Deep Vein Thrombosis Screening with Three-Dimensional Deep Learning on Lower Extremity Computed Tomography Studies

Brianna Kozemzak[†], Anoop Manjunath[†], Trevor Tsue[†], and Jonathan X Wang[†]

Department of Biomedical Data Science, Stanford University,
Stanford, CA, 94305, USA
kozemzak@stanford.edu, amanjuna@stanford.edu, ttsue@stanford.edu, jonxwang@stanford.edu
www.stanford.edu
[†]*Equal Contributors*

Every year, 60,000-100,000 Americans die of complications arising from deep vein thrombosis (DVT), blood clots that form in deep lower extremity veins. DVT is the leading cause of preventable hospital death for trauma patients, and despite 0.1% of the population receiving a new diagnosis annually, DVT is underdiagnosed due to nonspecific symptoms, diagnostic difficulty, and overburdened radiologists, especially when examining trauma patients with other urgent conditions. Developing computer-aided detection (CAD) systems would alleviate these concerns; however, current systems classify computed tomography (CT) slices individually and yield many false positives from DVT-like dark spots in poorly contrasted areas. Recent advances in deep neural networks (DNNs) allow us to leverage spatial dependencies between slices in imaging studies to identify false positives and ultimately deploy DNN systems that lighten physicians workloads while not exacerbating alarm fatigue. To train our DNN, we have acquired 119 lower-body CT imaging studies labeled by radiologists for DVT at the pixel level. Using these studies, we have developed a DNN-based CAD system that will (1) segment targeted deep veins in a CT slice, (2) classify whether a DVT is present within multiple slices given segmentations of deep veins, and (3) evaluate different deep learning approaches for handling 3D datasets for DVT detection. For segmentation, we use a 2D U-Net, 2D VGG encoder-decoder, and 3D U-Net and find that VGG performs best (Dice: 0.07, IoU: 0.48, AUROC: 0.78). For classification, we use a 2D ResNet, CNN-RNN, and 3D Inception model with and without segmentation masks. We find that our CNN-RNN without masks performs best in AUROC (Average Precision: 0.31, AUROC 0.64) and 3D Inception with masks performs best in average precision (Average Precision: 0.33, AUROC: 0.62). By developing more effective detection algorithms, we hope to ensure more frequent and accurate diagnosis of DVT, thereby reducing its high mortality rate.

Keywords: Deep Vein Thrombosis; Computer Vision; Deep Learning

1. Introduction

Deep vein thrombosis (DVT) is a blood clot commonly found in deep veins of the lower extremities. Approximately 2.5-5% of individuals in the United States will be affected by DVT during their lifetime.[1] Pulmonary Embolism (PE) is a fatal complication that can arise when DVT goes undetected or is left untreated, and is caused by clots breaking off and traveling into

© 2016 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

the lungs.[2] According to some studies, it is the leading cause of preventable hospital death for trauma patients[3,4] and the leading cause of maternal mortality,[5] resulting in 60,000-100,000 American deaths annually.[6]

Despite its high prevalence, DVT continues to be under-diagnosed in the clinical setting.[6] One reason for this is the nonspecific physical symptoms of DVT, including swelling, pain, tenderness, and redness in the area of the thrombosis.[7] Another reason arises due to the diagnostic difficulty. Common risk factors include injury and trauma, where DVT can be overshadowed by the primary diagnoses.[4] A study of severe trauma patients found that three cases of DVT and three cases of PE were missed when 205 CT scans were reviewed by an independent radiologist.[4] Finally, physician burnout has been on the rise, and even though half of blood clots form following a clinical visit, the sheer number of cases and immediate risks commonly associated with DVT (e.g. obesity, pregnancy, trauma) overburden physicians.[8,9]

One potential solution to these problems are computer-aided detection (CAD) systems whereby patients at risk are automatically screened for blood clots (and potentially other diagnoses). Although Ultrasound (US) is most commonly used for the diagnosis of DVT, the results of recent studies suggest that Computed Tomography (CT) venography studies can have equal or greater diagnostic power and may be preferable for patients who are also being examined for PE, in which a CT angiography is required, rather than performing two separate diagnostic tests.[10,11]

A major challenge in developing such systems is that the quality of a CT venography study relies on proper diffusion and timing of a venous contrast injection prior to the imaging study being performed. In areas of confluence and turbulent blood flow, hypodense streaks may occur on the resulting CT images.[12] These streaks may resemble a DVT when visualizing a single slice of a CT study and can therefore lead to false positives when analyzing the slices individually. Fortunately, the false positive streaks tend to span more slices than a true DVT, and therefore can be distinguished from DVT by leveraging spatial information between adjacent slices, just as radiologists do when they analyze CT studies for DVT.

Methods for incorporating 3D spatial context have only recently become available. Increased computational power and larger datasets have led to a resurgence of deep neural networks (DNNs) and progress

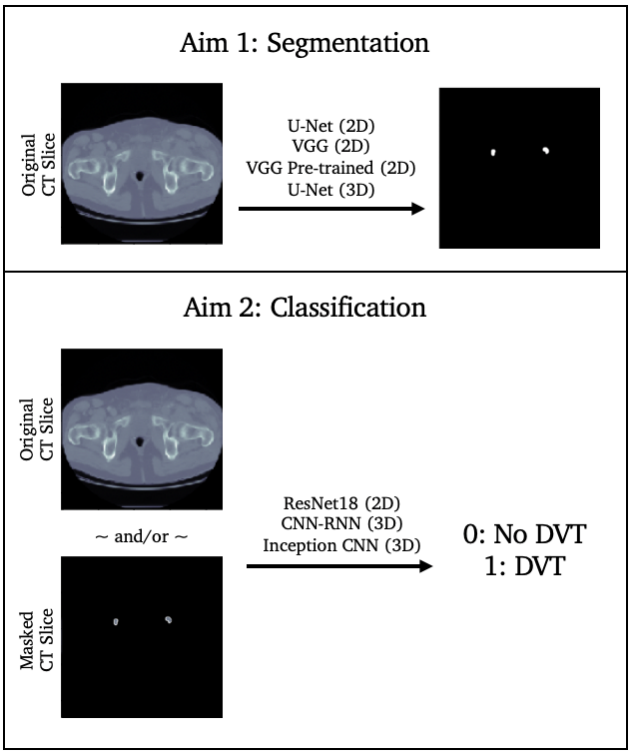


Fig. 1: Inputs, outputs, and models used to accomplish the two primary tasks in this study: segmentation and classification.

in the field of computer vision. Many networks have been trained on millions of images and are easily adapted for new tasks. We are optimistic that this new technology, coupled with innovative architectures to process 3D spatial information, will resolve many of the diagnostic difficulties arising from the low contrast areas.

In this study, we aim to (1) segment targeted deep veins in a CT slice, (2) classify whether a DVT is present within multiple slices given segmentations of deep veins, and (3) evaluate different deep learning approaches for handling 3D datasets for DVT detection. In our segmentation task, a CT slice or set of slices is passed as input to our model and a binary segmentation mask that highlights deep veins of interest is outputted from our model. For classification, we input a CT slice or set of slices, and we output a single binary value indicating whether the center slice contains a DVT. For both tasks, we employ 2D models as a baseline, and compare their performance to various 3D approaches. An overview of our workflow is depicted in Figure 1.

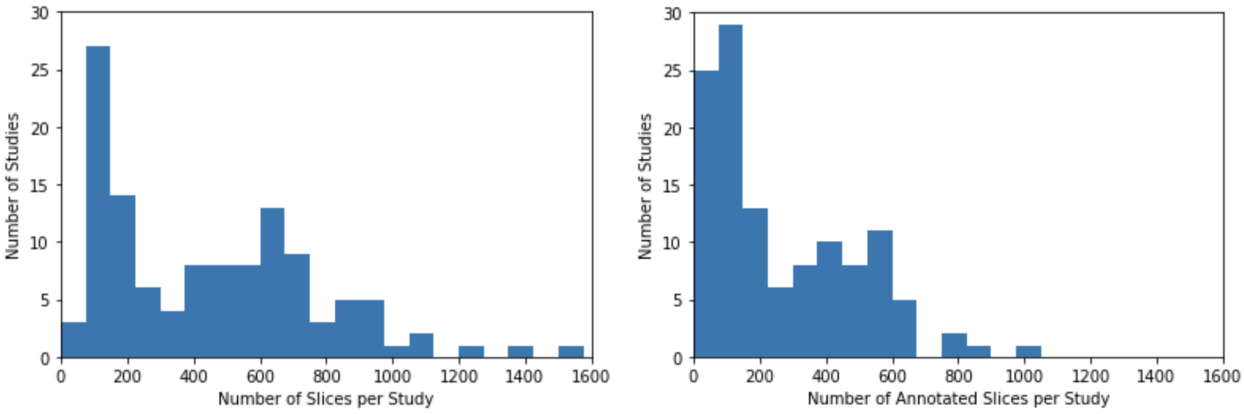
2. Related Work

To the best of our knowledge, no computer-aided detection (CAD) systems have been developed for the automated detection of deep vein thrombosis (DVT) from computed tomography (CT) studies. However, 3D convolutional approaches have demonstrated superior performance to 2D approaches for the related task of detecting pulmonary embolism in chest CT studies, achieving a sensitivity of 83% with 2 false positives per volume. This is the current state of the art for computer-aided pulmonary embolism detection.[13]

In 2016, two major approaches for leveraging 3D spatial information in computer vision tasks emerged: (1) 3D convolutional neural networks (CNNs) such as 3D U-Net,[14] and (2) 2D CNN and recurrent neural network (RNN) hybrids such as the FCN-RNN hybrid developed by Chen et al.[15] 3D U-Net has already been demonstrated to outperform U-Net,[16] its 2D counterpart, in many segmentation tasks.[17–19] Likewise, the RNN-CNN hybrid outperformed U-Net in segmenting 3D neuronal and fungal structures.[15] However, no studies have compared the effectiveness of the two approaches, so we utilize architectures from both approaches in an effort to compare their performance.

3. Data

We were granted access to a dataset consisting of 119 lower extremity computed tomography (CT) studies from 97 Stanford patients, 22 of whom had two CT studies performed. Each study consisted of 29 to 1,519 grayscale slices with pixel values represented by 16-bit unsigned integers. Anywhere from 0 to 1,031 slices were annotated by a radiologist. See Figure 2 for distributions of slices per study and annotated slices per study. Slices without an annotation were assumed to be negative for DVT in our classification task based on our knowledge of how the data was annotated. Our dataset was split randomly by patient ID into training, validation, and testing subsets at a ratio of 60% - 20% - 20%.



(a) Slices (b) Annotated Slices

Fig. 2: Distribution of slices per CT study and of annotated slices per CT study.

4. Methods

4.1. *Data Pre-processing*

For both the segmentation and classification tasks, CT slices were resized to 224 × 224 pixels from 512 × 512 pixels in an effort to make their dimensions compatible with weights pre-trained on ImageNet for our models. The pixel values were mean-centered and standardized according to the mean and standard deviation of the pixel values of images in our training set. For the classification task, pixel values were clipped to the interval $[0, 4095]$.

Depending on the dimension and task, our definition of an example from our dataset changed. For our 2D baselines, individual slices were treated as separate examples for input to the models with their corresponding mask/label the targeted output. For the 3D segmentation task, chunks of 8 consecutive CT slices were used as input to the models, with the corresponding 8 consecutive binary deep vein masks used as the targeted output. Thus, our input and output dimensions for 3D segmentation were 8 × 224 × 224 pixels. For the 3D classification task, the input consisted of chunks of 11 consecutive CT slices, and the targeted output was a single binary indicator of whether DVT was present in the center slice of the chunk. The number of consecutive slices to include was selected based on its ability to capture the entirety of a true DVT, allowing the model to identify false positives that typically span more slices than a true DVT. Images on the edges of the input volume were included in our examples by padding the chunks with empty images.

4.2. *Data Augmentation*

The training set of reference images and masks was augmented in order to combat overfitting. A generator was developed for the segmentation task that applied simple transformations to both 2D and 3D training examples. Rotations up to 72°, shifts up to 5% of the width and height, reflections over the x-axis, and scaling up to 5% were permitted. The parameters for image augmentation were selected based on a brief hyperparameter search according to segmentation performance on the validation set. For classification, a training generator that permitted reflections over both the x-axis and y-axis with a probability of 0.5 was used.

4.3. *Segmentation*

4.3.1. *Models*

We implement three convolutional neural network (CNN) architectures for the deep vein segmentation task. All of these networks have architectures consisting of an encoder, which takes an image and generates a high dimensional feature matrix, and a decoder, which takes a high dimensional feature matrix and generates a segmentation mask.

The first model is a 2D U-Net. Convolutional layers in the encoder are separated by max pooling operations while the second half are separated by upsampling operations.[16] U-Net is known for its increased precision in segmentation that results from concatenating high resolution features to outputs in the decoder. It yields state of the art performance on image segmentation tasks in biological contexts with fast train and test speeds. This is especially important for the potential application of real-time DVT image segmentation. Our implementation of U-Net involves 10 convolutional layers, selected according to a brief hyperparameter search (infrastructure especially inspired by[20]). The inputs to this model are individual 224 $\times$ 224 pixel images, and the outputs are equivalently shaped masks.

A modified VGG16 architecture was implemented for our second segmentation model. This model uses the encoder from the original VGG16 model,[21] but adds a decoder that mirrors the encoder's max pooling layers with upsampling. The major advantage of this architecture is that it allows for transfer learning in the encoder by using weights that were pre-trained on ImageNet. By using weights that were trained on a larger data set, we decrease the likelihood of over-fitting to the training set. Like our 2D U-Net, this model takes 224 $\times$ 224 pixel images as input and outputs an equivalently shaped mask.

The third CNN architecture employed for the segmentation task was a 3D implementation of U-Net. This model is capable of performing volumetric segmentation.[22] The model architecture is very similar to that of U-Net, with 2D convolutions and max pooling operations replaced by 3D variants. The input to this model consists of a volume of 8 adjacent images, each 224 $\times$ 224 pixels. Its outputted masks maintain this shape at 8 $\times$ 224 $\times$ 224 pixels.

4.3.2. *Loss Function*

The choice of loss function is especially important with the presence of class imbalance, as is the case in in this study. The ratio of pixels representing deep veins to those that do not is approximately 1:536 in our training set. We experimented with dice loss and weighted binary cross-entropy loss. Ultimately, we found that dice loss had the best performance on our validation set:

$$\mathcal{L} = -\left(\frac{2}{N}\right)\left(\frac{\sum_{i=1}^{N} t_i p_i + 1}{\sum_{i=1}^{N} t_i + \sum_{i=1}^{N} p_i + 1}\right)$$

where N is the number of pixels in the image, t_i is the true label for a pixel, and p_i is the probability output by the model. The 1 is added for smoothing in case the denominator would ever become zero.

4.3.3. *Evaluation*

Model performance was evaluated using traditional metrics of segmentation performance, including intersection over union (IoU),[23] and dice coefficient.[24] We perform an additional pixel-wise analysis where we treat each pixel as a binary classification task of containing deep vein tissue (1) or not (0). We then plot a receiver operating characteristic (ROC) curve, which demonstrates the trade-off between sensitivity and specificity across a range of probability thresholds. Hyperparameters including learning rate, number of layers, and layer activation function were determined according to model performance on these metrics on the validation set.

4.4. *Classification*

4.4.1. *Models*

To approach the DVT classification task, we utilized both 2D and 3D images. Our 2D images used a single CT slice in the study. For the 2D images, we utilized a standard 2D CNN. The 3D images utilized 11 slices in the study: 5 of the slices before, the center slice, and 5 slices after. Our 3D models were trained to predict DVT for the center slice. For the 3D volumes, we utilized a CNN-RNN and a 3D CNN. In all cases, we also experimented with providing the labeled segmentation masks of the veins as an extra channel.

Our 2D CNN utilized the ResNet18 architecture[25] pretrained on ImageNet. This architecture presents residual connections that act as shortcuts, skipping one or more layers in the network. Consequently, this network becomes easier to optimize, addresses the vanishing gradient problem, and also has groundbreaking performance on the ImageNet dataset. The final layers are a global average pooling and a fully connected layer with a softmax activation to output a probability over the two possible classes.

The 3D CNN-RNN also utilized the ResNet18 architecture. For the 11 slices, we passed them individually through the CNN, creating a feature vector for each one of the slices. Then, we pass each of the feature vectores through a bidirectional LSTM and a final fully connected layer, predicting the output for the middle slice.

The Kinetics 3D Inception model[26] utilizes a 3D version of Google's Inception v3 network. This network utilizes different convolutional layers in parallel and concatenates the results together. Unlike the CNN-RNN, this model uses a 3D kernel to run over multiple slices, incorporating this spatial information in the convolution calculations. In recent work, this has been shown to produce state of the art results for lung cancer screening on CT scans.[27]

4.4.2. *Loss Function*

For our classification problem, we used the Cross Entropy Loss on the softmax probability outputs:

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} y_c^{(i)} \log(\hat{y}_c^{(i)})$$

where M is the batch size, C is the number of classes (i.e., 2), y is the ground truth label, and $\hat{y}$ is the probability output of the model. Additionally, we added weight decay regularization to our loss function to penalize the model for having large weights, thereby reducing over-fitting.

4.4.3. *Evaluation*

To analyze our classification performance, we examined area under the receiver operator characteristic (AUROC) and average precision (AP). Because we have a unbalanced dataset with more negative examples, accuracy is not the best metric, as a null classifier could have very high accuracy (e.g., if 80% of the data were negative, then the null classifier would classify 80% correctly by always guessing the majority class).

AUROC shows the true positive rate ($\frac{TP}{P}$) as a function of false positive rate ($\frac{FP}{N}$), where the different values are a result of changing the classification threshold of the model. Essentially, this metric allows us to choose a false positive rate (what percent of negatives we miss classify as positives) that we accept and tells us how correctly we label the positives.

AP shows precision ($\frac{TP}{TP+FP}$) as a function of recall ($\frac{TP}{P}$). Essentially, we can choose the percent of true case that we want to catch (recall) and then show how many true positives compare to false positives. Therefore, for our problem, the average precision is the strongest metric, as we want to minimize the number of false positives resulting from poor CT contrast.

5. Results

5.1. *Segmentation*

Table 1 illustrates the performance of each deep vein segmentation model according to the putative metrics of segmentation performance: dice coefficient and intersection over union. The 2D VGG with encoder pre-trained on ImageNet had the best performance according to both dice coefficient and intersection over union. The performance of the 2D VGG architecture without weights closely followed in performance. The 2D U-Net model had the worst performance on the validation and test set despite middling performance on the training set indicating over-fitting. This over-fitting problem is especially present for the 3D U-Net architecture, which despite its superior performance on the training set, performed worse than the 2D VGG architecture with and without weights on the validation and test sets.

Figure 3 depicts the performance of each model according to pixel-wise binary classification metrics of specificity and sensitivity. The ranking of the area under the receiver operating characteristic (AUROC)

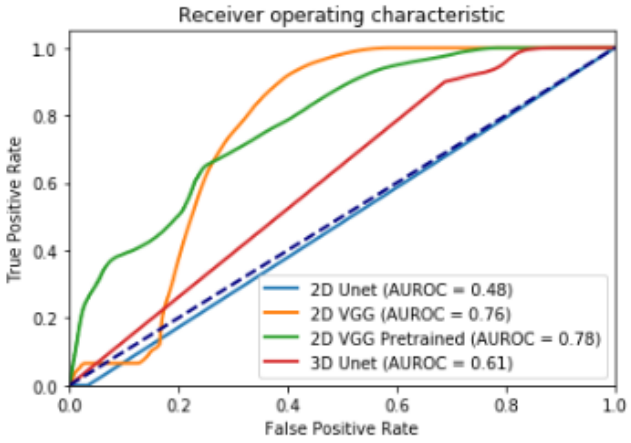


Fig. 3: Segmentation test ROC curve.

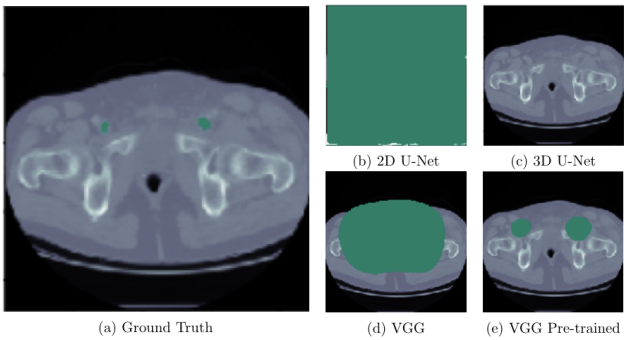


Fig. 4: Example deep vein segmentations from a random test set image on all four models.

recapitulates the results presented by the segmentation metrics. In particular, the VGG network with a pre-trained encoder has the best performance followed closely by the VGG network without pre-training. This model's performance is followed by that of the 3D U-Net, and finally by the 2D U-Net. With an AUROC of 0.48, the 2D U-Net has almost no prediction power. One interesting feature is that the pre-trained encoder appears to regularize the ROC curve of the VGG network. In particular, transfer learning improves performance when the false positive rate is set to low values, whereas the curve for non-pre-trained encoder hugs the diagonal at these low false positive values.

Table 1: Segmentation performance on intersection over union and dice coefficient metrics.

Model	Metric	Training	Validation	Test
2D U-Net	Dice	0.12	0.01	0.01
	IoU	0.23	0.05	0.04
2D VGG	Dice	0.31	0.05	0.05
	IoU	0.56	0.43	0.46
2D VGG Pre-trained	Dice	0.38	0.12	**0.07**
	IoU	0.67	0.55	**0.48**
3D U-Net	Dice	0.46	0.09	0.04
	IoU	0.72	0.41	0.42

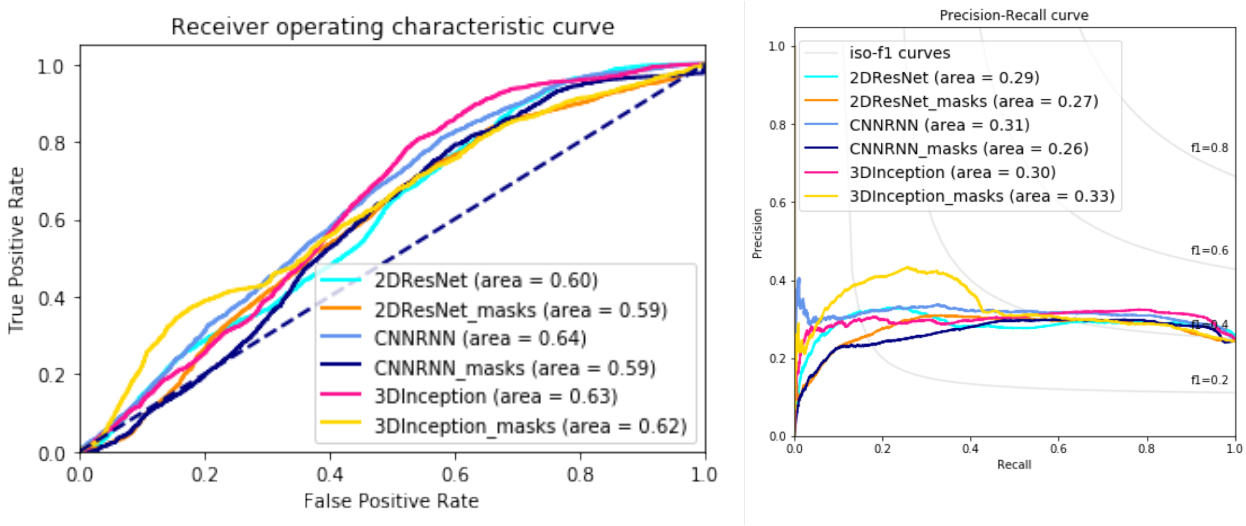
Finally, Figure 4 presents a qualitative analysis of the relative ability of all four models to segment the deep veins from a randomly selected image from the test set (ground truth shown on the left). The 2D U-Net in this case predicts almost all the pixels in the image to belong to a deep vein. The VGG with pre-trained encoder is able to segment slightly large circles around each of the deep veins, whereas the VGG model without pre-trained weights appears to have a general sense of the area of the CT scan. Finally, the 3D U-Net appears to have functioned as essentially a null classifier.

5.2. *Classification*

Table 2 shows the classification results for both the validation and test sets on all models. Figure 5 shows the ROC and precision-recall curves for our models, which demonstrate the performance of the model across a range of probability thresholds. In the validation dataset, the 2D ResNet with masks performs best in both the AUROC and average precision. In the test set, the CNN-RNN has the best AUROC and the 3D inception with masks has the best average precision. Overall, we do see a slight increase in performance on the test set when we use the 3D models: CNN-RNN and 3D Inception. However, adding the segmentation masks to the models does not necessarily help the model.

Table 2: Classification performance for our models on AUROC and average precision evaluation metrics.

Model	Metric	Validation	Test
2D ResNet	AUROC	0.80	0.60
	Avg. Precision	0.74	0.29
2D ResNet, Masks	AUROC	**0.82**	0.59
	Avg. Precision	**0.77**	0.27
CNN-RNN	AUROC	0.81	**0.64**
	Avg. Precision	0.77	0.31
CNN-RNN, Masks	AUROC	0.79	0.59
	Avg. Precision	0.70	0.26
3D Inception	AUROC	0.81	0.63
	Avg. Precision	0.70	0.30
3D Inception, Masks	AUROC	0.73	0.62
	Avg. Precision	0.70	**0.33**



(a) Receiver Operating Characteristic (ROC)

(b) Precision-Recall Curve

Fig. 5: Performance of DVT classification models on the test set.

6. Discussion

In segmentation, we find our models perform rather poorly on the task, with a highest dice coefficient of 0.07 and intersection over union of 0.48 for the VGG model with pre-trained weights. One factor that may contribute to this poor performance is missing deep veins in the segmentation masks provided by the labeling radiologist (Figure 6). Additionally, these missing masks seem to be disproportionately impacting the 3D model as a particular vein may not be traced fully through the volume, thus the model learns to be more careful in mask prediction. We believe one potential solution to mitigate this problem would be to ensure that masks are symmetric during data pre-preprocessing. We are also interested in restructuring the 3D model, such that the segmentations predicted from a volume is the single center slice rather than a volume of slices. We believe this may help the 3D model become more robust to the label issue as well.

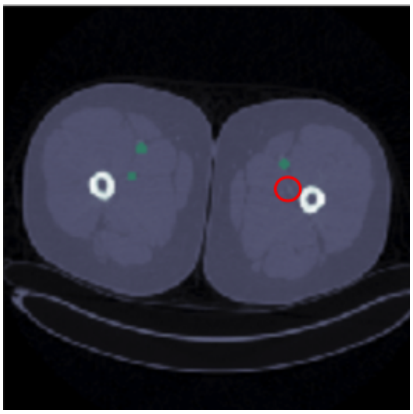


Fig. 6: Example of ground truth annotation missing vein in red circle.

Interestingly, our VGG outperforms U-Net, which traditionally has been shown to generate state of the art results on biological applications. The major functional difference between the models is that U-Net concatenates layers from the encoder in the decoder. We hypothesize that this concatenation is contributing to the over-fitting problem, as it increases the variance of the model. Additionally, our results validate the importance of pre-training, as the VGG model had better performance with pre-trained weights than without pre-trained weights.

For classification, our main issue seems to arise from our train, validation, and test split (Table 2). With limited tuning, our validation set results are disproportionately greater than

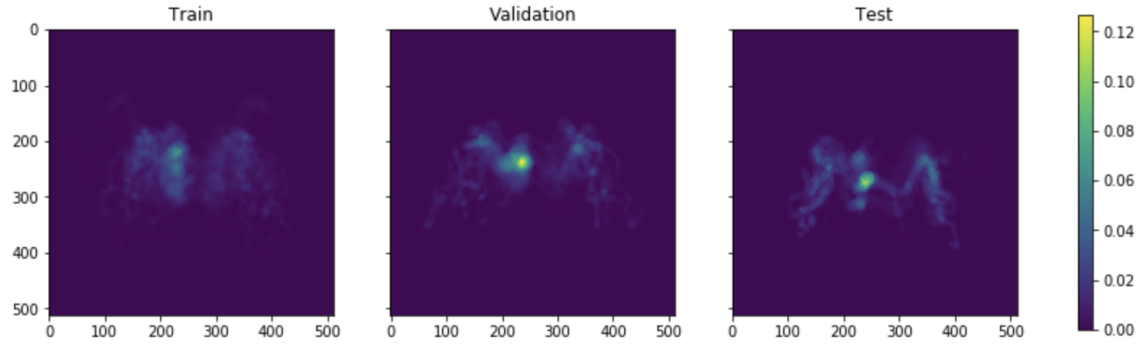


Fig. 7: Distribution of deep veins in the train, validation, and test set images. Each pixel represents the proportion of slices that had a deep vein annotation that included the pixel.

our test sets, with an AUROC of 0.82 and average precision of 0.77 in comparison to an AUROC of 0.64 and average precision of 0.33 on the test. We believe our small dataset size is leading to significant differences in the characteristics of the data splits. We randomly split our data into train, validation, and test sets based on patient ID. However, if the distribution of the test set randomly happens to be different from that of our validation set, our model would not be able to generalize to this different distribution, especially since we have limited labeled data. For example, in the validation set, 15% of slices were positive for DVT, while in the test only 11% were positive for DVT. Without these issues, it does seem that the algorithms are performing well based on the validation metrics. Interestingly, the dataset splitting issue seems to be less of a problem in the segmentation task than in the classification task, since the difference between validation and test performance is less. However, the locations of the ground truth deep veins appear fairly different between all dataset splits (Figure 7).

As with most medical imaging studies, our work is limited to data from a single institution. Thus, it is likely that we are over-fitting to certain institutional practices, such as imaging instrument quality and patient admission diagnoses. The dataset also comes from patients seeing an interventional radiologist, rather than general trauma patients, which is the intended use case. Unfortunately, our models currently do not perform well enough to be used as part of a computer aided detection system in clinical practice.

Our main interest is further iteration on our models to get interpretable and clinically relevant performance. Of primary importance is fixing the data issues with missing labels, as well as generating a more balanced data split through some form of matching based on CT study statistics. Additionally, we had limited time for hyperparameter tuning, and strongly believe that a more extensive hyperparameter search could yield much better model performance. Finally, we would be interested in performing ensembling, as this has also been shown to improve performance, especially in cases of over-fitting.[28]

Acknowledgments

We would like to thank Dr. Lawrence Hofmann for granting us access to the dataset used in this project and for his guidance throughout. We would also like to thank Andre Souffrant, David Cohn, and Jared Dunnmon for introducing us to the dataset and for providing high-level algorithmic advising.

References

1. K. Burnand, A. Irvine, N. Wilson *et al.*, Deep vein thrombosis: pathology, in *Diseases of the veins*, (Arnold London, 1999) pp. 249–289.
2. H. Decousus, A. Leizorovicz, F. Parent, Y. Page, B. Tardy, P. Girard, S. Laporte, R. Faivre, B. Charbonnier, F.-G. Barral *et al.*, A clinical trial of vena caval filters in the prevention of pulmonary embolism in patients with proximal deep-vein thrombosis, *New England Journal of Medicine* **338**, 409 (1998).
3. G. A. Maynard, *Preventing hospital-associated venous thromboembolism: a guide for effective quality improvement* (Agency for Healthcare Research and Quality, US Department of Health and , 2016).
4. S. Hamada, C. Espina, T. Guedj, R. Buaron, A. Harrois, S. Figueiredo and J. Duranteau, High level of venous thromboembolism in critically ill trauma patients despite early and well-driven thromboprophylaxis protocol, *Annals of intensive care* **7**, p. 97 (2017).
5. C. J. Berg, J. Chang, L. Elam-Evans, L. Flowers, J. Herndon, K. A. Seed and C. J. Syverson, Pregnancy-related mortality surveillance–united states, 1991–1999 (2003).
6. M. G. Beckman, W. C. Hooper, S. E. Critchley and T. L. Ortel, Venous thromboembolism: a public health concern, *American journal of preventive medicine* **38**, S495 (2010).
7. F. A. Anderson, H. B. Wheeler, R. J. Goldberg, D. W. Hosmer, N. A. Patwardhan, B. Jovanovic, A. Forcier and J. E. Dalen, A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism: the worcester dvt study, *Archives of internal medicine* **151**, 933 (1991).
8. N. McLeod and G. Montane, The radiologist assistant: the solution to radiology workforce needs, *Emergency radiology* **17**, 253 (2010).
9. A. P. Kiraly, C. L. Novak, D. P. Naidich, I. Vlahos, J. P. Ko and G. T. Brusca-Augello, A comparison of 2d and 3d evaluation methods for pulmonary embolism detection in ct images, in *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, 2006.
10. S. Thomas, S. Goodacre, F. Sampson and E. Van Beek, Diagnostic value of ct for deep vein thrombosis: results of a systematic review and meta-analysis, *Clinical radiology* **63**, 299 (2008).
11. M. D. Cham, D. F. Yankelevitz, D. Shaham, A. A. Shah, L. Sherman, A. Lewis, J. Rademaker, G. Pearson, J. Choi, W. Wolff *et al.*, Deep venous thrombosis: detection by using indirect ct venography, *Radiology* **216**, 744 (2000).
12. W.-Y. Shi, L.-W. Wang, S.-J. Wang, X.-D. Yin and J.-P. Gu, Combined direct and indirect ct venography (combined ctv) in detecting lower extremity deep vein thrombosis, *Medicine* **95** (2016).
13. N. Tajbakhsh, M. B. Gotway and J. Liang, Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.
14. F. Milletari, N. Navab and S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016.
15. J. Chen, L. Yang, Y. Zhang, M. Alber and D. Z. Chen, Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation, in *Advances in neural information processing systems*, 2016.
16. O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

17. J. Patravali, S. Jain and S. Chilamkurthy, 2d-3d fully convolutional neural networks for cardiac mr segmentation, in *International Workshop on Statistical Atlases and Computational Models of the Heart*, 2017.
18. X. Zhou, K. Yamada, T. Kojima, R. Takayama, S. Wang, X. Zhou, T. Hara and H. Fujita, Performance evaluation of 2d and 3d deep learning approaches for automatic segmentation of multiple organs on ct images, in *Medical Imaging 2018: Computer-Aided Diagnosis*, 2018.
19. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in *International conference on medical image computing and computer-assisted intervention*, 2016.
20. Z. Hao, Implementation of deep learning framework – unet, using keras (Github.
21. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
22. Ö. iek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, 3d u-net: Learning dense volumetric segmentation from sparse annotation, in *MICCAI*, 2016.
23. M. A. Rahman and Y. Wang, Optimizing intersection-over-union in deep neural networks for image segmentation, in *International symposium on visual computing*, 2016.
24. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin and M. J. Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, (Springer, 2017) pp. 240–248.
25. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *CoRR* **abs/1512.03385** (2015).
26. J. Carreira and A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, *CoRR* **abs/1705.07750** (2017).
27. D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado *et al.*, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, *Nature medicine* , p. 1 (2019).
28. Y. Wu and J. I. Arribas, Fusing output information in neural networks: Ensemble performs better, in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, 2003.