Random forest-based out-of-distribution detection for robust lung cancer segmentation

Aneesh Rangnekar and Harini Veeraraghavan

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA

ABSTRACT

Accurate detection and segmentation of cancerous lesions from computed tomography (CT) scans is essential for automated treatment planning and cancer treatment response assessment. Transformer-based models with self-supervised pretraining can produce reliably accurate segmentation from in-distribution (ID) data but degrade when applied to out-of-distribution (OOD) datasets. We address this challenge with RF-Deep, a random forest classifier that utilizes deep features from a pretrained transformer encoder of the segmentation model to detect OOD scans and enhance segmentation reliability. The segmentation model comprises a Swin Transformer encoder, pretrained with masked image modeling (SimMIM) on 10,432 unlabeled 3D CT scans covering cancerous and non-cancerous conditions, with a convolution decoder, trained to segment lung cancers in 317 3D scans. Independent testing was performed on 603 3D CT public datasets that included one ID dataset and four OOD datasets comprising chest CTs with pulmonary embolism (PE) and COVID-19, and abdominal CTs with kidney cancers and healthy volunteers. RF-Deep detected OOD cases with a FPR95 of 18.26%, 27.66%, and $< 0.1\%$ on PE, COVID-19, and abdominal CTs, consistently outperforming established OOD approaches. The RF-Deep classifier provides a simple and effective approach to enhance reliability of cancer segmentation in ID and OOD scenarios.

Keywords: Swin, lung cancer segmentation, out-of-distribution detection

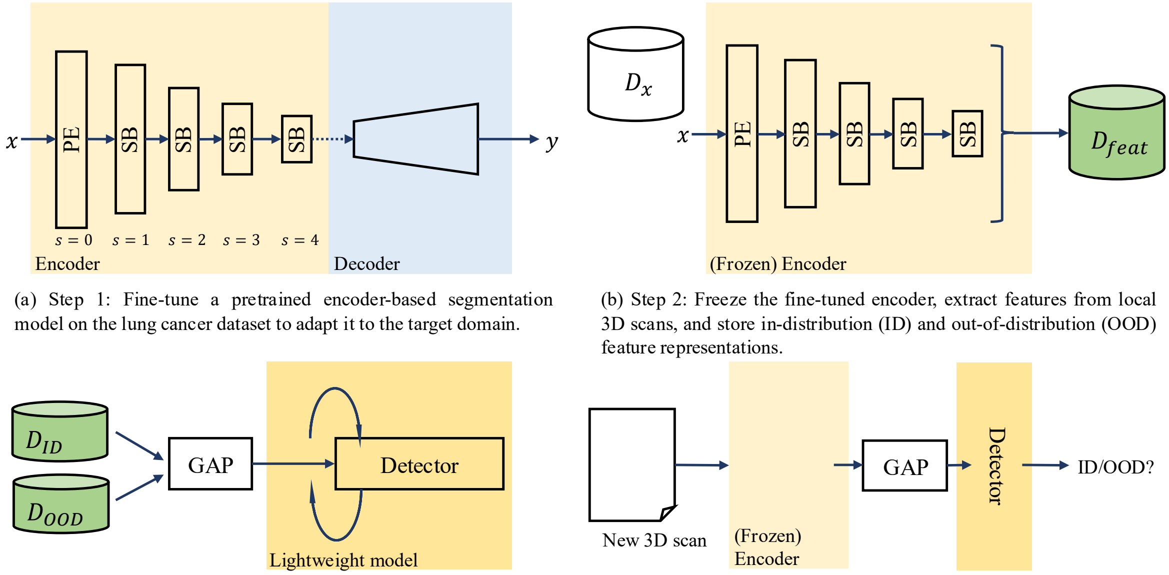
1. PURPOSE

Pretrained transformer based deep learning (DL) methods combined with convolutional decoders have demonstrated capability to segment organs and tumors from radiological images.[1–6] A major challenge in deploying DL segmentation models at scale in research and clinical settings is the potential accuracy degradation when the same models are applied to real world scenarios that differ from those seen during training. For example, models trained to segment malignant lung cancers from chest CTs for treatment planning may also be applied to scans with benign nodules from lung cancer screening or scans containing unrelated diseases such as pulmonary embolisms. Although all of these involve chest CTs, models trained for cancer segmentation can produce unanticipated and incorrect results when applied outside their intended scope.

Traditional segmentation evaluation metrics such as the Dice similarity coefficient (DSC) and Hausdorff distance at 95th percentile (HD95) are designed to assess model performance on in-distribution (ID) datasets, but are insufficient indicators of the model's robustness to out-of-distribution (OOD) data. OOD detection approaches relying on model confidence scores, such as MaxSoftmax,[7] often fail in scenarios where models produce confidently incorrect segmentations, occurring in both medical[8] and natural image analyses.[9] These methods have been applied to distinctly different disease sites such as the lung and abdomen,[10,11] a relatively easier task compared to detecting OOD cases within the same site. Alternative approaches using secondary models, such as VQ-GANs,[12,13] require substantially large secondary training data, are computationally intensive, and lack interpretability, limiting their practical utility in high-throughput clinical workflows. Finally, radiomics feature-based methods that leverage standardized features have been shown to be effective at distinguishing scans of organs but have not generalized to tumors.[14,15] To address these limitations, we developed a random forest classifier combining deep features (RF-Deep) of a model trained to segment lung cancers to detect OOD scans.

We evaluated our OOD detection approach on both, far-OOD scans consisting of diseases occurring in different anatomic site, and the more challenging near-OOD scans where different diseases occur in the same anatomic site

Send correspondence to Aneesh Rangnekar (Email: rangnea@mskcc.org)



(a) Step 1: Fine-tune a pretrained encoder-based segmentation model on the lung cancer dataset to adapt it to the target domain.

(b) Step 2: Freeze the fine-tuned encoder, extract features from local 3D scans, and store in-distribution (ID) and out-of-distribution (OOD) feature representations.

(c) Step 3: Train a lightweight OOD detector on stored ID and OOD features using random forest classifier.

(d) Step 4: Given a new 3D scan, extract features using the frozen encoder, compute MAP-based scores, and classify as ID or OOD using the trained detector.

Figure 1: Random forest (RF-Deep) out-of-distribution (OOD) detection framework. The lung tumor segmentation model developed on ID cases is used to extract features from the patch embedding (PE) layer and Swin blocks (SB) across all four stages to be used with a random forest classifier.

with similar appearance statistics as the training datasets. In summary, our contributions are: **(a)** a lightweight, random forest classifier based OOD detection framework utilizing deep features trained to extract useful feature representations for ID cases. The deep features use the encoder features in order to leverage the robustness of the pretrained encoder to imaging distribution variations inherent even in ID datasets. The classifier is trained on a small set of ID and representative OOD examples, following the outlier exposure paradigm,[16] **(b)** a demonstration of the effectiveness of our approach on lung cancer segmentation from 3D CT scans, including both near-OOD scenarios in chest CTs (pulmonary embolism and COVID-19) and far-OOD scenarios in abdominal CTs (kidney cancers and non-cancerous pancreas), and **(c)** a systematic comparison of our RF-Deep approach against common OOD detection methods (like MaxSoftmax) and a radiomics feature–based classifier (RF-Radiomics), using five public datasets of 603 patient scans.

2. METHOD

OOD detection task definition. Let $\mathcal{D}_{\text{in}}$ denote the distribution of lung cancer CT scans, closely matching the training dataset used to create the lung tumor segmentation model, and $\mathcal{D}_{\text{out}}$ represents the scans that differ in pathology and anatomic site. The objective is to determine whether a new scan x belongs to $\mathcal{D}_{\text{in}}$ or $\mathcal{D}_{\text{out}}$, at the scan level, via a scoring mechanism $S(x)$. Since tumor segmentation is binary, aggregation is restricted to the predicted tumor regions, leveraging model-relevant spatial context while excluding irrelevant anatomy.

Segmentation model architecture. A hybrid transformer encoder-convolutional decoder architecture was employed. The encoder extracts global and spatially local contexts through a hierarchical Swin Transformer[17] encoder. The decoder leverages local spatial precision of a U-Net–based[18] convolutional network to accurately delineate anatomical boundaries. The encoder uses a depth configuration of $2-2-12-2$ across four stages with $4 \times 4 \times 4$ patch size, enabling multi-scale feature aggregation, while windowed self-attention reduces computational complexity yet preserves global context for large volumetric inputs ($128 \times 128 \times 128$ voxels). Encoder was initialized with pretrained weights with self-supervised learning performed using SimMIM[19] method that uses masked image modeling (MIM) approach to predict and reconstruct masked image patches. Pretraining used 10,412 unlabeled 3D CT scans sourced from public and institutional datasets (cite MedPhys paper) with MIM

Table 1: Out-of-distribution (OOD) detection performance comparing RF-Deep and other methods. Results are reported as AUROC (↑) and FPR95 (↓), wiith best values per dataset highlighted in bold.

Method	Pulmonary Embolism		COVID-19		Kidney Cancer		Abdomen	
	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)
MaxSoftmax	88.74	37.01	89.53	53.25	96.72	12.34	97.26	14.29
MaxLogits	90.88	34.42	91.30	37.01	97.51	7.140	98.12	6.490
Energy	90.97	35.06	91.05	37.01	97.50	7.140	98.12	6.490
Entropy	90.59	33.81	92.47	53.24	96.97	12.23	97.55	12.95
RF-Radiomics	88.26	40.13	90.98	36.36	95.94	19.87	95.90	15.06
RF-Deep (Ours)	**95.16**	**18.26**	**92.88**	**27.66**	**99.81**	**0.110**	**99.89**	**0.720**

task performed by randomly masking 75% of 3D patches in the image. Note that SSL pretraining is a unsupervised pretraining approach that does not require labeled image datasets for pretraining. Pretrained encoder was used to extract features that are robust to CT image acquisition features to accurately identify ID cases despite imaging differences and detect OOD cases.

OOD Detection with RF-Deep classifier. Unlike object centric photographic images, medical images may contain varying anatomic extent, making a global image-based OOD assessment inefficient and potentially less accurate. Hence, we leveraged the target task-relevant regions or tumors as extracted by the segmentation model to focus OOD detection on relevant image regions. Our approach (Fig. 1) consists of four steps: **(i)** fine-tune a segmentation model for lung tumor segmentation, **(ii)** obtain tumor-centered 3D image regions, **(iii)** extract an aggregate feature representation within the generated tumor regions from the multi-scale transformer encoder layers and train a RF-Deep classifier using the deep features using ID and OOD examples, and **(iv)** the extracted RF-Deep classifier can then be used to distinguish ID from OOD samples.

Implementation details. The segmentation model was trained from a public dataset containing non-small cell lung cancers[20] (N=317). It was implemented using PyTorch[21] and MONAI,[22] fine-tuned using cross-entropy and Dice loss with a batch size of 16 across 4 NVIDIA GPUs. A learning rate of 2×10^{-4} was used, with linear warm-up and cosine annealing over 1000 epochs. Augmentations included flips, rotations, affine, and intensity shifts. Inputs were normalized (HU $[-400, 400]$), resampled (1mm^3), and cropped to 128^3. Sliding window with 50% overlap was used for inference. RF-Deep classifier consisted of a random forest (1000 trees, max depth 20, balanced weights) trained via the scikit-learn library. Features were extracted from 8 tumor-centered crops per scan, and predictions were averaged at inference. Experiments to detect OOD utilized a held-out ID dataset containing 140 lung cancers was used[23] and four datasets consisting of pulmonary embolism (PE)[24] (N=120), COVID-19[25] (N=120), kidney cancers[26] (N=120), and healthy abdominal CT scans[27] (N=82).

OOD comparison methods. The RF-Deep classifier was compared against standard OOD methods including MaxSoftmax,[7] MaxLogits,[28] energy,[29] and entropy measures. In addition, a RF-radiomics classifier was created using 293 IBSI-compliant radiomic features[30] extracted from detected tumor regions using PyCERR radiomics library.[31,32] In order to prevent accuracy degradation from correlated features, recursive feature elimination was employed. RF used 1000 trees, max depth of 20, and balanced weights.

OOD experiment protocol. RF-based approaches used a fixed 40/60 patient-level train–test split, repeated over 100 seeds. Other baselines used the full cohorts as they require no auxiliary data.

Evaluation metrics. We report AUROC to measure ID–OOD separability[33] and FPR95 as a threshold-based metric.[34]

3. RESULTS

Our results demonstrate that RF-Deep provides a robust scan-level OOD detection for lung cancer segmentation, outperforming established approaches including MaxSoftmax and MaxLogits (1). SimMIM-pretrained encoder features, paired with RF, consistently yielded higher AUROC and lower FPR95 on our lung cancer segmentation task, with large margin gains in the abdominal cohorts (FPR95 $\leq$ 0.1, second best is 6 points away). Fig. 2 visually shows RF-Deep's ability to detect scans from completely different anatomical sites and different disease cases, with some limitations. In pulmonary embolism cases, we observed that the segmentation model often

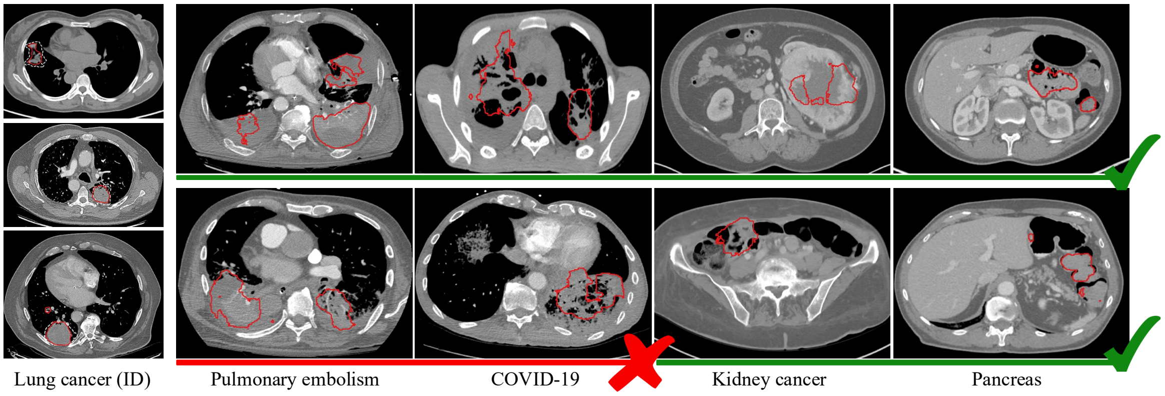
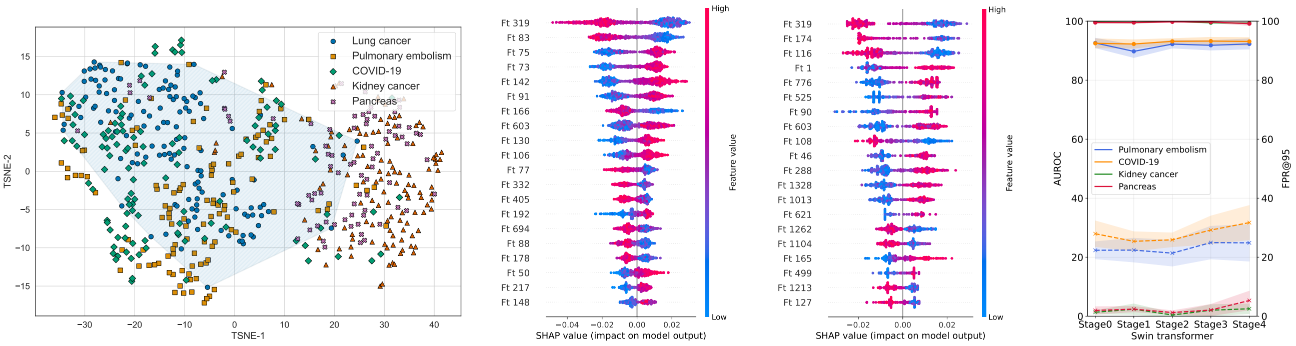


Figure 2: Representative results with green checkmarks indicating correctly detected OOD scans and the red cross highlighting missed cases. The red contours denote tumor segmentations generated by the fine-tuned segmentation model.



(a) (b) (c) (d)

Figure 3: Supporting analyses of our approach: (a) t-SNE visualization shows ID/OOD separation in encoder representations (the convex hull denotes extent of ID), (b,c) SHAP analysis highlighting feature importance and interpretability in RSNA-Pulmonary Embolism and KiTS-23 kidney cancer datasets respectively, and (d) Stage-wise ablation shows individual RF-Deep performance from Swin Transformer encoder features.

highlights tumor-like structures rather than the emboli; nevertheless, our contextual features enable RF-Deep to correctly classify these scans as OOD most of the times, surpassing non-contextual approaches.

Additionally, we performed t-SNE clustering of the features[35] across all cohorts (Fig. 3a) and observed partially distinct clusters for in-distribution (ID) and OOD cohorts, making them favourably separable with the random forest classifier. Fig. 3b and Fig. 3c show the SHAP analysis on the RF-Deep providing partial interpretability in identifying features that are most influential in classifying a scan as OOD. Finally, stage-wise ablation (Fig. 3d), wherein only features corresponding to a particular stage of the transformer were used for training RF-Deep, demonstrated that mid- and early-deep level Swin Transformer features are most effective, aligning with our SHAP findings.

4. CONCLUSION

We performed a comprehensive evaluation of OOD detection approaches applied to the clinical task of lung cancer auto-segmentation. Our analysis show that RF-Deep, built on SimMIM-pretrained encoder features with a random forest, achieved consistent improvements over the established approaches, surpassing in far-OOD scenarios. The approach is lightweight, interpretable, and improves the reliability of segmentation models by highlighting potential cases for manual intervention in clinical settings. Future work involves scaling it to more disease sites with larger cohorts, multi-class tasks, and generalizing across other diverse pretraining approaches.

ACKNOWLEDGMENTS

This research was partially supported by NCI R01CA258821 and Memorial Sloan Kettering (MSK) Cancer Center Support Grant/Core Grant NCI P30CA008748.

REFERENCES

[1] Willemink, M., Roth, R., and Sandfort, V., "Toward foundational deep learning models for medical imaging in the new era of transformer networks," *Radiol Artif Intell* **4**(6) (2022).

[2] Jiang, J., Tyagi, N., Tringale, K., Crane, C., and Veeraraghavan, H., "Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (SMIT)," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 556–566, Springer (2022).

[3] Nguyen, D. M. H., Nguyen, H., Mai, T. T. N., Cao, T., Nguyen, B. T., Ho, N., Swoboda, P., Albarqouni, S., Xie, P., and Sonntag, D., "Joint self-supervised image-volume representation learning with intra-inter contrastive clustering," in [*Proc. 37th AAAI*], AAAI Press (2023).

[4] Yan, X., Naushad, J., Sun, S., Han, K., Tang, H., Kong, D., Ma, H., You, C., and Xie, X., "Representation recovering for self-supervised pre-training on medical images," in [*2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 2684–2694 (2023).

[5] Qayyum, A., Razzak, I., Mazher, M., Khan, T., Ding, W., and Niederer, S., "Two-stage self-supervised contrastive learning aided transformer for real-time medical image segmentation," *IEEE Journal of Biomedical and Health Informatics* , 1–10 (2023).

[6] Gu, H., Dong, H., Yang, J., and Mazurowski, M. A., "How to build the best medical image segmentation algorithm using foundation models: a comprehensive empirical study with segment anything model," *arXiv preprint arXiv:2404.09957* (2024).

[7] Hendrycks, D. and Gimpel, K., "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations* (2017).

[8] Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.-B., and Yang, G., "Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation," *Journal of Digital Imaging* **36**(2), 739–752 (2023).

[9] Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D., "Scaling out-of-distribution detection for real-world settings," *ICML* (2022).

[10] Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T., "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging* **39**(12), 3868–3878 (2020).

[11] Zimmerer, D., Full, P. M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., et al., "Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images," *IEEE Transactions on Medical Imaging* **41**(10), 2728–2738 (2022).

[12] Graham, M. S., Tudosiu, P.-D., Wright, P., Pinaya, W. H. L., Jean-Marie, U., Mah, Y. H., Teo, J. T., Jager, R., Werring, D., Nachev, P., et al., "Transformer-based out-of-distribution detection for clinically safe segmentation," in [*International Conference on Medical Imaging with Deep Learning*], 457–476, PMLR (2022).

[13] Pinaya, W. H., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J., "Unsupervised brain imaging 3d anomaly detection and segmentation with transformers," *Medical Image Analysis* **79**, 102475 (2022).

[14] Vasiliuk, A., Frolova, D., Belyaev, M., and Shirokikh, B., "Limitations of out-of-distribution detection in 3d medical image segmentation," *Journal of Imaging* **9**(9), 191 (2023).

[15] Konz, N., Chen, Y., Gu, H., Dong, H., Chen, Y., and Mazurowski, M. A., "Rad: A metric for medical image distribution comparison in out-of-domain detection and other applications," (2024).

[16] Hendrycks, D., Mazeika, M., and Dietterich, T., "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606* (2018).

[17] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 10012–10022 (2021).

[18] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 234–241, Springer (2015).

[19] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H., "Simmim: A simple framework for masked image modeling," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 9653–9663 (2022).

[20] Aerts, H., Velazquez, E. R., Leijenaar, R., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al., "Data from nsclc-radiomics," *The cancer imaging archive* (2015).

[21] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems* **32** (2019).

[22] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al., "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701* (2022).

[23] Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Zhang, W., Leung, A., Kadoch, M., et al., "Data for nsclc radiogenomics collection," *The Cancer Imaging Archive* **10**, K9 (2017).

[24] Colak, E., Kitamura, F. C., Hobbs, S. B., Wu, C. C., Lungren, M. P., Prevedello, L. M., Kalpathy-Cramer, J., Ball, R. L., Shih, G., Stein, A., et al., "The rsna pulmonary embolism ct dataset," *Radiology: Artificial Intelligence* **3**(2), e200254 (2021).

[25] Tsai, E. B., Simpson, S., Lungren, M. P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B. J., Shih, G., Stein, A., Kalpathy-Cramer, J., Shen, J., Hafez, M. A., John, S., Rajiah, P., Pogatchnik, B. P., Mongan, J. T., Altinmakas, E., Ranschaert, E., Kitamura, F. C., Topff, L., Moy, L., Kanne, J. P., and Wu, C., "Medical imaging data resource center (MIDRC) - RSNA international COVID open research database (RICORD) release 1b - chest ct covid-," (2021).

[26] Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., and Weight, C., "The KiTS21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct," (2023).

[27] Roth, H. R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., and Summers, R. M., "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*], 556–564, Springer (2015).

[28] Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Mostajabi, M., Steinhardt, J., and Song, D. X., "Scaling out-of-distribution detection for real-world settings," in [*International Conference on Machine Learning*], (2022).

[29] Liu, W., Wang, X., Owens, J., and Li, Y., "Energy-based out-of-distribution detection," *Advances in neural information processing systems* **33**, 21464–21475 (2020).

[30] Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al., "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology* **295**(2), 328–338 (2020).

[31] Deasy, J. O., Blanco, A. I., and Clark, V. H., "CERR: a computational environment for radiotherapy research," *Medical physics* **30**(5), 979–985 (2003).

[32] Apte, A. P., Iyer, A., Crispin-Ortuzar, M., Pandya, R., Van Dijk, L. V., Spezi, E., Thor, M., Um, H., Veeraraghavan, H., Oh, J. H., et al., "Extension of CERR for computational radiomics: A comprehensive MATLAB platform for reproducible radiomics research," *Medical physics* **45**(8), 3713–3720 (2018).

[33] Davis, J. and Goadrich, M., "The relationship between precision-recall and roc curves," in [*Proceedings of the 23rd international conference on Machine learning*], 233–240 (2006).

[34] Liang, S., Li, Y., and Srikant, R., "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690* (2017).

[35] Masarczyk, W., Ostaszewski, M., Imani, E., Pascanu, R., Miłoś, P., and Trzcinski, T., "The tunnel effect: Building data representations in deep neural networks," *Advances in Neural Information Processing Systems* **36** (2024).