AutoDVT: Joint Real-Time Classification for Vein Compressibility Analysis in Deep Vein Thrombosis Ultrasound Diagnostics

Ryutaro Tanno[1,2], Antonios Makropoulos[1], Salim Arslan[1,3], Ozan Oktay[1,3], Sven Mischkewitz[1], Fouad Al-Noor[1], Jonas Oppenheimer[1], Ramin Mandegaran[1,4], Bernhard Kainz[1,3(✉)], and Mattias P. Heinrich[1,5]

[1] ThinkSono Ltd., London, UK
{ryutaro,antonios,salim,ozan,sven,fouad,jonas,ramin,bernhard,mattias}@thinksono.com
[2] Department of Computer Science, University College London, London, UK
[3] Department of Computing, Imperial College London, London, UK
[4] Department of Radiology, Guy's and St Thomas' NHS Foundation Trust, London, UK
[5] Institute of Medical Informatics, University of Lübeck, Lübeck, Germany

Abstract. We propose a dual-task convolutional neural network (CNN) to fully automate the real-time diagnosis of deep vein thrombosis (DVT). DVT can be reliably diagnosed through evaluation of vascular compressibility at anatomically defined landmarks in streams of ultrasound (US) images. The combined real-time evaluation of these tasks has never been achieved before. As proof-of-concept, we evaluate our approach on two selected landmarks of the femoral vein, which can be identified with high accuracy by our approach. Our CNN is able to identify if a vein fully compresses with a F1 score of more than 90% while applying manual pressure with the ultrasound probe. Fully compressible veins robustly rule out DVT and such patients do not need to be referred to further specialist examination. We have evaluated our method on 1150 5–10 s compression image sequences from 115 healthy volunteers, which results in a data set size of approximately 200k labelled images. Our method yields a theoretical inference frame rate of more than 500 fps and we thoroughly evaluate the performance of 15 possible configurations.

1 Introduction

Deep vein thrombosis (DVT) is caused by the formation of a blood clot within a deep vein, that most commonly takes place in the leg. If left untreated, DVT may lead to serious complications, including pulmonary embolism, which develops when pieces of blood clot break loose into the bloodstream and block vessels in the lungs. Typically, an average of one in a thousand people will be affected by DVT and related conditions during their lifetime. 20% of patients die because of DVT-related complications. Worldwide, approximately 10 million people suffer from DVT or related conditions, estimates suggest that 100,000 Americans alone

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 905–912, 2018.
https://doi.org/10.1007/978-3-030-00934-2_100

die of DVT each year [1]. Patients are referred for DVT-specific tests by front line medical professionals, or following surgery. As the risk of DVT leading to serious complications, including death, is high, development of DVT-focused point-of-care diagnostics is of great importance.

There are two major challenges in diagnosing DVT. First, DVT does not necessarily show evident symptoms and the symptoms may overlap with other less serious conditions, making it impossible to discover DVT without clinical tests. Second, the clinical routine method used to diagnose DVT is a D-dimer blood test [2], which determines the concentration of a small protein fragment in the blood that occurs when a blood clot is degraded by fibrinolysis. The D-dimer blood test can with high certainty rule out a pulmonary embolism (PE), the main DVT-related complication leading to death. However, while this test shows a high sensitivity to PE it has a low specificity to early DVT and returns a high number of false positives. Patients with false-positive D-dimer results are referred to unnecessary further expert examinations, which is costly and time consuming. Furthermore, the consequently introduced workload is very large relative to the number of specialist DVT radiologists.

A more accurate screening method is the manual evaluation of vein compressibility during ultrasound (US) examination of standardised anatomical locations (usually at three specified landmarks on the femoral and popliteal veins [3], C-US method). However, for front line medical professionals, it is currently difficult to assess DVT using US since training is required to navigate to anatomically defined landmarks on the veins where compressibility has to be evaluated.

Contribution: To solve this problem we propose an automatic, point-of-care ultrasound (POCUS) image-based method to make DVT diagnostics accessible for front-line non-specialists. We propose a dual-task convolutional neural network (CNN) that jointly classifies the anatomical landmark plane in the current field-of-view and scores vein compressibility. Thus, the proposed AutoDVT network can intrinsically learn to interpret video data to perform localisation, segmentation, local deformation estimation and classification from weak global labels. Furthermore, it is designed to require few floating point operations to enable real-time performance and its accuracy is thoroughly evaluated on over 100 real, manually annotated ultrasound sequences from DVT examinations.

Related Work: Automated methods for DVT screening using ultrasound have been subject to some research. They can be roughly categorised in image segmentation and tracking approaches. Early approaches used specialised ultrasound probe extensions to provide external tracking and pressure measurements [4]. Vessel segmentation has been achieved using semi-automatic initialisation (seed-points) and heuristic intensity-based algorithms [5]. These approaches need manual intervention, suffer from lack of robustness and real-time capabilities. More recently Doppler flow measurements have been added to heuristically define compressibility parameters derived from vessel segmentation masks [6]. While providing a high sensitivity of over 90% they generally suffer from low specificity around 50% and are not fully automatic.

Machine learning, especially deep learning, has recently shown to be highly useful for ultrasound image analysis [7,8]. End-to-end training from clinical data, high accuracy and real-time performance during model evaluation are essential for tasks like POCUS DVT diagnostics. While vessel localisation has been shown to be achievable through deep learning [9] we propose the first, fully automatic deep learning vessel compressibility evaluation that enables a semantic understanding of anatomy based on only weakly-labelled ultrasound images.

2 Method

AutoDVT aims to automate the compression-based examination of DVT [10], in which predefined landmarks on the femoral and popliteal veins are examined with regards to their compressibility by manually applying pressure with the US probe. DVT is suspected if any of the landmark veins is not fully compressible, indicating the potential presence of a clot in the vein. We propose AutoDVT, an end-to-end multi-task deep learning approach, which processes a stream of freehand US images in real-time, and simultaneously determines the type of relevant vessel landmark in the current frame and inspects its compressibility during the exam. The synopsis of the proposed method is shown in Fig. 1.

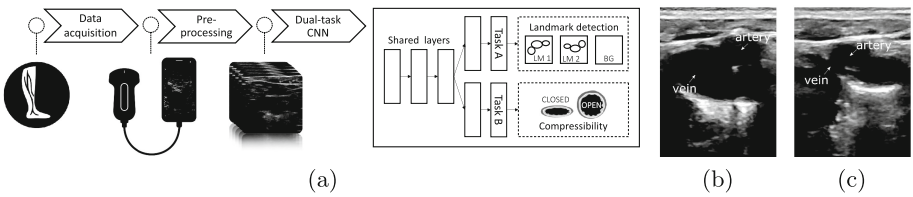


Fig. 1. Overview of the proposed approach (a): A dual-task CNN evaluates a stream of free-hand US frames in real-time and determines landmark type and compressibility. Example for an uncompressed (b) and compressed (c) vein as seen at a landmark position during diagnosis. Note that arteries do not compress during examination as seen in (c).

Preprocessing: Our network is trained on ultrasound sequences acquired with probes from different vendors. This makes inference more robust and more widely applicable, but requires preprocessing. We automatically remove text information from the frames, pad or crop to a common size of 600×600 pixels, and downsample by a factor of 4 in each dimension to the size of 150×150 pixels.

The **AutoDVT model** is a deep CNN that jointly solves two classification problems: (1) *landmark (LM) detection i.e.* a 3-way classification that discriminates, whether a given frame shows either one of the two major landmarks located on the femoral veins (FOVs), called LM1 and LM2 or any other anatomy, which we refer to as BG (background); (2) *open or closed (O/C) classification*

i.e. a binary classification if the present vein is open or closed. Based on the predictions of the two classification tasks from an ultrasound video, AutoDVT evaluates the vein compressibility of the two key landmarks in FOVs.

There are two important architectural choices that enable AutoDVT to accurately perform these tasks. Firstly, the network operates on a stack of consecutive frames and takes temporal information in a sequence into account through 3D convolutions. The network is thereby able to learn motion and deformation features, which enhances the temporal consistency of the predicted landmarks and open/closed state labels, particularly in the presence of noise and artifacts, ubiquitous issues with handheld US probes. Secondly, we employ a multi-task learning approach that shares the majority of convolutional layers across landmark localisation and vein compressibility. It branches out into task specific layers only for the last layers (see Fig. 1). This enables us to leverage sequences with partially missing labels (i.e. only for one task) and increases the amount of overall training data. Our joint training therefore improves generalisation via inductive transfer, where cues from one task regularise and improve the representation for another related task [11].

The details of the shared layers and task-specific layers are given in Table 1. All convolution layers (both 3D and 2D) are followed by rectified linear (ReLU) non-linearity. We apply batch-normalisation after each convolution for improved convergence and accuracy. In the task specific branches, dropout is used for every convolution layer with a rate of 0.5 for regularisation. For each task, the feature maps of the last convolution layers are spatially averaged, and fed into a linear classifier with softmax output. We use relatively few feature map channels and an aggressive progression of strided convolutions to limit both network capacity and floating point operations. This enables us to avoid overfitting and realise realtime performance for inference.

Table 1. Million floating point operations per second (MFlops, fused multiply-adds) performance for each layer in our network model and Global average pooling (GAP).

Layer	Shared layers					Task-specific layers			
	Conv3d	Conv3d	Conv3d	Conv3d	Conv2d	Conv2d	Conv2d	GAP	Linear
Channels	32	32	64	64	128	128	128	128	#classes
Kernel	$3 \times 3 \times 3$	$3 \times 3 \times 3$	$3 \times 3 \times 3$	$3 \times 3 \times 3$	3×3	3×3	3×3	2×2	1×1
Stride	(2,2,1)	(2,2,1)	(2,2,1)	(2,2,1)	(1,1)	(1,1)	(1,1)	N/A	N/A
Size	$150 \times 150 \times 9$	$74 \times 74 \times 7$	$36 \times 36 \times 5$	$17 \times 17 \times 3$	8×8	6×6	4×4	2×2	1
MFlops	23.7	107.5	5.3	2.4	2.7	2.4	0.6	0.1	0.0

Model Training. Unless otherwise stated, we employ a common protocol for the training of networks and determine best performing parameters experimentally. The loss is defined as the sum of the cross-entropy from O/C and LM classification, plus the L2 weight norm (weight decay) with a factor of 10^{-5}. We optimise the parameters by minimising the loss using ADAM for 50 epochs with

an initial learning rate of 10^{-3}, exponential decay every epoch, and a momentum parameter $\beta = [0.9, 0.999]$. ADAM was chosen because it is known for its stable and efficient optimisation performance. Values of β are chosen close to 1 to provide robustness to sparse gradients. The model from the last epoch is used for evaluation. We adopt a data-augmentation scheme where training images are randomly scaled, rotated, horizontally flipped and intensity-augmented.

3 Data and Experiments

Data Collection and Annotation: We have trained our model on manually labelled data from 115 healthy volunteers. Images have been acquired using a Clarius L7 Handheld Wireless Ultrasound Scanner, a Phillips iU22, a GE Logiq E9, and a Toshiba Aplio 500. The dataset is comprised of 1150 videos of length 5–10 s. Each sequence contains between 100 to 200 frames. In this paper we focus on landmark labels from a subset of 240 annotated sequences from the groin area. In each sequence, images have been labelled by 25 skilled annotators (medical students) and one radiologist as one of four landmark labels. We sample background from random frames without labels in these 240 videos and from additional 340 sequences that have been acquired from areas surrounding the femoral vein. We use two of the four available locations, thus, landmarks are located at the saphenofemoral junction (LM1) and great saphenous vein (LM2). Open/close binary labels are manually obtained for every frame by labelling and counting the number of vein pixels, i.e. measuring the area and a defined threshold. The O/C labels have been reviewed by an experienced radiologist. 60% of the volunteer examinations have been used for training, 20% for validation and 20% for testing. We split the data on subject level and not per sequence to avoid unfair testing.

Classification Performance Experiments: We evaluate the predictive performance of AutoDVT in O/C and LM image classification (O/C, LM1, LM2 vs. background BG). We measure the performance on standard metrics used for classification tasks: precision, recall and F1 score, and perform an ablation study to quantify the effects of the main three proposed features of our approach: data-augmentation, modelling temporal information with 3D convolutions and dual-task formulation. Table 2 summarises the results of this analysis.

To demonstrate the benefits of the dual-task architecture of AutoDVT, we constructed two task-specific baseline networks for O/C classification and LM classification by only retaining the shared layers and the individual branch for the chosen task. The two baseline networks have the same target classes for respective tasks as the AutoDVT model, enabling a direct evaluation of the regularisation effect gained from shared representation.

We also assessed the effect of modelling temporal information on classification performance. We implemented variants of the dual-task AutoDVT architecture and the task-specific networks in which every 3D convolution layer is replaced by a 2D convolution with appropriately increased number of kernels (keeping overall numbers of parameters the same).

4 Results and Discussion

Table 2 shows that the best average F1 scores (the harmonic mean of precision and recall) of 91% and 78% are achieved by the dual-task model with +Aug. +Temp. on O/C and LM classification, respectively. It is evident that our dual-task approach outperforms the single-task baselines (+Aug. +Temp.), in particular for the clinically most relevant compressibility analysis (O/C).

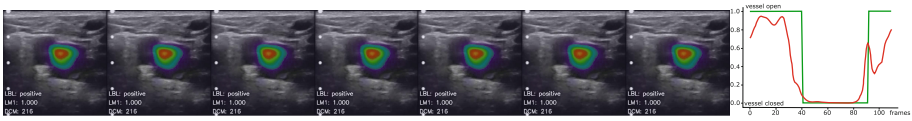
It is also clear that analysing a stack of nine consecutive frames with 3D convolutions (denoted by +Temp.) improves the F1 score compared to a static 2D image analysis, which also results in temporally more consistent classification in the presence of artefacts and noise. This indicates that the temporal network alleviates these challenges by augmenting the missing information using temporal context from the previous frames and estimates useful deformation features.

Table 2. Average performance of different models with best results in bold. +Temp. denotes the integration of spatial-temporal convolutions on nine frames (*i.e.* 3D convolutions) and +Aug. denotes the use of data augmentation during training. For single-task baselines, the results for two task-specific networks (one for O/C classification and the other for LM classification) are shown in each row of the table. AutoDVT corresponds to the dual-task model +Aug.+Temp.

	Precision				Recall				F1 score			
	O/C	LM1	LM2	BG	O/C	LM1	LM2	BG	O/C	LM1	LM2	BG
Separate models	0.78	**0.70**	0.54	0.94	0.92	0.60	0.35	**0.97**	0.84	0.64	0.43	0.95
+Aug.	0.85	0.63	0.53	0.93	0.88	0.86	**0.67**	0.94	0.88	0.73	0.59	0.96
+Aug.+Temp.	0.85	0.69	**0.63**	0.96	0.92	0.82	0.61	0.96	0.88	0.75	**0.62**	0.96
Dual-task models	0.77	0.41	0.41	0.97	**0.97**	**0.91**	0.34	0.90	0.86	0.57	0.37	0.93
+Aug.	0.83	0.66	0.53	0.95	0.92	0.68	0.41	0.96	0.88	0.67	0.46	0.96
+Aug.+Temp.	**0.89**	0.67	0.56	**0.97**	0.92	0.87	0.64	0.96	**0.91**	**0.76**	0.60	**0.97**

Runtime Performance: Table 1 shows the number of million floating point operations (MFlops) for AutoDVT for each layer. Current mobile GPUs can provide up to 384 GFlops (e.g. PowerVR GT7600 Plus), while AutoDVT only requires 148 MFlops, which would result in a theoretical computational overhead of 0.002 s when accounting for approximately 60% overhead for memory transfers and caching. Practically the frame-rate is limited by the image acquisition rate of the ultrasound probe, which is ∼20 fps. Thus, using our K80 GPU yields an application performance of ∼20 fps and ∼14 fps when using a Xeon E5-2686v4 CPU without GPU. We observed real-time frame rates for the implementations of AutoDVT model in two deep-learning frameworks, Theano and PyTorch. The dual-task architecture requires 50% fewer computations than running two separate task-specific networks (which amounts to 290 MFlops vs. 148 MFlops of AutoDVT).

Visual Exploration of Results: We used Gradient-weighted Class Activation Mapping (Grad-CAM) [12] to gain insight into which areas are considered by AutoDVT to make predictions. Figure 2 shows the obtained saliency maps, which qualitatively confirms that the model focuses on the relevant anatomy and makes the correct O/C and LM classifications in the example frames.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Visualisation of saliency maps of AutoDVT (MT+A.+T.) for O/C classification, obtained by applying Grad-CAM [12] to sample frames in a compression sequence at LM1. The super-imposed heatmaps (red is high, blue is low) indicate that AutoDVT prediction is most excited by the anatomically relevant areas around the femoral vein. (a–c) shows uncompressed state, (d, e) show compressed state, (f, g) show release of manual pressure. (h) shows a summary for a random example compression sequence with the predicted vessel compression state in red and the ground truth label in green. Note that, when the vessel is entirely closed, i.e. invisible, our method predicts the 'closed' state from spatio-temporal information and other structures in the image.

Limitations: This work focuses on DVT diagnosis in the groin area. Further evaluation will be required to confirm similar performance for all possible landmarks used during DVT examination especially in the area of the popliteal veins. Like most deep learning methods, domain shift is still a challenge despite having trained on data from four different devices. We are aware of the option to use Doppler ultrasound as additional source of information. However, the aim of this work is to make DVT diagnostics available for point-of-care applications. Doppler reduces the image acquisition frame rate significantly, needs to be adjusted by an experienced operator and does not necessarily increase detection rate of asymptomatic DVT [10]. Furthermore, not all POCUS transducers support Doppler imaging. A pure image-based approach is therefore desirable for POCUS DVT diagnostics and potential applications in ODE countries.

Conclusion: We have proposed a novel dual-task CNN to assist the real-time diagnosis of DVT. Our approach enables non-expert health practitioners to reliably support DVT diagnosis using machine learning guidance. Previous work [6] shows that 100% sensitivity and specificity can be reached for semi-automatic DVT classification and that also venous pressure can be determined [13]. Our approach can automate several of the currently required manual steps to achieve this level of diagnostic accuracy on a range of different devices. AutoDVT evaluates vascular compressibility at two anatomically classified landmark positions on the femoral vein. Landmark detection and compression state inference can be combined in a joint dual-path network and the tasks can be trained end-to-end. Our approach shows promising performance at accuracies greater than 90%,

which is well within the expected performance of expert examinations [3]. Three architectural choices: joint learning of open/close discrimination and landmark detection; data augmentation and careful restriction of model capacity as well as spatio-temporal convolutions, enabled substantial improvements in accuracy compared to baseline models. Our approach provides real-time performance and the potential to be used directly on mobile devices for POCUS diagnostics with significant impact on patient care and health care costs. In future work, we will evaluate our method on a more comprehensible dataset including patients with pathologies and generalise landmark classification to all relevant areas along the femoral vein.

References

1. Beckman, M.G., et al.: Venous thromboembolism: a public health concern. Am. J. Prev. Med. **38**(4, Supplement), S495–S501 (2010)
2. Stein, P.D., et al.: D-dimer for the exclusion of acute venous thrombosis and pulmonary embolism: a systematic review. Ann. Intern. Med. **140**(8), 589–602 (2004)
3. Schellong, S.M., et al.: Complete compression ultrasonography of the leg veins as a single test for the diagnosis of deep vein thrombosis. Thromb. Haemost. **89**(2), 228–234 (2003)
4. Guerrero, J., et al.: System for deep venous thrombosis detection using objective compression measures. IEEE Trans. Biomed. Eng. **53**(5), 845–854 (2006)
5. Friedland, N., Adam, D.: Automatic ventricular cavity boundary detection from sequential ultrasound images using simulated annealing. IEEE Trans. Med. Imaging **8**(4), 344–353 (1989)
6. Guerrero, J., et al.: Deep venous thrombosis identification from analysis of ultrasound data. Int. J. Comput. Assist. Radiol. Surg. **10**(12), 1963–1971 (2015)
7. Baumgartner, C.F., et al.: SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging **36**(11), 2204–2215 (2017)
8. Prevost, R., Salehi, M., Sprung, J., Bauer, R., Wein, W.: Deep learning for sensorless 3D freehand ultrasound imaging. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 628–636. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_71
9. Smistad, E., Løvstakken, L.: Vessel detection in ultrasound images using deep convolutional neural networks. In: Carneiro, G. (ed.) LABELS/DLMIA-2016. LNCS, vol. 10008, pp. 30–38. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_4
10. Lensing, A., et al.: A comparison of compression ultrasound with color doppler ultrasound for the diagnosis of symptomless postoperative deep vein thrombosis. Arch. Intern. Med. **157**(7), 765–768 (1997)
11. Caruana, R.: Multitask learning: a knowledge-based source of inductive bias. In: International Conference on Machine Learning (ICML) (1993)
12. Selvaraju, R.R., et al.: Grad-Cam: visual explanations from deep networks via gradient-based localization. CoRR abs/1610.02391 v3 **7**(8) (2016)
13. Crimi, A., et al.: Automatic measurement of venous pressure using B-mode ultrasound. IEEE Trans. Biomed. Eng. **63**(2), 288–299 (2016)