


REVIEW ARTICLE

Prognosis of COVID-19 Using Artificial Intelligence: A Systematic Review and Meta-Analysis

Saeed Reza Motamedian^{1,2}, Negin Cheraghi², Sadra Mohaghegh^{1,2*} , Elham Babadi Oregani², Mahrsa Amjadi², Parnian Shobeiri¹, Niusha Solouki², Nikoo Ahmadi², Yassine Bouchareb³, Arman Rahmim^{4,5}

¹Topic Group Dental Diagnostics and Digital Dentistry, ITU/WHO Focus Group AI on Health, Berlin, Germany

²Dental Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

³Sultan Qaboos University, College of Medicine and Health Sciences, Radiology and Molecular Imaging, Muscat, PO Box 35, PC 123, Oman

⁴Department of Radiology, University of British Columbia, Vancouver, BC, Canada

⁵Department of Physics, University of British Columbia, Vancouver, BC, Canada

*Corresponding Author: Sadra Mohaghegh
Email: mohaghegh.sa77@gmail.com

Received: 05 September 2024 / Accepted: 09 March 2025

Abstract

Purpose: Artificial Intelligence (AI) techniques have been extensively utilized for diagnosing and prognosing several diseases in recent years. This study identifies, appraises, and synthesizes published studies on the use of AI for the prognosis of COVID-19.

Materials and Methods: Electronic search was performed using Medline, Google Scholar, Scopus, Embase, Cochrane, and ProQuest. The systematic approach followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure comprehensive reporting. Studies that examined machine learning or deep learning methods to determine the prognosis of COVID-19 using Computed Tomography (CT) or chest X-Ray (CXR) images were included. Polled sensitivity, specificity, accuracy, Area Under the Curve (AUC), and diagnostic odds ratio were calculated.

Results: A total of 36 articles were included; various prognosis-related issues, including disease severity, mechanical ventilation, or admission to the intensive care unit, and mortality, were investigated. Several AI models and architectures were employed, such as the Siamense model, support vector machine, Random Forest, Extreme Gradient Boosting, and convolutional neural networks. The models achieved 71%, 88%, and 67% sensitivity for mortality, severity assessment, and need for ventilation, respectively. The specificities of 69%, 89%, and 89% were reported for the aforementioned variables.

Conclusion: Based on the included articles, machine learning and deep learning methods used for COVID-19 patients' prognosis using radiomic features from CT or CXR images can help clinicians manage patients and allocate resources more effectively. These studies also demonstrate that combining patient demographics, clinical data, laboratory tests, and radiomic features improves model performance.

Keywords: Artificial Intelligence; Deep Learning; Machine Learning; COVID-19; Prognosis.

1. Introduction

COVID-19 began in early December 2019 and spread rapidly worldwide [1]. The pandemic caused significant shortcomings, abrasion, and burnout in primary and tertiary care healthcare institutions [2]. The increase in hospital admissions has led to a remarkable increase in human errors [3, 4]. Consequently, the care needed for many patients during peak periods could not be adequately provided. Rapid diagnosis of COVID-19 and determination of the severity of infection enable healthcare professionals to better control the virus spread and manage increased hospital overloads, aiming to improve the quality of treatments [5]. Despite the recent ease in the COVID-19 situation, the lessons learned will help better manage future pandemics.

Triage is essential in patient management, alleviating the pressure on medical departments [6]. COVID-19 patients indicate various presentations and outcomes, ranging from asymptomatic to critical situations that may lead to death [7]. Based on the severity of the infection, it is essential to determine whether patients can receive care at home or should be admitted to COVID wards or Intensive Care Units (ICU). It is also important to diagnose patients who require mechanical ventilation (whether non-invasive or via intubation) [8]. Accordingly, prediction models can help triage systems by automatically combining predictors to estimate the severity, ventilation, or intensive care needed and the possibility of death, hence allocating adequate resources [5]. Indeed, determining these factors at the early stage helps clinicians prioritize patients during peak periods [9].

Conventional methods have the advantage that, since they mainly rely on rule-based scoring systems, they can be coherently understood and deployed. However, in practice, they are often time-consuming and prone to human errors. By contrast, in the case of Artificial Intelligence (AI)-based methods, rapid and large-scale analyses with higher reproducibility and accuracy can be developed. Artificial intelligence (AI) is the science of making intelligent programs or applications that mimic human intelligence, perform rapid assessments, and make accurate decisions [10]. AI can analyze a large amount of data in a short time and potentially provide accurate outcomes [11]. The recent deployment of AI models can be justified by

considering its merits. AI can alleviate the need of doing some repetitive tasks by physicians and technical staff, accelerate time-consuming processes, enhance quantification and interpretation, improve diagnostic reproducibility, and provide clinically relevant information [12]. Accordingly, AI techniques have been widely used for clinical purposes such as diagnosis, analysis of medical images, extensive data collection, research and clinical trials, management of intelligent health records, and prediction of outbreaks [13-16]. To better understand AI, it is important to explore its subfields, especially machine learning and deep learning. Machine learning is concerned with creating algorithms that let computers learn from data and make predictions, while deep learning uses multi-layered artificial neural networks to analyze complex datasets and has shown remarkable results in domains like image recognition and natural language processing [17]. On the other hand, concerns about the reproducibility, generalizability, and explainability of AI models remain to be solved, presently hindering AI translation and implementation in clinical practice [18, 19].

Several studies have already used prediction methods (e.g., rule-based scoring systems or advanced machine learning models) to accelerate patient assessment and ease pressure on frontline departments [20, 21]. Conventionally, radiomic features are extracted from the previously segmented Region Of Interest (ROIs) [21]. This procedure usually involves manually or semi-automatically defining the ROI, then extracting hand-crafted features that characterize the ROI's form, texture, and intensity, among other attributes. Deterministic approaches, which use mathematical formulas to quantify characteristics like volume, compactness, and surface area, are used to generate these characteristics. To identify complex patterns in the imaging data, higher-order statistics may also be used. To help in diagnosis and treatment planning, the collected attributes are then correlated with clinical results [22]. Using deep learning models, features can be implicitly derived from images without the necessity of defining a region of interest. Quantitative features extracted from images can help identify relevant disease biomarkers, impact the clinical decision-making process, and provide means of predicting lesions' growth and characteristics [23].

Several studies examined the performance of these models in prognosis using Computed Tomography (CT) or Chest X-Ray (CXR) images [24-27]. Systematic reviews on the application of AI for screening or diagnosis, prevention, and treatment planning of COVID-19 have been performed recently [28-31]. Unlike these reviews, this systematic review focuses on the use of AI for the prognosis of COVID-19 and quantitatively analyzes the performance of the deployed models using variables such as sensitivity, specificity, and Area Under the Curve (AUC) [32]. Building on the existing knowledge from several recent studies on this topic, we attempted to provide an updated review with a special focus on prognosis, infection severity, need for ventilation or ICU, and mortality, and also report on the most commonly used performance parameters, including AUC, accuracy, sensitivity, and specificity.

2. Materials and Methods

2.1. Protocol and Registration

The question of this study, according to PICO format, was as follows: To compare the function (O) of AI models (I) in determining the prognosis of COVID-19 patients (P) with the specified ground truth (C). The study was carried out according to the preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA) guidelines [33]. The review was registered in Prospero with the number of CRD42022351594.

2.2. Eligibility Criteria

The inclusion and exclusion criteria used for selecting the articles are presented in Table 1.

2.3. Information Sources

The electronic search was conducted in PubMed, Scopus, Embase, Web of Science, Cochrane, and ProQuest databases for English articles published before March 2023. Additionally, Google Scholar was utilized as a search engine to identify scholarly literature.

2.4. Search

The queries are indicated in Table 2. English articles were included, and no restriction was set on the publication date. Also, no filter was used for the type of study.

2.5. Study Selection

The electronic search results were entered into EndNote 20 software, and duplicate papers were omitted. Next, four authors (M.A., N.Ch, N.A, and N.S) screened the titles and abstracts of the remaining studies according to the abovementioned inclusion and exclusion criteria. For the final decision, the full text of the selected studies was assessed. Any uncertainty over the final decision was resolved by an independent expert (E.B.).

2.6. Data Collection

Four authors (M.A., N.Ch, N.A, and N.S) performed the data extraction. They tabulated the data as follows: author and year of publication, procedure (disease severity, prognosis, need for ICU, ventilation requirement, mortality and segmentation), dataset size, age of patients, imaging modality (CT or CXR images), task (classification or segmentation), pre-processing and augmentation of images, model architectures and their performance.

2.7. Risk of Bias

The included articles were assessed according to the quality assessment of diagnostic accuracy studies (QUADAS-AI) tool [34], which has been widely used in the AI systematic reviews [35-38]. The following domains were used to evaluate the Risk Of Bias (ROB): patient selection, index test, reference standard, and flow and timing. Studies with three or more items with a low risk of bias were considered overall low. Those with only one item at low ROB were evaluated as overall high ROB; others were deemed unclear ROB.

2.8. Synthesis of Results and Meta-Analysis

The accuracy of the AI models in predicting the need for ventilation, severity assessment, and

Table 1. Inclusion and Exclusion Criteria

	Inclusion Criteria	Exclusion Criteria
Population= Covid-19	Studies analyzing patients suffering from COVID-19 infection.	None
Intervention= AI	Literature used artificial intelligence, deep learning, or machine learning techniques based on radiographic images, including CXR and CT images.	If the results were not reported merely based on radiographic images, and combined with clinical and laboratory information.
Comparison= Gold standard (actual condition of the patients)	None	Studies that did not specify the ground truth
Outcome: Prognosis	Studies were performed to determine the severity, prognosis, recurrence, mortality, and survival rate of the COVID-19 disease. Also, studies that assessed the treatment outcomes were included.	None

Table 2. Search Queries

Motor Engine	Search Query	Result
PubMed	("artificial intelligence"[MeSH] OR "AI" OR "machine learning"[MeSH] OR "ML" OR "deep learning"[MeSH] OR "DL" OR "big data"[MeSH] OR "computer aided" OR "diagnosis, computer assisted"[MeSH Terms] OR "neural network") AND ("COVID-19"[MeSH] OR "SARS-CoV-2"[MeSH] OR "coronavirus"[MeSH] OR "covid-19" OR "sars-cov-2" OR "coronavirus") AND ("prognosis"[MeSH] OR "mortality"[MeSH] OR "prognostic" OR "prediction" OR "severity" OR "predict" OR "Treatment Outcome"[MeSH] OR "mortality" OR "survival" OR "recurrence")	782
Google Scholar	("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "big data" OR "computer aided") AND ("COVID-19" OR "SARS-CoV-2" OR "coronavirus") AND ("prognosis" OR "prognostic" OR "severity")	~21000
Scopus	("artificial intelligence" OR "AI" OR "machine learning" OR "ML" OR "deep learning" OR "DL" OR "big data" OR "computer aided" OR "diagnosis, computer assisted" OR "neural network") AND ("COVID-19" OR "SARS-CoV-2" OR "coronavirus") AND ("prognosis" OR "prognostic" OR "prediction" OR "severity" OR "predict" OR "Treatment Outcome" OR "mortality" OR "survival" OR "recurrence")	620
Embase	("artificial intelligence" OR "AI" OR "machine learning" OR "ML" OR "deep learning" OR "DL" OR "big data" OR "computer aided" OR "diagnosis, computer assisted" OR "neural network") AND ("COVID-19" OR "SARS-CoV-2" OR "coronavirus") AND ("prognosis" OR "prognostic" OR "prediction" OR "severity" OR "predict" OR "treatment Outcome" OR "mortality" OR "survival" OR "recurrence")	148
Web of Science	('artificial intelligence' OR 'ai' OR 'machine learning' OR 'deep learning') AND ('covid-19' OR 'sars-cov-2' OR 'coronavirus') AND ('prognosis' OR 'severity' OR 'mortality')	216
Cochrane	("artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "DL" OR "big data" OR "computer aided" OR "neural network") AND ("COVID-19" OR "SARS-CoV-2" OR "coronavirus") AND ("prognosis" OR "prognostic" OR "predict" OR "Treatment Outcome" OR "mortality" OR "survival" OR "recurrence")	68
ProQuest	SU.X("deep learning") AND "covid 19" AND "prognosis"	271

mortality in COVID-19 patients was determined using Receiver Operator Characteristic (ROC) curves, as evaluated by the AUC value and sensitivity and specificity (True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) values), if available [39]. The meta-analysis included studies that

evaluated the sensitivity and specificity of various AI models for predicting the need for ventilation, severity assessment, and mortality in COVID-19 patients. The heterogeneity of included studies was evaluated using I² and χ^2 statistics and was deemed significant if I² was more than 50% or the p-value was less than 0.05.

To account for the predicted heterogeneity of investigations [40], a random-effects model (DerSimonian-Laird method) was used. Deeks and colleagues, on the other hand, performed a simulation study of tests for publication bias in Diagnostic Test Accuracy (DTA) reviews in 2005 [41]. Hence, Deeks' test is suggested and should be preferred for DTA meta-analyses. Furthermore, Diagnostic Odds Ratio (DOR) is defined as the ratio of the chances of testing positive for the target condition to the odds of testing positive without the target condition [42] (Equation 1):

$$\text{DOR} = \frac{TP/FN}{FP/TN} = \frac{TP \cdot TN}{FP \cdot FN} = \frac{LR^+}{LR^-} \quad (1)$$

Using STATA version 17 (StataCorp LP, College Station, TX, USA), all plots were generated. In addition, all analyses were conducted using STATA 17.0 software. Accordingly, “midas” and “metandi” were utilized.

3. Results

3.1. Study Selection

After analyzing the titles and abstracts of the 1528 studies, the full texts of 193 articles were assessed for eligibility. Ultimately, 36 articles were retained and included for full subsequent analysis (Figure 1).

3.2. Study Characteristics

The results of data extraction are presented in Table 3. Among the included articles, 24 studies with a total sample size of 358181 examined the severity of the disease [15, 24, 25, 43–63]. Amongst them, 20 studies used CT [8, 24, 25, 43, 45, 46, 48, 52–54, 56, 58, 59, 61–67] images and 16 studies used CXR images [26, 27, 44, 47, 49–51, 55, 57, 60, 68–73]. They examined the images of the patients based on the image features, the extent of the infection and lung involvement and then classified the patients into two [47, 49, 50, 55–57, 61, 63, 74], three [24, 43, 58], four [48, 52, 54, 60] or five [53, 59] groups. Three studies differentiated only critical patients admitted to the ICU or deaths occurring before or after ICU admission [26, 51, 62]. Studies used different models for the classification of the severity, including Supported Vector Machine (SVM) [43, 48, 49, 74], Random Forest (RF) [49, 58], COV-CAF [53], LungDoc [63], COVID-Net CXR-S

[50], ResNet-50 and Inception models [44, 47], Siemens healthiness algorithms [24, 62] and different neural networks [26, 45, 46, 48, 51, 52, 54–56, 60, 61].

Nine studies examined the need for ICU or mechanical ventilation based on CT [8, 63, 64] or CXR [27, 68, 69, 71–73] images with a total sample size of 8239 patients. They reported their classification results as a binary outcome (e.g., whether ventilation or ICU was needed or not). They used different DL- or ML-models including LungDoc [63], Siemens healthcare [8], DenseNet121 [71], Balanced Random Forest (BRF) [68], RF and Convolutional Neural Network (CNN) [27, 64, 72, 73].

Nine studies classified patients according to mortality (whether the patients survived or not). Six studies examined CXR [26, 27, 68–70, 72, 73], and three studies examined CT images [64–67]. The total sample size of these studies was 18993 patients. The evaluated architectures were Qure.ai Technologies [16], VGG [70], Extreme Gradient Boosting (XGB) [65, 68], RF, and neural networks [27, 64, 66].

3.3. Risk of Bias and Applicability

Two included studies were at high risk of bias (46, 50), seven had unclear risk of bias, and others were at low risk of bias (Figure 2). The index test was the most problematic domain. Also, some studies did not mention the time taken to read CT or CXR records.

3.4. Results of Individual Sources of Studies

Determining the prognosis of COVID-19 disease can generally be classified into three groups: disease severity, mechanical ventilation or need for ICU, and mortality.

3.4.1. Disease Severity

Among articles that examined the disease severity, a range of 0.65 – 0.98 was reported for the area AUC. The DOR of the studies included in this category was between 7.3 and 297.6. The best AUC was reported by Irmak *et al.* [60], who proposed an automated CNN

Table 3. Results of data extraction

Author, Year	Desired outcome	Dataset Size (train/test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Ortiz 2022 [43]	Severity (mild, moderate, and severe)	596 Test: 596	Public	NA	CT	Classification	NA	SVM	Infection histogram	Accuracy	0.81
								SVM with Latent Dirichlet Allocation (LDA)			0.95
								DenseNet121			0.81
								ResNet50			0.76
								InceptionNet			0.75
Dinh 2022 [44]	Severity	2629 (2479/150)	Public	NA	CXR	Classification	P: NA A: height_shift_range, rotation_range, horizontal_flip, brightness_flip, width_shift_range, and rescale	Hybrid EfficientNet-DOLG	Image features	Micro-average	0.82
Chamberlin 2022 [45]	Severity	93 M= 57 F= 36	Private	59	CT	segmentation	P: + A: NA	dcCNN	Lesion region	Opacity score	0.89
Balaha 2022 [46]	Severity	15535 (5,159/10,376)	Private	NA	CT	classification	*No augmentation	CNN	Image features	AUC	0.99
							Normal augmentation				0.65
							CC-GAN				0.99
							CycleGAN				0.99
								ResNet-50			0.85
Ahmad 2022 [47]	Severity (Improved, Deterioration)	877 (582/295)	Private	NA	CXR	classification	NA	InceptionV3	Severity score	AUC	0.88
								InceptionResNetV2			0.89
								CheXNet			0.92
								EfficientNet			0.87
								An ensemble of five pre-trained CNN models with SVM			0.81
Shalbat 2022 [48]	Severity (normal, mild, moderate, and severe)	683 Train: 547 Validation: 68 Test: 68 F: 275 M:408	Private	NA	CT	Classification	P: extracting the central region of the images and converting to binary images A: NA	An ensemble of five pre-trained CNN models with fine-tuning and softmax	Image features	Accuracy	0.85

Author, Year	Desired outcome	Dataset Size (train/test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)		
Abbsai 2022 [49]	Severity prediction (More or less severe)	278 (114/164)	Public	NA	CXR	Classification	P: image resizing (313 × 313 pixel(s), de-noising, and contrast stretching A: NA	SVM	Severity Score	AUC	0.96		
								RF*		F1	0.90		
								XGBoost		AUC	0.84		
										F1	0.96		
Aboutalebi 2021 [50]	Severity (two levels)	258 (208/150) M:161 F:97	Private	59.11	CXR	segmentation	P: Cropped, resampled to 480*480. A: Translation, rotation, horizontal flip, zoom, and intensity shift	COVID-Net CXR-S	Opacity	Sensitivity	0.92		
Gouda 2020 [24]	Severity (mild, severe, critical)	120 Test: 120 F: 22 M: 98	Private	52.63	CT	Classification	NA	Siemens Healthineers	Total severity score	AUC	0.94		
Li, Z 2021 [52]	Lung and lesion segmentation	Lung: 5750CT Lesion:1117CT (train/test=4:1)	Public	NA	CT	Classification	P: Lung segmentation, Lesion segmentation, 3D visualization A: Left-right flip, Top-bottom flip, Top-bottom and Left-right flip, ± 15-degree rotation, ± 30-degree rotation with 12.5% probability	U-net++	Lung regions	DSC	0.97		
											Lesion regions		
													0.84

Author, Year	Desired outcome	Dataset Size (train/ test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Shan 2021 [25]	Lesion segmentation Severity (severe or non-severe)	549 Train: 249 Test: 300 F: NA M: NA	Private	>18	CT	Segmentation Classification	NA	VB-Net neural network SVM, C-SVM	Lesion regions	DSC	0.91
									MOI	Accuracy	0.73
									POI		0.72
Ibrahim 2021 [53]	Severity	1252 Train: 1126 Test: 126 F: NA M: NA	Private	NA	CT	Classification	P: Converting the 3D-CT volumes to 2D-slices A: NA	A novel computer- aided framework (COV-CAP)	Amount of ground glass, consolidation, and lung involvement	AUC	0.97
										Accuracy	0.96
										Sensitivity	0.96
										specificity	0.98
Qiblawey 2021 [54]	Lung segmentation Lesion segmentation	1139 F: NA M: NA	Public	NA	CT	Segmentation Classification	P: Normalizing the image intensity values and mapping to pixel values, changing the intensity interval to create consistent image and resizing the images to 256×256 A: applying rotations of 90, -90, and 180 degrees for CT images and ground truth masks	DenseNet 121 UNet DenseNet 201 FPN ED-CNN	Lung region Lesion region Infection percentage	DSC	0.97
											0.94
										Accuracy	0.97
										Sensitivity	0.94
										Specificity	0.95
Imak 2021 [60]	Severity (mild, moderate, severe, & critical)	3260 images Train: 1956 Validation: 652 Test: 652	Public	NA	CXR	Classification	NA	CNN	Opacity degree and lung involvement	AUC	0.98
										Accuracy	0.95
										Sensitivity	0.96
										Specificity	0.98

Author, Year	Desired outcome	Dataset Size (train/test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Jiao 2021 [55]	Severity (critical or non-critical)	1834 Train: 1285 Validation: 183 Test: 366 F:1177 M:1132	Private	56	CXR	Classification	P: Normalizing CXRs to the range 0–1 A: NA	EfficientNet-B0	Image features	AUC	0.75
Lassau 2021 [56]	Severity (low-moderate- or high-risk)	1626 Train: 646 Test: 980	Private	62.6	CT	Classification	P: Resizing the CT scans to a fixed pixel spacing of (0.7 mm, 0.7 mm, 10 mm), Applying a specific windowing on the HU intensities A: NA	A neural network containing two submodels: Resnet50 EfficientNet B0	Infection percentage	AUC	0.76
Elsharkawy 2021 [57]	Severity (low severity or high severity)	200 Test: 200 F: NA M: NA	Public	NA	CXR	Classification	P: Segmentation of the lung region, Enhancement of contrast, Extracting candidates of abnormal tissues A: Scale, Rotation, Translation	MGRF and a neural network-based fusion system	Gibbs energy CDF	Accuracy	0.98
										Sensitivity	1.00
										Specificity	0.97

Author, Year	Desired outcome	Dataset Size (train/test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Cai 2020 [58]	Lung and lesion segmentation	99 Test: 99 F: 41 M: 58	Private	54.5	CT	Segmentation	P: Lung CT window level setting, Batch normalization A: Adding noise, Random rotation, Random shift, Random shear, Random zoom	U-Net	Lung region	DSC	0.98
	Severity (model I: moderate vs severe+critical, model II: severe vs critical)					Lesion region			0.77		
						Lung involvement moderate vs severe+critical			AUC		0.82
									Accuracy		0.75
									Sensitivity		0.79
									Specificity		0.70
AUC	0.78										
Accuracy	0.72										
Sensitivity	0.79										
Specificity	0.66										
Purkayastha 2020 [59]	Lung segmentation	981 Train: 784 Test: 197 F: 475 M: 506	Private	48.9	CT	Segmentation	NA	A deep convolutional neural network	Lung region	DSC	NA
	Severity					Classification			AUC		0.74
									Accuracy		0.79
									Sensitivity		0.62
									Specificity		0.85
									AUC		0.81
Ho 2021 [61]	Severity (high risk or low risk)	297 Test: 297 F: 169 M: 128	Private	60	CT	Classification	P: Lung segmentation, Removing background, Lesion classification A: NA	CNN	Image features	Accuracy	0.91
	Sensitivity									0.51	
	Specificity									0.92	
	TN, FN, TP, FP									50, 4, 3, 0	

Author, Year	Desired outcome	Dataset Size (train/ test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Gieraens 2020 [62]	Severity (critical patient)	250 Test: 250 F: 133 M: 117	Private	66.6	CT	Segmentation Classification	N/A	Siemens Healthineers	Lung involvement CT score	AUC	0.87
											0.88
									Severity score	AUC	0.80
											0.79
											0.78
Li, Y 2020 [63]	Severity (non-severe and in progress-to-severe)	123 Test: 123 F: 61 M: 62	Private	64.43	CT	Classification	N/A	LungDoc	Consolidation volume	Specificity	0.71
										AUC	0.80
										Accuracy	0.75
										Sensitivity	0.74
										Specificity	0.70
Mushaq 2020 [26]	Severity (critical patient)	697 Test: 697 F: 232 M: 465	Private	62	CXR	Classification	N/A	qXR,v2.1 c2, Qure.ai Technologies	Lung involvement	AUC	0.77
											0.77
											0.66
											0.77
											0.70

Author, Year	Desired outcome	Dataset Size (train/ test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Kohli 2021 [8]	Ventilation requirement ^t	740 Test: 740 F : 257 M: 482	Private	59	CT	Segmentation Classification	NA	Siemens Healthcare version 2.5.2	OS1	AUC	0.92
										Sensitivity	0.61
										Specificity	0.98
									OS2	AUC	0.91
										Sensitivity	0.58
									OP	Specificity	0.97
										AUC	0.91
										Sensitivity	0.58
										Specificity	0.97
										AUC	0.97
Kulkarni 2021 [71]	Ventilation requirement ^t	528 Train: 410 Test: 118 F:170 M:358	Private	54.4	CXR	Classification	P: Resizing to 224×224 pixels, Centre cropped A: Random combination of right or left rotation (maximum 30°), Random cropping, Random lighting	DenseNet121	Image features	TN, FN, TP, FP	92, 13, 30, 18
										Accuracy	0.97
										Sensitivity	0.70
										Specificity	0.84
										AUC	0.88
Mumera 2022 [72]	Need of ICU Mortality	582 Train: 408 Validation: 105 Test: 69 F: NA M: NA	Private	NA	CXR	Classification	P: Take all images to the same dynamic range and remove elements that were not part of the image A: NA	CNN	Image features	Sensitivity	0.85
										Specificity	0.81
										AUC	0.75
										Sensitivity	0.71
										Specificity	0.76
Jordan 2022 [73]	Need for ICU Mortality	2456 Train: 2000 Test: 456 F: NA M: NA	Private	55.3	CXR	Classification	P: rescaling images to an isotropic resolution, resampling, and normalizing A: NA	CNN	Geographical extent of airspace opacities	AUC	0.87
											0.82

Author, Year	Desired outcome	Dataset Size (train/ test), Sex	Dataset Classification	Age	Imaging Modality	Task	Pre-Processing (P) Augmentation (A)	Model Architecture	Parameters/ Features	Reported Prognostic Performance	Performance (%)
Bae 2021 [27]	Ventilation requirement †	691 (N/A) F:328 M:363	Private	56	CXR	Classification	P: Segmentation of lung and artifact, Average, histogram matching, Automatic cropping A: Flipping, Rotation, Translation	Machine learning (RF, LDA and QDA)	Image features	AUC	0.78
								CNN		Sensitivity	0.72
								Specificity		0.72	
								AUC		0.75	
								Sensitivity		0.64	
								Specificity		0.73	
								AUC		0.78	
								Sensitivity		0.70	
								Specificity		0.73	
								AUC		0.75	
								Sensitivity		0.59	
								Specificity		0.74	
Aljouie 2021 [68]	Ventilation requirement †	1508 Train: 1208 Test: 300 F:651 M:857	Private	54.8	CXR	Classification	P: Feature normalization, Feature selection A: SMOTE, ADASYN sampling approach, RUS	BRF	Image features	AUC	0.76
				Accuracy				0.52			
	Mortality	1513 Train: 1212 Test: 301 F: 653 M:860		54.8		XGB + ADASYN		AUC		0.74	
				Accuracy	0.71						
Bernjejo 2022 [64]	Mortality						P: clipping the intensities outside the range, rescaling A: NA		Lesion region		AUC
	Admission to the Intensive Care Units (ICU)	103 (60/93) M=39 F=64	Private	64.83	CT	segmentation		CNN			
	Need for mechanical ventilation										
											0.68

[illegible][illegible]

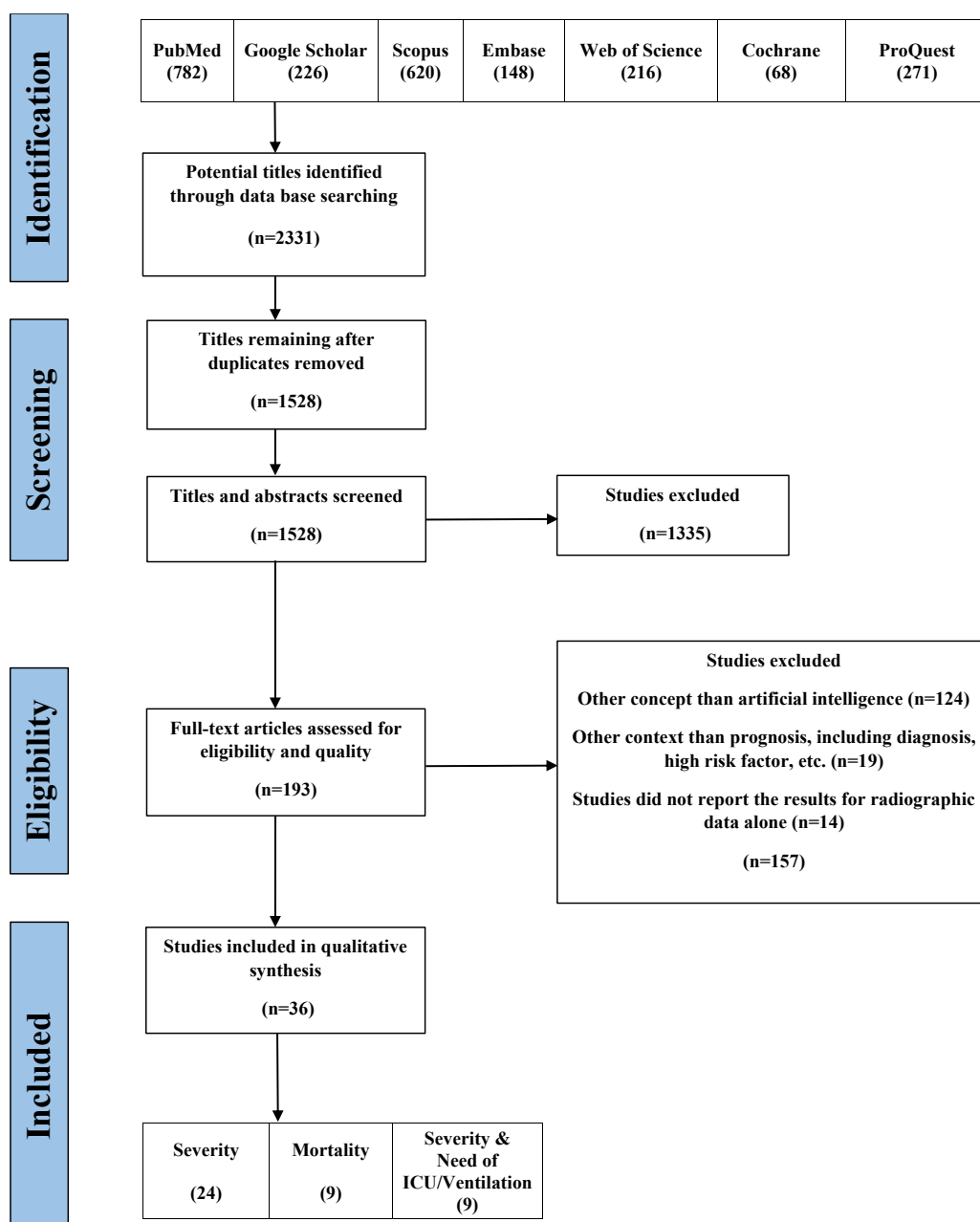


Figure 1. PRISMA flow diagram (literature search strategy and study selection)

model for severity classification into four groups: mild, moderate, severe, and critical. The amount of ground glass, consolidation, and lung involvement from 3260 chest X-ray images was evaluated in the study. They reported an average accuracy of 0.95, a sensitivity of 0.98, and a specificity of 0.96.

The lowest AUC was related to Balaha *et al.* [46], which used a CNN model with normal augmentation to analyze the image features of the 15535 CT images. It was reported that altering the augmentation approach or even eliminating it can increase the AUC significantly.

Among reported accuracies, a range of 0.72 to 0.98 was obtained. The highest accuracy was achieved by Elsharkawy *et al.* [57]. They developed a model named Markov-Gibbs Random Field (MGRF) to detect the severity of infection (low severity or high severity) using 200 chest X-ray images. They achieved accuracy, sensitivity, and specificity of 0.98, 1.00, and 0.97, respectively, by two-fold cross-validation.

An accuracy of 0.72 was reported in two studies by Shan *et al.* [25] and Cai *et al.* [58]. Shan *et al.* [25] used SVM for severity classification (severe or non-severe) based on the quantified radiological features,

	Patient Selection	Index Test	Reference Standard	Flow and Timing	Overall
Shalbfaf 2022	Low	High	High	Unclear	High
Li, Z 2021	High	High	Low	Unclear	High
Ahmad 2022	Low	High	Low	Low	Low
Aljouie 2021	Low	High	Low	Low	Low
Shan 2021	Low	High	Low	Low	Low
Jiao 2021	Low	High	Low	Low	Low
Lassau 2021	Low	High	Low	Low	Low
Li, M 2020	Low	High	Low	Low	Low
Gieraerts 2020	Low	High	Low	Low	Low
Gouda 2020	High	Low	Low	Low	Low
Cai 2020	High	Low	Low	Low	Low
Li, Y 2020	High	Low	Low	Low	Low
Ahmed T 2022	Low	Low	Low	Low	Low
Aslam 2022	Low	Low	Low	Low	Low
Bermejo 2022	Low	Low	Low	Low	Low
Chamberlin 2022	Low	Low	Low	Low	Low
Jordan 2022	Low	Low	Low	Low	Low
Munera 2022	Low	Low	Low	Low	Low
Spagnoli 2022	Low	Low	Low	Low	Low
Bae 2021	Low	Low	Low	Low	Low
Ho 2021	Low	Low	Low	Low	Low
Purkayastha 2020	Low	Low	Low	Low	Low
Balaha 2022	Low	Low	Low	Unclear	Low
Ortiz 2022	Low	Low	Low	Unclear	Low
Shiri 2022	Low	Low	Low	Unclear	Low
Kulkarni 2021	Low	Low	Low	Unclear	Low
Shiri 2021	Low	Low	Low	Unclear	Low
Qiblawey 2021	Low	Low	Low	Unclear	Low
Irmak 2021	Low	Low	Low	Unclear	Low
Kohli 2021	High	Low	Low	High	Unclear
Mushtaq 2020	High	High	Low	Low	Unclear
Abbasi 2022	Low	High	Low	Unclear	Unclear
Aboutalebi 2022	Low	High	Low	Unclear	Unclear
Dinh 2022	Low	High	Low	Unclear	Unclear
Elsharkawy 2021	Low	High	Low	Unclear	Unclear
Ibrahim 2021	High	Low	Low	Unclear	Unclear

Figure 2. Risk of bias of the included studies

including the Percentage Of Consolidation (POC), the Percentage Of Infection (POI) and Mass Of Infection (MOI), which were extracted from 549 CT scans. The best prediction accuracy was 0.73 and 0.72 when using MOI and POI, respectively. Also, they concluded that the quantified radiological features are more informative than the pneumonia severity index (PSI), which is a clinical prediction rule. Cai *et al.* [58] built RF models for severity classification into three groups, moderate, severe, and critical, using 99 CT scans. The defined model I radiomics as moderate vs. (severe + critical) and model II radiomics as severe vs. critical, and checked RF performance in each model. The AUC, accuracy, sensitivity, and specificity in model I were 0.82, 0.75, 0.79, and 0.70, respectively, and in

model II were 0.78, 0.72, 0.79, and 0.66, respectively. Also, they concluded that the hybrid models that combined the radiomics features and clinical data had better performance than those using radiomic features alone.

3.4.2. Mechanical Ventilation or Need for ICU

Among studies that reported AUC for the performance of the AI structures, a range of 0.68 to 0.98 was obtained. The DOR was between 4.8 and 76.6. The best AUC was related to Aslam *et al.* [69], in which CXR images of 1508 patients were analyzed with a combination of DL models and Explainable Artificial Intelligence (EAI). The CXR images were

segmented based on features such as opacity, and patients were classified accordingly. The authors reported an accuracy of 97% for their model.

A range of 0.52 to 0.97 was reported among studies examining model accuracies. The best accuracy was related to the study by Aslam *et al.* [69], and the lowest accuracy was reported by Aljouie *et al.* [68]. They used four classifiers, including linear SVM, RF, Linear Regression (LR), and XGB on 1508 CXR images to classify patients into mechanical ventilation, non-invasive ventilation, and no ventilation groups. They also used some techniques, including Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic (ADASYN), and Random Under-Sampling (RUS), to improve the performance of the models. The best-achieved performance was an accuracy of 0.52 and an AUC of 0.76 for the BRF using X-ray features. Also, the authors reported that combining X-ray features with clinical and laboratory tests showed better performance.

3.4.3. Mortality

Among articles that examined mortality prediction, a range of 0.74 to 0.99 was reported for AUC. The calculated DOR ranged from 2.16 to 22.6. The highest AUC was reported by Aslam *et al.* [69] based on the CXR images of 1513 patients. The study reported an accuracy of 98%, the highest among the included papers. The lowest accuracy and AUC were reported by Aljouie *et al.* [68]. They examined four classifiers on 1513 CXR images for ventilation requirement and mortality. For mortality prediction, XGB + ADASYN had the best performance (AUC of 0.72 and accuracy of 0.71). The accuracy of the included studies for mortality was 0.71 to 0.83.

3.5. Synthesis of Results

Figure 3 shows the accuracy, sensitivity, and specificity of the included studies. The results of meta-analyses are shown in Table 4.

3.5.1. Mortality

Four studies, which consisted of seven individual AI models, were included in the meta-analysis of mortality prediction using AI in COVID-19 patients. The overall sensitivity and specificity of the included

studies were 71% (95% CI: 65%, 77%) and 69% (95% CI: 61%, 76%), respectively (Figure 4A). Moreover, the funnel plot (Figure 5A) was symmetric, and the asymmetry test p-value of 0.19 was derived using Egger's test, a common statistical method for assessing publication bias in meta-analyses. A p-value greater than 0.05 typically indicates no significant evidence of publication bias. In this case, the p-value of 0.19 suggests that there is no evidence of publication bias affecting the results of our analysis. The area under the HSROC curve was 0.76 (95% CI: 0.72–0.80) (Figure 6A), indicating moderately accurate optical diagnostic performance of AI in predicting mortality. The DOR value for this outcome was 6 with a 95% confidence interval of 3–10, suggesting that patients identified as high-risk by the AI models had a significantly higher likelihood of mortality compared to those identified as low-risk.

3.5.2. Severity Assessment

In the meta-analysis of the assessment of severity using AI in COVID-19 patients, nine investigations, including 13 different AI models, were considered. Overall, the included studies demonstrated a sensitivity of 88% (95%CI: 77%, 94%) and a specificity of 89% (95%CI: 82%, 94%) (Figure 4B). In addition, the funnel plot (Figure 5B) was symmetric, and the asymmetry test p-value of 0.07 suggested that publication bias was not present. The area under the HSROC curve was 0.95 (95% CI: 0.92–0.96) (Figure 6B), showing highly accurate optical diagnostic performance of AI in severity assessment. The DOR value was 59 (95% CI: 18–197), demonstrating the models' effectiveness in distinguishing between different levels of severity.

3.5.3. Need for Ventilation

Four studies comprising six AI models were included in the meta-analysis of AI for predicting ventilation requirements in COVID-19 patients. Overall, the studies that were considered showed a pooled sensitivity of 67% (95% CI: 61%, 73%) and a pooled specificity of 89% (95% CI: 75%, 95%) (Figure 4C). In addition, there was no evidence of publication bias as shown by the symmetric funnel plot (Figure 5C) and a p-value of 0.92 for the asymmetry test. The area under the HSROC curve was 0.77 (95% CI: 0.73–0.80) (Figure 6C), demonstrating

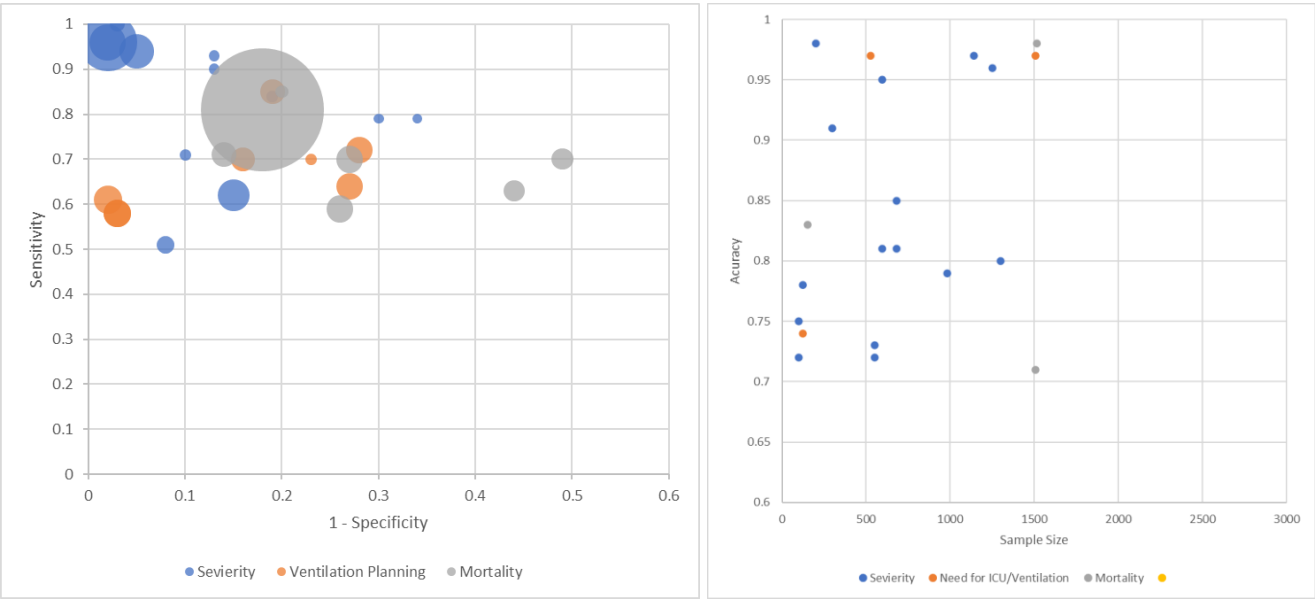


Figure 3. Distribution of the specificity, sensitivity, and accuracy of the included studies categorized based on the study aim

Table 4. Summary of the Meta-analysis Statistics

Parameter	Mortality	Severity Assessment	Need for Ventilation
No. Studies	4	9	4
No. Models	7	13	6
Pooled Sensitivity	71%; 95% CI [65%, 77%]	88%; 95% CI [77%, 94%]	67%; 95% CI [61%, 73%]
Pooled Specificity	69%; 95% CI [61%, 76%]	89%; 95% CI [82%, 94%]	89%; 95% CI [75%, 95%]
Positive Likelihood Ratio	2.3; 95% CI [1.7, 3.1]	8.2; 95% CI [4.6, 14.5]	5.9; 95% CI [2.7, 13.1]
Negative Likelihood Ratio	0.41; 95% CI [0.31, 0.55]	0.14; 95% CI [0.07, 0.28]	0.37; 95% CI [0.32, 0.43]
Diagnostic Odds Ratio (DOR)	6; 95% CI [3, 10]	59; 95% CI [18, 197]	16; 95% CI [7, 36]
Area under the HSROC curve	0.76; 95% CI [0.72, 0.80]	0.95; 95% CI [0.92, 0.96]	0.77; 95% CI [0.73, 0.80]
Heterogeneity (Chi-square)	94.059	102.071	365.338
df	2.00	2.00	2.00
p-value	0.000	0.000	0.000
Inconsistency (I-square) – I ²	98%; 95% CI [97%- 99%]	98%; 95% CI [97%- 99%]	99%; 95% CI [99%- 100%]
Proportion of heterogeneity likely due to threshold effect	0.13	0.51	0.49
Deek's Funnel Plot asymmetry test p-value	0.19	0.07	0.92

that AI's optical diagnostic performance in predicting the requirement for ventilation was reasonably accurate. The measured DOR was 16 (95% CI: 7-36), which indicates that patients identified as requiring mechanical ventilation by the AI models are 16 times more likely to actually need it compared to those not classified as such.

4. Discussion

Machine learning and deep learning methods facilitate the extraction and identification of body tissue characteristics from images and thus speed up patient triage and allow timely treatment plans for patients. Therefore, in the current study, we reviewed the studies that analyzed the performance of AI

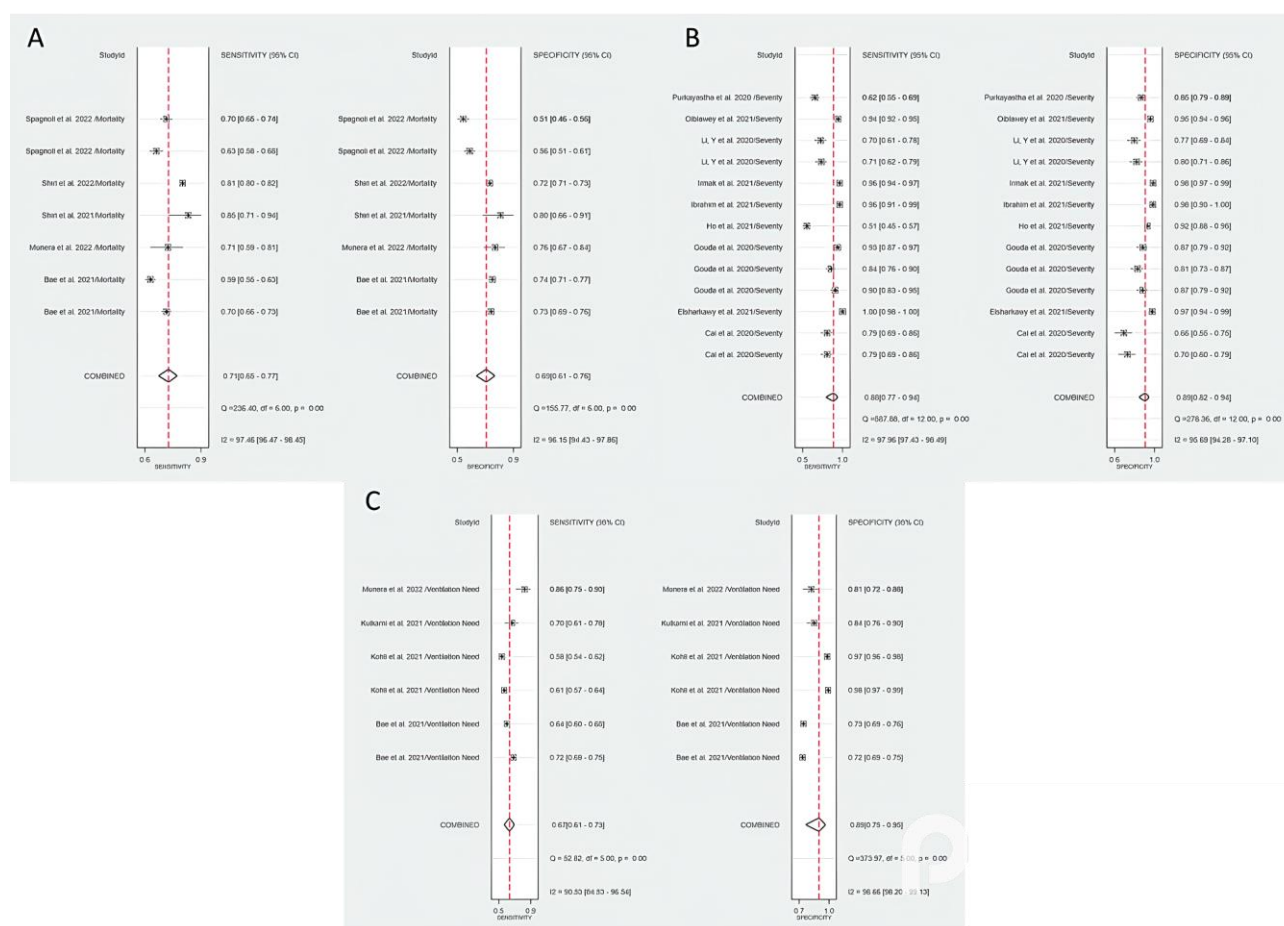


Figure 4. Forest plot of sensitivity and specificity of AI in predicting mortality (A), severity assessment (B) and predicting the need for ventilation (C)

models for predicting disease severity, ventilation requirement, need for ICU, and mortality using standard of care CT or CXR images.

CXR and CT imaging modalities were used in the included studies. Chest radiography is a quick and easy test and is usually requested due to low cost and fast data acquisition compared to CT [75]. However, it was reported that CXR has restrictions for the accurate detection of COVID-19 infection compared to CT. On the other hand, CT images are better options for disease severity analysis and patient monitoring, and they have shown higher sensitivity compared to CXR [43]. Another possible source of bias from CXR is that AI methods may evaluate images taken from different views, leading to an inaccurate outcome [19]. For instance, instead of the posteroanterior view, in severe cases an anteroposterior projection is used. Having mentioned the above points, based on our results, machine learning models that were applied to CXR showed accuracy of 95-98% had comparable

accuracy with CT (i.e., 72-97 % based on 10 studies) to evaluate the severity of the disease. However, there was no study aimed at comparing the results of these two data acquisition methods.

The biggest limitation behind using CT and CXR images for diagnosis and evaluating the prognosis of the disease is the lack of COVID-related experience among radiologists concerning the COVID-19 infection pathways and spread. Besides, there is always the possibility of error when human vision is used to analyze the images. In the early stage of the pandemic, the progression patterns of the disease were not completely recognized and showed different behaviors in each region. Besides, considering the variations in the health and triage systems in different regions, data regarding the virus behavior in one region cannot be generalized to all countries. Initially, due to a decrease in errors, it was recommended to design scoring systems to evaluate images objectively. This has resulted in more accurate decision-making

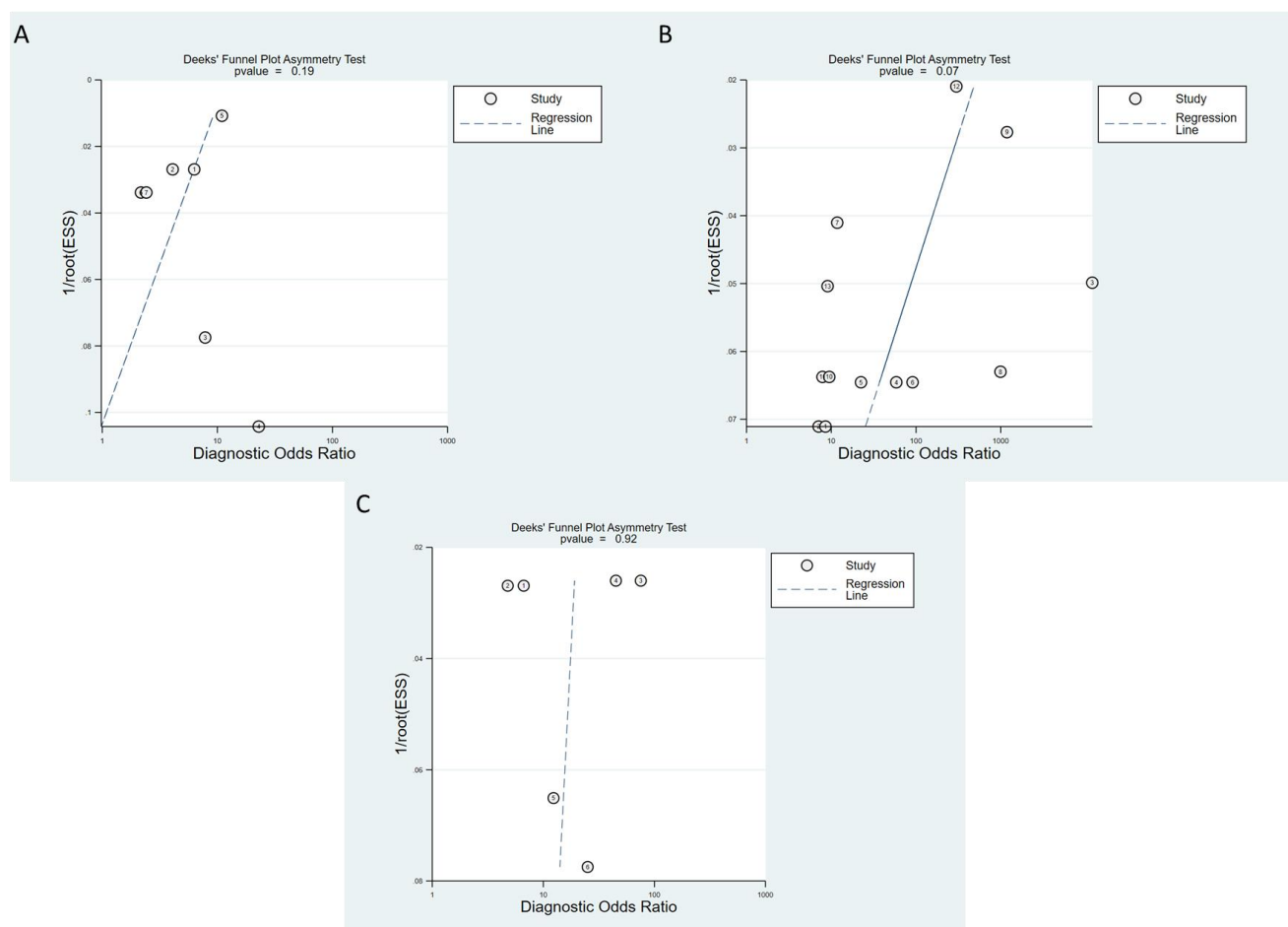


Figure 5. Deeks' funnel plot to evaluate publication bias of studies in predicting mortality (A), severity assessment (B), and predicting the need for ventilation (C). The vertical axis displays the inverse of the square root of the effective sample size ($1/\text{root}(\text{ESS})$). The horizontal axis displays the diagnostic odds ratio (DOR). All p-values indicated a symmetrical funnel plot

and increased efficacy. However, manual segment scoring is still time-consuming and may not be optimal for daily clinical practice. Thus, AI-based methods have the potential to decrease workload and improve patient safety [76].

In the included studies, the severity of the COVID-19 infection was assessed using different approaches. One of the most common methods was whole lungs/lesions segmentation and evaluation based on the extent of the affected tissue. Most included studies used UNet models to segment the lungs and lesion areas. They similarly obtained a dice similarity coefficient (DSC) of about 0.98 for lesion segmentation, with a range of 0.77-0.99. Li, Z *et al.* [52] used a Feature Pyramid Network (FPN) to achieve the best DSC. They reported that although FPN did not improve the results compared with the UNet model in lung segmentation, it showed better results in lesion segmentation. The lowest DSC was reported by Cai *et al.* [58] using the UNet model.

Compared to the study performed only lung assessment [59], all studies that performed both lung and lesion segmentation had higher accuracy in severity assessment, except one [58].

Studies used different categorizing methods to classify the severity among the patients [77]. Having more classes will increase the precision of the patient categorization and will improve the treatment response [44]. However, this can complicate the data processing, which can decrease the model's performance. Most of the studies that analyzed the severity level categorized patients into two groups, and the best performance belonged to one of these models.

It must be noted that factors such as age, sex, and body mass can impact the response of the human body to the infection. Indeed, patients with different abovementioned features may have different prognoses even with the same initial infection stage.

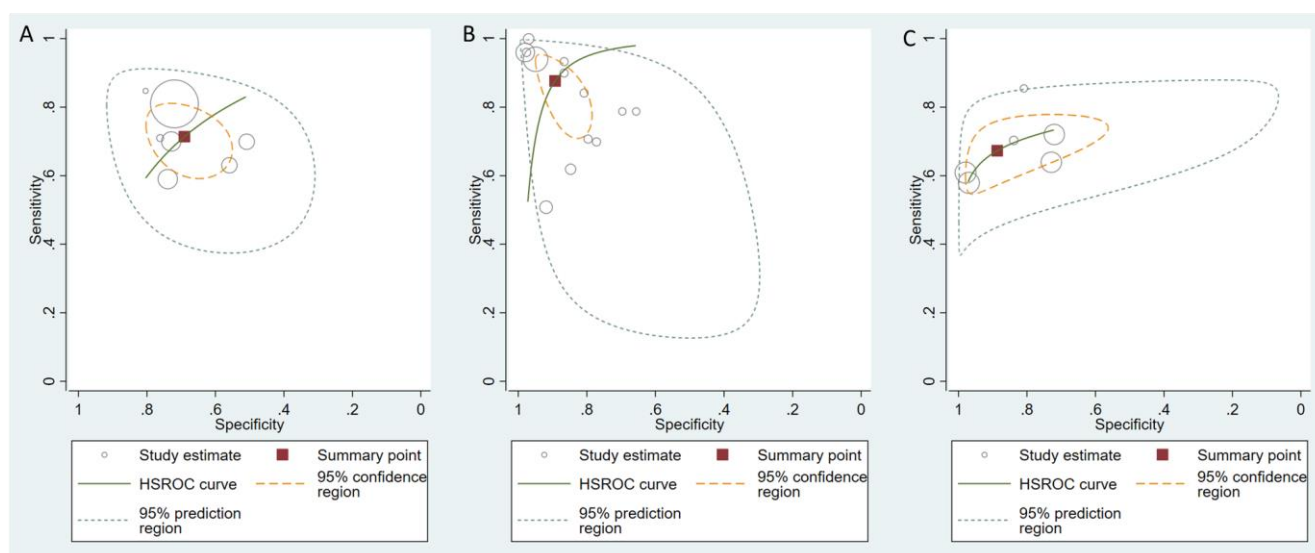


Figure 6. Hierarchical summary receiver-operating characteristic (HSROC) curve for the diagnostic performance of AI in predicting mortality (A), severity assessment (B), and predicting the need for ventilation (C). The size of the gray circles indicates the number of samples in the individual studies. The summary sensitivity and specificity are shown with a dark red square, and the 95% confidence region is plotted in short lines

Thus, neglecting these variables in evaluating the function of the AI models in some of the included studies can be considered a significant drawback. This issue is even more important in studies that evaluate more advanced outcomes, such as the need for ventilation or intensive care and mortality [45].

The AI model learn from the available training data. Thus, non-curated data could include some inaccuracies; hence lowers the performance of the AI model. Therefore, a more reliable outcome can be expected in the case of analyzing models in which the ground truth is accurate and reliable, since the machine is trained based on the imported ground truth. The gold standard may be less accurate in the case of analyzing severity with AI models since it depends on the practitioner's assessment, which may differ from site to site and expert to expert. On the other hand, as mortality and the need for ventilation are variables that have a binary condition (i.e., will happen or not), the gold standard of the models developed for these two variables can be considered ground truth, which is a critical advantage for this model.

In clinical routine practice, AI methods can accelerate the triage, aid decision makers in stressful situations, and enable practitioners to help people in a broader area. It is recommended that, based on their scoring system, the necessity of ventilation, intensive care, and the possibility of mortality in each of the mentioned situations can be discussed with the patient,

which can significantly help them in decision making [45]. Besides, this scoring system can estimate the length of stay and the duration of high-level care.

4.1. Limitation

Public databases of CT and CXR images of patients with COVID-19 provide a valuable source for AI research. Although these studies have been performed at different institutions across the globe, almost all AI systems are not open and are unavailable to the research community. Besides, to evaluate the generalizability of the models, Individual Participant Data (IDP) from different regions can be used, which can significantly increase the applicability and robustness of the models in daily routine care [78, 79]. Accordingly, the World Health Organization has designed a platform for sharing anonymized COVID-19 clinical data [3].

In case of combining public data sets to train or test the model, it has to be considered that most of them have no restrictions on the imported data, hence the possibility of using duplicated images or even those that are not correctly diagnosed with COVID-19. Besides, since not all of the images are in the DICOM format, a decreased image quality can be expected [19]. This can cause a serious problem for machine learning models since the amount of decrease is not the same among the images. Neglecting the demographics of patients and adding pediatric images

in public data sets has increased the bias of using them in analyses [19].

Moreover, the included studies did not provide complete data concerning the function of their models, such as sensitivity, specificity, and accuracy. This prevented other researchers from reproducing and comparing the achieved performance. Image modalities used in different studies had different features and setups. Despite the extensive efforts in developing ML models using different feature extraction, selection, and classification algorithms and DL models using different architectures and topologies, the comparison of their performances and applicability is, at least, challenging at this stage.

Imaging scans were acquired at various institutions using different scanners and data acquisition and image reconstruction protocols. Accordingly, the obtained images should be pre-processed to ensure consistency of the input [80]. Imaging systems and scanning protocols for acquiring images use different acquisition parameters, and so are CT image reconstruction methods. These factors can significantly impact the robustness and reliability of AI applications and lead to misdiagnosis.

Included studies lack an independent external validation. Thus, although the majority of the included studies were at low risk of bias, it should be noted that we cannot recommend any model to be used in daily practice, specifically considering that recent publications about COVID-19 prediction models are entering the literature quickly.

Furthermore, several studies reviewed here did not mention the imaging study duration, despite being an essential factor in determining the prognosis, and the timely determination of the prognosis leads to appropriate treatment. Although most images are acquired during admission, it has not been evaluated whether the models will have the same predictive values about the need for intensive care or mortality if images are taken at other time points.

To use prediction models for decision-making, the studies need to assess the performance of a diagnostic tool to specify the target population, enabling users to know which category of patients can be evaluated using a given model [3]. However, this data was not comprehensively provided in the included studies, which made users doubt whether to use the model for

their intended population. Considering the variabilities in the target population can justify the discrepancies in the results reported by different studies, the difference in the relative frequency between the population necessitates some alterations in the prediction model in other settings [3].

With regards to the choice of predictors in the prediction model, it is recommended to consider the expert opinion and published literature rather than choosing only the data-driven ones. For prediction models, the following variables are recommended: age, sex, C-reactive protein, lactic dehydrogenase, lymphocyte count, CT-scoring, albumin (or albumin/globin), direct bilirubin, and red blood cell distribution width.

Despite all recent progress, the proposed methods commonly face challenges for implementation in routine practice due to the following reasons: (1) the bias due to small datasets; (2) the variations observed in large internationally sourced datasets; (3) the poor integration of multistream data, particularly imaging data; (4) the difficulty of the task of prognosis; and (5) the necessity for clinicians and data analysts to work together to ensure the developed algorithms are clinically relevant and applicable into routine clinical care. Overall, there is a significant need for the creation of trustworthy ecosystems towards routine deployment of AI techniques [81].

5. Conclusion

Machine learning and deep learning models can help clinicians predict the severity of disease, ventilation requirement or need for ICU, and mortality, and to subsequently manage COVID-19 patients more effectively. Based on evidence from included studies, the models using imaging data extracted from CT or CXR reported adequate levels of performance. However, the proposed methods commonly face challenges for deployment in routine practice due to issues concerning data curation, harmonization of imaging protocols, reproducibility, external validation, explainability, robustness and applicability, and overall lack of following best practices in AI development and validation. Furthermore, it is essential to provide statistical analysis of model performances, including sensitivity,

specificity and accuracy, enabling researchers to compare models more objectively.

References

- 1- H. Swapnarekha, H. S. Behera, J. Nayak, and B. Naik, "Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review." (in eng), *Chaos Solitons Fractals*, Vol. 138p. 109947, Sep (2020).
- 2- Hiba Abdelsadig Mohammed, Shahd Abubaker Elamin, Alla El-Awaisi, and Maguy Saffouh El Hajj, "Use of the job demands-resource model to understand community pharmacists' burnout during the COVID-19 pandemic." *Research in Social and Administrative Pharmacy*, (2022).
- 3- L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal." (in eng), *Bmj*, Vol. 369p. m1328, Apr 7 (2020).
- 4- F. Abdulla, Z. Nain, M. Karimuzzaman, M. M. Hossain, and A. Rahman, "A Non-Linear Biostatistical Graphical Modeling of Preventive Actions and Healthcare Factors in Controlling COVID-19 Pandemic." (in eng), *Int J Environ Res Public Health*, Vol. 18 (No. 9), Apr 23 (2021).
- 5- Y. Bouchareb *et al.*, "Artificial intelligence-driven assessment of radiological images for COVID-19." (in eng), *Comput Biol Med*, Vol. 136p. 104665, Sep (2021).
- 6- Buddhisha Udugama *et al.*, "Diagnosing COVID-19: the disease and tools for detection." *ACS nano*, Vol. 14 (No. 4), pp. 3822-35, (2020).
- 7- Wei-jie Guan *et al.*, "Clinical characteristics of coronavirus disease 2019 in China." *New England journal of medicine*, Vol. 382 (No. 18), pp. 1708-20, (2020).
- 8- A. Kohli, T. Jha, and A. B. Pazhayattil, "The value of AI based CT severity scoring system in triage of patients with Covid-19 pneumonia as regards oxygen requirement and place of admission." (in English), *Indian Journal of Radiology and Imaging*, Article Vol. 31 (No. 5), pp. S61-S69, (2021).
- 9- Qin Sun, Haibo Qiu, Mao Huang, and Yi Yang, "Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province." *Annals of intensive care*, Vol. 10 (No. 1), pp. 1-4, (2020).
- 10- Pavel Hamet and Johanne Tremblay, "Artificial intelligence in medicine." *Metabolism*, Vol. 69pp. S36-S40, (2017).
- 11- Hossein Mohammad-Rahimi, Mohadeseh Nadimi, Azadeh Ghalyanchi-Langeroudi, Mohammad Taheri, and Soudeh Ghafouri-Fard, "Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review." *Frontiers in cardiovascular medicine*, Vol. 8p. 638011, (2021).
- 12- T. J. Bradshaw *et al.*, "Nuclear Medicine and Artificial Intelligence: Best Practices for Algorithm Development." (in eng), *J Nucl Med*, Vol. 63 (No. 4), pp. 500-10, Apr (2022).
- 13- Chen Ma, Zhihao Yao, Qinran Zhang, and Xiufen Zou, "Quantitative integration of radiomic and genomic data improves survival prediction of low-grade glioma patients." *Mathematical Biosciences and Engineering*, Vol. 18 (No. 1), pp. 727-44, (2021).
- 14- Hossein Mohammad-Rahimi *et al.*, "Deep Learning for Caries Detection: A Systematic Review: DL for Caries Detection." *Journal of Dentistry*, p. 104115, (2022).
- 15- Hossein Mohammad-Rahimi *et al.*, "Deep learning in periodontology and oral implantology: A scoping review." *Journal of Periodontal Research*, (2022).
- 16- Hossein Mohammad-Rahimi, Mohadeseh Nadimi, Mohammad Hossein Rohban, Erfan Shamsoddin, Victor Y Lee, and Saeed Reza Motamedian, "Machine learning and orthodontics, current trends and the future opportunities: a scoping review." *American Journal of Orthodontics and Dentofacial Orthopedics*, Vol. 160 (No. 2), pp. 170-92. e4, (2021).
- 17- Ş Busnatu *et al.*, "Clinical Applications of Artificial Intelligence-An Updated Overview." (in eng), *J Clin Med*, Vol. 11 (No. 8), Apr 18 (2022).
- 18- I. Buvat and F. Orlhac, "The T.R.U.E. Checklist for Identifying Impactful Artificial Intelligence-Based Findings in Nuclear Medicine: Is It True? Is It Reproducible? Is It Useful? Is It Explainable?" (in eng), *J Nucl Med*, Vol. 62 (No. 6), pp. 752-54, Jun 1 (2021).
- 19- Michael Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans." *Nature Machine Intelligence*, Vol. 3 (No. 3), pp. 199-217, 2021/03/01 (2021).
- 20- Mostafa Nazari, Isaac Shiri, and Habib Zaidi, "Radiomics-based machine learning model to predict risk of death within 5-years in clear cell renal cell carcinoma patients." *Computers in Biology and Medicine*, Vol. 129p. 104135, 2021/02/01/ (2021).
- 21- I. Shiri *et al.*, "Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: Test-retest and image registration analyses." (in eng), *Med Phys*, Vol. 47 (No. 9), pp. 4265-80, Sep (2020).
- 22- B. Koçak, EŞ Durmaz, E. Ateş, and Ö Kılıçkesmez, "Radiomics with artificial intelligence: a practical guide for beginners." (in eng), *Diagn Interv Radiol*, Vol. 25 (No. 6), pp. 485-95, Nov (2019).
- 23- M. R. Tomaszewski and R. J. Gillies, "The Biological Meaning of Radiomic Features." (in eng), *Radiology*, Vol. 298 (No. 3), pp. 505-16, Mar (2021).
- 24- W. Gouda and R. Yasin, "COVID-19 disease: CT Pneumonia Analysis prototype by using artificial intelligence, predicting the disease severity." (in English), *Egyptian Journal of Radiology and Nuclear Medicine*, Article Vol. 51 (No. 1), (2020).

- 25- F. Shan *et al.*, "Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction." (in eng), *Med Phys*, Vol. 48 (No. 4), pp. 1633-45, Apr (2021).
- 26- J. Mushtaq *et al.*, "Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients." (in English), *European Radiology*, Article Vol. 31 (No. 3), pp. 1770-79, (2021).
- 27- J. Bae *et al.*, "Predicting mechanical ventilation and mortality in covid-19 using radiomics and deep learning on chest radiographs: A multi-institutional study." (in English), *Diagnostics*, Article Vol. 11 (No. 10), (2021).
- 28- M. R. H. Mondal, S. Bharati, and P. Podder, "Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review." (in eng), *Curr Med Imaging*, Vol. 17 (No. 12), pp. 1403-18, (2021).
- 29- H. Wang, S. Jia, Z. Li, Y. Duan, G. Tao, and Z. Zhao, "A Comprehensive Review of Artificial Intelligence in Prevention and Treatment of COVID-19 Pandemic." (in eng), *Front Genet*, Vol. 13p. 845305, (2022).
- 30- M. Khan *et al.*, "Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review." (in eng), *Expert Syst Appl*, Vol. 185p. 115695, Dec 15 (2021).
- 31- L. Wang *et al.*, "Artificial Intelligence for COVID-19: A Systematic Review." (in eng), *Front Med (Lausanne)*, Vol. 8p. 704256, (2021).
- 32- Jawad Rasheed *et al.*, "A survey on artificial intelligence approaches in supporting frontline workers and decision makers for the COVID-19 pandemic." *Chaos, Solitons & Fractals*, Vol. 141p. 110337, (2020).
- 33- Matthew DF McInnes *et al.*, "Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement." *Jama*, Vol. 319 (No. 4), pp. 388-96, (2018).
- 34- Simar Singh Bajaj, Alister Francois Martin, and Fatima Cody Stanford, "Health-based civic engagement is a professional responsibility." *Nature Medicine*, Vol. 27 (No. 10), pp. 1661-63, (2021).
- 35- Soroush Sadr *et al.*, "Deep Learning for Detection of Periapical Radiolucent Lesions: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy." *Journal of Endodontics*, 2022/12/21/ (2022).
- 36- Hossein Mohammad-Rahimi *et al.*, "Deep learning in periodontology and oral implantology: A scoping review." *Journal of Periodontal Research*, Vol. 57 (No. 5), pp. 942-51, (2022).
- 37- Hossein Mohammad-Rahimi *et al.*, "Deep learning for caries detection: A systematic review." *Journal of Dentistry*, Vol. 122p. 104115, 2022/07/01/ (2022).
- 38- Hossein Mohammad-Rahimi, Mohadeseh Nadimi, Mohammad Hossein Rohban, Erfan Shamsoddin, Victor Y. Lee, and Saeed Reza Motamedian, "Machine learning and orthodontics, current trends and the future opportunities: A scoping review." *American Journal of Orthodontics and Dentofacial Orthopedics*, Vol. 160 (No. 2), pp. 170-92.e4, 2021/08/01/ (2021).
- 39- Jane V Carter, Jianmin Pan, Shesh N Rai, and Susan %J Surgery Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves." Vol. 159 (No. 6), pp. 1638-45, (2016).
- 40- SS Mahid, CA Hornung, KS Minor, M Turina, and S %J Journal of British Surgery Galandiuk, "Systematic reviews and meta-analysis for the surgeon scientist." Vol. 93 (No. 11), pp. 1315-24, (2006).
- 41- Jonathan J Deeks, Petra Macaskill, and Les Irwig, "The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed." *Journal of clinical epidemiology*, Vol. 58 (No. 9), pp. 882-93, (2005).
- 42- MMG Leeflang, "Systematic reviews and meta-analyses of diagnostic test accuracy." *Clinical Microbiology and Infection*, Vol. 20 (No. 2), pp. 105-13, (2014).
- 43- S. Ortiz, F. Rojas, O. Valenzuela, L. J. Herrera, and I. Rojas, "Determination of the Severity and Percentage of COVID-19 Infection through a Hierarchical Deep Learning System." (in eng), *J Pers Med*, Vol. 12 (No. 4), Mar 28 (2022).
- 44- Tuan Le Dinh, Suk-Hwan Lee, Seong-Geun Kwon, and Ki-Ryong Kwon, "COVID-19 Chest X-ray Classification and Severity Assessment Using Convolutional and Transformer Neural Networks." *Applied Sciences*, Vol. 12 (No. 10), p. 4861, (2022).
- 45- J. H. Chamberlin *et al.*, "An Interpretable Chest CT Deep Learning Algorithm for Quantification of COVID-19 Lung Disease and Prediction of Inpatient Morbidity and Mortality." (in eng), *Acad Radiol*, Vol. 29 (No. 8), pp. 1178-88, Aug (2022).
- 46- H. M. Balaha, E. M. El-Gendy, and M. M. Saafan, "A complete framework for accurate recognition and prognosis of COVID-19 patients based on deep transfer learning and feature classification approach." (in eng), *Artif Intell Rev*, Vol. 55 (No. 6), pp. 5063-108, (2022).
- 47- J. Ahmad *et al.*, "Disease Progression Detection via Deep Sequence Learning of Successive Radiographic Scans." (in eng), *Int J Environ Res Public Health*, Vol. 19 (No. 1), Jan 2 (2022).
- 48- P. Gifani, A. Shalbaf, and M. Vafaezadeh, "Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans." (in eng), *Int J Comput Assist Radiol Surg*, Vol. 16 (No. 1), pp. 115-23, Jan (2021).
- 49- Wajid Arshad Abbasi, Syed Abbas, and Dr Saiqa Andleeb, COVIDX: Computer-aided diagnosis of Covid-19 and its severity prediction with raw digital chest X-ray images. (2020).

- 50- Hossein Aboutaleb, Maya Pavlova, Mohammad Javad Shafiee, Ali Sabri, Amer Alaref, and Alexander Wong, "COVID-Net CXR-S: Deep Convolutional Neural Network for Severity Assessment of COVID-19 Cases from Chest X-ray Images." *Diagnostics*, Vol. 12 (No. 1), p. 25, (2022).
- 51- M. D. Li *et al.*, "Automated Assessment and Tracking of COVID-19 Pulmonary Disease Severity on Chest Radiographs using Convolutional Siamese Neural Networks." (in eng), *Radiol Artif Intell*, Vol. 2 (No. 4), p. e200079, Jul (2020).
- 52- Z. Li *et al.*, "A deep-learning-based framework for severity assessment of COVID-19 with CT images." (in English), *Expert Systems with Applications*, Article Vol. 185(2021), Art no. 115616.
- 53- M. R. Ibrahim, S. M. Youssef, and K. M. Fathalla, "Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment." (in English), *Journal of Ambient Intelligence and Humanized Computing*, Article (2021).
- 54- Y. Qiblawey *et al.*, "Detection and severity classification of COVID-19 in CT images using deep learning." (in English), *Diagnostics*, Article Vol. 11 (No. 5), (2021).
- 55- Z. Jiao *et al.*, "Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study." (in eng), *Lancet Digit Health*, Vol. 3 (No. 5), pp. e286-e94, May (2021).
- 56- N. Lassau *et al.*, "Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients." (in English), *Nature Communications*, Article Vol. 12 (No. 1), (2021).
- 57- M. Elsharkawy *et al.*, "Early assessment of lung function in coronavirus patients using invariant markers from chest X-rays images." (in eng), *Sci Rep*, Vol. 11 (No. 1), p. 12095, Jun 8 (2021).
- 58- W. Cai *et al.*, "CT Quantification and Machine-learning Models for Assessment of Disease Severity and Prognosis of COVID-19 Patients." (in English), *Academic Radiology*, Article Vol. 27 (No. 12), pp. 1665-78, (2020).
- 59- S. Purkayastha *et al.*, "Machine Learning-Based Prediction of COVID-19 Severity and Progression to Critical Illness Using CT Imaging and Clinical Data." (in English), *Korean journal of radiology*, Article Vol. 22 (No. 7), pp. 1213-24, (2021).
- 60- E. Irmak, "COVID-19 disease severity assessment using CNN model." (in eng), *IET Image Process*, Vol. 15 (No. 8), pp. 1814-24, Jun (2021).
- 61- T. T. Ho *et al.*, "Deep Learning Models for Predicting Severe Progression in COVID-19-Infected Patients: Retrospective Study." (in eng), *JMIR Med Inform*, Vol. 9 (No. 1), p. e24973, Jan 28 (2021).
- 62- Christopher Gieraerts *et al.*, "Prognostic Value and Reproducibility of AI-assisted Analysis of Lung Involvement in COVID-19 at Low-Dose Submillisievert Chest CT: Sample Size Implications for Clinical Trials." *Radiology: Cardiothoracic Imaging*, Vol. 2 (No. 5), p. e200441, (2020).
- 63- Y. Li *et al.*, "Prediction of disease progression in patients with COVID-19 by artificial intelligence assisted lesion quantification." (in English), *Scientific Reports*, Article Vol. 10 (No. 1), (2020), Art no. 22083.
- 64- D. Bermejo-Peláez *et al.*, "Deep learning-based lesion subtyping and prediction of clinical outcomes in COVID-19 pneumonia using chest CT." (in eng), *Sci Rep*, Vol. 12 (No. 1), p. 9387, Jun 7 (2022).
- 65- Isaac Shiri *et al.*, "Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients." *Computers in biology and medicine*, Vol. 132p. 104304, (2021).
- 66- I. Shiri *et al.*, "COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: Analysis of a multi-institutional dataset of 14,339 patients." (in eng), *Comput Biol Med*, Vol. 145p. 105467, Jun (2022).
- 67- Lorenzo Spagnoli *et al.*, "Outcome Prediction for SARS-CoV-2 Patients Using Machine Learning Modeling of Clinical, Radiological, and Radiomic Features Derived from Chest CT Images." *Applied Sciences*, Vol. 12 (No. 9), p. 4493, (2022).
- 68- A. F. Aljouie *et al.*, "Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory, and clinical data with machine learning." (in English), *Journal of Multidisciplinary Healthcare*, Article Vol. 14pp. 2017-33, (2021).
- 69- Nida Aslam, "Explainable Artificial Intelligence Approach for the Early Prediction of Ventilator Support and Mortality in COVID-19 Patients." *Computation*, Vol. 10 (No. 3), p. 36, (2022).
- 70- T. U. Ahmed, M. N. Jamil, M. S. Hossain, R. U. Islam, and K. Andersson, "An Integrated Deep Learning and Belief Rule Base Intelligent System to Predict Survival of COVID-19 Patient under Uncertainty." (in eng), *Cognit Comput*, Vol. 14 (No. 2), pp. 660-76, (2022).
- 71- A. R. Kulkarni *et al.*, "Deep learning model to predict the need for mechanical ventilation using chest X-ray images in hospitalised patients with COVID-19." (in English), *BMJ Innovations*, Article Vol. 7 (No. 2), pp. 261-70, (2021).
- 72- N. Munera *et al.*, "A novel model to predict severe COVID-19 and mortality using an artificial intelligence algorithm to interpret chest radiographs and clinical variables." (in eng), *ERJ Open Res*, Vol. 8 (No. 2), Apr (2022).

- 73- J. H. Chamberlin *et al.*, "Automated diagnosis and prognosis of COVID-19 pneumonia from initial ER chest X-rays using deep learning." (in eng), *BMC Infect Dis*, Vol. 22 (No. 1), p. 637, Jul 21 (2022).
- 74- F. Shan *et al.*, "Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction." *Medical Physics*, Vol. 48 (No. 4), pp. 1633-45, Apr (2021).
- 75- Lucas de Pádua Gomes de Farias *et al.*, "Imaging findings in COVID-19 pneumonia." *Clinics*, Vol. 75(2020).
- 76- Prabira Kumar Sethy, Santi Kumari Behera, Komma Anitha, Chanki Pandey, and MR Khan, "Computer aid screening of COVID-19 using X-ray and CT scan images: An inner comparison." *Journal of X-ray Science and Technology*, Vol. 29 (No. 2), pp. 197-210, (2021).
- 77- Pegah Moradi Khaniabadi *et al.*, "Two-step machine learning to diagnose and predict involvement of lungs in COVID-19 and pneumonia using CT radiomics." *Computers in biology and medicine*, Vol. 150p. 106165, (2022).
- 78- Yan Li, Zhenlu Yang, Tao Ai, Shandong Wu, and Liming Xia, "Association of "initial CT" findings with mortality in older patients with coronavirus disease 2019 (COVID-19)." *European Radiology*, Vol. 30 (No. 11), pp. 6186-93, (2020).
- 79- Qiang Li *et al.*, "A simple algorithm helps early identification of SARS-CoV-2 infection patients with severe progression tendency." *Infection*, Vol. 48 (No. 4), pp. 577-84, (2020).
- 80- Ilker Ozsahin, Boran Sekeroglu, Musa Sani Musa, Mubarak Taiwo Mustapha, and Dilber Uzun Ozsahin, "Review on diagnosis of COVID-19 from chest CT images using artificial intelligence." *Computational and Mathematical Methods in Medicine*, Vol. 2020(2020).
- 81- N. Hasani *et al.*, "Trustworthy Artificial Intelligence in Medical Imaging." (in eng), *PET Clin*, Vol. 17 (No. 1), pp. 1-12, Jan (2022).