Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Virtual Imaging Trials Improved the Transparency and Reliability of AI Systems in COVID-19 Imaging

Fakrul Islam Tushar[1,2], Lavsen Dahal[1,2], Saman Sotoudeh-Paima[1,2], Ehsan Abadi[1,2], William P. Segars[1,2], Ehsan Samei[1,2], Joseph Y. Lo[1,2]

[1]Center for Virtual Imaging Trials, Carl E. Ravin Advanced Imaging Laboratories, Dept. of Radiology Duke University, Durham, 27705, USA
[2]Department of Electrical & Computer Engineering, Pratt School of Engineering, Duke University, Durham, 27705, USA

Corresponding author: Fakrul Islam Tushar (e-mail: tushar.ece@ duke.edu).

This work was funded in part by the Center for Virtual Imaging Trials, NIH/NIBIB P41-EB028744.

ABSTRACT The credibility of Artificial Intelligence (AI) models in medical imaging, particularly during the COVID-19 pandemic, has been challenged by reproducibility issues and obscured clinical insights. To address these concerns, we propose a Virtual Imaging Trials (VIT) framework, utilizing both clinical and simulated datasets to evaluate AI systems. This study focuses on using convolutional neural networks (CNNs) for COVID-19 diagnosis using computed tomography (CT) and chest radiography (CXR). We developed and tested multiple AI models, 3D ResNet-like and 2D EfficientNetv2 architectures, across diverse datasets. Our evaluation metrics included the area under the curve (AUC). Statistical analyses, such as the DeLong method for AUC confidence intervals, were employed to assess performance differences. Our findings demonstrate that VIT provides a robust platform for objective assessment, revealing significant influences of dataset characteristics, patient factors, and imaging physics on AI efficacy. Notably, models trained on the most diverse datasets showed the highest external testing performance, with AUC values ranging from 0.73 to 0.76 for CT and 0.70 to 0.73 for CXR. Internal testing yielded higher AUC values (0.77 to 0.85 for CT and 0.77 to 1.0 for CXR), highlighting a substantial drop in performance during external validation, which underscores the importance of diverse and comprehensive training and testing data. This approach enhances model transparency and reliability, offering nuanced insights into the factors driving AI performance and bridging the gap between experimental and clinical settings. The study underscores the potential of VIT to improve the reproducibility and clinical relevance of AI systems in medical imaging.

INDEX TERMS Virtual Imaging trials, COVID-19, Computed tomography.

I. INTRODUCTION

For effective development and optimization, artificial intelligence (AI) models typically require massive amounts of data. Even when large datasets are available for training, AI models often struggle to generalize, resulting in limited clinical applicability. This crisis of reproducibility was starkly evident during the COVID-19 pandemic when chest radiography (CXR) and computed tomography (CT) were initially employed for detecting and managing lung infections [1, 2]. In the rush to develop AI aides for radiologists, however, many studies reported unrealistic, near-perfect performances that dropped almost to chance upon external testing [3-9]. Failure of medical imaging AI models to generalize is a pervasive problem. There is a pressing need for an evaluation framework for medical imaging AI models that can assess the true performance. The 4th COV19D Competition, as part of the CVPR 2024 Workshop, highlights the ongoing efforts by the machine learning community to address challenges in COVID-19 diagnosis through medical imaging [10]. When these AI models fail to generalize, we need to understand what out-of-distribution factors (e.g., patient normal anatomy and disease, or imaging physics conditions) are driving the model performance.

Although imaging is no longer used for the primary diagnosis of COVID-19, this disease remains a relevant and valuable case study for several reasons. In an unprecedented effort, many large public datasets of medical images were released [11-18]. and an ongoing coordination effort continues to be led by the Medical Imaging and Data

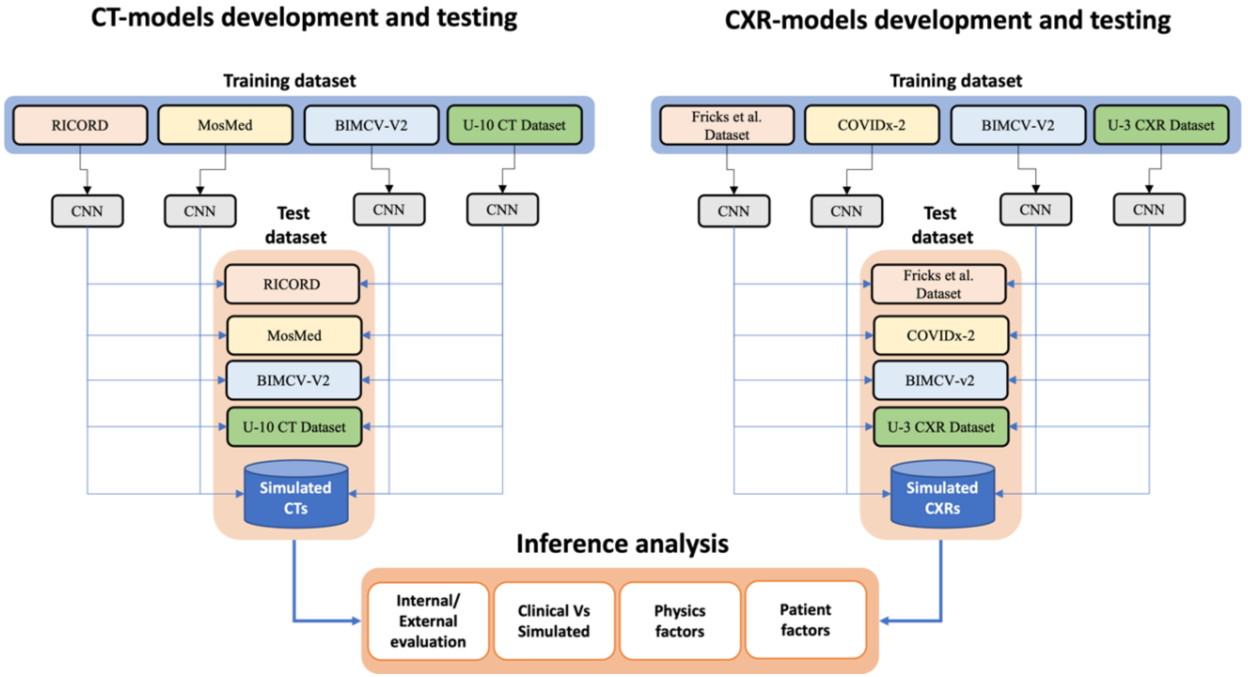


FIGURE 1. Study design overview. 12,844 CT scans and 25,219 CXR images for COVID-19 diagnosis were drawn from 13 clinical datasets comprising single or multiple centers (Supplement Fig. 1-2). Multiple deep-learning-based models were developed using these clinical datasets. All models underwent internal testing (held-out from the same training dataset) and external testing (all other datasets). Further external testing was performed using virtually simulated CT and CXR images to analyze effect of patient and imaging physics factors.

Resource Center (MIDRC) [19]. Image data from up to thousands of patients led to a plethora of AI models for the diagnosis of COVID-19 [18, 20-25], yet a review of 62 studies asserted that none of these models were fit for clinical use due to methodological flaws and underlying biases [18]. The lack of experimental controls in the clinical data also precluded further analysis for model explainability. The rare combination of so much data accompanied by widespread problems in reproducibility offers our field a rare opportunity to glean important lessons, which may inform not only our response to future health crises but also routine clinical practice.

A promising solution to this challenge lies in the use of the Virtual Imaging Trial (VIT) approach, which simulates three key components of an imaging trial: patients, scanners, and readers [26]. VITs provide practical opportunities to quantify the effects of imaging technologies and patient factors on radiological diagnosis. VITs allow the controlled comparison of alternative imaging modalities or the optimization of acquisition protocols. Previously, the VIT approach was demonstrated in simulated CT images using computational anatomical models of patients with and without COVID-19 pneumonia [27, 28]. To address reproducibility, a VIT framework can generate virtual image data with pixel-level ground truth for truly independent external validation, which helps tackle the ongoing reproducibility crisis in AI by allowing rigorous and unbiased evaluation of model performance across diverse scenarios. To provide transparency, the VIT approach can simulate a range of imaging technologies and patient characteristics, thus elucidating which factors drive model performance. In our study, virtual patients are modeled using computational anatomical phantoms, which replicate a diverse range of anatomies and pathologies, including various manifestations of COVID-19. These virtual patients are then imaged using simulated scanners, replicating the physical and technical characteristics of actual medical imaging devices, ensuring realistic imaging conditions. The readers, representing radiologists, are simulated to interpret the images, allowing us to evaluate the diagnostic performance of AI models under consistent conditions. By simulating these components, VITs enable independent validation and detailed analysis of the factors affecting AI model performance, addressing reproducibility issues effectively. This approach ensures that the models are evaluated in a variety of scenarios, leading to more reliable and generalizable AI systems in clinical practice.

Previously, we performed external validations of open-source deep-learning models for case-level COVID-19 detection with CT and CXR images [29, 30]. The current study builds upon that prior work by including simulated CT and CXR exams from the same virtual patients at effective doses ranging over multiple orders of magnitude that overlapped between the two modalities. Augmented with twice as many clinical datasets and multiple AI models compared to the prior study, we aimed to:

- Unpack the interplay of dataset-model matching and mismatching on the results.
- Compare model performance on virtual and clinical data.
- Systematically assess CT and CXR from the same cases.
- Evaluate the influence of patient- and physics-based factors on the generalizability of the results.

Our approach leverages the VIT framework to provide a controlled environment for evaluating AI models, ensuring robust and reproducible results. By simulating patients, scanners, and readers, VITs allow us to quantify the effects of various factors on radiological diagnosis and enhance the transparency and reliability of AI systems in medical imaging.

II. LITERATURE REVIEW

The development of AI models in medical imaging faces challenges in generalizability and reproducibility. Studies like Rubin et al.[1] and Kanne et al. [2] highlighted imaging protocol variability and early AI model limitations for COVID-19. Gunraj et al. [3] and Harmon et al. [4] showed high initial accuracy but significant performance drops in external validation. Roberts et al. [18] found none of the reviewed AI models fit for clinical use due to methodological flaws. Previous work involved VITs for validating deep-learning models for COVID-19 detection using clinical and simulated datasets [29, 30]. Arun et al. [31] highlighted the limitations of Grad-CAM due to repeatability and reproducibility issues.

TABLE I
Summary of Key Literature on AI in COVID-19 Imaging

Study	Focus	Findings	Limitations
Rubin et al. [1]	Role of chest imaging in COVID-19	Multinational consensus on imaging protocols	Lack of AI-specific insights
Kanne et al. [2]	COVID-19 imaging overview	Summary of known and unknown aspects	General overview, not AI-focused
Gunraj et al. [3]	COVIDNet-CT	AI model design for COVID-19 detection	Performance drop in external testing
Harmon et al. [4]	AI for COVID-19 detection	Multinational dataset evaluation	Limited generalizability
Javaheri et al. [5]	CovidCTNet	Open-source AI model for COVID-19	Small cohort limitations
Jin et al. [6]	AI system development for COVID-19	Performance evaluation	Lack of external validation
Bai et al. [7]	AI augmentation of radiologist performance	Comparative study	Variability in external datasets

III. METHOD

Institutional Review Board approval was obtained for this exempt study that used only anonymized image data and simulated phantom data. We briefly outline our study design that is necessary to understand the experimental results and analysis.

Multiple lightweight convolutional neural network (CNN) models (detailed in Section III.D) with residual connections were developed to process CT or CXR images efficiently. These lightweight CNNs, designed to reduce computational complexity while maintaining high accuracy, were used to classify cases as positive or negative for COVID-19. Multiple clinical datasets were acquired [11, 13-15, 17, 22, 32-35]. which vary in size, diversity, demographics, and class definitions. In addition, we simulated image data from a population of 4D-XCAT

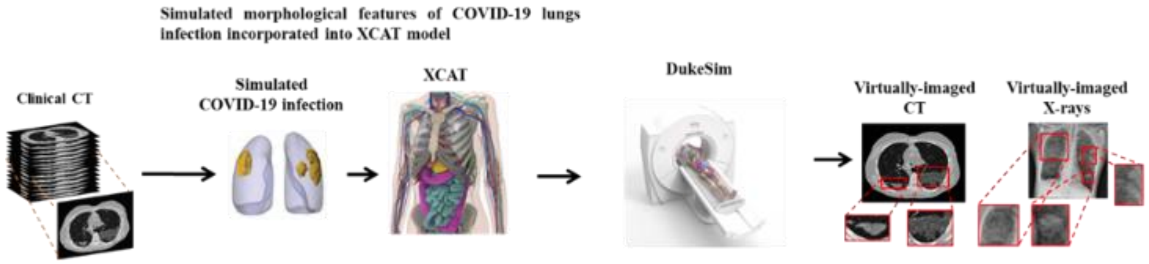

Simulated morphological features of COVID-19 lungs infection incorporated into XCAT model
Clinical CT
Simulated COVID-19 infection
XCAT
DukeSim
Virtually-imaged CT
Virtually-imaged X-rays

FIGURE 2. An overview COVID-19 computation phantoms development and simulated CT and CXR images.

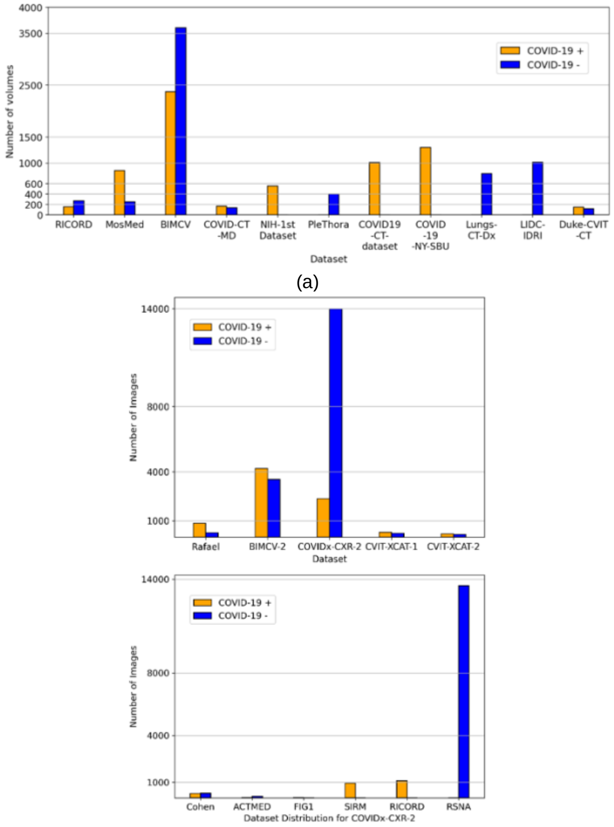
models with varying COVID-19 size and distribution, developed in a previous study then generating images using virtual CT and CXR scanners (DukeSim, CVIT, Duke University) [27]. The CNN models were trained using single and various combinations of clinical datasets. In parallel experiments, CT or CXR clinical data were analyzed for internal and external performance shift. The simulated data were reserved as a separate external validation. Finally, by varying the virtual imaging trial parameters, we explored how performance may be affected by factors pertaining to the patients (i.e., infection size) or imaging physics (i.e., effective dose and modalities). An illustration of the overall workflow of the analysis is presented in Figure 1.

A. CLINICAL DATASET COMPILATION

Define abbreviations and The clinical CT data included a total of 12,844 volumes of 7,452 patients from 10 datasets: RICORD [17], MosMed [14] BIMCV-COVID-19 +/- (BIMCV-V2) [13], COVID-CT-MD [11], CT Images in COVID-19 [12], PleThora [35], COVID19-CT-dataset [32], Stony Brook University COVID-19 Positive Cases (COVID-19-NY-SBU) [15], A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) [33], and Lung Image Database Consortium / Image Database Resource Initiative (LIDC-IDRI) [36]. These datasets had different prevalences of COVID-19 positive and negative images (Figure 3a) and demographics. Summary statistics regarding the datasets are detailed in Table 2.

Furthermore, all ten clinical CT datasets above were combined to create the U-10 CT dataset, which provides a more diverse dataset for factors such as patient population and demographics, disease appearances, CT systems, and imaging protocols. Figure 4 shows the inclusion and exclusion criteria followed in the curation of the clinical CT data.

CXR analysis included 25,219 clinical CXR images collected from 3 datasets: Fricks et al. [22], BIMCV [13], and COVIDx-CXR-2 [34]. These datasets also had different prevalences of COVID-19 positive and negative images (Figure 3b) and demographics. All three clinical CXR datasets were also combined to form the U-3 CXR dataset. In one of the datasets, COVIDx-CXR-2, positive images were from different sources, but the negative class was much larger and mainly from one source, namely the RSNA Pneumonia Detection Challenge [37] (Figure 3b). To ensure a balanced training and validation process for the unified U-3 dataset, the negative cases were randomly subsampled to achieve an equal distribution between the two classes.



(a)

(b)

FIGURE 3. Histograms showing distribution of COVID-19 positive (+) and negative (-) cases among different datasets (clinical and simulated) (a) CTs and (b) CXRs. In the latter, COVID-CXR-2 is further decomposed into its subsets. Log scale is used to show the large variation in numbers of exams. Note that the prevalence varies greatly, and some datasets contain only one class.

B. SIMULATED DATASET CONSTRUCTION

The XCAT computational phantoms used in this study are based on the method described in detail by Abadi et al. [27] An overview of the method is illustrated in Figure 2. Creating computational phantoms for COVID-19 is a

TABLE II
Clinical CT patient datasets utilized in model development and testing. The combination of all ten constitutes the U-10 CT dataset. Demographic values are reported as the percentage of patient sex and mean of patient age.

No	Dataset	Source	Demographics	Category	Train*	Validation*	Test*
1.	RICORD [17] (1b,1b)	Turkey, USA, Canada, Brazil	44% women Age 54 ±17	COVID+	66 (90)	22 (32)	22 (33)
				COVID-	70 (72)	23 (23)	24 (25)
				Total	**136 (162)**	**45 (55)**	**46 (58)**
2.	MosMed [14]	Russia	56% women Age 47	COVID+	512 (512)	170 (170)	174(174)
				COVID-	152 (152)	50 (50)	52 (52)
				Total	**664 (664)**	**220 (220)**	**226 (226)**
3.	BIMCV-V2 [13]	Spain	42% women. Age 64 ±16	COVID+	455 (1421)	152(484)	152(470)
				COVID-	728 (2077)	239(706)	268(823)
				Total	**1183 (3498)**	**391 (1190)**	**420 (129)**
4.	COVID-CT-MD [11]	Iran	40% women. Age 51 ±16	COVID+	101(101)	33 (33)	35 (35)
				COVID-	81 (81)	27 (27)	28 (28)
				Total	**182 (182)**	**60 (60)**	**63 (63)**
5.	An et al. [12]	Multi-center	N/A	**COVID+**	379 (391)	126 (129)	127 (130)
6.	PleThora[35]	USA	31% women. Age 68 ± 10	**COVID-**	241 (241)	80 (80)	81 (81)
7.	COVID19-CT [32]	Iran	39.1% women Age: 47 ± 16	**COVID+**	604 (604)	201 (201)	202 (202)
8.	COVID-19-NY-SBU[15]	USA	43% women. (Age: ranges between 18-90 years)	**COVID+**	251 (739)	84 (278)	84 (282)
9.	Lungs-CT-Dx [33]	China	46% women, Age 61 ± 10	**COVID-**	207 (479)	69 (154)	70 (164)
10.	LIDC-IDRI [36]	USA	N/A	**COVID-**	606 (611)	202 (204)	202 (203)
	Total / U-10 CT				**4453 (7571)**	**1478 (2571)**	**1521 (2702)**

Note-* Number of patients (number of scans), COVID+= COVID-19 positive, COVID-= COVID-19 negative, COVID-19-NY-SBU = Stony Brook University COVID-19 Positive Cases, Lungs-CT-Dx= A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis.

detailed process that unfolds in three distinct stages: constructing the body framework, detailing the morphological characteristics of lung abnormalities, and replicating the texture and composition of affected lung tissues.

Body Framework Construction: The process begins with the development of the body's framework using the 4D extended cardiac-torso (XCAT) model from the group at Duke University [38, 39], The XCAT model provides a comprehensive foundation with detailed anatomy, dynamic organ motions, and textured tissues, built from real patient data, ideal for diverse COVID-19 patient simulations.

Detailing Lung Abnormalities: The second stage involves the meticulous detailing of lung abnormalities typical of COVID-19, such as ground-glass opacities (GGO) and consolidations. This is achieved by examining CT scans from clinically confirmed 20 patients of COVID-19, where the abnormalities were manually segmented and modeled in a series of surfaces mimicking the morphology [27], These modeled features are then integrated into the XCAT phantoms, ensuring a match in body dimensions, gender, and age, to accurately represent the disease's manifestations within the computational models.

Replicating Lung Tissue Composition: The last phase involves fine-tuning the phantom's lung textures and materials to mirror the properties of the lung tissues affected by COVID-19 within the phantoms. This involves adjusting the lung parenchyma's texture in the computational model to reflect the changes observed in actual CT images, such as the addition of fluids in the case of GGO or the uniform texture seen in consolidations. These adjustments ensure that the simulated lung tissues closely mimic the radiological features of COVID-19, allowing for realistic simulation outcomes.

Simulated Images: These COVID-19 XCAT computational phantoms were imaged using DukeSim [27, 28], a validated radiographic simulator that combines ray tracing and Monte Carlo simulation to produce realistic CT and chest radiographs, tailored to specific scanner protocols and physics.

TABLE III

Clinical CXR Patient Cohorts utilized in model development and testing. Demographic values are reported as the percentage of patient sex and mean of patient age.

No	Dataset	Source	Demographics	Category	Train	Validation	Test
1.	Fricks et al.[22]	Iran, Italy,	N/A	COVID+	544	136	171
		USA		COVID-	174	44	55
				Total	**718**	**180**	**226**
2.	BIMCV-V2 [13]	Spain	46% Women	COVID+	2694	674	843
			Age 63 ± 17	COVID-	2265	566	708
				Total	**4959**	**1240**	**1551**
3.	COVIDx-CXR-2	Multi-center	N/A	COVID+	1727	431	200
	[34]			COVID-	11034	2759	200
				Total	**12761**	**3190**	**400**
	Total		N/A		**18438**	**4610**	**2177**
4	**U-3 CXR dataset**		N/A	COVID+	4965	1241	1214
				COVID-	4965	1241	963
				Total	**9930**	**2482**	**2177**

Simulated CT and CXR datasets were generated in three steps using a virtual imaging trial (VIT) framework [26, 27], These virtual patient models with or without the disease were imaged using an x-ray image acquisition simulator (DukeSim, CVIT, Duke University) [27, 28] Virtual scans were repeated at different effective doses (0.01, 0.1, 0.3, 1.6, 5.6, and 11.2 mSv). The dose settings were selected to represent a wide range of clinical applicability, as well as a direct comparison of CT and CXR images at the same hypothetical dose. For the CXR acquisitions, two commercial post-processing algorithms (denoted as Algorithm A and B to maintain. For the CXR acquisitions, two commercial post-processing algorithms (denoted as algorithms A and B to maintain confidentiality) were applied to examine the effects of vendor heterogeneity. Table 4 shows the characteristics of the generated CT and CXR images.

The concept of a simulated dataset is integral to our study, providing a robust alternative to conventional datasets. These datasets offer precise control over imaging parameters, including patient anatomy, disease characteristics, and imaging conditions, ensuring consistency and reproducibility. Unlike conventional datasets, which often suffer from variability in patient demographics and imaging protocols, simulated datasets enable a controlled and repeatable generation of imaging data. As shown in Table 5, simulated datasets possess advanced features such as comprehensive patient-level, slice-level, and pixel-level annotations, and the ability to image the same virtual patient with both CT and CXR at multiple doses. These features facilitate rigorous evaluation and validation of AI models, allowing for systematic studies of the effects of various factors on model performance. By integrating simulated datasets with clinical datasets, we aim to enhance the generalizability and reliability of AI systems in medical imaging, ensuring their applicability in diverse clinical scenarios.

C. PRE-PROCESSING

Standard preprocessing was performed on both CT and CXR images. Each CT volume was resampled to voxel dimensions of 2 mm × 2 mm × 5 mm (w, h, d). Intensities were clipped between -1000 to 500 HU, then standardized to a mean of 0 and standard deviation of 1. To reduce computational cost and the influence of

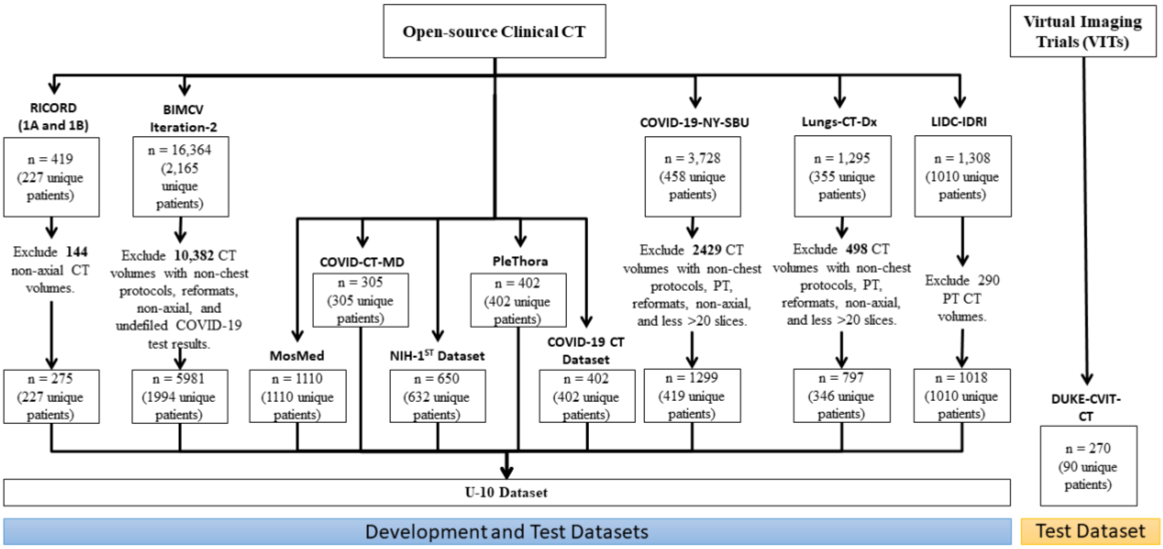

Open-source Clinical CT
Virtual Imaging Trials (VITs)
RICORD (1A and 1B)
n = 419 (227 unique patients)
Exclude 144 non-axial CT volumes.
n = 275 (227 unique patients)
BIMCV Iteration-2
n = 16,364 (2,165 unique patients)
Exclude 10,382 CT volumes with non-chest protocols, reformats, non-axial, and undefiled COVID-19 test results.
n = 5981 (1994 unique patients)
MosMed
n = 1110 (1110 unique patients)
COVID-CT-MD
n = 305 (305 unique patients)
NIH-1ST Dataset
n = 650 (632 unique patients)
PleThora
n = 402 (402 unique patients)
COVID-19 CT Dataset
n = 402 (402 unique patients)
COVID-19-NY-SBU
n = 3,728 (458 unique patients)
Exclude 2429 CT volumes with non-chest protocols, PT, reformats, non-axial, and less >20 slices.
n = 1299 (419 unique patients)
Lungs-CT-Dx
n = 1,295 (355 unique patients)
Exclude 498 CT volumes with non-chest protocols, PT, reformats, non-axial, and less >20 slices.
n = 797 (346 unique patients)
LIDC-IDRI
n = 1,308 (1010 unique patients)
Exclude 290 PT CT volumes.
n = 1018 (1010 unique patients)
DUKE-CVIT-CT
n = 270 (90 unique patients)
U-10 Dataset
Development and Test Datasets
Test Dataset

FIGURE 4. **Flowchart of inclusion and exclusion criteria for the chest CT scans. n= number of CT volumes. A total of 16,949 CT scans of 11,166 patients were used for model development and testing. There were ten clinical datasets: RICORD [17], MosMed [14], BIMCV-COVID-19 +/- (BIMCV-V2),[13] COVID-CT-MD [11], CT Images in COVID-19 [12], PleThora [35], COVID19-CT-dataset,[32] Stony Brook University COVID-19 Positive Cases (COVID-19-NY-SBU) [15], A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lungs-CT-Dx) [33], and Lung Image Database Consortium / Image Database Resource Initiative (LIDC-IDRI) [36], These ten clinical datasets were united into the U-10 CT Dataset. Additionally, simulated data were from the Center for Virtual Imaging Trials CT Dataset, Duke-CVIT-CT [27].**

background organs, three-dimensional (3D) patches of size 160×160×96 (w, h, d) were centered about the lungs. The patch size was based on average lung size plus a margin to allow for patient variability. CXR images were resized and randomly cropped to a size of 300x384 pixels, then standardized to 0.5 mean and 0.5 standard deviation to maintain consistency with the pre-trained dataset.

TABLE IV

Simulated (CVIT-COVID) dataset.

	Number of virtual exams	
Effective dose (mSv)	**COVID-19**	**Negative**
	CVIT-COVID-CT	
0.3	50	40
1.6	50	40
5.6	50	40
11.2	50	40
Total (CT)	**200**	**160**
	CVIT-COVID-CXR	
0.01	50	40
0.10	50	40
0.3	50	40
Total (CXR)	**150**	**120**

D. MODEL DEVELOPMENT AND TRAINING

We previously confirmed that complex deep learning models can reproduce the near-perfect performance reported in previous studies. Due to fundamental problems with the data, however, that performance would drop in external testing to chance [29, 30]. To minimize overtraining, we intentionally selected lightweight ResNet-like models [40, 41] and trained four separate CT-based models using the RICORD, MosMed, BIMCV, and U-10 CT datasets. Additionally, the ResNet architecture has consistently proven to perform well in various medical imaging tasks [29, 40, 41]. Similarly, for CXR, we trained four different EfficientNetv2 [42] models using the data from Fricks et al., BIMCV, COVIDx-CXR-2, and U-3 CXR datasets, respectively. Each dataset was randomly divided by the patient into subsets of training (60%), validation (20%), and testing (20%). No cross-validation was performed; instead, we utilized a train-validation-test split. As we aimed to assess the utility of virtual data to assess clinically trained algorithms, for virtual data, no training was applied – the model as trained by clinical data was applied to the entire dataset for testing**.**

The clinical datasets were selected to encompass a range of study samples. Specifically, limited datasets included RICORD for CT and Fricks et al. for CXR, while U-10 CT and U-3 CXR were more diverse. Detailed descriptions of the models and training processes can be found in the Methods (Figure 1).

CT models used a simple 3D CNN inspired by ResNet [43], the architecture is shown in Figure 5. After initial convolution, features were learned across two resolution scales, then halved by max-pooling (pooling size 2×2×2) while doubling the number of filters. The last R-block features underwent batch normalization, rectified linear unit (ReLu), global max-pooling, dropout (dropout rate 0.5), and finally, a dense classification layer with sigmoid activation for binary case-level COVID-19 detection. Additionally, we applied L2 regularization with a coefficient of 0.001 to prevent overfitting. The stochastic gradient descent (SGD) optimizer was used to optimize the weights with decay learning rate, and weighted binary cross-entropy was used as the loss function. Weights were initialized to a uniform distribution. To retain the natural prevalence, no class balancing was performed during training. The hyperparameters for the CT models were set as follows: initial learning rate of 1e-6, maximum learning rate of 1e-4, learning rate decay of 1e-2, batch size of 24, and 300 training epochs. CXR models were based on Efficientnetv2 with the original architecture [42], SGD was selected as the optimizer with the learning rate scheduler, [44] initial learning rate of 0.01, and cross-entropy loss. All models were developed using Python TensorFlow v2.6 and PyTorch deep learning frameworks. All model weights, initial hyper-parameters, and code are made publicly accessible [45].

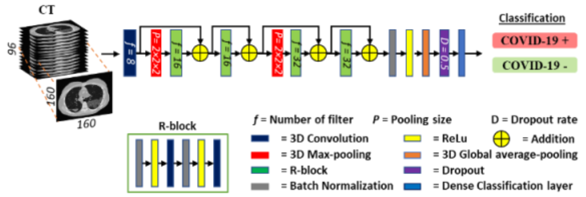


FIGURE 5. **3D CNN architecture for CT classification of COVID-19. The classification module is a 3D Resnet-like model with 2 R-Blocks in each resolution. The number of filters is denoted as f. The final output is a tensor of the probability of being COVID-19 positive or negative.**

E. EVALUATION AND STATISTICAL ANALYSIS

We conducted a series of studies to assess the model performance on clinical datasets and simulated datasets. For the simulated dataset, we further evaluated the influence of physics factors, i.e., the imaging modality and the acquisition of effective dose. acquisition-based evaluation performed to assess model performance based on different imaging protocols and effective radiation doses used during image acquisition. We additionally evaluated the effect of the patient factor of infection size, to understand the impact of infection severity on model performance. The simulated COVID-19 pneumonia cases were divided into two groups: "higher" infection (above the median value of 2.6% of total lung volume) and "lower" infection (below this median value). This approach helps in assessing how well the AI models perform across a spectrum of disease severity and identifying any performance biases or limitations. Other classifications in our study are based solely on the presence or absence of COVID-19, allowing us to assess the models' ability to distinguish between COVID-19 positive and negative cases. To support our findings and assess the significance of the results, all performances were evaluated using the receiver operating characteristic area under the curve (AUC) with 95% confidence interval (CI) calculated by the DeLong method as implemented by pROC 1.16.2 in R 3.6.1 with 2000 bootstrapping samples [46].

IV. RESULTS

A. EVALUATION OF THE MODELS ON CLINICAL DATASETS

As depicted in Figure 6, clinical CT and CXR models exhibited a consistent drop in performance from internal to external testing, and those differences often exceeded the confidence intervals. While some loss of performance is expected in external testing, these remarkably consistent differences indicate systemic differences across these datasets. The CT models showed an internal validation AUC range of 0.69 to 0.85, whereas external testing consistently dropped to between 0.54 and 0.76. Similarly, for CXR models, internal performance ranged from an AUC of 0.77 to 1, while external testing AUC again dropped to a range of 0.51 to 0.73. Models trained on the most diverse datasets (U-3 CXR and U-10 CT) consistently yielded a testing performance that was the highest or second highest. Notably, despite its size, the COVIDx-CXR-2 dataset for CXR was very biased, resulting in perfect internal validation and near-perfect external testing even for the U-3 model that was trained on all three datasets.

B. EVALUATION OF THE MODELS ON SIMULATED DATASETS

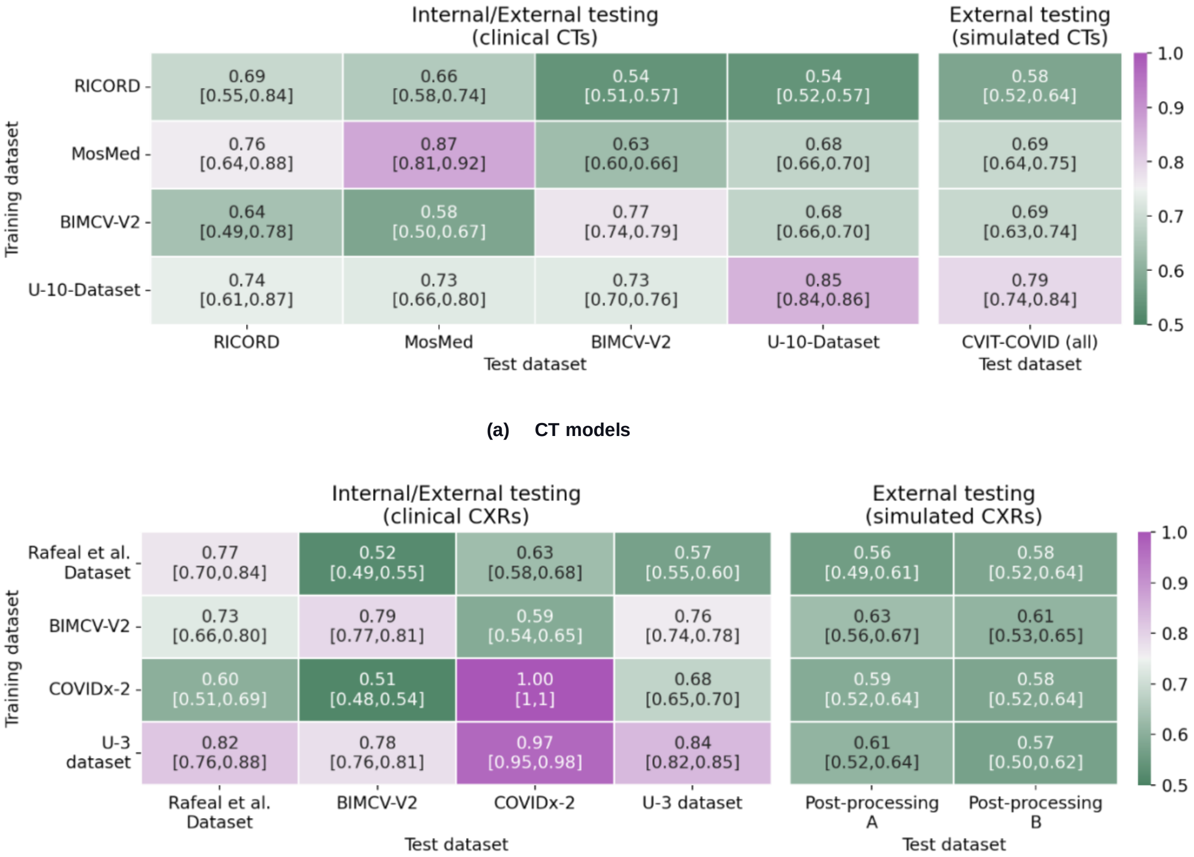
As shown in Figure 6, compared to clinical data, all CT models performed consistently with intermediate AUC values on these simulated data. In other words, simulated data was closer to the training clinical data than some of the actual clinical data, which suggests the simulated data is adequately realistic and often may be less biased. Among the CT models, training with the most diverse U-10 CT dataset yielded the highest testing performance on the simulated CT images, outperforming all three of the clinical datasets. This is remarkable since those three clinical datasets contributed to the U-10 CT training dataset, whereas the simulated data is completely independent. Conversely, all CXR models displayed comparably poor performance on the simulated CXR images.

C. PATIENT-BASED EVALUATION

Assessing the effect of infection size on the performance of models, Figure 7 shows all models performed better on both CT and CXR images with higher infection compared to images with lower infections.

TABLE V
Attributes of CT and CXR datasets. Note that simulated data are the only ones that contain all attributes, including the advanced features where the same virtual patient can be imaged with both CT and CXR at multiple doses, with multiple CXR post-processing. X= available.

Datasets	Class Type		Label Level			Advanced features
	COVID-19 positive	COVID-19 negative	Patient-level	Slice-level	Pixel-level	
CT datasets						
RICORD[17]	**X**	**X**	**X**			
MosMed[14]	**X**	**X**	**X**			
BIMCV-Iteration 2[13]	**X**	**X**	**X**			
COVID-CT-MD[11]	**X**	**X**	**X**	**X**		
An et al. dataset[12]	**X**		**X**			
COVID19-CT-dataset[32]	**X**		**X**			
COVID-19-NY-SBU[15]	**X**		**X**			
Lungs-CT-Dx[33]		**X**	**X**			
LIDC-IDRI[36]		**X**	**X**			
Duke-CVIT-CT	**X**	**X**	**X**	**X**	**X**	**X**
CXR datasets						
Fricks *et al.* dataset[22]	**X**	**X**	**X**	N/A		
BIMCV-2[13]	**X**	**X**	**X**	N/A		
COVIDx-CXR-2[34]	**X**	**X**	**X**	N/A		
Duke-CVIT-CXR	**X**	**X**	**X**	N/A	**X**	**X**



(a) CT models

(b) CXR models

FIGURE 6. Confusion matrix of case-level COVID-19 detection performance of (a) CT and (b) CXR models. Training dataset is shown in rows and testing dataset in columns; diagonal represents internal validation, while off-diagonal entries are external testing. Additional external testing on simulated images is shown on the right. Performance is reported as receiver operating characteristic area under the curve with 95% confidence interval. All models generally performed worse on external testing with both clinical and simulated data. However, models trained with the union datasets (U-10 CT and U-3 CXR) consistently yielded the highest external testing performance. Furthermore, simulation testing consistently provided intermediate results that may be more indicative of true performance.

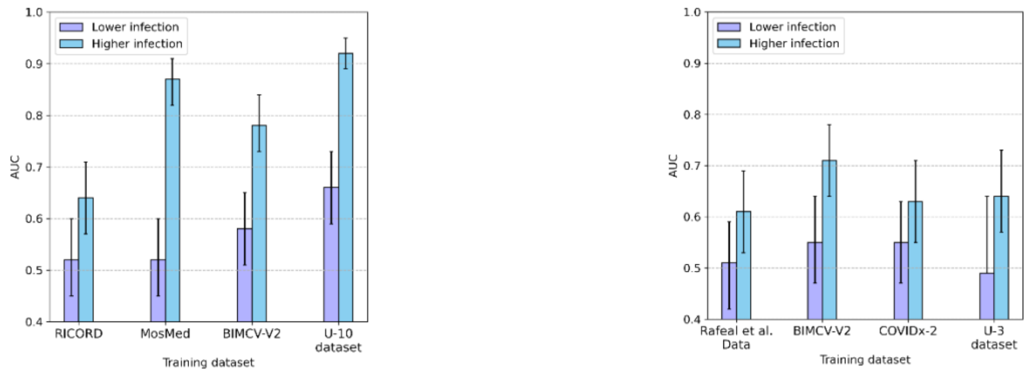
D. *ACQUISITION-BASED EVALUATION*

Assessing the effect of infection size For the same virtual patients, we assessed the performance of models over a wide, overlapping range of effective doses for the virtual CT and CXR acquisitions. As shown in Figure 8, the 3D CT models consistently outperformed the 2D CXR models, but the confidence intervals for the AUCs overlapped. Within each modality, although the effective dose (mSv) varied by 30-fold to represent the widest possible range of clinical use, there was no statistically significant change in performance [39, 47].

V. Discussion

There has been considerable research to develop AI models to improve radiology diagnosis. However, the practical application of these models in clinical practice has been hindered by two related challenges. First, models often underperform when applied to a new dataset with different attributes such as patient demographics, acquisition protocols, or scanner vendors. Second, most models function as "black boxes" that lack interpretability, making it difficult to determine which factors account for the poor performance. These issues became particularly evident during the urgent scientific response to COVID-19, when many early studies reported high performances that did not generalize [20, 23, 25, 29, 30, 48]. Although biases in AI models for healthcare may be unavoidable, a comprehensive understanding of such factors, supported by effective external testing, can raise the confidence that such models are trustworthy [18, 20, 49]. This study addresses the problem of biases in medical imaging AI models by leveraging clinical and simulated data for independent testing, thus enabling the evaluation of both generalizability and interpretability. The clinical relevance of these findings is substantial, as improving the generalizability and transparency of AI models can enhance their reliability in diverse clinical settings. By ensuring that

AI models perform consistently across various datasets and provide interpretable results, this research has the potential to significantly impact medical practice. It can lead to more accurate diagnoses, personalized treatment plans, and ultimately better patient outcomes, thereby integrating AI more effectively into routine clinical workflows.



(a) CT models

(b) CXR models

FIGURE 7. **Both (a) CT and (b) CXR models each trained on four datasets (represented on the x-axis), consistently demonstrated superior performance in "higher infection" cases, where the pneumonia volume exceeded the median, compared to "lower infection" cases that fell below the median. For CXR, results were almost identical for the two post-processing algorithms, so only algorithm A is shown. Error bars represent the 95% confidence interval.**

We compiled a large cohort of clinical CT and CXR images from dataset resources representing over 22,000 patients. Despite the large amount of training data, however, model performance was still impaired due to class imbalance and confounding issues such as radiographic markers, incorrect image orientation, and collimator edges [20, 25, 50]. Proper data curation is time-consuming and requires domain expertise in medical imaging, rendering this process prohibitively costly.[51] Therefore, external validation of AI models is essential to rule out biases [18, 20]. To address these needs, this study utilized a VIT simulation platform [26, 27]. Using simulated CT and CXR images provides two crucial advantages. First, simulated image data enables external validation that is not only truly independent but also controlled. Second, the VIT framework allows the evaluation of the models under different patient- and physics-based factors, which offers interpretability and reveals clinical or technical insights. VIT simulations facilitate conducting medical imaging studies in a trustworthy, reproducible, and practicable manner.

Our primary objective was to analyze the impact of dataset variability on model development. Unlike most studies, we intentionally chose to use very lightweight networks to minimize overfitting [40, 41]. Nevertheless, all models still dropped in performance substantially from internal to external testing, which was in line with previously reported studies [20, 25, 29, 30]. Since model performance reflects the underlying data, this generalizability gap suggests the lack of diversity in the existing datasets with regard to institution bias, patient demographics, disease appearances, and image quality [20, 23, 50]. To minimize such bias, we trained on the combination of multiple diverse datasets, U-10 CT and U-3 CXR, and the resulting models outperformed the single-dataset models in independent testing. The model trained on the diverse U-10 CT dataset demonstrated very consistent performance across all three clinical datasets with an AUC of approximately 0.73. Unlike the individual testing results showing considerable high and low bias, this moderate result is more credible and may indicate a more representative performance for this challenging clinical task. These general trends were also observed for the CXR datasets but with considerable residual bias due to the disproportionate influence of the COVIDx-CXR-2 dataset, which is much larger than other datasets and leads to confounding bias as its positive and negative cases come from different institutions. This quandary shows that despite rigorous training and external testing, AI models can still be affected by fundamental data biases. The VIT process proved to deliver a more realistic portrayal of true clinical performance. When many models were tested on simulated images, performances fell consistently within the middle of the range of external testing on clinical datasets, suggesting that the simulations presented data with an appearance that was realistic and relevant. This is highly encouraging considering the models were applied to the virtual data without even being trained on them, highlighting the potential generalizability of simulated datasets to evaluate AI-based diagnosis algorithms. Unlike clinical datasets, the simulated images are further free of institutional bias or other confounding factors, because the VIT framework offers precisely reproducible controls in terms of patient sampling as well as physical image formation. This enabled us to compare identical virtual patients with and without COVID-19 and also to conduct virtual imaging of each patient using

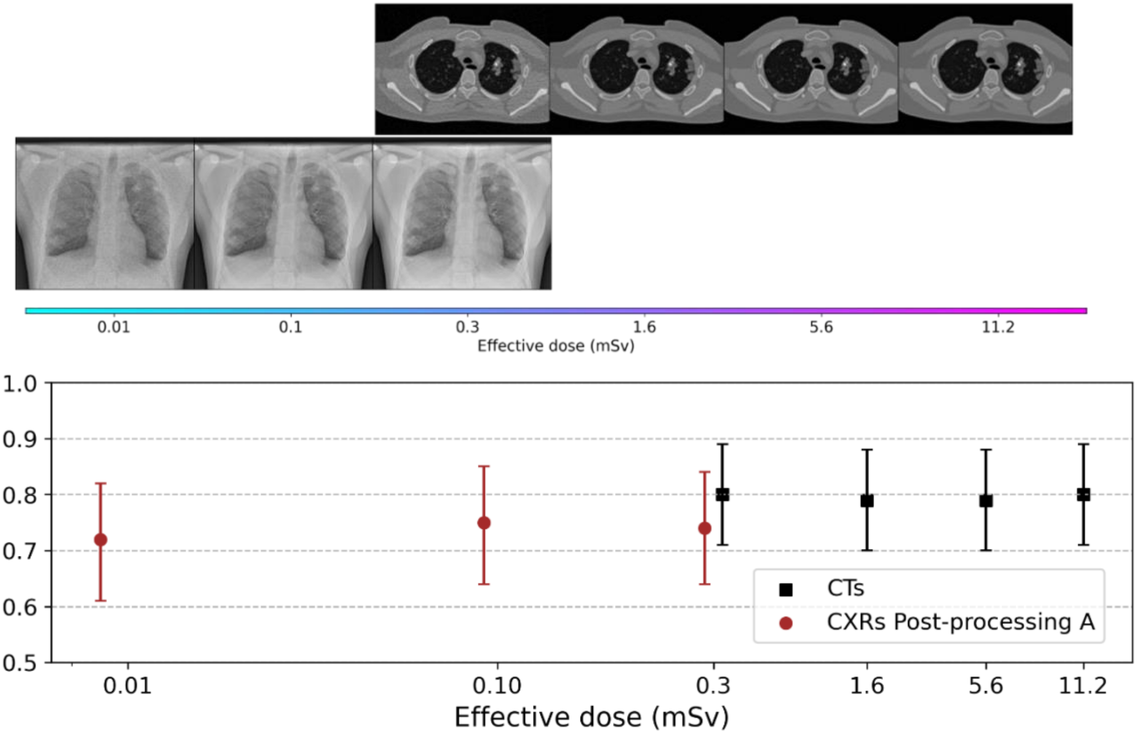


FIGURE 8. **Simulated images were used to evaluate physics-based factors. Although models consistently performed better on CT over CXR, the differences were not significant at the shared dose of 0.3 mSv. Within each modality, performances were also not significantly different across a wide range of effective dose. Error bars correspond to 95% confidence interval.**

both CT and CXR. The degree of experimental control provided by VITs is not physically possible in real clinical trials. The VIT process proved to deliver a more realistic portrayal of true clinical performance. When many models were tested on simulated images, performances fell consistently within the middle of the range of external testing on clinical datasets, suggesting that the simulations presented data with an appearance that was realistic and relevant. This is highly encouraging considering the models were applied to the virtual data without even being trained on them, highlighting the potential generalizability of simulated datasets to evaluate AI-based diagnosis algorithms. Unlike clinical datasets, the simulated images are further free of institutional bias or other confounding factors, because the VIT framework offers precisely reproducible controls in terms of patient sampling as well as physical image formation. This enabled us to compare identical virtual patients with and without COVID-19 and also to conduct virtual imaging of each patient using both CT and CXR. The degree of experimental control provided by VITs is not physically possible in real clinical trials.

Our VIT analysis further provided intriguing insights into the effects of patient- and physics-based factors driving AI performance. Regardless of the training datasets for both the CT and CXR models, there was a noticeable increase in performance when the COVID-19 infection size was larger than the median value. For both imaging modalities, performances stayed consistent even across a 30-fold range in effective dose (which well exceeds the range in clinical practice), suggesting that dose may not be as relevant for the AI detection of diffuse diseases such as pneumonia. In stark contrast to the model evaluation on clinical data, our analysis confirmed that CT outperformed CXR, which was consistent with expectations since 3D CT scans provide superior spatial information over 2D CXR images.

This study had several limitations. Although the simulated CT and CXR images realistically reproduced both anatomical and physical processes, they were generated from a pool of fifty virtual patients with variable anatomy and severity of the disease. Consequently, simulation testing showed consistent trends but with large confidence intervals. The minimal impact of imaging dose observed in our study might be influenced by down-sampling during the preprocessing. Additionally, the study did not account for potential variability in scanner-specific imaging characteristics, which could affect model performance in real-world settings. Future work will increase the number of computational phantoms to represent even larger and more diverse patient populations and explore the inclusion of additional imaging parameters to improve realism. In terms of the network architectures, each modality was analyzed using a single lightweight design, foregoing extension experiments with other networks. Expanding the model evaluation to include more complex architectures could provide insights into generalizability across different network types. Models were developed only to conduct case-level detection, which is the only annotation available in almost all datasets. Furthermore, the label of COVID-19 as negative or positive was defined by each dataset, and those standards varied widely, including radiologist assessment or different diagnostic tests [1]. Some datasets included both COVID-19 pneumonia and other types of pneumonia, which may not be readily differentiated by imaging alone. Finally, future work should also aim to address these limitations by incorporating more detailed multi-class annotations and evaluating model performance under different disease classification scenarios.

VI. CONCLUSION

AI-based diagnosis models hold the potential to revolutionize healthcare. However, factors contributing to model bias remain underexplored, especially in the medical imaging domain. An essential prerequisite to clinical deployment is a robust external evaluation. The VIT framework plays a crucial role in addressing the ongoing reproducibility crisis in AI models by providing the necessary image data that is objective and controlled. By enabling consistent evaluation across diverse scenarios, VIT not only helps to identify bias but also facilitates improvements in model robustness and generalizability. By studying patient- or physics-based factors influencing model performance, these procedures also offer interpretability and opportunities for model refinement. Through these contributions, virtual imaging trials can enhance clinical trials, making them faster, more rigorous, and more reproducible.

APPENDIX A
DATA AVAILABILITY

The clinical data utilized in this study are open-source and can be referenced via the citation in Table 2. The authors are committed to promoting transparency and open science. Reasonable requests for access to an anonymized version of the private datasets (Duke-CVIT-CT and Duke-CVIT-CXR) can be made by contacting the corresponding author. Upon publication, all model weights, initial hyper-parameters, and code will be publicly accessible at https://gitlab.oit.duke.edu/cvit-public/cvit_revicovid19

ACKNOWLEDGMENT

This work was funded in part by the Center for Virtual Imaging Trials, NIH/NIBIB P41- EB028744.

REFERENCES

[1] G. D. Rubin *et al.*, "The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society," *Radiology*, vol. 296, no. 1, pp. 172-180, Jul 2020, doi: 10.1148/radiol.2020201365.

[2] J. P. Kanne *et al.*, "COVID-19 Imaging: What We Know Now and What Remains Unknown," *Radiology*, vol. 299, no. 3, pp. E262-E279, Jun 2021, doi: /10.1148/radiol.2021204522.

[3] H. Gunraj, L. Wang, and A. Wong, "COVIDNet-CT: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases From Chest CT Images," *Front Med (Lausanne)*, vol. 7, p. 608525, 2020, doi: /10.3389/fmed.2020.608525.

[4] S. A. Harmon *et al.*, "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets," *Nat Commun*, vol. 11, no. 1, p. 4080, Aug 14 2020, doi: /10.1038/s41467-020-17971-2.

[5] T. Javaheri *et al.*, "CovidCTNet: an open-source deep learning approach to diagnose covid-19 using small cohort of CT images," *NPJ Digit Med*, vol. 4, no. 1, p. 29, Feb 18 2021, doi: /10.1038/s41746-021-00399-3.

[6] C. Jin *et al.*, "Development and evaluation of an artificial intelligence system for COVID-19 diagnosis," *Nat Commun*, vol. 11, no. 1, p. 5088, Oct 9 2020, doi: /10.1038/s41467-020-18685-1.

[7] H. X. Bai *et al.*, "Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT," *Radiology*, vol. 296, no. 3, pp. E156-E165, Sep 2020, doi: /10.1148/radiol.2020201491.

[8] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images," *Pattern Recognition Letters*, vol. 138, pp. 638-643, 2020, doi: /10.1016/j.patrec.2020.09.010.

[9] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet," *Chaos, Solitons & Fractals*, vol. 138, p. 109944, 2020, doi: /10.1016/j.chaos.2020.109944.

[10] D. Kollias, A. Arsenos, and S. Kollias, "Domain adaptation, explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans," *arXiv preprint arXiv:2403.02192*, 2024.

[11] P. Afshar *et al.*, "COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning," *Sci Data*, vol. 8, no. 1, p. 121, Apr 29 2021, doi: /10.1038/s41597-021-00900-3.

[12] P. An *et al.*, "CT Images in COVID-19 [Data set]," *The Cancer Imaging Archive*, 2020, doi: /10.7937/TCIA.2020.GQRY-NC81.

[13] M. d. l. I. Vayá *et al.*, "Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients," *arXiv preprint arXiv:2006.01174*, 2020.

[14] S. P. Morozov *et al.*, "MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic," *Digital Diagnostics*, vol. 1, no. 1, pp. 49-59, 2020-05-13 2020.

[15] J. Saltz *et al.*, "Stony Brook University COVID-19 Positive Cases [Data set]," *The Cancer Imaging Archive.*, 2021, doi: /10.7937/TCIA.BBAG-2923.

[16] S. Shakouri *et al.*, "COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis," *BMC Research Notes*, vol. 14, no. 1, 2021, doi: /10.1186/s13104-021-05592-x.

[17] E. B. Tsai *et al.*, "The RSNA International COVID-19 Open Radiology Database (RICORD)," *Radiology*, vol. 299, no. 1, pp. E204-E213, 2021, doi: 10.1148/radiol.2021203957.

[18] M. 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, doi: /10.1038/s42256-021-00307-0.

[19] "Medical Imaging and Data Resource Center (MIDRC). https://www.midrc.org/," no. 6 July, 2023, 2023. [Online]. Available: https://www.midrc.org/.

[20] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610-619, 2021, doi: 10.1038/s42256-021-00338-7.

[21] D. Driggs *et al.*, "Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise," *Radiol Artif Intell*, vol. 3, no. 4, p. e210011, Jul 2021, doi: 10.1148/ryai.2021210011.

[22] R. B. Fricks *et al.*, "Deep learning classification of COVID-19 in chest radiographs: performance and influence of supplemental training," *J Med Imaging (Bellingham)*, vol. 8, no. 6, p. 064501, Nov 2021, doi: 10.1117/1.JMI.8.6.064501.

[23] J. Sun *et al.*, "Performance of a Chest Radiograph AI Diagnostic Tool for COVID-19: A Prospective Observational Study," *Radiol Artif Intell*, vol. 4, no. 4, p. e210217, Jul 2022, doi: /10.1148/ryai.210217.

[24] D. Khemasuwan and H. G. Colt, "Applications and challenges of AI-based algorithms in the COVID-19 pandemic," *BMJ Innovations*, vol. 7, no. 2, pp. 387-398, 2021, doi: /10.1136/bmjinnov-2020-000648.

[25] D. Nguyen *et al.*, "Deep Learning-Based COVID-19 Pneumonia Classification Using Chest CT Images: Model Generalizability," (in English), *Front Artif Intell*, Original Research vol. 4, no. 87, p. 694875, 2021-June-29 2021, doi: /10.3389/frai.2021.694875.

[26] E. Abadi *et al.*, "Virtual clinical trials in medical imaging: a review," *Journal of Medical Imaging*, vol. 7, no. 4, p. 042805, 2020. [Online]. Available: https://doi.org/10.1117/1.JMI.7.4.042805.

[27] E. Abadi, W. Paul Segars, H. Chalian, and E. Samei, "Virtual Imaging Trials for Coronavirus Disease (COVID-19)," *AJR Am J Roentgenol*, vol. 216, no. 2, pp. 362-368, Feb 2021, doi: /10.2214/AJR.20.23429.

[28] E. Abadi, B. Harrawood, S. Sharma, A. Kapadia, W. P. Segars, and E. Samei, "DukeSim: A Realistic, Rapid, and Scanner-Specific Simulation Framework in Computed Tomography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 6, pp. 1457-1465, 2019, doi: /10.1109/tmi.2018.2886530.

[29] F. I. Tushar *et al.*, *Virtual vs. reality: external validation of COVID-19 classifiers using XCAT phantoms for chest computed tomography* (SPIE Medical Imaging). SPIE, 2022.

[30] L. Dahal *et al.*, *Virtual versus reality: external validation of COVID-19 classifiers using XCAT phantoms for chest radiography* (SPIE Medical Imaging). SPIE, 2022.

[31] N. Arun *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021.

[32] S. Shakouri *et al.*, "COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis," *BMC Res Notes*, vol. 14, no. 1, p. 178, May 12 2021, doi: /10.1186/s13104-021-05592-x.

[33] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang, "A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis," 2020, doi: /10.7937/TCIA.2020.NNC2-0461.

[34] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci Rep*, vol. 10, no. 1, p. 19549, Nov 11 2020, doi: /10.1038/s41598-020-76550-z.

[35] K. J. Kiser *et al.*, "Data from the Thoracic Volume and Pleural Effusion Segmentations in Diseased Lungs for Benchmarking Chest CT Processing Pipelines PleThora) [Data set]," *The Cancer Imaging Archive*, 2020, doi: /10.7937/tcia.2020.6c7y-gq39.

[36] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915-931, 2011 2011, doi: 10.1118/1.3528204.

[37] "Radiological Society of North America. RSNA pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data," 2019. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data.

[38] W. P. Segars *et al.*, "Population of anatomically variable 4D XCAT adult phantoms for imaging research and optimization," (in eng), *Med Phys*, vol. 40, no. 4, p. 043701, Apr 2013, doi: 10.1118/1.4794178.

[39] Y. Zhang, X. Li, W. P. Segars, and E. Samei, "Comparison of patient specific dose metrics between chest radiography, tomosynthesis, and CT for adult patients of wide ranging body habitus," *Medical Physics*, vol. 41, no. 2, p. 023901, 2014.

[40] F. I. Tushar *et al.*, "Classification of Multiple Diseases on Body CT Scans Using Weakly Supervised Deep Learning," *Radiology: Artificial Intelligence*, vol. 4, no. 1, p. e210026, 2022, doi: /10.1148/ryai.210026.

[41] F. I. Tushar, V. D'Anniballe, G. Rubin, E. Samei, and J. Lo, *Co-occurring diseases heavily influence the performance of weakly supervised learning models for classification of chest CT* (SPIE Medical Imaging). SPIE, 2022.

[42] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, 2021: PMLR, pp. 10096-10106.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[44] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558-567.

[45] F. I. Tushar *et al.* "CVIT COVID-19 AI Model Code." Duke University. https://gitlab.oit.duke.edu/cvit-public/cvit_revicovid19 (accessed.

[46] X. Robin *et al.*, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 2011/03/17 2011, doi: /10.1186/1471-2105-12-77.

[47] M. Fujita *et al.*, "Lung cancer screening with ultra-low dose CT using full iterative reconstruction," *Japanese Journal of Radiology*, vol. 35, no. 4, pp. 179-189, 2017/04/01 2017, doi: /10.1007/s11604-017-0618-y.

[48] J. Dhont, C. Wolfs, and F. Verhaegen, "Automatic coronavirus disease 2019 diagnosis based on chest radiography and deep learning - Success story or dataset bias?," *Med Phys*, vol. 49, no. 2, pp. 978-987, Feb 2022, doi: /10.1002/mp.15419.

[49] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," in *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, 2020: World Scientific, pp. 232-243.

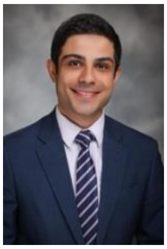
[50] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat Med*, vol. 27, no. 12, pp. 2176-2182, Dec 2021, doi: /10.1038/s41591-021-01595-0.

[51] M. J. Willemink *et al.*, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4-15, Apr 2020, doi: /10.1148/radiol.2020192224.

FAKRUL ISLAM TUSHAR is currently pursuing his PhD degree in electrical and computer engineering at Duke University, NC, USA.

LAVSEN DAHAL is currently pursuing his PhD degree in electrical and computer engineering at Duke University, NC, USA.

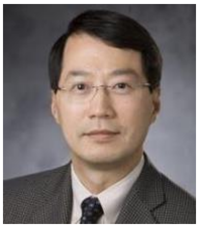
SAMAN SOTOUDEH-PAIM is currently pursuing his PhD degree in electrical and computer engineering at Duke University, NC, USA.



EHSAN ABADI, PHD is an imaging scientist at Duke University. He serves as an Assistant Professor in the departments of Radiology and Electrical & Computer Engineering, a faculty member in the Medical Physics Graduate Program and Carl E. Ravin Advanced Imaging Laboratories, and a co-lead in the Center for Virtual Imaging Trials. Ehsan's research focuses on quantitative imaging and optimization, CT imaging, lung diseases, computational human modeling, and medical imaging simulation. He is actively involved in developing computational anthropomorphic models with various diseases such as COPD, and scanner-specific simulation platforms (e.g., DukeSim) for imaging systems. Currently, his work is centered on identifying and optimizing imaging systems to ensure accurate and precise quantifications of lung diseases.

WILLIAM P. SEGARS, **PHD** is a distinguished professor at Duke University's Department of Radiology, where he specializes in developing innovative imaging techniques. His work significantly contributes to the field of radiology through both educational endeavors and pioneering research. Dr. Segars is actively involved in enhancing imaging technologies to improve patient diagnostics and treatment planning. His expertise and dedication make him a key figure in the academic and healthcare communities.

EHSAN SAMEI, PHD is a Persian-American medical physicist. He is the Reed and Martha Rice Distinguished Professor of Radiology, and Professor of Medical Physics, Biomedical Engineering, Physics, and Electrical and Computer Engineering at Duke University. He serves as the Chief Imaging Physicist for Duke University Health System, the Director of the Carl E Ravin Advanced Imaging Laboratories, and the Center for Virtual Imaging Trials (CVIT), and co-PI of one of the five Centers of Excellence in Regulatory Science and Innovation (CERSI), Triangle CERSI. He is certified by the American Board of Radiology, recognized as a Distinguished Investigator by the Academy of Radiology Research, and awarded a Fellow by five professional organizations. He founded/co-founded the Duke Medical Physics Program, the Duke Imaging Physics Residency Program, the Duke Clinical Imaging Physics Group, the Center for Virtual Imaging Trials, and the Society of Directors of Academic Medical Physics Programs (SDAMPP). He has held senior leadership positions in the AAPM, SPIE, SDAMPP, and RSNA, including election to the presidency of the SEAAPM (2010-2011), SDAMPP (2011), and AAPM (2023).



JOSEPH Y. LO, PHD is a prominent professor at Duke University, holding positions in the Departments of Radiology, Electrical and Computer Engineering, and Biomedical Engineering. He also serves as the Vice Chair for Research in Radiology and Associate Director of the Medical Physics Graduate Program. A member of the Duke Cancer Institute, Dr. Lo's research spans across various facets of medical imaging, focusing on enhancing diagnostic techniques and technologies.