Performance of a deep learning-based lung nodule detection system as an alternative reader in a Chinese lung cancer screening program

Cui, Xiaonan; Zheng, Sunyi; Heuvelmans, Marjolein A; Du, Yihui; Sidorenkov, Grigory; Fan, Shuxuan; Li, Yanju; Xie, Yongsheng; Zhu, Zhongyuan; Dorrius, Monique D

Published in:
European Journal of Radiology

DOI:
10.1016/j.ejrad.2021.110068

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Performance of a deep learning-based lung nodule detection system as an alternative reader in a Chinese lung cancer screening program

Xiaonan Cui a,b,1, Sunyi Zheng c,d,e,1, Marjolein A. Heuvelmans g, Yihui Du g, Grigory Sidorenkov b, Shuxuan Fan a, Yanju Li a, Yongsheng Xie e, Zhongyuan Zhu a, Monique D. Dorrius b, Yingru Zhao a, Raymond N.J. Veldhuis b, Geertruida H. de Bock g, Matthijs Oudkerk f, Peter M.A. van Ooijen e,f, Rozemarijn Vliegenthart b, Zhaoxiang Ye a,g,h

a Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Centre of Cancer, Key Laboratory of Cancer Prevention and Therapy, Department of Radiology, Tianjin, People’s Republic of China
b Westlake University, Artificial Intelligence and Biomedical Image Analysis Lab, School of Engineering, Hangzhou, People’s Republic of China
c University of Groningen, University Medical Center Groningen, Department of Radiology, Groningen, the Netherlands
d Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou, People’s Republic of China
e University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, the Netherlands
f University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, the Netherlands
g University of Groningen, University Medical Center Groningen, Department of Radiology, Groningen, the Netherlands
h University of Twente, Faculty of Electrical Engineering Mathematics and Computer Science, the Netherlands

Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou, People’s Republic of China

* Corresponding author at: Tianjin Medical University Cancer Institute and Hospital, Huan-Hu-Xi Road, Ti-Yuan-Bei, He Xi District, Tianjin 300060, People’s Republic of China.
E-mail address: zye@tmu.edu.cn (Z. Ye).
1 These authors contributed equally to this work.

Objective: To evaluate the performance of a deep learning-based computer-aided detection (DL-CAD) system in a Chinese low-dose CT (LDCT) lung cancer screening program.

Materials and methods: One-hundred-and-eighty individuals with a lung nodule on their baseline LDCT lung cancer screening scan were randomly mixed with screenees without nodules in a 1:1 ratio (total: 360 individuals). All scans were assessed by double reading and subsequently processed by an academic DL-CAD system. The findings of double reading and the DL-CAD system were then evaluated by two senior radiologists to derive the reference standard. The detection performance was evaluated by the Free Response Operating Characteristic curve, sensitivity and false-positive (FP) rate. The senior radiologists categorized nodules according to nodule diameter, type (solid, part-solid, non-solid) and Lung-RADS.

Results: The reference standard consisted of 262 nodules ≥ 4 mm in 196 individuals; 359 findings were considered false positives. The DL-CAD system achieved a sensitivity of 90.1% with 1.0 FP/scan for detection of lung nodules regardless of size or type, whereas double reading had a sensitivity of 76.0% with 0.04 FP/scan (P = 0.001). The sensitivity for detection of nodules ≥ 4 - ≤ 6 mm was significantly higher with DL-CAD than with double reading (86.3% vs. 58.9% respectively; P = 0.001). Sixty-three nodules were only identified by the DL-CAD system, and 27 nodules only found by double reading. The DL-CAD system reached similar performance compared to double reading in Lung-RADS 3 (94.3% vs. 90.0%, P = 0.549) and Lung-RADS 4 nodules (100.0% vs. 97.0%, P = 1.000), but showed a higher sensitivity in Lung-RADS 2 (86.2% vs. 65.4%, P < 0.001).

Conclusions: The DL-CAD system can accurately detect pulmonary nodules on LDCT, with an acceptable false-positive rate of 1 nodule per scan and has higher detection performance than double reading. This DL-CAD system may assist radiologists in nodule detection in LDCT lung cancer screening.

https://doi.org/10.1016/j.ejrad.2021.110068
Received 6 April 2021; Received in revised form 3 October 2021; Accepted 22 November 2021
Available online 24 November 2021
0720-048X/© 2021 Elsevier B.V. All rights reserved.
1. Introduction

Lung cancer is the leading cause of cancer mortality worldwide [1,2]. Due to the fact that lung cancer is often diagnosed at an advanced stage, the long-term survival rate is still low. Low-dose computed tomography (LDCT) scanning plays a pivotal role as an early-detection method of pulmonary nodules to improve lung cancer outcomes [3–7]. The application of LDCT in practice also presents formidable challenges to radiologists. Millions of individuals are expected to undergo LDCT screening every year because of an increasing number of lung cancer screening programs [5,6,8–10], which leads to an overwhelming reading workload. Compared with conventional CT, low-dose CT can greatly reduce the radiation dose, but also sacrifices the image quality, which may lead to missing nodules by a radiologist. In addition, radiologists are prone to overlook small nodules. Previous research has shown that 20–30% of lung nodules were missed by a single reader [11,12]. In the large National Lung Screening Trial (NLST), 6.2% of lung cancers were missed [13]. A potential approach to solve this problem is by using computer-aided detection (CAD) systems for lung nodule detection as an assistant reader.

In recent years, an increasing number of CAD systems to assist radiologists in nodule detection have been developed. Some studies indicate that the use of a CAD system can improve sensitivity of lung nodule detection beyond double reading or increase the mean percentage of agreement between scan readings [14,15]. However, until now the performance of CAD has still been considered suboptimal for clinical implementation [16]. Reasons include a high false-positive rate [11,17,18], and varying performance of systems when evaluated on scans from different vendors [19]. Recently, the public database Lung Image Database Consortium and Image Database Resource Initiative, LIDC/IDRI [20,21] was released in order to facilitate the development of CAD systems. The LIDC/IDRI consists of anonymized regular-dose CT images with marked lung nodules. Different deep learning CAD systems [11,22–31] were developed using this dataset; these achieved performances varying from 76.0% with 1.0 FP/per scan to 98.2% with 15.1 FPs/per scan. With the same dataset, we designed and validated an academic deep learning-based computer-aided detection (DL-CAD) system that reached a sensitivity of 94.3% with 2.4 FP findings on a regular-dose CT [31]. However, whether this DL-CAD system can be utilized in low-dose CT (LDCT) lung cancer screening is unknown. Therefore, the purpose of this study was to compare the performance of the academic DL-CAD system to double reading in a LDCT lung cancer screening program and to investigate the feasibility of applying the DL-CAD for automatic lung nodule detection.

2. Materials and methods

2.1. Study population

With approval from the Institutional Review Board, a low-dose CT (LDCT) screening study was conducted at the Tianjin Medical University Cancer Institute and Hospital as part of the NELCIN-B3 study from 2017 to 2021 [8]. Individuals were invited by Tianjin Medical University Cancer Institute and Hospital to undergo LDCT for lung cancer screening; inclusion criteria were age 40–74 years, and no self-reported history of any malignant tumor. All individuals signed informed consent. The detailed study design of the NELCIN-B3 study in Tianjin has been described previously [8]. In the current study, 180 individuals with one or more nodules (according to the double reading) on the baseline LDCT scan were randomly mixed with screenees without lung nodules in a 1:1 ratio.

2.2. CT protocol and image data

CT examinations included the entire chest and were performed on a Somatom Sensation 64 CT system (Siemens Medical Solutions, Forchheim, Germany). CT scans were performed at inspiration under breath-hold with a spiral scan mode with kVp setting of 120 and reference tube current of 35 mAs. The maximum radiation dose (CTDIsvol) was 2 mGy. Data were reconstructed using three reconstruction kernels, namely, D45F and B80F at 1.0/0.7 mm slice thickness/reconstruction increment, and B30 at 2.0/1.0 mm slice thickness/reconstruction increment. The reconstruction kernel B80F was used for the DL-CAD system.

2.3. Double reading by radiologists

According to the standard protocol of NELCIN-B3 study [8], all CT examinations were independently evaluated by double reading. Five junior radiologists were involved in the first reading, while three senior radiologists performed the second reading. The double reading procedure included: first use of a maximum intensity projection technique and then evaluation of the image slice by slice in three reconstruction increments, namely, D45F and B80F at 1.0/0.7 mm slice thickness/reconstruction increment, and B30 at 2.0/1.0 mm slice thickness/reconstruction increment. The reference standard was based on a consensus panel of two senior radiologists are prone to overlook small nodules. Previous research has shown that 20–30% of lung nodules were missed by a single reader [11,12]. In the large National Lung Screening Trial (NLST), 6.2% of lung cancers were missed [13]. A potential approach to solve this problem is by using computer-aided detection (CAD) systems for lung nodule detection as an assistant reader.

In recent years, an increasing number of CAD systems to assist radiologists in nodule detection have been developed. Some studies indicate that the use of a CAD system can improve sensitivity of lung nodule detection beyond double reading or increase the mean percentage of agreement between scan readings [14,15]. However, until now the performance of CAD has still been considered suboptimal for clinical implementation [16]. Reasons include a high false-positive rate [11,17,18], and varying performance of systems when evaluated on scans from different vendors [19]. Recently, the public database Lung Image Database Consortium and Image Database Resource Initiative, LIDC/IDRI [20,21] was released in order to facilitate the development of CAD systems. The LIDC/IDRI consists of anonymized regular-dose CT images with marked lung nodules. Different deep learning CAD systems [11,22–31] were developed using this dataset; these achieved performances varying from 76.0% with 1.0 FP/per scan to 98.2% with 15.1 FPs/per scan. With the same dataset, we designed and validated an academic deep learning-based computer-aided detection (DL-CAD) system that reached a sensitivity of 94.3% with 2.4 FP findings on a regular-dose CT [31]. However, whether this DL-CAD system can be utilized in low-dose CT (LDCT) lung cancer screening is unknown. Therefore, the purpose of this study was to compare the performance of the academic DL-CAD system to double reading in a LDCT lung cancer screening program and to investigate the feasibility of applying the DL-CAD for automatic lung nodule detection.

2. Materials and methods

2.1. Study population

With approval from the Institutional Review Board, a low-dose CT (LDCT) screening study was conducted at the Tianjin Medical University Cancer Institute and Hospital as part of the NELCIN-B3 study from 2017 to 2021 [8]. Individuals were invited by Tianjin Medical University Cancer Institute and Hospital to undergo LDCT for lung cancer screening; inclusion criteria were age 40–74 years, and no self-reported history of any malignant tumor. All individuals signed informed consent. The detailed study design of the NELCIN-B3 study in Tianjin has been described previously [8]. In the current study, 180 individuals with one or more nodules (according to the double reading) on the baseline LDCT scan were randomly mixed with screenees without lung nodules in a 1:1 ratio.

2.2. CT protocol and image data

CT examinations included the entire chest and were performed on a Somatom Sensation 64 CT system (Siemens Medical Solutions, Forchheim, Germany). CT scans were performed at inspiration under breath-hold with a spiral scan mode with kVp setting of 120 and reference tube current of 35 mAs. The maximum radiation dose (CTDIsvol) was 2 mGy. Data were reconstructed using three reconstruction kernels, namely, D45F and B80F at 1.0/0.7 mm slice thickness/reconstruction increment, and B30 at 2.0/1.0 mm slice thickness/reconstruction increment. The reconstruction kernel B80F was used for the DL-CAD system.

2.3. Double reading by radiologists

According to the standard protocol of NELCIN-B3 study [8], all CT examinations were independently evaluated by double reading. Five junior radiologists were involved in the first reading, while three senior radiologists performed the second reading. The double reading procedure included: first use of a maximum intensity projection technique and then evaluation of the image slice by slice in three reconstruction increments, namely, D45F and B80F at 1.0/0.7 mm slice thickness/reconstruction increment, and B30 at 2.0/1.0 mm slice thickness/reconstruction increment. The reference standard was based on a consensus panel of two senior radiologists. The consensus panel reviewed the results of the DL-CAD system and the
results obtained from the double reading blindly. The consensus panel did not check for potentially missed nodules that were not reported in either reading. Nodules were evaluated based on reported slice numbers and coordinates. The two senior radiologists identified relevant nodule (≥4 mm) and irrelevant nodules (<4 mm), and non-nodules. The final reference standard was determined by the concurrent evaluation of nodules in case of agreement, and by consensus of the two senior radiologists in case of disagreement. They measured the nodule diameter independently (average diameter was used as the final result). The rationale of conducting this reference standard was comparable to prior studies [14,36].

2.6. Evaluation metrics and statistical analysis

The FROC curve, measuring the sensitivity at various false-positive rates at the x-axis (i.e. 0.125, 0.25, 0.5, 1, 2, 4, and 8 FP/scan), was used to present the performance of the DL-CAD system on LDCT scans. Sensitivity of double reading and the DL-CAD system was calculated by comparison to the reference standard. Nodule detection performance per nodule category based on size, solidity, Lung-RADS category and screenee-based false negative rate stratified by Lung-RADS Category were compared using the McNemar test for double reading and the DL-CAD system. A P-value < 0.05 indicated statistical significance. Statistical analyses were performed by SPSS software version 20.0 (IBM, New York, US).

3. Results

The flowchart of the study is shown in Fig. 1. Age, gender, smoking history, and presence of emphysema did not differ between screenees with and without lung nodules (Table 1).

3.1. Detection performance of double reading and the DL-CAD system

Two-hundred-and-sixty-two true nodules were detected on 196 out of 360 scans according to the reference standard. There were 159 Lung-RADS 2 nodules, 70 Lung-RADS 3 nodules and 33 Lung-RADS 4 nodules. The FROC curve showed the performance of the academic DL-CAD system on the LDCT scans in this study and on the training data of the public dataset consisting of CT scans at regular dose, as displayed in Fig. 2. The nodule detection performance of the DL-CAD system on the LDCT dataset showed a sensitivity of 90.1%, with 1.0 FP/scan, when using the degree of suspicion $f$. Nodule detection rate of the DL-CAD system on the 360 LDCT scans (sensitivity 90.1%, [95% CI: 86.4–93.7%]) was significantly higher than that of double reading.

<table>
<thead>
<tr>
<th>Category</th>
<th>Nodule(s) present</th>
<th>Nodules absent</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>62.2 ± 6.6</td>
<td>61.1 ± 6.0</td>
<td>0.754</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td>0.681</td>
</tr>
<tr>
<td>Male</td>
<td>95 (52.8%)</td>
<td>98 (54.4%)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>85 (47.2%)</td>
<td>82 (45.6%)</td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td>0.520</td>
</tr>
<tr>
<td>Never</td>
<td>115 (63.9%)</td>
<td>112 (62.2%)</td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>40 (22.2%)</td>
<td>37 (20.6%)</td>
<td></td>
</tr>
<tr>
<td>Former</td>
<td>25 (13.9%)</td>
<td>33 (18.3%)</td>
<td></td>
</tr>
<tr>
<td>Emphysema presence</td>
<td>20 (11.1%)</td>
<td>12 (6.7%)</td>
<td>0.138</td>
</tr>
</tbody>
</table>

Numbers are shown as absolute values with percentages or as mean and standard deviation.

Fig. 1. Flowchart of the study. From July 2017 to December 2017, a total of 360 baseline LDCT scans in 360 screenees were included in the study.
suspected to be intrapulmonary lymph nodes. In Table 3, false negative baseline scan results (based on largest lung nodule) by double reading and the DL-CAD system were described and stratified by Lung-RADS Category. The double reading showed more false negative results than DL-CAD for Lung-RADS 2 screenees (22.2% vs. 20.2%; \( P = 0.878 \)) and Lung-RADS 3 screenees (4.7% vs. 1.6%; \( P = 0.625 \)) with no significantly difference.

3.2. False positive findings

False-positive findings detected by either double reading or the DL-CAD system are described in Table 4. More FPs were found by the DL-CAD system than by the radiologists (359 versus 13). FPs included pulmonary vessels (160, 44.6%), fibrosis (59, 16.4%), gastric mucosa (44, 12.2%), and irrelevant small nodules (< 4 mm) (40, 11.1%). The radiologists in double reading misclassified 13 findings that were fibrosis (4, 30.8%), a pulmonary vessel (1, 7.7%), and irrelevant small nodules (< 4 mm) (8, 61.5%). Interestingly, all FPs found by double reading were also detected by the DL-CAD system.

4. Discussion

In this study, we compared the performance of a DL-CAD system on LDCT scans in a lung cancer screening trial to radiologist double reading. The DL-CAD system achieved a higher sensitivity than double reading (90.1% [95% CI: 86.4–93.7] versus 76.0 [95% CI: 70.7–81.2], \( P = 0.001 \)) with a false-positive rate of 1.0 per scan. Sixty-three out of 262 nodules were identified by the DL-CAD system but missed by double reading, and 27 nodules only found by double reading, which implies that a DL-CAD system can improve the nodule detection performance of radiologists.

Before comparing the detection performance between DL-CAD and radiologists, the most challenging issue is the choice of the reference standard. According to our experience, even experienced radiologists may miss nodules, especially small nodules. If we use the results of experienced radiologists as a reference standard, it seems unfair to DL-CAD, because, when we used the public LIDC/IDRI dataset, we found that many small nodules were missed by the expert consensus, which resulted in increase of the false positive rate of DL-CAD. Therefore, we used a consensus panel of two senior radiologists to review the results of the DL-CAD system and the results obtained by double reading. The consensus panel did not check for potentially missed nodules that were not reported in either reading, and only discriminated relevant nodules, irrelevant nodules, and non-nodules. The consensus panel was fair for both double reading and DL-CAD, despite the possibility of missing false negative nodules (according to a previous study [37], it is very rare that nodules are miss-detected by both double reading and DL-CAD).

Our studies have demonstrated that the use of CAD systems can improve the sensitivity of nodule detection for radiologists. In an analysis from the NELSON study including 400 baseline LDCT examinations, CAD had a high nodule detection rate of 96.7% with 1.9 FP/scan compared to 78.1% for double reading with 0.5 FP/scan [14]. Jacobs et al. showed that although only 82% of nodules were found by the CAD system, the system was able to detect 45 (5.5%) nodules that were not identified by radiologists [38]. Our DL-CAD system confirms the value in improvement of nodule detection and achieved a higher sensitivity than double reading for nodules < 6 mm, solid nodules and Lung-RADS 2 nodules.

In lung cancer screening, it is important to not miss potential cancers at early stage. In the National Lung Screening Trial, 6.2% of lung cancers were missed by reading radiologists [13]. In our study, radiologists missed 55 (34.6%) Lung-RADS 2 nodules, with 40% suspected to be intrapulmonary lymph nodes. This may be because some radiologists did not consider intrapulmonary lymph nodes as relevant nodules and ignored them. Intrapulmonary lymph nodes are considered benign, so missed diagnosis will not cause harm to the patient. Radiologists also
Fig. 3. Examples of nodules missed by double reading. A-D: Solid nodule, diameter $\geq 8$ mm. E-H: Solid nodule, diameter $< 6$ mm.

Fig. 4. Examples of nodules missed by the DL-CAD system. A. The right middle lobe shows a nodular opacity adjacent to the pleural surface. B. The left lower lobe shows a solid nodule adjacent to the pleural surface. C. The right lower lobe shows a solid nodule with cavitation adjacent to the pleural surface. D. The left lower lobe shows a solid nodule adjacent to the pleural surface.
Table 3
False negative scan results (based on largest lung nodule) by double reading and the DL-CAD system with results stratified by Lung-RADS Category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Reference Standard</th>
<th>False negative scan result</th>
<th>Double reading</th>
<th>DL-CAD</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung-RADS 2</td>
<td>99</td>
<td>22 (22.2%)</td>
<td>20 (20.2%)</td>
<td>0.878</td>
<td></td>
</tr>
<tr>
<td>Lung-RADS 3</td>
<td>64</td>
<td>3 (4.7%)</td>
<td>1 (1.6%)</td>
<td>0.625</td>
<td></td>
</tr>
<tr>
<td>Lung-RADS 4</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Characteristics of false positive lesions reported by double reading and/or the DL-CAD system.

<table>
<thead>
<tr>
<th>False positives</th>
<th>Double reading</th>
<th>DL-CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>13</td>
<td>359</td>
</tr>
<tr>
<td>Pulmonary vessel</td>
<td>1 (7.7%)</td>
<td>160 (44.6%)</td>
</tr>
<tr>
<td>Fibrosis</td>
<td>4 (3.0%)</td>
<td>59 (16.4%)</td>
</tr>
<tr>
<td>Gastric mucosa</td>
<td>–</td>
<td>44 (12.3%)</td>
</tr>
<tr>
<td>Small nodule (&lt;4mm)</td>
<td>8 (61.5%)</td>
<td>40 (11.1%)</td>
</tr>
<tr>
<td>Rib</td>
<td>–</td>
<td>20 (5.6%)</td>
</tr>
<tr>
<td>Inflammatory lesion</td>
<td>–</td>
<td>17 (4.7%)</td>
</tr>
<tr>
<td>Bronchial wall</td>
<td>–</td>
<td>19 (5.3%)</td>
</tr>
</tbody>
</table>

missed 7 (10.0%) Lung-RADS 3 nodules and 1 (3.0%) Lung-RADS 4 nodule, that potentially could be malignant. Reasons for detection errors could be a high workload, various levels of observer alertness, fatigue, and experience of readers. In contrast, the CAD system is not affected by subjective factors, and its reading is standardized, consistent, and fast. Our DL-CAD system only missed 4 (5.7%) Lung-RADS 3 nodules, and none of the Lung-RADS 4 nodules. In clinical practice, nodules smaller than 6 mm tend to be overlooked due to its small size and complex background (bronchi and vasculature) in the slice. The DL-CAD system still found 86.3% of these, which is significantly higher than the sensitivity of double reading (58.9%, P = 0.001). This indicates that the DL-CAD system may have the ability to provide assistance with complementary detection results for radiologists. Despite the high sensitivity for lung nodule detection, DL-CAD systems generally suffer from a higher false-positive rate compared to radiologists [14,39]. Three-hundred-and-fifty-nine false positives were found by our DL-CAD system on 360 scans, whereas double reading resulted in only thirteen FP findings. One of the most common misclassified structures in our system was pulmonary vasculature (45%). This issue might be solved by using the method proposed by Filho et al. [40] who applied 3D shape analysis to reduce the number of false positives. In the future, we will apply this method to further improve our model. In addition, the DL-CAD system identified some nodules smaller than 4 mm. If these clinically irrelevant nodules are excluded, the FP rate can be further reduced. Therefore, we need to establish a more accurate nodule segmentation algorithm in the future, in order that small nodules are effectively filtered out.

Currently, there are several factors that limit the widespread use of CAD systems in clinical practice. One is that the system often has either a low sensitivity or a high FP rate for lung nodule detection. For example, in the study conducted by Jacobs et al. the CAD had a relatively low false positive rate of 3.1/scan, but only 82% of the present nodules were detected [38]. Therefore, experienced radiologists might prefer to view scans by themselves rather than use the CAD. A study by Cascarino et al. based on 84 LDCT scans had a high sensitivity of 97%, but the FP number was as high as 6 per scan [41]. This FP rate will lead to extra effort for radiologists to check the scans again and discriminate between nodules and non-nodules. For many CAD systems, this trade-off still needs to be optimized. The second challenge before widespread implementation is the lack of standard clinical data to evaluate the CAD performance. The CAD sensitivity for nodule detection can be affected by the scanning and reconstruction protocol used; this may influence nodule size, nodule density and nodule position as determined by the CAD system. These factors were found to cause a large variation in performance of different CAD systems, with sensitivity ranging from 38% to 100% and a FP rate ranging from 1.0 to 15.1 per scan [11,22-31,42-45]. Hence, it is difficult to compare the performance of published CAD systems without a standard database. In recent years, the public LIDC/IDRI dataset collected from seven academic centers has been widely used to facilitate the validation of CAD systems for lung nodule detection [20,21]. Different from previous CAD systems, our DL-CAD was developed based on the maximum intensity projection that is used by radiologists as one of the clinical routine procedures for lung nodule detection. This provides a new insight, namely that deep learning could accurately and effectively find nodules using the same clinical method as radiologists. However, the LIDC/IDRI dataset on which our DL-CAD was developed mainly consists of regular-dose CT scans, while in lung cancer screening LDCT protocols are used with higher noise levels. Thus, results from CAD systems based on the aforementioned dataset might not be extrapolated to a screening setting. This was the motivation for the current study.

The study has some limitations. First, external evaluation was based only on single-center data. The generalizability of the DL-CAD system should be further assessed on data from various centers and CT scan protocols/vendors. Second, we used 4 mm as the cutoff of relevant nodules in our study. This modified threshold of LungRADS1.1 will lead to nodules < 4 mm being classified as irrelevant nodules. When including smaller nodules in the Lung-RADS 2 category, more false positive results are expected for the DL-CAD, and lower sensitivity for the double reading. Third, we used one degree of suspicion f for the model, this was primarily for the purposes of comparing reader and model performance. Degree of suspicion f affects the sensitivity and FP rates. Pursuing high sensitivity will lead to a high FP rate, and radiologists will need to spend more time to identify the nodules. It is still an ongoing area of research to evaluate costs and outcomes to properly tradeoff between sensitivity and FP rates. Last, manual measurement of nodules was applied in this study which may cause inter-observer differences, therefore, semi-automatic volume measurement needs further research in the future.

In conclusion, our DL-CAD system achieved a higher sensitivity in nodule detection compared to double reading, with an acceptable false-positive rate (1.0 FP/scan) in a low-dose chest CT lung cancer screening dataset. The use of the DL-CAD system may assist radiologists to achieve a higher nodule detection rate in LDCT lung cancer screening.

CRediT authorship contribution statement


Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Matthijs Oudkerk has a financial interest in the company iDNA B.V. for Xray real-time monitoring software].
Sunyi Zheng is an advisor to IDNA B.V. All other authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Acknowledgments

This work was supported by a grant from Ministry of Science and Technology of the People’s Republic of China, National Key R&D Program of China (no.2018YFC1315600 no.2017YFC1308700 and no.2016YFE0103000), the Royal Netherlands Academy of Arts and Sciences (grant number. PSA_SA BD_01), and the University Center Groningen PhD Scholarship program.

References