Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Diagnostic Test Accuracy of Deep Learning Detection of COVID-19: A Systematic Review and Meta-Analysis

Temitope Emmanuel Komolafe, PHD, Yuzhu Cao, MS, Benedictor Alexander Nguchu, PHD, Patrice Monkam, MS, Ebenezer Obaloluwa Olaniyi, MS, Haotian Sun, MS, Jian Zheng, PHD, Xiaodong Yang, PHD

Rationale and Objective: To perform a meta-analysis to compare the diagnostic test accuracy (DTA) of deep learning (DL) in detecting coronavirus disease 2019 (COVID-19), and to investigate how network architecture and type of datasets affect DL performance.

Materials and Methods: We searched PubMed, Web of Science and Inspec from January 1, 2020, to December 3, 2020, for retrospective and prospective studies on deep learning detection with at least reported sensitivity and specificity. Pooled DTA was obtained using random-effect models. Sub-group analysis between studies was also carried out for data source and network architectures.

Results: The pooled sensitivity and specificity were 91% (95% confidence interval [CI]: 88%, 93%; I^2 = 69%) and 92% (95% CI: 88%, 94%; I^2 = 88%), respectively for 19 studies. The pooled AUC and diagnostic odds ratio (DOR) were 0.95 (95% CI: 0.88, 0.92) and 112.5 (95% CI: 57.7, 219.3; I^2 = 90%) respectively. The overall accuracy, recall, F1-score, LR^+ and LR^- are 89.5%, 89.5%, 89.7%, 23.13 and 0.13. Sub-group analysis shows that the sensitivity and DOR significantly vary with the type of network architectures and sources of data with low heterogeneity are (I^2 = 0%) and (I^2 = 18%) for ResNet architecture and single-source datasets, respectively.

Conclusion: The diagnosis of COVID-19 via deep learning has achieved incredible performance, and the source of datasets, as well as network architectures, strongly affect DL performance.

Key Words: Chest computed tomography; COVID-19; Diagnostic test accuracy; Deep learning; Meta-analysis.

Abbreviations: CI confidence interval, **COVID-19** coronavirus disease 2019, **CT** computed tomography, **DOR** diagnostic odds ratios, **GGOs** ground glass opacities, LR^+ positive likelihood ratio, LR^- negative likelihood ratio, **PRISMA** preferred reporting items for systematic reviews and meta-analyses, **QUADAS-2** quality assessment of diagnostic accuracy studies-2, **RE** random effect, **RT-PCR** reverse transcriptase-polymerase chain reaction, **SROC** summary receiver operating characteristic, **WHO** world health organization

INTRODUCTION

Coronavirus disease 2019 (COVID-19) outbreak which was officially reported in the Wuhan city of China in December 2019 (1), is now found in all countries of the world. The disease exponentially grew into a global pandemic by March 11, 2020 as declared by the World Health Organization (WHO) (2). Although China, where the first cases of the disease were reported, is gradually recovering from this global pandemic, most countries are still being ravaged by this lethal virus pneumonia. The report on the WHO global pandemic website on 2nd June 2021, shows the total confirmed cases, deaths, new cases and vaccinated as 170,812,850, 3,557,586, 371,489 and 1,581,509,628 respectively (3).

Despite the rollout of vaccines across the world, there are still new cases and new deaths recorded daily in some

Acad Radiol 2021; 28:1507–1523

From the School of Biomedical Engineering (Suzhou) (T.E.K.,Y.C., H.S.), Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230026, China. Department of Medical Imaging (T.E.K.,Y.C., H.S., J.Z., X.Y.), Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, 215163, China. Hefei National Lab for Physical Sciences at the Microscale and Centres for Biomedical Engineering (B.A.N.), University of Science and Technology of China, Hefei, 230026, China. EasySignal Group, Department of Automation (P.M.), Tsinghua University, Beijing 100084, China. Department of Biomedical Engineering (E.O. O.), Shenzhen University, Shenzhen, 518060, China. Jinhua Laboratory (X.Y.), Foshan, 528000, China. Received April 26, 2021; revised June 18, 2021; accepted August 12, 2021. **Address correspondence to:** X.Y. e-mail: xiaodong.yang@sibet.ac.cn

https://doi.org/10.1016/j.acra.2021.08.008

countries of the world. Most new cases reported are due to the second wave necessitating immediate and long-term solutions for early detection. This is crucial to the management of the disease to prevent the third wave due to the highly contagious nature of the disease. The gold standard diagnostic test for COVID-19 is the reverse transcriptase-polymerase chain reaction (RT-PCR) (4) but the time required for the result to be available is considerably long. This perceived shortcoming led to the development of a non-invasive assessment of COVID-19 patients. Radiologic assessment of the chest via plain chest radiography and chest computed tomography have been found useful in the management of COVID-19. Chest CT is capable of revealing some image features in patients with COVID-19 that do not show any detectable abnormalities on a plain radiograph (5). Radiologists' studies revealed that the imaging features of patients with the COVID-19 are bilateral, peripheral, multifocal ground-glass opacity, and consolidation, predominantly located at subpleural and peri-bronchovascular regions, were the typical features (1,5). However, other kinds of viral pneumonia can also mimic COVID-19 pneumonia thus making it difficult to differentiate (6).

The field of machine learning (ML) cuts across multiple statistics-based techniques useful for radiologists in disease diagnosis which complements the currently adopted deep learning (DL) approach (7). The incorporation of ML into deep learning and artificial intelligence (AI) has shown great potentials in assisting decision-making for assessing severity and prediction of clinical outcomes of disease in COVID-19 patients (8,9). Li *et al.* (10) conducted a systematic and meta-analysis review on machine learning diagnosis of COVID-19 on 151 published studies and reported the sensitivity and specificity of 92.5% and 97.9% respectively on the XGBoost model. Recently, Li et al (11) carried out a multi-reader study for the grading of COVID-19 in chest radiography and observed that the AI system improved radiologist performance. Since the deep learning technique has been found useful in the diagnosis of COVID-19 (12,13), combining radiologist interpretation with the DL approach gives a promising result for the detection of COVID-19 (13). To this effect, the potential use of deep learning suggests a better future in the clinical diagnosis of COVID-19 as supported by Islam *et al.* (14). Li *et al.* (13) performed a multi-center retrospective study using a deep learning COVID-19 detection neural network (COVNet) to extract visual features from volumetric CT scans for detection of COVID-19. Accurate detection of distinct features of COVID-19, other than those of community-acquired pneumonia (CAP) and other lung infections, was achieved (13). A study by Javo *et al.* (15) to test the diagnostic accuracy of convolutional neural network (ResNet-50) on public chest CT datasets revealed that while the diagnostic accuracy achieved by a deep learning model showed no significant difference with that of radiologists at rule-in thresholds, differences were significant at rule-out suggestive of better results of deep learning with public datasets (9). Moezzi *et al.* (16) summarized the evidence evaluated using the meta-analysis approach on prediction of the accuracy of AI assisted CT scanning for COVID-19 using 36 studies. The study compared DL, ML, and AI systems. The result shows that AI systems performed slightly better than their corresponding DL and ML counterparts which implies that the AI systems will be useful in identifying COVID-19 symptoms. This study did not consider the effect of training data or how the network architectures affect DL detection ability. A systematic analysis with meta-analysis of the effect of deep learning network architectures and data types will provide a means to bridge this evidence gap which is the aim of this systematic review. This systematic review and meta-analysis aimed to summarize, all the available evidence to quantitatively evaluate the diagnostic test accuracy (DTA) of a deep learning algorithm for detection of COVID-19 in chest CT. In doing so, the review provides crucial new information on how network architecture and data types affect the performance of the DL algorithm in COVID-19 diagnosis.

MATERIALS AND METHODS

This systematic review and meta-analysis was prospectively registered at PROSPERO with the registration number CRD: 42020223202 (17) The systematic review was performed by two independent reviewers (TEK and YC or PM and EOO using a well-established review protocol known as Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (18). The discrepancies between the two results were discussed by the two reviewers, and a more experienced third reviewer (XY or JZ) was consulted in case consensus was not reached.

We conducted a meticulous search that focused on deep learning diagnosis of COVID-19 using chest CT images patients who reported well-documented information on diagnosis accuracy test or at least 2 × 2 confusion matrix that is the sensitivity and specificity. Our search includes clinical trials, cohort, prospective, and retrospective studies based on deep learning detection of COVID-19. It is important to note that most studies fall in the retrospective studies because the nature of deep learning requires a large number of datasets, and all literature reviews were excluded.

Data Sources and Searches

PubMed, Inspec, Web of Science, and other biomedical databases were searched from inception with additional hand searched to unravel relevant literature from 1st January 2020 to 3rd December 2020. The same keywords were used for PubMed, Inspec, and Web of Science databases, which includes the following search terms: "computed tomography", "COVID-19″, "2019-novel corona-virus", "SARS-COV2", or "Diagnosis of COVID-19 based on Deep Learning or Artificial Intelligence (AI)". The complete search path algorithm which follows Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) is shown in Figure 1.

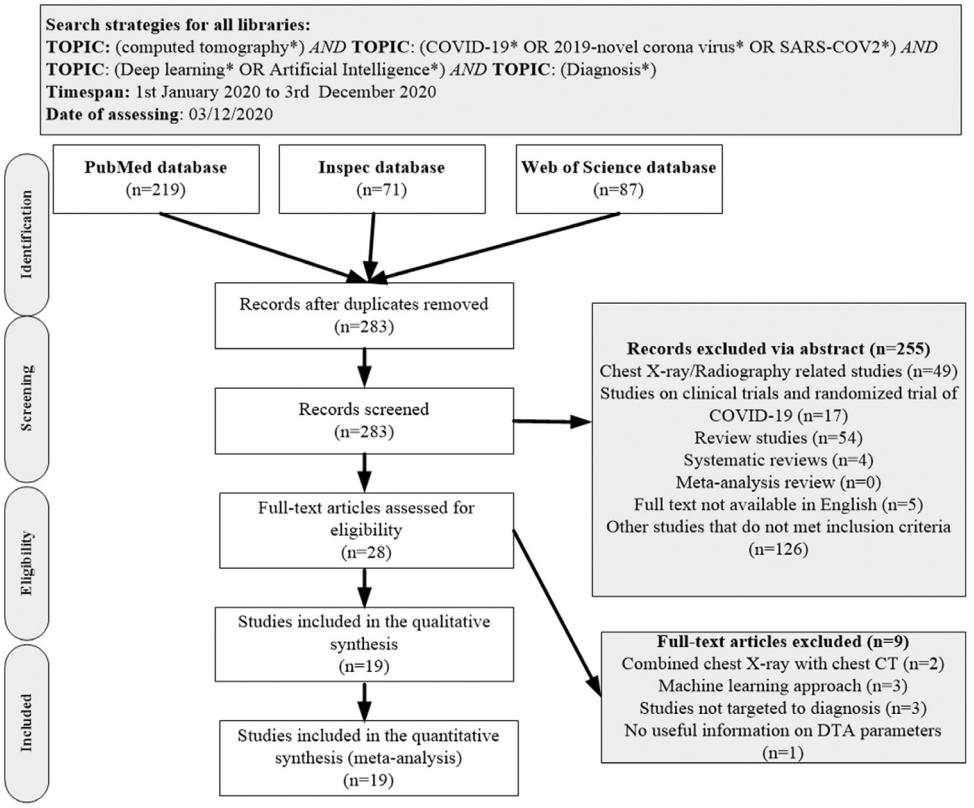

Figure 1. Study of inclusion and exclusion flowcharts adapted from the Preferred Reporting Items for PRISMA. n: number of literature and PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses, CT: Computed tomography, DTA: diagnostic test accuracy.

Eligibility Criteria

Literature was included in the study if it was based on deep learning diagnosis of COVID-19 using chest CT images both in screening and diagnostic protocol; well-documented information on diagnosis accuracy test at least sensitivity and specificity or 2 × 2 confusion matrix to compute other diagnostic test accuracy parameters. The included studies are composed majorly of retrospective with few prospective studies, an observer performance study, clinical trial, and comparative studies. The exclusion criteria comprised studies that involved literature reviews, studies on RT-PCR, other machine learning detection-based algorithms; detection using chest X-ray datasets or a combination of both chest CT and chest X-ray. Besides, studies devoid of useful information to compute the DTA and multiple publications were also excluded. For studies that reported the same study cohort or sub-set of the study, the most detailed one in terms of data availability was used.

Study Selection

Articles retrieved were manually sorted and duplicates were removed using titles/abstracts, then followed by full text according to the predefined search criteria and final eligible studies were selected.

Data Collection Process

We developed a standard extraction sheet which was consensually agreed upon by two independent reviewers team (TEK and YC or BAN and HS), to extract the information needed and resolve the conflict by consensus from eligible studies which includes: Nationality, data source, data partitioning, training model, deep learning techniques, training parameters, the total number of positive (cohort) vs control (negative) and other valuable information. Also, we extracted quantitative data for the meta-analysis which include (2 × 2) confusion matrix (True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP)) needed to compute the required DTA like sensitivity, specificity, diagnostic odds ratios (DOR), recall, accuracy, precision, F1- Score, the positive and negative likelihood ratios and the AUC (19,20). The expressions of these assessment measures are given below:

$$\text{Positive likelihood ratio } (LR^{+}) = \frac{\text{Sensitivity}}{1 - \text{Specificity}} \quad (1)$$

$$\text{Negative likelihood ratio } (LR^{-}) = \frac{1 - \text{Sensitivity}}{\text{Specificity}} \quad (2)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (3)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{False Negative}} \quad (5)$$

$$\text{F1} - \text{Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (6)$$

Risk of Bias and Quality Appraisal

The quality of included studies was assessed using a modified QUADAS-2 to ensure appropriateness for COVID-19 screening (21). The domains assessed were Patient Selection, Index Tests, Reference Standard, Flow and Timing, and Applicability. Two reviewers (TEK and YC) performed an independent quality assessment and the final result was based on consensus. The overall study quality pipeline is shown in Fig. 2

Statistical Data Analysis

A univariate meta-analysis was performed separately for sensitivity, and specificity to estimate the diagnostic accuracy of each modality using the DerSimonian-Laird method of random effects model (RE) (22). We chose the RE model due to suspicion of high rates of heterogeneity that might be occasioned by differences in the network architecture used for the training, differences in training data across age, sex, and so on. The primary outcomes were sensitivity, specificity, summary receiver operating characteristic (SROC) curve, and diagnostic odds ratios (DOR). We calculated point estimates and 95% confidence intervals (CI) for each study to ensure consistency in sensitivity and specificity. To obtain a SROC curve, we used a bivariate meta-analysis of sensitivity and specificity using R version 3.6.2 with RStudio version 1.2.5042 implementing R-packages "mada" and "meta", following which mean AUC of SROC was estimated (23). Secondary outcomes included positive likelihood and negative likelihood ratios, accuracy, precision and F1-score.

Statistical heterogeneity between studies was evaluated with Cochran's Q test and the I^2 statistic (19). For the Q statistic, values range 0%−40% imply insignificant heterogeneity, 30%−60% connote moderate heterogeneity, 75%−100% mean considerable heterogeneity. Publication bias was evaluated and visualized by constructing a funnel plot (25). All p-values were based on two-sided tests and p-value <0.05 was considered to represent statistical significance. We conducted sub-group analysis by screening based on the deep learning techniques and training model (transfer learning and customized method).

Quality Assessment

Quality assessment studies were rated as being of the moderate overall assessment of quality according to QUADAS2 (Fig. 2). About 5% of the included studies did not give details about patient selection, 5% provided unclear information about patient selection leading to high and unclear biases in patient selection as others are considered as having a low risk of bias. Also, three

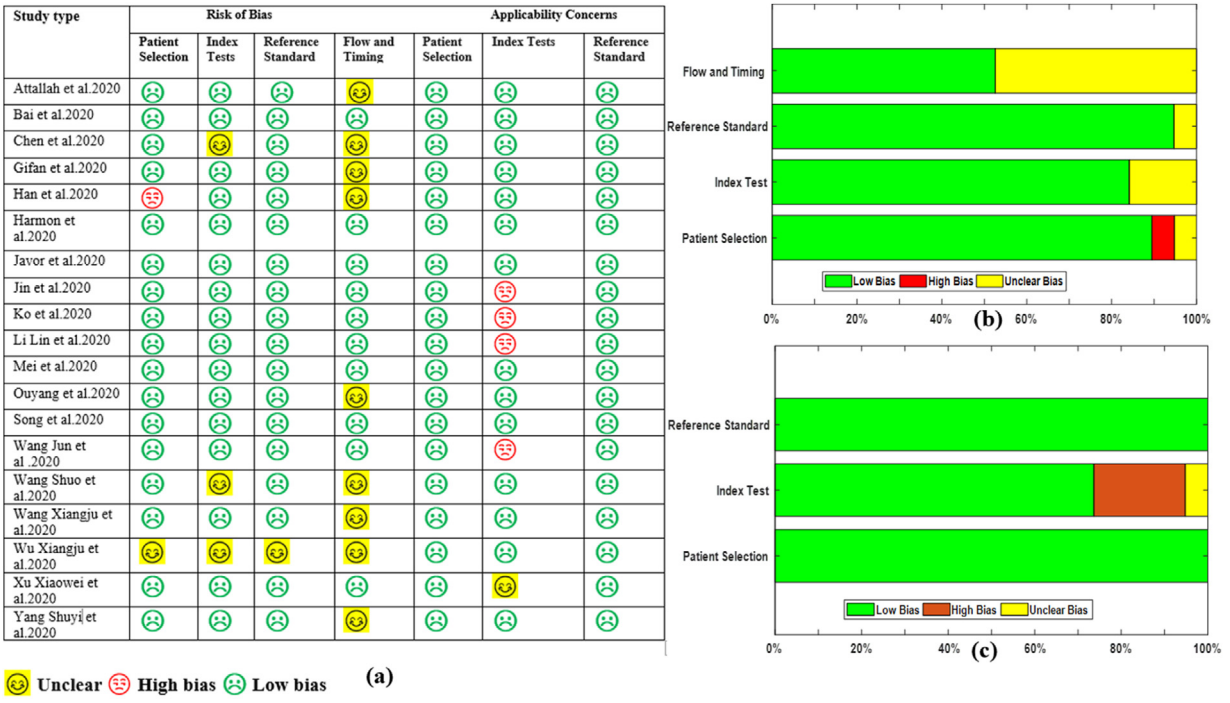

Figure 2. Assessment of quality of all included studies using the QUADAS-2 tool (a) Summary of risk of bias for each studies (b) Proportion of risk of bias for all domains (c) Proportion of applicability concerns in three domains. (Color print). (Color version of figure is available online.)

studies (16%) gave unclear information about how the index test was performed thus leading to an unclear risk of bias in the index test. Four studies (21%) focused on detection of COVID-19 from other pneumonia thus making the review question not match exactly the index test and about 5% of the studies did not give clear information on whether the review question matched the targeted condition thus proving high and unclear applicability concerns for the index test respectively. Eight studies (42%) provided no clear information about the interval between index test and reference standard test and how they were performed leading to unclear bias in flow and timing as others are considered as having a low risk of bias. A funnel plot was used to also assess the publication bias for the 19 studies that met the inclusion criteria. There is low publication bias in the study according to Liu (25) the points will be symmetrically distributed around the true effect in the shape of an inverted funnel when publication bias is very low as shown in Figure 3. This was also supported by the QUADAS-2 assessment in Figure 2.

RESULTS

Overview of the Included Studies

The database search retrieved 283 publications. After the duplicates were removed, and the publications screened using title and abstracts, a total of 255 publications were screened out (Fig. 1). Twenty-eight full-text articles were assessed for eligibility. Nineteen articles were found worthy to meet inclusion criteria (12,13,26–41). Three articles applied the machine learning approach, three articles combined datasets of chest X-ray and chest CT and three studies did not provide useful information on parameters to estimate the diagnostic test accuracy (DTA), hence these nine studies were exempted as shown in Figure 1. The included studies are from eight different countries: Austria (5.3%), China (63.2%), Iran (5.3%), Korea (%5.3%), Egypt (5.3%), China, and U.SA. (10.5%), China, U.S.A., Japan, and Italy (5.3%). In these studies, two data sources were identified, namely, single source (26,27,38,41) and multiple sources (12,13,28–30,33,34,36,37,40). The deep learning networks were classified into three categories ResNet models (15,31,33,39), Hybrid of ResNet architectures like Alex, GoogleNet, FCoNet, UNet++, Ensembed deep learning; 3D DensNet-121, ConvNet, UNet (13,26,32,34,36,40) and finally other models that do not fall into the above two categories (7,22–25,30,32,33,36). Some of the studies included in the quantitative synthesis (meta-analysis) have reported a higher DTA performance for deep learning algorithms compared with radiologist interpretation (31,37,41), while others have shown that deep learning algorithm did aid DTA performance (12,30,33,39). Other studies reported higher sensitivity over specificity (26–29,33,35,39,41), while some reported higher specificity (13,15,30,31,32,34,36,37,40).

Diagnostic Test Accuracy of all Included Studies

This is overall of all diagnostic test accuracy (DTA), the pooled sensitivity of univariate analysis of nineteen studies was 0.908 (95% CI:0.879 to 0.931, I^2 = 81.6% for 19 studies) as shown in Figure 4. The pooled specificity of univariate analysis was 0.916 (95% CI:0.877 to 0.944, I^2 = 82.2% for 19 studies) as shown in Figure 5. The pooled diagnostic odd

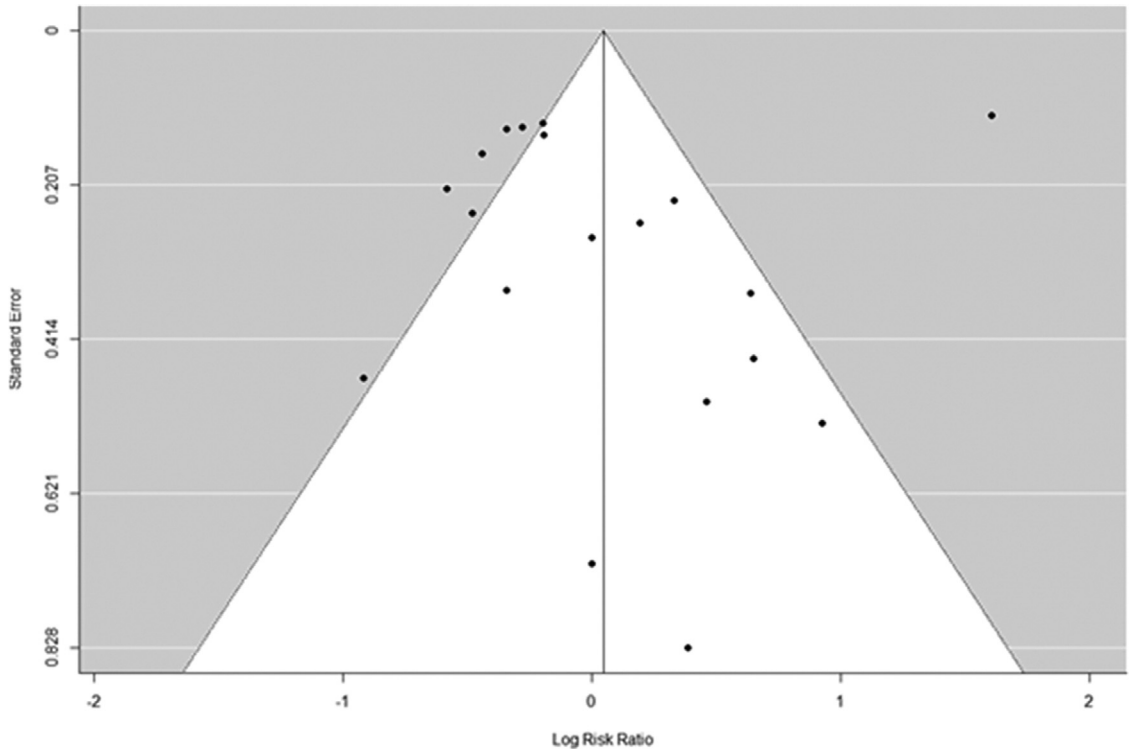

Figure 3. Funnel plot showing the low likelihood of publication bias in all included studies.

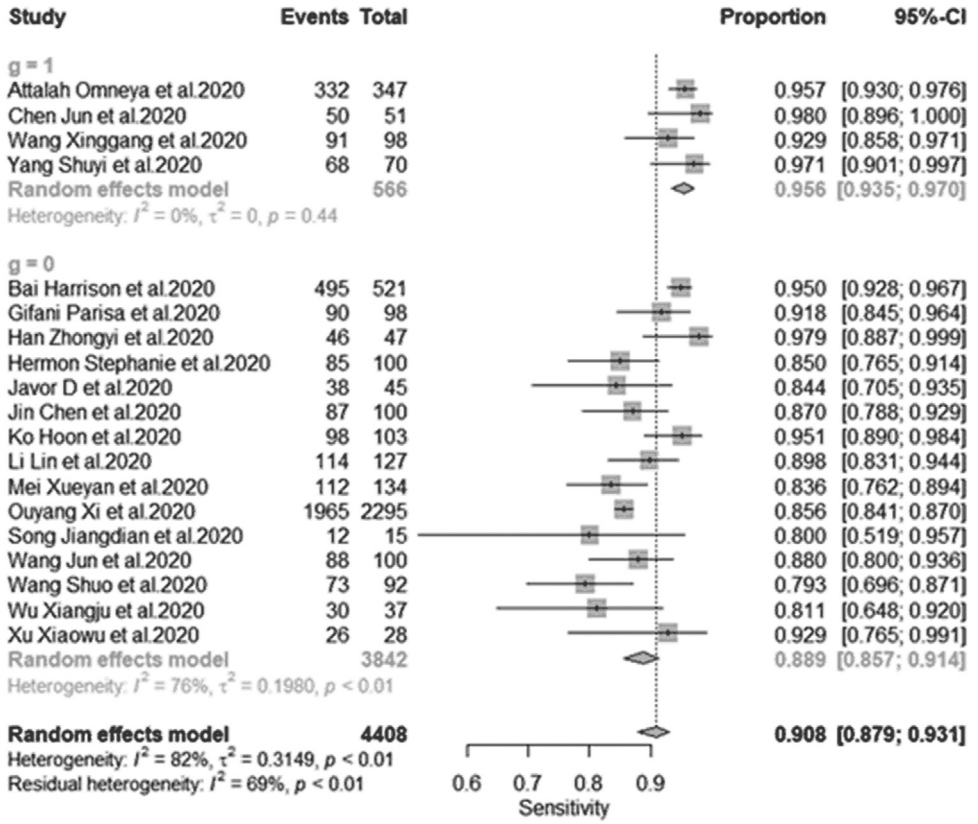

Figure 4. Univariate sub-group analysis of sensitivity with random model based on data source. g represents sub-group analysis of data, when g = 1(single-source datasets) and g = 0 (multi-source datasets).

ratios (DOR) was 112.5 (95% CI:57.7 to 219.3, I^2 = 90.7% for 19 studies) as shown in Figure 6. The positive likelihood ratio (LR^+) ranges from 2.11 to 177.60 with pooled mean of 23.13 (Table 2), likewise the negative likelihood ratio (LR^-) spans from 0.021 to 0.31 with pooled mean of 0.13. The SROC of the bivariate model has an AUC of 0.95 (Fig. 7). The accuracy of all included studies ranges from 0.7600 to 0.9879 with a mean of 0.8948 (Table 2), while the precision ranges from 0.7059 to 0.9703 with the mean of 0.8966 (Table 2), the F1-score has a mean of 0.8966 and ranges from 0.7500 to 0.9787 and finally, the recall ranges from 0.7935 to 0.9804 with mean of 0.8949 (Table 2).

DTA for the Sub-Group Analysis Based on Training Data

We decided to check the effect of different training datasets on the diagnostic performance. The analysis with multi-datasets has a sensitivity of 0.889 (95% CI:0.857 to 0.914, I^2 = 75.6% for 15 studies) while that of single-source datasets was 0.956 (95% CI:0.935 to 0.970, I^2 = 0.0% for four studies), indicating no significant heterogeneity between the sensitivity. The random effect model shows a slightly significant difference in the sensitivity of studies with single-source and multiple source datasets with (p-value <0.001) (Fig. 4). In addition, the specificity is 0.917(95% CI:0.866 to 0.949, I^2 = 90.2% for 15 studies) for multi-source datasets, while the specificity of single-source datasets is 0.923(95% CI:0.894 to 0.945, I^2 = 18.0% for 15 studies). The result indicates that there was a slightly significant difference in the specificity of single-source and multi-source datasets (Fig. 5). Furthermore, the single-source dataset has a pooled DOR of 282.7 (95% CI:168.9 to 473.1, I^2 = 0.0% for four studies), while the multi-source datasets has DOR of 88.8 (95% CI:41.9 to 188.2, I^2 = 91.3% for 15 studies). The result of DOR indicates that there is a significant difference between DOR of single and multi-source datasets during training as shown in Figure 6.

DTA for the Sub-Group Analysis Based on Network Training Model

The whole process was subdivided into two models namely pre-trained or customized network based on the way the network was trained. For the pre-trained model, the pooled sensitivity is 0.905 (95% CI:0.875 to 0.929, I^2 = 81.6% for 15 studies), while the sensitivity of customized network is 0.931 (95% CI:0.795 to 0.979, I^2 = 67.7% for four studies). There was no significant difference between the customized training datasets and pre-trained data for analysis with p-value = 0.6008 as shown in Fig. 8. Likewise, the specificity for pre-trained datasets is 0.925(95% CI:0.887 to 0.952, I^2 = 89.7% for 15 studies), while that of customized datsets is 0.862 (95% CI:0.639 to 0.956, I^2 = 77.0% for four studies) as shown in

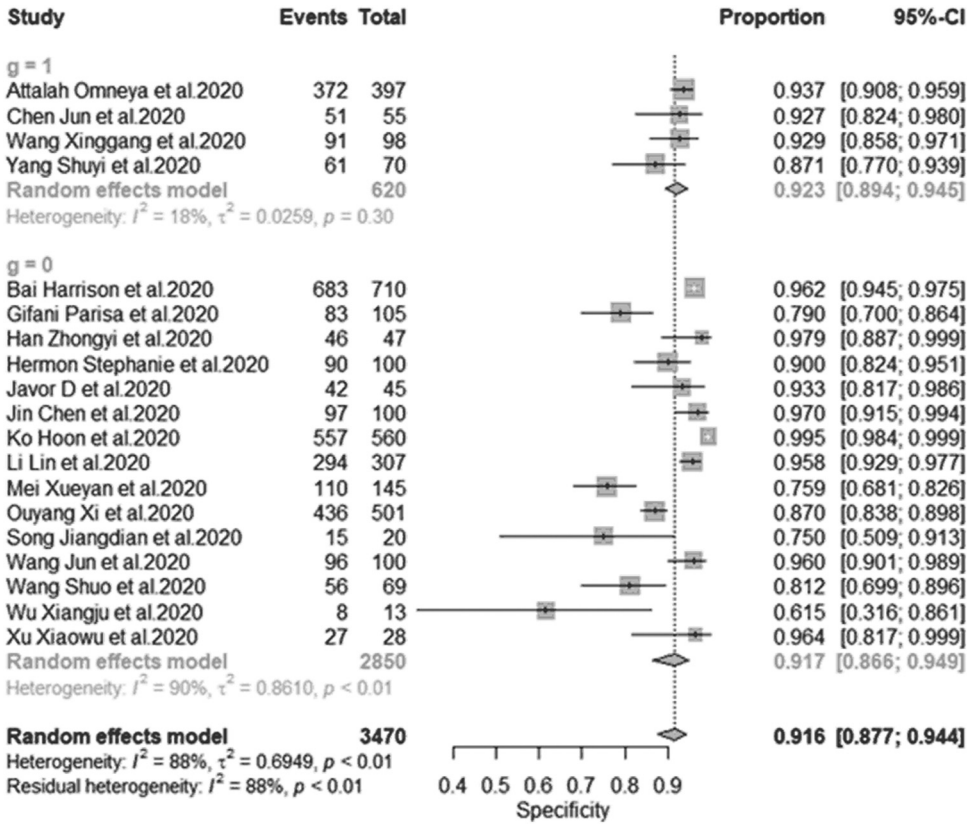

Figure 5. Univariate sub-group analysis of specificity with random model based on data source. g represents sub-group analysis of data when g = 1(single-source datasets) and g = 0 (multi-source datasets).

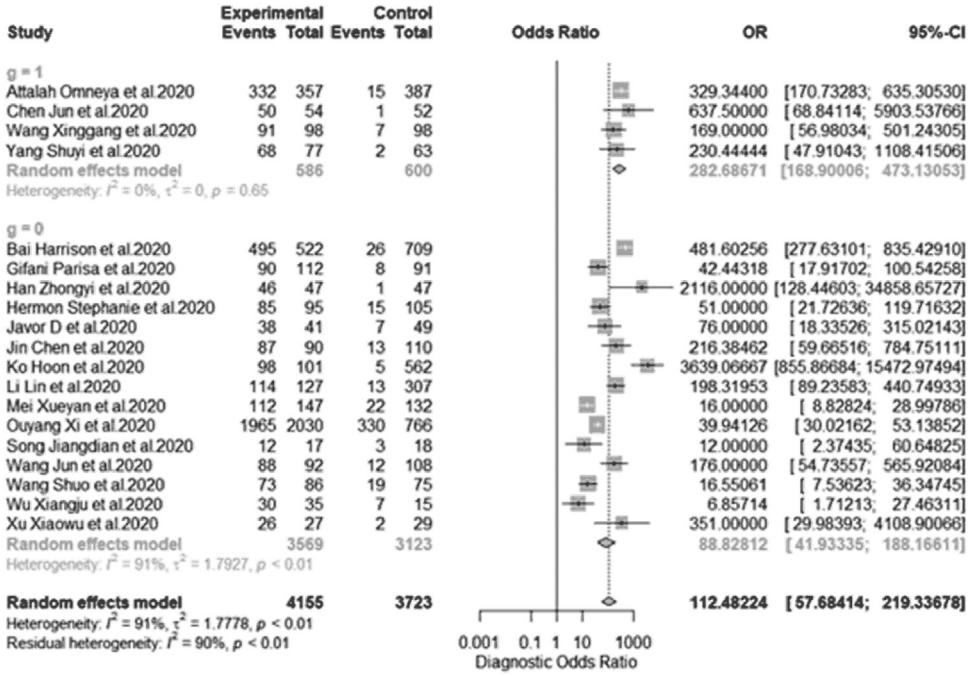

Figure 6. Univariate sub-group analysis of DOR based on data source. DOR: diagnostic odds ratio, g represents sub-group analysis of data when g = 1(single-source datasets) and g = 0 (multi-source datasets).

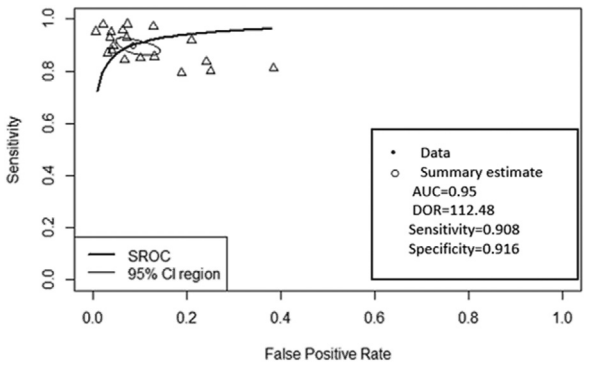

Figure 7. The SROC bivariate model curve for diagnostic test accuracy. SROC: summary receiver operating characteristic curve (bivariate model) for diagnostic test accuracy. CI: Confidence interval; AUC: area under the curve.

Figure 9. These results indicate that there was no significant difference in the sub-group analysis between the pre-trained and customized models with *p*-value of 0.3134. The pooled DOR of the pre-trained model is 125.3 (95% CI:61.6 to 254.7, I^2 = 91.7% for 15 studies), while that of customized model is 83.8 (95% CI:55.7 to 219.3, I^2 = 86.2% for four studies) as shown in Figure 10. There is no statistical difference between the DOR of the pre-trained and customized models.

DTA for the Sub-Group Analysis Based on Network Architecture

We sub-divided the analysis into three categories of network for convenience and for ease analysis during the meta-analysis as ResNet, ResNet Hybrid, and other networks as shown in Table 1. The network with ResNet architectures has a sensitivity of 0.845 (95% CI:0.801 to 0.881, I^2 = 0.0% for four studies), while the Hybrid ResNet has sensitivity of 0.916 (95% CI:0.872 to 0.945, I^2 = 77.4% for six studies) and other network architecture has sensitivity of 0.927(95% CI:0.878 to 0.957, I^2 = 77.7% for nine studies). The sensitivity values indicate that there is a significant difference between the three sub-groups of the network used for the training of deep learning with (*p*-value = 0.0068) as shown in Figure 11. Looking in-depth into the specificity also, the ResNet model has 0.868 (95% CI:0.665 to 0.956, I^2 = 85.9% for four studies), the pooled specificity of ResNet Hybrid is 0.957 (95% CI:0.912 to 0.980, I^2 = 89.9% for 6 studies) and the specificity of other network models is 0.896(95% CI:0.825 to 0.940, I^2 = 85.5% for nine studies).

These results reveal that there is no significant difference in the specificity of the three categories of network architecture with a *p*-value of 0.1011 as shown in Figure 12. For the ResNet architecture, the pooled DOR is 35.4 (95% CI:8.8 to 143.3, I^2 = 83.9% for four studies) while the ResNet Hybrid architecture has DOR of 109.7(95% CI:37.4 to

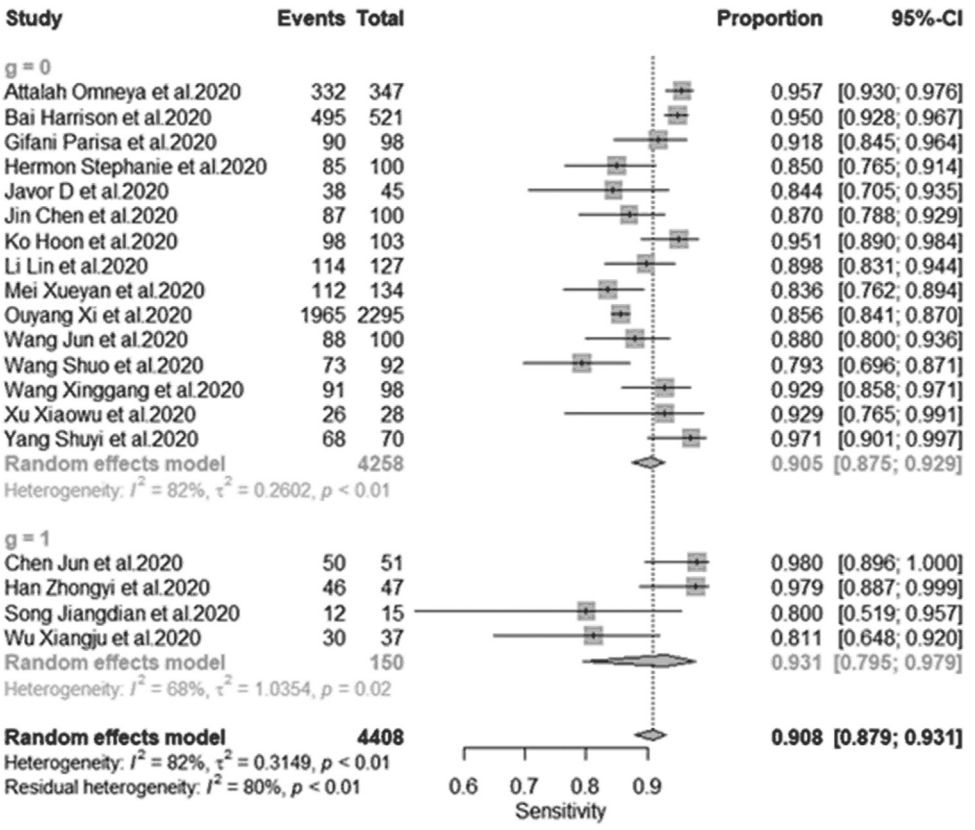

Figure 8. Univariate sub-group analysis of sensitivity with random model based on the training model. g represents sub-group analysis of the network model when g = 1(pre-trained model) and g = 0 (customized model).

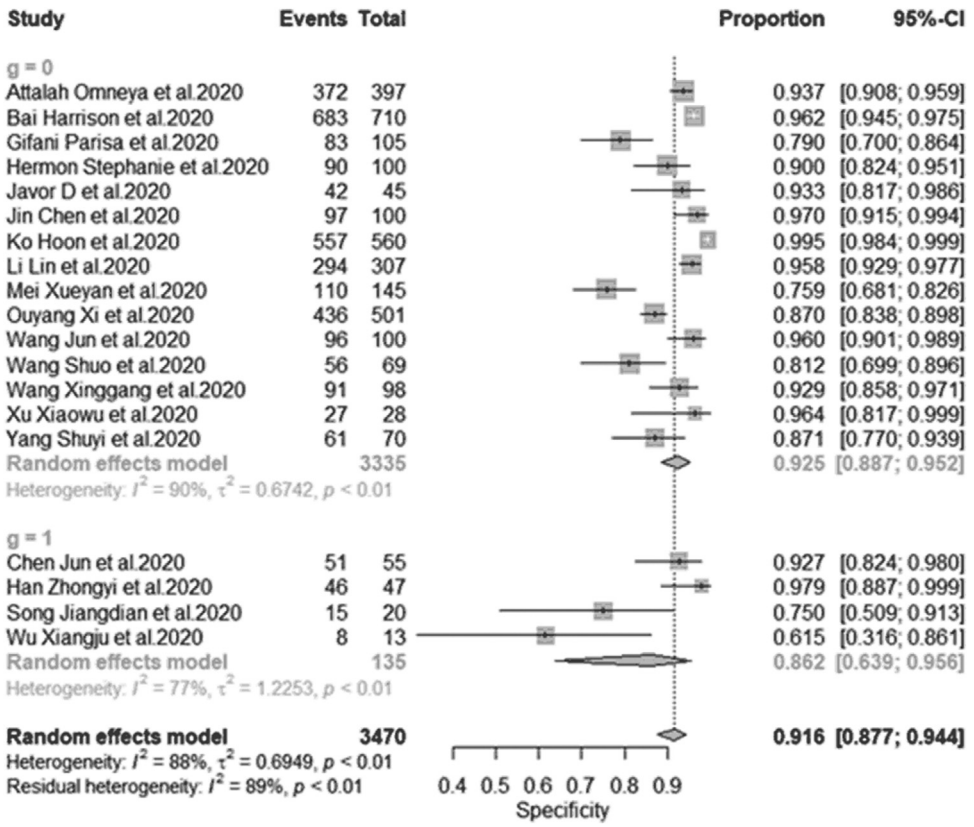

Figure 9. Univariate sub-group analysis of specificity with random model based on the training model. g represents sub-group analysis of the network model when g = 0 (pre-trained model) and g = 1 (customized model).

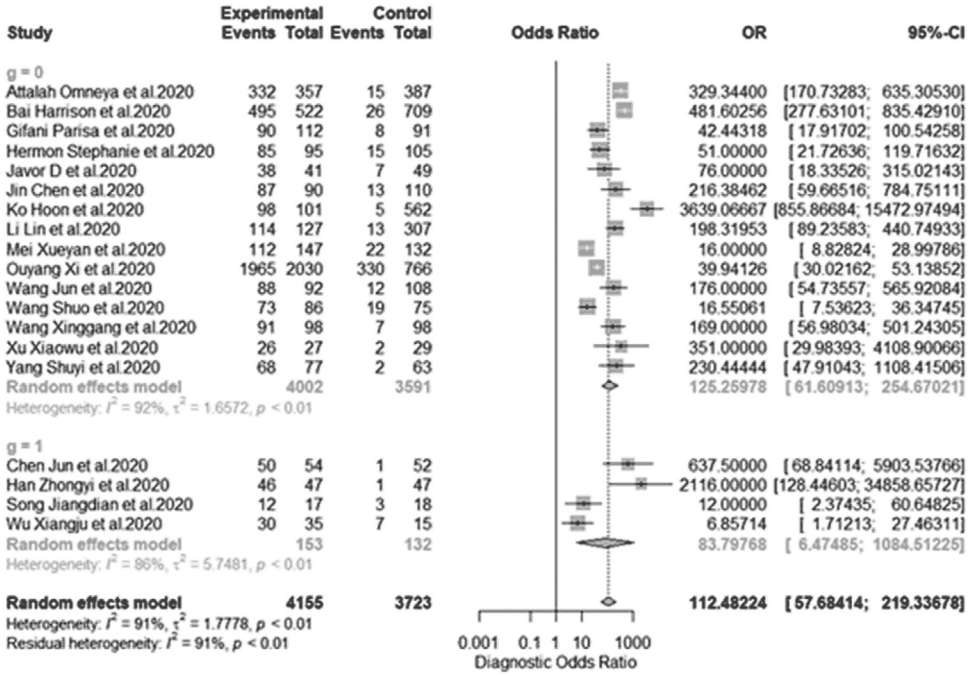

Figure 10. Univariate sub-group analysis of DOR based on the training model. DOR=diagnostic odds ratio, g represents sub-group analysis of the network model when g = 0 (pre-trained model) and g = 1 (customized model).

TABLE 1. Summary of Study Characteristics of all Included Studies

S/N	Authors	Nationality	Data Sources (Single/Multiple) or benchmark or real-time data	Training model (Pre-trained (transfer) or customized model)	Data partitioning Training data (TD) Validation data (VD), Test data (TTD)	Deep learning Techniques Eg GAN, CNN, VGG etc	Number of training epochs/ Learning rate/ Regularization	Batch size	Threshold for the classifiers	Brief summary of the study	Total no. of patients COVID-19 positive vs Negative (Control)	Any other vital information about the study
1	Attallah *et al.* (26)	Egypt	Single/ benchmark	Pre-trained		Alex, GoogleNet, ResNet, ShuffleNet	20/1 × 10^{-4}/ 5 × 10^{-4}	10	Cubic SVM classifier	4 pre-trained CNNs were used individually to detect COVID-19 and distinguish them from non-COVID-19 cases	347 vs 397	Retrospective, stochastic gradient descent with momentum is used for optimization with 5-fold cross validation.
2.	Bai *et al.*(12)	China/USA	Multiple/ benchmark	Pre-trained	TD = 830, VD = 237, TTD = 119	ImageNet	NA/1 × 10^{-4}/ 1 × 10^{-4}	64	EfficientNet B4	AI+Radiologist vs Radiologist vs Without AI	521 vs 665	Retrospective
3.	Chen *et al.* (27)	China	Single/ benchmark	Customized	-	UNet++	NA/1 × 10^{-4}/NA	~	Prediction box	Compare DL detection in chest CT using UNet++with radiologist efficiency	51 vs 55	Retrospective
4.	Gifan *et al.* (28)	Iran	Multiple/ benchmark	Pre-trained	TD = 232, VD=58, TTD = 97	Ensembled deep transfer +CNN architecture	50/1 × 10^{-4}/NA	32	Softmax	Performance of ensemble deep transfer learning for COVID-19 detection	349 vs 397	Retrospective
5.	Han *et al.*(29)	China	Multiple/ benchmark	Customized	TD = 276, VD = 92, TTD = 92	3D CNN	100/1 × 10^{-5}/NA	~	~	Proposed a weakly-supervised learning framework for screening of COVID-19	230 vs 230	5-fold cross validation
6.	Harmon *et al.* (30)	China, Italy, Japan, U.S.A.	Multiple/ benchmark	Pre-trained	TD = 984, VD = 296, TTD = 1337	3D DenseNet-121	~	~	Grad-CAM method	Develop and evaluate a DL algorithm for the detection of COVID-19 on chest CT with hybrid 3D and full 3D models	922 vs 1695	Lung segmentation +data augmentation
7.	Javor *et al.* (15)	Austria	Multicenter/ benchmark	Pre-trained	TD = 328, VD = 66	ResNet-50	17/NA/NA	32	~	Compare the robustness of DL in classification of COVID-19 to experienced radiologists.	856 vs 254	Data augmentation
8.	Jin *et al.* (31)	China	Multicenter/ benchmark		TD = 751, TTD = 751	ResNet-152	~	~	Grad-CAM +LASSO	AI vs Radiologist detection	~	Lung segmentation
9.	Ko *et al.*(32)	Korea	Multicenter/ benchmark	Pre-trained	TD = 955, TTD = 239	FCoNet+ResNet-50	~	~	Softmax		~	No lung segmentation + data augmentation
10.	Li *et al.*(13)	China	Multicenter/ benchmark	Pre-trained	TD = 3918, TTD = 434	CovNet+ResNet-50	~	~	Softmax (Grad-Cam)	AI to detect COVID-19 from CAP	468 vs 2854	Lung segmentation

(*continued on next page*)

TABLE 1. (Continued)

S/N	Authors	Nationality	Data Sources (Single/Multiple) or benchmark or real-time data	Training model (Pre-trained (transfer) or customized model)	Data partitioning Training data (TD) Validation data (VD), Test data (TTD)	Deep learning Techniques Eg GAN, CNN, VGG etc	Number of training epochs/ Learning rate/ Regularization	Batch size	Threshold for the classifiers	Brief summary of the study	Total no. of patients COVID-19 positive vs Negative (Control)	Any other vital information about the study
11.	Mei *et al.* (33)	China/ U.S.A.	Multicenter/ benchmark	Pre-trained	TD = 534, VD = 92, TTD = 279	ResNet-18	40/1 × 10^{-3}/NA	16	~	AI Vs Radiologist detection	419 vs 486	Lung segmentation
12.	Ouyang *et al.* (34)	China	Multicenter/ benchmark	Pre-trained	TD = 2186, TTD = 2796	ResNet-34	20/2 × 10^{-4}/ 1 × 10^{-4}	20	~	Diagnosis of COVID-19 vs CAP		Lung segmentation
13.	Song *et al.* (35)	China	Multi-centre/ benchmar	Customized	TD = 161, VD = 20, TTD = 20	Big-BIGAN	120/NA/NA	16	~	Diagnosis of COVID-19 vs Other viral pneumonia	98 vs 103	~
14.	Wang *et al.* (36)	China	Multicenter/ benchmark	Pre-trained		3D UNet + 3D ResNet	300/NA/0.95	~	Binary classifier	Applied a novel multi-task prior-attention residual learning strategy for COVID-19 screening	1315 vs 3342	Lung segmentation +Data augmentation
15.	Wang *et al.* (37)	China	Multicenter/ benchmark	Pre-trained	TD = 709, TTD = 342	DenseNet-121 +COVID-19 Net	~	~	~	Propose a fully automatic DL system for COVID-19 diagnostic	~	Lung Segmentation
16.	Wang *et al.* (38)	China	Single source/ benchmark	Pre-trained	TD = 449, TTD = 131	UNet +DeCovNet	100/1 × 10^{-5}/ 1 × 10^{-4}	1	Prob threshold of 0.8 using binary crossentropy	Classify COVID-19 and Non-COVID-19	313 vs 229	~
17.	Wu *et al.*(39)	China	Multicenter/ benchmark	Customized	TD = 294, TTD = 50	ResNet-50	NA/1 × 10^{-5}/NA	4	~	Classify into 3 COVID-19, non-COVID-19 and other influenza	~	Lung segmentation +Data augmentation
18.	Xu *et al.* (40)	China	Multicenter/ benchmark	Pre-trained	TD = TTD =	VNet + ResNet-18	~		~	~	219 vs 399	Lung segmentation +Data augmentation
19.	Yang *et al.* (41)	China	Single/ benchmark	Pre-trained	TD = 135, VD = 20, TTD = 140	DenseNet	20/NA/NA	32	~	Pilot study on COVID-19 diagnosis	~	~

CAP, community acquired pneumonia; COVID, coronavirus disease 2019; DL, deep learning

TABLE 2. DTA Estimated From all Included Studies Using the (2 × 2) Truth Table

Authors	Sensitivity TP	FN	Specificity TN	FP	LR^+	LR^-	Accuracy	Precision	F1-Score	Recall
Attallah *et al.* (26)	332	15	372	25	15.1935	0.0461	0.9462	0.9300	0.9432	0.9568
Bai *et al.* (12)	495	26	638	27	23.4005	0.0520	0.9553	0.9483	0.9492	0.9501
Chen *et al.* (27)	50	1	51	4	13.4804	0.0211	0.9528	0.9259	0.9524	0.9804
Gifani *et al.* (28)	90	8	83	22	46.0000	0.1033	0.8522	0.8036	0.8571	0.9184
Han *et al.* (29)	46	1	46	1	8.5000	0.0217	0.9787	0.9787	0.9787	0.9787
Harmon *et al.* (30)	85	15	90	10	12.6667	0.1667	0.8750	0.8947	0.8718	0.8500
Javor *et al.* (15)	38	7	42	3	29.0000	0.1667	0.8889	0.9268	0.8837	0.8444
Jin *et al.* (31)	87	13	97	3	177.6052	0.1340	0.9200	0.9667	0.9159	0.8700
Ko *et al.* (35)	98	5	557	3	21.11981	0.0488	0.9879	0.9703	0.9808	0.9515
Li *et al.* (13)	114	13	294	13	3.4627	0.1069	0.9401	0.8976	0.8976	0.8976
Mei *et al.* (33)	112	22	110	35	6.5994	0.2164	0.7957	0.7619	0.7972	0.8358
Ouyang *et al.* (34)	1965	330	436	65	3.2000	0.1652	0.8587	0.9680	0.9087	0.8562
Song *et al.* (35)	12	3	15	5	22.0000	0.2667	0.7714	0.7059	0.7500	0.8000
Wang *et al.* (36)	88	12	96	4	4.2115	0.1250	0.9200	0.9565	0.9167	0.8800
Wang *et al.* (37)	73	19	56	13	13.0000	0.2545	0.8012	0.8488	0.8202	0.7935
Wang *et al.* (38)	91	7	91	7	2.1081	0.0741	0.9286	0.9286	0.9286	0.9286
Wu *et al.* (39)	30	7	8	5	26.0000	0.3074	0.7600	0.8571	0.8333	0.8108
Xu *et al.* (40)	26	2	27	1	26.0000	0.0741	0.9464	0.9630	0.9455	0.9286
Yang *et al.* (41)	68	2	61	9	7.5556	0.0328	0.9214	0.8831	0.9252	0.9714
				Min.	**2.1081**	**0.0211**	**0.7600**	**0.7059**	**0.7500**	**0.8949**
				Max.	**177.6052**	**0.3074**	**0.9879**	**0.9787**	**0.9787**	**0.9804**
				Avrg.	**23.1350**	**0.1256**	**0.8948**	**0.9008**	**0.8966**	**0.7935**

DTA, diagnostic test accuracy; FN, false negative; FP, false positive; LR^+, positive likelihood ratio; LR^-, negative likelihood ratio; TP, true positive; TN, true negative.

The bold values represent the minimum, maximum and average value for the computed DTA.

321.6, I^2 = 89.0% for nine studies). Finally, the category of other networks has pooled DOR of 363.2 (95% CI:73.6 to 741.7, I^2 = 93.3% for six studies) revealing there is no statistical difference between these types of networks used during the training process (Fig. 13).

DISCUSSION

The systematic review of diagnostic test accuracy (DTA) of deep learning (DL) detection of COVID-19 in nineteen studies has been carried out. The pooled DTA for all the 19 studies at a 95% Confidence Interval (CI) had a sensitivity of 0.908 (0.879 to 0.931, I^2 = 81.6%), specificity of 0.916 (0.877 to 0.944, I^2 = 82.2%), DOR of 112.5 (57.7 to 219.3, I^2 = 90.7%), LR^+ of 23.13(2.11 to 177.60), LR^- of 0.13(0.021 to 0.31), accuracy of 0.8948 (0.7600 to 0.9879), recall of 0.8949 (0.7935 to 0.9804), precision of 0.8966 (0.7059 to 0.9703), F1-score of 0.8966 (0.7500 to 0.9787) and AUC of 0.95. From this high pooled DTA of the DL algorithm, it is evident that the deep-learning algorithm can distinguish between patients with and without COVID-19 successfully. A previous study on DL detection of COVID-19 by Moezzi *et al.* (16) on 23 studies had recorded similar results with sensitivity of 0.91, specificity of 0.88, AUC of 0.96, and DOR of 99.4, although our pooled specificity and DOR increased by 4% and 13% respectively. Other meta-analysis studies on DTA performance of chest CT on COVID-19 detection were also compared (42–48). Comparing the pooled sensitivity with the work of Mahmoud *et al.* (42) on DTA of chest CT for the detection of COVID-19 on 7 studies, it was found that the pooled sensitivity is 0.89, and the result is similar to the pooled sensitivity of 0.89 recorded by Komolafe *et al.*(43) on DTA of chest CT using 36 studies. It was observed from our results that deep learning detection achieved higher sensitivity. This implies that deep learning algorithms has the capacity to detect more COVID-19 compared to radiologist result of Mahmoud *et al.*(42) and Komolafe *et al.*(43). Also, our pooled sensitivity achieved a slight increase of about 0.9% when compared to the pooled result of seven studies by Vafea *et al.* (44), and this was also in line with the result of 13 studies done by Bao *et al.*(45) with a pooled sensitivity of 0.904. Similarly, Kim *et al.*(46) performed a meta-analysis on 63 studies with chest CT pooled sensitivity of 0.94 which represents approximately 3.5% increment above our deep learning result. This is in good agreement with the conclusive remarks by Duarte *et al.* (47) that pooled only two studies with a sensitivity of 0.953. Similarly, Boger *et al.* (48) worked on six studies with a pooled sensitivity of 0.92 that shows a slightly higher sensitivity over our result.

Critically examining the specificity, most studies on meta-analysis detection of COVID-19 using chest CT seldomly report specificity, our pooled specificity was 0.916 with a high proportionate increment over the pooled specificity of 0.37 in 63 studies of Kim *et al.*(46). Likewise, there was a significantly higher increment over that of Duarte *et al.* (47) with pooled specificity of 0.44.

Even Boger *et al.* (48) who previously recorded higher sensitivity had an extremely lower pooled specificity of 0.251.

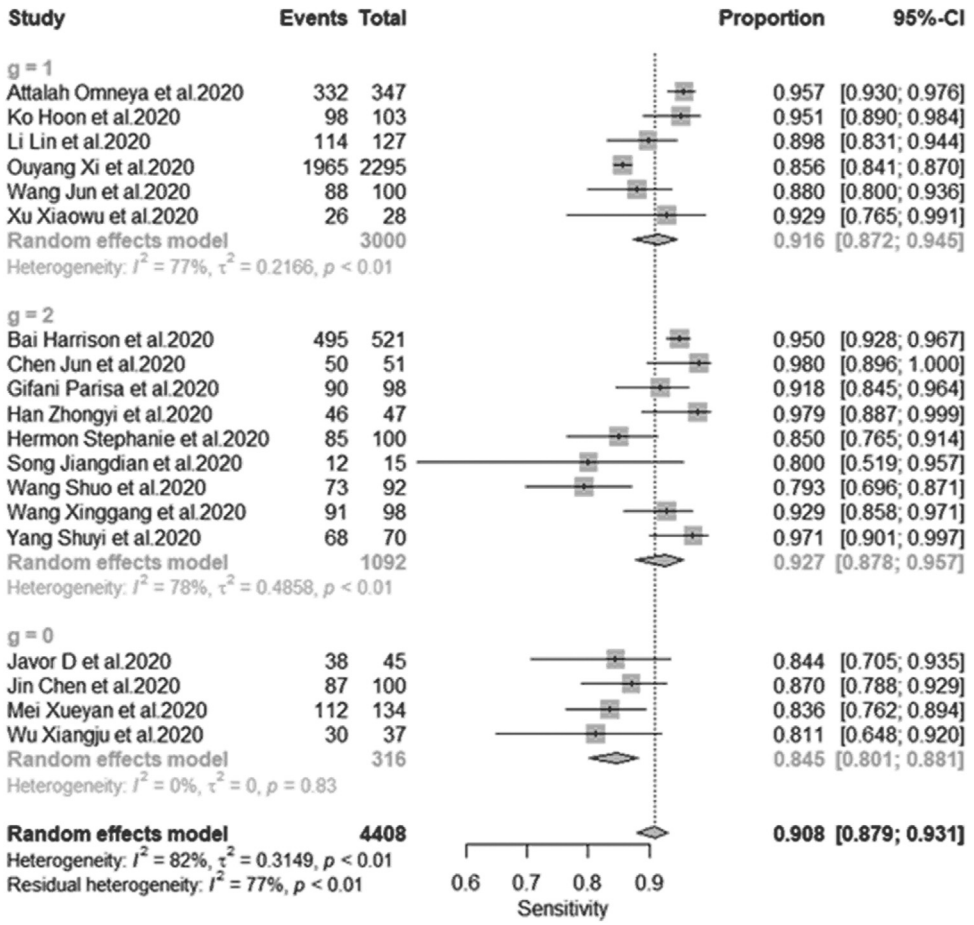

Figure 11. Sub-group analysis of sensitivity with random model based on deep learning architecture. g represents sub-group analysis of the network architecture when g = 0 (ResNet), g = 1 (ResNet Hybrid) and g = 2 (Other networks).

According to this result, it is extremely important to see the efficacy of the DL algorithm in classifying patients with and without COVID-19. Good diagnostic equipment must have significantly high sensitivity and specificity, that is, the ability to classify disease suspects as disease and non-disease as non-disease. The AUC of this study was 0.95 which is significantly high. This is evident that the deep-learning algorithm can distinguish between patients with and without COVID-19. All the considered meta-analysis (42−48) on DTA of chest CT in COVID-19 detection did not report AUC.

One of the major DTA parameters is the likelihood ratios, since relying only on sensitivity and specificity may lead to overestimation of the benefits of the test (15). The likelihood ratios are useful over a range of disease frequencies and could help to improve clinical judgments. The pooled positive (LR^+) and negative (LR^-) likelihood ratios are 23.13 and 0.13, respectively. The LR^+ of 23 means that COVID-19 positive using DL algorithm is 23 times more likely to occur in patients with COVID-19 than without COVID-19, likewise the LR^- of 0.13 means COVID-19 negative has a higher likelihood of negative test for DL algorithm than patients without COVID-19. According to Jaeschke *et al.* (49), LR^+ greater than 10 produces a greater pretest probability, and the LR^- less than 0.1 produces conclusive changes in the post-test probability. Juxtaposing the LR^+= 1.194 and LR^- = 0.301 recorded by Boger *et al.* (48), our deep learning produced a significantly higher likelihood and thus reveals and detects more COVID-19 cases. Our meta-analysis had a DOR of 112.5 for 19 studies which means the odds ratio to positive result among persons with COVID-19 was approximately 113 times higher than the odds ratio for positive result among patients without COVID-19. Besides, it is noteworthy that none of the comparison studies of COVID-19 detection with radiologist perspective documented DOR for their studies (37-43). The overall accuracy of the 19 studies is 0.8948, this value is significantly higher than that reported by Boger *et al.*(48) on COVID-19 using chest CT. The overall precision of 0.896 was estimated by our meta-analysis for all studies, which signifies how accurate or precise the deep-leaning model was compared to the total predicted positive value, as this is helpful to determine when the cost of false positive is high. In the same manner, overall recall is 0.8949, and this value estimated how many of the true COVID-19 positives the model was able to classify relative to the total actual COVID-19 positives. Finally, the overall F1-score of 0.8966 recorded provides better DTA

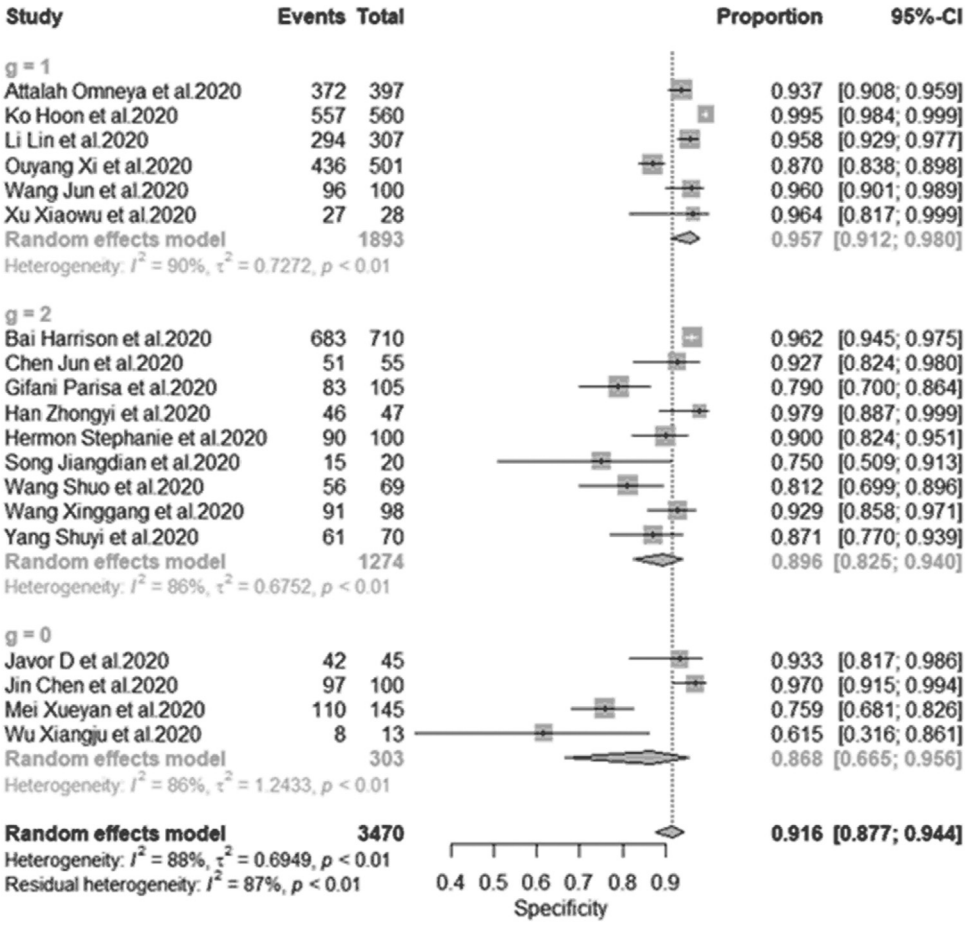

Figure 12. Sub-group analysis of specificity with random model based on deep learning architecture. g represents sub-group analysis of the network architecture when g = 0 (ResNet), g = 1 (ResNet Hybrid) and g = 2 (Other networks).

than the overall accuracy explained above because it seeks a balance between precision and recall rate of the diagnostic model which is typically applied with a large number of actual negatives. Therefore, taking a clue from the general result of the meta-analysis of DTA, it can be deduced that the deep learning algorithm has a significantly higher DTA compared to the radiologist performance and is more likely to reduce the number of false negatives and false positives since that is the main goal of good diagnostic equipment.

To effectively understand how a different deep learning model, deep learning architectures, and nature of datasets will influence the performance of the algorithm for COVID-19 detection, we did a sub-group analysis based on the type of data source, deep learning model, and type of network architecture. In terms of data type, the sensitivity of 0.889 and 0.956 was recorded for multi-source and single-source datasets respectively. This indicates that a single-source had both slightly higher non-significant differences over that of overall sensitivity and that of multi-source datasets. For the specificity, both multi-source and single-source datasets showed higher results over the overall specificity, but single-source datasets exhibited higher non-significant over multi-source data. In the pooled estimate of DOR, the value of 282.7 and 88.8 was recorded for single-source and multi-source, respectively. This implies that single-source had slightly higher non-significant differences over that of overall DOR but significant difference from that of multi-source datasets with a p-value of 0.013. For the sub-group analysis based on the training model, the sensitivity of the customized model shows a slightly statistically non-significant difference over that of pre-trained and overall, while the specificity of the pre-trained model shows a statistically non-significantly higher value than the customized and pre-trained models. This means there is no significant difference in terms of the training model. We also did a sub-group analysis based on the type of network architecture. The algorithm trained on ResNet alone had the least sensitivity with no heterogeneity compared to the overall, which had higher sensitivity than the rest. The significantly low heterogeneity indicates the consistency of ResNet for detection. The highest sensitivity was discovered in other variants of network architecture apart from ResNet and its hybrid. The sensitivity of this sub-group shows a slightly significant difference between ResNet, ResNet Hybrid, and other network variants with (p-value = 0.007). For the specificity, there is a slight non-significant difference among the three categories of network used.

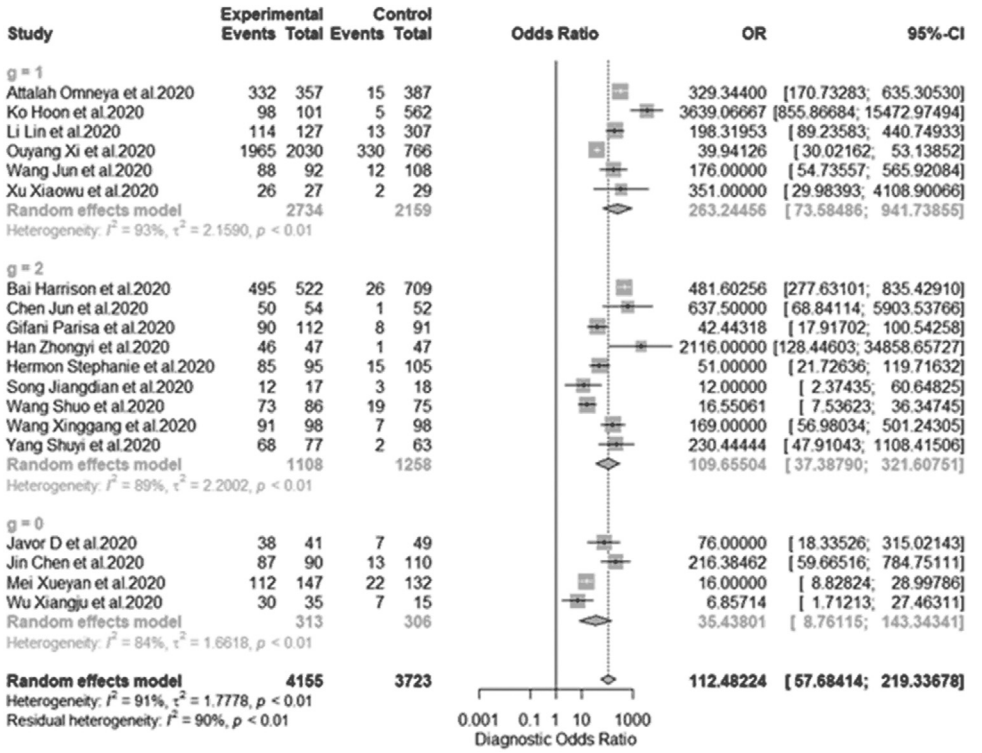

Figure 13. Sub-group analysis of DOR based on deep learning architecture. DOR=diagnostic odds ratio, g represents sub-group analysis of the network architecture when g = 0 (ResNet),g = 1 (ResNet Hybrid) and g = 2 (Other networks).

Similarly, there is a visible non-significant difference in DOR using ResNet, ResNet Hybrid, and other network variants. This result showed that there is a correlation between diagnostic accuracy, which is a function of sensitivity, and network architecture.

There was substantial heterogeneity in all the studies because different Countries were included in the meta-analysis (Austria, Iran, Korea, Egypt, China, U.S.A, Japan, and Italy). These differences in data collected could be a source of potential heterogeneity. One of the potential sources of heterogeneity is the different DL architectures used ranging from ResNet and its variant, ResNet Hybrid and other architectures like AlexNet, GoogleNet, UNet++ and other ensembled DL networks. This can be ascertained by the sensitivity of the sub-group analysis when considering only ResNet architecture for detection with the heterogeneity of (I^2 = 0%). In terms of data source, the result of single-source sub-group analysis shows extremely low heterogeneity in sensitivity and DOR with (I^2 = 18%) and (I^2 = 0%) respectively. This simply means that multi-source datasets serve as a potential source of heterogeneity in DL detection.

Apart from heterogeneity in data type and DL architecture, most of the model's function is based on radiologist performance to serve as the reference standard. It would therefore be very difficult to conclude that DL outperforms its correspondence radiologist interpretation but rather aid and speed up the detection since a good quality image is needed to estimate accurately the DTA of any equipment. Also, most of the DL detection on chest CT only documented sensitivity and specificity, which may lead to overestimation of the benefits of DTA, hence it is recommended that other DTA likelihood ratios and DOR be estimated alongside sensitivity and specificity. The DL algorithm is regarded as a black box because there is no established mathematical formulation to support its performance making it difficult to replicate, and this might also be another source of concern for a wide range of acceptance. Advances in computing hardware and software will lead to better data acquisition and storage with increase quality, enabling further research into how this model behaves and allowing for complete automation of the detection of diseases like COVID-19.

In conclusion, the meta-analysis on DTA of DL detection of COVID-19 was carried out. The results show the high performance of the DL model to detect COVID-19 while establishing that factors such as the source of datasets and DL architectures strongly affect the detection performance of DL algorithms.

ACKNOWLEDGMENTS

The authors acknowledged Professor Sung Ryul Shim at the Department of Preventive Medicine, Korea University College of Medicine, Seoul, Korea for providing statistical guidance during analysis; Dr. Kayode Charles Komolafe for proof reading the article.

FUNDING

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0104505, in part by the National Natural Science Foundation of China under Grant 61701492, in part by the Jiangsu Science and Technology Department under Grant BK20170392, in part by the Suzhou Municipal Science and Technology Bureau under Grant SYG201825. Temitope E. Komolafe receives support from the Chinese Government Scholarship for his doctoral program (CSC No: 2017GXZ021382).

ETHICAL APPROVAL

Institutional Review Board approval was not required because it is a review

INFORMED CONSENT

Written informed consent was not required for this study because the study is a literature review.

REFERENCES

1. Chen Z, Fan H, Cai J, et al. High-resolution computed tomography manifestations of COVID-19 infections in patients of different ages. Eur J Radiol 2020; 126:108972.
2. Li M. Chest CT features and their role in COVID-19. Radiol Infect Dis 2020; 7(2):51–54.
3. World Health Organization (WHO). *Coronavirus Disease (COVID-19) Pandemic*; World Health Organization, Geneva, Switzerland. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed on 2 June 2021.
4. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance 2020; 25 (3):2000045.
5. Guan CS, Lv ZB, Yan S, et al. Imaging features of coronavirus disease 2019 (COVID-19): evaluation on thin-section CT. Acad Radiol 2020; 27 (5):609–613.
6. Lin L, Fu G, Chen S, et al. CT manifestations of coronavirus disease (COVID-19) pneumonia and influenza virus pneumonia: A comparative study. Am J Roentgenol 2020; 216:1–9.
7. Borstelmann SM. Machine learning principles for radiology investigators. Acad Radiol 2020; 27(1):13–25.
8. Cai W, Liu T, Xue X, et al. CT Quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients. Acad Radiol 2020; 27(12):1665–1678.
9. Mader C, Bernatz S, Michalik S, et al. Quantification of COVID-19 opacities on chest CT—evaluation of a fully automatic AI-approach to noninvasively differentiate critical versus noncritical patients. Acad Radiol 2021; 28(8):1048–1057.
10. Li WT, Ma J, Shende N, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2020; 20(1):1–13.
11. Li MD, Little BP, Alkasab TK, et al. Multi-radiologist user study for artificial intelligence-guided grading of COVID-19 lung disease severity on chest radiographs. Acad Radiol 2021; 28(4):572–576.
12. Bai HX, Wang R, Xiong Z, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 2020; 296(3):E156–E165.
13. Li L, Qin L, Xu Z, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology 2020; 296(2). doi:10.1148/radiol.2020200905.
14. Islam MM, Karray F, Alhajj R, et al. A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). IEEE Access 2021; 9:30551–30572.
15. Javor D, Kaplan H, Kaplan A, et al. Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography. Eur J Radiol 2020; 133:109402.
16. Moezzi M, Shirbandi K, Shahvandi HK, et al. The diagnostic accuracy of Artificial Intelligence-Assisted CT imaging in COVID-19 disease: a systematic review and meta-analysis. Inform Med Unlocked 2021; 24:100591.
17. T.E. Komolafe, B.A. Nguchu, H.Sun, et al. (2020). Diagnostic accuracy of deep learning detection of COVID-19: a systematic review and meta-analysis. PROSPERO 2020 CRD42020223202 Available at: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020223202
18. McInnes MD, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. Jama 2018; 319(4):388–396.
19. Dansana D, Kumar R, Bhattacharjee A, et al. Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm. Soft comput. 2020:1-9. doi:10.1007/s00500-020-05275-y. Epub ahead of print. PMID: 32904395; PMCID: PMC7453871.
20. Manikandan R, Dorairajan LN. How to appraise a diagnostic test. Indian J Urol 2011; 27(4):513–519. doi:10.4103/0970-1591.91444.
21. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155:529–536.
22. Shim SR, Kim SJ, Lee J. Diagnostic test accuracy: application and practice using R software. Epidemiol Health 2019; 41.
23. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009; 28(21):2653–2668. doi:10.1002/sim.3631.
24. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21(11):1539–1558. doi:10.1002/sim.1186.
25. Liu JL. The role of the funnel plot in detecting publication and related biases in meta-analysis. Evid Based Dent 2011; 12(4):121–122.
26. Attallah O, Ragab DA, Sharkas M. MULTI-DEEP: a novel CAD system for coronavirus (COVID-19) diagnosis from CT images using multiple convolution neural networks. PeerJ 2020; 8:e10086.
27. Chen J, Wu L, Zhang J, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. Sci Rep 2020; 10(1):1–11.
28. Gifani P, Shalbaf A, Vafaeezadeh M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. Int J Comput Assist Radiol Surg 2021; 16 (1):115–123. doi:10.1007/s11548-020-02286-w. Epub 2020 Nov 16. PMID: 33191476; PMCID: PMC7667011.
29. Han Z, Wei B, Hong Y, et al. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. IEEE Trans Med Imaging 2020; 39(8):2584–2594.
30. Harmon SA, Sanford TH, Xu S, et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 2020; 11(1):1–7.
31. Jin C, Chen W, Cao Y, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020; 11 (1):1–14.
32. Ko H, Chung H, Kang WS, et al. COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation. J Med Int Res 2020; 22 (6):e19569.
33. Mei X, Lee HC, Diao KY, et al. Artificial intelligence—enabled rapid diagnosis of patients with COVID-19. Nat Med 2020; 26(8):1224–1228.
34. Ouyang X, Huo J, Xia L, et al. Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia. IEEE Trans Med Imaging 2020; 39(8):2595–2605.
35. Song J, Wang H, Liu Y. End-to-end automatic differentiation of the coronavirus disease 2019 (COVID-19) from viral pneumonia based on chest CT. Eur J Nucl Med Mol Imaging 2020; 47(11):2516–2524.
36. Wang J, Bao Y, Wen Y. Prior-attention residual learning for more discriminative COVID-19 screening in CT images. IEEE Trans Med Imaging 2020; 39(8):2572–2583.
37. Wang S, Zha Y, Li W, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Resp J 2020; 56 (2):2000775–1—2000775—11.
38. Wang X, Deng X, Fu Q, et al. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans Med Imaging 2020; 39(8):2615–2625.

39. Wu X, Hui H, Niu M, et al. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: a multicentre study. Eur J Radiol 2020; 128:109041.
40. Xu X, Jiang X, Ma C, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering 2020; 6(10):1122–1129.
41. Yang S, Jiang L, Cao Z, et al. Deep learning for detecting corona virus disease 2019 (COVID-19) on high-resolution computed tomography: a pilot study. Ann Trans Med 2020; 8(7):450.
42. Mahmoud H, Taha MS, Askoura A, et al. Can chest CT improve sensitivity of COVID-19 diagnosis in comparison to PCR? A meta-analysis study. Egypt J Otolaryngol 2020; 36:49. doi:10.1186/s43163-020-00039-9.
43. Komolafe TE, Agbo J, Olaniyi EO, et al. Prevalence of COVID-19 diagnostic output with chest computed tomography: a systematic review and meta-analysis. Diagnostics 2020; 10(12):1023.
44. Vafea MT, Atalla E, Kalligeros M, et al. Chest CT findings in asymptomatic cases with COVID-19: a systematic review and meta-analysis. Clin Radiol 2020; 75(11):876–e33.
45. Bao C, Liu X, Zhang H, et al. Coronavirus disease 2019 (COVID-19) CT findings: a systematic review and meta-analysis. J Am Coll Radiol 2020; 17(6):701–709.
46. Kim H, Hong H, Yoon SH. Diagnostic performance of CT and reverse transcriptase polymerase chain reaction for coronavirus disease 2019: a meta-analysis. Radiology 2020; 296(3):E145–E155.
47. Duarte ML, Santos LRD, Contenças ACDS, et al. Reverse-transcriptase polymerase chain reaction versus chest computed tomography for detecting early symptoms of COVID-19. A diagnostic accuracy systematic review and meta-analysis. Sao Paulo Med J 2020; 138 (5):422–432.
48. Böger B, Fachi MM, Vilhena RO, et al. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. Am J Infect Control 2020; 49(1):21–29.
49. Jaeschke R, Guyatt GH, Sackett DL, et al. Users' guides to the medical literature: III. How to use an article about a diagnostic test B. what are the results and will they help me in caring for my patients? JAMA 1994; 271(9):703–707. doi:10.1001/jama.1994.0351033008103.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.acra.2021.08.008.