High Performance Embolism Detection in Real-World CT Pulmonary Angiography Examinations Using Deep Learning

Ali Teymur Kahraman[1]*, Tomas Fröding[2]*, Dimitris Toumpanakis[3,4], Christian Jamtheim Gustafsson[5,6]**, Tobias Sjöblom[1]**+

[1] Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden.

[2] Department of Radiology, Nyköping Hospital, Nyköping, Sweden

[3] Consultant, Neuroradiology, Karolinska University Hospital, Stockholm, Sweden

[4] PhD fellow, Department of Surgical Sciences, Uppsala University

[5] Radiation Physics, Department of Hematology, Oncology, and Radiation Physics, Skåne University Hospital, Lund, Sweden.

[6] Department of Translational Medicine, Medical Radiation Physics, Lund University, Malmö, Sweden.

* Co-first authors

** Co-Senior authors

+ Corresponding author: tobias.sjoblom@igp.uu.se

Abstract

Purpose: To develop a pipeline that automatically classifies patients for pulmonary embolism (PE) in CT pulmonary angiography (CTPA) examinations with high sensitivity and specificity.

Materials and Methods: Seven hundred non-ECG-gated CTPA examinations from 652 patients (median age 72 years, range 16-100 years; interquartile range 18 years; 353 women) performed at a single institution between 2014 and 2018, of which 149 examinations contained PE, were used for model development. The nnU-Net deep learning-based segmentation framework was trained and validated in 5-fold cross-validation. To enhance classification, we applied logical rules based on PE volume and probability thresholds. External model testing was then performed in 770 and 34 CTPAs from two independent datasets.

Results: For patient-level classification, a threshold PE volume of 20 mm^3 resulted in the best balance between sensitivity and specificity. In internal cross-validation and test set, the trained model correctly classified 123 of 128 examinations as positive for PE (sensitivity 96%; 95% C.I. 91-98%) and 521 of 551 as negative (specificity 95%; 95% C.I. 92-96%). In the first external test dataset, the trained model correctly classified 31 of 32 examinations as positive (sensitivity 97%; 95% C.I. 84-99%) and 2 of 2 as negative (specificity 100%; 95% C.I. 34-100%). In the second external test dataset, the trained model correctly classified 379 of 385 examinations as positive (sensitivity 98%; 95% C.I. 97-99%) and 346 of 385 as negative (specificity 90%; 95% C.I. 86-93%).

Conclusion: Beyond state-of-art classification for PE in CTPA was achieved using nnU-Net for deep learning-based segmentation in combination with volume- and probability-based classification.

Introduction

Pulmonary embolism (PE) is a potentially life threatening blockage of pulmonary arteries caused by blood clotting and is associated with significant morbidity and mortality (1). PE affects >400,000 patients in Europe (2) and between 300,000 and 600,000 patients in the US (3) causing an estimated >100,000 deaths annually (4). PE is the leading cause of preventable hospital deaths in the world (5), demanding rapid clinical management (6). The computed tomography pulmonary angiography (CTPA) imaging method is the current gold standard for PE diagnosis (7). The CTPA is a CT scan performed after intravenous injection of iodinated contrast medium. As the emboli do not absorb contrast medium they can be recognized as dark filling defects in the pulmonary arteries (8). Identification of PE in CTPA is time-consuming for the radiologist and requires considerable training and attentiveness, and the inter-observer variability is high for small, sub-segmental emboli (9). An automated solution for detection of PE in CTPA has potential to assist the radiologist by reducing reading times and the risk of emboli being overlooked.

Developing a general solution for automatic detection of PE has proven challenging because of anatomical variation, motion and breathing artifacts, inter-patient variability in contrast medium concentration, and concurrent pathologies. Over the past two decades, automated PE detection has been attempted using deterministic models, such as image processing and analysis techniques (11,12), or probabilistic/statistical models such as machine learning (13–15) and deep convolutional neural networks (16,17). Yet, the accuracies of these solutions have been insufficient for clinical use due to low sensitivity (11,14,16) and high false positive rate (11,12,14,15), potentially caused by training on small datasets (11,12,14–16). The state-of-art is a residual neural network (ResNet) classification architecture on 1465 CTPA examinations with sensitivity of 92.7% and specificity of 95.5% (18). To mitigate dataset size obstacles of the classification problem, a fine-tuned U-Net-like architecture could be used as a semantic segmentation model which improved its performance in several medical image segmentation tasks (19). The no-new U-Net framework (nnU-Net) successfully addresses challenges of finding the best U-net model and fine-tuning its hyperparameters (20). Here, we sought to take advantage of the segmentation performance of the nnU-Net framework in a pipeline that automatically classifies routine patient CTPA examinations as having PE or not with higher sensitivity and specificity than the state-of- art.

Materials and Methods

Internal dataset

The single-institution (Nyköping Hospital, Sweden) retrospective dataset consisted of 700 non-ECG-gated CTPA examinations performed between 2014 and 2018 (n=149 positive for PE); 383 CTPA examinations from 353 women (age range 16-97 years; median age 73 years; interquartile range 20 years) and 317 from 299 men (age range 19-100 years; median age 71 years; interquartile range 15 years) (21). The CTPAs were clinical routine examinations exported from a list in chronological order, with time gaps to include a larger number of CT scanners. The CTPAs were acquired on 5 different CT scanners (Somatom Definition Flash, Siemens Healthcare, Erlangen, Germany; LightSpeed VCT, General Electric (GE) Healthcare Systems, Waukesha, WI, USA; Brilliance 64, Ingenuity Core and Ingenuity CT, Philips Medical Systems, Eindhoven, the Netherlands). As contrast medium, Omnipaque 350 mg I/ml (GE Healthcare Systems, Waukesha, WI, USA) was used. The most frequently used CT image acquisition parameters were slice thickness 0.625 mm (range 0.625 mm - 2.0 mm), pixel spacing 0.7 mm (range 0.59 mm - 0.98 mm), and tube voltage 100 kV (range 80 kV - 120 kV). Collection and analysis of CTPA examinations was approved by the Swedish Ethical Review Authority (EPN Uppsala Dnr 2015/023 and 2015/023/1). The CTPA data was anonymized and exported in Digital Imaging and Communications in Medicine (DICOM) format, using a hardware solution (Dicom2USB). The CTPAs were reviewed and annotated using the open-source software Medical Imaging Interaction Toolkit (MITK) (22) by two radiologists (DT and TF) with 6 and 16 years of experience. Each CTPA was annotated by either DT or TF. All blood clots in 149 CTPA examinations positive for PE were manually segmented in axial view, image by image, resulting in 36,471 segmentations.

External datasets

Two publicly available datasets were used for external testing; the Ferdowsi University of Mashhad's PE dataset (FUMPE) (23) and the RSNA-STR Pulmonary Embolism CT (RSPECT) Dataset (24). The FUMPE dataset contains 35 CTPAs with voxel-level PE annotation by radiologists. Of the 35 CTPAs, 2 were negative for PE, 32 were positive and one was excluded for lack of ground truth annotation. The RSPECT dataset consisted of a training (n=7279) and a test (n=2167) set and image-level annotations were provided for the training set by several subspecialist thoracic radiologists. From the RSPECT training dataset were selected 385 CTPAs of the total 398 with central PE, and 13 examinations were excluded because of DICOM to nifti file conversion error. Of the 4877 CTPAs without PE or other true filling defect, 385 examinations were randomly selected. An overview of our internal and external datasets is shown in Figure 1.

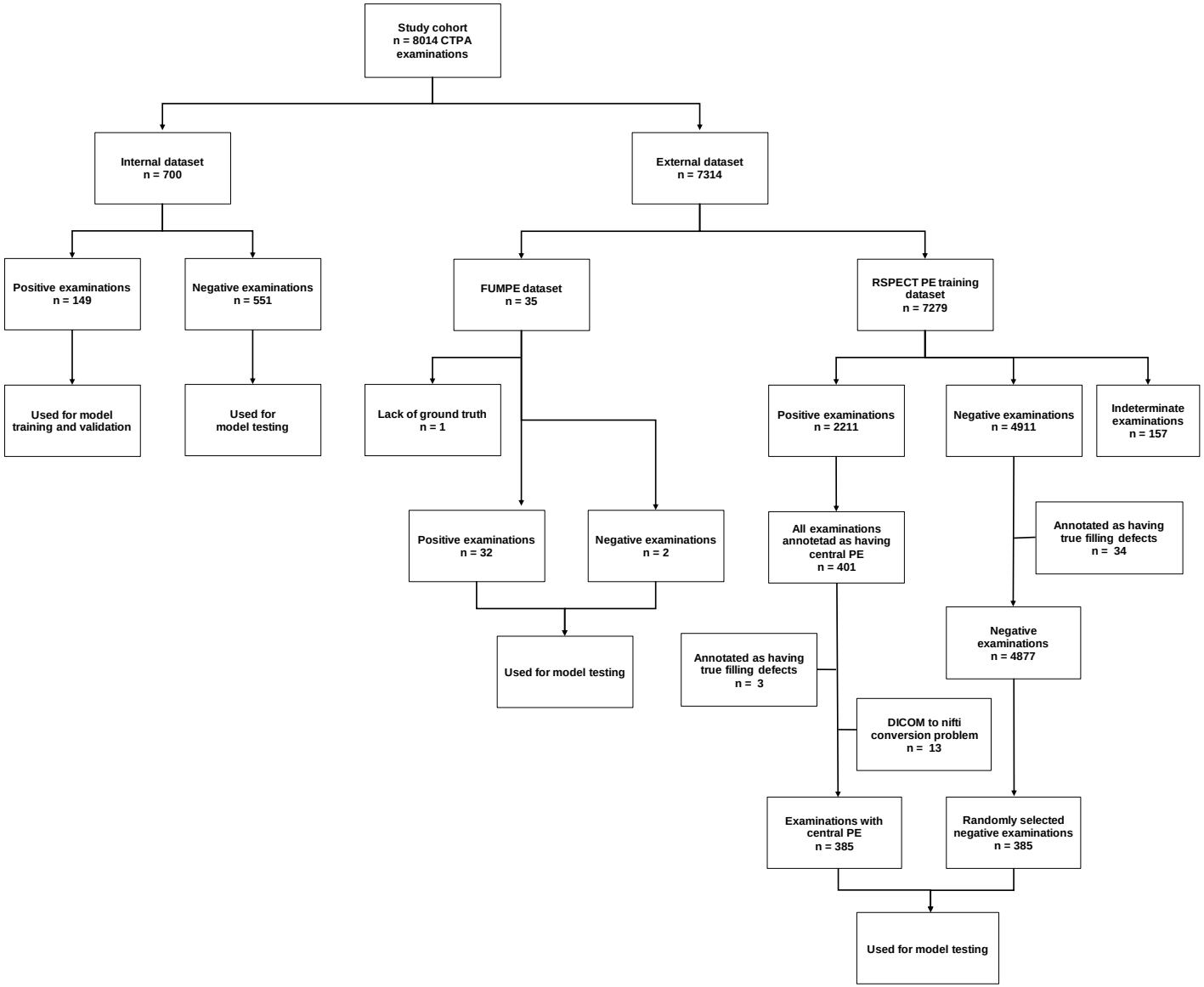
Study cohort
n = 8014 CTPA examinations
Internal dataset
n = 700
External dataset
n = 7314
Positive examinations
n = 149
Negative examinations
n = 551
FUMPE dataset
n = 35
RSPECT PE training dataset
n = 7279
Used for model training and validation
Used for model testing
Lack of ground truth
n = 1
Positive examinations
n = 2211
Negative examinations
n = 4911
Indeterminate examinations
n = 157
Positive examinations
n = 32
Negative examinations
n = 2
All examinations annotetad as having central PE
n = 401
Annotated as having true filling defects
n = 34
Used for model testing
Annotated as having true filling defects
n = 3
Negative examinations
n = 4877
DICOM to nifti conversion problem
n = 13
Examinations with central PE
n = 385
Randomly selected negative examinations
n = 385
Used for model testing

Figure 1. Internal and external datasets for training and testing of a segmentation-based classification model for pulmonary embolism detection. Positive examinations refer to the patient having pulmonary embolism (PE) and negative examinations are patients without PE. True filling defect refers to tumor invasion, stump thrombus, catheter, embolized wire, or other obvious non-PE condition as defined in the RSPECT dataset.

nnU-Net Model training and validation

For model training, the nnU-net DL open-source framework implemented in a Docker container (Docker Inc., Palo Alto, California, USA) was used (20). Given a training dataset, nnU-Net automatically configures an end-to-end experimental pipeline. The PE positive examinations from the internal dataset (n=149) were randomly assigned to training (80%, n=119) and validation sets (20%, n=30) in 5-fold cross-validation during model training.

Automated Classification Pipeline

After model training and validation, the validated model was embedded in a classification pipeline consisting of three steps, pre-processing, segmentation inference, and post-processing (Figure 2). As nnU-Net requires the Nifti file format for model inference, all DICOM data were converted to Nifti format in the pre-processing step. Next, the nnU-Net model inference was performed. Since the nnU-Net model is a volumetric segmentation model, the inference results in a segmentation mask of predicted pulmonary emboli. By thresholding these predicted segmentations based on total predicted emboli volume, a patient-level PE and non-PE classification was obtained. Multiple PE cutoff values from 0 to 200 mm^3 were explored to convert the PE segmentation output of the trained nnU-Net model to accurate classification at the patient level. Fine-tuning between PE and non-PE voxel classes was implemented to improve our PE/non-PE class separation. The *softmax* activation function of the final layer of the U-Net like architecture can be used to scale network output into probabilities. Hence, the segmentation mask can be converted into probabilities via the *softmax* function. Finally, we applied rules based on different *softmax* probability thresholds (0.75 - 0.95) and volumetric thresholding (experimenting with cutoff values from 0 mm^3 to 200 mm^3 in 10 mm^3 intervals) to transform the model segmentation output into a classifier (Supp. materials). Considering these rules, two strategies were developed, where the best balance between sensitivity and specificity is referred to as Strategy 1, and the highest specificity is referred to as Strategy 2 (Supp. materials).

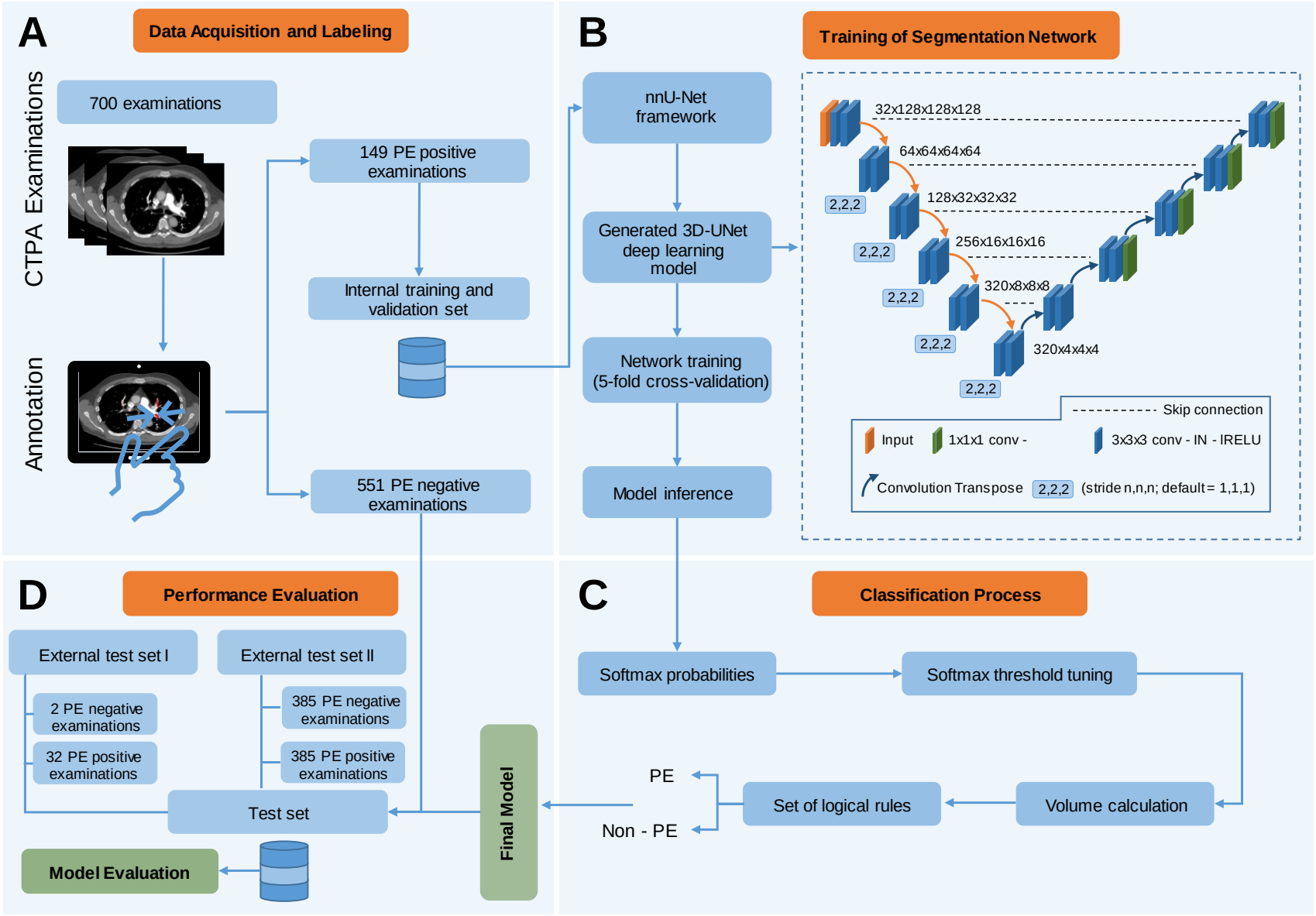
A
Data Acquisition and Labeling
CTPA Examinations
Annotation
700 examinations
149 PE positive examinations
Internal training and validation set
551 PE negative examinations
B
Training of Segmentation Network
nnU-Net framework
Generated 3D-UNet deep learning model
Network training (5-fold cross-validation)
Model inference
32x128x128x128
64x64x64x64
128x32x32x32
256x16x16x16
320x8x8x8
320x4x4x4
2,2,2
Skip connection
Input
1x1x1 conv -
3x3x3 conv - IN - IRELU
Convolution Transpose
2,2,2
(stride n,n,n; default = 1,1,1)
C
Classification Process
Softmax probabilities
Softmax threshold tuning
Volume calculation
Set of logical rules
PE
Non - PE
D
Performance Evaluation
External test set I
External test set II
2 PE negative examinations
32 PE positive examinations
385 PE negative examinations
385 PE positive examinations
Test set
Model Evaluation
Final Model

Figure 2. Training and testing of a segmentation-based classification model for pulmonary embolism detection. A. 700 CTPA examinations were collected and annotated by two radiologists. Of these, all 149 PE positive examinations were used for training and the PE negative were kept for later evaluation. B. The 3D U-Net deep learning model, which is generated by the nnU-Net framework, was trained with the 149 PE-positive CTPAs using 5-fold cross-validation. The convolution layer used a 3×3×3 filter size by default, followed by an instance normalization (IN) layer and a leaky rectified Linear Unit (lRELU) layer. C. The *softmax* probabilities were obtained from the model inference for fine-tuning classification into PE or non-PE voxel classes and for calculating the predicted PE volume. By thresholding the predicted volumes and applying a set of logical rules, accurate patient-level classification for PE was achieved. D. The final model was tested on 804 external CTPA examinations from two publicly available datasets.

Statistical Analysis

Sensitivity and specificity of our trained model for binary classification for PE/non-PE were assessed on a per-patient basis. Matthew's correlation coefficient (MCC) was used to find the optimal balance between sensitivity and specificity. The area under the receiver operating characteristic (AUROC) curve for model training, validation, and testing was used to determine classification performance. Statistical analysis was performed with Microsoft Office Excel (Microsoft Corporation, Washington, USA, Office Professional Plus 2016) and statsmodels package (version 0.13.5) in Python (version 3.8.10; Python Software Foundation). A *p*-value less than .05 was defined as statistically significant and for C.I., the Wilson score interval was used.

Results

Model training and performance evaluation on the internal dataset

For model training, 2,439,000 voxels of 1497 PE were annotated by two radiologists in all 149 PE positive CTPAs of the internal dataset (Table 1). An nnU-net model was trained with 5-fold cross-validation with 119 training and 30 validation CTPAs per set in 4 sets and 120 training and 29 validation CTPAs in the fifth set without overlap between the validation sets. To assess model performance, 21 PE positive exams with small PEs with a total volume of less than 50 mm³ were excluded. The remaining 128 PE positive CTPAs and 551 PE negative CTPAs constituted the internal cross-validation and test set. Training and validation were performed on a single Nvidia RTX 2080 TI GPU card which took ~1 week for all cross-validation folds. The classification performance of the trained nnU-Net model on internal and external test datasets was explored over different threshold volumes and with and without post-processing strategies. With a threshold volume of 20 mm³ and no post-processing, a Matthew's correlation coefficient score (MCC) of 0.64 was obtained with 128 of 128 positive examinations correctly classified as having PE, and 434 of 551 negative examinations correctly classified as non-PE. With the post-processing strategy 1 (Supp. materials) and threshold volume of 20 mm³, the best MCC (0.85) was obtained with 123 of 128 positive examinations correctly classified as PE, and 521 of 551 negative examinations correctly classified as non-PE. Further, the model achieved an AUROC of 0.97 and 0.95 with and without post-processing respectively (Figure 3). The trained nnU-Net model thus achieved a sensitivity of 0.96 (95% C.I. 91-98%, $P < .05$) and 1.0 (95% C.I. 97-100%, $P < .05$), and a specificity of 0.95 (95% C.I. 92-96%, $P < .05$) and 0.79 (95% C.I. 75-82%, $P < .05$) in the internal dataset with and without the post-processing strategies, respectively (Table 2).

Table 1. Ground truth annotation of 149 internal CTPAs with PE.

Component	Blood Clots	Average Volume (mm^3)	Min Volume (mm^3)	Max Volume (mm^3)
3D (Volume)	1497	682	0.21	36510
2D (Area)	36471	16	0.21	830
1D (Voxel)	2439400			

Note. — The total PE volume in all examinations was 744783 mm^3. PE = Pulmonary embolism, 3D = 3-dimensional, 2D = 2-dimensional, 1D = 1-dimensional. Min = minimum, Max = maximum. 3D components comprise 2D components, and 2D components comprise 1D components where a 1D component is equal to 1 voxel.

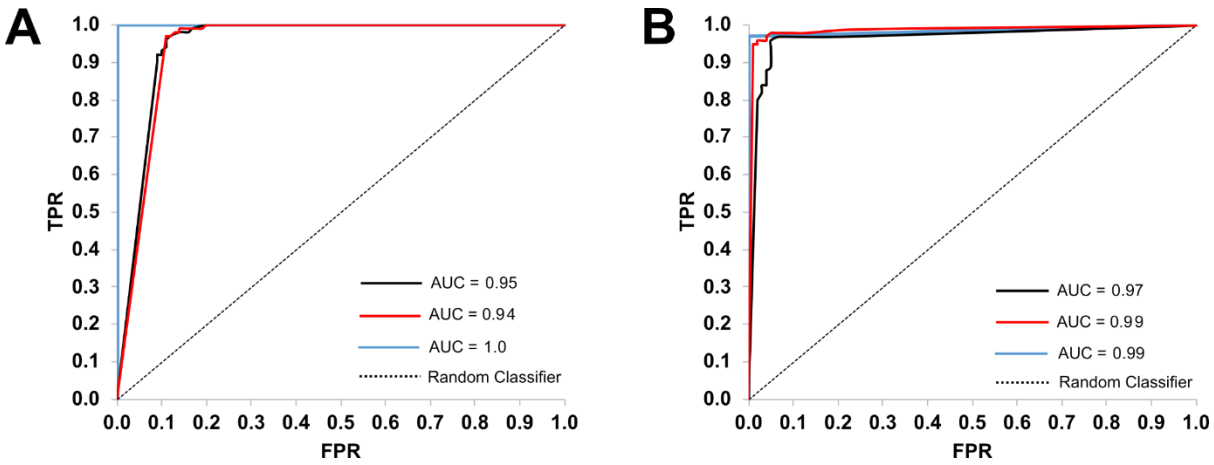

Figure 3 Classification performance of the trained nnU-Net model. Receiver operating characteristic (ROC) curves without (A) and with (B) post-processing. Black, internal dataset (n = 679, 128 PE and 551 non-PE); Blue, the FUMPE datasets (n = 34, 32 PE and 2 non-PE); Red, the RSNA PE dataset (n = 770, 385 PE and 385 non-PE). TPR, true positive rate; FPR, false positive rate

Model performance on external datasets

For external testing, the trained model was applied to a total of 804 CTPAs from two publicly available datasets. First, 34 PE positive CTPAs and 2 PE negative CTPAs from the FUMPE dataset were analyzed. With post-processing strategy 1, an MCC score of 0.80 was obtained with 31 of 32 positive examinations correctly classified as PE, and 2 of 2 negative examinations correctly classified as non-PE. The trained model achieved AUROC 0.99 (Figure 3) with sensitivity 0.97 (95% C.I. 84-99%, $P < .05$) and specificity 1.0 (95% C.I. 34-100%, $P < .05$) (Table 2, Supp. Table 5). Focusing on central PE, where the annotations can be assumed to be more consistent, we used 385 CTPAs annotated as having at least one central PE and 385 PE negative CTPAs from the RSPECT pulmonary embolism CT dataset were used for testing. With the post-processing strategy 1 (Supp. Materials), an MCC of 0.89 was obtained with 379 of 385 positive examinations correctly classified as PE, and 346 of 385 negative examinations correctly classified as non-PE. The trained model achieved an AUROC of 0.99 (Figure 3) with sensitivity of 0.98 (95% C.I. 97-99%, $P < .05$) and a specificity of 0.9 (95% C.I. 86-93%, $P < .05$) (Table 2, Supp. Table 6). Without the post-processing strategy and by setting the threshold volume to 20 mm^3, MCC of 1.0 and 0.73 were obtained with 32 (n=32) and 385 (n=385) positive examinations correctly classified as PE, and 2 (n=2) and 269 (n= 385) negative examinations correctly classified as non-PE in the first and second external datasets, respectively (Table 2, Supp. Table 2 and 3). Moreover, the model achieved an AUROC of 1.0 and 0.94 (Figure 3) with a sensitivity of 1.0 (95% C.I. 89-100%, $P < .05$) and 1.0 (95% C.I. 99-100%, $P < .05$), and a specificity of 1.0 (95% C.I. 34-100%, $P < .05$) and 0.70 (95% C.I. 65-74%, $P < .05$) in the first and second datasets, respectively (Table 2, Supp. Table 2 and 3).

Table 2. Diagnostic performance of the trained model

	Without the Post-Processing			With the Post-Processing Strategy 1		
Parameter	Internal Dataset	FUMPE External Dataset	RSPECT External Dataset	Internal Dataset	FUMPE External Dataset	RSPECT External Dataset
No. of CTPAs	679	34	770	679	34	770
No. of TN	434	2	269	521	2	346
No. of FP	117	0	116	30	0	39
No. of TP	128	32	385	123	31	379
No. of FN	0	0	0	5	1	6
MCC	0.64	1.00	0.73	0.85	0.80	0.89
Sensitivity	1.0 (97-100)	1.0 (89-100)	1.0 (99-100)	0.96 (91-98)	0.97 (84-99)	0.98 (97-99)
Specificity	0.79 (75-82)	1.0 (34-100)	0.70 (65-74)	0.95 (92-96)	1.0 (34-100)	0.90 (86-93)
Accuracy	0.83	1.0	0.85	0.95	0.97	0.94
Balanced Accuracy	0.89	1.0	0.85	0.95	0.98	0.94
AUC	0.95	1.0	0.94	0.97	0.99	0.99

Note. — The threshold volume is set to 20 mm^3. Data in parentheses are 95% CIs in percentages. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient, AUC = area under the receiver operating characteristic curve.

The output of the automated classification pipeline is shown (Figure 4). Running model inference within the nnU-net framework for a single CTPA volume examination took 25-45 minutes on a single Nvidia RTX 2080 TI GPU card.

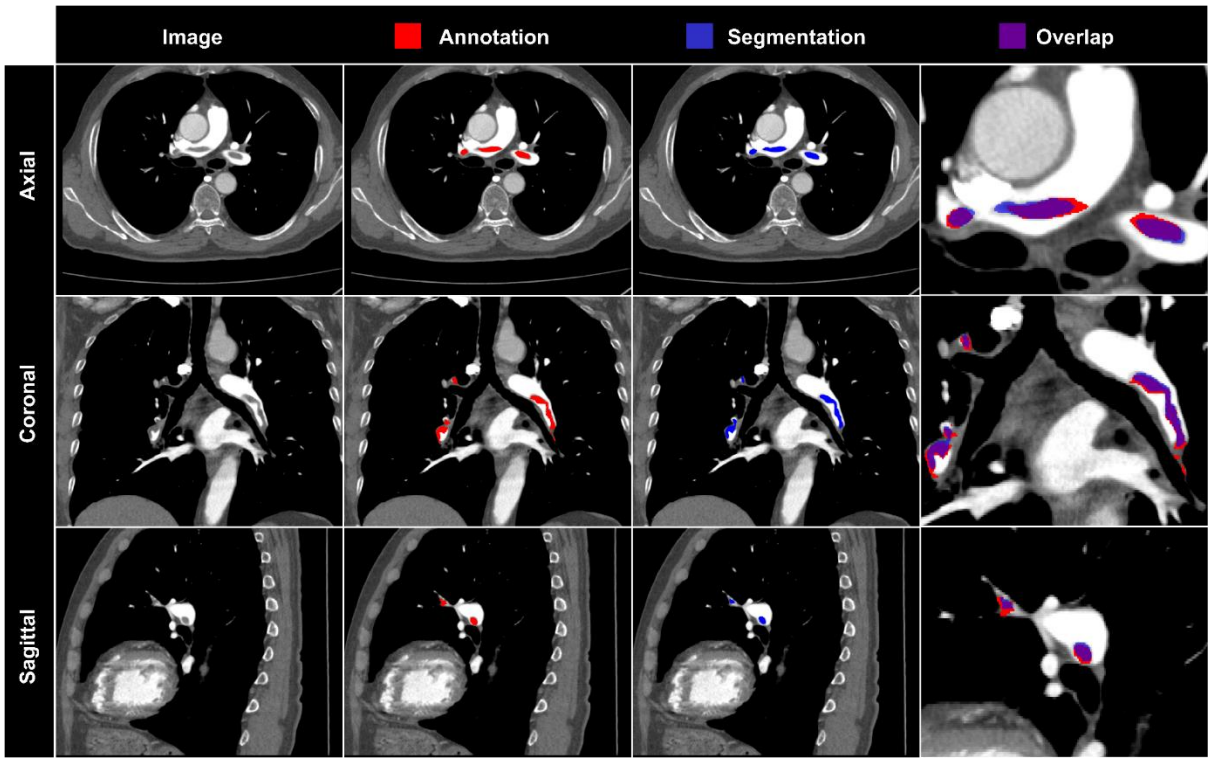

Figure 4. Representative segmentation results of the trained model. Axial, coronal, and sagittal planes from the same CTPA examinations from the external FUMPE dataset with the same window setting (width = 800 HU, level = 100 HU) are shown. Red, pulmonary embolism annotation; blue, model segmentation; purple, overlay of annotation and model segmentation

Benchmarking of model performance

As mentioned above, post-processing strategy 1 was used to find out the best balance between sensitivity and specificity and 20 mm^3 was determined as the optimal threshold volume. Aiming for the highest specificity and the lowest patient level false positive rate, we used post-processing strategy 2 where 50 mm^3 was determined as optimal threshold volume. For size comparison, 20 mm^3, 50 mm^3, and other threshold volumes (Figure 5A) are compared to a segmented reference pulmonary artery (Figure 5B). In the internal test dataset, the highest specificity (0.97; 95% C.I. 95-98%, $P < .05$) was obtained with a sensitivity of 0.88 (95% C.I. 81-92%, $P < .05$) with post-processing strategy 2 (Supp. materials, Supp. Table 7) and threshold volume of 50 mm^3. Without post-processing, the highest specificity (0.91; 95% C.I. 88-93%, $P < .05$) was obtained with a sensitivity of 0.92 (95% C.I. 86-96%, $P < .05$) at a threshold volume of 130 mm^3 (Supp. Table 1). For the external datasets, the highest specificity (1.0; 95% C.I. 34-100%, $P < .05$ and 0.97; 95% C.I. 95-98%, $P < .05$) was obtained with a

sensitivity of 0.91 (95% C.I. 76-97%, $P < .05$) and 0.97 (95% C.I. 94-98%, $P < .05$) in the FUMPE and RSPECT datasets, respectively (Supp. Table 8 and 9). For reference, the voxel-wise ground truth PE from the internal training set and the FUMPE external test set are shown (Figures 5C and D). Moreover, Moreover, we examined the source of FPs by setting the post-processing strategy 2 which resulted in a minimum number of FPs per dataset. The most frequent false positives were due to low contrast medium in pulmonary arteries (Table 3, Figures 5E and F). Whereas 18% of FPs occurred on the outside of the thoracic cavity, the upper abdomen, or the superior vena cava. and the remaining FPs occurred within or close to the pulmonary vessel network.

Table 3. Sources of false positives in PE negative CTPA examinations from internal and external datasets

Source	Internal Dataset (n=18)	RSPECT External Dataset (n=12)
Flow artifact	1	1
Upper abdomen (in left colon)	1	0
Outside the thoracic cavity	3	0
Low contrast medium in PT	6	2
Pulmonary vein	1	2
Superior vena cava	1	0
Intrafissural fluid / atelectasis	0	1
Multiple metastasis	1	1
Tumor	4	2
True pulmonary emboli	0	3

Note. — CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, PE = pulmonary embolism, PT = pulmonary trunk, RSPECT = RSNA Pulmonary Embolism CT Dataset. The cause of false positives in a total of 30 CTPAs is shown, 18 in internal and 12 in external datasets.

We next compared model performance to those of previous studies (Table 4). With post-processing strategy 2, the proposed pipeline achieved a sensitivity of 0.96 (95% C.I. 94-98%, $P < .05$) and a specificity of 0.97 (95% C.I. 95-98%, $P < .05$) on the combined (internal and external) testing set (Supp. Table 10 and 11). While investigating the causes of false positives, we observed that 3 CTPAs from the RSPECT dataset that were annotated as PE negative were actually PE positive. Considering this correction, the proposed pipeline achieved a specificity of 97.1%.

Table 4. Model performance comparison for patient-level classification for PE in CTPA examinations.

Author	Year	Method	Classification level	PE location	AUC (%)	Sensitivity (%)	Specificity (%)	Testing size	
								PE positive CTPAs	PE negative CTPAs
PIOPED II (27)	2006	Radiologists	patient-level	M, L, S, s	N/A	83	96	181	592
Maizlin et al (28)	2007	IPAT	patient-level	M, L, S, s	N/A	53.3	77.5	15	89
Wittenberg et al (9)	2010	IPAT	patient-level	M, L, S, s	N/A	94	21	68	210
Wittenberg et al (10)	2012	IPAT	patient-level	M, L, S, s	N/A	96	22	51	158
Lahiji et al (29)	2014	IPAT	patient-level	L, S, s	N/A	97.5	26.9	40	26
Rajan et al (17)	2020	2D U-Net + LSTM	patient-level	M, L	85	N/A	N/A	385	127
Rajan et al (17)	2020	2D U-Net + LSTM	patient-level	S, s	70	N/A	N/A	385	127
Weikert et al (18)	2020	DCNN	patient-level	M, L, S, s	N/A	92.7	95.5	232	1233
Weikert et al (18)	2020	DCNN	patient-level	M, L, (S, s) *	N/A	95.7	95.5	232	1233
Weikert et al (18)	2020	DCNN	patient-level	S, (s)*	N/A	93.3	95.5	232	1233
Weikert et al (18)	2020	DCNN	patient-level	s	N/A	85.7	95.5	232	1233
Huang et al (26)	2020	3D CNN	patient-level	M, L, S	85	75	81	94	106
Huhtanen et al (30)	2022	CNN	patient-level	M, L, S, s	N/A	86.6	93.5	97	107
Proposed pipeline	**2023**	**nnU-Net + DPPS1**	patient-level	M, L, S, s	**98.2**	**98.3**	**92.6**	**417**	**938**
Proposed pipeline	**2023**	**nnU-Net + DPPS2**	patient-level	M, L, S, s	**98.2**	**96.2**	**96.8**	**417**	**938**

Note. — * can possibly have pulmonary emboli in these segments , PE = pulmonary embolism, CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, IPAT = image processing and analysis techniques, N/A = not available, M = left, right and main pulmonary arteries-level PE, L = lobar level PE, S = segmental level PE, s = sub-segmental level PE, LSTM = long short-term memory, CNN = convolutional neural network, DCNN = deep CNN, PIOPED = Prospective Investigation of Pulmonary Embolism Diagnosis II, DPPS= deterministic post-processing strategy.

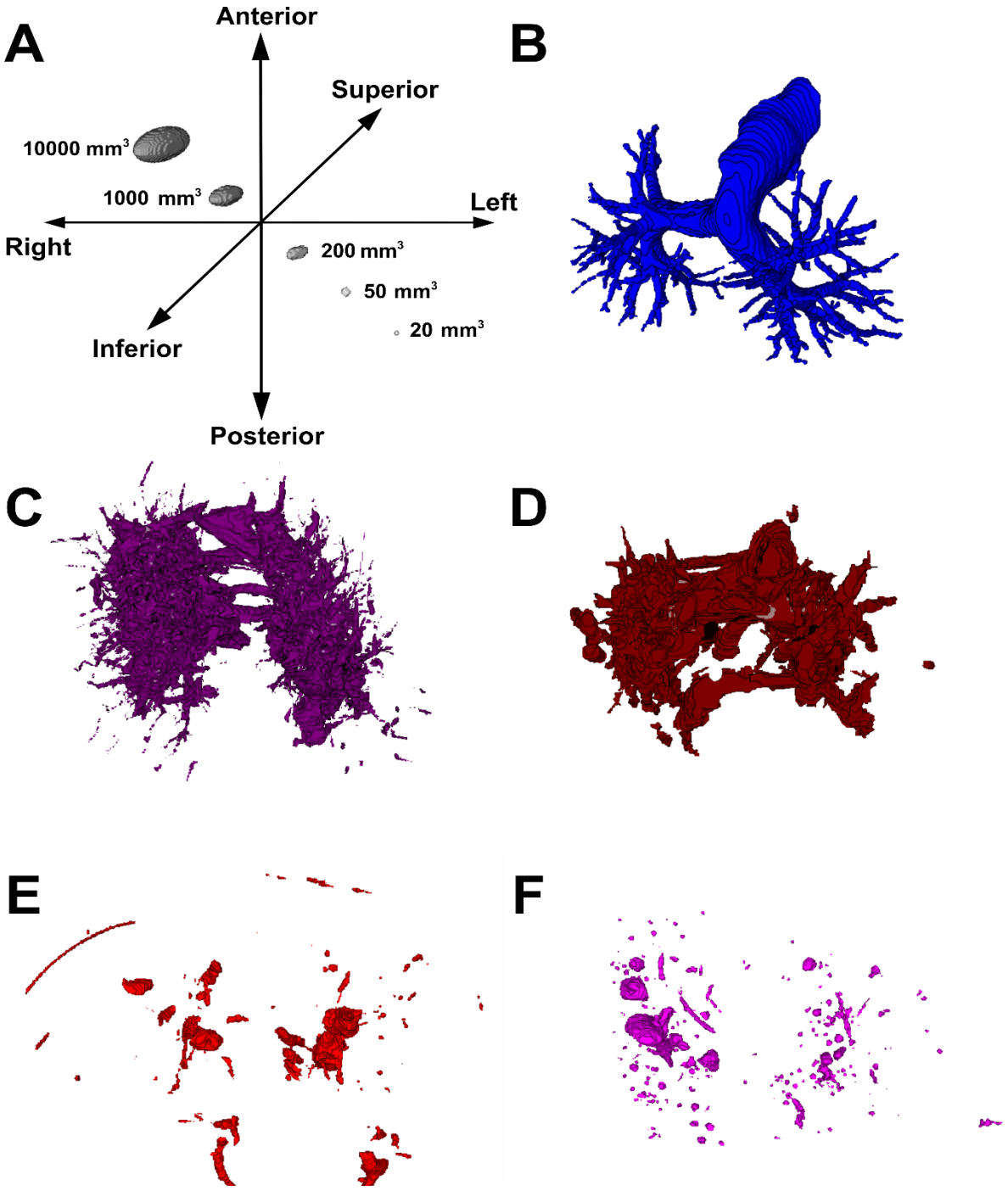

Figure 5. Ground truth and false positive pulmonary emboli. The PEs from ground truth or false positive AI annotations of the indicated cohort were projected in a single volume. A. Patient orientation of 3D volumes with reference threshold volumes (20, 50, 200, 1000, and 10000 mm³). B. Manually segmented reference pulmonary artery (volume of 113 cm³) from a male patient without PE. C. Ground truth PE from the 149 positive examinations of the internal training set shown in a single volume (total PE volume = 745 cm³). D. Ground truth PE from 32 positive exams of the FUMPE external test set shown in a single volume (total PE volume = 466 cm³). E. False positive PE from the internal test set (18 PE negative examinations with total FP volume = 59 cm³). F. False positive PE from the RSNA external test set (12 PE negative examinations with total FP volume of 28 cm³). False positives were obtained from PE negative examinations by aiming for the highest specificity using post-processing strategy 2. All volumetric images are isotropic (1 mm × 1 mm × 1 mm).

Discussion

Automatic detection of PE has potential to assist the radiologist in the time-consuming reading of CTPA examinations. Preferentially, such a system could highlight PE positive examinations in the work list, helping radiologists identify high-priority cases for rapid review (10). Detection of PE in CTPA using DCNN was first demonstrated by Tajbakhsh et al. with a sensitivity of 83% and 34.6% at 2 FPs/examinations on 121 internal and 20 external test examinations, respectively (25). Rajan et al. proposed a two-stage solution where a 2D U-Net model was used for PE candidate generation, followed by a convolutional long short-term memory (LSTM) network coupled with multiple instance learning to detect PE lesions, with AUROC of 0.70 for subsegmental and segmental PE and 0.85 for saddle and main pulmonary artery PE on a test dataset of 512 CTPA examinations (17). In a study by Huang et al. (26), a 3D CNN model termed PENet was developed which achieved a sensitivity of 75% and specificity of 81% on an external test dataset of 200 CTPA examinations. However, all these studies have major limitations such as small testing dataset sizes or low specificity rates. The current state-of-art results were recently achieved using the Resnet architecture on 1465 CTPA examinations with a sensitivity of 92.7% and specificity of 95.5% at the patient level (18). Taken together, the performance of AI systems for PE detection is now at a point where clinical utility can be expected, but further gains in sensitivity and specificity are still warranted.

Here, we developed a pipeline that classifies CTPA examinations for PE consisting of two main stages, PE candidate selection and post-processing. For PE candidate selection, we trained and validated a semantic segmentation model, nnU-Net, on our internal dataset. The nnU-Net is a medical image segmentation framework based on the U-Net architecture and has outperformed state-of-the-art models by competing in 53 segmentation tasks from 11 international biomedical image segmentation challenges and taking first place in 33 of them (20). To our knowledge, this is the first use of nnU-Net for classification for PE. To transform the segmentation model into a classification model, we developed rules based on probability and minimum volume thresholds as a post-processing stage. We defined two post-processing strategies, one for the best trade-off between sensitivity and specificity and one for achieving the highest specificity. At the best trade-off between sensitivity and specificity, the patient-level classification performance of the trained model achieved a sensitivity of 98.3% and specificity of 92.6% on the combined testing dataset using a threshold volume of 20 mm^3, compared to specificity of 75.2% with sensitivity of 100% without post-processing. Thus, by sacrificing 1.7% of sensitivity, the model gained 17.4% in specificity using post-processing. The model outperformed the current state-of-art using the strategy of highest specificity, achieving 96.2% sensitivity and a specificity of 96.8% on the combined testing dataset of 1355 CTPA examinations with a threshold total emboli volume of 50 mm^3.

Although the nnU-Net based model presented here is superior to the state-of-art, there are some limitations and opportunities for future enhancement. First, the model was trained on data from a single institution, although derived from five different CT scanners. Second, the RSPECT validation dataset lacks voxel level annotation of PE by radiologists, which precludes final determination of sensitivity and specificity until a review has been completed. Finally, by disabling the test data augmentation the model inference can take 3 to 5 minutes in a CTPA examination. However, disabling the test data augmentation induces an accuracy loss of close to %5. To avoid loss of accuracy, the model inference should run with the test data augmentation which took between 30-45 minutes in a CTPA examination. Therefore, the model inference time needs to be accelerated for future clinical applications. Taken together, we have obtained promising results with the nnU-Net deep learning architecture for binary classification for PE/non-PE.

References

1. Sista AK, Kuo WT, Schiebler M, Madoff DC. Stratification, Imaging, and Management of Acute Massive and Submassive Pulmonary Embolism. Radiology. 2017;284:5–24. doi: 10.1148/radiol.2017151978.

2. Raskob GE, Angchaisuksiri P, Blanco AN, Buller H, Gallus A, Hunt BJ, Hylek EM, Kakkar A, Konstantinides SV, McCumber M, et al. Thrombosis: A Major Contributor to Global Disease Burden. ATVB. 2014;34:2363–2371. doi: 10.1161/ATVBAHA.114.304488.

3. Elenizi K, Alharthi R, Galinier M. Pulmonary embolism originating from germ cell tumor causes severe left ventricular dysfunction in a healthy young adult with full recovery: a case report. BMC Cardiovasc Disord. 2021;21:260. doi: 10.1186/s12872-021-02066-7.

4. Sista AK, Horowitz JM, Tapson VF, Rosenberg M, Elder MD, Schiro BJ, Dohad S, Amoroso NE, Dexter DJ, Loh CT, et al. Indigo Aspiration System for Treatment of Pulmonary Embolism. JACC: Cardiovascular Interventions. 2021;14:319–329. doi: 10.1016/j.jcin.2020.09.053.

5. Jha AK, Larizgoitia I, Audera-Lopez C, Prasopa-Plaizier N, Waters H, Bates DW. The global burden of unsafe medical care: analytic modelling of observational studies. BMJ Qual Saf. 2013;22:809–815. doi: 10.1136/bmjqs-2012-001748.

6. Rivera-Lebron B, McDaniel M, Ahrar K, Alrifai A, Dudzinski DM, Fanola C, Blais D, Janicke D, Melamed R, Mohrien K, et al. Diagnosis, Treatment and Follow Up of Acute Pulmonary Embolism: Consensus Practice from the PERT Consortium. Clin Appl Thromb Hemost. 2019;25:107602961985303. doi: 10.1177/1076029619853037.

7. Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing G-J, Harjola V-P, Huisman MV, Humbert M, Jennings CS, Jiménez D, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS): The Task Force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC). European Heart Journal. 2020;41:543–603. doi: 10.1093/eurheartj/ehz405.

8. Hendriks BMF, Eijsvoogel NG, Kok M, Martens B, Wildberger JE, Das M. Optimizing Pulmonary Embolism Computed Tomography in the Age of Individualized Medicine: A Prospective Clinical Study. Invest Radiol. 2018;53:306–312. doi: 10.1097/RLI.0000000000000443.

9. Wittenberg R, Peters JF, Sonnemans JJ, Prokop M, Schaefer-Prokop CM. Computer-assisted detection of pulmonary embolism: evaluation of pulmonary CT angiograms performed in an on-call setting. Eur Radiol. 2010;20:801–806. doi: 10.1007/s00330-009-1628-7.

10. Wittenberg R, Berger FH, Peters JF, Weber M, van Hoorn F, Beenen LFM, van Doorn MMAC, van Schuppen J, Zijlstra IjAJ, Prokop M, et al. Acute Pulmonary Embolism: Effect of a Computer-assisted Detection Prototype on Diagnosis—An Observer Study. Radiology. 2012;262:305–313. doi: 10.1148/radiol.11110372.

11. Zhou C, Chan H-P, Hadjiiski LM, Chughtai A, Patel S, Cascade PN, Sahiner B, Wei J, Ge J, Kazerooni EA. Automated detection of pulmonary embolism (PE) in computed tomographic pulmonary angiographic (CTPA) images: multiscale hierachical expectation-maximization segmentation of vessels and PEs. In: Giger ML, Karssemeijer N, editors. San Diego, CA; 2007 [cited 2023 Feb 16]. p. 65142F. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.713769.

12. Özkan H, Osman O, Şahin S, Boz AF. A novel method for pulmonary embolism detection in CTA images. Computer Methods and Programs in Biomedicine. 2014;113:757–766. doi: 10.1016/j.cmpb.2013.12.014.

13. Myers MH, Beliaev I, Lin K-I. Machine Learning Techniques in Detecting of Pulmonary Embolisms. 2007 International Joint Conference on Neural Networks [Internet]. Orlando, FL, USA: IEEE; 2007 [cited 2023 Feb 16]. p. 385–390. Available from: http://ieeexplore.ieee.org/document/4370987/.

14. Wang X, Song X, Chapman BE, Zheng B. Improving performance of computer-aided detection of pulmonary embolisms by incorporating a new pulmonary vascular-tree segmentation algorithm. In: van Ginneken B, Novak CL, editors. San Diego, California, USA; 2012 [cited 2023 Feb 16]. p. 83152U. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.911301.

15. Ozkan H, Tulum G, Osman O, Sahin S. Automatic Detection of Pulmonary Embolism in CTA Images Using Machine Learning. ElAEE. 2017;23:63–67. doi: 10.5755/j01.eie.23.1.17585.

16. Tajbakhsh N, Shin JY, Gotway MB, Liang J. Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and discriminative image representation. Medical Image Analysis. 2019;58:101541. doi: 10.1016/j.media.2019.101541.

17. Rajan D, Beymer D, Abedin S, Dehghan E. Pi-PE: A Pipeline for Pulmonary Embolism Detection using Sparsely Annotated 3D CT Images. Proceedings of the Machine Learning for Health NeurIPS Workshop [Internet]. PMLR; 2020 [cited 2023 Feb 16]. p. 220–232. Available from: https://proceedings.mlr.press/v116/rajan20a.html.

18. Weikert T, Winkel DJ, Bremerich J, Stieltjes B, Parmar V, Sauter AW, Sommer G. Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. Eur Radiol. 2020;30:6545–6553. doi: 10.1007/s00330-020-06998-0.

19. 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;9:82031–82057. doi: 10.1109/ACCESS.2021.3086020.

20. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–211. doi: 10.1038/s41592-020-01008-z.

21. Kahraman AT, Fröding T, Dimitrios T, Natasa S, Sjöblom T. Automated Detection, Segmentation and Measurement of Major Vessels and the Trachea in CT Pulmonary Angiography. Manuscript submitted for publication; 2022.

22. Wolf I, Vetter M, Wegner I, Böttger T, Nolden M, Schöbinger M, Hastenteufel M, Kunert T, Meinzer H-P. The Medical Imaging Interaction Toolkit. Medical Image Analysis. 2005;9:594–604. doi: 10.1016/j.media.2005.04.005.

23. Masoudi M, Pourreza H-R, Saadatmand-Tarzjan M, Eftekhari N, Zargar FS, Rad MP. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. Sci Data. 2018;5:180180. doi: 10.1038/sdata.2018.180.

24. Colak E, Kitamura FC, Hobbs SB, Wu CC, Lungren MP, Prevedello LM, Kalpathy-Cramer J, Ball RL, Shih G, Stein A, et al. The RSNA Pulmonary Embolism CT Dataset. Radiology: Artificial Intelligence. 2021;3:e200254. doi: 10.1148/ryai.2021200254.

25. Tajbakhsh N, Gotway MB, Liang J. Computer-Aided Pulmonary Embolism Detection Using a Novel Vessel-Aligned Multi-planar Image Representation and Convolutional Neural Networks. In: Navab N, Hornegger J, Wells WM, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015 [Internet]. Cham: Springer International Publishing; 2015 [cited 2023 Feb 16]. p. 62–69. Available from: http://link.springer.com/10.1007/978-3-319-24571-3_8.

26. Huang S-C, Kothari T, Banerjee I, Chute C, Ball RL, Borus N, Huang A, Patel BN, Rajpurkar P, Irvin J, et al. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. npj Digit Med. 2020;3:61. doi: 10.1038/s41746-020-0266-y.

27. Stein PD, Fowler SE, Goodman LR, Gottschalk A, Hales CA, Hull RD, Leeper KV, Popovich J, Quinn DA, Sos TA, et al. Multidetector Computed Tomography for Acute Pulmonary Embolism. N Engl J Med. 2006;354:2317–2327. doi: 10.1056/NEJMoa052367.

28. Maizlin ZV, Vos PM, Godoy MB, Cooperberg PL. Computer-aided Detection of Pulmonary Embolism on CT Angiography: Initial Experience. Journal of Thoracic Imaging. 2007;22:324–329. doi: 10.1097/RTI.0b013e31815b89ca.

29. Lahiji K, Kligerman S, Jeudy J, White C. Improved Accuracy of Pulmonary Embolism Computer-Aided Detection Using Iterative Reconstruction Compared With Filtered Back Projection. American Journal of Roentgenology. 2014;203:763–771. doi: 10.2214/AJR.13.11838.

30. Huhtanen H, Nyman M, Mohsen T, Virkki A, Karlsson A, Hirvonen J. Automated detection of pulmonary embolism from CT-angiograms using deep learning. BMC Med Imaging. 2022;22:43. doi: 10.1186/s12880-022-00763-z.

Data sharing statement: Data generated or analyzed during the study are available from the corresponding author by request.

Funding information
The project was supported by a grant from Analytic Imaging Diagnostic Arena (AIDA), https://medtech4health.se/aida-en/, to Tobias Sjöblom. Tomas Fröding and Dimitrios Toumpanakis were supported by clinical fellowships from AIDA. Tomas Fröding was supported by the Centre for Clinical Research Sörmland, Uppsala University, Eskilstuna, Sweden.

Acknowledgements. The project was supported by a grant from Analytic Imaging Diagnostic Arena (AIDA) to Tobias Sjöblom. Tomas Fröding and Dimitrios Toumpanakis were financially supported by clinical fellowships from AIDA. Tomas Fröding was financially supported by the Centre for Clinical Research Sörmland, Uppsala University.

Supplementary Materials

Post-processing step

The nn-Unet *softmax* activation function of the final layer of the U-Net architecture can be used to scale network output into probabilities. Hence, the probabilities could be gathered, and not only final pixel class values. We developed a set of logical rules based on different *softmax* probability thresholds (0.75 - 0.95) and threshold volumes (0 mm^3 to 200 mm^3 in 10 mm^3 intervals) to reduce false positives (FPs) and convert nnU-Net inference segmentation output into a patient-level classification output. By setting different *softmax* probability thresholds, we obtained different predicted PE volumes. If the model is well-trained to distinguish between PE and non-PE classes, the number of predicted voxels (false positive voxels) that do not belong to the PE class will decrease when the *softmax* probabilities are set to higher thresholds. Therefore, we developed the formulas below to decide whether the total predicted PE volume is sufficient to designate the patient as PE positive/negative.

Proposition 1:

$$R = \begin{cases} if \ \dfrac{(P_{0.75} - P_{0.90})}{P_{0.90}} > r \ , & Non-PE \\ otherwise \ , & PE \end{cases}$$

where $P_{0.75}$ is the volume of total PE predicted by the trained model at a softmax probability of 0.75, $P_{0.90}$ is the volume of total PE predicted by the trained model at a *softmax* probability of 0.90, and r is the ratio factor, which was fixed at 15. The *softmax* probability value range and the ratio factor were optimized by systematic exploration.

Proposition 2:

$$Q = \begin{cases} \left(\sum_{i=0.75}^{0.95} \begin{pmatrix} 1, & if \ (P_i < v) \\ 0, & otherwise \end{pmatrix} \right) \geq k \ , & Non-PE \\ otherwise \ , & PE \end{cases}$$

where P_i is the volume of total PE predicted by the trained model at a *softmax* probability of *i* between 0.75 to 0.95 with 0.05 intervals, *v* is the threshold volume between 0 and 200 mm^3 at 10 mm^3 intervals and k is the condition factor (min value is 0, max value is 4) that refers to the total number of true conditions satisfying $P_i < v$ equation.

Then, the final decision is made as follows:

$$R \vee Q = \begin{cases} Patient\ without\ PE, & True \\ Patient\ with\ PE, & False \end{cases}$$

According to the propositions above, we defined two post-processing strategies. Strategy 1 (Rule-in classification for PE) aimed to find the exact threshold volume value and k value for the best trade-off between sensitivity and specificity by checking the Matthew's correlation coefficient (MCC) value. And strategy 2 (Rule-out classification for PE) aimed to find the exact threshold volume and k values for the highest specificity alongside the highest MCC value.

Strategy 1:

By systematic exploration, setting the threshold volume value to 20 mm^3 and the k value to 1 gives the highest MCC value (0.85, Supplementary Table 4).

Strategy 2

By systematic exploration, setting the threshold volume value to 50 mm^3 and the k value to 0 gives the highest specificity alongside the highest MCC value (0.84, Supplementary Table 7).

Supplementary Table 1. Diagnostic performance of the trained model without post-processing in the internal dataset

	Internal Dataset (CTPAs = 679)																				
	Without Post-Processing																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	213	389	434	459	465	476	486	488	490	492	493	496	497	499	501	501	502	502	502	503	503
No. of FP	338	162	117	92	86	75	65	63	61	59	58	55	54	52	50	50	49	49	49	48	48
No. of TP	128	128	128	127	126	126	124	123	123	121	120	119	118	118	118	117	117	116	116	115	115
No. of FN	0	0	0	1	2	2	4	5	5	7	8	9	10	10	10	11	11	12	12	13	13
MCC	0.33	0.56	0.64	0.69	0.70	0.73	0.74	0.74	0.75	0.74	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75	0.75	0.74	0.74
Sensitivity	1.00	1.00	1.00	0.99	0.98	0.98	0.97	0.96	0.96	0.95	0.94	0.93	0.92	0.92	0.92	0.91	0.91	0.91	0.91	0.90	0.90
Specificity	0.39	0.71	0.79	0.83	0.84	0.86	0.88	0.89	0.89	0.89	0.89	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
Accuracy	0.50	0.76	0.83	0.86	0.87	0.89	0.90	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
Balanced Accuracy	0.69	0.85	0.89	0.91	0.91	0.92	0.93	0.92	0.93	0.92	0.92	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient.

Supplementary Table 2. Diagnostic performance of the trained model without post-processing in the external FUMPE dataset

	FUMPE External Dataset (CTPAs = 34)																				
	Without Post-Processing																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
No. of FP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
No. of TP	32	32	32	32	32	32	32	32	32	32	31	31	30	30	29	29	29	29	29	29	29
No. of FN	0	0	0	0	0	0	0	0	0	0	1	1	2	2	3	3	3	3	3	3	3
MCC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.80	0.68	0.68	0.60	0.60	0.60	0.60	0.60	0.60	0.60
Sensitivity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.97	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91	0.91
Specificity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Accuracy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.97	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91	0.91
Balanced Accuracy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.97	0.97	0.95	0.95	0.95	0.95	0.95	0.95	0.95

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew’s correlation coefficient. FUMPE = Ferdowsi University of Mashhad's PE dataset.

Supplementary Table 3. Diagnostic performance of the trained model without post-processing in the external RSPECT Dataset

	RSPECT External Dataset (CTPAs = 770)																				
	Without Post-Processing																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	131	249	269	287	297	303	308	312	314	318	321	326	328	330	333	333	335	337	338	341	343
No. of FP	254	136	116	98	88	82	77	73	71	67	64	59	57	55	52	52	50	48	47	44	42
No. of TP	385	385	385	385	385	385	384	383	382	382	382	382	382	380	376	376	376	375	374	374	374
No. of FN	0	0	0	0	0	0	1	2	3	3	3	3	3	5	9	9	9	10	11	11	11
MCC	0.45	0.69	0.73	0.77	0.79	0.81	0.81	0.82	0.82	0.83	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.86	0.87
Sensitivity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.97	0.97	0.97	0.97
Specificity	0.34	0.65	0.70	0.75	0.77	0.79	0.80	0.81	0.82	0.83	0.83	0.85	0.85	0.86	0.86	0.86	0.87	0.88	0.88	0.89	0.89
Accuracy	0.67	0.82	0.85	0.87	0.89	0.89	0.90	0.90	0.90	0.91	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.93	0.93
Balanced Accuracy	0.67	0.82	0.85	0.87	0.89	0.89	0.90	0.90	0.90	0.91	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.93	0.93

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient. RSPECT = RSNA Pulmonary Embolism CT Dataset

Supplementary Table 4. Diagnostic performance of the trained model with post-processing strategy 1 in the internal dataset

	Internal Dataset (CTPAs = 679)																				
	With Post-Processing Strategy 1																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	439	511	521	523	526	529	529	530	530	530	531	533	534	535	536	536	537	538	538	538	538
No. of FP	112	40	30	28	25	22	22	21	21	21	20	18	17	16	15	15	14	13	13	13	13
No. of TP	124	124	123	120	114	113	113	111	109	108	108	108	106	106	106	105	105	102	102	102	101
No. of FN	4	4	5	8	14	15	15	17	19	20	20	20	22	22	22	23	23	26	26	26	27
MCC	0.63	0.82	0.85	0.84	0.82	0.83	0.83	0.82	0.81	0.80	0.81	0.82	0.81	0.81	0.82	0.81	0.82	0.81	0.81	0.81	0.80
Sensitivity	0.97	0.97	0.96	0.94	0.89	0.88	0.88	0.87	0.85	0.84	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.80	0.80	0.80	0.79
Specificity	0.80	0.93	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
Accuracy	0.83	0.94	0.95	0.95	0.94	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.95	0.94	0.94	0.94	0.94
Balanced Accuracy	0.88	0.95	0.95	0.94	0.92	0.92	0.92	0.91	0.91	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.89	0.89	0.89	0.88

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient.

Supplementary Table 5. Diagnostic performance of the trained model with post-processing strategy 1 in the external FUMPE Dataset

	FUMPE External Dataset (CTPAs = 34)																				
	With Post-Processing Strategy 1																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
No. of FP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
No. of TP	31	31	31	31	30	29	29	29	29	28	28	28	28	28	28	28	28	28	28	27	27
No. of FN	1	1	1	1	2	3	3	3	3	4	4	4	4	4	4	4	4	4	4	5	5
MCC	0.80	0.80	0.80	0.80	0.68	0.60	0.60	0.60	0.60	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.49	0.49
Sensitivity	0.97	0.97	0.97	0.97	0.94	0.91	0.91	0.91	0.91	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.84	0.84
Specificity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Accuracy	0.97	0.97	0.97	0.97	0.94	0.91	0.91	0.91	0.91	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.85	0.85
Balanced Accuracy	0.98	0.98	0.98	0.98	0.97	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.92	0.92

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew’s correlation coefficient. FUMPE = Ferdowsi University of Mashhad's PE dataset.

Supplementary Table 6. Diagnostic performance of the trained model with post-processing strategy 1 in the external RSPECT Dataset

	RSPECT External Dataset (CTPAs = 770)																				
	With Post-Processing Strategy 1																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	294	336	346	356	361	364	368	370	370	371	372	374	376	377	377	377	377	378	380	380	380
No. of FP	91	49	39	29	24	21	17	15	15	14	13	11	9	8	8	8	8	7	5	5	5
No. of TP	382	379	379	379	377	376	373	372	372	371	369	369	368	367	367	366	365	364	364	364	363
No. of FN	3	6	6	6	8	9	12	13	13	14	16	16	17	18	18	19	20	21	21	21	22
MCC	0.78	0.86	0.89	0.91	0.92	0.92	0.92	0.93	0.93	0.93	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Sensitivity	0.99	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
Specificity	0.76	0.87	0.90	0.92	0.94	0.95	0.96	0.96	0.96	0.96	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99
Accuracy	0.88	0.93	0.94	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.96
Balanced Accuracy	0.88	0.93	0.94	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.96

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient. RSPECT = RSNA Pulmonary Embolism CT Dataset

Supplementary Table 7. Diagnostic performance of the trained model with post-processing strategy 2 in the internal Dataset

	Internal Dataset (CTPAs = 679)																				
	With Post-Processing Strategy 2																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	466	519	524	529	531	533	534	536	536	537	538	538	538	540	540	540	540	541	541	541	541
No. of FP	85	32	27	22	20	18	17	15	15	14	13	13	13	11	11	11	11	10	10	10	10
No. of TP	124	120	116	114	113	112	109	108	108	106	106	106	106	105	103	102	102	101	100	99	99
No. of FN	4	8	12	14	15	16	19	20	20	22	22	22	22	23	25	26	26	27	28	29	29
MCC	0.69	0.83	0.82	0.83	0.83	0.84	0.83	0.83	0.83	0.82	0.83	0.83	0.83	0.83	0.82	0.82	0.82	0.82	0.81	0.80	0.80
Sensitivity	0.97	0.94	0.91	0.89	0.88	0.88	0.85	0.84	0.84	0.83	0.83	0.83	0.83	0.82	0.80	0.80	0.80	0.79	0.78	0.77	0.77
Specificity	0.85	0.94	0.95	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Accuracy	0.87	0.94	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94
Balanced Accuracy	0.91	0.94	0.93	0.93	0.92	0.92	0.91	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.89	0.89	0.89	0.89	0.88	0.88	0.88

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient.

Supplementary Table 8. Diagnostic performance of the trained model with post-processing strategy 2 in the eternal FUMPE Dataset

	FUMPE External Dataset (CTPAs = 34)																				
	With Post-Processing Strategy 2																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
No. of FP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
No. of TP	31	31	31	31	29	29	29	28	28	28	28	28	28	28	28	27	27	27	27	27	27
No. of FN	1	1	1	1	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5
MCC	0.80	0.80	0.80	0.80	0.60	0.60	0.60	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.49	0.49	0.49	0.49	0.49	0.49
Sensitivity	0.97	0.97	0.97	0.97	0.91	0.91	0.91	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.84	0.84	0.84	0.84	0.84	0.84
Specificity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Accuracy	0.97	0.97	0.97	0.97	0.91	0.91	0.91	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.85	0.85	0.85	0.85	0.85	0.85
Balanced Accuracy	0.98	0.98	0.98	0.98	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.92	0.92	0.92	0.92	0.92	0.92

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew’s correlation coefficient. FUMPE = Ferdowsi University of Mashhad's PE dataset.

Supplementary Table 9. Diagnostic performance of the trained model with post-processing strategy 2 in the external RSPECT Dataset

	RSPECT External Dataset (CTPAs = 770)																				
	With Post-Processing Strategy 2																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	310	349	358	366	371	373	373	375	376	377	379	379	379	380	381	382	382	382	382	382	382
No. of FP	75	36	27	19	14	12	12	10	9	8	6	6	6	5	4	3	3	3	3	3	3
No. of TP	380	378	374	374	372	372	370	369	367	366	366	365	364	362	362	362	362	362	362	361	360
No. of FN	5	7	11	11	13	13	15	16	18	19	19	20	21	23	23	23	23	23	23	24	25
MCC	0.81	0.89	0.90	0.92	0.93	0.94	0.93	0.93	0.93	0.93	0.94	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Sensitivity	0.99	0.98	0.97	0.97	0.97	0.97	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
Specificity	0.81	0.91	0.93	0.95	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Accuracy	0.90	0.94	0.95	0.96	0.96	0.97	0.96	0.97	0.96	0.96	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.96	0.96
Balanced Accuracy	0.90	0.94	0.95	0.96	0.96	0.97	0.96	0.97	0.96	0.96	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.96	0.96

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient. RSPECT = RSNA Pulmonary Embolism CT Dataset

Supplementary Table 10. Diagnostic performance of the trained model with post-processing strategy 1 in the combined testing dataset

	Testing Dataset (CTPAs = 1355)																				
	With Post-Processing Strategy 1																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	735	849	869	881	889	895	899	902	902	903	905	909	912	914	915	915	916	918	920	920	920
No. of FP	203	89	69	57	49	43	39	36	36	35	33	29	26	24	23	23	22	20	18	18	18
No. of TP	413	410	410	410	407	405	402	401	401	399	397	397	396	395	395	394	393	392	392	391	390
No. of FN	4	7	7	7	10	12	15	16	16	18	20	20	21	22	22	23	24	25	25	26	27
MCC (%)	71.7	85.0	87.8	89.6	90.2	90.8	90.9	91.2	91.2	91.0	90.9	91.6	91.9	92.0	92.2	92.0	92.0	92.2	92.5	92.3	92.2
Sensitivity (%)	99.0	98.3	98.3	98.3	97.6	97.1	96.4	96.2	96.2	95.7	95.2	95.2	95.0	94.7	94.7	94.5	94.2	94.0	94.0	93.8	93.5
Specificity (%)	78.4	90.5	92.6	93.9	94.8	95.4	95.8	96.2	96.2	96.3	96.5	96.9	97.2	97.4	97.5	97.5	97.7	97.9	98.1	98.1	98.1
Accuracy (%)	84.7	92.9	94.4	95.3	95.6	95.9	96.0	96.2	96.2	96.1	96.1	96.4	96.5	96.6	96.7	96.6	96.6	96.7	96.8	96.8	96.7
Balanced Accuracy (%)	88.7	94.4	95.5	96.1	96.2	96.3	96.1	96.2	96.2	96.0	95.8	96.1	96.1	96.1	96.1	96.0	95.9	95.9	96.0	95.9	95.8

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient. RSPECT = RSNA Pulmonary Embolism CT Dataset, FUMPE = Ferdowsi University of Mashhad's PE dataset, Testing Dataset = 551 PE negative CTPAs from internal testing set + 32 PE positive and 2 PE negative from FUMPE + 385 PE positive and 385 PE negative CTPAs from RSPECT

Supplementary Table 11. Diagnostic performance of the trained model with post-processing strategy 2 in the combined testing dataset

	Testing Dataset (CTPAs = 1355)																				
	With Post-Processing Strategy 2																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	778	870	884	897	904	908	909	913	914	916	919	919	919	922	923	924	924	925	925	925	925
No. of FP	160	68	54	41	34	30	29	25	24	22	19	19	19	16	15	14	14	13	13	13	13
No. of TP	411	409	405	405	401	401	399	397	395	394	394	393	392	390	390	389	389	389	389	388	387
No. of FN	6	8	12	12	16	16	18	20	22	23	23	24	25	27	27	28	28	28	28	29	30
MCC (%)	76.2	87.8	89.1	91.1	91.5	92.1	91.9	92.2	92.0	92.2	92.7	92.5	92.4	92.5	92.7	92.7	92.7	92.9	92.9	92.7	92.5
Sensitivity (%)	98.6	98.1	97.1	97.1	96.2	96.2	95.7	95.2	94.7	94.5	94.5	94.2	94.0	93.5	93.5	93.3	93.3	93.3	93.3	93.0	92.8
Specificity (%)	82.9	92.8	94.2	95.6	96.4	96.8	96.9	97.3	97.4	97.7	98.0	98.0	98.0	98.3	98.4	98.5	98.5	98.6	98.6	98.6	98.6
Accuracy (%)	87.7	94.4	95.1	96.1	96.3	96.6	96.5	96.7	96.6	96.7	96.9	96.8	96.8	96.8	96.9	96.9	96.9	97.0	97.0	96.9	96.8
Balanced Accuracy (%)	90.8	95.4	95.7	96.4	96.3	96.5	96.3	96.3	96.1	96.1	96.2	96.1	96.0	95.9	96.0	95.9	95.9	95.9	95.9	95.8	95.7

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient. RSPECT = RSNA Pulmonary Embolism CT Dataset, FUMPE = Ferdowsi University of Mashhad's PE dataset, Testing Dataset = 551 PE negative CTPAs from internal testing set + 32 PE positive and 2 PE negative from FUMPE + 385 PE positive and 385 PE negative CTPAs from RSPECT

Supplementary Table 12. Diagnostic performance of the trained model without post-processing strategy in the combined testing dataset

	Testing Dataset (CTPAs = 1355)																				
	Without Post-Processing Strategy																				
Metric \ Threshold	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
No. of TN	346	640	705	748	764	781	796	802	806	812	816	824	827	831	836	836	839	841	842	846	848
No. of FP	592	298	233	190	174	157	142	136	132	126	122	114	111	107	102	102	99	97	96	92	90
No. of TP	417	417	417	417	417	417	416	415	414	414	413	413	412	410	405	405	405	404	403	403	403
No. of FN	0	0	0	0	0	0	1	2	3	3	4	4	5	7	12	12	12	13	14	14	14
MCC (%)	39.0	63.1	69.4	74.0	75.8	77.8	79.4	79.9	80.2	80.9	81.2	82.3	82.5	82.6	82.3	82.3	82.7	82.7	82.7	83.2	83.5
Sensitivity (%)	100.0	100.0	100.0	100.0	100.0	100.0	99.8	99.5	99.3	99.3	99.0	99.0	98.8	98.3	97.1	97.1	97.1	96.9	96.6	96.6	96.6
Specificity (%)	36.9	68.2	75.2	79.7	81.4	83.3	84.9	85.5	85.9	86.6	87.0	87.8	88.2	88.6	89.1	89.1	89.4	89.7	89.8	90.2	90.4
Accuracy (%)	56.3	78.0	82.8	86.0	87.2	88.4	89.4	89.8	90.0	90.5	90.7	91.3	91.4	91.6	91.6	91.6	91.8	91.9	91.9	92.2	92.3
Balanced Accuracy (%)	68.4	84.1	87.6	89.9	90.7	91.6	92.3	92.5	92.6	92.9	93.0	93.4	93.5	93.5	93.1	93.1	93.3	93.3	93.2	93.4	93.5

Note. — The thresholds are in mm^3. CTPAs = computed tomography (CT) pulmonary angiography (CTPA) examinations, TN = true-negative CTPAs, FP = false-positive CTPAs, TP = true-positive CTPAs, FN = false-negative CTPAs, MCC = Matthew's correlation coefficient. RSPECT = RSNA Pulmonary Embolism CT Dataset, FUMPE = Ferdowsi University of Mashhad's PE dataset, Testing Dataset = 551 PE negative CTPAs from internal testing set + 32 PE positive and 2 PE negative from FUMPE + 385 PE positive and 385 PE negative CTPAs from RSPECT