AI and Ethics (2023) 3:1193–1201
https://doi.org/10.1007/s43681-022-00227-8

ORIGINAL RESEARCH

Turing test-inspired method for analysis of biases prevalent in artificial intelligence-based medical imaging

Satvik Tripathi[1,2,3,5] · Alisha Augustin[3,4] · Farouk Dako[5] · Edward Kim[1,3]

Received: 7 June 2022 / Accepted: 5 October 2022 / Published online: 24 October 2022

Abstract

Due to the growing need to provide better global healthcare, computer-based and robotic healthcare equipment that depend on artificial intelligence has seen an increase in development. In order to evaluate artificial intelligence (AI) in computer technology, the Turing test was created. For evaluating the future generation of medical diagnostics and medical robots, it remains an essential qualitative instrument. We propose a novel methodology to assess AI-based healthcare technology that provided verifiable diagnostic accuracy and statistical robustness. In order to run our test, we used a state-of-the-art AI model and compared it to radiologists for checking how generalized the model is and if any biases are prevalent. We achieved results that can evaluate the performance of our chosen model for this study in a clinical setting and we also applied a quantifiable method for evaluating our modified Turing test results using a meta-analytical evaluation framework. His test provides a translational standard for upcoming AI modalities. Our modified Turing test is a notably strong standard to measure the actual performance of the AI model on a variety of edge cases and normal cases and also helps in detecting if the algorithm is biased towards any one type of case. This method extends the flexibility to detect any prevalent biases and also classify the type of bias.

Keywords Artificial intelligence · Turing test · Diagnostic tests · Healthcare · Fairness

1 Introduction

Artificial intelligence is a computer science discipline that can analyze complicated medical data. In many clinical contexts, their ability to exploit a relationship with data collection can be employed in the diagnosis, treatment, and prediction of results [1–3].

Artificial intelligence systems are computer programs that allow computers to operate in ways that make them appear intelligent. Alan Turing (1950), a British mathematician, was one of the pioneers of modern computer science and artificial intelligence [4]. He characterized that the intelligent behavior in a computer has the capacity to exhibit human-level performance in cognitive activities, subsequently known as the "Turing test" [5, 6]. The Turing test is one of the most debatable issues in artificial intelligence and cognitive science, as some machines might not pass their test but they may still be intelligent. Alan Turing proposed the Turing test (TT) in his 1950 Mind article 'Computing Machinery and Intelligence' [7] replacing the question "Can machines think?" The goal of Turing's work is to provide a mechanism for determining whether or not a computer can

Satvik Tripathi
st3263@drexel.edu

Alisha Augustin
aia43@drexel.edu

Farouk Dako
farouk.dako@pennmedicine.upenn.edu

Edward Kim
ek826@drexel.edu

[1] Department of Computer Science, College of Computing, Drexel University, Philadelphia, PA, USA

[2] Department of Psychological and Brain Sciences, College of Arts and Sciences, Drexel University, Philadelphia, PA, USA

[3] Drexel Society of Artificial Intelligence, Drexel University, Philadelphia, PA, USA

[4] Department of Electrical and Computer Engineering, College of Engineering, Drexel University, Philadelphia, PA, USA

[5] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

think. His paper has been seen as the "starting point" of artificial intelligence (AI), whereas the TT has been regarded as its final objective. He further proposes the Imitation Game to give this idea a concrete form [8–10].

Researchers have been investigating the possible uses of intelligent techniques in every sector of medicine since the last century. Medical AI has witnessed a rise in popularity during the previous two decades. AI systems can consume, analyze, and report vast amounts of data from various modalities to diagnose disease and guide healthcare choices. In addition to diagnosis, AI can aid in the prediction of cancer patient survival rates, such as lung cancer patients. In the field of radiology, artificial intelligence (AI) is being utilized to diagnose disorders in patients using CT scans, MR imaging, and X-rays [1, 4, 11–13]. Alongside, the question of fairness and ethics has also become very crucial as more and more techniques are getting ready to be implemented in a clinical setting [14–18].

2 Problem statement

Our prime question in regards to the advancements of the state-of-the-art AI-based medical imaging algorithms and devices, how do we compute the performance of the algorithm before actually deploying and decide if it is better or at least as good as a clinician in a real-life medical setting or not [19–21]? Which also raises the concern of whether can we completely trust an AI and give it the status of an individual entity or do we need a clinician in the loop to oversee the predictions made by the AI algorithm [22]? In addition, we can examine if the current state-of-the-art techniques are good enough for clinical use or if we need more advancements in the development by comparing if they have the same level of preciseness and accuracy on a diverse cohort of patients as a professional clinician with years of experience [23].

2.1 Aim and objective

With the rise of AI-based radiological devices and algorithms providing clinical, diagnostic, and prognostic predictions, along with accuracy we need to look beyond the performance of the model in certain cases and think about whether these modalities are ethically sound and free of biases or not [24–26]. Therefore, with our proposed test, we can deeply analyze the predictions made by the algorithm and compare them against humans and see if it is safe enough to be implemented in a medical institution while considering the prevalent biases it may have [27–29]. The article draws its inspiration from A.M. Turing's classic Turing test. We propose a modified Turing test which serves as a metric to discover the AI-models true performance in the real-life clinical setting and can also help in detecting any possible biases.

3 Methodology

3.1 Dataset

For this project, we used two different datasets to train and test our dataset. For the training of our models, we used the publicly available Medical Imaging Data Resource Center (MIDRC)—RSNA International COVID-19 Open Radiology Database (RICORD) [30]. In partnership with the Society of Thoracic Radiology (STR) and the American Society of Nuclear Medicine, the MIDRC-RICORD dataset 1a was developed. For all COVID-positive thoracic computed tomography (CT) imaging studies, pixel-level volumetric segmentation with clinical annotations by thoracic radiology subspecialists was performed according to a labeling schema that was coordinated with other international consensus panels and COVID data annotation efforts.

Database 1a of the MIDRC-RICORD is comprised of 120 thoracic computed tomography (CT) images from four international sites, each of which has been annotated with precise segmentation and diagnostic labeling. For our model training process, we employed 120 Chest CT tests (axial series) as input. The data were retrieved using Cancer Imaging Archive [31]. A CT scan sample of lungs infected with COVID-19 is shown in Fig. 1.

To test our model, we used the COVID-19 CT Lung and Infection Segmentation Dataset publically available at Zenodo [32]. This dataset contains 20 COVID-19 CT scans that have been labeled and annotated. The left lung, right lung, and disease areas were labeled by two radiologists and checked by an expert radiologist before being sent to the pathology lab for testing. The dataset completely fits our research interests because of the additional human-annotated segmentation along with the ground truth.

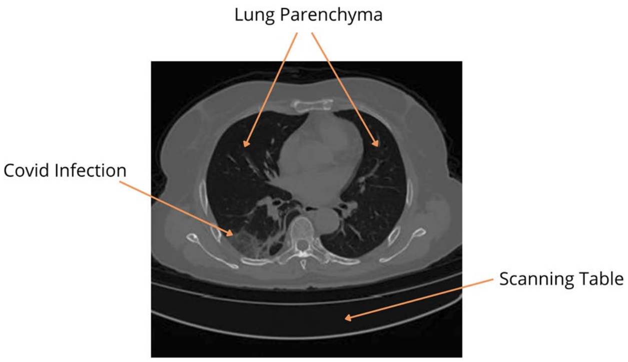


Fig. 1 Semantic segmentation of normal and edge cases lung infection produced by our proposed UNet model

3.2 Data prepossessing and training

The volumes included in this set of data have a resolution of 512 by 512 pixels, and there is no consistent number of slices throughout the cohort of patients. The input pictures from the CT include some information that is not necessary for the procedure. As a consequence of this, preprocessing is required in order to get rid of the unnecessary information included in the volumes. In order to reduce memory consumption, the dimensions of each slice in the 3D CT data were shrunk to 256 by 256.

A considerable quantity of GPU memory is essential due to the fact that the inputs are 3D volumes. To do this, we used a strategy that was based on 3D patches. Each input volume is then randomly segmented into 16 patches with dimensions of 128 by 128 by 32. A CNN that has had enough training should be insensitive to changes in translation, size, and perspective. In order to accomplish this goal, a substantial quantity of data must first be entered. We used augmentation to attain a substantial quantity of data, and as a result, we were successful in obtaining data invariance. The augmentation was carried out on each of the 128 by 128 by 32 patches in a manner that was completely arbitrary. This piece makes use of a variety of different augmentation methods, including zooming in and out, shearing, horizontal and vertical translations, and a ninety-degree rotation.

Further, all the data were divided into a training set, validation set, and test set in the ratio of 60:20:20, respectively.

3.3 Segmentation model

Semantic segmentation of the lung CT scans was performed using a VGG16-UNet model and compared its performance to other models such as UNet, UNet++, UNet3+, and Attention UNet [33–36], shown in Fig. 1. The choice of VGG16-UNet is because of its similarity to UNet's contracted layer and its number of parameters is also less than UNet [37].

The left-hand side of the network is an encoder and incorporates the 13 convolutional layers from the original VGG16. After each convolution layer, the MaxPooling operation which reduces the dimensions of the image by 2 × 2 is performed. On the right-hand side of the network, is a decoder. UpSampling operation which restores the dimensions of the image. Each UpSampling operation repeats the rows and columns of the image by 2 × 2. The skip connections are used to restore the dimensions of the image. These skip connections are implemented using the concatenate operation to combine the corresponding feature maps. Since this is a variant of the fully convolutional neural network, FCN for semantic segmentation, the spatial dimension information of the image needs to be retained hence we use the skip connections. The last convolutional layer has only one filter which is similar to a final Dense layer in most other neural networks and gives the binary mask prediction. In total, the network has about 29 convolutional layers which are followed by a PReLU activation. The PReLU has an alpha parameter that is learned during training. In addition, the last convolutional layer has a sigmoid activation function.

The semantic segmentation produced by our proposed UNet-VGG16 is shown in Fig. 2. We trained the model on multiple edge cases (artifact scans or complex patient cases) for producing a more generalized segmentation and the model performed really well in various of these cases.

3.4 Modified Turing test

This study analyzes the Turing test's possible usage in healthcare informatics, intending to highlight the broader use of diagnostic accuracy approaches for the Turing test in the present and future AI situations. As a response, we aim to create a model for a measurable diagnostic accurate scoring approach for the Turing test (how distinct are clinician and AI models?). In diagnostic accuracy testing, we adapted the Turing test to account for false positives and true negatives (Fig. 3).

As shown in Fig. 4, Examiner (A) (blinded) attempts to differentiate between a human control (B) and a computer test subject (C) versus a human test subject (D). The examiner does not know whether the test subject is human or a machine, therefore (C) vs (D) provides the diagnostic assessment. As a diagnostic test, the redesigned Turing test will now be assessed using a diagnostic accuracy technique and can provide the fast feedback of a human examiner—a method for determining if a computer "(C)" is indistinguishable from its homolog.

The findings of this test may be compared to the results of a gold-standard reference test, namely whether or not the test subject is a computer. The segmentation done by the expert radiologist (D) is shown in Fig. 3. The radiologist did not see the ground truth while doing the human-annotated segmentation.

The participants (radiologists) of the test were asked to make the prediction on the basis of how accurate the given segmentation is when compared to the ground truth, and based on this the individual may classify whether the segmentation is absolutely accurate and has details and done by a professional radiologist or if it is done by an AI model and it has some missing features. The motive of this study is not to see who does more neat segmentation rather it focuses on whether or not the machine learning algorithms pick up on the clinically important features in the scan.

Consequently, each computer may be evaluated numerous times by the same human and compared to find how biased or accurate the algorithm is. This allows us to obtain several diagnostic evaluation parameters such as sensitivity,

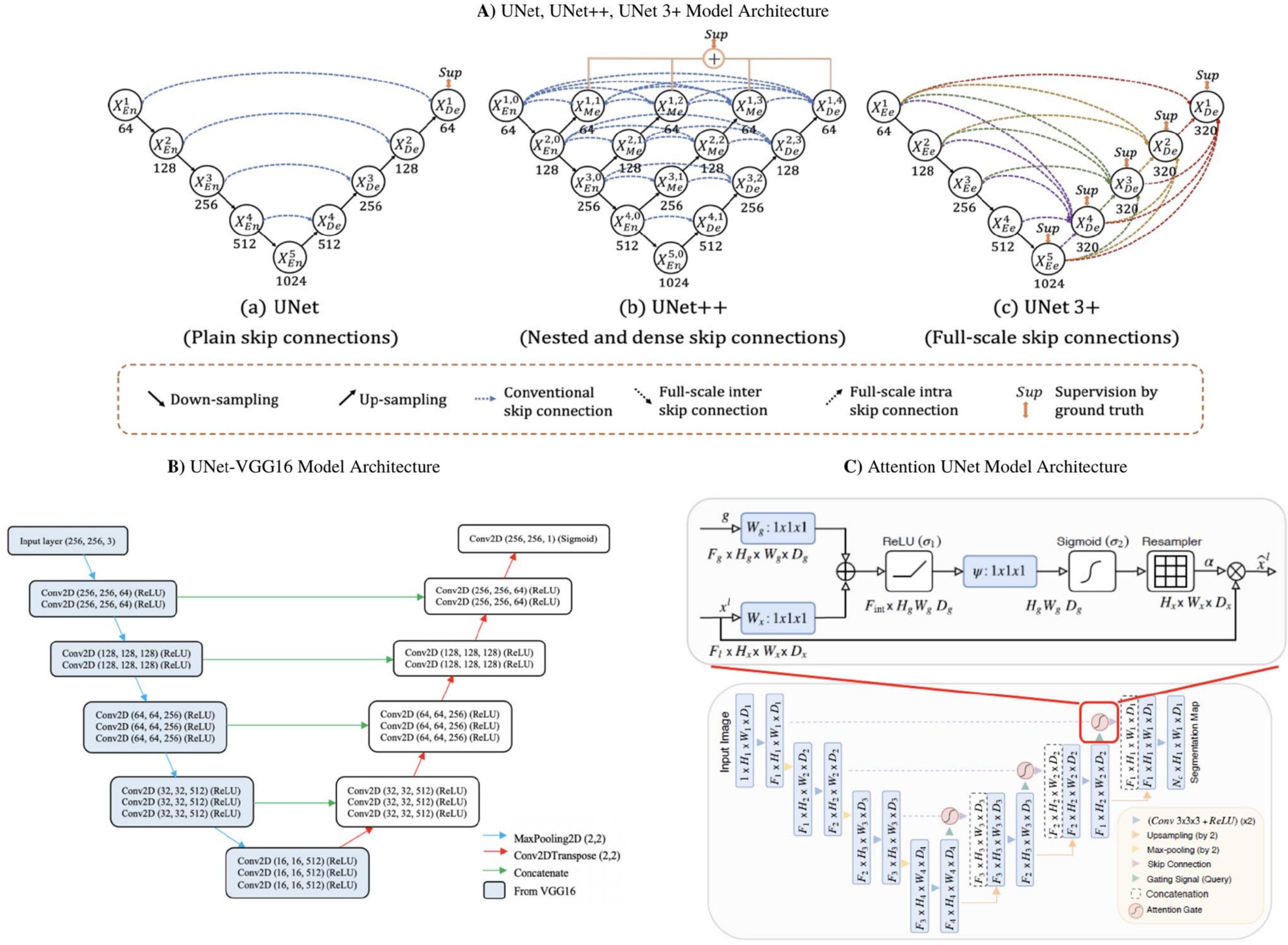

A) UNet, UNet++, UNet 3+ Model Architecture
Sup
X^1_{En} 64
X^2_{En} 128
X^3_{En} 256
X^4_{En} 512
X^5_{En} 1024
X^4_{De} 512
X^3_{De} 256
X^2_{De} 128
X^1_{De} 64
(a) UNet
(Plain skip connections)
$X^{1,0}_{En}$ 64
$X^{2,0}_{En}$ 128
$X^{3,0}_{En}$ 256
$X^{4,0}_{En}$ 512
$X^{5,0}_{En}$ 1024
$X^{1,1}_{Me}$ 64
$X^{1,2}_{Me}$ 64
$X^{1,3}_{Me}$ 64
$X^{2,1}_{Me}$ 128
$X^{2,2}_{Me}$ 128
$X^{3,1}_{Me}$ 256
$X^{4,1}_{De}$ 512
$X^{3,2}_{De}$ 256
$X^{2,3}_{De}$ 128
$X^{1,4}_{De}$ 64
(b) UNet++
(Nested and dense skip connections)
X^1_{Ee} 64
X^2_{Ee} 128
X^3_{Ee} 256
X^4_{Ee} 512
X^5_{Ee} 1024
X^4_{De} 320
X^3_{De} 320
X^2_{De} 320
X^1_{De} 320
(c) UNet 3+
(Full-scale skip connections)
Down-sampling
Up-sampling
Conventional skip connection
Full-scale inter skip connection
Full-scale intra skip connection
Sup Supervision by ground truth
B) UNet-VGG16 Model Architecture
Input layer (256, 256, 3)
Conv2D (256, 256, 64) (ReLU)
Conv2D (256, 256, 64) (ReLU)
Conv2D (128, 128, 128) (ReLU)
Conv2D (128, 128, 128) (ReLU)
Conv2D (64, 64, 256) (ReLU)
Conv2D (64, 64, 256) (ReLU)
Conv2D (64, 64, 256) (ReLU)
Conv2D (32, 32, 512) (ReLU)
Conv2D (32, 32, 512) (ReLU)
Conv2D (32, 32, 512) (ReLU)
Conv2D (16, 16, 512) (ReLU)
Conv2D (16, 16, 512) (ReLU)
Conv2D (16, 16, 512) (ReLU)
Conv2D (256, 256, 1) (Sigmoid)
MaxPooling2D (2,2)
Conv2DTranspose (2,2)
Concatenate
From VGG16
C) Attention UNet Model Architecture
g
W_g: 1x1x1
F_g x H_gx W_gx D_g
x^l
W_x: 1x1x1
F_l x H_xx W_xx D_x
ReLU (σ_1)
F_{int} x $H_g W_g D_g$
ψ: 1x1x1
$H_g W_g D_g$
Sigmoid (σ_2)
Resampler
α
H_xx W_xx D_x
$\hat{x}^l$
Input Image
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_4 \times W_4 \times D_4$
$F_4 \times H_4 \times W_4 \times D_4$
$F_3 \times H_3 \times W_3 \times D_3$
$F_2 \times H_3 \times W_3 \times D_3$
$F_2 \times H_2 \times W_2 \times D_2$
$F_1 \times H_2 \times W_2 \times D_2$
$F_1 \times H_1 \times W_1 \times D_1$
$N_c \times H_1 \times W_1 \times D_1$
Segmentation Map
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Max-pooling (by 2)
Skip Connection
Gating Signal (Query)
Concatenation
Attention Gate

Fig. 2 **a** The model architecture outlining the workflow of UNet, UNet++, and UNet3+. The notable difference between the three models is the skip connections. UNet is using plain skip connections, UNet++ has nested and dense skip connections which have the downside of not being able to explore a sufficient amount of information from full scales. UNet3+, however, uses full-scale skip connections so more information can be obtained during upsampling. **b** A VGG16-UNet is comprised of an encoder that is based on a VGG16 model and a decoder that is based on a UNet model. **c** Attention UNet uses attention mechanisms, compared to a standard UNet model, by focusing on the varying size and shape of target structures

specificity, positive predictive value (PPV), and false predictive value (FPV), and we can also generate a receiver operating characteristic (ROC) curve. The proposed diagnostic metrics could be made using the principles of the confusion matrix [38], as shown in Fig. 5.

The AI model would be considered more accurate and reliable if the AI predictions make the radiologists believe that the segmentation is done by a real human being in terms of preciseness, picking of the area of interest, and if any important considerations are needed in a scenario of an edge case [39, 40].

This is a technique that has never been implemented before and thus is highly novel. The Turing test modification can provide verifiable diagnostic precision and statistical effect–size resilience in the evaluation of AI for computer-based and robotic healthcare and clinical solutions.

4 Results and discussion

4.1 Segmentation results

In this study, we used multiple metrics to evaluate the performance of the model: dice coefficient (DSC), mean intersection over union (mIoU), recall (RE), precision (PR), specificity (SP), and $F1$-score ($F1$). The expressions of the metrics are described as follows:

$$\mathrm{DSC}(Y, \hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y + \hat{Y}|} \tag{1}$$

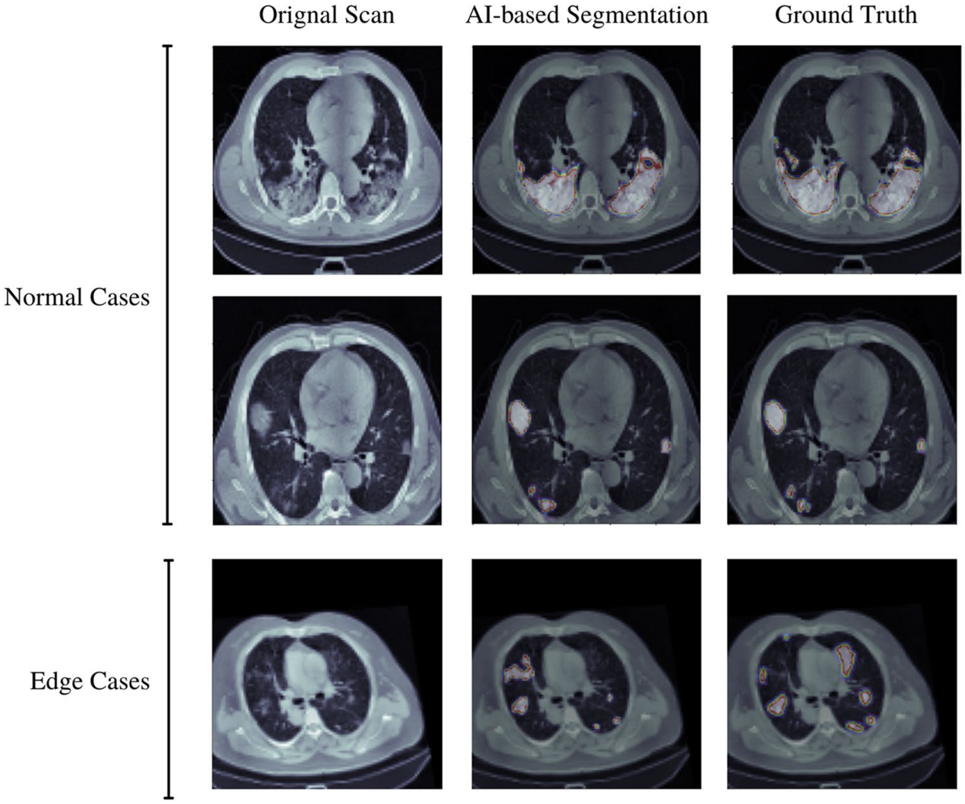


Fig. 3 Semantic segmentation of normal and edge cases' lung infection produced by our proposed UNet model

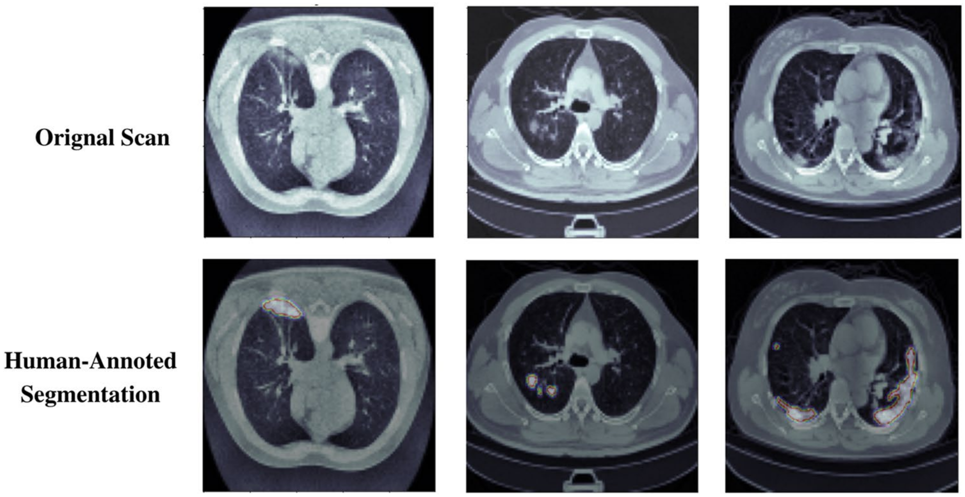


Fig. 4 Human-annotated segmentation produced by expert radiologist

$$\mathrm{mIoU}(Y,\hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} \tag{2}$$

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{5}$$

$$F1 = 2 \times \frac{\mathrm{PR} \times \mathrm{RE}}{\mathrm{PR}+\mathrm{RE}}. \tag{6}$$

Table 1 compares the segmentation results of the UNet, UNet++, UNet3+, Attention UNet, and UNet-VGG16 models in terms of all metrics used in our experiments. Our proposed UNet-VGG16 model achieved the highest

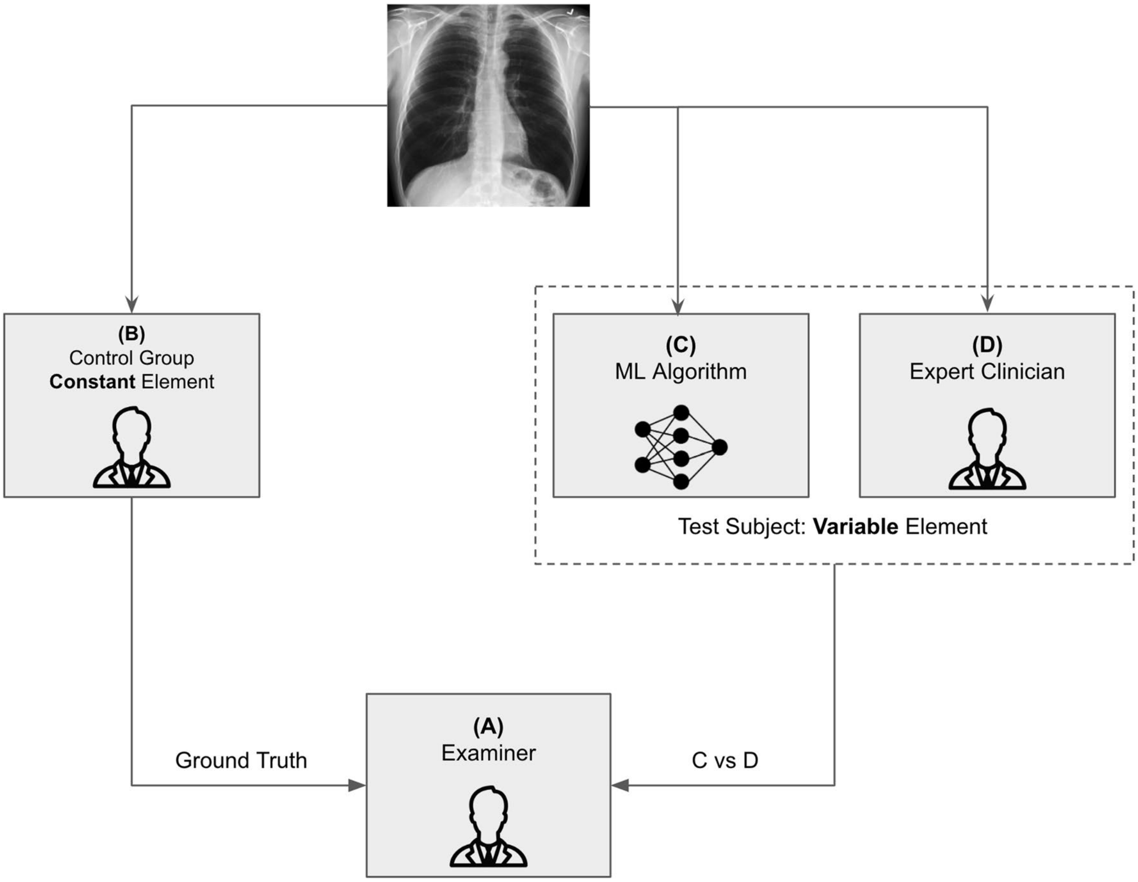


Fig. 5 A systemic layout of the modified Turing test

Table 1 Segmentation metrics results among various UNet models trained

Model	DSC	F1	mIoU	RE	PR	SP
UNet	93.42	94.65	88.36	95.31	93.28	98.71
UNet++	93.56	94.89	87.89	94.7	93.27	97.99
UNet3+	95.89	96.06	88.92	95.4	94.63	98.65
Attention UNet	94.86	95.64	88.45	95.46	94.12	98.19
UNet-VGG16	**96.73**	**97.94**	**89.21**	**96.98**	**95.56**	**99.41**

accuracy among all other models in all the matrices, represented in bold.

In addition, during the testing, the model was examined on various edge cases and cases with complex or rare infections to check whether the UNET is biased or not, but the results are very promising, our model achieved a dice score of 94.76% on these critical cases.

4.2 Modified Turing test results

For this study, 10 board-certified radiologists with more than 10 years of experience each in interpreting cardio-thoracic imaging reviewed 20 sets of medical images and give out their predictions of whether the segmentation is done by a human or AI based on the preciseness and accuracy of the segmentation. All the radiologists were given the same platform and time to give their predictions. The predictions analysis of each radiologist is given in Table 2.

The true positive (TP) denotes that the tester was able to detect the AI-based segmentation and the true negative (TN) denotes that the radiologist was able to detect the human-based segmentation. False negative (FN) represents that the tester thought it was an AI while the segmentation was done by a human, whereas in true negative (TN) the tester thought the segmentation is done by the AI while it was done by a human.

We would consider the TP and FN as our most important metrics here as they reveal the most context about the

Table 2 Analysis of prediction derived from the test results of the participants

Radiologist	TP	FP	FN	TN	Total cases
1	7	7	4	2	20
2	1	6	9	4	20
3	6	7	6	1	20
4	4	5	8	3	20
5	2	5	11	2	20
6	4	7	5	4	20
7	4	5	10	1	20
8	5	6	6	3	20
9	0	7	8	5	20
10	1	3	12	4	20
Total	34	58	79	29	200

performance of the AI algorithms in a clinical setting. TP score reveals the reliability of AI-based segmentation, participants reported when the segmented scan did not include not-so-obvious infections or overdid some of the areas, it made it easier to say for them that the segmentation was done by an AI because a professional radiologist can never do such segmentation [41]. Therefore, having a high TP score is not a good metric for the AI because it means that the segmentation generated is not clinically relevant enough. FN score is what makes an algorithm come closer to an "expert radiologist." If the model earns more TN scores that means that the AI system is as good as a professional radiologist and is very hard to distinguish whether the segmentation is done by a human or a machine.

To furthermore understand the metrics we have calculated evaluation matrices like accuracy, recall, precision, and others to understand the overall performance and behavior of participants as well as the AI model, the results are shown in Table 3.

Overall, our UNet model did exceedingly well in this test, where not only it achieves a high FN score but also received a low TP ratio. This data distribution explains that the model is compatible enough to get implemented in a clinical setting but at the same time there was also a considerable portion of the FP and TN cases, where the participant distinguished between AI and humans, so taking that into the account we would still need a clinician in the loop to safe-gourd patient care and false prediction making by the algorithm. We also incorporated 5 out of 20 to be edge cases and the model, and the AI-based segmentation was picked most of the time as an FP.

Table 3 Evaluation metrics results among all the participants

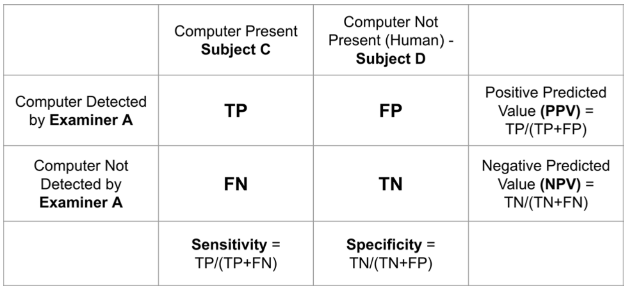
Diagnostic evaluation metrics	Score (%)
Accuracy	31.3
True-positive rate (recall)	30.08
True-negative rate (TNR)	33.4
False-negative rate (FNR)	69.91
False-positive rate (FPR)	66.7
Precision	36.95

Finally, we compared the working and performance of the actual Turing test proposed by Alan Turing and the modified Turing test proposed in this study using bivariate meta-analysis [42], and the results are closely similar to what we expected. In Fig. 6, we have shown how our modified Turing test works exactly like the original test and even the UNet model's performance could analyze using the plot.

5 Conclusion and future work

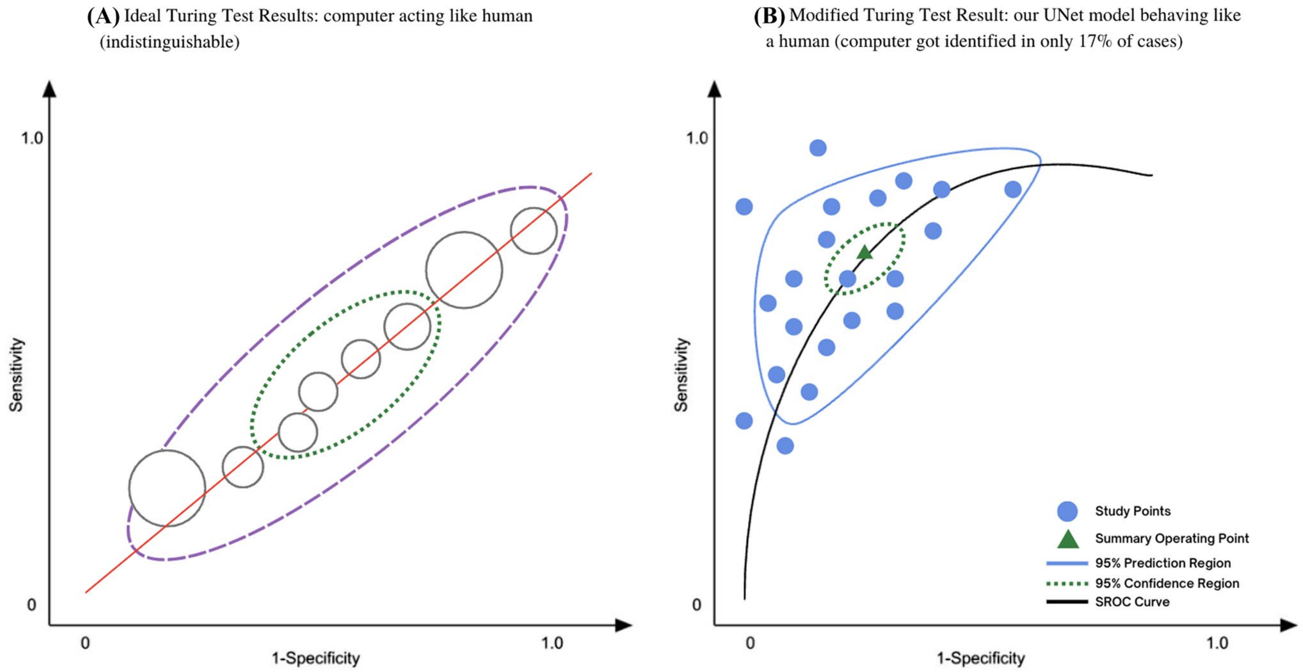
The number of AI-based medical imaging devices is getting increased every single day and it is crucial to think about the potential bias it may inherit. This test would be a transnational standard for upcoming AI modalities. We learned through the study we conducted that the use of our modified Turing test is a notably strong standard to measure the actual performance of the AI model on a variety of edge cases and normal cases and also helps in detecting if the algorithm is biased towards any one type of case. Not just we can detect biases but also classify the type of bias and can work towards resolving it (Fig. 7).

Since artificial intelligence systems in healthcare can be utilized for both diagnosis and treatment of diseases, even a tiny error can result in diagnostic inaccuracy and, as a result, increased morbidity and death rates. As a result, it is critical to conduct a comprehensive verification and validation of each artificial intelligence system prior to using it for diagnosis. Consequently, distinguishing between computers and humans (Turing test or modified Turing test) should not detract from the importance of diagnostic accuracy in disease detection and healthcare provision provided by each computer-based AI system, which should be independently appraised for its healthcare safety, precision, and utility. Therefore, as we proceed towards the upcoming ages of AI



	Computer Present **Subject C**	Computer Not Present (Human) - **Subject D**	
Computer Detected by **Examiner A**	**TP**	**FP**	Positive Predicted Value **(PPV)** = TP/(TP+FP)
Computer Not Detected by **Examiner A**	**FN**	**TN**	Negative Predicted Value **(NPV)** = TN/(TN+FN)
	Sensitivity = TP/(TP+FN)	**Specificity** = TN/(TN+FP)	

Fig. 6 Diagnostic evaluation metrics generated through test results


(A) Ideal Turing Test Results: computer acting like human (indistinguishable)

(B) Modified Turing Test Result: our UNet model behaving like a human (computer got identified in only 17% of cases)

Fig. 7 Bivariate meta-analysis of actual Turing test and our modified Turing test

in medicine, this technique would still be applicable in not only segmentation but also in various other prediction and detection models as well. The modified Turing test provides us trust in the AI algorithm and helps us if not look then predict what is inside the black box of the algorithm.

The future of this subject lies in the application of diagnostic accuracy methods to the modified Turing test, which will spur the development of enhanced technology that can closely replicate human behavior in the process of development. This has the potential to produce healthcare computers and other artificial intelligence-based technologies that can improve human health and quality of life while also igniting the next generation of human–technological conversation.

References

1. Hamet, P., Tremblay, J.: Artificial intelligence in medicine. Metabolism **69**, 36–40 (2017)
2. Ramesh, A., Kambhampati, C., Monson, J.R., Drew, P.: Artificial intelligence in medicine. Ann. R. Coll. Surg. Engl. **86**(5), 334 (2004)
3. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. Nat. Rev. Cancer **18**(8), 500–510 (2018)
4. Tripathi, S.: Artificial intelligence: a brief review. In: Analyzing Future Applications of AI, Sensors, and Robotics in Society, pp. 1–16 (2021)
5. Pinar Saygin, A., Cicekli, I., Akman, V.: Turing test: 50 years later. Mind. Mach. **10**(4), 463–518 (2000)
6. Turing, A.M.: Computing Machinery and Intelligence. Parsing the Turing Test. Springer, Berlin (2009)
7. Turing, A.M.: Mind. Mind **59**(236), 433–460 (1950)
8. Moor, J.H.: An analysis of the Turing test. Philos. Stud. **30**(4), 249–257 (1976)
9. Marcus, G., Rossi, F., Veloso, M.: Beyond the Turing test. AI Magz. **37**(1), 3–4 (2016)
10. Oppy, G., Dowe, D.: The Turing test (2003)
11. Tripathi, S., Augustin, A., Kim, E.: Longitudinal neuroimaging data classification for early detection of Alzheimer's disease using ensemble learning models. https://doi.org/10.36227/techrxiv.19295120.v1 (2022)
12. Tripathi, S.: Early diagnostic prediction of covid-19 using gradient-boosting machine model. arXiv preprint arXiv:2110.09436 (2021)
13. Yu, K.-H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. Nat. Biomed. Eng. **2**(10), 719–731 (2018)
14. Wegner, L., Houben, Y., Ziefle, M., Calero Valdez, A.: Fairness and the need for regulation of AI in medicine, teaching, and recruiting. In: International Conference on Human–Computer Interaction, pp. 277–295. Springer (2021)
15. Dori-Hacohen, S., Montenegro, R., Murai, F., Hale, S.A., Sung, K., Blain, M., Edwards-Johnson, J.: Fairness via AI: bias reduction in medical information. arXiv preprint arXiv:2109.02202 (2021)
16. Park, Y., Jackson, G.P., Foreman, M.A., Gruen, D., Hu, J., Das, A.K.: Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open **3**(3), 326–331 (2020)
17. Tripathi, S., Musiolik, T.H.: Fairness and ethics in artificial intelligence-based medical imaging. In: Ethical Implications of Reshaping Healthcare With Emerging Technologies, pp. 71–85. IGI Global (2022)
18. Szolovits, P.: Artificial Intelligence in Medicine. Routledge, New York (2019)
19. Holzinger, A., Haibe-Kains, B., Jurisica, I.: Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. Eur. J. Nucl. Med. Mol. Imaging **46**(13), 2722–2730 (2019)

20. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
21. Fihn, S., Saria, S., Mendonça, E., et al.: Deploying AI in clinical settings. In: Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril, 145 (2019)
22. Balagurunathan, Y., Mitchell, R., El Naqa, I.: Requirements and reliability of AI in the medical context. Phys. Med. **83**, 72–78 (2021)
23. Patel, V.L., Shortliffe, E.H., Stefanelli, M., Szolovits, P., Berthold, M.R., Bellazzi, R., Abu-Hanna, A.: The coming of age of artificial intelligence in medicine. Artif. Intell. Med. **46**(1), 5–17 (2009)
24. Asan, O., Bayrak, A.E., Choudhury, A., et al.: Artificial intelligence and human trust in healthcare: focus on clinicians. J. Med. Internet Res. **22**(6), 15154 (2020)
25. Pesapane, F., Volonté, C., Codari, M., Sardanelli, F.: Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging **9**(5), 745–753 (2018)
26. Brady, A.P., Neri, E.: Artificial intelligence in radiology-ethical considerations. Diagnostics **10**(4), 231 (2020)
27. Recht, M.P., Dewey, M., Dreyer, K., Langlotz, C., Niessen, W., Prainsack, B., Smith, J.J.: Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur. Radiol. **30**(6), 3576–3584 (2020)
28. Mazurowski, M.A.: Artificial intelligence in radiology: some ethical considerations for radiologists and algorithm developers. Acad. Radiol. **27**(1), 127–129 (2020)
29. Banja, J.: AI hype and radiology: a plea for realism and accuracy. Radiol. Artif. Intell. **2**(4) (2020)
30. Tsai, E.B., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: The rsna international covid-19 open radiology database (ricord). Radiology **299**(1), 204–213 (2021)
31. Tsai, E., Simpson, S., Lungren, M., Hershman, M., Roshkovan, L., Colak, E., Erickson, B., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: data from medical imaging data resource center (midrc)-rsna international covid radiology database (ricord) release 1—chest X-ray, covid+(midrc-ricord-1c). The Cancer Imaging Archive. https://doi.org/10.7937/91ah-v663 (2021)
32. Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., Shucheng, C., Hao, W., Sen, M., Xiaoyu, Y., Ziwei, N., Chen, L., Lu, T., Yuntao, Z., Qiongjie, Z., Guoqiang, D., Jian, H.: COVID-19 CT Lung and Infection Segmentation Dataset. https://doi.org/10.5281/zenodo.3757476
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
34. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018)
35. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059. IEEE (2020)
36. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
37. Pravitasari, A.A., Iriawan, N., Almuhayar, M., Azmi, T., Fithriasari, K., Purnami, S.W., Ferriastuti, W., et al.: Unet-vgg16 with transfer learning for MRI-based brain tumor segmentation. Telkomnika **18**(3), 1310–1318 (2020)
38. Caelen, O.: A Bayesian interpretation of the confusion matrix. Ann. Math. Artif. Intell. **81**(3), 429–450 (2017)
39. Vought, R.T.: Re: Guidance For Regulation of Artificial Intelligence Applications (2020)
40. Rubin, D.L.: Artificial intelligence in imaging: the radiologist's role. J. Am. Coll. Radiol. **16**(9), 1309–1317 (2019)
41. Meuli, R., Hwu, Y., Je, J.H., Margaritondo, G.: Synchrotron radiation in radiology: radiology techniques based on synchrotron sources. Eur. Radiol. **14**(9), 1550–1560 (2004)
42. Van Houwelingen, H.C., Zwinderman, K.H., Stijnen, T.: A bivariate approach to meta-analysis. Stat. Med. **12**(24), 2273–2284 (1993)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.