Chin Med Sci J
September 2021

Vol. 36, No. 3 P. 210-217
doi:10.24920/003968

ORIGINAL ARTICLE *Artificial Intelligence & Precision Oncology*

External and Internal Validation of a Computer Assisted Diagnostic Model for Detecting Multi-Organ Mass Lesions in CT images

Lianyan Xu[1#], Ke Yan[2#], Le Lu[2], Weihong Zhang[1], Xu Chen[3], Xiaofei Huo[3], Jingjing Lu[3*]

[1]Department of Radiology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China
[2]PAII Inc., Bethesda, MD 20817, USA
[3]Department of Radiology, Beijing United Family Hospital, Beijing 100015, China

Key words: lesion detection; computer-aided diagnosis; convolutional neural network; deep learning

Objective We developed a universal lesion detector (ULDor) which showed good performance in in-lab experiments. The study aims to evaluate the performance and its ability to generalize in clinical setting via both external and internal validation.

Methods The ULDor system consists of a convolutional neural network (CNN) trained on around 80K lesion annotations from about 12K CT studies in the DeepLesion dataset and 5 other public organ-specific datasets. During the validation process, the test sets include two parts: the external validation dataset which was comprised of 164 sets of non-contrasted chest and upper abdomen CT scans from a comprehensive hospital, and the internal validation dataset which was comprised of 187 sets of low-dose helical CT scans from the National Lung Screening Trial (NLST). We ran the model on the two test sets to output lesion detection. Three board-certified radiologists read the CT scans and verified the detection results of ULDor. We used positive predictive value (PPV) and sensitivity to evaluate the performance of the model in detecting space-occupying lesions at all extra-pulmonary organs visualized on CT images, including liver, kidney, pancreas, adrenal, spleen, esophagus, thyroid, lymph nodes, body wall, thoracic spine, *etc*.

Results In the external validation, the lesion-level PPV and sensitivity of the model were 57.9% and 67.0%,

Received July 10, 2021; accepted August 30, 2021; published online September 21, 2021.
[#]Authors contributed equally to this work.
*Corresponding author Tel: 86-13501398512; E-mail: cjr.lujingjing@vip.163.com

respectively. On average, the model detected 2.1 findings per set, and among them, 0.9 were false positives. ULDor worked well for detecting liver lesions, with a PPV of 78.9% and a sensitivity of 92.7%, followed by kidney, with a PPV of 70.0% and a sensitivity of 58.3%. In internal validation with NLST test set, ULDor obtained a PPV of 75.3% and a sensitivity of 52.0% despite the relatively high noise level of soft tissue on images.

Conclusions The performance tests of ULDor with the external real-world data have shown its high effectiveness in multiple-purposed detection for lesions in certain organs. With further optimisation and iterative upgrades, ULDor may be well suited for extensive application to external data.

WHEN IBM purchased Merge Healthcare, a company possessing a large data set of radiological images with the associated reports, AI stepped into medical imaging.[1] Since then, various AI algorithms in different fields have surprisingly sprung up. There are algorithms developed to perform different tasks as detection, prediction, segmentation. For example, algorithms were reported being capable of detecting universal trauma on pelvic radiographs,[2] predicting genetic variations in low-grade gliomas,[3] detecting multiple clinically important pathologies with chest radiographs,[4, 5] segmenting and quantifying the traumatic brain injury lesions on head CT,[6] identifying mammographic masses,[7] and detecting pancreatic cancer or cystic lesions,[8] etc.

Deep learning offers considerable promise for medical imaging diagnosis. However, the current existing lesion detection productions mostly focus on specific type of lesions and organs.[8, 9] In clinical circumstances, different type of lesions in visualized organs are often found in a single patient, and radiologists are required to report all the lesions, especially those with clinical significance.

Since DeepLesion[10] dataset was released in 2018, scientists have started to work to develop automated universal lesion detection algorithm. It is designed to assess all the organs and detect all abnormalities present on CT images to perform fully automated image interpretation. Based on convolutional neural network (CNN), we developed a universal lesion detector (ULDor) to help radiologists find all potential lesions within one unified computing framework[11] and trained it on six public lesion datasets including DeepLesion dataset. The model has shown good performance in previous internal validation,[10-14] but the generalization needs external validation. The study aims to evaluate the model's ability to generalize with unseen data. If the model performs as well in the test samples as in the training samples, it is likely to be applied in clinical practice in future.

MATERIALS AND METHODS

Algorithm development

The universal lesion detection framework we proposed is a computer-aided diagnosis (CAD) system based on deep learning. It consists of a detection network and a 3D convolutional neural network (CNN) classifier for false positive reduction. The detection network is developed from CNNs, blending feature pyramid module and 3D feature fusion mechanism to improve the accuracy of detecting subtle lesions.

First, we trained the network on the DeepLesion dataset, which contains over 32,000 lesions on various body parts in CT scans, including lung nodules, liver lesions, enlarged lymph nodes, kidney lesions, bone lesions, etc. It is large-scale but partially-labeled, produced by mining hospital archives. However, the mined clinical annotations are incomplete with many lesions unlabeled, which subsequently degrades the accuracy of the trained detector. For this reason, we added other 5 datasets to tackle this problem and proposed three strategies to harvest/complete these missing annotations. They are 3D annotation completion strategy, intra-patient lesion matching strategy, and cross-dataset knowledge transfer strategy.[11] The 5 organ-specific datasets we used are small-scale but full-labeled public single-type lesion datasets, *i.e.*, LUNA (LUng Nodule Analysis),[15] LiTS (Liver Tumor Segmentation Benchmark),[16] NIH-LN (NIH Lymph Node),[17] Decathlon-lung,[18] and Decathlon-Hepatic Vessel,[18] respectively.

The proposed detection framework is exhibited in **Figure 1**. Given an image, the algorithm can predict several groups of lesion proposals that respectively match the semantics of each dataset, and these proposals are complementary. By filtering the overlapped parts and fusing the remaining, the model generates the joint lesion proposals decision, like human

generalists and specialists cooperate to combine their knowledge and get a final diagnosis.

The external and internal validation datasets

We retrospectively selected 188 consecutive plain chest and upper abdomen CT scans from both in-patients and outpatients of Beijing United Family Hospital (UFH) between February 8th and March 31st in 2021. All the CT scans were retrieved from the hospital's Picture Archiving and Communication System regardless of the disease. Each CT scan is composed of hundreds of thin axial images. This test set has never been seen by the model, serving as the external validation dataset of this study. The data collection had been approved by the ethic review committee of the UFH.

The internal validation dataset consists of 216 low-dose helical CT scans from the National Lung Screening Trial (NLST) of the US, which is publicly available from the Cancer Data Access System of NIH.[19] It consists of randomly selected patients with reported cancer or metastasis, from which we have chosen 36 separate scans to fine tune the model to reduce common false positives (FPs) before internal validation.

Radiologists browsed through all the CT images of both external and internal validation datasets, marked images with poor quality (great noise or incomplete images) for exclusion. Consequently, 164 plain chest and upper abdomen CT scans from UFH and 187 low-dose helical CT scans from NLST formed the final test sets.

Implementation of the universal lesion detection model

We ran the ULDor on these test sets to output lesion detection. All types of mass-like abnormalities in liver, kidney, pancreas, thyroid, lymph nodes, body wall, thoracic spine (T-spine) and other organs were supposed to be detected. We did not consider lung lesions in this study because they have been extensively studied with other single-organ lesion detection algorithms. The ULDor recorded the following indices for each abnormal lesion: location, size, CT value, and confidence score. The model produced a confidence score for each lesion along with its diameter and the organ name. Confidence score is the average score of the detection and classification networks. We kept findings with score > 0.6 and removed low-confidence ones. The score setting was based on experience of our scientists who developed the algorithm. As Lymph node analysis is of crucial importance for patients, especially cance

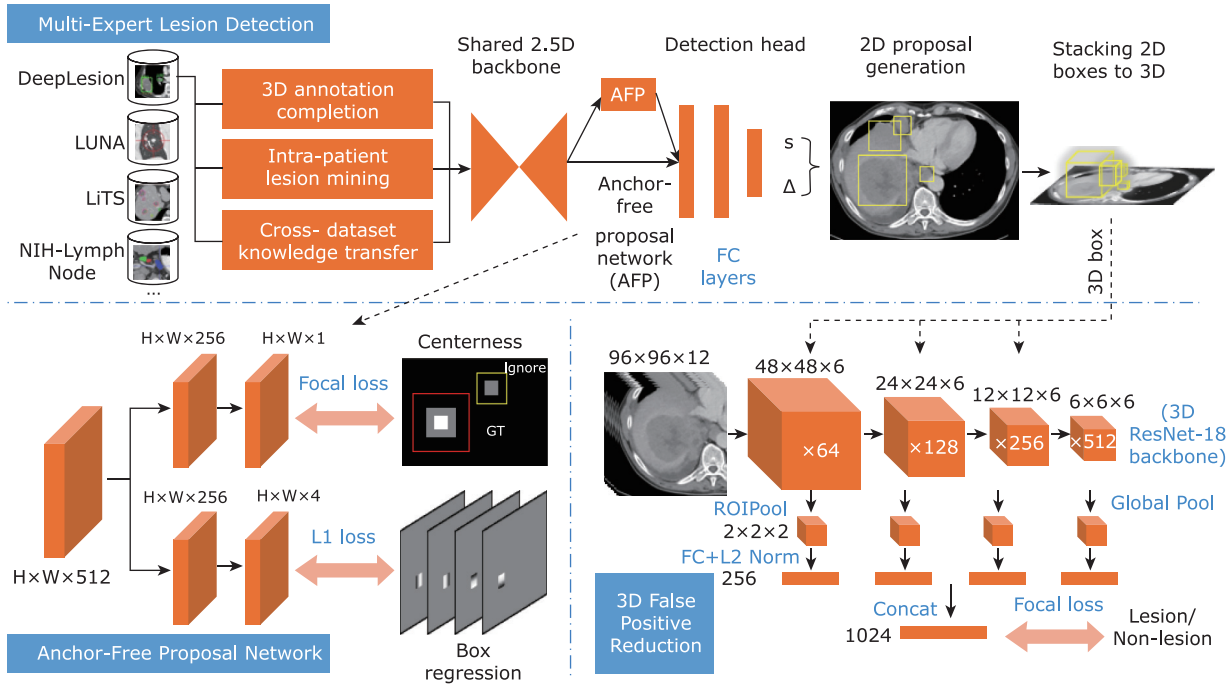

Figure 1. Framework of the proposed algorithm model

The anchor-free proposal (AFP) network and the 3D false positive reduction (FPR) network. AFP works as the backbone to generate initial proposals. FPR further classifies the 3D lesion proposals of the detector. The detector jointly learns from multiple datasets.

patients, we consider any lymph node with short diameter greater than 0.8 cm as lymphadenopathy in the study.[20,21] The model recorded up to 5 lesions per organ or 5 enlarged lymph nodes per nodal region.

Verification by radiologists

Three radiologists including 2 independent board-certified radiology experts with average experience of 10 years (Lu JJ and Chen X) and 1 senior radiology resident (Xu LY) read the CT scans and verified the detection results. Radiologists were blinded to patients' clinical information. They labeled and revised each CT image set only rely on imaging characteristics. The main work of radiologists included to check if the detected findings were real lesions and lesion proposals were accurate or not, and to add positive findings that were missed by the detector. When encountering uncertain lesions, they discussed to reach a consensus.

Performance evaluation of the ULDor model

Performance of ULDor was evaluated by positive predictive value (PPV) and sensitivity. PPV is the percentage of TP examples to all that ULDor proposed as positive, calculated as TP/ (TP + FP); sensitivity is the proportion of TP examples to the total true-lesion examples, calculated as TP/(TP+FN). The TP, FP, FN are true positive, false positive, and false negative, respectively. TPs are lesions both model and radiologists determined as positive; FPs are lesions those ULDor detected as lesions but radiologists disagreed; FNs are those ULDor missed but radiologists added.

RESULTS

External validation results

The validation results for detecting multi-organ lesions are exhibited in **Table 1**. The examples of TP, FP, FN of each organ are presented in **Figure 2**. The model detected 340 abnormal findings, among them 197 were TPs and 143 were FPs; radiologists found 97 FNs. The lesion-level PPV and sensitivity of the model were 57.9% and 67.0%, respectively. On average, the model detected 2.1 findings per image set, of which 0.9 were FPs. The model worked well at liver lesions and thoracic-spine (T-spine) lesions, with 78.9% PPV and 92.7% sensitivity for liver, and 85.7% PPV and 100.0% sensitivity for T-spine, followed by kidney, with 70.0% PPV and 58.3% sensitivity. Among FPs of liver lesions, 85% (23/27) were portal vein and hepatic vein, and 15% (4/27) were microcalcification. There were 8 uncertain small nodules with slightly heterogeneous density missed by the model.

Both sensitivity and PPV for thoracic lymphadenopathy were around 50%. Among FPs of lymph nodes, cervical vessels and mediastinal vessels accounted for 83% (59/71), and clustering of small lymph nodes which were misidentified as enlargement accounted for 17% (12/71).

For kidney lesions**,** all FPs were the upper pole of kidney, which was subject to scan range; the FNs consisted of 3 angiomyolipomas and 2 hyperattenuating renal cysts.

Among FPs in thyroid, 75% (6/8) were adjacent common carotid artery and internal jugular vein, 25% (2/8) were thyroids with heterogeneous density.

The numbers of T-spine lesions were relatively small, all of which were detected by ULDor with high PPV and sensitivity. The only 1 FP was a bonespot.

For lesions in adrenal gland, detector achieved 85.7% PPV and 33.3% sensitivity. The 1 FP was a retroperitoneal abdominal lymph node, and 12 FNs were adrenal gland hyperplastic nodules with homogenous density.

For spleen lesions, 83% (13/16) FPs were splenules which were focus of normal splenic tissue and do not usually require treatment, the other FPs were contents of stomach.

There were 3 FPs in pancreas and 13 FPs in esophagus. These lesions were interpreted by radiologists as adjacent coiled small intestine, but further work-up or follow-up were recommended to exclude real soft tissue mass.

In addition, it is worthy to note that the detector discovered the only lesion in body wall in all cases, which was so small that doctors almost missed it.

Internal validation results

In the NLST test set, the model detected 227 abnormal findings, among which 171 findings were TPs and 56 were FPs. There were 158 FNs detected by radiologists. The model acquired 75.3% (171/227) PPV and 52.0% (171/329) sensitivity, despite the relatively high noise level of the soft tissue on images.

DISCUSSIONS

In our previous work, we developed a universal lesion detection algorithm using deep learning. From

Table 1. Performance of the ULDor model in detection of multi-organ mass lesions in external validation

Organ	Model detected (*n*)	Radiologists detected (*n*)	TP (*n*)	FP (*n*)	FN (*n*)	PPV (%)	Sensitivity (%)
Liver	128	109	101	27	8	78.9	92.7
LN thorax	131	109	60	71	49	45.8	55.1
Kidneys	10	12	7	3	5	70.0	58.3
Thyroid	14	12	6	8	6	42.9	50.0
Thoracic spine	7	6	6	1	0	85.7	100.0
Adrenal	7	18	6	1	12	85.7	33.3
Spleen	20	5	4	16	1	20.0	80.0
Pancreas	6	6	3	3	3	50.0	50.0
Esophagus	16	3	3	13	0	18.8	100.0
Body wall	1	1	1	0	0	100.0	100.0
Breasts	0	8	0	0	8	NaN	0
Gallbladder	0	5	0	0	5	NaN	0
In total	340	294	197	143	97	57.9	67.0

TP, true positive; FP, false positive; FN, false negative; PPV, positive predictive value; LN, lymph node. NaN: not a number (cannot be calculated).

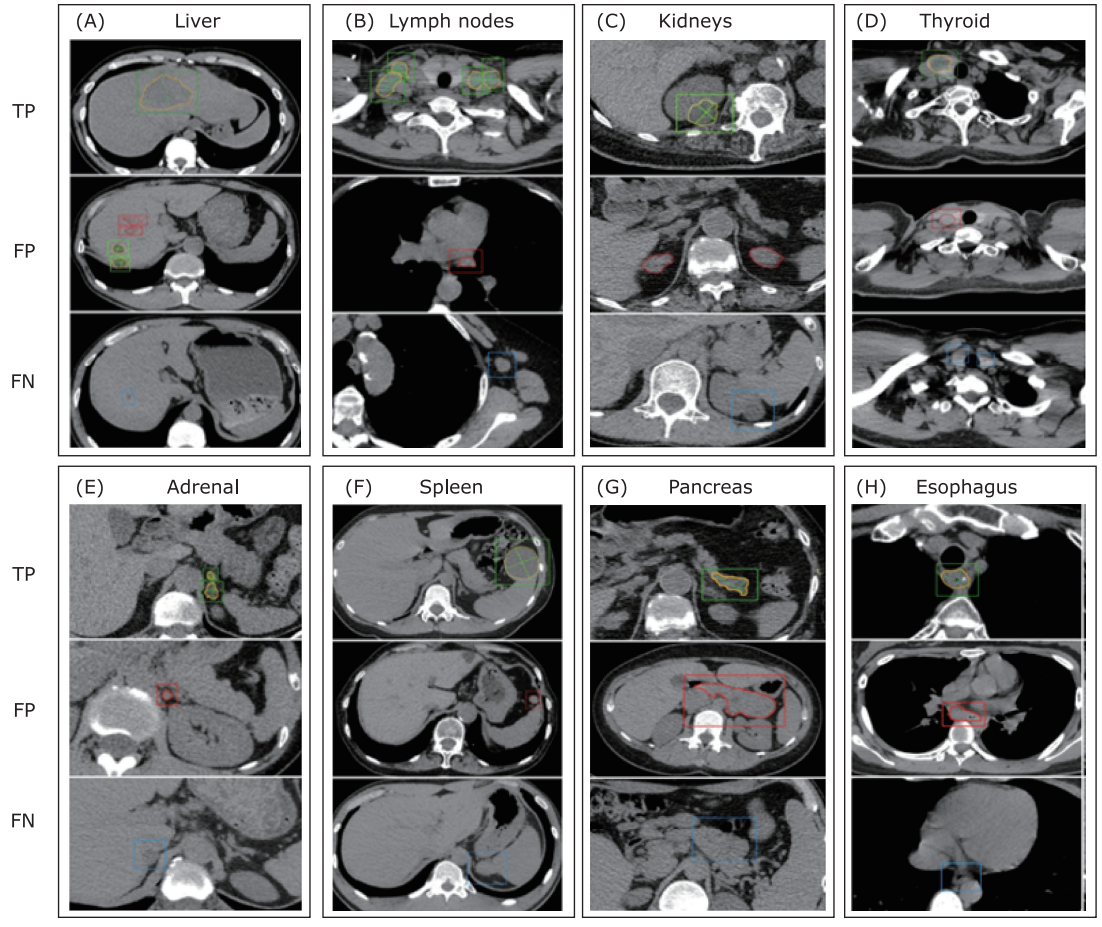

Figure 2. Examples of lesion detection and segmentation results of the ULDor model on the real-world CT image test sets. For detection, boxes in green, red, and blue are TPs, FPs and FNs, respectively. For segmentation, the green lines are ground-truth diameters, the yellow and red contours show lesions' masks. TP, true positive; FP, false positive; FN, false negative.

the algorithm development to model establishment, the model testing limited to laboratory or public data. ULDor has never been verified externally. In this study, we demonstrated the reproducibility and generalizability of the model by testing through internal and external validation. To the best of our knowledge, there have been some AI algorithms developed that are capable of identifying multiple abnormalities,[22-24] but validations in real-world practice have been scarce, and this study is the first one to develop universal lesion detection model with internal and external validation.

In the current study, ULDor detected 2.1 findings per scan, of which 0.9 were FPs. It significantly outperformed the previous works on this topic in 2018-2020, with a sensitivity of 59.6% and a false-positive rate of 2 per sub-volume.[11] We also found that our detector achieved good performance in some organs, especially in liver, kidney and T-spine, with relatively high PPVs and sensitivities.

ULDor was designed to identify space-occupying lesions of all organs in one CT scan with description of location, size, and density together, and provide a confidence score for comprehensive reference. Unlike organ-specific or single-type lesion algorithms, it was expected to achieve detection in a wider range of organs and tissue. The clinical needs arose from the facts that regular CT screening is of great significance and has been widely used worldwide for early detection of cancer, as well as regular imaging follow-ups of patients after anti-tumor treatment are common, especially in the aging of population. These undoubtedly increase routine workload of image interpretation for radiologists. The multi-organ lesion detecting model was expected to help radiologists in daily work to avoid missing lesions, as well as to reduce workload by pre-selecting suspicious lesions and prioritized patients with significant abnormality or in urgent situations that need to be treated firstly. With the increasing workload faced by radiologists, it may play an invaluable role in instantaneous proposal supports.

In clinical practice, generation of a radiological report usually can be divided into two processes: junior doctors read firstly and compile a diagnostic report, then senior doctors verify and revise the results. ULDor could serve as a screening tool to assist junior doctors or radiology residents to ascertain the characterization of a focus and support intepretation and diagnosis. It helps to find common and evident abnormalities so that radiologists can devote more time to differential diagnosis or look into rare cases. Besides, the whole detection process takes up about 80 seconds, which is less time-consuming than human interpreting and finishing a report of a CT scan.

The validation of multi-organ universal lesion detection model on external real-world data has shown good performance in liver, thoracic lymph nodes, kidneys, T-spin, but unsatisfied performance in some organs. For instance, the sensitivity was low for adrenal lesions, the PPVs were relatively low for lesions of spleen and esophagus, and the FNs were high for lesions of breast and gallbladder, which meant high possibility of missing lesions in these organs. But clinically, CT is not the routine imaging modality for detecting disease of breast and gallbladder, for its relatively low sensitivity compared to the optimal imaging modalities of these organs, such as mammography and ultrasound.

There are some limitations in the current study. Because the training dataset was mainly DeepLesion, which is abundant in high-quality annotations about tumor-like lesions, the detector was targeted on space-occupying lesions and not sensitive to non-tumor-like lesions, such as inflammatory lesions, infectious disease, congenital malformation, anatomic variation, traumatic or degenerative diseases. The external test set was of general patient population, different from the training dataset of cance patients. This difference has an unneglectable impact on the validation result, which may cause the unsatisfactory overall results in the external validation. Secondly, due to the single-center clinical setting and the small sample size, the clinical implication and generalizability of this study were not fully appreciated. The unbalanced distribution of lesions in organ (*e.g.*, more lesions in liver but fewer in adrenal gland) and nature of lesions (*e.g.*, more benign tumors but less malignant ones, more cystic but fewer solid ones) hindered ULDor's performance to be fully verified. Therefore, a large-scale multi-centric investigation should be arranged to provide more potent statistical power. Thirdly, both training and validation of the model were blind to patients' history and clinical examination information, which may affect accuracies of the model and the radiologists in this study. ULDor might perform better when it works with medical history and other clinical information. The current study focuses on lesions in chest and upper abdomen CT, the model need extended validation with pelvic CT in future.

To sum up, ULDor should be targeted at specific

applicable population, while the PPV and sensitivity of CT examination for general patient population are organ-specific. Through external and internal validation, it has been demonstrated good performance with acceptable sensitivity and PPV in multi-task CT image analyses for some organs. With further optimization and iterative upgrade of the algorithm, the detection range and capability of the model will be significantly improved. Efficient integration of radiologists and machines is clearly achievable in radiological clinical practice.

Conflict of interests

Dr. Lv and Dr. Yan reported personal fees from PAII Inc. outside the submitted work; they both held a patent "Device and method for universal lesion detection in medical images" (US62/962,271), with a US regular patent (US16/983,373) in pending. The remaining authors declared no competing interests.

REFERENCES

1. Dreyer KJ, Geis JR. When machines think: radiology's next frontier. Radiology 2017; 285(3):713-8. doi: 10.1148/radiol.2017171183.
2. Cheng CT, Wang YR, Chen HW, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun 2021; 12(1):1066. doi: 10.1038/s41467-021-21311-3.
3. Korfiatis P, Kline TL, Lachance DH, et al. Residual deep convolutional neural network predicts MGMT methylation status. J Digit Imaging 2017; 30(5):622-8. doi: 10.1007/s10278-017-0009-z.
4. Nam JG, Kim M, Park JC, et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. Eur Respirat J 2021; 57(5):2003061. doi: 10.1183/13993003.03061-2020.
5. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Med 2018; 15(11):e1002686. doi: 10.1371/journal.pmed.1002686.
6. Monteiro M, Newcombe VFJ, Mathieu F, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. Lancet Digit Health 2020; 2(6): e314-e322. doi: 10.1016/S2589-7500(20)30085-6.
7. Benndorf M, Burnside ES, Herda C, et al. External validation of a publicly available computer assisted diagnostic tool for mammographic mass lesions with two high prevalence research datasets. Med Physic 2015; 42(8):4987-96. doi: 10.1118/1.4927260.
8. Si K, Xue Y, Yu XZ, et al. Fully end-to-end deep-learning-based diagnosis of pancreatic tumors. Theranostics 2021; 11(4):1982-90. doi: 10.7150/thno.52508.
9. Masood A, Sheng B, Li P, et al. Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images. J Biomed Inform 2018; 79:117-28. doi: 10.1016/j.jbi.2018.01.005.
10. Yan K, Wang X, Lu L, et al. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imaging (Bellingham) 2018; 5(3):036501. doi: 10.1117/1.JMI.5.3.036501.
11. Yan K, Cai JZ, Zheng YJ, et al. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. IEEE Trans Med Imaging 2020; Dec 28, 2020; E-pub ahead of print. doi: 10.1109/TMI.2020.3047598.
12. Cai JZ, Yan K, Cheng CT, et al. Deep volumetric universal lesion detection using light-weight pseudo 3D convolution and surface point regression. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. p. 3–13. Springer, Cham. doi: 10.1007/978-3-030-59719-1_1.
13. Yan K, Tang YB, Peng YF, et al. Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation, In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. p. 194–202. Springer, Cham. doi: 10.1007/978-3-030-32226-7_22.
14. Tang YB, Yan K, Tang YX, et al. ULDor: A universal lesion detector for CT scans with pseudo masks and hard negative example mining. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019. p. 833-6. doi: 10.1109/ISBI.2019.8759478.
15. Setio AA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 2017; 42:1-13. doi: 10.1016/j.media.2017.06.015.
16. Bilic P, Christ FP, Vorontsov E, et al. The Liver Tumor

Segmentation Benchmark (LiTS). arXiv:1901.04056vi [preprint], 2019. Available from http://arxiv.org/abs/1901.04056. Accessed June 20, 2021.

17. Roth H, Lu L, Seff A, et al. (2015). A new 2.5 D representation for lymph node detection in CT [Data set]. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2015.AQIIDCNM.
18. Antonelli M, Reinke A, Bakas S, et al. The Medical Segmentation Decathlon. arXiv:2106.05735v1 [preprint], 2021. Available from https://arxiv.org/abs/2106.05735. Accessed June 20, 2021.
19. National Cancer Institute. Cancer Data Access System. The National Lung Screening Trial (NLST). Available from https://cdas.cancer.gov/nlst/. Accessed June 20, 2021.
20. Nin CS, de Souza VVS, Do Amaral RH, et al. Thoracic lymphadenopathy in benign diseases: a state of the art review. Respir Med 2016; 112:10-17. doi: 10.1016/j.rmed.2016.01.021.
21. Wu ZJ, Wang ZJ, Zheng ZY, et al. Risk factors for lymph node metastasis and survival outcomes in colorectal neuroendocrine tumors. Cancer Manag Res 2020; 12:7151-64. doi: 10.2147/CMAR.S256723.
22. Li ZH, Zhang S, Zhang JG, et al. MVP-Net: multi-view FPN with position-aware attention for deep universal lesion detection. In: Shen D et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. p.13-21. Springer, Cham. doi: 10.1007/978-3-030-32226-7_2.
23. Tao QY, Ge ZY, Cai JF, et al. Improving deep lesion detection using 3D contextual and spatial attention. In: Shen D et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. p.185-93. Springer, Cham. doi: 10.1007/978-3-030-32226-7_21.
24. Zhang S, Xu JC, Chen YC, et al. Revisiting 3D context modeling with supervised pre-training for universal lesion detection in CT slices, In: Martel AL et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12264. p.542-51. Springer, Cham. doi: 10.1007/978-3-030-59719-1_53.

专题：人工智能与精准肿瘤学

CT 图像中多器官占位性病变的计算机辅助检测模型的外部和内部验证

徐潋滟[1#]，闫轲[2#]，吕乐[2]，张伟宏[1]，陈旭[3]，霍晓菲[3]，陆菁菁[3*]

[1] 中国医学科学院 北京协和医学院 北京协和医院 放射科，北京 100730，中国
[2]PAII Inc.，贝塞斯达，马里兰州 20817，美国
[3] 北京和睦家医院 放射科，北京 100015，中国

摘要

目的 我们开发了一种在实验室测试中具有较佳表现的通用病变检测模型 ULDor。本研究旨在通过外部数据集和内部数据集对其检测性能进行测试并评估其在临床上的应用价值。

方法 通过卷积神经网络（convolutional neural network，CNN）构建通用病变检测模型（ULDor）。该模型经过 DeepLesion 数据集和其他 5 个特定器官的公共数据集对模型进行训练，其中 DeepLesion 数据集包括 12,000 多组 CT 扫描图像及其中 80,000 多个病变注释。验证测试集包括外部和内部验证数据集。外部验证数据集由一家综合医院回顾性收集的 164 组胸部（含上腹部）CT 平扫检查图像组成，内部验证数据集由来自美国国家肺筛查试验（NLST）的 187 组胸部低剂量螺旋 CT 扫描图像组成。我们运行 ULDor 对这两个测试集的图像进行病变检测，记录并测量模型所检测出的所有肺外器官（包括肝、肾、胰腺、甲状腺、淋巴结、体壁、胸椎，等）的占位性病变；另由三名经过资格认证的放射科医生对两个测试集进行人工阅片，以此为标准对 ULDor 的检测结果进行验证分析，采用阳性预测值和灵敏度来评价模型的检测性能。

结果 在外部验证中，模型对所有病变的整体阳性预测值和敏感性分别为 57.9% 和 67.0%，平均每组图像检测出 2.1 个病变，其中 0.9 个是假阳性。ULDor 检出肝脏病变的能力最佳，阳性预测值为 78.9%，敏感性为 92.7%，其次是肾脏，阳性预测值为 70.0%，敏感性为 58.3%。在内部验证中，尽管图像的软组织噪声水平较高，ULDor 仍实现了 75.3% 的阳性预测值和 52.0% 的灵敏度。

结论 ULDor 在外部真实数据的验证显示模型在多用途计算机辅助诊断方面对于某些器官占位病变具有较好的检测效能。通过进一步优化和迭代升级，ULDor 或许可以很好地推广应用到外部数据。

关键词：病变检测；计算机辅助诊断；卷积神经网络；深度学习