VDPF: Enhancing DVT Staging Performance Using a Global-Local Feature Fusion Network

Xiaotong Xie[1,3,†], Yufeng Ye[1,†], Tingting Yang[1,3], Bin Huang[2], Bingsheng Huang[2] (✉), and Yi Huang[1] (✉)

[1] The Affiliated Panyu Central Hospital, Guangzhou Medical University
13719231991@126.com
[2] Medical AI Lab, School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, China
huangb@szu.edu.cn
[3] School of Life Science, South China Normal University, Guangzhou, 510631, China

Abstract. Deep Vein Thrombosis (DVT) presents a high incidence rate and serious health risks. Therefore, accurate staging is essential for formulating effective treatment plans and enhancing prognosis. Recent studies have shown the effectiveness of Black-blood Magnetic Resonance Thrombus Imaging (BTI) in differentiating DVT stages without necessitating contrast agents. However, the accuracy of clinical DVT staging is still limited by the experience and subjective assessments of radiologists, underscoring the importance of implementing Computer-aided Diagnosis (CAD) systems for objective and precise DVT staging. Given the small size of thrombi and their high similarity in signal intensity and shape to surrounding tissues, precise staging using CAD technology poses a significant challenge. To address this, we have developed an innovative classification framework that employs a Global-Local Feature Fusion Module (GLFM) for the effective integration of global imaging and lesion-focused local imaging. Within the GLFM, a cross-attention module is designed to capture relevant global features information based on local features. Additionally, the Feature Fusion Focus Network (FFFN) module within the GLFM facilitates the integration of features across various dimensions. The synergy between these modules ensures an effective fusion of local and global features within the GLFM framework. Experimental evidence confirms the superior performance of our proposed GLFM in feature fusion, demonstrating a significant advantage over existing methods in the task of DVT staging. The code is available at https://github.com/xiextong/VDPF.

Keywords: Feature Fusion · Black-blood Magnetic Resonance Thrombus Imaging · Computer-aided Diagnosis · Deep Vein Thrombosis

† Xiaotong Xie and Yufeng Ye contribute equally to this work.
✉Corresponding authors: Bingsheng Huang and Yi Huang.

1 Introduction

Deep Vein Thrombosis (DVT) is a vascular condition characterized by a high risk of recurrence. Its incidence increases with age [6,3,7]. Clinically, DVT is categorized into acute, sub-acute, and chronic stages, with treatment response varying across these stages. Acute thrombosis requires rapid thrombolytic treatment due to its tendency to dislodge and cause pulmonary embolism. However, such therapy is less effective in the chronic stage. Therefore, precise staging of DVT is critical for determining optimal treatment and enhancing prognosis [12]. Digital Subtraction Angiography (DSA) is recognized as the gold standard for diagnosing DVT. But it requires the use of a contrast agent, which may cause renal damage. Recently, the Black-blood Magnetic Resonance Thrombus Imaging (BTI) has been shown to distinguish the stage of DVT without contrast agent, particularly in the chronic stage. In addition, BTI is non-invasive and capable of providing excellent soft tissue contrast. Nevertheless, the precision of clinical DVT staging is still limited by the radiologists' experience and subjectivity.

In practice, radiologists frequently use muscle signal intensity as a crucial reference for BTI-based DVT staging [11]. Thrombi usually display equal or greater signal intensity than adjacent muscle tissue. Nonetheless, delineating muscle tissue manually will significantly increase the workload [13,8].

At present, the implementation of Computer-aided Diagnosis (CAD) systems for DVT staging remains undocumented. However, certain researchers have investigated the potential applications of CAD systems for various issues associated with DVT. For example, a study successfully applies the Deep Semantic Segmentation Feature-Based Radiomics framework to predict the effectiveness of thrombolytic therapy for DVT. Its Area Under the Curve (AUC) value reaches 0.919 [5]. Additionally, an innovative generative adversarial network based on 3D U-net, specifically designed for the automatic segmentation of DVT in BTI, demonstrates excellent results. This method achieves good results on the test set and two other external test sets, with accuracy (ACC) values of 0.96, 0.94, and 0.95. It introduces a novel diagnostic tool for DVT [9].

These studies underscore the potential of CAD system applications in DVT image analysis. However, the application of CAD systems to DVT image analysis faces several challenges. As shown in Fig. 1, the high similarity in signal intensity and shape of DVT to other tissues, along with the small size of thrombi, hinders the network's focus. Furthermore, the prevalent practice of lesion-centered image cropping in small-target classification tasks may result in the loss of crucial information, such as muscle signals and edema, restricting the model's capacity to learn features pertinent to DVT staging. Therefore, it is paramount to retain essential diagnostic information for staging without overburdening physicians.

The contributions of this paper are summarized in three main aspects. (1) To our knowledge, this is the initial application of CAD technology in the field of DVT staging using BTI. (2) We have developed an innovative predictive framework for DVT staging tasks based on Vision Transformer (ViT) and Global-Local Feature Fusion (GLFM), aimed at enhancing the accuracy of DVT staging. This framework optimizes the staging prediction process by integrating information

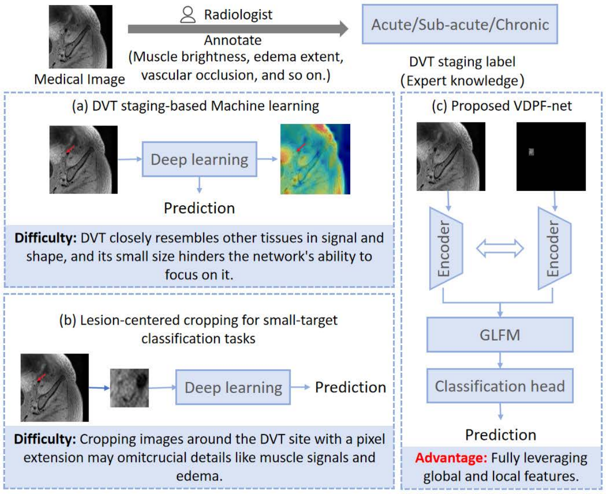
Radiologist
Annotate
(Muscle brightness, edema extent, vascular occlusion, and so on.)
Medical Image
Acute/Sub-acute/Chronic
DVT staging label
(Expert knowledge)
(a) DVT staging-based Machine learning
Deep learning
Prediction
Difficulty: DVT closely resembles other tissues in signal and shape, and its small size hinders the network's ability to focus on it.
(b) Lesion-centered cropping for small-target classification tasks
Deep learning
Prediction
Difficulty: Cropping images around the DVT site with a pixel extension may omitcrucial details like muscle signals and edema.
(c) Proposed VDPF-net
Encoder
Encoder
GLFM
Classification head
Prediction
Advantage: Fully leveraging global and local features.

Fig. 1. Challenges in staging DVT.

from both global imaging and lesion-focused local imaging. (3) Our proposed GLFM, leveraging a cross-attention module and the Feature Fusion Focus Network (FFFN), efficiently integrates global and local imaging features. This fusion strategy effectively combines global and local information. Experimental results demonstrate the superior performance of our framework, offering significant improvements over existing technological approaches in DVT staging tasks.

2 Methods

2.1 Overview of Framework

The proposed VDPF framework is shown in Fig. 2. In practice, radiologists analyze not only localized lesion information such as the degree of vascular occlusion but also global information like muscle signal intensity and edema conditions. Therefore, we first slice the 3D data to obtain global images I_G and local images I_L. The ViT-based dual-branch backbone network separately extracts features from I_G and I_L, obtaining global features F_G and local features F_L. Then, F_L and F_G are fed into the Global-Local Feature Fusion Module (GLFM) to fuse features. The fused features are processed through a fully connected layer to obtain the DVT staging prediction for that slice. Finally, the average of the predictions from each slice is calculated to obtain the overall staging prediction result.

2.2 Global-Local Feature Fusion Module

The GLFM is designed to fully leverage the complementary nature between global and local features, thereby enhancing the model's predictive accuracy

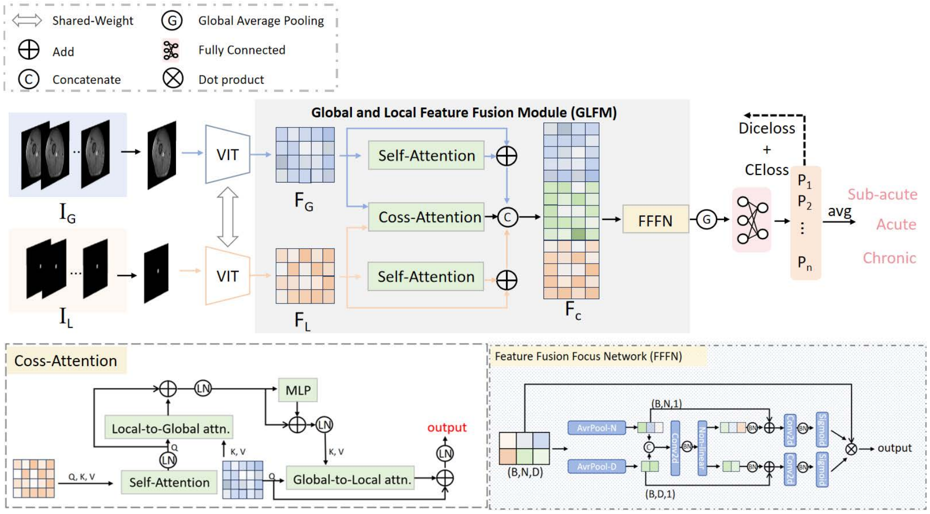

Fig. 2. The VDPF framework consists of the ViT-based dual-branch backbone network, GLFM, and a fully connected layer. The design of the GLFM aims to fully exploit the complementarity between global and local features, consisting of Self-Attention, Cross-Attention, and FFFN modules.

for DVT staging. GLFM initially inputs both F_L and F_G in parallel to self-attention and cross-attention block. The self-attention block enables the model to reinforce the internal relevance of features and highlight the essential ones. At the same time, F_L and F_G are integrated in a cross-attention block. To exchange information between F_L and F_G by calculating the impact of local features on global features. By concatenating the outputs of the self-attention and cross-attention block, we obtain the feature F_C. These features are then input into the Feature Fusion Focus Network (FFFN) and then undergo a series of pooling and convolution processes, enabling refined encoding and integration across spatial and depth dimensions. Features across dimensions are integrated and enhanced to better support accurate DVT staging.

Cross-Attention block consists of self-attention block, local-to-global attention block, multi-layer perceptron (MLP) block, and global-to-local attention block. In self-attention block, F_L is transformed into three vectors Q_L, K_L, V_L through three separate linear layers. Subsequently, Q_L and K_L are multiplied to obtain attention scores. A scaling factor is applied to the attention scores. The softmax function is used for normalization to generate attention weights for each element towards others. Finally, the output attention weights are multiplied by V_L to produce the final output $F_{L\text{-}S}$. In the local-to-global attention block, $F_{L\text{-}S}$ is transformed into $Q_{L\text{-}S}$ through a linear layer, while F_G is converted into K_G and V_G via two separate linear layers. The subsequent operations mirror those in the self-attention module. Finally, we obtain $F_{L\text{-}G}$ by adding the input and output of the local-to-global attention block together.

$$F_{L-G} = \text{Softmax}\left(\frac{Q_{L-S}K_G^T}{\sqrt{d}}\right)V_G + F_{L-S} \tag{1}$$

The MLP, consisting of two fully connected layers and a GELU function, is employed to uncover crucial features and enhance their non-linearity. $F_{L\text{-}G}$ is input into the MLP layer and undergo a residual operation, then yield $F_{L\text{-}G\text{-}M}$.

$$F_{L-G-M} = L_2(GELU(L_1(F_{L-G}))) \tag{2}$$

Here, L_1 and L_2 are two fully connected layers.

Through linear transformations, $F_{L\text{-}G\text{-}M}$ is transformed into $K_{L\text{-}G\text{-}M}$ and $V_{L\text{-}G\text{-}M}$, while F_G is transformed into Q_G. These vectors serve as inputs for the global-to-local attention module, which processes them into the output F_{Cross}.

$$F_{Cross} = \text{Softmax}\left(\frac{Q_G K_{L-G-M}^T}{\sqrt{d}}\right)V_{L-G-M} + F_G \tag{3}$$

In the FFFN block, the feature F_C with dimensions $B \times N \times D$ undergoes global average pooling across N and D to yield two vectors ($B \times N \times 1$ and $B \times D \times 1$). These are then combined, processed through convolutional and nonlinear layers to extract deeper-level features, resulting in two vector types, W_N and W_D. Both undergo normalization and convolution with a residual operation. Then they are transformed into weights using a Sigmoid function. Finally, F_C is multiplied by these weights to produce the output.

3 Experimental Results

3.1 Materials

We collect a DVT dataset from 196 patients at the Affiliated Panyu Central Hospital of Guangzhou Medical University, obtaining pelvis, thigh, and calf sections via BTI sequence. Considering the lower MRI sensitivity of the calf, it's excluded. The dataset is segmented into 223 training, 75 validation, and 75 test cases, resulting in 68,361 training slices, 18,901 validation slices, and 20,663 test slices. This segmentation allows for precise DVT evaluation and treatment planning. Image examples are in Fig. 3. I_L is obtained by retaining only the regions within the bounding boxes manually annotated by radiologists around thrombus areas. Dataset details are in Table 1. We use the consensus of two experienced radiologists as the ground truth for staging, aligning with clinical standards.

Implementation details are as follows: The data augmentation strategies encompass spatial transformations, color adjustments, noise addition, and resampling transformations. The batch size for training is set to 24. An early stopping strategy is employed to dynamically adjust the epoch value during training, with backpropagation occurring every five iterations. The Adam optimizer is

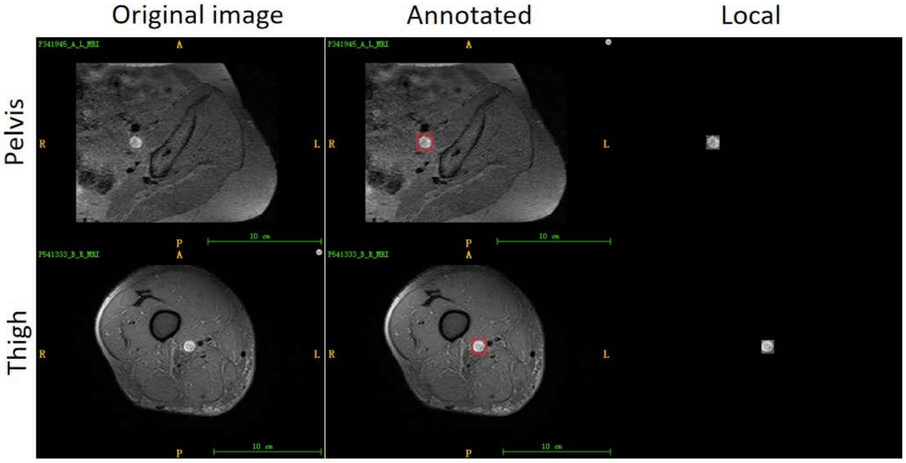

Fig. 3. Representative images of the pelvic and thigh. The original image represents the global image. The annotated image emphasizes the DVT area. The local image solely preserves the rectangular region of DVT.

Table 1. Classification Statistics for dataset. Number of cases in Parentheses.

	Acute	Sub-acute	Chronic	Total
Training	14641(43)	44231(139)	9489(41)	68361(223)
Validation	4335(15)	11895(47)	2671(13)	18901(75)
Test	4036(14)	13462(47)	3165(14)	20663(75)
Total	23012(72)	69588(223)	15325(68)	107925(373)

used to train VDPF, with a learning rate of 0.0001, a weight decay coefficient of 0.000001, and a momentum value of 0.99. Furthermore, we adopt a learning rate adjustment strategy to achieve more refined training control and better convergence outcomes. A warm-up strategy is implemented during the first 10 training epochs, followed by gradual adjustment of the learning rate using a cosine annealing strategy. We train comparison models using 2D images cropped around lesion centers, involving: 1) Slicing 3D volumes for 2D images; 2) Extending bounding boxes by 5 pixels; 3) Cropping and resizing to 224x224. Other training parameters match those of the VDPF. All experiments are conducted in PyTorch using an NVIDIA Corporation Device 2230. ACC and the F1 score are used to evaluate the performance of the VDPF multi-classification model.

3.2 Ablation Study

In this study, we examine the effects of Resnet and ViT backbones on VDPF model performance. Table 2 shows that ViT significantly boosts accuracy on both validation and test datasets, with validation set accuracy rising from 72.0% to 81.3% and test set accuracy from 69.3% to 80.0%. This improvement is likely due to ViT's ability to capture global contextual information through its attention mechanism, contrasting with Resnet's focus on local features through convolutions. ViT's global approach is especially advantageous for medical imaging, enhancing diagnostic accuracy by analyzing complex data patterns and regional relationships.

Table 2. Impact of different Blocks on Model Performance.

Methods	Validation		Test	
	ACC (%)	F1-Score (%)	ACC (%)	F1-Score (%)
Resnet-DPF	72.0	75.4	69.3	65.9
ViT-DPF(VDPF)	**81.3**	**80.8**	**80.0**	**80.0**
VDPF-CA	78.7	79.0	70.7	71.2
VDPF-CAT	52.0	53.1	45.3	46.5
VDPF-SC	80.0	79.1	74.7	74.5

Additionally, we investigate the impact of changing the input sequence of local and global features within the cross-attention module [4], which we refer to as VDPF-SC. The results on the test set show a 5.3% decrease in model ACC. This decrease could be due to the rich information content in the global images, which correlates with the local images. In contrast, when information from the local images relevant to the global context is extracted, the noise in the global images may cause interference, thus complicating the feature extraction process. In addition, we conduct two experiments to assess the FFFN's efficacy. Initially, we replace the FFFN module with the Coordinate Attention module, referenced in literature as VDPF-CA. Then, we explore inputting the combined feature F_c directly into a fully connected layer for classification, termed VDPF-CAT. The VDPF-CA model shows a 9.3% decrease in accuracy on the test set, which we attribute to the original setup's normalization across two dimensions, likely diminishing the impact of smaller features during processing. Similarly, the VDPF-CAT model records an accuracy of only 45.3%, affected by variations in feature data distribution. For example, features from local images with background noise removed may be smaller, whereas global features often have larger values, leading to a diminished contribution of local features in the final linear layer. These experiments highlight the superiority of our proposed module.

As is shown in Table 3. To delve deeper into the impact of distinct feature inputs on the effectiveness of the model, we experiment with several combinations of features. These include global features refined through a self-attention module(denoted as F_G), local features similarly processed (labeled as $F_{L\text{-}S}$), and F_{Cross}, which is honed using a cross-attention module. The table meticulously documents the outcomes of these experiments. An examination of each feature input indicates that $F_{L\text{-}S}$ often has a significant impact, aligning with our expectations, as the features derived from lesion-focused imaging tend to carry less redundant information and noise, thereby simplifying the model's interpretation process. Comparative analysis of the data from rows two, four, six, and seven of the table shows that adding $F_{G\text{-}S}$ to $F_{L\text{-}S}$ generally reduces model performance, while F_{Cross} minimally affects it. However, the concurrent input of all three types of features yields the best outcome, implying that $F_{G\text{-}S}$ and F_{Cross} may have a synergistic effect during feature fusion, possibly by mitigating noise from the global imaging background. This hypothesis gains further support when considering the comparative data in rows one, three, and five, where the combined input of $F_{G\text{-}S}$ and F_{Cross} enhances model performance beyond what is achievable with individual inputs. It is on this basis that the simultaneous incorporation of $F_{G\text{-}S}$, $F_{L\text{-}S}$, and F_{Cross} into the model achieves the most optimal performance.

Table 3. Impact of Feature Input Combinations on Model Performance.

Methods	Validation		Test	
	ACC (%)	F1-Score (%)	ACC (%)	F1-Score (%)
$\mathrm{F}_{\mathrm{G\text{-}S}}$	70.7	68.4	62.7	61.7
$\mathrm{F}_{\mathrm{L\text{-}S}}$	**81.3**	**81.9**	77.3	77.8
$\mathrm{F}_{\mathrm{Cross}}$	73.3	71.6	69.3	69.0
$\mathrm{F}_{\mathrm{G\text{-}S}}+\mathrm{F}_{\mathrm{L\text{-}S}}$	80.0	80.1	74.7	74.9
$\mathrm{F}_{\mathrm{G\text{-}S}}+\mathrm{F}_{\mathrm{Cross}}$	77.3	77.3	74.7	74.6
$\mathrm{F}_{\mathrm{L\text{-}S}}+\mathrm{F}_{\mathrm{Cross}}$	80.0	79.3	78.7	78.6
$\mathrm{F}_{\mathrm{G\text{-}S}}+\mathrm{F}_{\mathrm{L\text{-}S}}+\mathrm{F}_{\mathrm{Cross}}$	**81.3**	80.8	**80.0**	**80.0**

3.3 Comparison with Other Methods

We compare VDPF with several other classic classification methods, as shown in Table 4. VDPF outperforms the other methods on both the training and validation sets. It may effectively utilize the complementarity between global and local information, allowing the network to better focus on localized lesion information and related global information (such as muscle signal intensity).

Table 4. Results of different methods.

Methods	Validation		Test	
	ACC (%)	F1-Score (%)	ACC (%)	F1-Score (%)
Resnet50 [2]	78.7	79.7	69.3	69.8
ViT [1]	78.7	78.3	73.3	73.2
Efficientnet-b0 [10]	76.7	75.8	78.7	77.9
VDPF	**81.3**	**80.8**	**80.0**	**80.0**

4 Conclusions

We pioneer the exploration of applying CAD for BTI-based DVT staging. Facing the challenges of employing CAD in this domain, we specifically developed an innovative classification framework that integrates global imaging with lesion-focused local imaging for precise DVT staging. This framework utilizes a dual-branch structure with shared weights for image features extraction and employs a GLFM for efficient features integration. Crucially, the cross-attention block within GLFM allows for the capture of relevant global features information based on local features, while the FFFN further enhances features integration by amalgamating different dimensional features. Experimental validation demonstrates the significant superiority of our proposed method over existing technologies.

Acknowledgments. This work was supported by the Natural Science Foundation of China (No. 62371303), the Science and Technology Program of Guangzhou (202201011638), the Science and Technology Program of Panyu (2022-Z04-006), and the General Guidance Project of the Panyu District Bureau of Science, Industry, Commerce and Information Technology (2020-Z04-027, 2020-Z04-028).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
3. Heit, J.A., Mohr, D.N., Silverstein, M.D., Petterson, T.M., O'Fallon, W.M., Melton, L.J.: Predictors of recurrence after deep vein thrombosis and pulmonary embolism: a population-based cohort study. Archives of internal medicine **160**(6), 761–768 (2000)
4. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713–13722 (2021)
5. Huang, B., Tian, J., Zhang, H., Luo, Z., Qin, J., Huang, C., He, X., Luo, Y., Zhou, Y., Dan, G., et al.: Deep semantic segmentation feature-based radiomics for the classification tasks in medical image analysis. IEEE Journal of Biomedical and Health Informatics **25**(7), 2655–2664 (2020)
6. Labropoulos, N., Jen, J., Jen, H.: Recurrent deep vein thrombosis: Long-term incidence and natural history. Journal of Vascular Surgery **52**(5), 1420–1421 (2010)
7. Schulman, S., Lindmarker, P., Holmstrm, M., Lrfars, G., Carlsson, A., Nicol, P., Svensson, E., Ljungberg, B., Viering, S., Nordlander, S.a.: Post-thrombotic syndrome, recurrence, and death 10 years after the first episode of venous thromboembolism treated with warfarin for 6 weeks or 6 months. Journal of thrombosis and haemostasis : JTH **4**(4), 734–42 (2006)
8. Spritzer, C.E., Trotter, P., Sostman, H.D.: Deep venous thrombosis: gradient-recalled-echo mr imaging changes over time–experience in 10 patients. Radiology **208**(3), 631 (1998)

9. Sun, C., Xiong, X., Zhang, T., Guan, X., Mao, H., Yang, J., Zhang, X., Sun, Y., Chen, H., Xie, G.: Deep learning for accurate segmentation of venous thrombus from black-blood magnetic resonance images: A multicenter study. BioMed research international **2021**, 4989297 (2021)
10. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
11. Wu, G., Liu, L., Wang, T., Pan, C.: T1 mapping is useful for staging deep venous thrombosis in the lower extremities. Acta Radiologica **63**(4), 489–496 (2022)
12. Xiaona, Liu, Na, Li, Chaoyang, Wen: Effect of pathological heterogeneity on shear wave elasticity imaging in the staging of deep venous thrombosis. PLOS ONE **12**(6), e0179103 (2017)
13. Xie, G., Chen, H., He, X., Liang, J., Deng, W., He, Z., Ye, Y., Yang, Q., Bi, X., Liu, X.a.: Black-blood thrombus imaging (bti): a contrast-free cardiovascular magnetic resonance approach for the diagnosis of non-acute deep vein thrombosis. Journal of Cardiovascular Magnetic Resonance **19**(1) (2017)