. 560, Dec. 2023, Art. no. 126837.

[22] NVIDIA. (Aug. 5, 2025). *NVIDIA Grace Hopper Superchip Architecture Whitepaper*. [Online]. Available: https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

[23] J. Kobold, E. Lang, A. M. Tome, V. Vigneron, H. Maaref, D. Fourer, M. Aghasaryan, C. Alecu, N. Chausson, Y. L'Hermitte, and D. Smadja, *Stroke Thrombus Segmentation on SWAN With Multi-Directional U Nets*. Istanbul, Turkey: IEEE, Nov. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8936074/

[24] M. L. Tolhuisen, E. Ponomareva, A. M. M. Boers, I. G. H. Jansen, M. S. Koopman, R. Sales Barros, O. A. Berkhemer, W. H. Van Zwam, A. Van Der Lugt, C. B. L. M. Majoie, and H. A. Marquering, "A convolutional neural network for anterior intra-arterial thrombus detection and segmentation on non-contrast computed tomography of patients with acute ischemic stroke," *Appl. Sci.*, vol. 10, no. 14, p. 4861, Jul. 2020.

[25] M. Mojtahedi et al., "Fully automated thrombus segmentation on CT images of patients with acute ischemic stroke," *Diagnostics*, vol. 12, no. 3, p. 698, Mar. 2022.

[26] R. Zoetmulder, A. Bruggeman, I. Išgum, E. Gavves, C. Majoie, L. Beenen, D. Dippel, N. Boodt, S. den Hartog, P. van Doormaal, S. Cornelissen, Y. Roos, J. Brouwer, W. Schonewille, A. Pirson, W. van Zwam, C. van der Leij, R. Brans, A. van Es, and H. Marquering, "Deep-learning-based thrombus localization and segmentation in patients with posterior circulation stroke," *Diagnostics*, vol. 12, no. 6, p. 1400, Jun. 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/6/1400

[27] J. Liang, Y. Li, S. Chen, Z. Wang, C. Nie, Z. Kang, and B. Mei, "Cerebral thrombus segmentation in CT angiography using refinement segmentation network with context pooling," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–9.

[28] W. K. Soh and J. C. Rajapakse, "Hybrid UNet transformer architecture for ischemic stoke segmentation with MRI and CT datasets," *Frontiers Neurosci.*, vol. 17, Nov. 2023, Art. no. 1298514. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2023.1298514/full

[29] R. Berzoini, E. D'Arnese, and D. Conficconi, "On how to push efficient medical semantic segmentation to the edge: The SENECA approach," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, May 2022, pp. 104–111.

[30] R. Berzoini, E. D'Arnese, D. Conficconi, and M. D. Santambrogio, "NERONE: The fast way to efficiently execute your deep learning algorithm at the edge," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 3, pp. 1252–1260, Mar. 2024.

[31] M. Dörrich, M. Fan, and A. M. Kist, "Impact of mixed precision techniques on training and inference efficiency of deep neural networks," *IEEE Access*, vol. 11, pp. 57627–57634, 2023.

[32] J. Fornt, E. Reggiani, P. Fontova-Musté, N. Rodas, A. Pappalardo, O. S. Unsal, A. Cristal Kestelman, J. Altet, F. Moll, and J. Abella, "Mix-GEMM: Extending RISC-V CPUs for energy-efficient mixed-precision DNN inference using binary segmentation," *IEEE Trans. Comput.*, vol. 74, no. 2, pp. 582–596, Feb. 2025.

[33] V. Sisini, A. Miola, G. Minghini, E. Calore, A. U. Cavallo, S. F. Schifano, and C. Zambelli, "Benchmarking a DNN for aortic valve calcium lesions segmentation on FPGA-based DPU using the vitis AI toolchain," *Future Gener. Comput. Syst.*, vol. 175, Feb. 2026, Art. no. 108115. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X25004091

[34] M. Waqas, M. Mokin, C. T. Primiani, A. D. Gong, H. Hamid, F. Chin, A. Rai, E. I. Levy, and A. H. Siddiqui, "Large vessel occlusion in acute ischemic stroke patients: A dual-center estimate based on a broad definition of occlusion site," *J. Stroke Cerebrovascular Diseases*, vol. 29, no. 2, Feb. 2019, Art. no. 104504. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1052305719305889

[35] A. L. Czap, M. Bahr-Hosseini, N. Singh, J.-M. Yamal, M. Nour, S. Parker, Y. Kim, L. Restrepo, R. Abdelkhaleq, S. Salazar-Marioni, K. Phan, R. Bowry, S. S. Rajan, J. C. Grotta, J. L. Saver, L. Giancardo, and S. A. Sheth, "Machine learning automated detection of large vessel occlusion from mobile stroke unit computed tomography angiography," *Stroke*, vol. 53, no. 5, pp. 1651–1656, May 2022. [Online]. Available: https://www.ahajournals.org/doi/10.1161/STROKEAHA.121.036091

[36] S. Prabhakaran et al., "2026 guideline for the early management of patients with acute ischemic stroke: A guideline from the American heart association/American stroke association," *Stroke*, 2026. [Online]. Available: https://www.ahajournals.org/doi/10.1161/STR.0000000000000513

[37] L. L. L. Yeo, P. Paliwal, H. L. Teoh, R. C. Seet, B. P. Chan, E. Ting, N. Venketasubramanian, W. K. Leow, B. Wakerley, Y. Kusama, R. Rathakrishnan, and V. K. Sharma, "Assessment of intracranial collaterals on CT angiography in anterior circulation acute ischemic stroke," *Amer. J. Neuroradiology*, vol. 36, no. 2, pp. 289–294, Feb. 2015.

[38] I. Wolf, M. Vetter, I. Wegner, M. Nolden, T. Boettger, M. Hastenteufel, M. Schobinger, T. Kunert, and H.-P. Meinzer, "The medical imaging interaction toolkit (MITK)—A toolkit facilitating the creation of interactive software by extending VTK and ITK," *Proc. SPIE*, vol. 5367, May 2024, Art. no. 535112.

[39] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou, "TransUNet: Rethinking the U-net architecture design for medical image segmentation through the lens of transformers," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103280.

[40] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Comput. Vis. ECCV Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., Cham: Springer Nature Switzerland, 2021, pp. 205–218.

[41] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Stat. Society. Ser. B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[42] W. Luo, R. Fan, Z. Li, D. Du, H. Liu, Q. Wang, and X. Chu, "Dissecting the NVIDIA hopper architecture through microbenchmarking and multiple level analysis," 2025, *arXiv:2501.12084*.

[43] Y. Wei, Y. C. Huang, H. Tang, N. Sankaran, I. Chadha, D. Dai, O. Oluwole, V. Balan, and E. Lee, "9.3 NVLink-C2C: A coherent off package chip-to-chip interconnect with 40Gbps/pin single-ended signaling," in *Proc. IEEE Int. Solid- State Circuits Conf. (ISSCC)*, Chad, Feb. 2023, pp. 160–162.

[44] A. Carass, S. Roy, A. Gherman, J. C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, D. L. Pham, C. M. Crainiceanu, P. A. Calabresi, J. L. Prince, W. R. G. Roncal, R. T. Shinohara, and I. Oguz, "Evaluating white matter lesion segmentations with refined Sørensen-dice analysis," *Sci. Rep.*, vol. 10, no. 1, p. 8242, May 2020.

[45] S. Nikolov et al., "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," 2018, *arXiv:1809.04430*.

[46] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 205–229, Jul. 1989.

[47] A. Reinke et al., "Common limitations of image processing metrics: A picture story," 2023, *arXiv:2104.05642*.

[48] J. Park, I. Joo, S. K. Jeon, J. Kim, S. J. Park, and S. H. Yoon, "Automated abdominal organ segmentation algorithms for non-enhanced CT for volumetry and 3D radiomics analysis," *Abdominal Radiol.*, vol. 50, no. 3, pp. 1448–1456, 2024.

[49] M. Alkhodari, E. Alefishat, H. F. Jelinek, A. Kaabneh, and P. Liatsis, "Enhancing CCTA image quality: A review of deep learning approaches for advanced artifact correction and denoising," *Artif. Intell. Rev.*, vol. 58, no. 10, Aug. 2025.

[50] A. Rau, S. Elsheikh, P. Cimflova, T. Stein, C. A. Taschner, J. A. Hosp, M. Reisert, H. Urbach, and E. Kellner, "Contrast bolus timing in CT-angiography and CT-perfusion: Insights from a large clinical dataset," *Neuroradiology*, vol. 67, no. 5, pp. 1171–1181, May 2025.

[51] M. Kappelhof, H. A. Marquering, O. A. Berkhemer, J. Borst, A. van der Lugt, W. H. van Zwam, J. A. Vos, G. Lycklama Ã Nijeholt, C. B. L. M. Majoie, and B. J. Emmer, "Accuracy of CT angiography for differentiating pseudo-occlusion from true occlusion or high-grade stenosis of the extracranial ICA in acute ischemic stroke: A retrospective MR CLEAN substudy," *Amer. J. Neuroradiology*, vol. 39, no. 5, pp. 892–898, May 2018.

[52] W. S. Jung, J. S. Lee, S. Solander, and J. W. Choi, "Pseudo-occlusion of the internal carotid artery in acute ischemic stroke: Clinical outcome after mechanical thrombectomy," *Sci. Rep.*, vol. 10, no. 1, p. 2832, 2020.

[53] T. Shirasaka, A. Hiwatashi, K. Yamashita, M. Kondo, H. Hamasaki, Y. Shimomiya, Y. Nakamura, Y. Funama, and H. Honda, "Optimal scan timing for artery–vein separation at whole-brain CT angiography using a 320-row MDCT volume scanner," *Brit. J. Radiol.*, vol. 90, no. 1070, Feb. 2017.

[54] H. Kawano, A. Bivard, L. Lin, N. J. Spratt, F. Miteff, M. Parsons, and C. Levi, "Relationship between collateral status, contrast transit, and contrast density in acute ischemic stroke," *Stroke*, vol. 47, no. 3, pp. 742–749, 2016.

[55] A. Potreck, E. Scheidecker, C. S. Weyland, U. Neuberger, C. Herweh, M. A. Möhlenbruch, M. Chen, S. Nagel, M. Bendszus, and F. Seker, "RAPID CT perfusion–based relative CBF identifies good collateral status better than hypoperfusion intensity ratio, CBV-index, and time-to-maximum in anterior circulation stroke," *Amer. J. Neuroradiology*, vol. 43, no. 7, pp. 960–965, Jul. 2022.

[56] X. Guo, T. Liu, Y. Yang, J. Dai, L. Wang, D. Tang, and H. Sun, "Automatic segmentation of type a aortic dissection on computed tomography images using deep learning approach," *Diagnostics*, vol. 14, no. 13, p. 1332, Jun. 2024.

[57] Y. Zhu, L. Chen, W. Lu, Y. Gong, and X. Wang, "The application of the nnU-net-based automatic segmentation model in assisting carotid artery stenosis and carotid atherosclerotic plaque evaluation," *Frontiers Physiol.*, vol. 13, Dec. 2022.

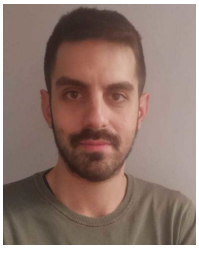
[58] European Commission. (2025). *Energy Prices and Costs in Europe*. [Online]. Available: https://energy.ec.europa.eu/data-and-analysis/energy-prices-and-costs-europe_en
[59] S. Nikolov et al. (2023). *NVML API Reference Guide*. [Online]. Available: https://docs.nvidia.com/deploy/nvml-api/index.html
[60] G. Brugnara et al., "Deep-learning based detection of vessel occlusions on CT-angiography in patients with suspected acute ischemic stroke," *Nature Commun.*, vol. 14, no. 1, p. 4938, Aug. 2023. [Online]. Available: https://www.nature.com/articles/s41467-023-40564-8
[61] S. A. Amukotuwa, M. Straka, H. Smith, R. V. Chandra, S. Dehkharghani, N. J. Fischbein, and R. Bammer, "Automated detection of intracranial large vessel occlusions on computed tomography angiography: A single center experience," *Stroke*, vol. 50, no. 10, pp. 2790–2798, Oct. 2019. [Online]. Available: https://www.ahajournals.org/doi/10.1161/STROKEAHA.119.026259



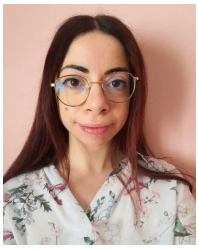
GIADA MINGHINI received the master's degree in biotechnology for the environment and health from the University of Ferrara, Ferrara, Italy, in 2021. Since 2022, she has been a Researcher with the University of Ferrara, and her activity focuses on the analysis of cardiovascular images to develop a high-performance automatic system to detect, segment, and quantify the volume of calcium in the aortic valve.



LUIGI BELLINI received the master's degree in medicine and surgery from the University of Rome "La Sapienza" in 2018, the Medical degree and the Specialization degree in diagnostic imaging and interventional radiology from the University of Rome "Tor Vergata," in 2019 and 2023, respectively, and the master's degree in interventional radiology from the University of Pisa, in 2025. He is currently pursuing the Ph.D. degree focusing on radiomics and artificial intelligence in neuroimaging. Further specialized through training with the IRCCS-Policlinico di Sant'Orsola, Bologna, focusing on the endovascular diagnosis and treatment of liver tumors. He was a Radiologist with the Policlinico Casilino of Rome, from 2023 to 2024. He is currently a Radiologist Consultant with the Policlinico "Tor Vergata" of Rome.



ARMANDO UGO CAVALLO received the degree in medicine and surgery from the University of Salerno, in 2014, the Specialization degree in radio diagnostics and the Ph.D. degree in medical biotechnology and translational medicine from the University of Rome "Tor Vergata," in 2019, with a major focus on radiomics and artificial intelligence applications in cardiovascular and thoracic imaging. From 2015 to 2019, he was a Radiology Resident with the University of Rome "Tor Vergata." In 2018, he was a Research Fellow in cardiovascular imaging with Cleveland University Hospital/Case Western Reserve University, Cleveland, OH, USA. Since 2022, he has been a Consultant Radiologist with the Istituto Dermopatico dell'Immacolata (IDI) IRCCS, Rome, Italy. His main research interests include artificial intelligence and radiomics in oncological and cardiovascular radiology, as well as quantitative imaging.



MARIAFRANCESCA TRULLI received the degree in medicine and surgery from the University of Rome "Tor Vergata," in 2022. She is currently a third-year Radiology Resident with the Policlinico Tor Vergata, Rome, Italy. Her clinical training is increasingly focused on interventional radiology, with a particular interest in neuro interventional procedures.



SIMONE DI CANDIA received the master's degree (Hons.) in medicine and surgery from the University of Rome "Tor Vergata", 110 cum laude. He is currently a third-year Resident in diagnostic imaging and interventional radiology with "Tor Vergata" University Hospital. His interests range from interventional radiology, neuro intervention, and neuroradiology.



FRANCESCO GIUSEPPE GARACI received the Medical degree (cum laude) from the University of Rome "Tor Vergata," in 1998, and the Specialization degree in radio diagnostics from the University of Rome "Tor Vergata" and subsequently specialized in nuclear medicine. His early career included a period as an In-training Specialist Registrar in neuroradiology with the Lysholm Department of Radiology, National Hospital for Neurology and Neurosurgery, London, U.K.; a Visiting-Fellow in neuroradiology with the Leids Universitair Medisch Centrum, Leiden, The Netherlands; and a Visiting-Fellow with the Department of Radiology, Boston Medical Center, Boston University, USA. He also held a Visiting-Student position with the Nuclear Medicine Section, National Institute of Health (NIH), Bethesda, Maryland, USA. He is currently a Full Professor of radiology with the University of Rome "Tor Vergata," where he is the Interim Director of the UOC Diagnostic Imaging and the Head of the UOSD Neuroradiology. Furthermore, he is the Director of the School of Specialization in radio diagnostics.



ANDREA MIOLA received the M.Sc. degree in astrophysics and cosmology and the second-level professional master's degree in space missions science, design, and applications (SPICES) from the Alma Mater Studiorum–University of Bologna, Italy, in 2019 and 2022, respectively, and the Ph.D. degree in physics from the University of Ferrara, Ferrara, Italy, in 2025, where his research focused on deep learning and high-performance computing (HPC). Since 2021, he has been with MEEO, Ferrara, where he began applying artificial neural networks to various research domains. His main research interests include computer-aided analysis of medical imaging and satellite-based remote sensing.



ENRICO CALORE received the degree in computer engineering from the University of Padua, Italy, in 2010, and the Ph.D. degree in computer science from the University of Milan, Italy, in 2014. Later, he was a Postdoctoral Researcher with Italian National Institute of Nuclear Physics (INFN), Ferrara, and with the University of Ferrara, where he was also an Adjunct Professor with the Mathematics and Computer Science Department. He is currently a Research Engineer with INFN Ferrara. He is the author of more than 60 articles and conference papers. His research interests include HPC, parallel, and distributed computing, focusing in particular on scientific computing, parallel programming, performance engineering, and energy efficiency.



SEBASTIANO FABIO SCHIFANO received the degree in computer science from the University of Pisa, Italy. He spent his early scientific career with IEI-CNR and INFN. He is currently a Professor of computer science with the University of Ferrara, Italy. He had a major role in several projects for the development of parallel systems optimized for lattice gauge theory, fluid dynamics, and spin glasses. In 2007, he was the co-author of the QPACE Project to develop a massively parallel system based on IBM Cell-BE processors, interconnected by a custom 3D-mesh network, awarded as the best Green500 system, in November and June. From 2005 to 2010, he contributed to the design and development of the Janus systems, an array of FPGAs for spin-glass simulations. He is the author of more than 100 articles and conference papers. More recently, his research activities have focused on the design and optimization of applications for multi- and many-core processor architectures.



CRISTIAN ZAMBELLI (Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from the University of Ferrara, Ferrara, Italy, in 2008 and 2012, respectively. He is currently an Associate Professor with the University of Ferrara. His current research interests include the electrical characterization, physics, and reliability modeling of different non-volatile memory technologies for data storage applications and emerging in-memory computing concepts. He is also interested in evaluating the reliability of solid-state drives' performance tradeoffs exposed by integrated memory technology to develop computational storage accelerators for the HPC scenario.



VALERIO DA ROS received the medical degree from the University of Rome "Tor Vergata," in 2008, the Specialization Diploma degree (Hons.) in diagnostic imaging from the University of Rome "Tor Vergata," in 2014, the master's degree in healthcare economics and management, in 2021, and the Diploma degree in interventional neuroradiology from the University of Limoges. He completed an Observership with the Memorial Sloan Kettering Cancer Center, New York, USA, in 2013. From 2015 to 2018, he pursued a European Ph.D. Label with NEURI Bicetre University Hospital, Paris, France, under the guidance of Professors Jacques Moret and Laurent Spelle. In 2016, he obtained the European Diploma of Neuroradiology and, in 2017. He was a Neurointerventionalist-Praticien Attaché with Bicetre University Hospital, Paris, from 2016 to 2017. He has been a Researcher with the University of Rome Tor Vergata, since 2019, where he was appointed an Associate Professor, in 2022. In 2025, he achieved the national habilitation as a Full Professor of radiology. He was appointed to the artificial intelligence and robotics Ad Hoc Committee of ESMINT. He has contributed extensively to the field with 99 publications in Scopus and PubMed-indexed journals and holds an H-index of 20.