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

Computers in Biology and Medicine 146 (2022) 105587

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

The effect of machine learning explanations on user trust for automated diagnosis of COVID-19

Kanika Goel [a,*], Renuka Sindhgatta [b], Sumit Kalra [c], Rohan Goel [d], Preeti Mutreja [e]

[a] School of Information Systems, Queensland University of Technology, Australia
[b] IBM Research AI, Bangalore, India
[c] Department of Computer Science, Indian Institute of Technology, Jodhpur, India
[d] COVID-19 Centre at Guru Teg Bahadur (GTB) Hospital, Delhi, India
[e] All India Institute of Medical Sciences (AIIMS), Jodhpur, India

ARTICLE INFO

Keywords:
Medical diagnosis
Machine learning explanations
Application-oriented evaluation
User trust

ABSTRACT

Recent years have seen deep neural networks (DNN) gain widespread acceptance for a range of computer vision tasks that include medical imaging. Motivated by their performance, multiple studies have focused on designing deep convolutional neural network architectures tailored to detect COVID-19 cases from chest computerized tomography (CT) images. However, a fundamental challenge of DNN models is their inability to explain the reasoning for a diagnosis. Explainability is essential for medical diagnosis, where understanding the reason for a decision is as important as the decision itself. A variety of algorithms have been proposed that generate explanations and strive to enhance users' trust in DNN models. Yet, the influence of the generated machine learning explanations on clinicians' trust for complex decision tasks in healthcare has not been understood. This study evaluates the quality of explanations generated for a deep learning model that detects COVID-19 based on CT images and examines the influence of the quality of these explanations on clinicians' trust. First, we collect radiologist-annotated explanations of the CT images for the diagnosis of COVID-19 to create the ground truth. We then compare ground truth explanations with machine learning explanations. Our evaluation shows that the explanations produced.

by different algorithms were often correct (high precision) when compared to the radiologist annotated ground truth but a significant number of explanations were missed (significantly lower recall). We further conduct a controlled experiment to study the influence of machine learning explanations on clinicians' trust for the diagnosis of COVID-19. Our findings show that while the clinicians' trust in automated diagnosis increases with the explanations, their reliance on the diagnosis reduces as clinicians are less likely to rely on algorithms that are not close to human judgement. Clinicians want higher recall of the explanations for a better understanding of an automated diagnosis system.

1. Introduction

There has been increasing interest and success in using deep learning methods for automated diagnosis using medical images. Studies for diagnosis of COVID-19 infections using computerized tomography (CT) images has shown encouraging results with a sensitivity of 98% ($\rho < 0.01$) [1]. At the core of the accurate automated diagnosis of CT images for COVID-19 are the various deep neural network models. For example, the use of a small number of COVID-19 CT images along with a large number of non-COVID- 19 CT images by self-supervised learning of features has been proposed [2]. A deep convolutional neural network (CNN) architecture has been evaluated for detection of COVID-19 via a machine-driven design exploration approach trained on a benchmark data set of 1489 patient cases [3]. The primary objective of these models have been high classification performance according to several standard measures such as F1-score, sensitivity, and specificity.

Supporting an accurate diagnosis of a deep learning method with the underlying reason for the diagnosis is necessary to increase the trust of medical experts in AI-based diagnostic systems [4,5]. Providing explanations or justifications is necessary with many regulations mandating

* Corresponding author.
E-mail addresses: k.goel@qut.edu.au (K. Goel), renuka.sr@ibm.com (R. Sindhgatta), sumitk@iitj.ac.in (S. Kalra), rohanpgoel@gmail.com (R. Goel), dr.preeti.mutreja@gmail.com (P. Mutreja).

https://doi.org/10.1016/j.compbiomed.2022.105587
Received 18 January 2022; Received in revised form 1 May 2022; Accepted 2 May 2022
Available online 8 May 2022

an automated system to justify a decision.[1] This has motivated the development of vari-ous algorithms for generating machine learning (ML) explanations. Gaining insights into the reasons for the diagnosis using ML explanations would also enhance clinicians trust in such AI-enabled diagnosis systems. In this work, we aim to generate ML explanations using existing algorithms and conduct a human-centred evaluation systematically via an application-grounded evaluation [6,7]. "Application-grounded evaluation involves conducting human experiments with a real application" [7]. Hence, conducting an application-grounded evaluation would require using ML explanations on real-world diagnosis of COVID-19 CT images to study how well they influence clinicians' trust in such an automated decision-making task. Conducting an application-grounded evaluation addresses multiple aspects:

A1: Evaluation of ML Explanations: Several state-of-the-art algorithms gen-erate explanations with intuitive visualisations elucidating regions of an im-age influencing the prediction [8–11]. These evaluations on deep-learning models use object recognition data sets such as ImageNet.[2] A quan-titative evaluation against a ground-truth baseline of CT lung images would measure the effectiveness of existing methods on homogeneous grey scale images having subtle variations caused by different lung infections.

A2: Human-centred evaluation of ML Explanations by Clinicians: ML explanations are used to confirm the core requisites of trust and *transparency* [7]. The congruence of ML explanations is measured with respect to human ground-truth baseline. The influence of ML explanations on the trust of clinicians for automated decision support of COVID-19 diagnosis can be studied systematically.

A3: Collection of Labelled Data: Comparing ML explanations with the observations of a domain expert (a radiologist in our case) for high stake decision support requires labelled data. For the purpose of evaluation, CT images labelled by three radiologists are collected and used as human-produced ex-planations.

In this work, we first review existing literature on various algorithms that generate ML explanations. We further examine and evaluate the correctness of ML explanations against expert produced explanations. The explanations vary in their relevance (or proximity) when compared to human experts. We then conduct a structured user study with clinicians aiming to test the influence of explanations on their trust in using such automated diagnosis systems for decision support. Further, we release the human-labelled expla-nation data set for further development towards such studies.

The main contributions of this work are:

- Quantitative evaluation of existing ML explanations methods for auto-mated diagnosis using CT lung images.
- Empirical study of the influence of trust on clinicians based on the ML explanations.
- Provision of a data set of human-annotated explanations of COVID-19 CT images for future study.

The remainder of this paper is organised as follows. We discuss the related work in Sec 2. The details of the proposed approach are presented in Sec 3, followed by a report on evaluating ML explanations generated by different algorithms in Sec 4. Details of the empirical study with clinicians is presented in Sec 5. We conclude the paper and discuss future work in Sec 6.

2. Related work

We now position our work with respect to the related literature, categorized in four main bodies of works.

2.1. ML explanations for image predictions

There has been considerable research in generating explanations of com-plex deep neural networks to understand the reason for predictions post-hoc (i.e. after the model has been trained). In particular, we focus on *local* explanations. A local explanation aims to provide the reasoning for a particular prediction in a locale specific to a given image. Class Activation Mapping (CAM) and its generalization, Gradient-weighted Class Activation Mapping (Grad-CAM), enables visualization of the later layer's class-specific feature maps of a convolution neural network (CNN). Grad-CAM uses the *gradient* information flowing into the last convolutional layer of a CNN to assign important values to neurons for a particular prediction [9]. Grad-CAM is class-discriminative and can be used to produce visual explanations for any CNN-based model.

Additionally, a few techniques assess the relationship between input image and the output by *perturbing* the input image and observing its effect on the output. These approaches are model agnostic and do not need to be aware of the underlying model architecture. One approach is to greedily grey out segments of an image until it is misclassified and visualize the drop in classification score [12]. Similarly, occluding portions of the input image help in revealing parts that are important for classification [13]. Randomized Input Sampling for Explanation (RISE) estimates pixel saliency by randomly generating small masks [10]. Fong et al. [11] introduce *extremal perturbation* which refers to computing the smallest mask that results in highest change to the model prediction.

Post-hoc model-agnostic approaches such as Local Interpretable Model-agnostic Explanations (LIME) [8] *approximate* a complex CNN-based model to a simpler interpretable linear model. LIME draws random samples around the instance to be explained and trains an approximate linear model. The saliency of LIME is based on super-pixels to produce coarse attention maps. All the attribution and approximation methods have been evaluated on image data sets that have large, yet distinct classes. In this work, we evaluate explanations for CT images where the difference between the classes can be subtle. Further, we assess different explanation generation methods that identify salient regions with human annotations.

2.2. ML explanations for medical image predictions

With increased used of machine learning for medical diagnosis, there has been an increased focus on using ML explanations for understanding pre-dictions [14]. Eitel et al. [15] quantitatively compare the robustness of attribution maps generated by different attribution methods on magnetic resonance imaging (MRI) images for Alzheimer's disease classification. In an-other study, two gradient-based explanation methods were used to identify the relevant features used by a CNN model trained for classifying estrogen receptor status from breast MRI. The explanations helped identify the use of irrelevant features that led to changing the pre-processing and training [16]. Attribution-based techniques have been used to generate explanations and detect spurious features learned by the model for melanoma detection [17]. To explain a deep learning model trained to detect COVID-19 from chest X-ray images [18], a new technique, GSInquire, is proposed to produce attributions [19]. GSInquire is evaluated with two new metrics: impact score and impact coverage. Impact score is defined as the percentage of features that strongly impact the model confidence, and impact coverage is defined as the coverage of the identified critical features based on the coverage of adversarial impacted features in the input image. GSInquire outperformed prior methods such as LIME when evaluated for impact coverage and impact score on non-medical data sets. A qualitative evaluation of GSInquire is performed on COVID-19 X-ray images. Zhu et al. [20] propose a Guideline-based additive explanation framework that first determines anatomical features on the basis of an expert guideline. Next, these features are perturbed in the CT images of lung to incorporate medical guidelines and generate understandable explanations. There has been significant progress in exploring ML explanations

[1] https://gdpr.eu/what-is-gdpr/.
[2] http://www.image-net.org/.

for various modalities in the medical domain that include MRI, CT images, skin images, and X-ray images. Wang et al. [21] propose a deep rank-based average pooling network for COVID-19 recognition for more ac-curate and precise COVID-19 diagnosis system. The technique provides an 18-way data augmentation technique to aid model from overfitting. A new Deep RAP Network (DRAPNet) is suggested and Grad-CAM is utilized to prove the explainable heat map that links with COVID-19 lesions. Another technique, Deep Stacked Sparse Autoencoder Analytical Model, is brought to the fore by Wang et al. [22] to diagnose COVID-19 on chest CT images. First features are extracted using two-dimensional fractional Fourier entropy. Next, a classifier is created using a custom deep stacked sparse autoencoder (DSSAE) model. Finally overfitting is resisted using an improved multiple-way data augmentation. These studies focus on identifying relevant features used by the model, thus enabling model designers to improve the reliability of ML models. In this work we compare state-of-the-art explanation genera-tion techniques for COVID-19 detecting using CT images and quantitatively compare their accuracy as compared to human explanations.

2.3. Evaluation of the quality of ML explanations

A recent study by Zhou et al. [6] thoroughly review existing approaches to assess the quality of ML explanations. The authors study existing quan-titative and human-centred evaluations. In the context of medical imaging, a large body of existing work on ML explanations focus on *attributed-based* explanations that rank the importance or influence of input on the output prediction. For such ML explanation methods, various metrics have been proposed that include *recall of important features*: the fraction of these ground-truth features that are recovered by the ML explanations or *sensitivity*: the degree to which the explanation is affected by insignificant perturbations of the input. An explanation with lower sensitivity is preferable. In addition to evaluating the features that influence the ML model, the understandibility of ML explanations to human users is important. Here, the approaches that involve humans for evaluating ML explanations are applicable. Doshi-Velez and Kim [7] introduce three categories of evaluation: i) *Application-grounded evaluation:* involving the evaluation of explanations in the context of a near real application or task. ii) *Human-grounded evaluation:* where evaluation involves humans on simpler tasks and is appropriate when testing general no-tions of quality of explanations. iii) *Functional-grounded evaluation:* where proxy tasks not involving human subjects are used. Similarly, the need to measure the quality of explanations using a *System Causability Scale (SCS)* has been proposed, ".. to quickly determine whether and to what extent an explainable user interface (human–AI interface), an explanation, or an explanation process itself is suitable for the intended purpose" [23].

Application-grounded and human-grounded evaluations have the ML explanations evaluated with humans. In these studies qualitative and quantiative metrics are used to evaluate explanation qualities. Qualitative metrics include rating provided by users via surveys or questionnaires. Quantitative metrics would require measuring performance of humans on tasks when assisted with ML explanations. In our work we use a qualitative approach by presenting a survey to clinicians for COVID-19 diagnosis.

2.4. User-centric evaluation of trust with ML explanations

A common qualitative evaluation approach is to study the effects of ML explanations on user trust via a set of tasks and a questionnaire designed to obtain user responses on the task and the ML explanations [24,25]. "Willingness to accept a computer generated recommendation" is an observable sign of user trust [26]. User trust evaluations measure the experience of users, using a questionnaire or based on observations. Nourani et al. [24] conducted a study with 60 undergraduate and graduate students as participants. The participants were asked to provide feedback on the perceived accuracy of the system after presenting them with varying types of explanations (i.e. no explanations, weak, and strong explanations). Papermeier et al. [25] conducted a user study that dealt with the task of classifying offensive tweets.

The study investigated the effects of model accuracy and fidelity of explanations on user trust. A low fidelity explanation did not provide any useful information about the underlying model. The study tested nine conditions (3 model accuracy levels $\times$ 3 explanation fidelity levels). Both the studies show that model accuracy and explanation fidelity or correctness influences user trust and that providing meaningless explanations harms user trust. In another study, Zhou et al. [27] investigated the influence of ML explanation on user trust. Here, the explanations are presented by referring to training data points that influence predictions. The study found that ML explanation enhanced trust but only for training data points with higher influence on prediction, and for high model performance.

A quantitative approach to measure the usefulness of ML explanations is by measuring users' understanding of the machine generated explanations [28]. The users' understanding is measured by considering a proxy measure such as the time taken by the user to do a certain task based on the explanation or by a subjective rating from the user. A study on enhancing the ability of users to predict the outcome of a machine learning model by providing explanations using saliency maps has been investigated [29]. The study indicates that the use of instance-level saliency maps does not enable users to predict the outcome of the model. There have been approaches that quantify the response time and accuracy in decision making tasks when presented with ML explanations as an indicator of the quality of ML explanations [30] and to measure user trust [31]. Additionally, recent study measuring physiological signals such as Galvanic Skin Response (GSR) and Blood Volume Pulse (BVP) showed significant differences when users were presented the ML explanations. Hence, physiological responses can be used as indicators of user trust and for assessment of the quality of ML explanations [32].

Existing work has extensively focused on non-medical domains and few of these studies have been carried out on simpler tasks that do not require domain expertise. We present an application-grounded evaluation by evaluating ML explanations for COVID-19 diagnosis, compare them with ground truth generated by radiologist, and conduct our user study on clinicians.

3. Approach

This section describes the approach used to i) evaluate ML explanations generated by different post-hoc methods for COVID-19 diagnosis, and ii) perform an application-grounded evaluation of user trust via a study design with clinicians as participants. As presented in Fig. 1, our approach starts with using a pre-trained deep learning model with high classification performance to diagnose COVID-19 using CT-images. ML Explanations for the diagnosis are generated using different post-hoc methods. Human explanations (or ground truth explanations) are collected from radiologists for a subset of 65 CT images. The human explanations are used to categorise ML explanations into *strong* and *weak* categories. ML explanations that are aligned or congruent with human explanations are considered strong. Finally, we perform a user study by presenting the diagnosis and different categories of ML explanations to evaluate clinicians' trust on the AI model.

3.1. CNN-based COVID-19 diagnosis model

Our first step is to diagnose COVID-19 infections using CT images automatically. We use a pre-trained deep convolutional neural network model for detection of COVID-19 cases [3] (COVIDNet-CT). As our study's objective is to evaluate model explanations and its influence on clinicians' trust, we do not focus on designing a deep learning model, but use a pre-trained model. However, we require a highly accurate model to evaluate the ma-chine explanations. Hence, we use a pre-trained

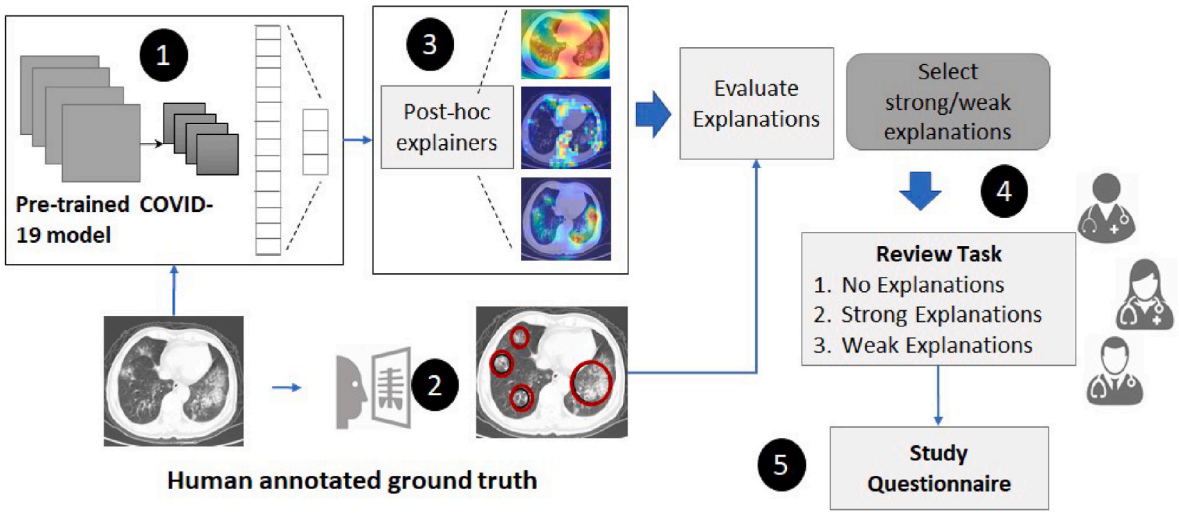

Fig. 1. An overview of the approach to evaluate visual explanations and clinician's trust.

model, COVIDNet-CT with the best model performance. The pre-trained CNN-based model architecture is designed via machine-driven design exploration. In a machine-driven design exploration strategy, the problem of identifying a tailored deep neural network architecture is formulated as a constrained optimization problem where the objective is to maximize an accuracy based score with constraints of having a recall of $\geq$ 95% and a COVID-19 precision of $\geq$ 95%. The network architecture produced consists of a heterogeneous composition of spatial convolution layers, pointwise convolutional layers, and depthwise convolution layers. As discussed by the authors [3], the architechture (see Fig. 2) uses the projection-replication-projection-expansion design pattern (denoted as PRPE and PRPE-S for unstrided and strides patterns, respectively). Here, a projection to a lower channel dimensionality is done via point wise convolutions, followed by a replication of projections. Depth wise convolutions are further followed by a point wise convolutions. The architecture also contains long-range connectivity enabling better representational capabilities than a densely-connected deep neural network architecture. This design pattern is explored and automatically discovered using the machine-driven design. The model is first trained on ImageNet data set and then trained on the data of 1489 patient cases.

The data used for training, validation, and test comprises 104,009 im-ages of 1489 patients collected by the China National Center for Bioinformation [33]. The data consists of CT images of three different infection types:

3.2. Human annotation baseline

We capture human annotation of salient features to create the ground truth for the evaluation of machine-generated model explanations. We select CT images of 65 different patients covering COVID-19 (35 CT images) and common pneumonia (30 CT images) classes. The data collection approach is detailed as follows:

- CT-images of 30 distinct patients with coronavirus pneumonia that were correctly predicted by the model. These were *true positives* and the prediction probability of the model was $\geq$ 0.98.
- CT-images of 30 distinct patients with normal pneumonia that were correctly predicted by the model. These were the *true negatives* and the prediction probability of the model was $\geq$ 0.98.
- CT-images of 5 patients with coronavirus pneumonia were chosen where the model predicted as normal pneumonia resulting in *false negatives.* The prediction probability of the model was $\geq$ 0.83.

The model has a single patient classified as a *false positive* and this was due to the bad resolution of the CT images. Hence, we did not consider a false positive for our annotations. Each CT-image was annotated by a minimum of two radiologists independently. A random set of 40 CT scans were annotated by 3 radiologists. Multiple annotations enables us to quantify the alignment of human explanations between experts in the domain. We further make the data set available for future research.[3]

The human annotation of CT images indicative of COVID-19 infection contains peripheral, bilateral ground-glass opacities as indicated in Fig. 3. These opacities may be visible with or without crazy-paving lines. The ground glass opacities have a rounded morphology (or shape) during the initial course of the disease but involves a larger part of the lung as the disease advances. Pneumonia images, on the other hand, have septal thickening, bronchovascular bundle thickening, interstitial nodules, and honeycombing. These differences are subtle when compared to other data sets that are used to evaluate and generate ML explanations for object detection.

3.3. ML explanations for COVID-19 diagnosis

We use current state-of-the-art attribution methods to generate local explanations for the CT images. Local explanations are specific to that particular image:

- **Grad-CAM** [9]: It is a class-discriminative attribution method that uses the gradients of any target class flowing into the final convolutional layer. The gradients are global-average-pooled to produce a coarse heat-map highlighting important regions in the image for predicting a class (COVID-19 and Common pneumonia in our case).
- **Occlusion** [13]: In occlusion, the attribution is computed by the change in the output prediction confidence when some part of the input image is "occluded" (i.e. set to zero or a constant value). Here, we occlude 2 $\times$ 2 pixel window and generate heat-maps based on their influence on the prediction probability. It is a perturbation-based approach for generating explanations.
- **RISE** [10]: It is a model-agnostic approach as the importance heat map is obtained with access to only the input and output of the deep learning model. The importance of pixels is determined by masking them in random combinations. Bilinear upsampling is used to avoid adversarial effects due to sharp edges of masks.

[3] https://sites.google.com/view/explainable-ai-user-trust/.

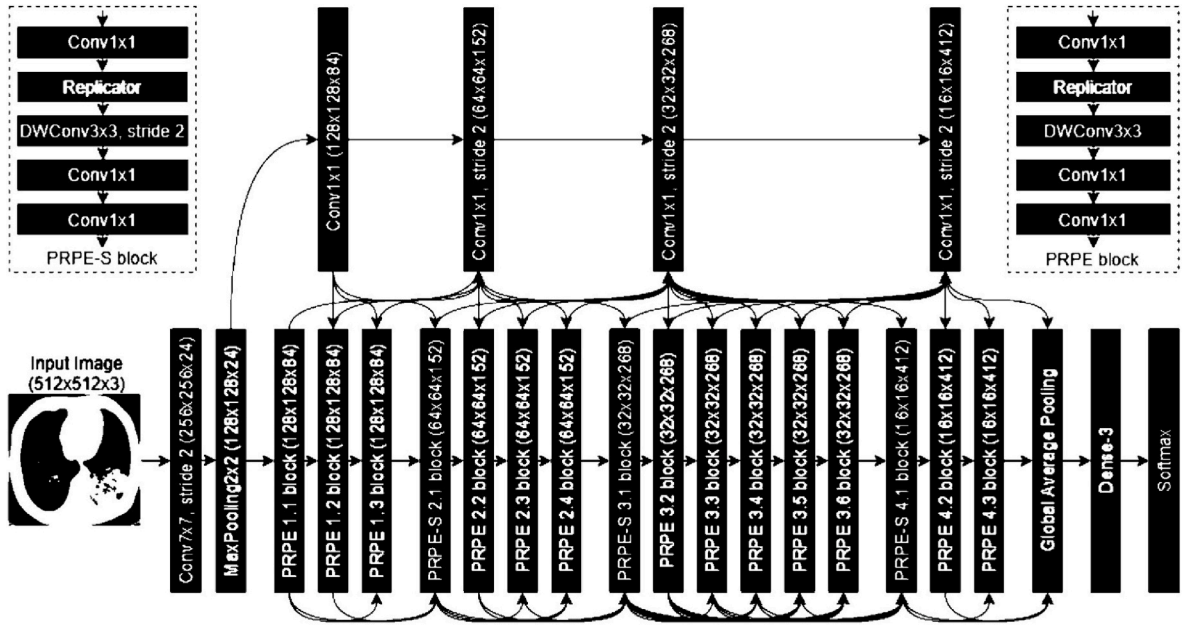

Fig. 2. The COVID-19 Pre-trained model architecture using machine-driven design exploration [3].*novel coronavirus pneumonia*, *common pneumonia*, and *normal controls*. The trained model achieves an overall accuracy of 99.1% with COVID-19 sensitivity (or recall) of 97.3% on the test data (20% of the data) which comprises of 120 COVID-19, 120 pneumonia, and 50 normal patient cases. The specificity of the model for COVID-19 is 99.9% and the positive predictive value (or precision) is 99.7%.

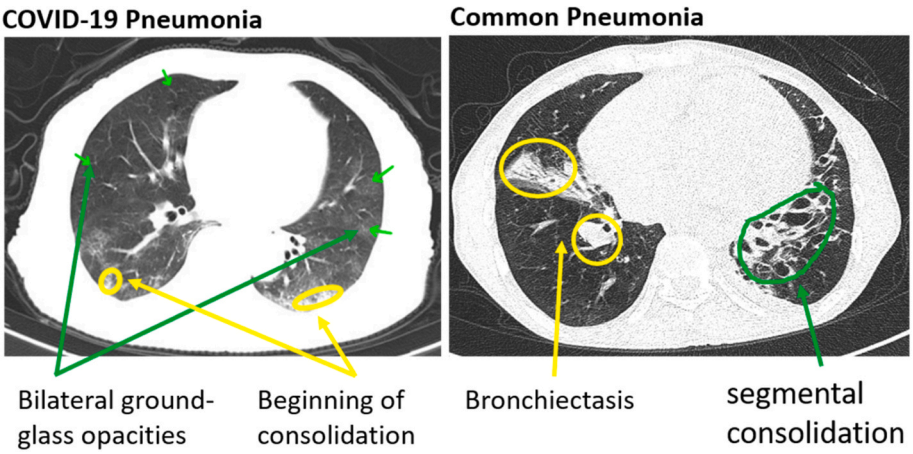

Fig. 3. Human annotated CT-images for COVID-19 pneumonia and common pneumo-nia.

- **LIME** [8]: It is another model agnostic approach that first creates samples by perturbation of super-pixels of an image. The data set

created using the perturbed samples is then used to train a linear model that is locally similar to the black-box model. The samples are weighted based on their distance from the image. The linear model provides explanations of super-pixels that influence the prediction.

We choose methods that are model-agnostic (RISE, LIME, Occlusion), and Grad-CAM, which is suitable for any CNN-based model. All the ML explana-tions are generated for the same underlying pre-trained COVID-19 diagnostic model.

3.4. Evaluating trust with a user study

As trust is an experience of each clinician, it is evaluated with a user study. Ideally, the trust in the system would be high if the human and machine reasoning are congruent [23]. However, given that existing research on generating ML explanations that provide insights into the reasoning of the model is still evolving, we set up an online study to assess the trust for automated COVID-19 diagnosis using explanations generated by the current state-of-the-art techniques. User trust is evaluated via a questionnaire representative of questions brought forth by Körber [34]. Three underlying dimensions that influence the trust in automation are postulated by Körber: Reliability/Competence, Understandability/Predictability, and Intention of Developers. In our work, we focus on the factors that are influenced by ML explanations: Trust, Understandability, and Reliability. In total, our questionnaire consists of eight questions related to understandability, reliability, and the user trust, as shown in Table 1.

The user survey consists of three blocks. In the first block, the user is presented with CT images with the prediction probability followed by the set of eight questions. Here, we would like to observe the trust in a system when the automated system provides minimal level of information. The second block has the user presented with CT images with ML explanations generated by one of the techniques mentioned in Section 3.3. ML Explanations.

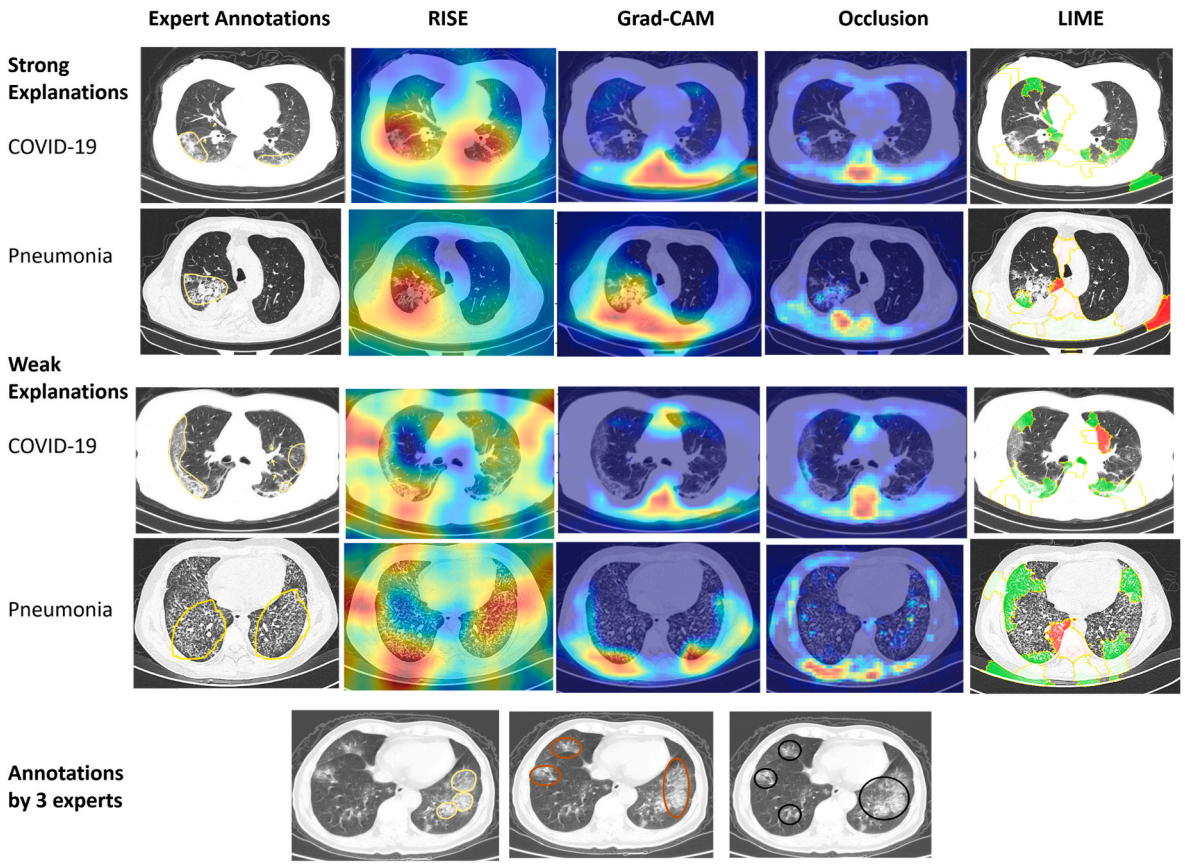

Fig. 4. Human annotations and explanations generated by RISE [10], Grad-CAM [9], Occlusion [13], and LIME [8]. A strong explanation has at least one technique with high precision and recall. Human annotations have different salient regions highlighted for a COVID-19 CT image. The salient regions responsible for the prediction are represented by red color when using RISE, Grad-CAM, and Occlusion. LIME presents pixels supporting the prediction with green color and pixels negating the prediction with red color.annotations was measured by the *pointing game* as used in the previous studies [10,11,35]. For the evaluation, the highest (or maximum) saliency points of each explanation is extracted. In Fig. 4, this would mean the dark red attributed regions of the explanations provided by RISE, Grad-CAM and Occlusion and green regions for LIME. If the point lies within the human-annotated region, it is counted as a hit. Otherwise, the point is considered as a miss. For example, if we consider the first explanation of RISE for COVID-19 CT image in Fig. 4, there are two high saliency points that lie in the region indicated by a human annotation. Hence, it has 2 hits and 0 miss. However, for LIME, it would result in 2 hits and 5 miss The pointing precision is then computed as $^{\#hit}$. In addition, we also compute the recall of the explanations with respect to the ground truth annotations as $^{\#hit}$. Each ML explanation is compared to at least $\#ground-truth$ two human annotations. Further, the pointing game precision and recall is used to compare alignment between human explanations for COVID-19 and pneumonia CT images. Hence, one expert is considered as ground-truth to compare other two human experts. The assumption is that humans could focus on different regions of the CT images to make a decision.

that closely match human judgement (strong explanations) are presented to the user. The questionnaire follows it. In the third block, CT images that reasonably match human judgement are presented, followed by the same set of eight questions. The ML explanations are chosen based on the precision and recall measures that we discuss in detail in Section 4. Our objective is to evaluate the variance in trust as the user navigates from *no explanations*, *strong explanations* to *weak explanations*. As medical diagnostics is a high stake domain, we assume that a deployed system would be used if it provides explanations that have some congruence with human explanations. Hence, we do not present explanations to users that have zero match to human judgement. Our study, therefore, reflects an optimistic scenario of ML ex-planations having reasonable match with human explanations. In the study, each user responds to 24 questions. Besides, we capture qualitative feedback from the user on the explanations provided to the user.

4. Explanation evaluation

We evaluated the ML explanations generated by different methods to measure their proximity to human annotations. Their proximity to human.

For evaluation, we ran RISE over 3 runs and took the best measure. For Occlusion, we used a mask size of (4×4), as running a smaller size mask was computationally expensive. We ran LIME for 5000 perturbations. The results for the 65 CT images are presented in Table 2. Fig. 4 illustrates four examples of explanations generated by the different methods and human annotations. The first two explanations are labelled as strong explanations as there is high alignment of at least one machine explanation with the human annotation (RISE). The two weak explanation have low alignment with the human annotations. Fig. 4 further illustrates one CT-image where the three radiologists have marked different regions.

The results indicate that RISE outperforms other methods with an aver-age precision of 0.73 and recall of 0.50 across all explanations in the context of the COVID-19 CT images that we used for the study. The other methods have significantly lower precision and recall. Occlusion, for example, is susceptible to identifying irrelevant attributions as masking triggers adversarial effects demonstrated by previous studies where specific inputs can generate unexpected outputs [36]. Hence, the results of Occlusion lead to low precision and lower recall. We observe lower prediction and recall for Grad-CAM too. This is inline with the recent study that indicates that Grad-CAM is not a reliable explanation method and can highlight locations that the model does not use [37]. Hence, Grad-CAM sometimes produces misleading explanations that

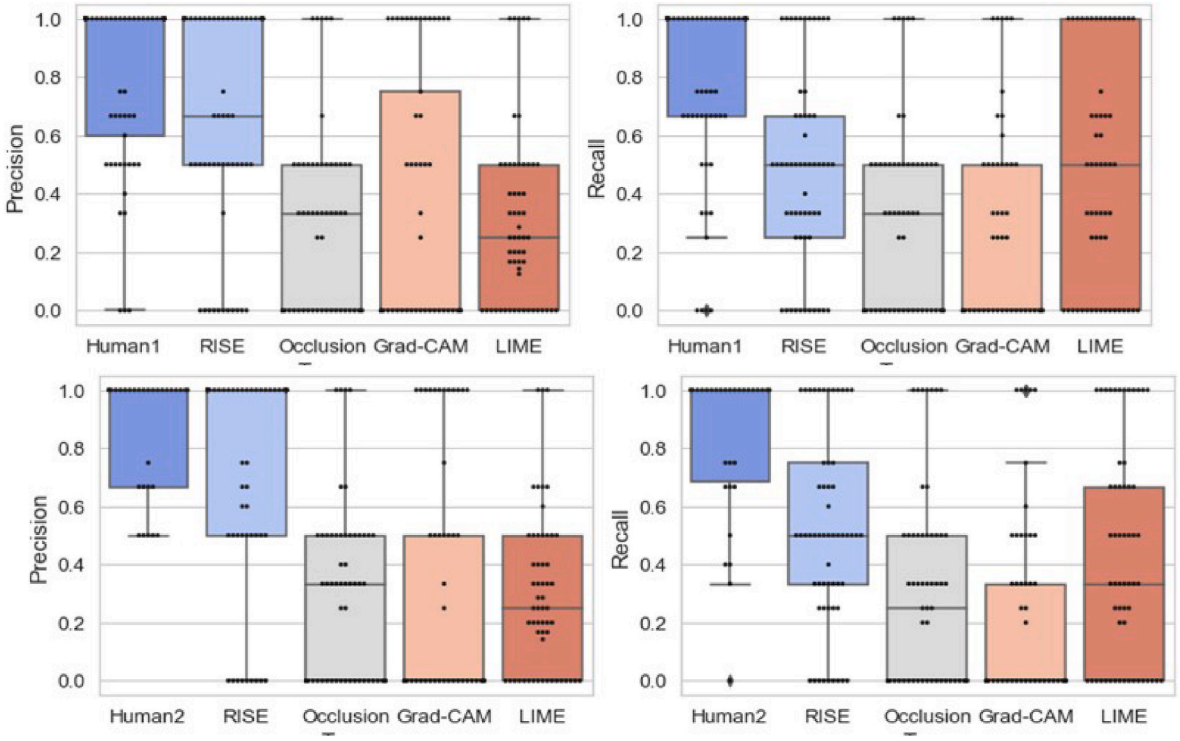

Fig. 5. Precision and Recall of pointing game.

Table 1
Questions and dimensions for the user study.

Question	Dimension
The system is capable of diagnosing correctly	Reliability
The system is reliable	Reliability
The system can make errors[a]	Reliability
The system is capable of making complicated diagnosis	Reliability
I was able to understanding why things happened	Understanding
The system makes unpredictable decisions[a]	Understanding
I can rely on the system	Trust in the system
I trust the system	Trust in the system

[a] = inverse item.

Table 2
Pointing game evaluation measure with two ground-truth (GT) human annotations and recall across the different methods (Fig. 5).

GT Human3	Prec.	Recall	GT Human1	Prec.	Recall
Human1	0.81	0.77	Human 2	0.86	0.83
RISE [10]	0.68	0.48	RISE	0.78	0.52
Grad-CAM [9]	0.33	0.28	Grad-CAM	0.31	0.30
Occlusion [13]	0.29	0.31	Occlusion	0.30	0.38
LIME [8]	0.30	0.47	LIME	0.30	0.48

highlight irrelevant locations. LIME uses a local approximation model that can lead to incorrect explanations [38].

It is further observed that human judgement varies with an average pre-cision ranging between 0.81 and 0.86, and a recall ranging between 0.77 and 0.83. The box plots present further details of the distribution of precision.

5. User evaluation

We conducted a user study to assess the influence of explanations on the trust of clinicians and gain an assessment on how reliable and understandable the explanations are for automated COVID-19 diagnosis. Therefore, the questions that the user study intended to answer were:

- Are the explanations perceived reliable by clinicians for automated COVID-19 diagnosis?
- Are the explanations understandable by clinicians for automated COVID- 19 diagnosis?
- Do the explanations induce trust in clinicians for automated COVID-19 diagnosis?

5.1. Design

To answer the above mentioned questions, a user survey with a within subject design was prepared. The survey consisted of four main blocks. Block 1 focused on obtaining demographic information related to the user such as years of practice and experience with assessing lung CT scans. The next three blocks focused on providing three different types of explanations as explained in Section 3.4.

Block 2 consisted of five images along with the prediction and the prediction probability of the model. Of the five images, three were true positive (i.e., the patient had COVID-19 and the model predicted COVID-19) with high prediction probability (> 98%), one was false negative (i.e., the patient had COVID-19 but the model predicted Pneumonia) with high prediction probability (= 0.98), and one true negative (i.e., the patient had Pneumonia and the model predicted Pneumonia) with high prediction probability.

(= 0.98). We presented all images and their prediction probability to understand if it influences the trust of users as well as reliability and understanding of the model, which were evaluated using the eight questions mentioned in Table 1.

In block 3 and 4, we presented five images with strong and weak explanations, respectively. Since RISE outperformed other methods, explanations generated by RISE were used in the survey. We displayed the original CT image and the explanation alongside it and mentioned the prediction along with its prediction probability. The pointing game metrics of precision and recall were used to choose the explanations.

Strong explanations had a precision of ($\geq$ 0.75) and a recall of ($\geq$ 0.75) when compared to at least one of the human annotation. A *weak* explanation had precision in the interval [0.5,0.75) and a recall $\geq$ 0.5 when compared to a human annotation. The pointing game metrics was not presented to the user. Some examples of strong and weak explanations are highlighted in Fig. 4. Block 3 had three COVID-19 images that were true positive and two Pneumonia images that were true negative. All these predictions had a high prediction probability ($\geq$ 0.98). Block 4 had three COVID-19 images that were true positive, one Pneumonia image that was true negative, and one COVID-19 image falsely predicted as Pneumonia. The false negative prediction had a prediction probability of 0.83, while the others had high prediction probability ($\geq$ 0.98). These blocks aimed at assessing the influence of strong and weak explanations on reliability and understanding of automated systems and subsequently trust on such a system.

The ordering of strong and weak explanations can affect user trust. We rely on the findings by Nourani et al. [39] which indicate that user trust is influenced by first impressions. Their study finds that errors early-on can cause negative first impressions for domain experts, negatively influencing their trust over the course of interactions. However, encountering correct information early would enable them to adjust their trust as they encounter errors. Based on their findings, we first present strong explanations in block 3 followed by weak explanations in block 4.

The survey ended with two questions that aimed at obtaining qualitative feedback on the usefulness of explanations.

5.2. Participants

The survey was sent via email to 50 clinicians experienced in reading CT images, using their publicly available email addresses. Convenience sampling was used to identify clinicians for this study. In the email the details related to study were explained. We asked the clinicians to respond to the email if they believed that they can contribute to the study. We received 30 responses, which were analysed for this study. The responses obtained are made available for future study.[4] All participants had experience in reading and analysing CT images. Fig. 6 provides an overview of the years of experience participants had in analysing CT images. The demographic summary conveys that all participants were well-equipped to answer the questions raised in this study.

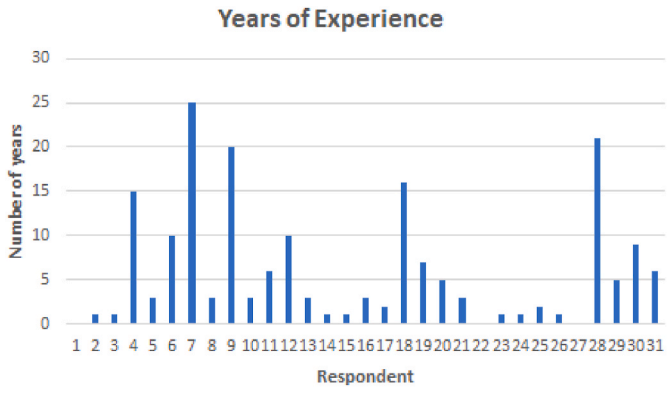

Fig. 6. Years of experience of participants.

[4] https://sites.google.com/view/explainable-ai-user-trust/home.

5.3. Analysis

We measure reliability, understanding, and trust in automation quantitatively using questions listed in Table 1, which were rated on a 5-point likert scale. We computed the average values of responses related to each dimension across the three blocks, denoted as *No explanation, Strong explanation, and Weak explanation* in the results subsequently. Non-parametric Friedman test [40] was used to analyse the results. Friedman test is an alternative to one-way ANOVA with repeated measures and is used to test the differences between groups when the dependent variable is quantitative or ordinal. It is also used for continuous data that violates assumptions necessary to run one-way ANOVA repeated measures. In this study, Friedman test was used to detect the differences across the three blocks.

Table 3 presents the results of responses to questions related to reliability and Fig. 7 provides an overview of the results on reliability for each block in the survey. The results are found to be statistically insignificant ($p = 0.415$). However, the mean ranks show that the reliability of the model reduced from when no explanations were provided blueto strong explanations and then weak explanations. This is also evident in Fig. 7 where we observe that the minimum and maximum reliability score is higher for string explanations than weak explanations. The results demonstrate that clinicians found the AI model reliable when no explanations were provided. The explanations reduced the perceived reliability of the model. The findings can be attributed to clinicians not finding the explanations specific as stated by them in their qualitative feedback. While the pointing game measure leads.

to high precision, the annotations by humans are more specific. Humans point to specific areas and in contrast, the heat-maps generated by attribution models highlight broader regions. The results convey that metrics of high precision and recall with pointing game does not lead to high reliability of the model in the medical domain. The observation is further reinforced through the feedback from participants, "*multiple regions are not highlighted properly*" and "*[need] more precise color coding.*"

Table 4 demonstrates the results related to questions on understanding and Fig. 8 provides an overview of the responses related to questions on understanding in each block. The results are statistically significant ($p = 0.004$), suggesting that the perceived understanding of the clinicians varied across.

the three blocks. The mean ranks convey that the understanding of the prediction reduced from Section 1 to 3. This is also reinstated from Fig. 8, which shows that the median and quartiles drop from when no explanations to provided to when weak explanations are presented. Explanations should.

have contributed to the understanding of the clinicians; however, this is not true. Furthermore, to investigate the difference among various blocks, we conducted Wilcoxon post-hoc test [41]. The results are present in Table 5. It is inferred that the difference is statistically significant for no explanations and weak explanations ($p = 0.002$) and strong explanations and weak explanations ($p = 0.017$). The results clearly demonstrate that the understanding is impacted the most when weak explanations are provided than when strong or no explanations are provided.

The results reinforce the fact that explanations require high precision and high recall to contribute to the understanding of the clinician. Missing certain regions or having spurious regions in explanations can be detrimental to the understanding of clinicians. This is reinstated by a

Table 3
Results for questions on reliability.

Block	Mean Rank	N	25th	50th	75th
No explanation	2.15	30	2.92	3.67	4.33
Strong explanation	2.00	30	2.92	3.33	4.00
Weak explanation	1.85	30	2.67	3.33	3.75

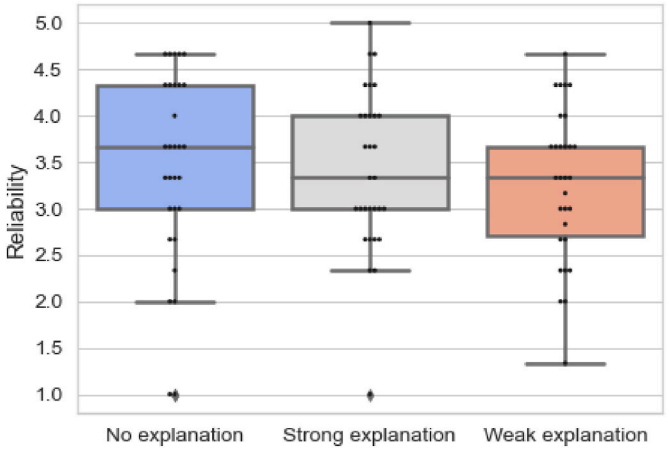

Fig. 7. Overview of responses for questions on reliability.

Table 4
Results for questions on understanding.

Block	Mean Rank	N	25th	50th	75th
No explanation	2.25	30	3.50	4.00	4.50
Strong explanation	2.17	30	3.50	4.00	4.50
Weak explanation	1.58	30	3.00	3.25	4.00

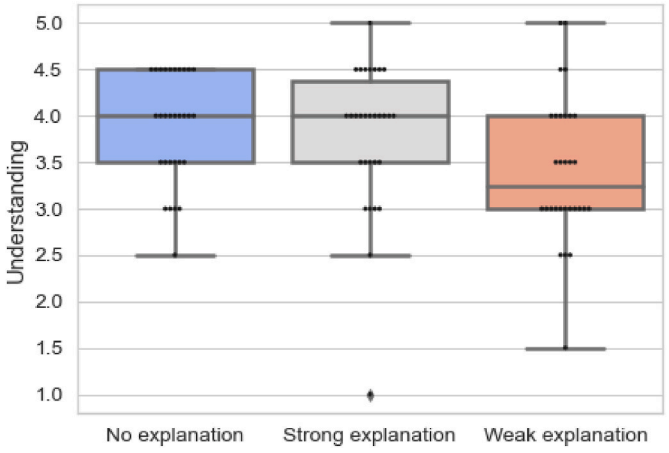

Fig. 8. Overview of responses for questions on understanding.

Table 5
Wilcoxon Post-hoc test for understanding.

Blocks Compared	p value
No explanation and Strong explanation	0.339
No explanation and Weak explanation	0.002
Strong explanation and Weak explanation	0.017

comment in a survey, *"all regions need to be highlighted for diagnosis"*, which communicates that high recall is crucial for the understanding of the diagnosis. In addition, the expectation is to highlight *"patterns of opacities."*

Next, we analysed the results presented in Table 6 which demonstrates the trust in the system. Fig. 9 presents the variation in trust from when no explanations are provided to strong and weak explanations. The results are statistically insignificant ($p = 0.240$). However, the mean rank reveals that explanations, in fact strong explanations (2.12), contribute to perceived trust of the clinician. The results convey that the

Table 6
Results for questions on trust.

Block	Mean Rank	N	25th	50th	75th
No explanation	2.10	30	3.00	4.00	4.50
Strong explanation	2.12	30	3.00	4.00	4.50
Weak explanation	1.78	30	3.00	3.50	4.50

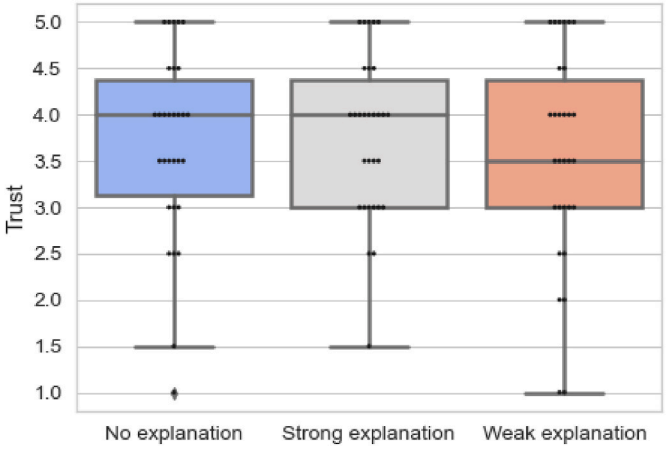

Fig. 9. Overview of responses for questions on trust.

trust increased from no explanations (2.10), to when strong explanations were provided (2.10), and then decreased again when weak explanations were presented (1.78). This is also reinforced from Fig. 9. The findings communicate that while explanations do not contribute to reliability and understanding of the model, they do contribute to the perceived trust. This finding was fortified by comments made by participants, such as "*it assists in diagnosis and makes the things easier*." Moreover, they also show that weak explanations are not useful and do not contribute to enhancing trust of clinicians with comments such as "*the system makes mistakes, in some cases made incorrect assessment* ". This can be related to the fact that medical domain requires completeness and precision to provide appropriate care to patients [42].

Although the results on trust were statistically insignificant ($p = 0.240$), a higher result for strong explanations was evidenced, which could be examined in future experiments. Furthermore, for explanations to enhance understanding and be perceived as reliable, the medical domain demands a high recall in addition to precise salient maps.

Analysis of Qualitative feedback: The survey also consisted of two optional questions to obtain qualitative feedback from clinicians: *Did the explanations help you? Why?* and *What else in addition to explanations do you require to use such systems?*

Out of 30 respondents, 13 clinicians indicated that they found the explanations helpful. However, the same 13 participants also commented that while they are helpful, there is a potential of they being more comprehensive. One participant stated, "*Prior knowledge is must*" and another remarked "*it helped but only to an extent.*" This is in accordance with our quantitative analysis, which communicated the need for explanations to have a high recall and have precise salient regions for them to be perceived as trustworthy by the clinicians. Further, a few clinicians suggested that additional details are necessary to enhance understanding of the overall system. This included providing additional information such as *the age of the patient, clinical history, and duration of illness during the imaging*. The findings demonstrate that while machine learning model explanations can assist doctors in making diagnosis, they are not considered comprehensive for enhancing user trust. As indicated by a respondent, "[the explanations] *save time for a clinician.*" The findings indicate that further advancements in techniques to generate

explanations with rigorous measures to evaluate their quality are required to enhance reliability, understanding, and trust in using imaging for medical diagnostics.

6. Conclusion

With recent advances in the use of deep learning for medical diagnostics, it is essential to understand the efficacy of explainable models. We presented our study on generating machine learning explanations using several state-of- the-art techniques to evaluate them with a human benchmark for CT images of COVID-19 and common pneumonia infections. Our data set can be used for quantitative evaluation of saliency maps based on human annotations. This evaluation enables us to quantify the quality of explanations generated by existing methods. We further performed an empirical study to evaluate the trust of clinicians based on explanations which had high and low precision and recall grounded on existing pointing game metrics. The results were inconclusive for trust and reliability, but conclusive for understanding. Evidence suggests that explanations increase trust on the system; however, they reduce the understanding and reliability. This is because clinicians require precise and complete explanations for a high-stake decision support system. Considering our study presented only ML explanations that had some overlap with human judgement, the results present an optimistic outcome.

As for our future work, we are interested in deriving metrics that would be effective in evaluating machine explanations for CT images. These measures would need to consider overlapping and non-overlapping salient regions of explanations and human annotations. As indicated in the qualitative feed-back, clinicians would like to consider additional information for better understanding and diagnosis. Here, multi-modal feature representation that considers images, text, or genomics data can be used to generate and present ML explanations as motivated by prior research [43]. In the user survey, we would further assess user trust by including additional factors such as familiarity and users' propensity to trust that could influence our current findings. Additional evaluation methods such as the System Causability Scale can be explored to evaluate the quality of explanations. Finally, the number of clinicians available to respond during the pandemic was a limitation in our current study. We would strive to extend our study to a larger number of clinicians, which would also assist in obtaining statistically significant results.

Funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare no conflict of interest for the work presented in this manuscript.

References

[1] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest ct for covid-19: comparison to rt-pcr, Radiology 296 (2020) E115–E117, https://doi.org/10.1148/radiol.2020200432, pMID: 32073353.

[2] Y. Li, D. Wei, J. Chen, S. Cao, H. Zhou, Y. Zhu, J. Wu, L. Lan, W. Sun, T. Qian, K. Ma, H. Xu, Y. Zheng, Efficient and effective training of covid-19 classification networks with self-supervised dual-track learning to rank, IEEE J. Biomed. Health Inform. 24 (2020) 2787–2797.

[3] H. Gunraj, L. Wang, A. Wong, Covidnetct: a tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images, Front. Med. 7 (2020) 1025, https://doi.org/10.3389/fmed.2020.608525.

[4] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What Do We Need to Build Explainable Ai Systems for the Medical Domain?, 2017 arXiv:1712.09923.

[5] A.B. Arrieta, N.D. Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Bar- bado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[6] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: a survey on methods and metrics,, Electronics 10 (2021), https://doi.org/10.3390/electronics10050593. URL, https://www.mdpi.com/2079-9292/10/5/593.

[7] F. Doshi-Velez, B. Kim, Towards a Rigorous Science of Interpretable Ma- Chine Learning, 2017, 08608 arXiv:1702.

[8] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?": ex- plaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.

[9] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vi- Sion, ICCV), 2017, pp. 618–626, 2017.

[10] V. Petsiuk, A. Das, K. Saenko, Rise: randomized input sampling for explanation of black-box models, in: Proceedings of the British Machine Vision Conference, BMVC), 2018.

[11] R. Fong, M. Patrick, A. Vedaldi, Understanding deep networks via extremal perturbations and smooth masks, 2019, in: IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2019, pp. 2950–2958, https://doi.org/10.1109/ICCV.2019.00304. URL: https://doi.org/10.1109/ICCV.2019.00304.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR, 2015, 2015.

[13] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D.J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Volume 8689 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 818–833.

[14] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, J. Imaging 6 (2020) 52, https://doi.org/10.3390/jimaging6060052. URL: https://doi.org/10.3390/jimaging6060052.

[15] F. Eitel, K. Ritter, Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support - Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Proceedings, Volume 11797 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 3–11.

[16] Z. Papanastasopoulos, R.K. Samala, H.-P. Chan, L. Hadjiiski, C. Para- magul, M. A. Helvie, C.H. Neal, Explainable AI for Medical Imaging: Deep-Learning CNN Ensemble for Classification of Estrogen Receptor Sta- Tus from Breast MRI, 11314, SPIE, 2020, pp. 228–235, https://doi.org/10.1117/12.2549298.

[17] K. Young, et al., Deep neural network or dermatologist?, in: Inter- Pretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, 2019, pp. 48–55.

[18] L. Wang, A. Wong, Covid-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images, CoRR abs/2003.09871, https://doi.org/10.1038/s41598-020-76550-z, 2020.

[19] Z.Q. Lin, M.J. Shafiee, S. Bochkarev, M.S. Jules, X. Wang, A. Wong, Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms, CoRR abs/1910.07387, https://doi.org/10.1109/re.2019.00032, 2019.

[20] P. Zhu, M. Ogino, Guideline-based additive explanation for computer- aided diagnosis of lung nodules, in: Interpretability of Machine in- Telligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, 2019, pp. 39–47.

[21] W. Shui-Hua, M.A. Khan, V. Govindaraj, S.L. Fernandes, Z. Zhu, Z. Yu-Dong, Deep rank-based average pooling network for covid-19 recognition, Comput. Mater. Continua (CMC) (2022) 2797–2813.

[22] S.-H. Wang, X. Zhang, Y.-D. Zhang, Dssae: deep stacked sparse autoencoder analytical model for covid-19 diagnosis by fractional fourier en- tropy, ACM Transact. Manage. Inform. Sys.(TMIS) 13 (2021) 1–20.

[23] A. Holzinger, A.M. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (SCS), Künstliche Intell. 34 (2020) 193–198.

[24] M. Nourani, S. Kabir, S. Mohseni, E.D. Ragan, The effects of meaning- ful and meaningless explanations on trust and perceived system accuracy in intelligent systems, AAAI (2019), 2019.

[25] A. Papenmeier, G. Englebienne, C. Seifert, How Model Accuracy and Explanation Fidelity Influence User Trust, CoRR Abs/1907, 2019, 12652. URL: http://arxiv.org/abs/1907.12652.

[26] E. Vorm, Assessing Demand for Transparency in Intelligent Systems Us- Ing Machine Learning, in: Innovations in Intelligent Systems and Applications, INISTA 2018, Thessaloniki, Greece, 2018, pp. 1–7. July 3-5, 2018, IEEE, 2018.

[27] J. Zhou, Z. Li, H. Hu, K. Yu, F. Chen, Z. Li, Y. Wang, Effects of influence on user trust in predictive decision making, ACM (2019), https://doi.org/10.1145/3290607.3312962.

[28] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, F. Doshi- Velez, An evaluation of the human-interpretability of explanation, CoRR abs/1902.00006, https://doi.org/10.2139/ssrn.3064761, 2019.

[29] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, N. Berthouze, Evaluating saliency map explanations for convolutional neural networks: a user study, in: IUI '20: 25th International Conference on Intelligent User Interfaces, ACM, 2020, pp. 275–285.

[30] H. Lakkaraju, S.H. Bach, J. Leskovec, Interpretable decision sets: a joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD, ACM, 2016, pp. 1675–1684.

[31] P. Schmidt, F. Bießmann, Quantifying Interpretability and Trust in Machine Learning Systems, CoRR Abs/1901, 2019, 08558 arXiv:1901.08558.
[32] J. Zhou, H. Hu, Z. Li, K. Yu, F. Chen, Physiological indicators for user trust in machine learning with influence enhanced fact-checking, in: Machine Learning and Knowledge Extraction - International Cross- Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26-29, 2019, Proceedings, Volume 11713 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 94–113, https://doi.org/10.1007/978-3-030-29726-8_7.
[33] K. Zhang, X. Liu, J. Shen, et al., Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid- 19 pneumonia using computed tomography, Cell 181 (2020) 1423–1433, https://doi.org/10.1016/j.cell.2020.04.045, e11.
[34] M. Körber, Theoretical considerations and development of a questionnaire to measure trust in automation, in: Congress of the International Ergonomics Association, Springer, 2018, pp. 13–30.
[35] J. Zhang, S.A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top- down neural attention by excitation backprop, Int. J. Comput. Vis. 126 (2018) 1084–1102, https://doi.org/10.1007/s11263-017-1059-x.
[36] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: high confidence predictions for unrecognizable images, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2015, pp. 427–436, https://doi.org/10.1109/CVPR.2015.7298640, 2015.
[37] R.L. Draelos, L. Carin, Explainable multiple abnormality classification of chest CT volumes with axialnet and hirescam, CoRR abs/2111.12215, https://doi.org/10.1016/j.media.2020.101857, 2021.
[38] P. Hase, H. Xie, M. Bansal, Search Methods for Sufficient, Socially- Aligned Feature Importance Explanations with In-Distribution Counter- Factuals, CoRR Abs/2106, 2021, 00786.
[39] M. Nourani, J.T. King, E.D. Ragan, The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Sys- Tems, 2008, 2020, 09100. URL: https://arxiv.org/abs/2008.09100.arXiv:2008.09100.
[40] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (1937) 675–701.
[41] A. Benavoli, G. Corani, F. Mangili, Should we really use post-hoc tests based on mean-ranks? J. Mach. Learn. Res. 17 (2016) 152–161.
[42] A. Nasir, V. Gurupur, X. Liu, A new paradigm to analyze data com- pleteness of patient data, Appl. Clin. Inf. 7 (2016) 745.
[43] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Inf. Fusion 71 (2021) 28–37, https://doi.org/10.1016/j.inffus.2021.01.008.