Artificial Intelligence in the Life Sciences 4 (2023) 100084

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Analysis of Swin-UNet vision transformer for Inferior Vena Cava filter segmentation from CT scans ☆

Rahul Gomes [a,*], Tyler Pham [b], Nichol He [a], Connor Kamrowski [a], Joseph Wildenberg [c,**]

[a] *Department of Computer Science, University of Wisconsin-Eau Claire, Eau Claire, 54701, WI, United States*
[b] *Department of Computer Science, University of Minnesota-Twin Cities, Minneapolis, 55455, MN, United States*
[c] *Interventional Radiology, Mayo Clinic Health System, Eau Claire, 54703, WI, United States*

ARTICLE INFO

Dataset link: https://github.com/rahulgomes19/IVC_3D

Keywords:
Deep learning
Medical imaging
Convolutional neural networks
SWIN transformer
UNet
ResNet
IVC filter

ABSTRACT

Purpose: The purpose of this study is to develop an accurate deep learning model capable of Inferior Vena Cava (IVC) filter segmentation from CT scans. The study does a comparative assessment of the impact of Residual Networks (ResNets) complemented with reduced convolutional layer depth and also analyzes the impact of using vision transformer architectures without performance degradation.
Materials and Methods: This experimental retrospective study on 84 CT scans consisting of 54618 slices involves design, implementation, and evaluation of segmentation algorithm which can be used to generate a clinical report for the presence of IVC filters on abdominal CT scans performed for any reason. Several variants of patch-based 3D-Convolutional Neural Network (CNN) and the Swin UNet Transformer (Swin-UNETR) are used to retrieve the signature of IVC filters. The Dice Score is used as a metric to compare the performance of the segmentation models.
Results: Model trained on UNet variant using four ResNet layers showed a higher segmentation performance achieving median Dice = 0.92 [Interquartile range(IQR): 0.85, 0.93] compared to the plain UNet model with four layers having median Dice = 0.89 [IQR: 0.83, 0.92]. Segmentation results from ResNet with two layers achieved a median Dice = 0.93 [IQR: 0.87, 0.94] which was higher than the plain UNet model with two layers at median Dice = 0.87 [IQR: 0.77, 0.90]. Models trained using SWIN-based transformers performed significantly better in both training and validation datasets compared to the four CNN variants. The validation median Dice was highest in 4 layer Swin UNETR at 0.88 followed by 2 layer Swin UNETR at 0.85.
Conclusion: Utilization of vision based transformer Swin-UNETR results in segmentation output with both low bias and variance thereby solving a real-world problem within healthcare for advanced Artificial Intelligence (AI) image processing and recognition. The Swin UNETR will reduce the time spent manually tracking IVC filters by centralizing within the electronic health record. Link to GitHub repository.

1. Introduction

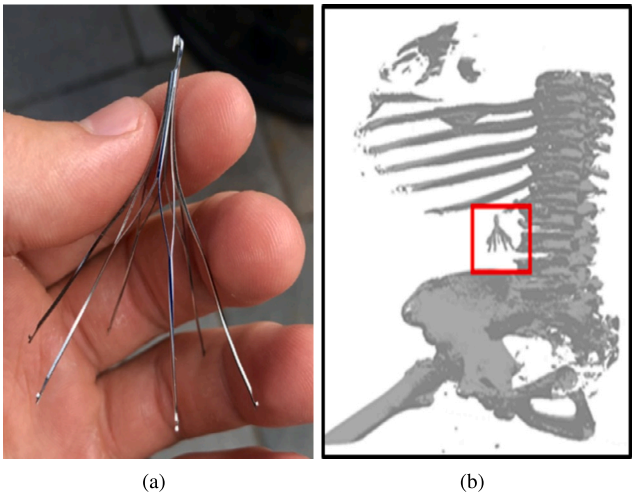
Inferior Vena Cava (IVC) filters are medical devices placed inside the IVC to reduce the morbidity and mortality of Venous Thromboembolism (VTE), specifically Pulmonary Embolism (PE). In the United States, there are about 60,000 to 100,000 deaths per year due to VTE [3]. IVC filtration is an alternative option for managing these conditions when anticoagulation, the first line treatment for VTE, cannot be used – usually due to a high risk of major bleeding [15]. The National Hospital Discharge Survey recorded a total of 803,00 IVC filters placed between 1985 to 2006, and about 259,000 filters were estimated to have been placed in 2012 [3].

Currently, the two types of IVC filters in use within the United States are permanent filters and retrievable filters. Retrievable filters originated in the 1990s and were designed with the option of being retrieved or left in place after the risk of PE subsided [18]. When re-

☆ This document is the results of the research project funded by the Mayo Clinic-UW Eau Claire Research Innovation Council, and National Science Foundation-Research Experience for Undergraduates OAC-2150191. The computational resources of the study were provided by the Blugold Center for High-Performance Computing under National Science Foundation grant CNS-1920220.

* Corresponding author.
** Principal corresponding author.
E-mail addresses: gomesr@uwec.edu (R. Gomes), Wildenberg.Joseph@mayo.edu (J. Wildenberg).

https://doi.org/10.1016/j.ailsci.2023.100084
Received 8 June 2023; Received in revised form 25 July 2023; Accepted 9 August 2023
Available online 18 August 2023


(a)

(b)

Fig. 1. (a) An example of IVC filter used in patients. (b) shows a 3D rendering of CT scan with the position of IVC filter in the patient.

trieved at the appropriate time, retrievable IVC filters provide an added benefit of reducing the occurrence of long-term complications associated with their permanent counterpart [5,26]. While retrievable IVC filters were advertised with the option of having it remain in a patient's body, serious long-term complications can occur. Leaving IVC filters in the body for more than 30 days increases the risk of Deep Vein Thrombosis (DVT), filter migration/embolization, filter fracture, and IVC perforation [10,4,2]. To address these potential complications, the FDA in 2010 recommended that physicians consider the removal of retrievable IVC filters as soon as the risk for PE subsided. Then, in 2014, the FDA issued guidance that, once the patient is no longer at risk for PE, the risks of keeping filters in the patient begin to outweigh the benefits between 29 to 54 days after implantation [8].

Although removal of these IVC filter is as simple as setting up an appointment with the clinician within 30 days of the procedure, most studies reported an average retrieval rate between 20% and 30% [28,23]. There are three factors relating to the underwhelming rate of IVC filter retrieval: (a) patient, (b) system, and (c) technical factors. Patient factors include socioeconomic status and medical comorbidities. Patients could have limited resources, such as healthcare coverage and transportation, which results in poor clinical follow-up. Patients with many comorbidities may have a higher periprocedural risk than the risks of leaving the filter in place. The system factors consist of poor tracking and patient follow-up [8] mostly arising due to patients moving to a different health care system due to job, education, and other factors and not transferring records from their past. The technical factors include the challenges encountered during retrievals, such as filter tilt or fragmentation. An easy way to track these IVC filters at a much later date is to have a lightweight filter detection algorithm that runs whenever a person gets a CT scan for other health reasons thereby uncovering patients who would otherwise be lost to follow-up. Fig. 1 shows an example of IVC filter used during the procedure. With these three issues leading to underwhelming retrieval rates, it is crucial that the process of IVC tracking, and retrieval be integrated in the general CT scan pipeline to enable rapid diagnosis. The proposed research explores the development of the segmentation algorithm which will be an integral part of this pipeline. There is also a need to ensure that this algorithm is very accurate with significantly less false positives thereby saving valuable time for clinicians.

Research has already been conducted on classifying the type of IVC filter found in radiographs using CNNs in the context of medical image analysis. In one study, radiographic pictures that had been manually cropped were utilized to classify 14 different IVC filters. This categorization was carried out using a 50-layer ResNet architecture with a modified final fully connected layer [24]. While this study was able to identify different types of IVC filters, it was not applicable to detecting IVC filters that were already inside the patients. Another study made an effort to improve on this strategy by developing an architecture that can categorize three different IVC filter kinds without requiring the radiographs to be cropped and adjusted to be centered on the filter [25]. An image-processing-only approach created in the past by Dr. Wildenberg can identify a filter from CT images with sensitivity and specificity of roughly 80% each. However, this study used a smaller dataset and mostly relied on a linear approach rendering it impossible to adjust to the heterogenous nature of patient CT scans. Given the low prevalence of IVC filters in the general community and the daily average of 400 suitable CT scans in Mayo Enterprise, this accuracy would put an unreasonably heavy burden on clinician analysis of the inevitable false positive results. A workable answer is offered by an automated deep learning method, and to our knowledge, real-time IVC filter presence/absence detection from segmentation is an uncharted area.

The success of the transformer architecture in the domain of natural language processing (NLP) has led researchers to examine its applicability in the domain of computer vision. The usage of attention in transformer models allows it to learn long-range relationships between image elements that are difficult to attain through convolutions. The Vision Transformer (ViT) is the first architecture that fully replaced convolutions with the original transformer model and garnered significant research attention [11,20]. However, inherent differences between NLP and computer vision presents significant challenges in adapting the transformer architecture. Visual elements are more variable in scale compared to words in a passage and images of higher resolution scales complexity beyond what is expected in NLP. The Swin Transformer utilizes shifting windows of attention and hierarchical layers to overcome these differences and achieve strong performance while maintaining linear computational complexity [21].

This research explores the feasibility and performance of a Swin UNet transformer to create a segmentation backbone for identification and tracking of patients with IVC filters and facilitate timely removal, even when they transition their care into or throughout a health system. In addition to the reduction in complications related to unintentionally long filter dwell times, timely removal will also result in shorter procedural times during retrieval [9,12].

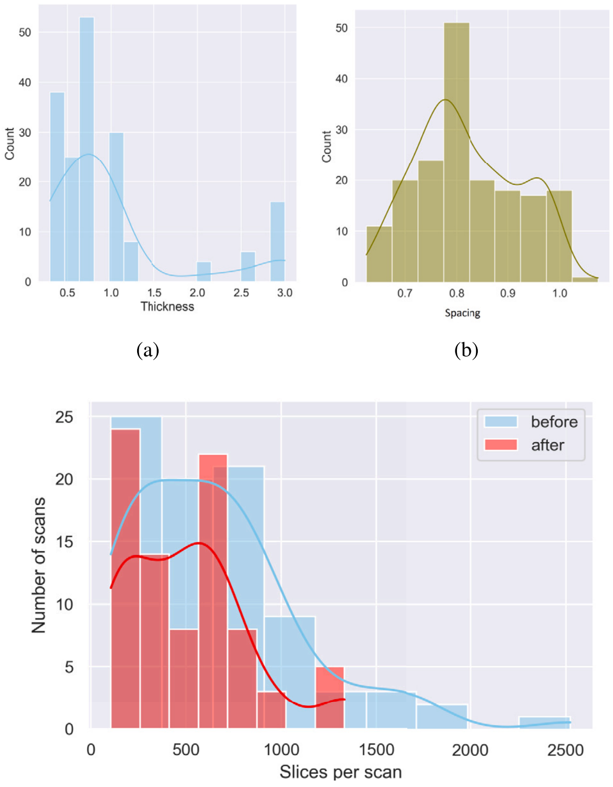
2. Materials and methods

2.1. Image datasets

All data processing was retrospective and IRB-exempt in the need for consent. Using an internal list of known patients with filters, a total of 84 CT scans having 54618 slices along the axial plane were provided by the health system. By eliminating any attached protected health information (PHI), these scans were anonymized. To make the model scanner agnostic, CT scans were purposely chosen from multiple different scanners used throughout the health system. A distribution of Hounsfield Units (HU) for scans is shown in Fig. 2. Fig. 2a shows the slice thickness and Fig. 2b shows the slice spacing of all the CT scans used for analysis. These metrics show the heterogeneity in the CT scans used for training the deep learning models. Fig. 2c shows how many slices are available per CT scan used in this research. The bar graphs in blue resemble a reduced version of CT scans to further remove slices not relevant to our research. This significantly reduces the dataset as it can be observed in the graph compared to the dimensions of the original scans shown in red. A further explanation about data preprocessing is discussed in the next section.

2.2. Data preprocessing

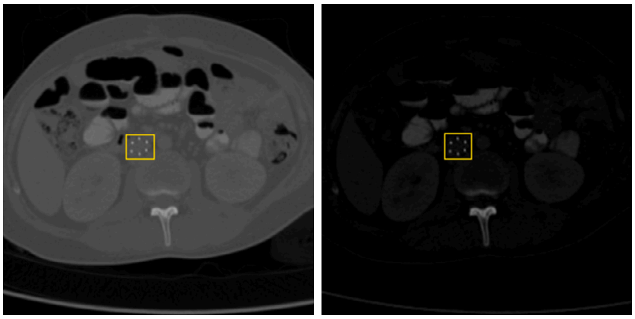
All CT scans were spatially cropped before training to remove as much background thereby giving the deep learning model less data for



(a) (b)

(c)

Fig. 2. Metrics of 84 CT scans used for model training. (a) Shows the slice thickness in mm, (b) Shows the slice pixel spacing in mm, and (c) shows the number of slices per scan before (blue) and after (red) application of 40 cm spatial cropping. It is observed that the number of slices reduces significantly thereby removing redundant data.



(a) (b)

Fig. 3. Slices (a) before and (b) after HU normalization.

better performance. The 20% spatial cropping was done equally to each of the four edges. As a result, the original (512 × 512) CT scan slices were downsized to (307 × 307) before resampling to (256 × 256). Following the spatial cropping, slices of each CT scan 40 cm below the cranial-most slice were discarded. A 40 cm cut-off was chosen because almost all IVC filters are placed within the upper to mid abdomen. The cut-off was able to decrease CT slices by 19.01% as shown in Fig. 2c followed by resampling to have 128 slices per scan. The minimum and maximum HU for the study was set at 1 HU and 2500 HU respectively. These scans were then normalized between 0 and 1 before deep learning was applied. This hard normalization scheme works well for IVC filter detection as shown in [14]. The segmentation masks were created under the supervision of a board-certified radiologist. Fig. 3 shows a slice before and after application of hard normalization scheme.

The training set consisted of 58 CT scans of size (256 × 256 × 128). This was 70% of the total CT scans split randomly. These scans were further split into smaller 3-D patches to be used in the training process. The selected patch size was (64 × 64 × 32). The desired patches were obtained using a function *view_as_blocks* [29] made available through the Scikit-image library. A total of 3712 patches were generated along with corresponding 3712 segmentation masks. The patches were split into IVC and non-IVC based on the presence of IVC filters in the mask. Hence, a total of 3557 patches and masks were generated with no IVC filters while 155 patches and masks were generated with IVC filters for training. The validation set consisted of 26 CT scans (30%) resulting in 1591 patches and masks with no IVC filters along with 73 patches and masks with IVC filters.

These IVC filter voxels constitute an extremely small portion of the entire volume of a CT scan. To remedy the bias towards background data, the ratio of IVC filter patches to normal patches were restricted to 1:1 while training. Thus, the total input size was 155 patches with IVC filters and 155 patches without IVC filters in each sub-epoch. A total of 22 sub-epochs were run each containing a different sequence of non-IVC filter CT scan patches. This procedure was repeated for 25 epochs resulting in 550 total iterations thereby training on the entire dataset. To address overfitting that may arise for using same IVC patches, dynamic augmentation techniques such as random crop, zoom, flipping, and rotation were applied on 3D-patches.

2.3. *Deep learning segmentation*

2.3.1. *UNet*

UNet is a Convolutional Neural Network (CNN) designed for image segmentation [27]. CNNs have been extensively used in healthcare for creating diagnostic and prognostic models [1,32,19,13]. The underlying structure is made up of two paths: a contracting path and an expansion path. The contracting path is an encoder consisting of convolution layers and pooling layers. The expansion path is a decoder consisting of transposed convolution layers and concatenations with the feature maps from the encoder. The 3D-UNet is an extension of the basic UNet structure, allowing for 3D volumetric processing [7]. The 3D-UNet still consists of the contraction and expansion paths, but all its convolution and pooling operations are implemented on a data cube instead of a 2D image. Changing the basic UNet model to fit 3D patches allows it to fully take advantage of spatially dependent features across all three dimensions [30].

2.3.2. *ResUNet*

ResUNets are similar to UNet in terms of the UNet architecture. However, the convolution blocks are replaced with residual blocks also known as ResNets [17]. ResNets use “skip connections” that allow information to bypass one or more layers in the network, and thus propagate more easily through the network. It is therefore able to address the vanishing gradient problem during backpropagation for higher accuracy. A comparison of a convolution block and a ResNet block used in this research is shown in Fig. 5. The concatenation of the input with the output from convolution layers increased prediction performance of ResNets.

2.3.3. *SWIN UNet*

The Swin UNet Transformer (Swin UNETR) is a novel architecture that integrates the Swin Transformer into the traditional UNet convolution architecture. Swin UNETR replaces the convolutional encoder of the original UNet design with the hierarchical design of the Swin Transformer. The outputs of the Swin Transformer are reconnected to the standard UNet decoder through residual block skip connections. Swin UNETR was originally designed for accurate and reproducible segmentation of brain tumors in MRI images [16]. The integration of the Swin Transformer into the UNet architecture was intended to allow the model to learn long-range dependencies for accurate segmentation of tumors that can appear in varying locations, shapes, and sizes. Model

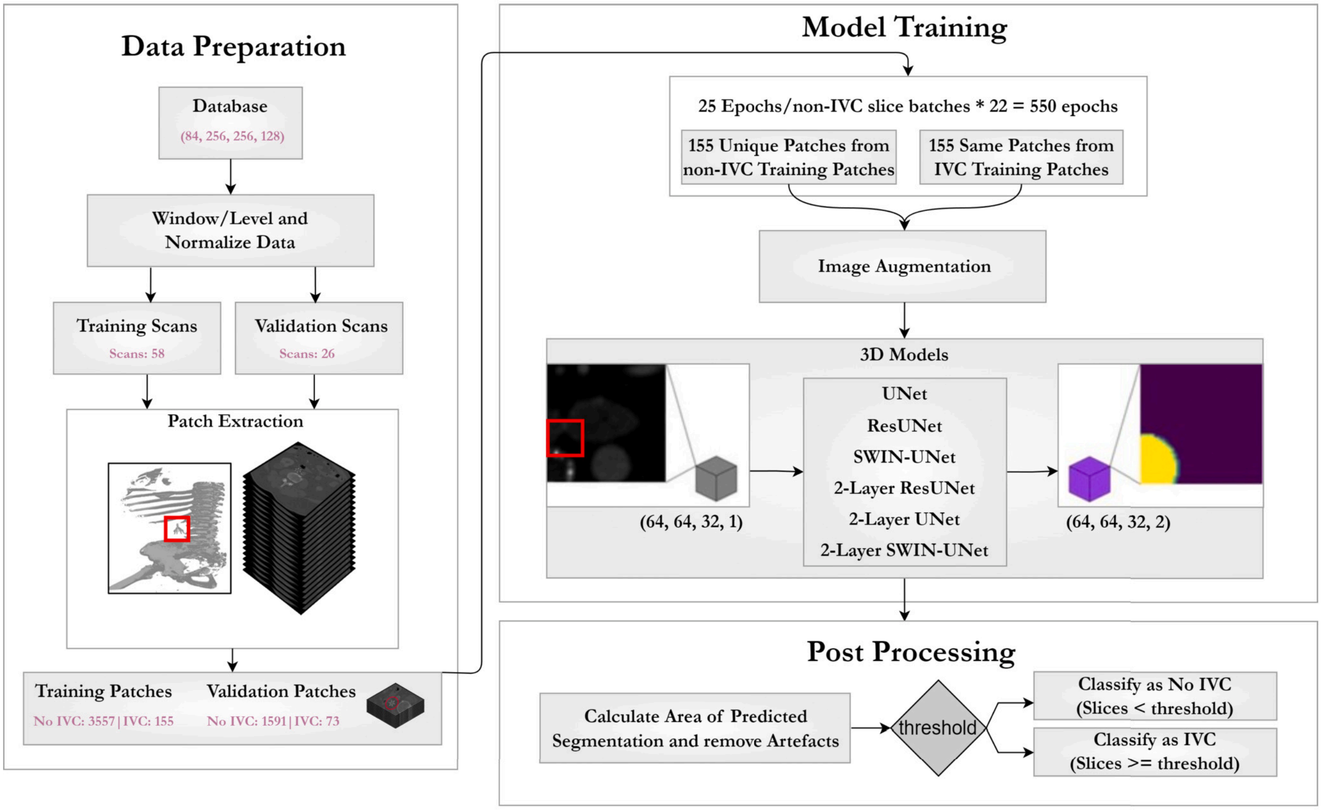


Fig. 4. Overview of the entire training and validation process of the IVC prediction pipeline. After data normalization, slices are split into training and validation. This is followed by 3D-patch extraction. Patches are fed to 3D-UNet variants for IVC filter segmentation. Finally, image processing algorithms remove artefacts to classify a scan having IVC filter.

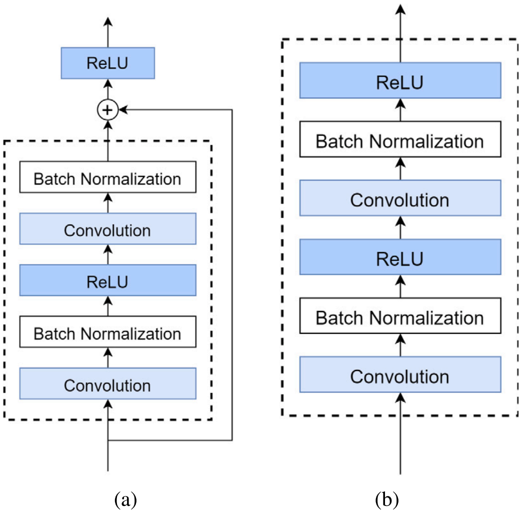


Fig. 5. Overview of a) ResNet blocks and b) UNet blocks used in the training process.

was trained on the dataset provided as part of the 2021 Multi-modal Brain Tumor Segmentation Challenge (BraTS). Compared to the winning methodologies of the 2021 BraTS, Swin UNETR outperformed all other models with at least 0.7%, 0.6%, and 0.5% greater Dice scores on Enhancing Tumor (ET), Whole Tumor (WT), and Tumor Core (TC) semantic classes respectively. The model also achieved a highly competitive performance in the testing phase. Swin UNETR has also been applied to head and neck primary tumors and lymph node segmentation using FGD-PET/CT images. 524 samples provided by the Head and Neck Tumor (HECKTOR) 2022 challenge. Swin UNETR was pre-trained on 5050 CT scans from publicly available datasets and transfer learned to the HECKTOR dataset. The model achieved an average of 0.707 and 0.582 aggregated Dice similarity coefficient on the primary tumors and lymph node gross tumor volume (GTV) respectively. The model performed at a more stable 0.633 (primary tumor) and 0.673 (lymph node) during the evaluation on the test dataset [6]. In a comparative study conducted by Wei et al. [33], Swin UNETR was used among other architectures to evaluate the effectiveness of the proposed high-resolution Swin Transformer network. In the BraTS 2021 dataset, the Swin UNETR achieved the best Hausdorff score amongst the transformer-based models while maintaining a comparable DICE score. Swin UNETR also achieved the best average validation DICE performance of 0.787 on the BTCV multi-organ segmentation task. Maurya et al. [22] applied Swin UNETR to the segmentation of pulmonary arteries using 200 samples from the PARSE 2022 Grand Challenge. All versions of the Swin UNETR trained and examined performed better than all versions of UNet, although by a small margin of around 0.1 DICE score. The authors observed that Swin UNETR was able to better segment arterial branches while UNet was able to better detect the main artery.

2.4. *Training process*

Fig. 4 summarizes this entire process. Six variants of UNet model were implemented. The plain UNet models had two convolutional layers per block. One model utilized four blocks and the other utilized two blocks. A similar two and four-block approach was used by replacing basic UNet blocks with ResNets. Batch normalization was applied to the output of the convolution layers. The output of the decoder blocks is fed through SoftMax activation to create a (64 × 64 × 32 × 2) patch containing prediction probabilities for background and IVC filter. The remaining two UNet models utilized SWIN transformers.

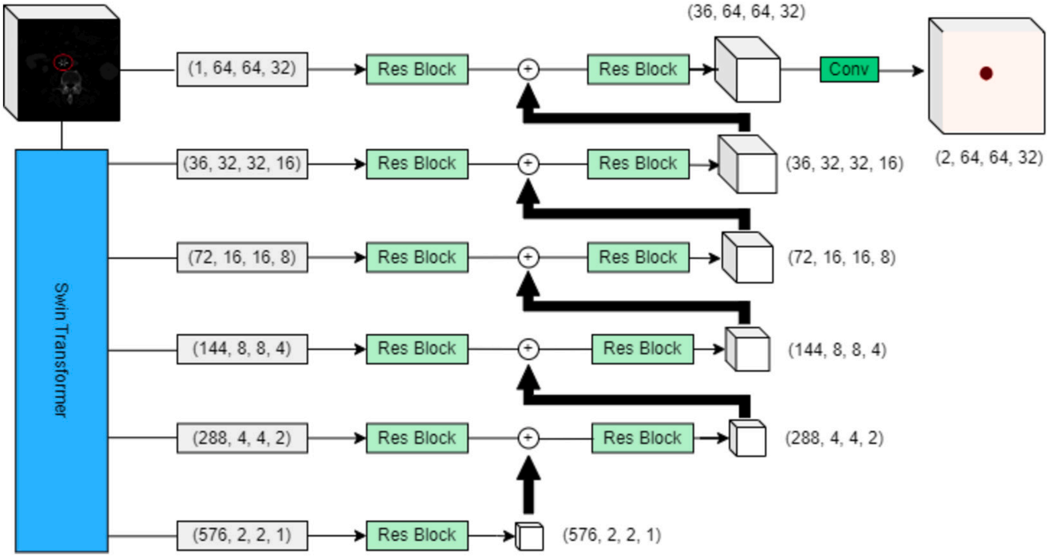


Fig. 6. Overview of SWIN-UNet used in the training process.

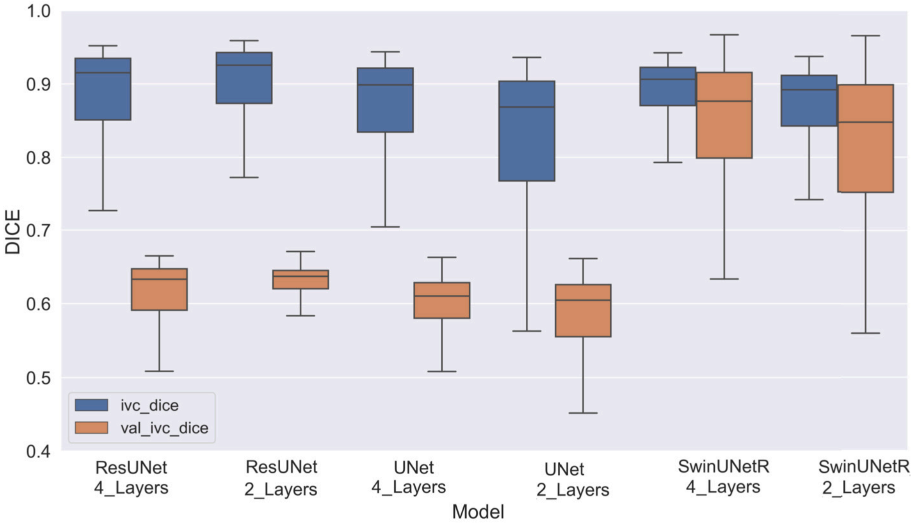


Fig. 7. Box plot showing the Dice Score distributions of the six model variants used in this study.

Table 1
Dice scores for different model variants derived from training and validation datasets.

		UNet Simple		ResNet blocks		Swin UNETR	
		2 Layer	4 Layer	2 Layer	4 Layer	2 Layer	4 Layer
Train	Median	0.87	0.89	0.93	0.92	0.89	0.90
	IQR	0.77-0.90	0.83-0.92	0.87-0.94	0.85-0.93	0.84-0.91	0.87-0.92
Valid	Median	0.6	0.61	0.64	0.63	0.85	0.88
	IQR	0.55-0.63	0.58-0.63	0.62-0.65	0.59-0.65	0.75-0.89	0.79-0.92

The two variants of the Swin UNETR architecture was implemented to allow for comparison with the four UNet variants. One utilizes the four stage design detailed by Hatamizadeh et al. [16]. The model was implemented using the MONAI library, which limits the starting number of filters to a multiple of 12. Therefore, both variants are implemented with a starting feature size of 36 to best match the other UNet variant implementations. Each subsequent stage doubles the number of features, to a max of 576 features in the 4-stage variant and 144 features in the 2-stage variant. The output of the Swin UNETR at each stage is fed into a residual block and then concatenated with the deconvolution of the previous stage in a standard UNet decoder. The output of the model is (64 x 64 x 32 x 2) patches of logits, which are then processed into inferences using the Argmax function. The 4-stage Swin UNETR contained 35,072,552 trainable parameters and completed training in 4 hours 40 minutes. The 2 stage Swin UNETR contained 6,612,392 trainable parameters and completed training in 4 hours 28 minutes. The diagram of the 4-stage Swin UNETR is shown in Fig. 6.

3. Results

Application of different UNet variants showed significant variation in segmentation performance. Dice Score was used as the evaluation

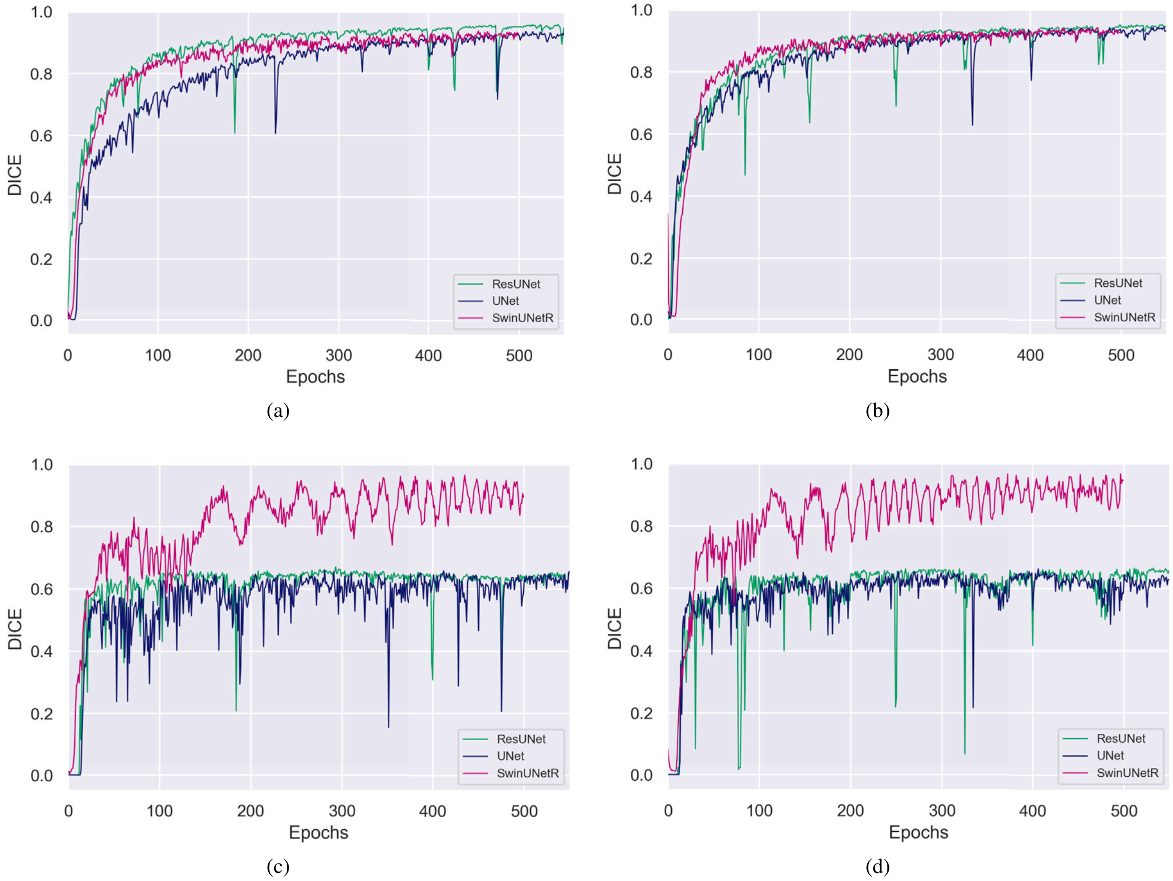


Fig. 8. Overview of IVC Dice for training data for a) Two stage and b) Four stage models and validation data for c) Two stage and d) Four stage models.

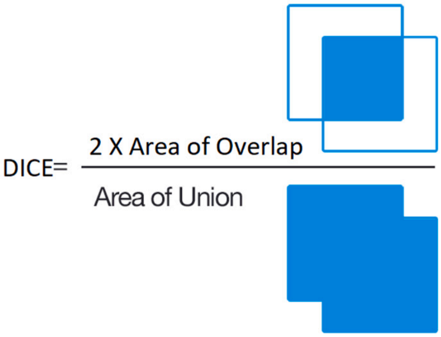


Fig. 9. Figure showing calculation of Dice Score adopted from image under a CC BY-SA 4.0 license.

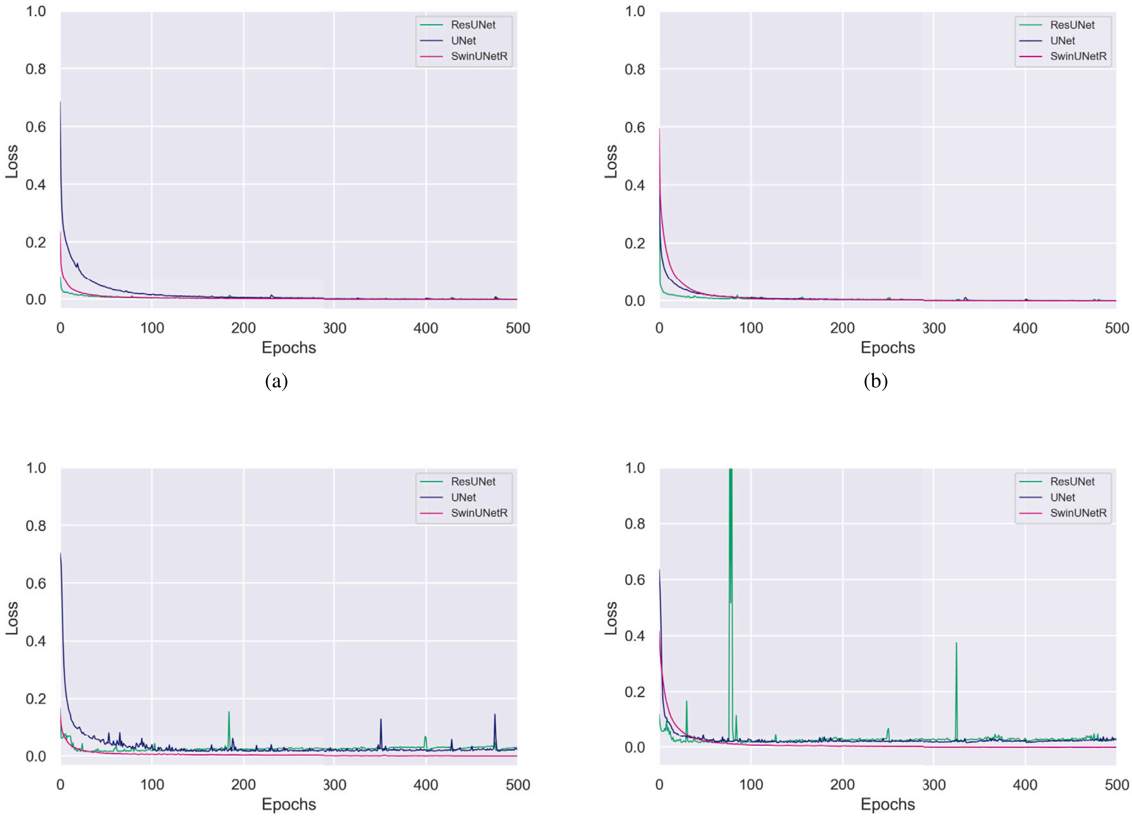
metric which is a common tool in computer vision tasks. In the context of image segmentation, Dice is used to assess the accuracy of a segmentation model's predictions compared to the ground truth segmentation masks as shown in Fig. 9. The Dice score ranges from 0 to 1 with 1 being a complete match between segmentation and ground truth. Considering the highly imbalanced nature of IVC pixels, the Dice provided a simple and intuitive way to investigate the segmentation performance across different models which would otherwise have been challenging if sensitivity and specificity were used due to such a large amount of background class. On the training dataset, the ResNet models performed significantly better compared to plain UNet versions. ResNet with two layers achieved the highest median Dice = 0.93 [IQR: 0.87, 0.94] while the UNet model with two layers achieved the lowest median Dice = 0.87 [IQR: 0.77, 0.90]. Results from applying the model on validation data also revealed a similar trend with ResNet variants achieving a higher accuracy. ResNet with two layers again achieved the highest median Dice = 0.64 [IQR: 0.62, 0.65] while the UNet model with two layers achieved the lowest median Dice = 0.60 [IQR: 0.55, 0.63].

The 4-stage Swin UNETR training recorded a median Dice = 0.90 [IQR: 0.87, 0.92]. The 2-stage Swin UNETR training recorded a median Dice = 0.89 [IQR: 0.84, 0.91]. Compared to results from plain UNet and UNet with ResNet blocks, it is noticeable that Swin UNETR training returns a more consistent and better performance. The effects are also visible with Swin UNETR performance on validation datasets. The 4-stage model returns median Dice = 0.88 [IQR: 0.79, 0.92] and 2-stage model reported median Dice = 0.85 [IQR: 0.75, 0.89]. These metrics are significantly higher than the other CNN-based implementations used earlier. The IVC Dice during training process along with the Dice distributions is shown in Fig. 7 and summarized in Table 1.

The training accuracy for all six model variants for the desired number of epochs were similar as noted in Fig. 8a and 8b. The graphs denote performance of 2-layer compared to 4-layer models. The validation accuracy showed a major improvement while using Swin UNETR across both two and four stage models. The validation accuracy is shown in Fig. 8c and 8d. The loss values of the training and validation process are shown in Fig. 10.

4. Discussion

A 3D-patch based segmentation solution has been proposed to IVC filter detection from CT scans along the axial plane. The use of 3D patches enabled retention of the spatial component of these filters thereby addressing the limitations of a 2D approach [14]. A smaller patch size also reduced the bias introduced by the sparseness of IVC filters in the data even though the scans used for training only constituted about 0.08% of IVC filters. The effect of this imbalance was further reduced during the training process when using a batch size with equal



(a)

(b)

(c)

(d)

Fig. 10. Overview of IVC Loss for training data for a) Two stage and b) Four stage models and validation data for c) Two stage and d) Four stage models.

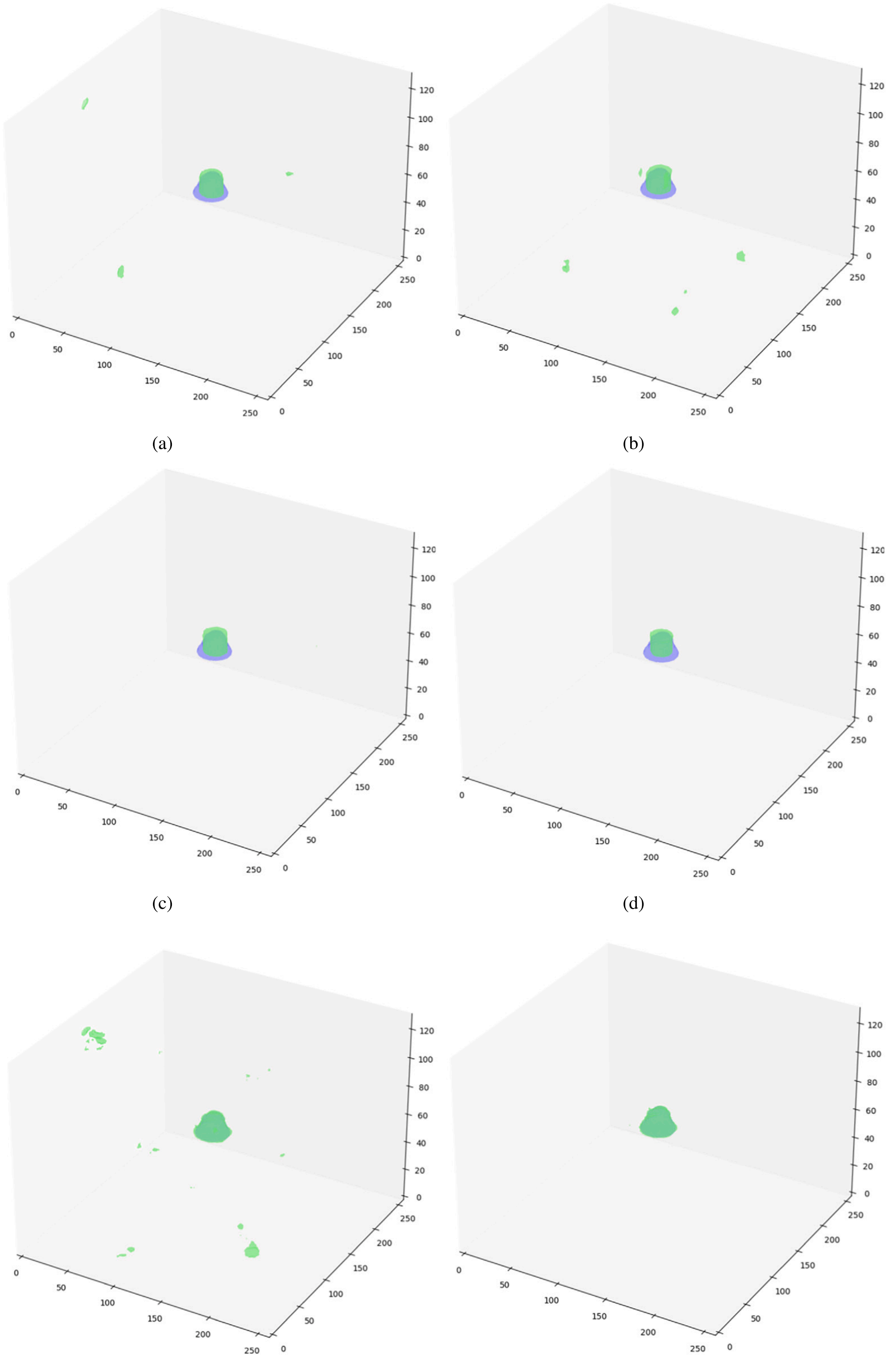
distribution of filter and no filter patches complemented with augmentation during training process. Six different segmentation approaches were compared to further ensure sparsity of filter signature would not impact the outcome. It was observed that using ResNet blocks outperformed basic convolution blocks in segmentation. The ResNet blocks also reduced the rate of false-positives. These false positives were comprised mostly of bones and calcification around the spine. A probable cause for this issue was the lack of skip-connections in basic UNet blocks that is present in ResNet. The skip connections enable a network to learn residual mappings, i.e., the difference between the input and output. Since the IVC filter signature is very low compared to background, the kernels used in deep layers find it challenging to detect features unique to filters. As skip connections bring back the input from shallow layers, the network learns more easily the underlying filter patterns.

Another very significant outcome of this research indicates that deep learning models like Swin UNETR with transformers as their backbone are capable of maintaining consistent performance across both training and validation datasets. This is in contrast to CNN based segmentation which shows a drop in segmentation accuracy for validation data. Fig. 11 further corroborates this fact. Notice how both Swin UNETR models are able to conform the segmentation to the actual shape of the filters whereas the CNN-derived UNet and ResUNet models lose the edges during the segmentation process. Also, the four layer models achieve better segmentation hence less post-processing as their are no spurious segmentation regions around the scans.

A comparative layer analysis resulted in an optimized model with a smaller footprint. While ResNet blocks increase the segmentation performance, they also require more parameters. For example, the two layer ResNet model used in this research utilized around three million parameters (3,155,810) compared to 53 million parameters (53,154,018) in the four layer ResNet models. The Swin UNETR also provided better results with very comparable usage of trainable parameters. While these numbers are higher than traditional UNet variants at 1.3 million (1,357,954) and 22 million (22,587,138) for two and four layer versions respectively, this research established that a transformer-based model with similar parameters as a CNN can produce better training and validation outcomes when integrated with existing medical information technology. The data augmentation strategies including random crop, random resize, random flip, and random rotation ensured that there was significant variation in the training dataset that consisted of 155 IVC and 155 non-IVC patches per sub-epoch. This augmentation strategy ensures that even if the same image is used repeatedly, there are significant variations introduced by randomness thereby reducing any chances of any model overfitting.

5. Conclusion

Future work in the IVC filter detection incorporates algorithm validation and classification pipeline on an estimated 1,500 clinical CT scans predicted to be acquired from the health system. IVC filter prediction pipeline will be developed to provide a report of the presence of an IVC filter and its location. After receiving a predicted 3D mask from the Swin UNETR model, the program will use scikit-image image processing library [31] in Python to analyze the region properties of the segmented image, removing any spurious segmentation that may occasionally arise. The entire algorithm will be tested on clinical CT scans performed in the health system to assess performance, with manual review of each scan. This may be supplemented with additional known-positive test cases, separate from the training set, if the number of true-positive clinical cases is too low to generate meaningful statistics. The results and statistics from this review will then be used to retrain the Swin UNETR, if necessary, with rapid re-deployment and re-evaluation.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 11. Overview of results a) UNet 2 Layer b) UNet 4 layer, c) ResUNet 2 Layer, d) ResUNet 4 layer, e) Swin UNETR 2 Layer, and f) Swin UNETR 4 Layer. Results indicate Swin UNETR 4 Layer is able to conform to the shape of IVC Filters without any false positives.

CRediT authorship contribution statement

Rahul Gomes: Conceptualization, Software, Methodology, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

Tyler Pham: Data curation, Software, Investigation, AI methodology, Writing – original draft.

Nichol He: Data curation, AI methodology, Visualization, Writing – original draft.

Connor Kamrowski: Data curation, Algorithm validation, Software, Visualization.

Joseph Wildenberg: Conceptualization, Software, Validation, Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

https://github.com/rahulgomes19/IVC_3D.

References

[1] Ahsan M, Gomes R, Denton A. Application of a convolutional neural network using transfer learning for tuberculosis detection. In: 2019 IEEE international conference on electro information technology (EIT). IEEE; 2019. p. 427–33.

[2] Angel LF, Tapson V, Galgon RE, Restrepo MI, Kaufman J. Systematic review of the use of retrievable inferior vena cava filters. J Vasc Interv Radiol 2011;22:1522–30.

[3] Ayad MT, Gillespie DL. Long-term complications of inferior vena cava filters. J Vasc Surg Venous Lymphat Disord 2019;7:139–44.

[4] Caplin DM, Nikolic B, Kalva SP, Ganguli S, Saad WE, Zuckerman DA, of Interventional Radiology Standards of Practice Committee S, et al. Quality improvement guidelines for the performance of inferior vena cava filter placement for the prevention of pulmonary embolism. J Vasc Interv Radiol 2011;22:1499–506.

[5] Charles HW, Black M, Kovacs S, Gohari A, Arampulikan J, McCann JW, et al. G2 inferior vena cava filter: retrievability and safety. J Vasc Interv Radiol 2009;20:1046–51.

[6] Chu H, De la Arévalo LR, Tang W, Ma B, Li Y, et al. Swin unetr for tumor and lymph node segmentation using 3d pet/ct imaging: a transfer learning approach. In: Head and neck tumor segmentation and outcome prediction: third challenge, HECKTOR 2022, held in conjunction with MICCAI 2022, proceedings. Springer; 2023. p. 114–20.

[7] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical image computing and computer-assisted intervention–MICCAI 2016: 19th international conference, proceedings, Part II 19. Springer; 2016. p. 424–32.

[8] Crumley KD, Hyatt E, Kalva SP, Shah H. Factors affecting inferior vena cava filter retrieval: a review. Vasc Endovasc Surg 2019;53:224–9.

[9] Desai KR, Laws JL, Salem R, Mouli SK, Errea MF, Karp JK, et al. Defining prolonged dwell time: when are advanced inferior vena cava filter retrieval techniques necessary? An analysis in 762 procedures. Circ Cardiovasc Interv 2017;10:e003957.

[10] Deso SE, Idakoji IA, Kuo WT. Evidence-based evaluation of inferior vena cava filter complications based on filter type. In: Seminars in interventional radiology. Thieme Medical Publishers; 2016. p. 093–100.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv preprint. arXiv:2010.11929, 2020.

[12] Durack JC, Westphalen AC, Kekulawela S, Bhanu SB, Avrin DE, Gordon RL, et al. Perforation of the ivc: rule rather than exception after longer indwelling times for the Günther tulip and celect retrievable filters. Cardiovasc Interv Radiol 2012;35:299–308.

[13] Gomes R, Kamrowski C, Langlois J, Rozario P, Dircks I, Grottodden K, et al. A comprehensive review of machine learning used to combat Covid-19. Diagnostics 2022;12:1853.

[14] Gomes R, Kamrowski C, Mohan PD, Senor C, Langlois J, Wildenberg J. Application of deep learning to ivc filter detection from ct scans. Diagnostics 2022;12:2475.

[15] Hann CL, Streiff MB. The role of vena caval filters in the management of venous thromboembolism. Blood Rev 2005;19:179–202.

[16] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 7th international workshop, BrainLes 2021, held in conjunction with MICCAI 2021, virtual event, revised selected papers, Part I. Springer; 2022. p. 272–84.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.

[18] Heit JA, Cohen AT, Anderson Jr FA. Estimated annual number of incident and recurrent, non-fatal and fatal venous thromboembolism (vte) events in the us. Blood 2005;106:910.

[19] Ismael SAA, Mohammed A, Hefny H. An enhanced deep learning approach for brain cancer mri images classification using residual networks. Artif Intell Med 2020;102:101779.

[20] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. ACM Comput Surv 2022;54:1–41.

[21] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 10012–22.

[22] Maurya A, Patil KD, Padhy R, Ramakrishna K, Krishnamurthi G. Parse challenge 2022: pulmonary arteries segmentation using swin u-net transformer (swin unetr) and u-net. ArXiv preprint. arXiv:2208.09636, 2022.

[23] Mismetti P, Rivron-Guillot K, Quenet S, Décousus H, Laporte S, Epinat M, et al. A prospective long-term study of 220 patients with a retrievable vena cava filter for secondary prevention of venous thromboembolism. Chest 2007;131:223–9.

[24] Ni JC, Shpanskaya K, Han M, Lee EH, Do BH, Kuo WT, et al. Deep learning for automated classification of inferior vena cava filter types on radiographs. J Vasc Interv Radiol 2020;31:66–73.

[25] Park BJ, Sotirchos VS, Adleberg J, Stavropoulos SW, Cook TS, Hunt SJ. Feasibility and visualization of deep learning detection and classification of inferior vena cava filters. 2020. medRxiv, 2020–06.

[26] Ray CE, Mitchell E, Zipser S, Kao EY, Brown CF, Moneta GL. Outcomes with retrievable inferior vena cava filters: a multicenter study. J Vasc Interv Radiol 2006;17:1595–604.

[27] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, proceedings, Part III 18. Springer; 2015. p. 234–41.

[28] Sarosiek S, Crowther M, Sloan JM. Indications, complications, and management of inferior vena cava filters: the experience in 952 patients at an academic hospital with a level I trauma center. JAMA Intern Med 2013;173:513–7.

[29] Scikit-image. skimage 0.21.0 documentation. https://scikit-image.org/docs/stable/api/skimage.util.html#skimage.util.view_as_blocks, 2023. [Accessed 22 July 2023].

[30] Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: a review of theory and applications. IEEE Access 2021;9:82031–57.

[31] Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in python. PeerJ 2014;2:e453.

[32] Wang G, Li W, Ourselin S, Vercauteren T. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 4th international workshop, BrainLes 2018, held in conjunction with MICCAI 201, revised selected papers, Part II 4. Springer; 2019. p. 61–72.

[33] Wei C, Ren S, Guo K, Hu H, Liang J. High-resolution swin transformer for automatic medical image segmentation. Sensors 2023;23:3420.