Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253134

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Gastrointestinal bleeding detection on digital subtraction angiography using convolutional neural networks with and without temporal information

Derek Smetanick[1]
Sailendra Naidu[2]
Alex Wallace[2]
M-Grace Knuttinen[2]
Indravadan Patel[2]
Sadeer Alzubaidi[2]

[1]The University of Arizona College of Medicine, Department of Interventional Radiology, Tucson, USA

[2]Mayo Clinic College of Medicine and Science, Department of Radiology, Phoenix, USA

Corresponding author: Derek Smetanick

E-mail: dereksmetanick@arizona.edu

Received 17 February 2025; revision requested 08 April 2025; last revision received 14 June 2025; accepted 22 June 2025.

Epub: 07.08.2025

DOI: 10.4274/dir.2025.253134

PURPOSE

Digital subtraction angiography (DSA) offers a real-time approach to locating lower gastrointestinal (GI) bleeding. However, many sources of bleeding are not easily visible on angiograms. This investigation aims to develop a machine learning tool that can locate GI bleeding on DSA prior to transarterial embolization.

METHODS

All mesenteric artery angiograms and arterial embolization DSA images obtained in the interventional radiology department between January 1, 2007, and December 31, 2021, were analyzed. These images were acquired using fluoroscopy imaging systems (Siemens Healthineers, USA). Thirty-nine unique series of bleeding images were augmented to train two-dimensional (2D) and three-dimensional (3D) residual neural networks (ResUNet++) for image segmentation. The 2D ResUNet++ network was trained on 3,548 images and tested on 394 images, whereas the 3D ResUNet++ network was trained on 316 3D objects and tested on 35 objects. For each case, both manually cropped images focused on the GI bleed and uncropped images were evaluated, with a superimposition post-processing (SIPP) technique applied to both image types.

RESULTS

Based on both quantitative and qualitative analyses, the 2D ResUNet++ network significantly outperformed the 3D ResUNet++ model. In the qualitative evaluation, the 2D ResUNet++ model achieved the highest accuracy across both 128 × 128 and 256 × 256 input resolutions when enhanced with the SIPP technique, reaching accuracy rates between 95% and 97%. However, despite the improved detection consistency provided by SIPP, a reduction in Dice similarity coefficients was observed compared with models without post-processing. Specifically, the 2D ResUNet++ model combined with SIPP achieved a Dice accuracy of only 80%. This decline is primarily attributed to an increase in false positive predictions introduced by the temporal propagation of segmentation masks across frames.

CONCLUSION

Both 2D and 3D ResUNet++ networks can be trained to locate GI bleeding on DSA images prior to transarterial embolization. However, further research and refinement are needed before this technology can be implemented in DSA for real-time prediction.

CLINICAL SIGNIFICANCE

Automated detection of GI bleeding in DSA may reduce time to embolization, thereby improving patient outcomes.

KEYWORDS

Convolutional neural networks, digital subtraction angiography, gastrointestinal bleeding, image segmentation, interventional radiology, machine learning

You may cite this article as: Smetanick D, Naidu S, Wallace A, Knuttinen MG, Patel I, Alzubaidi S. Gastrointestinal bleeding detection on digital subtraction angiography using convolutional neural networks with and without temporal information. *Diagn Interv Radiol.* 07 August 2025 DOI: 10.4274/dir.2025.253134 [Epub Ahead of Print].

Gastrointestinal (GI) bleeding involves active hemorrhaging from blood vessels within the GI tract. In 5%–10% of cases, patients require either surgery or transcatheter arterial embolization.[1] To perform transcatheter embolization, interventional radiologists often use digital subtraction angiography (DSA) to image the hemorrhage in real time. DSA works by visualizing contrast-opacified vessels and subtracting surrounding anatomical structures, such as soft tissues and bone, to provide a clearer view of the vascular system. The resulting images reveal areas where contrast "pools," indicating the site of bleeding to the interventional radiologist.[2] Although DSA offers a real-time method for locating bleeding, some sources may not be easily visible on angiograms. A neural network used as a decision support tool may assist radiologists in identifying bleeding sites prior to transcatheter arterial embolization.

Convolutional neural networks (CNNs) have demonstrated both accuracy and efficiency in object detection within images.[3] Ronneberger et al.[4] pioneered the U-Net architecture, an extension of the fully convolutional network, which includes a contracting path to capture image context and an expanding path to enable precise localization for segmentation. Neural networks based on the ResUNet architecture have addressed the high computational demands of three-dimensional (3D) convolutional networks.[5] Zhang et al.[6] implemented this design for road detection using a combination of upsampling and downsampling residual blocks. This model was further developed by Jha et al.[7], who proposed the residual neural networks (ResUNet++) architecture and tested it on a segmentation task to identify polyps in two-dimensional (2D) colonoscopy images. Given that ResUNet++ outperformed both the original ResUNet and U-Net models in image segmentation,[7] this architecture serves as the foundation for our model, which aims to segment GI bleeding on DSA images.

This study aims to investigate the utility of a deep learning approach for the automated detection of GI bleeding on DSA images, specifically by comparing 2D and 3D ResUNet++ architectures. We hypothesized that both models could identify bleeding sites, but that one may outperform the other. Our rationale for using a deep learning approach stems from the temporal variability and subtlety of GI bleeds, which may evade human detection on sequential angiographic images. Automated segmentation could assist radiologists by identifying bleeding pixels in real time, potentially reducing time to embolization. This study also evaluates a novel temporal consistency algorithm–superimposition post-processing (SIPP)–to determine whether incorporating temporal bleed memory improves segmentation performance across sequences. We address the following research questions. (1) Can deep learning accurately identify bleeding on DSA? (2) How does a 2D model compare with a 3D model in this context? (3) Does temporal information improve performance when integrated through post-processing?

It is also critical to consider the clinical impact of GI bleeding segmentation in DSA without introducing workflow delays. In practice, a supportive model must identify bleeding sites faster than the interventional radiologist to improve procedural outcomes. Earlier identification could reduce contrast volume, lower radiation exposure, and shorten procedure times.

Main points

- Automated image segmentation may play a beneficial role in detecting gastrointestinal (GI) bleeding in real time in digital subtraction angiography (DSA) prior to transarterial embolization.
- The three-dimensional (3D) residual neural networks use the temporal resolution from the DSA sequence to predict the bleeding location.
- The two-dimensional neural network outperformed the 3D neural network in segmenting GI bleeding on images.
- Increasing image resolution and using a graphics processing unit may improve both the accuracy of image segmentation and the processing speed, respectively.

Methods

Image datasets for training and testing

Mayo Clinic Phoenix approved this study as exempt on 01/31/2024 due to its retrospective nature (IRB application #: 24-000309). Between 2007 and 2021, a total of 96 patients underwent mesenteric artery angiography or arterial embolization DSA procedures for suspected GI bleeding. Of these, 70 patients showed no active extravasation on angiography and were excluded. The remaining 26 patients, who demonstrated confirmed active hemorrhage, were included in the study, as shown in Figure 1. No images were excluded based on patient age, motion artifacts, or image corruption. From the 26 patients, 39 unique image series positive for active hemorrhage were identified by an interventional radiologist and selected for neural network training. These cases involved hemorrhaging in the small and large intestines. On average, each series contained 11 bleeding images. To avoid inflated model performance, data were split at the patient level for training and testing. The bleeding images were cropped to highlight the hemorrhage in higher resolution. The dataset was then augmented by replicating each image nine times, systematically shifting the bleed location to the following regions: upper-left, upper-center, upper-right, middle-left, cen-

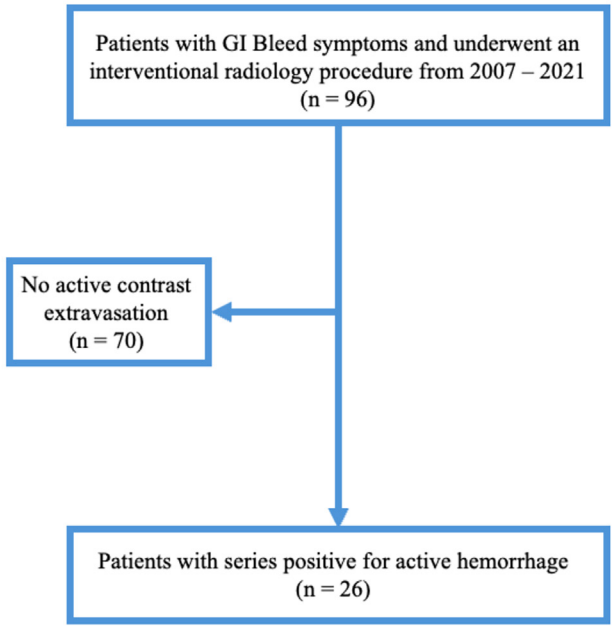

Figure 1. Criteria and number of patients from initial retrieval to the final study cohort. GI, gastrointestinal.

ter, middle-right, lower-left, lower-center, and lower-right. This approach increased the dataset size by 900%. Segmentation masks were created manually using Photoshop (Adobe Inc., San Jose, CA, USA) with a thresholding tool to isolate the bleeding. The segmentations displayed bleeding areas in white on a solid black background to produce binary images. The same augmentation technique was applied to the segmentation masks to ensure proper pixel alignment with the original images. Table 1 summarizes the number of GI bleeding-positive and-negative images in the test set after augmentation.

Although 70 patients had no visible extravasation, including all of their image sequences as negative controls would have created a heavily imbalanced dataset. Instead, non-bleeding frames from within the same DSA sequences of the 26 bleeding-positive patients were used. These frames provided sufficient negative control data for training and testing while preserving representative angiographic conditions and avoiding overrepresentation of non-bleeding cases. Moreover, the model's task was to identify where bleeding occurred, rather than whether bleeding was present. In this context, even within bleeding-positive images, the majority of pixels are negative for bleeding.

Both 128 × 128-pixel and 256 × 256-pixel images were used to train separate 2D CNNs, whereas only 128 × 128-pixel images were used to train a 3D CNN for image segmentation. A post-processing technique–superimposing all masks within a series into a single mask for final image segmentation–was applied to both 2D and 3D segmentations. In total, these three networks were evaluated across four distinct testing scenarios: (1) uncropped images from the DSA sequence, (2) cropped images focusing on the bleed, (3) uncropped images with the SIPP technique applied, and (4) cropped images with the SIPP technique.

Superimposition post-processing technique

The SIPP technique algorithm was developed to address the temporal inconsistency of GI bleeding predictions across angiographic image sequences. Bleeding may not be clearly visible in every frame. To mitigate this, SIPP enforces temporal continuity by propagating the presence of bleeding pixels forward through the predicted image sequence. For each frame in the sequence, the model produces a binary segmentation mask M_t, where each pixel is labeled either as 1 (bleeding present) or 0 (no bleeding). The mask $M_t \in \{0,1\}^{\{H \times W\}}$ is a 2D grid with the same height (H) and width (W) as the original image and represents the classification of each pixel. SIPP modifies these predictions by updating each new mask M_t to include any pixel that was previously marked as bleeding. This is defined as:

$$\widehat{M}_t = M_t \vee \widehat{M}_{\{t-1\}}$$

Where ∨ represents a logical OR operation performed on all pixels between the current prediction M_t and the accumulated mask from the previous frame $\widehat{M}_{t-1}$. This rule ensures that once a pixel is marked as bleeding, it remains labeled as such in all following frames of the DSA. This effectively preserves prior bleeding evidence even if the current frame is less confident. This simple yet effective mechanism improves temporal consistency and reduces missed detections due to frame-level variability. A flowchart is provided in Figure 2 to further explain the SIPP technique.

Deep neural network architecture

Both a 3D and a 2D ResUNet++ were constructed based on the architecture shown in Figure 3. A single DSA frame served as the input image for the 2D network, whereas the entire DSA series served as the input for the 3D network. After entering the network, the image passed through a series of convolutional layers with a 3 × 3 kernel size and increasing numbers of filters (16, 32, 48, and 64), referred to as the encoding phase. Each convolutional layer was followed by batch normalization to improve training speed and

Table 1. The number of images positive for gastrointestinal bleeding and the number of control images negative for gastrointestinal bleeding were tabulated for each of the different models

	GI bleed images	Control images
2D uncropped 128 × 128	343	95
2D cropped 128 × 128	321	73
2D uncropped 256 × 256	354	84
2D cropped 256 × 256	321	73
3D uncropped 128 × 128	343	281
3D cropped 128 × 128	273	287
2D uncropped 128 × 128 w/SIPP	343	281
2D cropped 128 × 128 w/SIPP	3195	2421
2D uncropped 256 × 256 w/SIPP	354	270
2D cropped 256 × 256 w/SIPP	3196	2420
3D uncropped 128 × 128 w/SIPP	319	241
3D cropped 128 × 128 w/SIPP	273	287

GI, gastrointestinal; SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional.

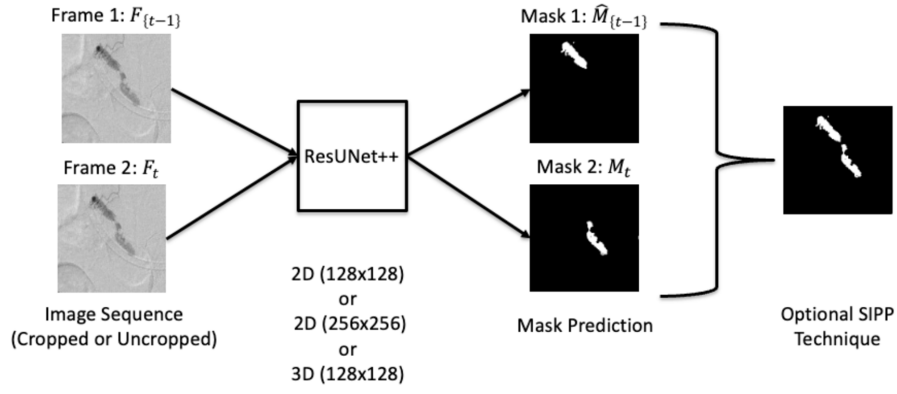

Figure 2. Schematic of the image segmentation pipeline with the optional superimposition post-processing technique. Each image sequence (cropped or uncropped) is passed through a ResUNet++ model configured as either 2D (128 × 128 or 256 × 256) or 3D (128 × 128) to generate frame-wise predicted masks. If applied, the SIPP technique performs a logical OR operation between the current and previous masks to enhance temporal consistency in bleeding detection. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

stability by standardizing the inputs. A rectified linear unit (ReLU) activation function was then applied to introduce non-linearity, enabling the network to learn complex patterns and shapes. The spatial dimensions of the feature maps were reduced through 2D max pooling after each convolutional layer, allowing the network to retain the most important features. After the encoding phase, the features were upsampled back to the original image size using transposed convolutions with a 3 × 3 kernel size and decreasing numbers of filters (64, 48, 32, and 16). Each layer was again followed by batch normalization and ReLU activation. At each step of the decoding path, the feature maps were concatenated with the corresponding feature maps from the encoding phase, allowing the network to leverage both low-level and high-level features for more accurate segmentation. The final output layer consisted of a 1 × 1 convolutional layer with a single filter and sigmoid activation. The resulting segmentation map assigned each pixel a predicted class. Both the 2D and 3D ResUNet++ models described in this study were deep learning architectures designed for semantic segmentation tasks. Although implemented as machine learning models during training and inference, their structural design–comprising convolutional layers, encoding–decoding paths, and feature concatenations–was fundamentally that of deep learning architectures.

The convolutional ResUNet++ networks were implemented using the Keras framework[8] with a TensorFlow backend (Google, Inc.),[9] using Python version 3.9. All experiments were performed on a computer with an Intel Core i7-8700 central processing unit (CPU) @ 3.20 GHz (Intel). To prevent overfitting, a smaller learning rate of 1.0×10^{-4} was used during training to avoid issues such as model instability or failure to converge. Data augmentation was also applied to artificially increase dataset variability, further helping to mitigate overfitting. The architecture was optimized using the Adam optimizer. A batch size of 20 and 20 training epochs were used for each experiment to maintain consistency. Binary cross-entropy loss was employed to optimize the segmentation task.

Quantitative evaluation

The MATLAB software (MathWorks, Natick, MA, USA) was used to quantify the results from predicted and actual masks by measuring mask overlap. A pixel-by-pixel analysis identified true positive pixels (TPP), true negative pixels (TNP), false positive pixels (FPP), and false negative pixels (FNP). TP and TN values were calculated by dividing TPP and TNP by the respective numbers of positive and negative pixels in the ground truth. FP and FN values were calculated by dividing FPP and FNP by the total number of pixels in the ground truth, respectively. These scores were computed for each of the 12 experiments. Dice similarity coefficients (DSCs) were calculated to quantitatively assess the spatial overlap between the predicted segmentation masks and the ground truth annotations. For each model and imaging configuration, the Dice coefficients were computed on a per-sample basis and summarized as mean values with corresponding 95% confidence intervals (CIs). All Dice analysis was performed as part of the quantitative evaluation.

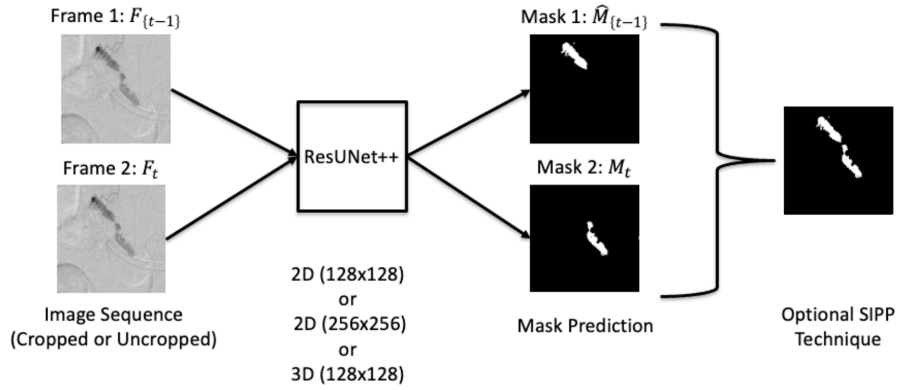

Figure 3. Neural network architecture used in both the 2D ResUNet++ and 3D ResUNet++ models. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

Qualitative evaluation

Although quantitative metrics provided objective measures of segmentation accuracy, a qualitative evaluation was also conducted to assess clinical relevance. This evaluation was performed by a single evaluator–a medical student–who visually compared the predicted segmentation masks with both the ground truth masks and the original DSA images. Each image was classified as TP, TN, FP, or FN using the same definitions applied in the quantitative evaluation. To aid in the classification process, a MATLAB script was used to help identify TN, FP, and FN images. A prediction was considered a TP if the white pixels in the predicted mask overlapped with those in the ground truth mask. This overlap was initially assessed visually and subsequently verified to ensure at least one pixel of overlap, which served as a safeguard to minimize human error in classification. This minimal overlap threshold was intentionally selected based on the model's intended clinical use: to serve as a real-time assistive tool during embolization procedures. In such settings, even a small correctly flagged area could be sufficient to prompt further investigation by an interventional radiologist. The model is not intended to deliver volumetric precision but rather to alert clinicians to potential regions of bleeding. Cases where the white pixels of the predicted and ground truth masks overlapped but also included some FP areas were generally classified as TPs, unless the FP region exceeded 10% of the image area. All ground truth segmentation masks were manually created using a thresholding method and validated by a team of fellowship-trained interventional radiologists to ensure accuracy before comparison.

Statistical analysis

A one-way analysis of variance single-factor test was conducted in MATLAB to determine the statistical significance within the TP results of the quantitative evaluation. An α value of 0.05 was selected, with the null hypothesis stating that there is no statistically significant difference among the various networks. If the *P* values obtained from the analysis were less than α, the null hypothesis was rejected, indicating a statistically significant difference between the networks. In cases where such a difference was detected, a Tukey–Kramer post-hoc test was performed to identify which networks exhibited this disparity.

Results

Quantitative evaluation

The accuracy, intersection over union (IoU), loss, and precision obtained during the initial training and testing of the 2D 128

× 128 ResUNet++, 2D 256 × 256 ResUNet++, and 3D 128 × 128 ResUNet++ models are presented in Table 2. The accuracy and precision scores were comparable across all three networks. The 3D 128 × 128 ResUNet++ exhibited the lowest IoU at 0.06.

Depicted in Figure 4 are the accuracy scores for the 12 different 2D and 3D ResUNet++ structures based on a DSA frame-wise basis. These results are summarized in Table 3, whereas outcomes of the statistical analysis are presented in Table 4. A statistically significant improvement in the accuracy score was observed using the SIPP technique for all six different ResUNet++ structures: 2D uncropped 128 × 128, 2D cropped 128 × 128, 2D uncropped 256 × 256, 2D cropped 256 × 256, 3D uncropped 128 × 128, and 3D cropped 128 × 128 compared with the control trial. Notably, there was no statistical significance between the 2D uncropped 128 × 128 model and the 2D uncropped 256 × 256 model, the 2D uncropped 128 × 128 model and the 3D uncropped 128 × 128 model, the 2D cropped 128 × 128 model and the 2D cropped 256 × 256 model, and the 2D uncropped 256 × 256 and the 3D uncropped 128 × 128 model when the SIPP method was not used. The largest mean accuracy values were 0.961 and 0.956 for the 2D cropped 128 × 128 with SIPP model and the 2D cropped 256 × 256 with SIPP model, respectively. There was no statistically significant difference between the accuracy values for these two different networks. Both models had a statistically significantly higher accuracy than the 3D cropped 128 × 128 model with SIPP. The 2D cropped 128 × 128 and the 2D cropped 256 × 256 models also maintained the highest accuracy for models without SIPP, with accuracy scores of 0.853 and 0.812, respectively. There was no statistically significant difference between these two models. These models had a statistically significantly higher accuracy than the 3D cropped 128 × 128 model. The 2D uncropped 256 × 256 with SIPP model had a statistically significantly higher accuracy than the 2D uncropped 128 × 128 with SIPP model and the 3D uncropped 128 × 128 with SIPP model. Meanwhile, the 2D uncropped 128 × 128 with SIPP model had a statistically significantly higher accuracy than the 3D uncropped 128 × 128 with SIPP model.

DSCs for each model configuration, with and without SIPP, are summarized in Table 5. Compared with their corresponding original models, the use of SIPP resulted in statistically significant reductions in Dice coefficients for the 2D uncropped 128 × 128 model [from 0.042 (95% CI: 0.0264–0.0575) to 0.019 (95% CI: 0.0133–0.0248)] and the 2D cropped 128 × 128 model [from 0.798 (95% CI: 0.7720–0.8232) to 0.190 (95% CI: 0.1839–0.1959)]. Similarly, the 2D cropped 256 × 256 model exhibited a substantial decrease in Dice score when SIPP was applied [from 0.797 (95% CI: 0.7708–0.8223) to 0.278

Table 2. The results from training the 2D ResUNet++ on 128 × 128-pixel and 256 × 256-pixel images, as well as the 3D ResUNet++ on 128 × 128-pixel images, are tabulated. The metrics of accuracy, intersection-over-union, and precision were included for all three neural networks

Method	Accuracy	IoU	Precision
2D 128 × 128 ResUNet++	**0.95**	0.62	0.99
2D 256 × 256 ResUNet++	**0.96**	0.61	0.98
3D 128 × 128 ResUNet++	**0.96**	0.06	0.95

2D, two-dimensional; 3D, three-dimensional; IoU, intersection-over-union; ResUNet++, residual neural networks.

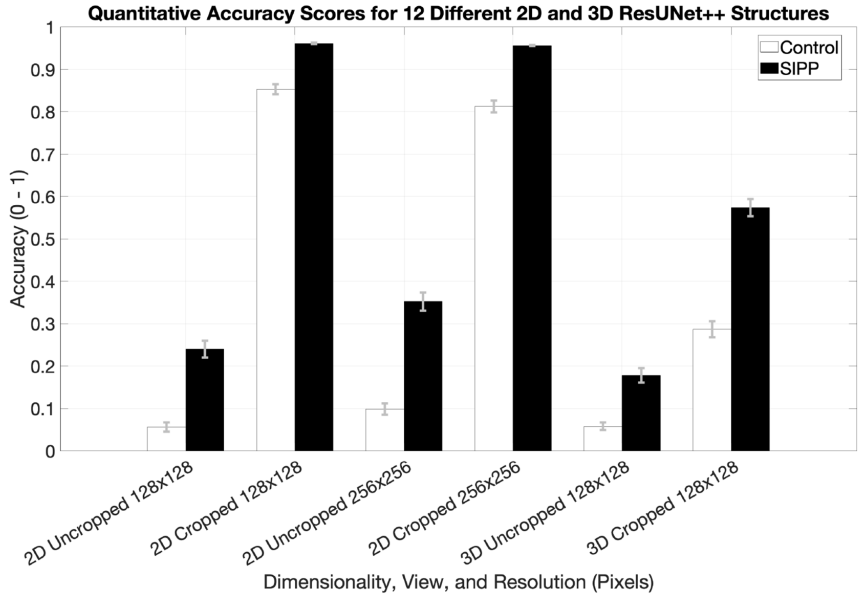

Figure 4. Bar graph showing differences in the quantitative accuracy of segmentation results for the 12 testing scenarios. The control represents cases without post-processing, whereas the other cases used the superimposition post-processing technique. Error bars indicate one standard deviation. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

Table 3. The true positive, true negative, false positive, and false negative rates were tabulated for the 12 different cases for the quantitative results

	True positive	True negative	False positive	False negative
2D uncropped 128 × 128	0.056	0.966	0.034	0.001
2D cropped 128 × 128	0.853	0.996	0.003	0.002
2D uncropped 256 × 256	0.099	0.978	0.022	0.001
2D cropped 256 × 256	0.812	0.997	0.002	0.003
3D uncropped 128 × 128	0.058	0.998	0.002	0.001
3D cropped 128 × 128	0.287	0.999	0.001	0.006
2D uncropped 128 × 128 w/SIPP	0.240	0.919	0.081	0.001
2D cropped 128 × 128 w/SIPP	0.961	0.898	0.098	0
2D uncropped 256 × 256 w/SIPP	0.352	0.972	0.028	0.001
2D cropped 256 × 256 w/SIPP	0.956	0.946	0.051	0.001
3D uncropped 128 × 128 w/SIPP	0.178	0.985	0.015	0.001
3D cropped 128 × 128 w/SIPP	0.573	0.982	0.018	0.003

SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional.

(95% CI: 0.2703–0.2858)]. In contrast, for the 2D uncropped 256 × 256, 3D uncropped 128 × 128, and 3D cropped 128 × 128 models, although minor changes in Dice coefficients were observed, the corresponding 95% CIs overlapped. Therefore, these changes are not statistically significant based on CI analysis. Overall, these results indicate that although SIPP altered segmentation performance, its effects were not uniformly beneficial across all models, and in some cases, it led to considerable declines in segmentation accuracy. These results are visually represented in Figure 5.

Qualitative evaluation

An example image from the 2D cropped 128 × 128 model, the 2D uncropped 256 × 256 model, and the 3D cropped 128 × 128 model is shown in Figure 6. The original image is on the left, the ground truth is in the middle, and the predicted image is on the right. Each image was reviewed manually for quality control to compare the ground truth with the predicted image. The results from the qualitative evaluation are displayed in Table 6 and plotted in Figure 7.

From the highest TP accuracy count to the lowest TP count, the twelve networks ranked as follows for the qualitative results: 2D cropped 256 × 256 with SIPP, 2D cropped 128 × 128 with SIPP, 2D cropped 128 × 128, 2D cropped 256 × 256, 3D cropped 128 × 128 with SIPP, 3D cropped 128 × 128, 2D uncropped 256 × 256 with SIPP, 2D uncropped 128 × 128 with SIPP, 3D uncropped 128 × 128 with SIPP, 2D uncropped 256 × 256, 3D uncropped 128 × 128, and 2D uncropped 128 × 128. The range of TP accuracy was from 0.999 to 0.122. The models using the SIPP technique had higher accuracy rates than their control counterparts. The ranking order was similar to the TP accuracies from the quantitative section. The main differences in the qualitative list compared with the quantitative list are that 2D cropped 256 × 256 with SIPP marginally outperformed 2D cropped 128 × 128 with SIPP, and 3D cropped 128 × 128 marginally outperformed 2D uncropped 256 × 256 with SIPP.

Discussion

The widely used U-Net architecture for medical image segmentation is leveraged in this study through the ResUNet++ variant. ResUNet preserves input dimensions and minimizes information loss, as described by Yousef et al.[10], whereas U-Net++ incorporates nested skip connections to enhance semantic segmentation, as detailed by Zhou et al.[11] The effectiveness of ResUNet++ has been validated by Jha et al.[7], supporting its use in segmentation tasks. This study evaluates segmentation accuracy using standard analyses similar to those employed in cone-beam CT acquisitions for prostate treatments.[12]

Using 2D ResUNet++ for DSA images offers distinct advantages over 3D ResUNet++. Although 3D ResUNet++ benefits from incorporating temporal information across image sequences, it did not outperform the 2D model. For uncropped DSA images, 3D ResUNet++ performed similarly to 2D ResUNet++, likely because downscaling the original 1064 × 1064-pixel images to 128 × 128 or 256 × 256 pixels led to a loss of crucial spatial detail. This limitation was addressed by manually cropping the images to focus specifically on bleeding regions, allowing the bleed to occupy approximately 5% of the image area and substantially improving training and testing resolution. This process improved segmentation accuracy for both 2D and 3D ResUNet++ models, emphasizing the importance of image resolution for accurate GI bleeding localization and favoring 2D model performance. These quantitative findings were further supported by qualitative assessments.

The Keras framework[8] was used to evaluate accuracy, IoU, loss, and precision metrics during the training of both 2D and 3D ResUNet++ models on cropped images. Across

Table 4. A one-way analysis of variance with a Tukey–Kramer post-hoc test was conducted, and the resulting *P* values were tabulated to compare different models. The significance level (α) was set at 0.05. Statistical differences in segmentation accuracy were observed for models with *P* values less than α

Model 1	Model 2	*P* value
2D uncropped 128 × 128	2D uncropped 128 × 128 w/SIPP	<0.001
2D cropped 128 × 128	2D cropped 128 × 128 w/SIPP	<0.001
2D uncropped 256 × 256	2D uncropped 256 × 256 w/SIPP	<0.001
2D cropped 256 × 256	2D cropped 256 × 256 w/SIPP	<0.001
3D uncropped 128 × 128	3D uncropped 128 × 128 w/SIPP	<0.001
3D cropped 128 × 128	3D cropped 128 × 128 w/SIPP	<0.001
2D uncropped 128 × 128	2D uncropped 256 × 256	0.098
2D uncropped 128 × 128	3D uncropped 128 × 128	1.000
2D cropped 128 × 128	2D cropped 256 × 256	0.192
2D uncropped 256 × 256	3D uncropped 128 × 128	0.143
2D cropped 128 × 128 w/SIPP	2D cropped 256 × 256 w/SIPP	0.995
2D cropped 256 × 256 w/SIPP	3D cropped 128 × 128 w/SIPP	<0.001
2D cropped 128 × 128 w/SIPP	3D cropped 128 × 128 w/SIPP	<0.001
2D cropped 256 × 256	3D cropped 128 × 128	<0.001
2D cropped 128 × 128	3D cropped 128 × 128	<0.001
2D uncropped 256 × 256 w/SIPP	2D uncropped 128 × 128 w/SIPP	<0.001
2D uncropped 256 × 256 w/SIPP	3D uncropped 128 × 128 w/SIPP	<0.001
2D uncropped 128 × 128 w/SIPP	3D uncropped 128 × 128 w/SIPP	0.001

SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional.

Table 5. Mean Dice similarity coefficients and corresponding 95% confidence intervals are reported for each model configuration, comparing results with and without superimposition post-processing. Statistically significant differences, identified by non-overlapping confidence intervals, are indicated in bold

Model	Original mean (95% CI)	SIPP mean (95% CI)
2D uncropped 128 × 128	0.042 [0.0264, 0.0575]	**0.019 [0.0133, 0.0248]**
2D cropped 128 × 128	0.798 [0.7720, 0.8232]	0.190 [0.1839, 0.1959]
2D uncropped 256 × 256	0.069 [0.0493, 0.0893]	**0.065 [0.0533, 0.0757]**
2D cropped 256 × 256	0.797 [0.7708, 0.8223]	0.278 [0.2703, 0.2858]
3D uncropped 128 × 128	0.054 [0.0381, 0.0694]	0.064 [0.0479, 0.0795]
3D cropped 128 × 128	0.334 [0.2955, 0.3731]	0.281 [0.2523, 0.3104]

CI, confidence interval; SIPP, superimposition post-processing.

all metrics, the 2D ResUNet++ outperformed its 3D counterpart. Higher IoU indicates superior segmentation, and although 3D ResUNet++ had a lower IoU, its performance improved following the application of the SIPP technique. SIPP accumulates bleeding-positive pixels across sequential frames, enhancing temporal consistency. Originally applied to 3D ResUNet++ to address intermittent bleeding visibility, SIPP also improved segmentation performance for 2D ResUNet++ models. However, quantitative analysis revealed that SIPP increased FP rates, as errors persisted across frames, whereas FN rates remained relatively unaffected by post-processing. The increase in FP rates resulting from the SIPP technique also contributed to a decrease in DSCs across most models. Since the Dice coefficient is sensitive to both FPs and FNs, the propagation of errors across sequential frames reduced overall spatial overlap precision, despite improvements in bleeding pixel continuity. This tradeoff highlights an important limitation of SIPP: although it enhances temporal consistency and bleed detection sensitivity, it may compromise segmentation specificity, as reflected in Dice score reductions.

Since transarterial embolization is performed in real time under fluoroscopy, model inference speed is critical. Doubling image resolution from 128 × 128 to 256 × 256 pixels nearly quadrupled the model runtime. Interestingly, there was no statistically significant difference in runtime between 2D ResUNet++ trained on 256 × 256 images and 3D ResUNet++ trained on 128 × 128 images, indicating that 3D models also demand substantial computational resources. Prior studies using graphics processing unit (GPU) hardware have demonstrated that 512 × 512-pixel images can be segmented in less than 1 second,[4] suggesting that GPU acceleration could greatly enhance model performance and enable the training of higher-resolution 3D networks. Although training on a GPU would have considerably expedited model development, cost constraints and limited institutional access to dedicated GPU hardware necessitated CPU-based training in this study. For future real-time deployment, GPU acceleration will be critical to support high-throughput inference and maintain clinical usability.

Ground truth segmentation quality greatly impacts machine learning model performance. To ensure reliable labeling, ground truth masks underwent rigorous validation. A chart review was conducted to confirm each bleeding episode's anatomical site, with

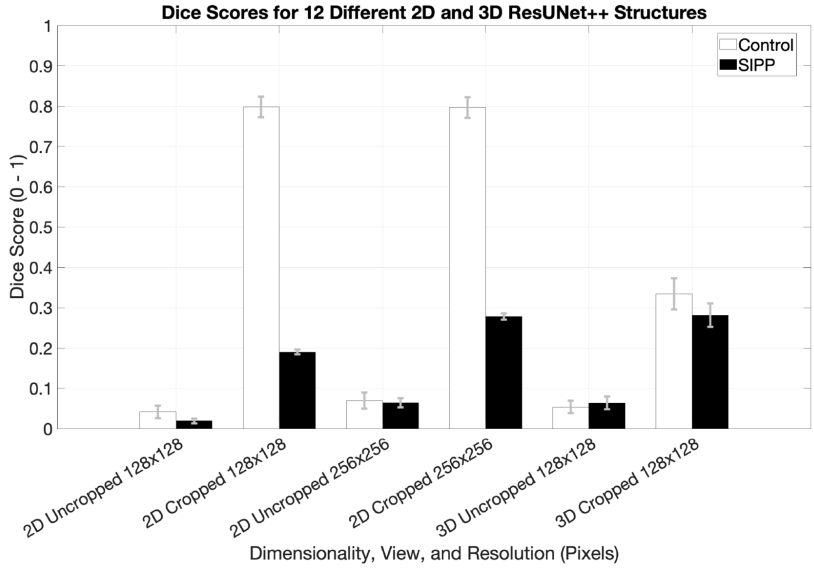

Figure 5. Mean Dice similarity coefficients and 95% confidence intervals for twelve different 2D and 3D ResUNet++ segmentation models, evaluated with and without SIPP. Bars indicate the mean DSC values, and error bars represent the corresponding 95% confidence intervals. Dice coefficients range from 0 (no overlap) to 1 (perfect overlap). SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks; DSC, Dice similarity coefficients.

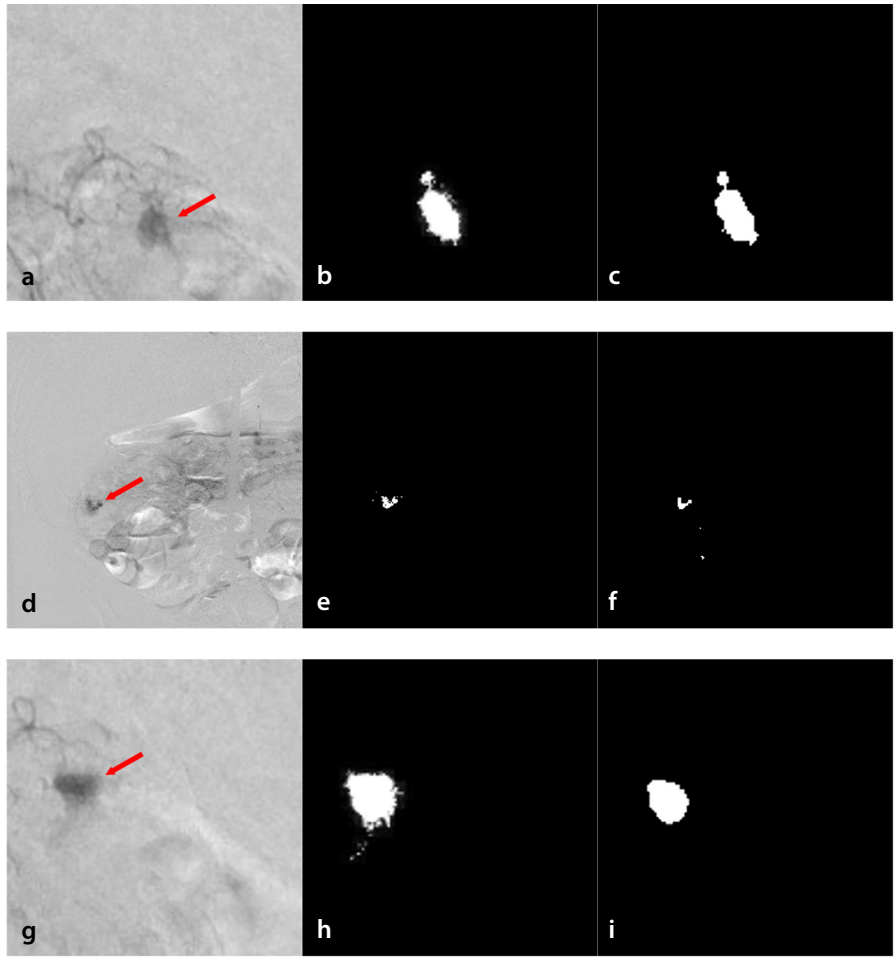

Figure 6. Results for the 2D and 3D ResUNet++ models: **(a)** original image tested on the 2D cropped 128 × 128 model; **(b)** ground truth; **(c)** predicted image; **(d)** image tested on the 2D uncropped 256 × 256 model; **(e)** corresponding ground truth; **(f)** predicted image; **(g)** image tested on the 3D cropped 128 × 128 model; **(h)** ground truth; **(i)** predicted image. Red arrows in **(a)**, **(d)**, and **(g)** point toward the GI bleed. GI, gastrointestinal; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

ambiguities resolved in consultation with fellowship-trained interventional radiologists. Manual image thresholding using MATLAB assigned white pixel values (255) to bleeding regions and black (0) to background areas, creating binary masks. Although manual segmentation is labor-intensive, it remains the gold standard for validation, as emphasized by Yepes-Calderon et al.[13] Potential operator bias was minimized by having a single individual perform all segmentations. Data augmentation techniques, including cropping and translation, as described by Shorten et al.,[14] expanded the training dataset. Cropping enhanced effective resolution, and systematic translations increased the dataset size by 900%. Due to image series grouping for 3D training, fewer images were available for the 3D models compared with the 2D models. Further research is needed to assess how expanded 3D datasets could impact model performance.

When comparing Tables 2 and 4, an apparent contradiction emerges because models such as 2D cropped 128 × 128 with SIPP and 2D cropped 256 × 256 with SIPP show high true positive rates (TPRs) and true negative rates (TNRs) in Table 2 yet exhibit a notable decrease in DSCs in Table 4. This discrepancy stems from fundamental differences in how these metrics are calculated. TPRs and TNRs incorporate TNs, which dominate pixel-based segmentation tasks and can inflate performance metrics, particularly when background regions vastly outnumber bleeding pixels. In contrast, the Dice coefficient is a spatial overlap metric that does not consider TNs and is highly sensitive to both FPs and FNs. Since the SIPP technique propagates predictions across frames, it can increase FPs and lead to reduced Dice scores despite stable or improved TPRs and TNRs. This tradeoff underscores a central tension in medical image segmentation: balancing sensitivity and temporal consistency with spatial specificity. Given the model's intended role as an assistive tool during real-time embolization, the slight increase in FPs introduced by SIPP may be clinically acceptable if it ensures that critical bleeding regions are not missed. Both methods were incorporated in this study for transparency.

In Table 6, some models display a TNR of 1.0 while still reporting a nonzero FPR. This discrepancy stems from differences in denominator definitions: TPs and TNs were calculated relative to the number of positive and negative pixels in the ground truth, whereas FPs and FNs were normalized over the total number of pixels in the image. As a result, even a small number of FPPs yields a measurable FPR despite a perfect TNR. This normalization strategy was chosen to consistently reflect prediction error impacts across images of varying sizes and class balances.

Recent studies have further demonstrated the potential of machine learning for DSA-based bleeding detection. Barash et al.[15] utilized a CNN to classify DSA images as either normal or containing active bleeding, achieving an area under the curve of 85.0% and an accuracy of 77.43%. Similarly, Liu et al.[16] introduced a method using parametric color imaging to enhance DSA sequences and better localize bleeding points. Additionally, Min et al.[17] developed a two-stage deep learning model, "InterNet," to detect active abdominal arterial bleeding on emergency DSA images. Their model considerably improved workflow efficiency, reducing radiologist interpretation time from 84.88 to 43.78 seconds. This highlights the potential of artificial intelligence tools to expedite bleeding detection during high-stakes procedures. Compared with these classification-based approaches, the present study focuses on semantic segmentation to directly identify and localize bleeding regions at the pixel level. Furthermore, our study introduces the SIPP technique to enhance temporal consistency.

Table 6. The true positive, true negative, false positive, and false negative rates were tabulated for the 12 different cases in the qualitative results

	True positive rate	True negative rate	False positive rate	False negative rate
2D uncropped 128 × 128	0.122	1	0.402	0.285
2D cropped 128 × 128	0.969	1	0.003	0.023
2D uncropped 256 × 256	0.22	1	0.338	0.292
2D cropped 256 × 256	0.953	1	0	0.038
3D uncropped 128 × 128	0.163	1	0.149	0.311
3D cropped 128 × 128	0.597	1	0.014	0.182
2D uncropped 128 × 128 w/SIPP	0.408	1	0.325	0
2D cropped 128 × 128 w/SIPP	0.997	1	0.002	0
2D uncropped 256 × 256 w/SIPP	0.571	1	0.244	0
2D cropped 256 × 256 w/SIPP	0.999	1	0	0
3D uncropped 128 × 128 w/SIPP	0.276	1	0.234	0.179
3D cropped 128 × 128 w/SIPP	0.853	1	0.018	0.054

SIPP, superimposition post-processing.

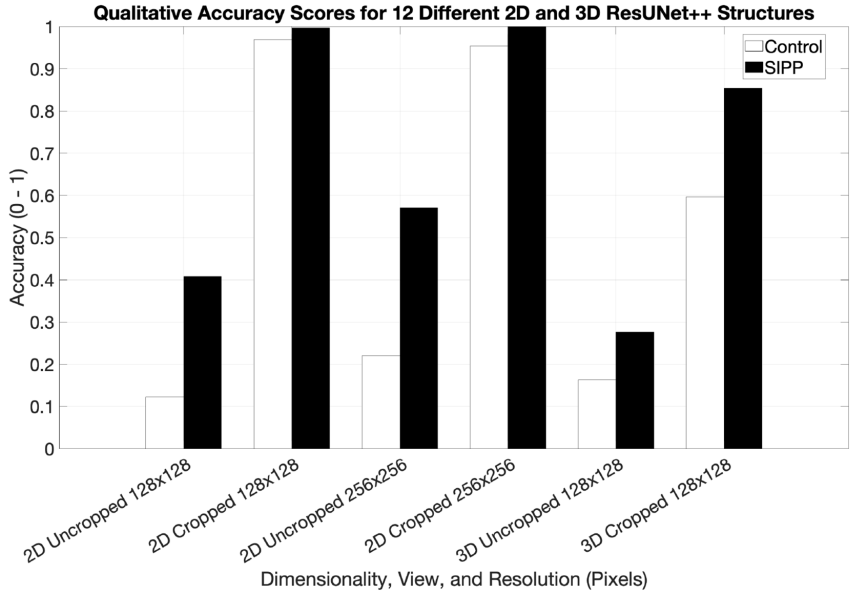

Figure 7. Bar graph showing differences in the qualitative accuracy of segmentation results for the 12 testing scenarios. The "control" represents cases without post-processing, whereas the other cases used the SIPP technique. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

Limitations

Several limitations must be acknowledged. First, the sample size was relatively small (26 patients), limiting statistical power and generalizability. Second, no external validation set from a separate institution was used, raising concerns about model robustness across different imaging protocols and vendors. Third, training was performed on a CPU rather than a GPU, which constrained image resolution, limited model complexity, slowed inference speeds, and necessitated manual cropping of bleeding regions to preserve resolution for training. Although necessary under computational constraints, manual cropping introduces potential bias and is not feasible for clinical deployment. In future work, GPU-accelerated training and inference will be pursued to allow the processing of entire uncropped DSA images at full resolution. Alternatively, a sliding window approach could be implemented, whereby the model systematically analyzes overlapping regions of the full image to detect bleeding without manual preselection. Fourth, the dataset included only bleeding-positive cases, limiting the ability to fully assess FPRs and overall specificity. Future studies can address these limitations by expanding datasets, incorporating external validation cohorts, utilizing GPU acceleration, and including negative control cases to better assess real-world model performance.

In conclusion, this study investigated the use of 2D ResUNet++ and 3D ResUNet++ neural network models to segment GI bleeding in DSA prior to transarterial embolization. Most notably, the 2D ResUNet++ outperformed the 3D ResUNet++ model. In qualitative analysis, the 2D ResUNet++ model achieved the highest accuracy, ranging from 95% to 97%, when enhanced with the SIPP technique. The highest DSC observed was 80% for the same model. Both quantitative and qualitative analyses highlight the potential feasibility of this model for real-time bleeding segmentation in the interventional radiology suite. Furthermore, training and testing with more 3D data are recommended to further refine the performance of the 3D ResUNet++ model. Incorporating GPU acceleration is also advised for faster processing. Future studies should evaluate the impact of these tools on DSA images in real time.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Shin JH. Refractory gastrointestinal bleeding: role of angiographic intervention. *Clin Endosc*. 2013;46(5):486-491. [Crossref]
2. Taslakian B, Ingber R, Aaltonen E, Horn J, Hickey R. Interventional radiology suite: a primer for trainees. *J Clin Med*. 2019;8(9):1347. [Crossref]
3. Ajit A, Acharya K, Samanta A. A review of convolutional neural networks. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE). IEEE; 2020:1-5. [Crossref]
4. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *Arxiv*. [Crossref]
5. Qiu Z, Yao T, Mei T. Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Trans Multimedia*. 2018;20(4):939-949. [Crossref]
6. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. *Arxiv*. [Crossref]
7. Jha D, Smedsrud PH, Riegler MA, et al. ResUNet++: an advanced architecture for medical image segmentation. *Arxiv*. [Crossref]
8. Chollet F. 2015. Keras. GitHub. [Crossref]
9. Keeton K. Proceedings of the 12th USENIX conference on operating systems design and implementation. USENIX Association; 2016. [Crossref]
10. Yousef R, Khan S, Gupta G, et al. U-Net-based models towards optimal MR brain image segmentation. *Diagnostics*. 2023;13(9):1624. [Crossref]
11. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a Nested U-Net architecture for medical image segmentation. deep learn med image anal multimodal learn clin decis support. 2018;11045:3-11. [Crossref]
12. Tegtmeier RC, Kutyreff CJ, Smetanick JL, et al. Custom-trained deep learning-based auto-segmentation for male pelvic iterative CBCT on C-arm linear accelerators. *Pract Radiat Oncol*. 2024;14(5):e383-e394. [Crossref]
13. Yepes-Calderon F, McComb JG. Eliminating the need for manual segmentation to determine size and volume from MRI. A proof of concept on segmenting the lateral ventricles. *PLoS One*. 2023;18(5):e0285414. [Crossref]
14. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60. [Crossref]
15. Barash Y, Livne A, Klang E, et al. Artificial intelligence for identification of images with active bleeding in mesenteric and celiac arteries angiography. *Cardiovasc Intervent Radiol*. 2024;47(6):785-792. [Crossref]
16. Liu J, Zhou X, Guan W, Gong S, Liu J. Research on detection method of bleeding point in two-dimensional DSA image based on parametric color imaging. *Comput Biol Med*. 2022;146:105496. [Crossref]
17. Min X, Feng Z, Gao J, et al. InterNet: detection of active abdominal arterial bleeding using emergency digital subtraction angiography imaging with two-stage deep learning. *Front Med (Lausanne)*. 2022;9:762091. [Crossref]