© Pol J Radiol 2025; 90: e384-e393
DOI: https://doi.org/10.5114/pjr/205344

Received: 13.04.2025
Accepted: 20.05.2025
Published: 05.08.2025

Polish Journal of Radiology

http://www.polradiol.com

Original paper

Optimising strategies for artificial intelligence-assisted classification of viral pneumonias on CT imaging: a comparative study of selective and default approaches

Francesco Rizzetto[1,A,B,C,D,E,F], Luca Berta[2,A,B,C,D,E,F], Giulia Zorzi[3,B,C,D,E], Francesca Travaglini[1,B,E], Diana Artioli[1,B,E], Luca Alessandro Carbonaro[1,A,B,E], Silvia Nerini Molteni[4,B], Chiara Vismara[4,B], Alberto Torresin[5,A,D], Paola Enrica Colombo[3,5,A,D], Angelo Vanzulli[1,6,A,D,E]

[1]Department of Radiology, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy
[2]Department of Oncology, University of Turin, Italy
[3]Department of Medical Physics, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy
[4]Department of Chemical-Clinical and Microbiological Analyses, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy
[5]Department of Physics, Universy of Milan, Italy
[6]Department of Oncology and Hemato-Oncology, Universy of Milan, Italy

Abstract

Purpose: To evaluate how different artificial intelligence (AI)-powered approaches affect human performance in a demanding chest computed tomography (CT) task, such as distinguishing between viral pneumonias.

Material and methods: Three radiologists blindly evaluated 220 chest CT scans of viral pneumonia cases (n = 151 COVID-19; n = 69 other viruses), classifying them with a probabilistic scoring system (COVID-19 Reporting and Data System – CO-RADS) in 2 phases: before (S1) and after (S2) receiving AI classifier results. Two S2 scenarios were investigated: a default approach, with AI predictions available for all cases, and a selective approach, with AI limited to equivocal S1 cases (CO-RADS = 3). Inter-reader agreement (Gwet's AC2) and diagnostic performance were analysed.

Results: Radiologists demonstrated good-to-excellent agreement across all scenarios (AC2 = 0.77-0.81). Evaluation changes between S1 and S2 occurred in 18% of cases, with 29% of cases initially classified as CO-RADS = 3. In these equivocal cases, AI led to an average correct classification rate of 85%. Conversely, when radiologists were confident in their S1 diagnoses (CO-RADS ≠ 3), classification changes in S2 occurred in 7% of cases, preventing incorrect diagnoses in 45% of patients but resulting in missed correct classifications in 55%. Regarding diagnostic performance, S1 accuracy was 78%, with 15% of CO-RADS = 3 cases. In S2, under the default approach, accuracy increased to 81%, with 16% of CO-RADS = 3 cases, whereas the selective approach achieved 79% accuracy with only 10% of CO-RADS = 3 cases. Only the selective approach significantly reduced the proportion of equivocal cases ($p < 0.009$).

Conclusions: A selective AI approach effectively reduces diagnostic uncertainty without introducing unnecessary complexity, emphasising its potential to optimise radiological workflows in challenging diagnostic scenarios.

Key words: artificial intelligence, lung, chest CT, classification, viral pneumonia.

Correspondence address:
Dr. Francesco Rizzetto, Department of Radiology, ASST Grande Ospedale Metropolitano Niguarda, Piazza dell'Ospedale Maggiore 3, 20162, Milan, Italy, e-mail: francesco.rizzetto@ospedaleniguarda.it

Authors' contribution:
A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection

This is an Open Access journal, all articles are distributed under the terms of the Creative Commons Attribution-Noncommercial-No Derivatives 4.0 International (CC BY-NC-ND 4.0). License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

In recent years, the healthcare community has faced unprecedented challenges, including the coronavirus disease 2019 (COVID-19) pandemic, increasing demand for medical imaging, and a growing shortage of radiologists [1,2]. These pressures have accelerated the development and adoption of artificial intelligence (AI) in medical imaging, aiming to improve efficiency and accuracy in tasks like classification, segmentation, detection, and clinical decision-making [3-5].

Numerous studies have applied AI systems to chest computed tomography (CT) [6-8], but many primarily focused on evaluating their technical and clinical feasibility, often limiting their analysis to stand-alone performance metrics [9]. Although some researchers have compared these models directly to the performance of radiologists [10,11], relatively few have investigated their influence on what is termed "diagnostic thinking efficacy" [12,13]. This concept extends beyond diagnostic accuracy to evaluate the added value AI brings to the diagnostic process, including its capacity to enhance radiological decision-making and impact clinical judgment [14]. Importantly, the integration of AI systems into clinical practice does not automatically translate into improved decision-making. Receiving AI-generated results can trigger complex cognitive processes in radiologists, involving confirmation or disconfirmation of their initial judgments, which may affect the final diagnostic outcomes [15].

As learned from the COVID-19 pandemic, distinguishing between different types of viral pneumonia serves as a prime example of a challenging classification task. Typical CT characteristics have been identified for various infectious agents, but significant overlap in findings reduces their specificity and complicates accurate diagnosis, especially in high-pressure clinical scenarios [16-19]. In this context, Meng *et al.* [12] recently demonstrated the benefits of AI-human interaction in improving diagnostic accuracy and confidence when distinguishing between COVID-19 and community-acquired pneumonia. However, they focused solely on the effect of presenting AI results before individual diagnosis, without addressing how different AI usage strategies might variably affect radiologists' performance.

Building on this foundation, our study investigates the impact of integrating an AI classifier into the complex diagnostic process of distinguishing between viral pneumonias. Specifically, we assessed how AI integration influenced diagnostic accuracy and decision-making efficiency, using the COVID-19 Reporting and Data System (CO-RADS) as a structured framework [20]. Furthermore, we compared two distinct AI usage strategies to determine the optimal approach: a "default approach", where AI outputs were available for all cases, and a "selective approach", where AI support was limited to equivocal cases. Our aim was to evaluate not only how diagnostic performance varies across these different strategies but also their different potential to reduce uncertainty and build trust in AI-assisted workflows.

Material and methods

Study design and imaging data

This study received approval from the Local Ethics Committee, and informed consent was waived because data were collected retrospectively and processed anonymously.

The analysis was performed on a random cohort of 220 patients with respiratory symptoms who underwent a CT scan within 15 days of serological evidence of infection from either SARS-CoV-2 (COVID-19, n = 151) or other respiratory viruses (non-COVID-19, n = 69). The CT scans of COVID-19 patients were acquired between March 2020 and April 2021, during the pandemic, while scans of non-COVID-19 patients were performed between January 2015 and October 2019, prior to the emergence of COVID-19. The inclusion of temporally distinct cohorts was essential to avoid misclassification between different viral pneumonia because all control cases were selected prior to the emergence of SARS-CoV-2.

The selected sample size enabled the detection of a minimum accuracy difference of 10%, with a minimum value of 80%, while maintaining a statistical power of 80% [21]. The proportion of COVID-19 to non-COVID-19 cases was chosen to simulate a pandemic-like scenario, with a higher prevalence of COVID-19 cases, to appropriately frame the reader evaluation task, as explained below.

Chest CT examinations were acquired using various CT scanners (Siemens SOMATOM Definition Edge, Siemens SOMATOM Sensation 64, Philips Brilliance) with the patients in supine position with arms over the head for a single breath-hold, in keeping with their compliance.

The acquisition parameters were as follows: tube voltage = 80-140 kV; automatic tube current modulation; pitch = 1; matrix = 512 × 512. All acquisitions were reconstructed with high-resolution thoracic kernels and a slice thickness of 1 mm.

Reader evaluation

Three radiologists with > 10-year experience (Readers 1-3) were enrolled to evaluate the 220 CT scans. The readers were blinded to the original radiologic report and all non-imaging data, including the acquisition date of the CT scans and serologic findings. They were asked to assign each case a CO-RADS score (1 to 5) to indicate the level of suspicion for COVID-19, both before (S1) and after (S2) being provided with an AI-generated probability (0-100%) that the CT scan belonged to a COVID-19 patient. The CO-RADS scoring system is an established framework used to evaluate the likelihood of COVID-19

based on CT imaging findings. Scores of 1 and 2 suggest a diagnosis other than COVID-19, a score of 3 is designated for equivocal cases with indeterminate findings, while scores of 4 and 5 indicate a high or very high suspicion of COVID-19. Before the test, the readers were informed that the AI classifier, based on a multi-layer perceptron architecture [22], had been previously trained and validated on a large cohort of patients, achieving an accuracy of 79% (95% CI: 73-84%) in distinguishing between viral pneumonias during the COVID-19 pandemic [23]. To simulate a comparable clinical scenario, readers were instructed to interpret the CT findings as if the patients were presenting with acute symptoms (e.g. arriving at an Emergency Department).

The test was conducted using a JavaScript-based program [24], which presented the anonymized CT series in a randomized sequence. During the first evaluation (S1), readers assigned an initial CO-RADS score using a dialogue box. The program then provided the AI-generated probability indicating whether the CT scan was likely to represent a COVID-19 case. Subsequently, the program reopened the same patient's CT scan for a second evaluation (S2), allowing readers to either confirm or adjust their initial CO-RADS score through the dialogue box. After completing the evaluation for each patient, the program automatically loaded the next anonymised CT scan. The system recorded all assigned CO-RADS scores and the time taken by readers to analyse each CT scan during both S1 and S2 phases.

Table 1. Demographic characteristics of the study population

Factor	All patients	COVID-19	non-COVID-19
Total	220 (100%)	151 (100%)	69 (100%)
Age	68 (59-78)	67 (59-79)	68 (60-75)
Sex			
Male	159 (72%)	111 (74%)	48 (70%)
Female	61 (28%)	40 (26%)	21 (30%)
Virus			
SARS-CoV-2	151 (69%)	151 (100%)	–
Adenovirus	2 (1%)	–	2 (3%)
Coronavirus 229E/NL63/OC43	4 (2%)	–	4 (6%)
Enterovirus	1 (0.01%)	–	1 (1%)
Influenza virus A/B	28 (13%)	–	28 (41%)
Bocavirus 1/2/3/4	2 (1%)	–	2 (3%)
Metapneumovirus	5 (2%)	–	5 (7%)
Parainfluenza virus 1/2/3/4	6 (3%)	–	6 (9%)
Rhinovirus A/B/C	15 (7%)	–	15 (22%)
Respiratory syncytial virus A/B	6 (3%)	–	6 (9%)

Data analysis

Continuous variables were reported as median values with 1st-3rd quartiles (Q1-Q3) of their distribution; categorical variables were expressed as counts and percentages, with corresponding 95% confidence interval (95% CI) using the Wilson method [25].

The chance-corrected inter-reader agreement for the assigned CO-RADS score was tested using Gwet's second-order agreement coefficient (AC2) with ordinal weights [26]. AC2 was chosen to correct for the partial agreement occurring when comparing ordinal variables with multiple readers and because it is less affected by prevalence and marginal distribution [27-29]. The level of agreement was interpreted following Altman's guidelines [30].

Sensitivity (SE), specificity (SP), accuracy (ACC), positive likelihood ratio (PLR), and negative likelihood ratio (NLR) of human readers in discriminating COVID-19 patients from non-COVID-19 patients were calculated for both S1 and S2. For the latter, two scenarios were examined: one with AI output available for all cases (default approach), and another simulating AI application exclusively to CO-RADS = 3 cases from S1 (selective approach). Performance metrics were computed for each radiologist and as an average among all readers.

The Mann-Whitney test was employed to assess the null hypothesis that readers' evaluation time for S1 and S2 under both default and selective approaches originate from the same distribution. Significant differences in diagnostic performance among these scenarios were assessed using the χ^2 test with post hoc analysis of adjusted residuals [31].

The data analysis was performed using the Real Statistics Resource Pack software (Release 6.8) (www.real-statistics.com) for Microsoft Excel (Microsoft Corporation, Redmond, Washington, USA) and GraphPad Prism 9.5.1 (GraphPad Software, La Jolla, CA).

Statistical significance was established at the $p < 0.050$ level.

Results

The demographic characteristics of the patient population are reported in Table 1. Out of the 220 patients, 159 (72%) were males and 61 (28%) females, with a median age of 68 years (Q1-Q3: 59-78 years). The median interval between CT scans and molecular swabs was of 1 day (Q1-Q3: 0-2 days) for COVID-19 and 3 days (Q1-Q3: 1-6 days) for non-COVID-19 patients.

In S1, the median time required for radiologists to distinguish between COVID-19 and non-COVID-19 pneumonia cases was 10 s (Q1-Q3: 7-14 s). When AI results were introduced in S2 with the default approach, the median evaluation time increased slightly but significantly to 14 s (Q1-Q3: 10-19 s, $p < 0.001$). For cases where readers altered their evaluations, the median additional time

compared to S1 was 4 s (Q1-Q3: 3-9 s). In S2 with the selective approach, the median evaluation time remained at 10 s (Q1-Q3: 7-15 s), showing no significant difference from S1 (p = 0.262). Evaluation time distributions are illustrated in Figure 1.

The CO-RADS scores assigned by each reader in S1 and in S2, considering both the default and selective approaches, are presented in Table 2.

The inter-reader agreement for assigning the CO-RADS score was good-to-excellent across all scenarios. Specifically, in S1, the ordinal-weighted AC2 was 0.77 (95% CI: 0.73-0.81; $p < 0.001$), in S2 with the default approach it was 0.81 (95% CI: 0.78-0.85; $p < 0.001$), and in S2 with the selective approach it was 0.79 (95% CI: 0.76-0.83; $p < 0.001$). Perfect agreement was obtained between the 3 readers in 73/220 (33%) cases in S1, 91/220 (41%) cases in S2 with the default approach, and 82/220 (37%) cases in S2 with the selective approach. For all the evaluations, the category with the highest level of perfect agreement between the readers was the CO-RADS = 2 (range: 58-63%).

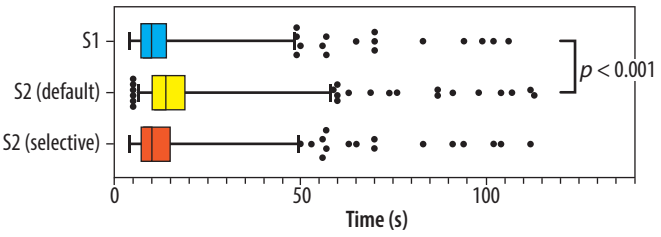

Figure 1. Distribution of readers' chest computed tomography evaluation times in S1 and S2 under default and selective approaches

Table 2. COVID-19 Reporting and Data System (CO-RADS) scores assigned to COVID-19 and non-COVID-19 patients by the readers

S1									
	COVID-19 patients				Non-COVID-19 patients				Total readings
CO-RADS (radiologists)	Reader 1	Reader 2	Reader 3	Total	Reader 1	Reader 2	Reader 3	Total	
1	1 (1%)	0 (0%)	0 (0%)	1 (0%)	9 (13%)	8 (12%)	9 (13%)	26 (13%)	27 (4%)
2	25 (17%)	44 (29%)	29 (19%)	98 (22%)	39 (57%)	52 (75%)	37 (54%)	128 (62%)	226 (34%)
3	25 (17%)	22 (15%)	20 (13%)	67 (15%)	9 (13%)	6 (9%)	16 (23%)	31 (15%)	98 (15%)
4	45 (30%)	56 (37%)	55 (36%)	156 (34%)	9 (13%)	2 (3%)	4 (6%)	15 (7%)	171 (26%)
5	55 (36%)	29 (19%)	47 (31%)	131 (29%)	3 (4%)	1 (1%)	3 (4%)	7 (3%)	138 (21%)
Dataset (true label)	151			453	69			207	660
S2 (default)									
	COVID-19 patients				Non-COVID-19 patients				Total readings
CO-RADS (radiologists)	Reader 1	Reader 2	Reader 3	Total	Reader 1	Reader 2	Reader 3	Total	
1	0 (0%)	1 (1%)	0 (0%)	1 (0%)	9 (13%)	6 (9%)	10 (14%)	25 (12%)	26 (4%)
2	23 (15%)	41 (27%)	28 (19%)	92 (20%)	40 (58%)	56 (81%)	39 (57%)	135 (65%)	227 (34%)
3	26 (17%)	22 (15%)	24 (16%)	72 (16%)	13 (19%)	4 (6%)	15 (22%)	32 (15%)	104 (16%)
4	41 (27%)	43 (28%)	54 (36%)	138 (30%)	5 (7%)	1 (1%)	2 (3%)	8 (4%)	146 (22%)
5	61 (40%)	44 (29%)	45 (30%)	150 (33%)	2 (3%)	2 (3%)	3 (4%)	7 (3%)	157 (24%)
Dataset (true label)	151			453	69			207	660
S2 (selective)									
	COVID-19 patients				Non-COVID-19 patients				Total readings
CO-RADS (radiologists)	Reader 1	Reader 2	Reader 3	Total	Reader 1	Reader 2	Reader 3	Total	
1	1 (1%)	0 (0%)	0 (0%)	1 (0%)	9 (13%)	8 (12%)	11 (16%)	28 (14%)	29 (4%)
2	27 (18%)	47 (31%)	29 (19%)	103 (23%)	42 (61%)	54 (78%)	39 (57%)	135 (65%)	238 (36%)
3	14 (9%)	10 (7%)	18 (12%)	42 (9%)	6 (9%)	4 (6%)	12 (17%)	22 (11%)	64 (10%)
4	54 (36%)	65 (43%)	57 (38%)	176 (39%)	9 (13%)	2 (3%)	4 (6%)	15 (7%)	191 (29%)
5	55 (36%)	29 (19%)	47 (31%)	131 (29%)	3 (4%)	1 (1%)	3 (4%)	7 (3%)	138 (21%)
Dataset (true label)	151			453	69			207	660

Upon reviewing the AI output in S2 with default approach, Reader 1 modified the CO-RADS score in 47/220 cases (21%), Reader 2 in 54/220 cases (25%), and Reader 3 in 16/220 cases (7%). From a clinical perspective, these changes in CO-RADS scores translated into an actual shift in classification between COVID-19, non-COVID-19, or equivocal case in 35/220 (16%) for Reader 1, 26/220 (12%) for Reader 2, and 15/220 (7%) for Reader 3. Notably, when the radiologists were confident in S1 diagnosis (i.e. CO-RADS ≠ 3), a classification change in S2 occurred in 7% of cases, on average. This prevented an incorrect diagnosis in 45% of cases, but in the remaining 55% a correct classification was missed. When considering only the score variations when a CO-RADS = 3 was initially assigned, corresponding to the selective approach, changes were observed in 14/220 (6%) for Reader 1, 14/220 (6%) for Reader 2, and 6/220 (3%) cases for Reader 3. In such a setting, the radiologists moved from an uncertain diagnosis (i.e. CO-RADS = 3) to a correct classification in 85% of the cases, on average. Some relevant examples are shown in Figure 2.

Regarding the diagnostic performance in identifying COVID-19 pneumonia, detailed results are provided in Table 3. Also, Figure 3 and Figure 4 provide a visual, quantitative comparison of the accuracy results between readers and different AI usage approaches.

Considering all the readers, SE = 74% (95% CI: 70-79%), SP = 88% (95% CI: 82-92%), ACC = 78% (95% CI: 75-82%), PLR = 5.95 (95% CI: (4.01-8.83), and NLR = 0.29 (95% CI: 0.25-0.35) were observed in S1, with 15% of cases assigned a CO-RADS = 3. In S2 with the default approach, SE = 76% (95% CI: 71-80%), SP = 91% (95% CI: 86-95%), ACC = 81% (95% CI: 77-84%), PLR = 8.85 (95% CI: 5.44-14.40), and NLR = 0.26 (95% CI: 0.22-0.32) were globally obtained, with 16% of cases assigned a CO-RADS = 3. On the other hand, in S2 with the selective approach, SE = 75% (95% CI: 70-79%), SP = 88% (95% CI: 83-92%), ACC = 79% (95% CI: 75-82%), PLR = 6.28 (95% CI: 4.23-9.34), and NLR = 0.29 (95% CI: 0.24-0.42) were observed, with 10% of cases assigned a CO-RADS = 3. These differences in the proportions of correctly classified, incorrectly classified, and equivocal cases were not statistically significant, except when applying AI results to CO-RADS = 3 cases in the selective approach in S2. Specifically, the number of cases classified as equivocal significantly decreased compared to S1 and S2 with the default AI approach, affecting both the global performance and the individual performance of 2 out of 3 readers. Notably, no statistically significant differences were observed in the ratios of correctly and incorrectly classified cases across the different scenarios. Full details are reported in Table 4 and Figure 5.

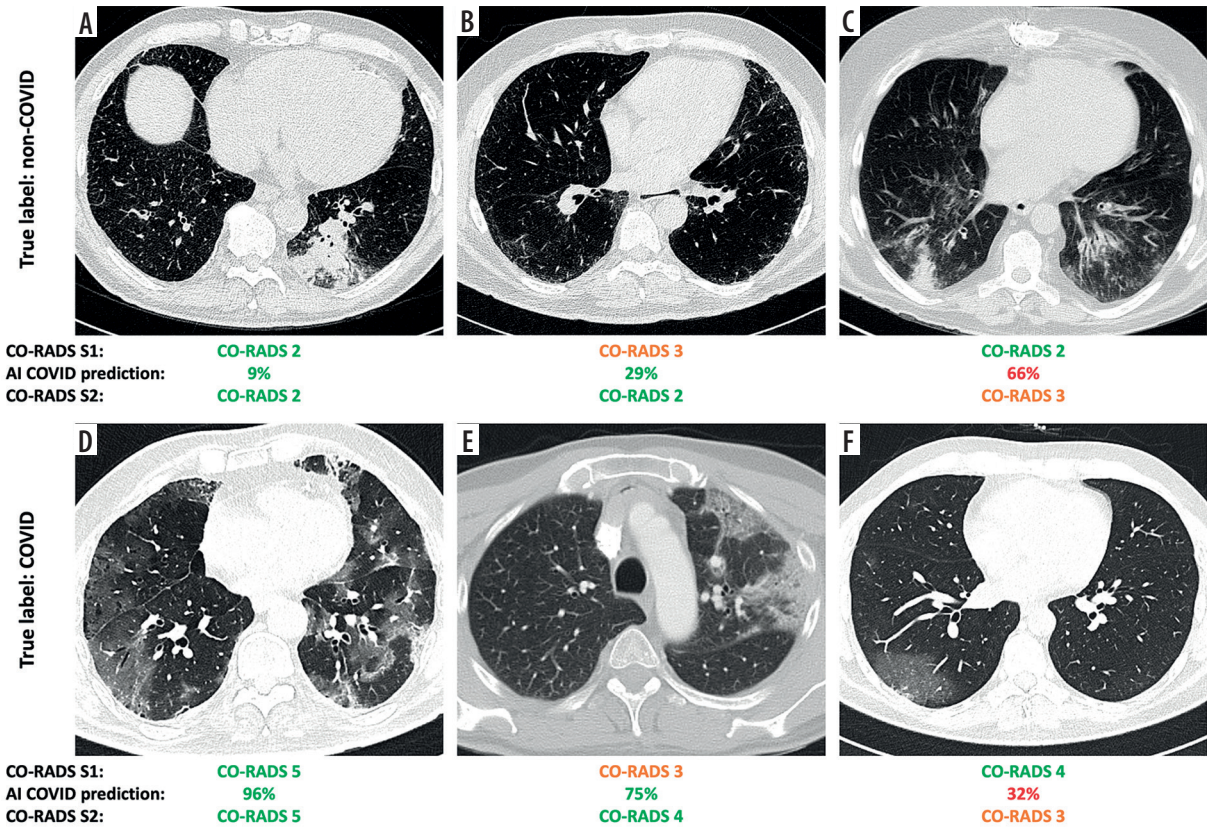

Figure 2. Representative cases illustrating different outcomes in the classification of respiratory viral pneumonias with the aid of an artificial intelligence (AI)-based classifier. For each case, the radiologist-assigned COVID-19 Reporting and Data System (CO-RADS) score is reported before (S1) and after (S2) reviewing the AI-generated COVID-19 probability. In cases where initial uncertainty (S1) was present, AI support often helped resolve the doubt. Conversely, in initially confident assessments, AI input occasionally introduced new uncertainty. The examples shown include pneumonias caused by (**A**) rhinovirus, (**B**) human coronavirus NL63 (HCoV-NL63), (**C**) parainfluenza virus type 3 (PIV3), and (**D-F**) SARS-CoV-2 infection

Table 3. Diagnostic performance of the readers in classifying the patients before (S1) and after (S2) being provided with the results of the artificial intelligence classifier, under both the default and selective approaches

	SE	SP	ACC	PLR	NLR
S1					
Reader 1	79% (71-86%)	80% (68-89%)	80% (73-85%)	3.97 (2.37-6.63)	0.26 (0.18-0.37)
Reader 2	66% (57-74%)	95% (87-99%)	76% (69-81%)	13.84 (4.55-42.04)	0.36 (0.28-0.46)
Reader 3	78% (70-85%)	87% (75-95%)	80% (74-86%)	5.9 (2.94-11.83)	0.26 (0.18-0.36)
Total	74% (70-79%)	88% (82-92%)	78% (75-82%)	5.95 (4.01-8.83)	0.29 (0.25-0.35)
S2 (default)					
Reader 1	82% (74-88%)	88% (76-95%)	83% (77-89%)	6.53 (3.25-13.12)	0.21 (0.14-0.31)
Reader 2	68% (59-76%)	95% (87-99%)	77% (71-83%)	14.78 (4.86-44.91)	0.33 (0.26-0.43)
Reader 3	78% (70-85%)	91% (80-97%)	82% (75-87%)	8.42 (3.63-19.5)	0.24 (0.17-0.34)
Total	76% (71-80%)	91% (86-95%)	81% (77-84%)	8.85 (5.44-14.4)	0.26 (0.22-0.32)
S2 (selective)					
Reader 1	80% (72-86%)	81% (69-90%)	80% (74-85%)	4.18 (2.49-7)	0.25 (0.18-0.36)
Reader 2	67% (58-74%)	95% (87-99%)	76% (69-81%)	14.44 (4.75-43.89)	0.35 (0.28-0.44)
Reader 3	78% (70-85%)	88% (76-95%)	81% (75-86%)	6.37 (3.16-12.82)	0.25 (0.18-0.35)
Total	75% (70-79%)	88% (83-92%)	79% (75-82%)	6.28 (4.23-9.34)	0.29 (0.24-0.34)

95% confidence intervals were reported in parentheses.
SE – sensitivity, SP – specificity, ACC – accuracy, PLR – positive likelihood ratio, NLR – negative likelihood ratio

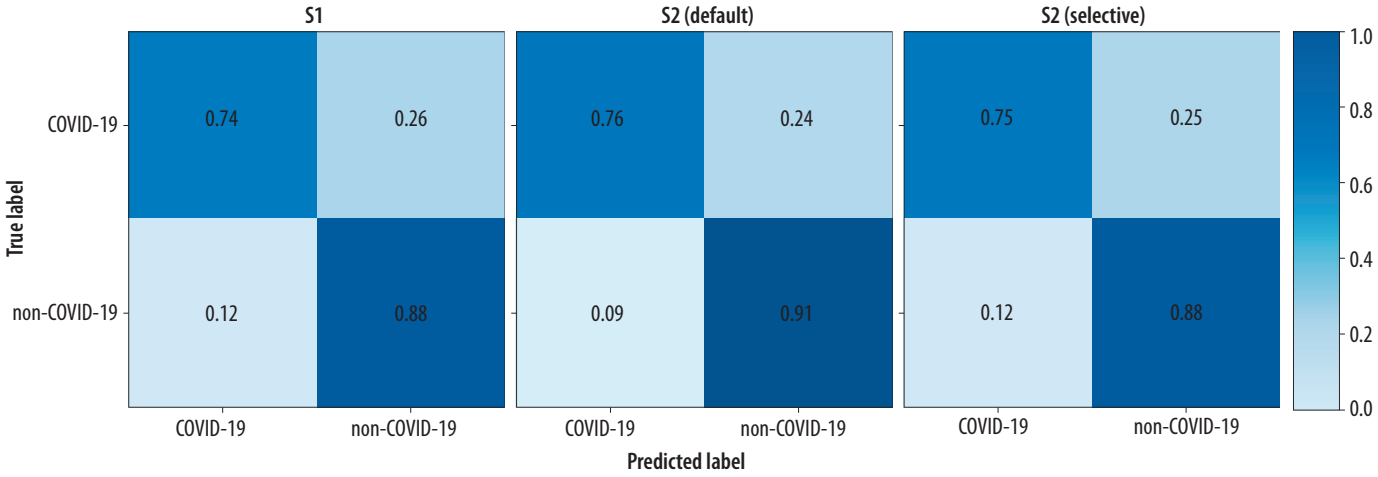

Figure 3. Confusion matrices of the global performance of the 3 readers before (S1) and after (S2) receiving the artificial intelligence classifier results, under both the default and selective approaches

Discussion

In this study, we examined how the performance of radiologists was affected when they received an independent classification from an AI-driven algorithm, focusing on a complex diagnostic task such as the differential diagnosis between viral pneumonias. Additionally, we simulated two usage scenarios: one where the AI tool was available to the readers by default for all cases, and another where AI was selectively applied to cases classified as equivocal.

A good-to-excellent inter-reader agreement (AC2 range: 0.77-0.81) in assigning the CO-RADS score was found in all scenarios. Notably, in S2, there was an increase in the rate of perfect concordance among the radiologists,

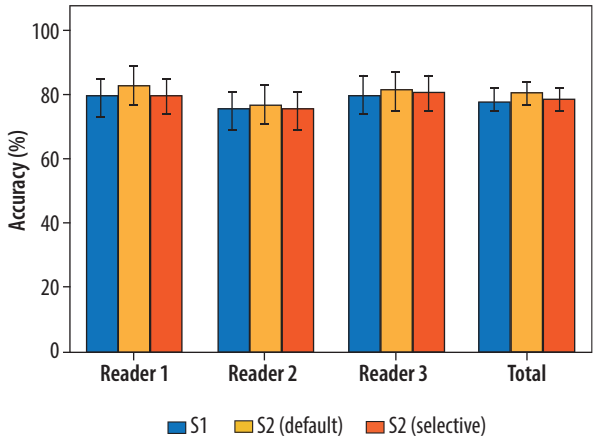

Figure 4. Comparison between the accuracy of the 3 readers before (S1) and after (S2) receiving the artificial intelligence classifier results, under both the default and selective approaches

Table 4. Distribution of correct, incorrect, and equivocal cases among the readers before (S1) and after (S2) being provided with the results of the artificial intelligence classifier, under both the default and selective approaches. The *p*-values from χ^2 test with post hoc analysis of adjusted residuals were reported for each class (significant values are reported in bold)

		Correct cases		Incorrect cases		Equivocal cases	
Reader 1	S1	67%	0.185	17%	0.672	15%	0.762
	S2 (default)	69%	0.360	14%	0.090	18%	0.971
	S2 (selective)	73%	0.895	18%	0.814	9%	0.005
Reader 2	S1	66%	0.172	21%	0.500	13%	0.926
	S2 (default)	68%	0.476	20%	0.273	12%	0.817
	S2 (selective)	71%	0.843	23%	0.727	6%	0.009
Reader 3	S1	67%	0.361	16%	0.589	16%	0.589
	S2 (default)	67%	0.361	15%	0.326	18%	0.817
	S2 (selective)	70%	0.761	16%	0.589	14%	0.129
Total	S1	67%	0.102	18%	0.646	15%	0.904
	S2 (default)	68%	0.328	16%	0.086	16%	0.984
	S2 (selective)	71%	0.957	19%	0.840	10%	< 0.001

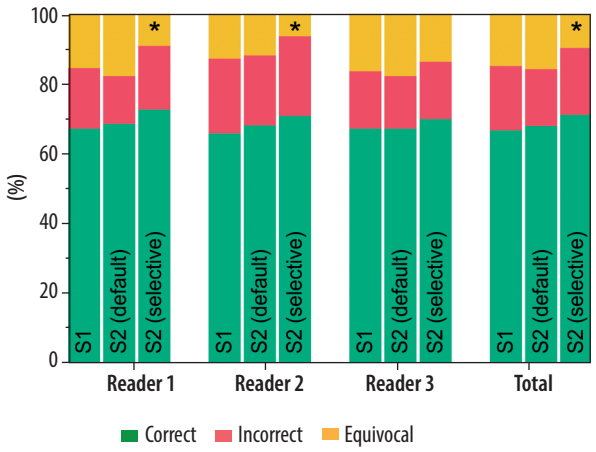

Figure 5. Comparison of the distribution of correct, incorrect, and equivocal cases among the 3 readers before (S1) and after (S2) receiving the artificial intelligence classifier results, under both the default and selective approaches. The asterisks indicate classes that show significant differences from the others ($p < 0.009$)

although the overall effect on inter-reader agreement was limited.

In terms of evaluation time, we observed that the introduction of AI output did not result in substantial delays in the radiological assessment. The default approach added a statistically significant but highly marginal amount of time to the overall process, even when the AI prompted the readers to adjust their scores. Conversely, no significant differences were observed for the selective approach compared to the evaluation without AI.

Interestingly, AI support did not lead to significant differences in the proportions of correctly and incorrectly classified cases, with all readers achieving very similar accuracy levels across the different scenarios. Previous studies have reported AI models achieving accuracy rates ranging from 80% to over 95% in distinguishing between different types of viral pneumonia [32-36]. However, these studies most often concentrated on assessing AI performance itself rather than evaluating how effectively it integrates into the decision-making process of radiologists [14]. Furthermore, the abovementioned results may be attributed to the use of CT datasets encompassing heterogeneous pulmonary conditions, including bacterial infections [10], whose distinct features can ease the classification task.

Meng *et al.* [12], who analysed a well-balanced CT dataset, considering the stage and severity of pneumonia, reported an accuracy level of around 80%, closely aligned with our observations. Differently from our findings, they noted a statistically significant increase in accuracy when incorporating the AI results, but this improvement was relatively slight (less than 3%). Indeed, examining the performance of individual readers in their study, some showed no significant benefit from AI support, regardless of their experience level. Our study yielded similar results, but we also provided full details about the readers' scores and how they changed after receiving AI output. This analysis revealed that clinically meaningful classification changes occurred in 7% to 16% of cases, revealing a fluctuating impact of AI on radiologists. Despite the varying rates of classification changes, the lack of significant performance differences among the readers can be traced back to the score adjustments made when the radiologists were initially confident in their evaluation. This introduced an element of randomness into the final performance, as AI advice prevented an incorrect diagnosis in nearly half of these cases but also led to missing the correct classification in the remaining patients.

Regarding equivocal cases, we confirmed that the AI support reduced the classification uncertainty as reported by Meng *et al.* [12], but only if applied when the radiologist was initially unsure about the diagnosis. In such instances, we achieved a nearly 40% reduction in equivocal cases while maintaining the same level of accuracy. Notably, we determined the confidence level not through self-assessment but by employing the CO-RADS score, thereby providing a more standardised scoring framework.

In previous research [24], we highlighted the potential of AI in assisting radiologists with the classification of equivocal cases, where the accuracy of human readers typically declines. The findings of the current study reinforce this insight, demonstrating that the selective application of AI yields the most significant benefit compared to a default approach applied to all cases. By restricting AI use to equivocal cases, this strategy minimises the risk of introducing new diagnostic uncertainties when radiologists are already confident in their initial assessments. Moreover, selective AI usage aligns with the principles of trustworthy AI by providing targeted, context-specific support, thereby enhancing diagnostic efficacy without overwhelming clinicians with superfluous information. These results highlight the value of integrating AI as an "on-demand" tool in clinical workflows, designed to assist radiologists specifically in situations of diagnostic uncertainty. This approach aligns with the recommendations of Fügener *et al.* [37], who emphasised that the most effective collaboration between AI and human expertise occurs when classification tasks are delegated to the actor best equipped to address them on a case-by-case basis. However, their study also revealed that overconfidence among users often hinders effective delegation to AI. By linking the decision to utilise AI with a predefined equivocal class (such as CO-RADS = 3 cases), this bias can be mitigated, ensuring AI is used effectively to support decision-making while maintaining radiologists' autonomy and confidence in the diagnostic process. Therefore, we believe our findings contribute to the broader goal of developing strategies for trustworthy AI integration into clinical practice, supporting radiologists in complex diagnostic scenarios while enhancing confidence in, and reliability of, AI-assisted workflows.

Our study has some limitations. Firstly, all participating radiologists had extensive exposure to CT imaging of COVID-19 patients during the pandemic. It is possible that less experienced readers would have derived more substantial benefit from the AI assistance, resulting in a greater performance improvement. Additionally, the retrospective design of the study, conducted within a single institution, may have introduced selection bias. For instance, the CT scans included in our study were acquired over several years, even though no substantial changes occurred in local imaging protocols during this period. All examinations were performed using standardised chest CT protocols routinely adopted in clinical practice, including consistent acquisition parameters and reconstruction settings. Moreover, readers were fully blinded to acquisition dates and scanner details, further minimising any influence of temporal factors on diagnostic performance.

Lastly, given the reduced impact of COVID-19 on healthcare systems, the significance of these findings might seem diminished. However, the landscape could swiftly shift with the emergence of new SARS-CoV-2 variants of concern or the outbreak of other rapidly spreading respiratory viruses [38]. Moreover, our focus on viral pneumonias served dual purposes: leveraging existing knowledge and resources, including a validated AI model, while also providing a representative example of a complex diagnostic challenge in chest CT imaging. Importantly, the insights gained from understanding how radiologists interact with AI in this demanding classification task can extend beyond viral pneumonias to other domains, such as non-infectious interstitial lung diseases. Regardless of the specific application, it is crucial to recognize that even the most advanced AI systems with exceptional diagnostic capabilities will not achieve successful adoption unless they enhance diagnostic thinking efficacy and improve clinical decision-making processes. In other words, this underscores the importance of going beyond stand-alone performance metrics to identify the most effective patterns of AI usage within clinical workflows. By developing tailored strategies for AI integration that address specific contexts and minimise potential errors, we can maximise its benefits and build greater trust and confidence in its application.

Conclusions

Our study showed that AI can effectively reduce uncertainty when distinguishing equivocal cases of viral pneumonias on CT imaging. However, its reliability decreases when radiologists are already confident in their diagnoses. By adopting a selective approach that limits AI analysis to equivocal cases, radiologists can maintain comparable diagnostic accuracy while avoiding exposure to additional and potentially misleading AI-generated information to review.

Disclosures

1. Institutional review board statement: The study was approved by the Ethics Committee of Milan Area 3 (22-04-2020/decision number 188).
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

References

1. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open 2021; 11: e045343. DOI: https://doi.org/10.1136/bmjopen-2020-045343.
2. Sagili C. Enhancing diagnostic accuracy with AI: a review of current applications and future directions. Int J Sci Res Comput Sci Eng Inf Technol 2024; 10: 796-805.
3. Hassan H, Ren Z, Zhao H, Huang S, Li D, Xiang S, et al. Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. Comput Biol Med 2022; 141: 105123. DOI: https://doi.org/10.1016/j.compbiomed.2021.105123.
4. Pennisi M, Kavasidis I, Spampinato C, Schinina V, Palazzo S, Proietto F, et al. An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. Artif Intell Med 2021; 118: 102114. DOI: https://doi.org/10.1016/j.artmed.2021.102114.
5. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020; 181: 1423-1433.e11. DOI: https://doi.org/10.1016/j.cell.2020.04.045.
6. Yang W, Sirajuddin A, Zhang X, Liu G, Teng Z, Zhao S, et al. The role of imaging in 2019 novel coronavirus pneumonia (COVID-19). Eur Radiol 2020; 30: 4874-4882.
7. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. Radiology 2020; 296: 172-180.
8. Tse ZTH, Hovet S, Ren H, Barrett T, Xu S, Turkbey B, et al. AI-assisted CT as a clinical and research tool for COVID-19. Front Artif Intell 2021; 4: 590189. DOI: https://doi.org/10.3389/frai.2021.590189.
9. Fusco R, Grassi R, Granata V, Setola SV, Grassi F, Cozzi D, et al. Artificial intelligence and covid-19 using chest ct scan and chest x-ray images: machine learning and deep learning approaches for diagnosis and treatment. J Pers Med 2021; 11. DOI: https://doi.org/10.3390/jpm11100993.
10. Kriza C, Amenta V, Zenié A, Panidis D, Chassaigne H, Urbán P, et al. Artificial intelligence for imaging-based COVID-19 detection: systematic review comparing added value of AI versus human readers. Eur J Radiol 2021; 145: 110028. DOI: https://doi.org/10.1016/j.ejrad.2021.110028.
11. Ni Q, Sun ZY, Qi L, Chen W, Yang Y, Wang L, et al. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. Eur Radiol 2020; 30: 6517-6527.
12. Meng F, Kottlors J, Shahzad R, Liu H, Fervers P, Jin Y, et al. AI support for accurate and fast radiological diagnosis of COVID-19: an international multicenter, multivendor CT study. Eur Radiol 2022; 33: 4280-4291.
13. Yang Y, Lure FYM, Miao H, Zhang Z, Jaeger S, Liu J, et al. Using artificial intelligence to assist radiologists in distinguishing COVID-19 from other pulmonary infections. J Xray Sci Technol 2021; 29: 1-17. DOI: https://doi.org/10.3233/XST-200735.
14. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021; 31: 3797-3804.
15. Jussupow E, Spohrer K, Heinzl A. Radiologists' usage of diagnostic AI systems: the role of diagnostic self-efficacy for sensemaking from confirmation and disconfirmation. Business and Information Systems Engineering 2022; 64: 293-309.
16. Parekh M, Donuru A, Balasubramanya R, Kapur S. Review of the chest CT differential diagnosis of ground-glass opacities in the COVID era. Radiology 2020; 297: E289-E302. DOI: https://doi.org/10.1148/radiol.2020202504.
17. Koo HJ, Choi SH, Sung H, Choe J, Do KH. Radiographics update: radiographic and CT features of viral pneumonia. Radiographics 2020; 40: E8-E15. DOI: https://doi.org/10.1148/rg.2020200097.
18. Rizzetto F, Perillo N, Artioli D, Travaglini F, Cuccia A, Zannoni S, et al. Correlation between lung ultrasound and chest CT patterns with estimation of pulmonary burden in COVID-19 patients. Eur J Radiol 2021; 138: 109650. DOI: https://doi.org/10.1016/j.ejrad.2021.109650.
19. Garrana SH, Som A, Ndakwah GS, Yeung T, Febbo J, Heeger AP, et al. Comparison of chest CT findings of COVID-19, influenza, and organizing pneumonia: a multireader study. Am J Roentgenol 2021; 217: 1093-1102.
20. Prokop M, van Everdingen W, van Rees Vellinga T, Quarles van Ufford H, Stöger L, Beenen L, et al. CO-RADS: a categorical CT assessment scheme for patients suspected of having COVID-19 – definition and evaluation. Radiology 2020; 296: E97-E104. DOI: https://doi.org/10.1148/radiol.2020201473.
21. Akoglu H. User's guide to sample size estimation in diagnostic accuracy studies. Turk J Emerg Med 2022; 22: 177. DOI: https://doi.org/10.4103/2452-2473.357348.
22. Bikku T. Multi-layered deep learning perceptron approach for health risk prediction. J Big Data 2020; 7: 50. DOI: https://doi.org/10.1186/s40537-020-00316-7.
23. Zorzi G, Berta L, Rizzetto F, De Mattia C, Felisi MMJ, Carrazza S, et al. Artificial intelligence for differentiating COVID-19 from other viral pneumonias on CT: comparative analysis of different models based on quantitative and radiomic approaches. Eur Radiol Exp 2023; 7: 3. DOI: https://doi.org/10.1186/s41747-022-00317-6.
24. Rizzetto F, Berta L, Zorzi G, Cincotta A, Travaglini F, Artioli D, et al. Diagnostic performance in differentiating COVID-19 from other viral pneumonias on CT Imaging: multi-reader analysis compared with an artificial intelligence-based model. Tomography 2022; 8: 2815-2827.
25. Wilson EB. Probable Inference, the law of succession, and statistical inference. J Am Stat Assoc 1927; 22: 209-212.
26. Tran D, Dolgun A, Demirhan H. Weighted inter-rater agreement measures for ordinal outcomes. Commun Stat Simul Comput 2020; 49: 989-1003.
27. Quarfoot D, Levine RA. How robust are multirater interrater reliability indices to changes in frequency distribution? Am Stat 2016; 70: 373-384.

28. Vial A, Assink M, Stams GJJM, van der Put C. Safety and risk assessment in child welfare: a reliability study using multiple measures. J Child Fam Stud 2019; 28: 3533-3544.
29. Berta L, Rizzetto F, De Mattia C, Lizio D, Felisi M, Colombo PEE, et al. Automatic lung segmentation in COVID-19 patients: impact on quantitative computed tomography analysis. Phys Med 2021; 87: 115-122.
30. Altman D. Practical statistics for medical research. London: Chapman and Hall; 1991.
31. MacDonald PL, Gardner RC. Type I error rate comparisons of post hoc procedures for I × j chi-square tables. Educational and Psychological Measurement 2000; 60: 735-754.
32. Wang M, Xia C, Huang L, Xu S, Qin C, Liu J, et al. Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. Lancet Digit Health 2020; 2: e506-e515. DOI: https://doi.org/10.1016/S2589-7500(20)30199-0.
33. Zhou M, Yang D, Chen Y, Xu Y, Xu JF, Jie Z, et al. Deep learning for differentiating novel coronavirus pneumonia and influenza pneumonia. Ann Transl Med 2021; 9: 111. DOI: https://doi.org/10.21037/ATM-20-5328.
34. Wang H, Wang L, Lee EH, Zheng J, Zhang W, Halabi S, et al. Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. Eur J Nucl Med Mol Imaging 2021; 48: 1478-1486.
35. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 2020; 296: E156-165. DOI: https://doi.org/10.1148/radiol.2020201491.
36. Mulrenan C, Rhode K, Fischer BM. A literature review on the use of artificial intelligence for the diagnosis of COVID-19 on CT and Chest X-ray. Diagnostics 2022; 12: 869. DOI: https://doi.org/10.3390/diagnostics12040869.
37. Fügener A, Grahl J, Gupta A, Ketter W. Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. Information Systems Research 2022; 33: 678-696.
38. Manirambona E, Okesanya OJ, Olaleke NO, Oso TA, Lucero-Prisno DE. Evolution and implications of SARS-CoV-2 variants in the post-pandemic era. Discover Public Health 2024; 21: 16. DOI: https://doi.org/10.1186/s12982-024-00140-x.