frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 07 December 2022
DOI 10.3389/fpubh.2022.1071673

OPEN ACCESS

EDITED BY
Yanwu Xu,
Baidu, China

REVIEWED BY
Muhammad Attique Khan,
HITEC University, Pakistan
Shanshan Wang,
Shenzhen Institutes of Advanced Technology (CAS), China
Jianjiang Feng,
Tsinghua University, China

*CORRESPONDENCE
Xiangfeng Meng
mengxiangfeng@nifdc.org.cn
Jiage Li
lijiage@nifdc.org.cn

SPECIALTY SECTION
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

RECEIVED 16 October 2022
ACCEPTED 21 November 2022
PUBLISHED 07 December 2022

CITATION
Wang H, Tang N, Zhang C, Hao Y, Meng X and Li J (2022) Practice toward standardized performance testing of computer-aided detection algorithms for pulmonary nodule. *Front. Public Health* 10:1071673. doi: 10.3389/fpubh.2022.1071673

Practice toward standardized performance testing of computer-aided detection algorithms for pulmonary nodule

Hao Wang[1], Na Tang[2], Chao Zhang[1], Ye Hao[1], Xiangfeng Meng[1]* and Jiage Li[1]*

[1]Division of Active Medical Device and Medical Optics, Institute for Medical Device Control, National Institutes for Food and Drug Control, Beijing, China, [2]School of Bioengineering, Chongqing University, Chongqing, China

This study aimed at implementing practice to build a standardized protocol to test the performance of computer-aided detection (CAD) algorithms for pulmonary nodules. A test dataset was established according to a standardized procedure, including data collection, curation and annotation. Six types of pulmonary nodules were manually annotated as reference standard. Three specific rules to match algorithm output with reference standard were applied and compared. These rules included: (1) "center hit" [whether the center of algorithm highlighted region of interest (ROI) hit the ROI of reference standard]; (2) "center distance" (whether the distance between algorithm highlighted ROI center and reference standard center was below a certain threshold); (3) "area overlap" (whether the overlap between algorithm highlighted ROI and reference standard was above a certain threshold). Performance metrics were calculated and the results were compared among ten algorithms under test (AUTs). The test set currently consisted of CT sequences from 593 patients. Under "center hit" rule, the average recall rate, average precision, and average F_1 score of ten algorithms under test were 54.68, 38.19, and 42.39%, respectively. Correspondingly, the results under "center distance" rule were 55.43, 38.69, and 42.96%, and the results under "area overlap" rule were 40.35, 27.75, and 31.13%. Among the six types of pulmonary nodules, the AUTs showed the highest miss rate for pure ground-glass nodules, with an average of 59.32%, followed by pleural nodules and solid nodules, with an average of 49.80 and 42.21%, respectively. The algorithm testing results changed along with specific matching methods adopted in the testing process. The AUTs showed uneven performance on different types of pulmonary nodules. This centralized testing protocol supports the comparison between algorithms with similar intended use, and helps evaluate algorithm performance.

KEYWORDS

computer-aided detection (CAD), algorithm testing, pulmonary nodule, test set, data curation

Introduction

Lung cancer has become the most common malignant tumor that threatens human health (1). Pulmonary nodules are common imaging signs in the early stage of lung cancer. Early detection of pulmonary nodules and timely medical intervention can improve the survival rate of patients (2), CT screening provides an effective method for early diagnosis, thereby accumulating tremendous amount of CT images for radiologists to read (3).

CAD algorithm for pulmonary nodules may help assist clinical decisions, and improve clinical work efficiency (4). Usually, the common function of a CAD system is to detect the location of the lesions (5). Afterwards, lesion type classification and lesion size measurement may also be covered. Therefore, lesion detection is fundamental and important. Many researchers have devoted to improving or innovating Artificial Intelligence (AI) -enabled algorithms for pulmonary nodule detection to improve the detection performance of the algorithms (6–21). Public datasets such as LIDC-IDRI (22), LNDb (23), ANODE09 (24) supported algorithm competitions and researches (25–37), which provided insights on how to arrange algorithm testing. However, the procedure of algorithm competitions is significantly different from product verification and validation, which should provide high quality evidence for regulation. In algorithm competitions, the training data and testing data may come from the same dataset (22) and have similar features. The annotation process may be conducted by clinicians from a limited number of hospitals, which may not reflect wide consensus of medical community. This situation may decrease the comparability of algorithm testing results among different labs. More effort is thus needed to improve algorithm testing procedure. And for relative research, details of the matching process between reference standard and algorithm predicted ROIs were not sufficiently described in the literature, but different matching methods may lead to differences in performance test results which limit the comparability among different algorithms.

AI-enabled CAD products for pulmonary nodules have been developed and marketed in many countries (38–40), and most of them are in the form of software as a medical device (SaMD). While stakeholders are interested to compare products and understand their quality, the verification and validation of such products is often conducted by manufacturers individually. Currently, there are differences in the performance metrics and verification methods claimed by different manufacturers, resulting in a lack of comparability between algorithms (41). There is also a lack of understanding of the common quality characteristics of these algorithms. Recently, standardization organizations start to establish the framework to conduct algorithm performance testing. It would be necessary to gain practical experience (42, 43).

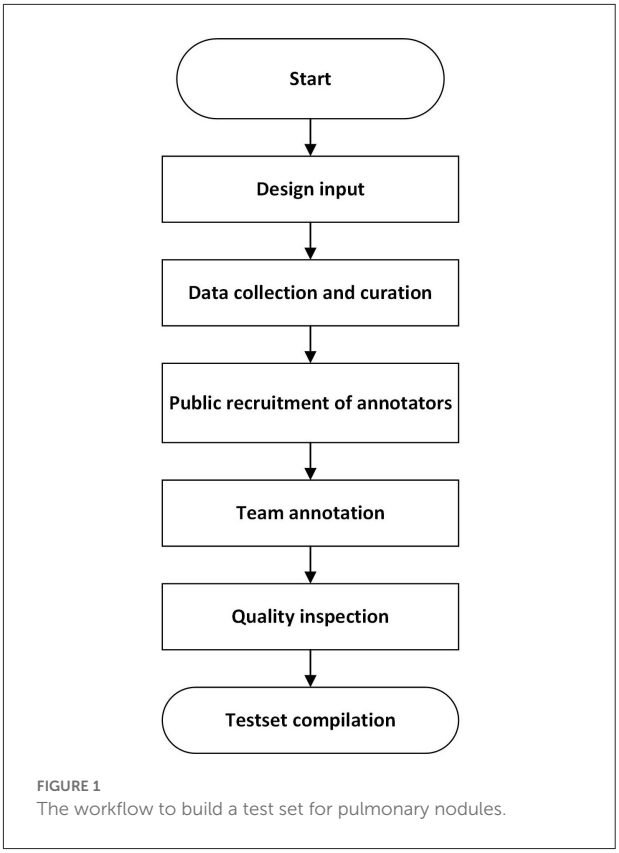

FIGURE 1
The workflow to build a test set for pulmonary nodules.

In this paper, a protocol is proposed to build datasets for algorithm testing from the perspective of third party and regulation. Three matching methods [center distance (44), center hit (45), area overlap (46)] are applied to test the detection performance of AUTs provided by ten different developers. The overall performance of ten AUTs and the differences in algorithm performance under different matching methods are analyzed. Algorithm errors are also analyzed to further understand the quality features of the AUTs. This work is aimed at establishing prototypes for standardized verification of such products and promote quality control. It may provide experience for development of technical standards on CAD products.

Materials and methods

Test set

The construction of a test set followed the procedure in Figure 1, which referred to the strategy in literature (47).

The design input clarified data collection requirement (CT equipment, imaging parameters such as tube voltage, slice thickness, slice spacing and reconstruction algorithm), data diversity (patient spectrum, geological distribution of data collection sites, proportion of different types of nodules) and

TABLE 1 The proportion of different types of nodules.

Types	Proportion (%)
Solid nodules	42.66
Part-solid nodules	5.06
Pure ground-glass nodules	19.58
Calcified nodules	7.40
Pleural nodules	23.06
Pleural calcified nodules	2.24

TABLE 2 The proportion of different sizes of nodules.

Sizes (diameter/mm)	Proportion (%)
<4	69.91
[4, 6)	19.63
[6, 10)	7.28
≥10	3.18

rules for unique data identification. More details can refer to literature (47).

CT images of patients with pulmonary nodules were collected retrospectively under local ethical approval and patient privacy protection requirements. They are stored as Dicom (Digital Imaging and Communications in Medicine) files. A data cleaning procedure was conducted to ensure the integrity and validity of CT images. Cases with discontinuous imaging sequences, missing slices, unreadable files, problematic field of view and irrelevant imaging position were excluded. Further examination was conducted to remove data which were duplicated internally or externally and ensure the uniqueness of each image sequence.

Annotation was conducted by groups of radiologists on a custom-built annotation software. CT sequences can be displayed at different angles. Multiple window level and width settings are provided. Annotators were asked to label the location (center of the bounding box), boundary and type of pulmonary nodules on each slice (usually cross-sectional view). The size of the nodule was also recorded, containing long diameter, short diameter, average diameter, length and width of the bounding box. The outputs were exported as csv files, which were used as reference standard during algorithm testing.

Annotators were publicly recruited through a qualification exam, which evaluated annotators' skills to detect and segment pulmonary nodules on 20 thoracic CT sequences in comparison with annotation results from a high-level expert panel. One hundred and eighty-five candidates from 112 hospitals participated in the test. The passing criteria is: precision >0.8, recall >0.8 and Dice coefficient >0.8. 24 junior radiologists and 15 senior radiologists passed the exam and received training. They were from 25 hospitals in 13 provinces in China. The junior radiologists have been engaged in image reading service in tertiary hospitals for more than 5 years and have a title of resident doctor or above. Every three junior radiologists form a team, and the team leader is a deputy chief physician with more than 10 years of work experience. The senior experts are chief physicians or deputy chief physicians with more than 15 years of work experience. They provide final review and arbitration.

The annotation consists of three steps. First, every team of junior radiologists independently highlight pulmonary nodules on a batch of CT image sequences back-to-back, and then a computer program automatically evaluates the consistency of the detection results. If they are consistent, the output will be the union of the results marked by the three junior radiologists. If results are inconsistent on certain slices, such slices will be highlighted to remind senior experts. Second, the same team gives the classification labels for the detected nodules from the previous step. Third, the outputs are reviewed by the leader from another team and arbitrated by arbitration experts. In case there are controversial results, the arbitration experts will discuss and determine the final annotation result.

The dataset contains a total of 593 cases from 22 hospitals in 9 provinces in China, with a total of 6,109 nodules. These nodules were counted from the perspective of type and size. The proportion of nodules of different types and sizes was shown in Tables 1, 2. A range of CT scanner manufacturers and models was represented (38% of scans from seven different Siemens Definition, Sensation, and Emotion scanner models, 36% of scans from three different Philips Brilliance, iCT scanner models, 12% of scans from three different GE Medical Systems LightSpeed, BrightSpeed scanner models, 9% of scans from UIH uCT scanners, 3% of scans from Toshiba Aquilion scanners, 2% of scans from other scanner models). Tube voltage ranged from 100 to 150 kV (mean: 117.4 kV). Tube current ranged from 17 to 544 mA (mean:189.8 mA). Conventional and enhanced CT accounted for 67% and low-dose screening CT accounted for 33%. The in-plane pixel size ranged from 0.5 to 0.9 mm. Slice thicknesses included 0.625 mm (1.2%), 0.75 mm (4.7%), 0.8 mm (14.0%), 1 mm (34.1%), 1.25 mm (13.8%), 1.5 mm (1.3%), 2 mm (28.5%), 2.5–6 mm (2.9%). 72.1% of CT scans were reconstructed using standard algorithm and lung algorithm, and 27.9% were reconstructed using high frequency algorithm and bone algorithm.

The definition of various types of nodules is as follows (48):

(1) Solid nodules: focal increased density shadows with clear borders in the lung parenchyma with a circular or quasi-circular (sphere or sphere-like) boundary, and the bronchi and blood vessel edges in the lesions cannot be identified. The maximum long diameter of the nodule is ≤3 cm.
(2) Part-solid nodules (mixed ground-glass density nodules): focal increased density shadows with clear borders in the lung parenchyma, round or round-like (sphere or sphere-like). In some lesions, the bronchi and blood vessel

edges can be identified, and the maximum long diameter is ≤3 cm.

(3) Pure ground-glass nodules: focal increased density shadows with clear borders in the lung parenchyma with a circular or quasi-circular (sphere or sphere-like) boundary, and the edges of the bronchi and blood vessels in the entire lesion can be identified, with a maximum long diameter of ≤3 cm.

(4) Calcified nodules: circular or quasi-circular (sphere or sphere-like) complete calcium deposits in the lung parenchyma with clear boundaries, the maximum long diameter is ≤3 cm, and the CT value is usually above 100 HU.

(5) Pleural nodules and pleural plaques: Pleural nodules are round and round-like (sphere and sphere-like) or irregular focal increased density shadows originating from the pleura, often connected to the broad base of the pleura, the maximum long diameter ≤3 cm. Pleural plaques are irregular flat protrusions of the pleura that are localized and broad-based, with an irregular surface. Hereinafter referred to as pleural nodules.

(6) Pleural calcified nodules: round or round-like (sphere or sphere-like) complete calcium deposition foci with clear borders originating from the pleura, the maximum long diameter is ≤3 cm, and the CT value is usually above 100 HU.

More details and examples of pulmonary nodule annotation refer to an expert consensus (48).

The dataset is managed as a sequestered test set according to the IEEE 2801–2022 standard (49).

Algorithm under test (AUT)

10 AUTs provided by 10 different developers are tested in this study. The developers are from medical device industry. They all use deep convolutional neural network to perform lung nodule detection. The AUTs are generally developed by transfer learning based on developers' self-developed training sets. The pre-training models referred to Faster-RCNN, YOLO and other public available algorithms (50–52). 9 AUTs are designed to detect 2-D targets on each CT image. 1 AUT is designed to detect 3D targets. To protect confidential information, details on the algorithm architecture and weights are not disclosed. According to the intended use, 3 AUTs detect only pulmonary nodules (4 types in total), and 7 AUTs detect 4 types of pulmonary nodules and 2 types of pleural nodules (6 types in total). The output of the AUTs highlight the region of predicted nodule in the form of a bounding box (B-Box) on each slice.

During the test, each developer provided a server to install and run their AUT. The configuration of the server was determined by the developer. No internet connection was allowed. The test set was imported to the AUT through an external hard drive.

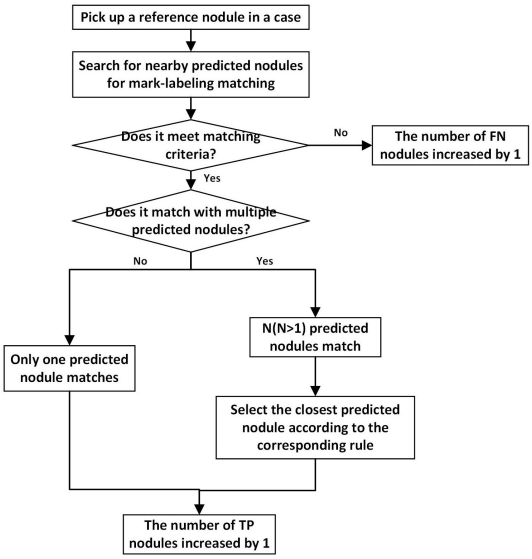

FIGURE 2
The flowchart depicting matching process for one reference nodule. Every reference nodule in each case goes through matching process. Unmatched predicted nodules result in a corresponding increase in the number of *FPs*.

Mark-labeling methods

To determine whether the predicted mark matches the reference standard (53) [this process is also called "mark-labeling" (54)], the flowchart of matching process is shown in Figure 2, and the specific details of mark-labeling methods are described as follows, and for a more intuitive schematic diagram, see Supplementary Figure A.1.

Center hit

In this study, the center point of the B-Box of the predicted nodule falls within the range of the corresponding reference nodule region as a successful detection match. If the predicted nodule has multiple slices, as long as one slice satisfies the above situation, the reference nodule can be recorded as successful detection. The successfully detected reference nodules and predicted nodules are recorded as true positive (*TP*) nodules. After all reference nodules participated in the mark matching process, the reference nodules and predicted nodules that are not successfully matched are recorded as false-negative (*FN*) and false-positive (*FP*) nodules, respectively. When a reference nodule meets the conditions for successful matching with

multiple predicted nodules at the same time, the distance between the predicted nodule and the largest slice of the reference nodule is further compared. The predicted nodule with the smallest distance is recorded as the *TP* nodule, and other predicted nodules need to participate in the matching process of the remaining reference nodules. If a predicted nodule can be matched with multiple reference nodules, the first reference nodule is selected to match the predicted nodule according to the matching order. With respect to the predicted nodule with multiple slices, the definition of *TP*, *FN*, *FP* nodules, and the matching of multiple reference nodules with a predicted nodule, are handled in the same way in the three mark-labeling methods.

Center distance

For center distance scheme, the distance between the center point of the predicted nodule B-Box and that of the corresponding reference nodule B-Box is compared with a threshold. Such a threshold is varied which is adaptively set to the average radius of each reference nodule under matching, which is the maximum of one quarter of the sum of the long and short diameters across slices. If the distance is less than threshold, reference nodule is considered successfully detected. When a reference nodule meets the conditions for successful matching with multiple predicted nodules at the same time, their distance is further compared, and the predicted nodule with the shortest center distance is recorded as the *TP* nodule, and the remaining predicted nodules participate in the matching process of other reference nodules.

Area overlap

For "area overlap" rule, it is stipulated that the proportion of the overlapping part of the predicted nodule B-Box area and the reference nodule B-Box area to the reference nodule B-Box area is greater than a threshold as a successful detection, and the threshold is set to 0.5 empirically. When looking for a matching predicted nodule for a certain reference nodule, if there are multiple predicted nodules that can satisfy the aforementioned description of successful matching, their proportions are further compared, and the predicted nodule with the highest proportion is considered to match the reference nodule.

Performance evaluation metrics

Recall, precision, and F_1 score were selected to evaluate algorithm performance. The recall reflects the proportion of the correct nodules detected by the algorithm to the reference nodules, that is, whether the algorithm can find out as many reference nodules as possible. The accuracy reflects the proportion of the correct nodules detected by the algorithm to the nodules predicted by the algorithm itself, that is, whether the algorithm can predict the reference nodules as accurately as possible. The F_1 score reflects an overall performance. For the definitions of *TP*, *FP* and *FN*, see the description in section Mark-labeling Methods. It can be seen that different mark-labeling methods may affect the judgement and quantities of *TP*, *FP*, and *FN*, thereby influencing the recall rate, precision, and F_1 score values.

Results

Comparison between three mark-labeling methods

Test the detection performance of ten algorithms according to the three matching methods specified in chapter Mark-labeling Methods, and comprehend the impact of different testing methods on the metrics. The overall performance of the ten algorithms is represented by the mean ± standard deviation (%) of the metric values (Figure 3).

In general, the algorithm performance under center hit and center distance is higher than the area overlap, and the test results of center hit and center distance are very close (the mean of recall, precision, and F_1 Score differ by 0.75, 0.5, and 0.57%, respectively). Compared with the center hit, the average value of the three indicators all drop by more than 10%. Among them, the recall dropped the most by 14.33%, the precision dropped by 10.44%, and the F_1 score dropped by 11.26%.

Further, different matching methods are regarded as different groups, and then the results under different matching methods are analyzed by the analysis of variance (ANOVA) and *t*-test. When $P > 0.05$, it is considered that there is no significant difference between groups. For recall (Figure 4A), significant differences are seen among the three groups ($P = 0.0033$), and through *t*-test analysis, there are significant differences between center hit and area overlap ($P = 0.0075$), and between center distance and area overlap ($P = 0.0054$). For center hit and center distance, there is no significant difference

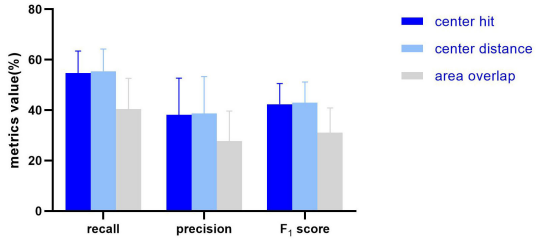

FIGURE 3
The histogram showing the average and standard deviation of the recall, precision, F_1 score of the 10 AUTs according to three mark-labeling methods.

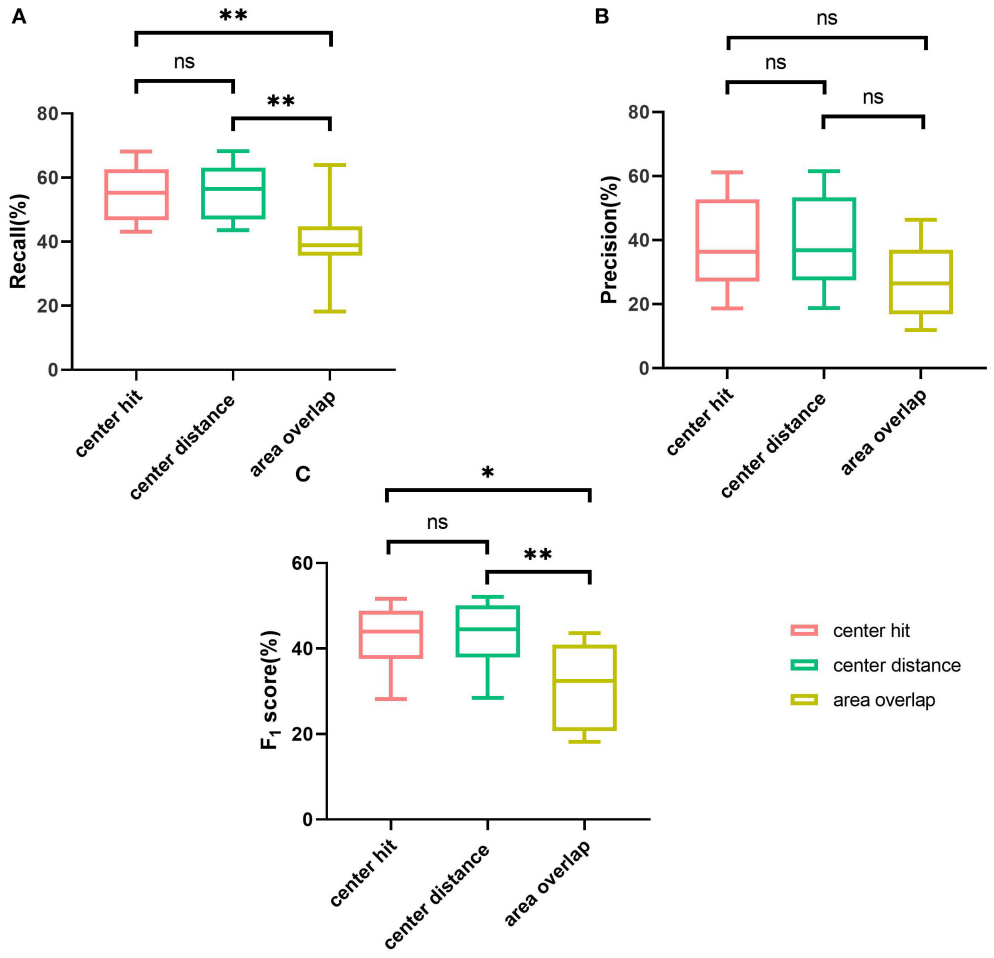

FIGURE 4
Comparison of three mark-labeling methods by boxplot. **(A)** The boxplot showing the range of the recall of the AUTs according to three mark-labeling methods. **(B)** The boxplot showing the range of the precision of the AUTs according to three mark-labeling methods. **(C)** The boxplot showing the range of the F1 score of the AUTs according to three mark-labeling methods. Note: **means *P* value is no more than 0.01, *means *P* value is no more than 0.05, ns means P value is greater than 0.05.

between two groups ($P = 0.8503$). Regarding the precision (Figure 4B), although the mean precision for the area overlap is lower than the other two matching methods, from the analysis of variance, there is no significant difference in the mean precision under the three matching methods ($P = 0.152$). Also, there is no significant difference between any two groups under the *t*-test ($P = 0.9394$ for center hit and center distance, $P = 0.0954$ for center hit and area overlap, $P = 0.0829$ for center distance and area overlap). As for F_1 score (Figure 4C), the same as the variance analysis of recall, significant differences are seen among the three groups ($P = 0.0080$). And similarly, *t*-test results for any two groups show statistically significant differences between area overlap and center hit ($P = 0.0120$) and between area overlap and center distance ($P = 0.0090$). In the meanwhile, *P*-value between center hit and center distance is 0.8782 which means there is no significant difference between two groups.

In order to have a more intuitive impression of the specific performance differences of different AUTs under the three matching methods, the relative differences of ten AUTs under the three matching rules are further explored from the perspective of *TP* nodules. Specifically, test results under the "center hit" rule are chosen as the baseline. The relative changes for the other two matching rules are calculated, respectively (Figure 5). For example, the calculation formula of the relative difference (RD) of the recall under "center distance" rule is as follows.

$$RD_{dist} = \frac{TP_{dist} - TP_{hit}}{TP_{hit}} \quad (1)$$

In Figure 5, the ten AUTs show varying degrees of difference for the three matching methods. The performance of the 90% (9/10) AUTs under "area overlap" rule is relatively lower than the other two matching methods. There is one AUT (AUT 5)

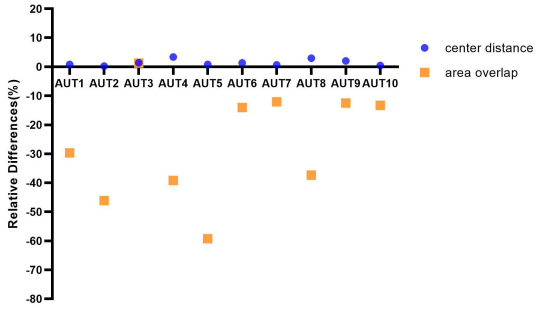

FIGURE 5
The graph depicting ten AUTs' relative differences of center distance and area overlap with respect to center hit, respectively.

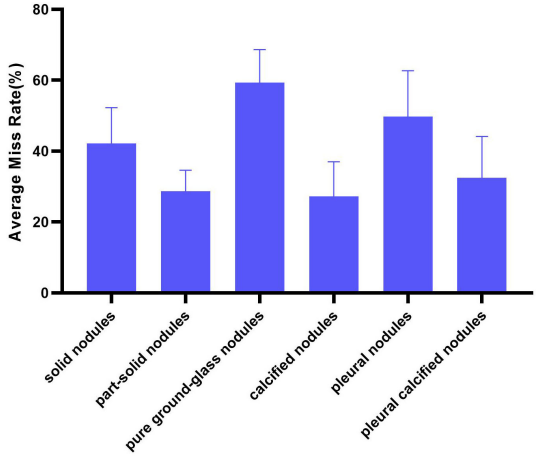

FIGURE 6
The average miss rate of AUTs on the six types of nodules under the center hit.

that is particularly unaccustomed to the test method of area overlap, with a relative drop of as high as 59%. And it is worth mentioning that there is also an AUT (AUT 3) that maintains a high degree of consistency under three mark-labeling rules, and its performance under "area overlap" rule is slightly higher than that under "center hit" rule, which is fundamentally different from other AUTs. In addition, the performance of all AUTs under "center distance" rule is slightly higher than that under "center hit" rule, and the relative increase is as high as 3.34%. All AUTs show relatively optimal test results under "center distance" rule.

Analysis of FN nodules

After comparing the differences in the evaluation metrics of AUTs under three mark-labeling rules objectively, in order to assess the quality of AUTs and try to find out the reasons for the general performance of the AUTs using this sequestered test set, further research is carried out to count the erroneous results of ten AUTs under "center hit" rule. The type and size of nodules may affect the performance test results, which were taken into account in experimental design. Focus on the reference nodules that were not successfully detected, namely *FN* nodules, features of *FN* nodules in the two dimensions of type and size were observed, and the common quality characteristics of AUTs were investigated. In addition, we selected the images of FN and TP nodules of four types of nodules with stronger medical significance to indicate the nodules that are difficult to detect and relatively easy to detect by the algorithms, see Supplementary Figures A.2–A.5.

Six types of nodules

Divide the number of each type of *FN* nodules by the number of this type of reference nodules as the miss rate of AUT on this type of nodules. Then, the miss rate of all AUTs for a certain type of nodules is expressed in the form of mean ± standard deviation (%) to illustrate the overall situation of AUTs' miss detection of this type of nodules (Figure 6). And overall miss situation of different nodule types can be compared.

In terms of nodule types, AUTs can detect part-solid nodules and calcified nodules more accurately than other types, with an average miss rate of 28.64 and 27.28%, respectively. Especially for part-solid nodules, the fluctuation of detection performance of different AUTs on such nodules is also the smallest (standard deviation is 5.96%). For these AUTs, the most challenging nodule type is pure ground-glass nodule, with an average miss detection rate of over 50% (as high as 59.32%, twice as many as solid nodules), followed by pleural nodules and solid nodules, with an average miss rate of 49.80 and 42.21%.

Combine size and type

The dimension of size was added to further refine the types that are likely to be missed in different size ranges. The specific method is to plot the miss rate of each type of nodules against the range of average diameter, including 4 bins [<4, (4, 6), (6, 10), ≥10 mm] for each AUT, respectively. Within each bin, nodule types with highest miss rate for each AUT are counted and shown as a heatmap (Figure 7).

In general, among all nodule diameter ranges, pure ground-glass nodules are the most likely to be missed. For nodules <6 mm in diameter, all AUTs have the highest miss detection rate for pure ground-glass nodules among all types. In the diameter range from 6 to 10 mm, pure ground-glass nodules are the most difficult type to be detected for 7 AUTs. Part-solid

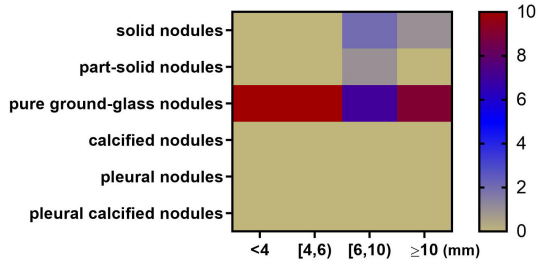

FIGURE 7
Heatmap depicting nodule types and diameters most likely to be missed by 10 algorithms. The x axis represents nodule size range. The y axis represents nodule type. For example, when the nodule diameter is between 6 and 10 mm, 7 AUTs are most likely to miss pure ground-glass nodules; 2 AUTs are most likely to miss part-solid nodules; 1 AUT is most likely to miss solid nodules.

nodules are the most difficult type to be detected for one AUT. Solid nodules in this size range are most difficult for two AUTs. For large nodules larger than 10 mm in size, 90% (9/10) of AUTs need to continue to make efforts to detect pure ground-glass nodules, there is also an AUT that needs to detect large solid nodules more accurately.

Discussion

With the development of AI technology, more computer-aided detection products for pulmonary nodules may enter the market in the future. While verification and validation activities are mainly conducted by manufacturers, it is important for regulators and public stakeholders to understand the algorithm performance in a more comparable manner. From the perspective of third-party testing, it may be helpful to explore a pathway to build test set and compare products directly, objectively and quantitatively.

This study demonstrated a centralized method to build test set and conduct algorithm performance testing. Data was randomly sampled from diverse hospitals and regions according to the design input. Annotators were recruited publicly through qualification exams and randomly grouped to conduct annotation. The whole process relies on the same standardized procedure. The workflow is different from conventional multicenter study and may decrease variation of annotation among different hospitals.

Using the same test set, different AUTs are tested and compared quantitatively. The testing results indicated that specific mark-labeling method would affect interpretation of algorithm performance. This study shows that, among the three mark-labeling methods adopted in the experiments, ten AUTs under the method of center distance showed the highest precision and recall, as performance metrics of computer-aided detection. The "center hit" method showed intermediate results. The "area overlap" method showed the worst results. From the perspective of clinical application, the mark-labeling methods are associated with follow-up operations after image analysis. For example, if robotically assisted surgery needs information from computer-aided detection of pulmonary nodules, it would be necessary for the AUT to export the position of predicted nodule, so the "center distance" approach is favorable. Under the context of radiotherapy, the "area overlap" approach is preferred. It may be helpful to choose mark-labeling rule according to the intended use and usage scenarios of the product.

In this study, it seems that the average recall of the ten AUTs is lower than results reported in other literature (25–27). There are several underlying reasons. First, the test set used in this study is independent and isolated from the training or tuning process of AUTs. In other literature, AUTs may be trained and tuned on a subset of a large data set and then tested on another subset. The correlation between training and testing data may facilitate the model to achieve better testing results. Therefore, the test set in this paper seems more challenging and helps reflect the generalizability of algorithm. Second, the annotation of the test set is based on a centralized and relatively strict procedure, which requires intense support from experienced radiologists. For developers of the AUTs, however, their training sets and tuning sets are prepared spontaneously. Annotation activities and results may have difference.

For the evaluation metrics, the FROC curve was not adopted (other researchers may have chosen) because each developer has fixed the optimal detection threshold before providing the algorithm. In order to be more in line with the actual clinical use scenarios of the product, evaluation metrics such as recall on its specified detection threshold were calculated.

Based on the above results, the algorithm errors were further compared among different AUTs and analyzed the trend. False-negative nodules are chosen as the target, and we observed what type and size of nodules are more likely to be missed by AUTs, resulting in a lower recall by the algorithm. Based on the research results, manufacturers are encouraged to pay attention to the accuracy in the detection of solid nodules, pure ground-glass nodules and pleural nodules in the development stage. Especially for pure ground-glass nodules of various sizes, the low detection accuracy of such nodules is the main reason for the poor performance of AUTs. It is necessary to consider taking related technical methods (such as increasing its proportion in the training set, using better tuned algorithms, etc.) to improve the detection performance of the product on these three types of nodules and small nodules. At the same time, how to reduce the number of false positive nodules also needs to be considered in the development process, and there is also a trade-off between recall and precision.

In summary, this paper systematically described a standardized workflow to build test sets and conduct algorithm testing in a third-party manner. The test set construction covered data collection, data curation, annotation, annotator management, which followed data quality management standards (49). If organizations follow this workflow, the comparability of data sets may be improved, since each step is well defined and the rule is relatively transparent. The algorithm testing section compared both performance metrics and trend of algorithm errors among different products, which also provided useful evidence to enrich the perspective of product evaluation and comparison.

There are several limitations in this study. First, while three mark-labeling rules are compared, it is difficult to assign the specific threshold of "center distance" rule or "area overlap" rule. The thresholds used in the experiments are selected empirically. It may not represent the requirement from clinical users' perspective. More discussion on the threshold selection should be made in the future. Consensus is needed to propose clear requirement on how computer-aided detection products should present the algorithm output. Second, this study compared different algorithm outputs on a sequestered test set in a black box manner, providing barely no clue to evaluate the process of algorithm design. It is difficult to further discuss the advantages and disadvantages of model design according to such test results, since there may be implicit discrepancies in the training sets, parameter settings, and training methods among developers in the research and development stage. Third, efficiency is not compared among different AUTs since they operated on separate servers that were provided by developers. Since medical device manufacturers would claim their own requirement on computation resource during premarket application, it may be helpful to define a benchmark to further evaluate algorithm efficiency based on consensus of manufacturers and clinical users.

In the future, more work will be conducted to evaluate the quality and comparability of test sets, which may further support standardization of testing methods and provide technical reference for the regulation of such products.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

HW and NT performed algorithm testing experiments and wrote the manuscript draft. CZ and YH analyzed the data. XM and JL contributed to the data collection. All authors read, contributed to the research design, and approved the final manuscript.

Funding

This research is sponsored by National Key Research and Development Program of China with a Grant Number of 2019YFB1404805.

Acknowledgments

The authors want to show their gratitude to all annotators participating the study and technical support from Chinese Society of Radiology, especially Dr. Kai Liu, Prof. Li Fan, Prof. Yi Xiao, and Prof. Shiyuan Liu from Shanghai Changzheng Hospital.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2022.1071673/full#supplementary-material

References

1. Gao S, Li N, Wang S, Zhang F, Wei W, Li N et al. Lung Cancer in People's Republic of China. *J Thorac Oncol.* (2020) 15:1567–76. doi: 10.1016/j.jtho.2020.04.028

2. Aberle DR, Adams AM, Berg CD et al. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med.* (2011) 365:395–409. doi: 10.1056/NEJMoa1102873

3. Binczyk F, Prazuch W, Bozek P, Polanska J. Radiomics and artificial intelligence in lung cancer screening. *Transl Lung Cancer Res.* (2021) 10:1186–99. doi: 10.21037/tlcr-20-708

4. Zhang G, Jiang S, Yang Z, Gong L, Ma X, Zhou Z et al. Automatic nodule detection for lung cancer in CT images: a review. *Comput Biol Med.* (2018) 103:287–300. doi: 10.1016/j.compbiomed.2018.10.033

5. Weikert T, Akinci D'Antonoli T, Bremerich J, Stieltjes B, Sommer G, Sauter AW. *Evaluation of an AI-Powered Lung Nodule Algorithm for Detection and 3D Segmentation of Primary Lung Tumors*. Hoboken, NJ: Contrast Media & Molecular Imaging (2019). doi: 10.1155/2019/1545747

6. Gu Y, Lu X, Yang L, Zhang B, Yu D, Zhao Y et al. Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs. *Comput Biol Med.* (2018) 103:220–31. doi: 10.1016/j.compbiomed.2018.10.011

7. Jin H, Li Z, Tong R, Lin L. A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. *Med Phys.* (2018) 45:2097–107. doi: 10.1002/mp.12846

8. Kim BC, Yoon JS, Choi JS, Suk HI. Multi-scale gradual integration CNN for false positive reduction in pulmonary nodule detection. *Neural Netw.* (2019) 115:1–10. doi: 10.1016/j.neunet.2019.03.003

9. Zuo FZ, He Y. An embedded multi-branch 3D convolution neural network for false positive reduction in lung nodule detection. *J Digit Imag.* (2020) 33:846–57. doi: 10.1007/s10278-020-00326-0

10. Majidpourkhoei R, Alilou M, Majidzadeh K. Babazadehsangar A. A novel deep learning framework for lung nodule detection in 3d CT images. *Multimed Tools Appl.* (2021) 80:30539–55. doi: 10.1007/s11042-021-11066-w

11. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* (2019) 25:954–961. doi: 10.1038/s41591-019-0447-x

12. Zhang J, Xia Y, Zeng H, Zhang Y. NODULE: combining constrained multi-scale LoG filters with densely dilated 3D deep convolutional neural network for pulmonary nodule detection. *Neurocomputing.* (2018) 317:159–67. doi: 10.1016/j.neucom.2018.08.022

13. Han C, Kitamura Y, Kudo A, Ichinose A, Rundo L, Furukawa Y et al. Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. In: *2019 International Conference on 3D Vision (3DV).* (2019). p. 729–37. doi: 10.1109/3DV.2019.00085

14. Shi Z, Hao H, Zhao M, Feng Y, He L, Wang Y et al. A deep CNN based transfer learning method for false positive reduction. *Multimed Tools Appl.* (2019) 78:1017–33. doi: 10.1007/s11042-018-6082-6

15. Huang X, Sun W, Tseng TL Li C, Qian W et al. Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks. *Comput Med Imag Graphics.* (2019) 74:25–36. doi: 10.1016/j.compmedimag.2019.02.003

16. Tan M, Wu F, Yang B, Ma J, Kong D, Chen Z et al. Pulmonary nodule detection using hybrid two-stage 3D CNNs. *Med Phys.* (2020) 365:395–409. doi: 10.1002/mp.14161

17. Hesamian MH, Jia W, He X, Wang Q, Kennedy PJ. Synthetic CT images for semi-sequential detection and segmentation of lung nodules. *Appl Intell.* (2021) 51:1616–28. doi: 10.1007/s10489-020-01914-x

18. Zhu L, Gao J. Adoption of computerized tomography images in detection of lung nodules and analysis of neuropeptide correlative substances under deep learning algorithm. *J Supercomput.* (2021) 77:7584–97. doi: 10.1007/s11227-020-03538-x

19. Wozniak M, Połap D, Capizzi G, Sciuto GL, Kośmider L, Frankiewicz K. Small lung nodules detection based on local variance analysis and probabilistic neural network. *Comput Methods Programs Biomed.* (2018) 161:173–80. doi: 10.1016/j.cmpb.2018.04.025

20. Su Y, Li D, Chen X. Lung nodule detection based on faster R-CNN framework. *Comput Methods Programs Biomed.* (2021) 200:105866. doi: 10.1016/j.cmpb.2020.105866

21. Zheng S, Cornelissen LJ, Cui X, Jing X, Veldhuis RN, Oudkerk M et al. Deep convolutional neural networks for multiplanar lung nodule detection: Improvement in small nodule identification. *Med Phys.* (2020) 48:733–44. doi: 10.1002/mp.14648

22. Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys.* (2011) 38:915–31. doi: 10.1118/1.3528204

23. Pedrosa J, Aresta G, Ferreira C, Atwal G, Phoulady HA, Chen X et al. LNDb challenge on automatic lung cancer patient management. *Med Image Anal.* (2021) 70:102027. doi: 10.1016/j.media.2021.102027

24. Van Ginneken B, Armato III SG, de Hoop B, van Amelsvoort-van de Vorst S, Duindam T, Niemeijer M et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Med Image Anal.* (2010) 14:707–22. doi: 10.1016/j.media.2010.05.005

25. Zhu W, Liu C, Fan W, Xie X. DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV).* Lake Tahoe, NV: IEEE (2018). pp. 673–81. doi: 10.1109/WACV.2018.00079

26. Liu J, Cao L, Akin O, Tian Y. *3DFPN-HS2: 3D Feature Pyramid Network Based High Sensitivity and Specificity Pulmonary Nodule Detection*. arXiv e-prints. arXiv: 1906.03467 (2019). doi: 10.1007/978-3-030-32226-7_57

27. Xie ZL. 3D Region Proposal U-Net With Dense and Residual Learning for Lung Nodule Detection. In: *LUNA*. Nijmegen (2017). Available online at: http://rumc-gcorg-p-public.s3.amazonaws.com/f/challenge/71/fbf2ae95-0781-45b2-aa5e-160b1eb80905/20171011_032308_report_Xie.pdf

28. Zheng S, Guo J, Cui X, Veldhuis RN, Oudkerk M, Van Ooijen PM. Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection. *IEEE Trans Med Imag.* (2020) 39:797–805. doi: 10.1109/TMI.2019.2935553

29. Hu Z, Muhammad A, Zhu M. Pulmonary nodule detection in CT images via deep neural network: nodule candidate detection. In: *Proceedings of 2nd International Conference Graphics Signal Processing.* (2018). p. 79–83. doi: 10.1145/3282286.3282302

30. Xie H, Yang D, Sun N, Chen Z, Zhang Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recogn.* (2019) 85:109–19. doi: 10.1016/j.patcog.2018.07.031

31. Wang B, Qi G, Tang S, Zhang L, Deng L, Zhang Y. Automated pulmonary nodule detection: high sensitivity with few candidates. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* (2018). p. 759–67. doi: 10.1007/978-3-030-00934-2_84

32. Gruetzemacher AG. D Paradice, 3D deep learning for detecting pulmonary nodules in CT scans. *J Am Med Inf Assoc.* (2018) 25:1301–10. doi: 10.1093/jamia/ocy098

33. Wen C, Hong M, Yang X, Jia J. Pulmonary nodule detection based on convolutional block attention module. In: *2019 Chinese Control Conference (CCC).* (2019). p. 8583–7. doi: 10.23919/ChiCC.2019.8865792

34. Wang Q, Shen F, Shen L, Huang J, Sheng W. Lung nodule detection in CT images using a raw patch-based convolutional neural network. *J Digit Imag.* (2019) 32: 971–9. doi: 10.1007/s10278-019-00221-3

35. Rafael-Palou X, Aubanell A, Bonavita I, Ceresa M, Piella G, Ribas V et al. *Re-Identification and Growth Detection of Pulmonary Nodules without Image Registration Using 3D Siamese Neural Networks*. arXiv preprint arXiv:1912.10525 (2019).

36. Gong L, Jiang S, Yang Z, Zhang G, Wang L. Automated pulmonary nodule detection in CT images using 3D deep squeeze-and-excitation networks. *Int J Comput Assisted Radiol Surg.* (2019) 14:1969–79. doi: 10.1007/s11548-019-01979-1

37. Jacobs C, van Rikxoort EM, Murphy K, Prokop M, Schaefer-Prokop CM, van Ginneken B. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. *Eur Radiol.* (2015). doi: 10.1007/s00330-015-4030-7

38. Food and Drug Administration [EB/OL]. Available online at: https://www.accessdata.fda.gov/cdrh_docs/pdf20/K201560.pdf (accessed April 18, 2012).

39. Food and Drug Administration[EB/OL]. Available online at: https://www.accessdata.fda.gov/cdrh_docs/pdf16/K162484.pdf (accessed April 18, 2012).

40. Food and DrugAdministration [EB/OL]. Available online at: https://www.accessdata.fda.gov/cdrh_docs/pdf20/K202300.pdf (accessed April 18, 2012).

41. Wang L, Wang H, Xia C, Wang Y, Tang Q, Li J, Zhou XH. Toward standardized premarket evaluation of computer aided diagnosis/detection products: insights from FDA-approved products. *Expert Rev Med Dev*. (2020) 17:899–918. doi: 10.1080/17434440.2020.1813566

42. PWI 62-3. *Artificial Intelligence/Machine Learning-enabled Medical Device-Performance Evaluation Process*. Geneva: International Electrotechnical Commission (2021).

43. PNW 62-411 ED1. *Testing of Artificial Intelligence/Machine Learning-enabled Medical Devices*. Geneva: International Electrotechnical Commission (2022).

44. Setio AA, Traverso A, De Bel T, Berens MS, Van Den Bogaard C, Cerello P 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

45. Li FLi, Doi K. Computerized detection of lung nodules in thin section CT images by use of selective enhancement filters and an automated rule-based classifier. *Acad Radiol.* (2008) 15:165–75. doi: 10.1016/j.acra.2007.09.018

46. Kim KG, Goo JM, Kim JH, Lee HJ, Min BG, Bae KT et al. Computer-aided diagnosis of localized ground-glass opacity in the lung at CT: Initial experience. *Radiology*. (2005) 237:657–61. doi: 10.1148/radiol.2372041461

47. Jin Z. Expert consensus on the construction and quality control of thoracic CT datasets for pulmonary nodules. *Chin J Radiol*. (2021) 55:104–10. doi: 10.3760/cma.j.cn112149-20200713-00915

48. Jin Z. Expert consensus on the rule and quality control of pulmonary nodule annotation based on thoracic CT. *Chin J Radiol*. (2019) 4:9–15. doi: 10.3760/cma.j.issn.1005-1201.2019.01.004

49. IEEE 2801-2022. *Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence*. New Jersy: IEEE Standard Association (2022).

50. Ren S, He K, Girshick R, Sun J, Faster R-CNN. Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* (2017) 39:1137–49. doi: 10.1109/TPAMI.2016.2577031

51. Redmon J, Farhadi A, YOLO9000: Better, Faster, Stronger. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (2017). p. 6517–25. doi: 10.1109/CVPR.2017.690

52. Valente IR, Cortez PC, Neto EC, Soares JM, de Albuquerque VH, Tavares JM. Automatic 3D pulmonary nodule detection in CT images: A survey. *Comput Methods Prog Biomed*. (2016) 124:91–107. doi: 10.1016/j.cmpb.2015.10.006

53. Kallergi GMC, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. *Med Phys*. (1999) 26:267–75. doi: 10.1118/1.598514

54. Petrick N, Sahiner B, Armato III SG, Bert A, Correale L, Delsanto S et al. Evaluation of computer-aided detection and diagnosis systems. *Med. Phys*. (2013) 40:87001. doi: 10.1118/1.4816310