Neural Computing and Applications (2023) 35:22087–22100
https://doi.org/10.1007/s00521-022-08021-7

S.I.: DEEP LEARNING IN MULTIMODAL MEDICAL IMAGING FOR CANCER DETECTION

Explaining COVID-19 diagnosis with Taylor decompositions

Mohammad Mehedi Hassan[1] · Salman A. AlQahtani[1] · Abdulhameed Alelaiwi[1] · João P. Papa[2]

Received: 28 May 2022 / Accepted: 26 October 2022 / Published online: 17 November 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
The COVID-19 pandemic has devastated the entire globe since its first appearance at the end of 2019. Although vaccines are now in production, the number of contaminations remains high, thus increasing the number of specialized personnel that can analyze clinical exams and points out the final diagnosis. Computed tomography and X-ray images are the primary sources for computer-aided COVID-19 diagnosis, but we still lack better interpretability of such automated decision-making mechanisms. This manuscript presents an insightful comparison of three approaches based on explainable artificial intelligence (XAI) to light up interpretability in the context of COVID-19 diagnosis using deep networks: Composite Layer-wise Propagation, Single Taylor Decomposition, and Deep Taylor Decomposition. Two deep networks have been used as the backbones to assess the explanation skills of the XAI approaches mentioned above: VGG11 and VGG16. We hope that such work can be used as a basis for further research on XAI and COVID-19 diagnosis for each approach figures its own positive and negative points.

Keywords Explainable artificial intelligence · Deep Taylor expansion · COVID-19 · Machine learning

1 Introduction

At the end of 2019, the pandemic provoked by the new coronavirus (COVID-19) has its probable emergence in Wuhan, China, with similar symptoms to those of viral pneumonia [22]. A deeper analysis of the respiratory tract in humans highlighted that the problem's origin concerns a new virus from the family *Coronaviridae*, termed SARS-CoV-2, supposedly originating from bats of species *Phinolophus* [1]. Four months later, SARS-CoV-2 ended up in a world-scale public health crisis with no precedents, killing more than 239, 000 people and infecting around 3, 435, 800 persons by May 2020.

The scientific community has dedicated the past months to the fast-run development of vaccines to mitigate and control the number of victims worldwide. However, the disease transmission is still high and with new variants scaling up as time goes by. Computer-aided approaches are paramount to help in such a scenario, either in the automatic disease identification or simulations, to better understand the rationale behind COVID-19 infection behavior among people. A common approach to diagnose/confirm a possible infection by the new coronavirus regards thoracic X-ray images [3, 14] — most patients affected by COVID-19 figure anomalies in the lungs, which are primarily used for diagnostic purposes. However, such alterations can be visible to the human eye to some extent only, for their shape, color, and texture may face subtle changes. Besides, human fatigue is another important fact that may lead to judgments prone to errors.

Research on machine learning-driven approaches to cope with the automatic COVID-19 identification has

✉ Mohammad Mehedi Hassan
mmhassan@ksu.edu.sa

Salman A. AlQahtani
salmanq@ksu.edu.sa

Abdulhameed Alelaiwi
aalelaiwi@ksu.edu.sa

João P. Papa
joao.papa@unesp.br

[1] College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[2] Department of Computing, São Paulo State University, Bauru, Brazil

flourished in the past months, with particular attention to those methods based on deep learning [8, 16–18, 29, 30]. Wang et al. [43] presented COVID-Net, a convolutional neural network (CNN) designed to detect COVID-19 from X-ray images, with promising results over a public image repository [11]. However, the model proposed by the authors figures a relatively high number of parameters that lead to considerable training time.

Santos et al. [37] showed the importance of normalizing images when dealing with COVID-19 automatic diagnosis using X-ray images and an EfficientNet-B6 architecture. Although it is consensus that normalizing images is usually satisfying for image classification purposes, no study focused on such a statement had been considered for the context addressed here. Song et al. [41] proposed an approach based on deep learning to detect COVID-19 accurately in computed tomography (CT) images. The new architecture, termed DRENet, first detects potential lesion regions by blending a pre-trained ResNet50 with a feature pyramid network. Further, ResNet50 is applied once more to the detected regions for local feature extraction, which is then combined with the global features learned in the former step. The model achieved a precision of 0.79 and a recall (sensitivity) of 0.95. However, precision and recall were boosted when using three types of CT images to 0.86 and 0.93, respectively.

Most works pay little attention to explaining the decisions learned by the intelligent models [2, 20]. The scientific community has broadly used the term “computer-aided” diagnosis in the past decades. Still, algorithms are primarily used to make decisions and not help humans to learn from their knowledge. Explainable artificial intelligence (XAI) has come to help that end so that models have been developed to light up why a specific decision has been taken apart from others[15, 26].

Pennisi et al. [32] proposed an approach based on deep learning to segment lung parenchyma and lobes for further using them as input for classification networks. The accuracies were compared against the ones provided by three expert radiologists on a dataset composed of 166 CT scans. The interpretation of the trained AI models’ outcomes revealed that most regions supporting COVID-19 identification are closely related to those areas clinically relevant. Ye et al. [46] described an initial study concerning XAI and deep learning for COVID-19 automatic diagnosis using CT scan volumes. The authors compared the proposed approach against class activation maps (CAM) [47], arguing that the latter can be used as a post-processing procedure only, i.e., the network should be trained first. The proposed approach incorporates LIME [33] in its classification module to estimate each image’s superpixel contribution in the final prediction.

Montavon et al. [27] came up with the idea of representing the contribution of each input neuron to the model’s explainability as a decomposition of functions. The work investigated the applicability of Single Taylor Decomposition to encode the relevance of each neuron during inference, and they proposed the Deep Taylor Decomposition. The authors also highlighted its similarities with Layer-wise Relevance Propagation.

The primary contribution of this manuscript is to compare some XAI-based approaches in the context of computer-aided COVID-19 identification, i.e., Composite Layer-wise Propagation [35], Single Taylor Decomposition [27], and Deep Taylor Decomposition [27]. As far as we know, no work aimed at such an analysis up to date. The techniques used for explainability purposes are compared considering three factors: (i) explainability continuity, (ii) explainability selectivity, and (iii) input perturbation.

In short, the main contributions are:

- To compare Single Taylor Decomposition, Deep Taylor Decomposition, and Layer-wise Propagation for explainability in computer-aided COVID-19 identification;
- Different scenarios are considered to evaluate the degree of explainability; and
- To foster research on XAI applied to COVID-19 diagnosis.

The remainder of this paper is organized as follows. Sects. 2 and 3 present a review of the literature and the background theory, respectively. Sect. 4 describes the methodology, and the experiments are discussed in Sect. 5. Last but not least, Sect. 6 states conclusions and future works.

2 Related works

Pennisi et al. [32] presented a recent and interesting study concerning XAI applied to computer-assisted COVID-19 diagnosis using CT scans. Their research was focused on comparing the results obtained by experts with the ones achieved by computers. However, they did not use XAI tools to assess the regions that matter for computer-driven classification purposes. DeGrave et al. [13] showed that deep learning systems used to detect COVID-19 from chest radiographs rely on several factors other than medical pathology only. The authors argued that machine learning models trained straightforwardly present undesired effects in real-world scenarios. Their work recommends further examination of the results by experts and reporting the outcomes using XAI tools. Moreover, we should remain skeptical of high performances without external validation.

Serte and Demirel [39] also considered CT images but at a tridimensional scale. A ResNet50 was applied to classify each patient's slide for further fusing image-level predictions from the 3D scan to take the final decision. Al-Waisy et al. [4] proposed a hybrid framework to cope with COVID-19 diagnosis using deep learning called COVID-CheXNet. The approach combines predictions from both ResNet34 and HRNet architectures to make the final decision. Dansana et al. [12] considered X-ray and CT images to distinguish between COVID-19 and pneumonia using CNNs. The well-known VGG-16 and Inception-V2 deep architectures and a decision tree have been considered for classification purposes, with VGG-16 achieving the top results.

Wang et al. [44] introduced a PatchShuffle Stochastic Pooling Neural Network to detect COVID-19, which outcomes were further analyzed by Gradient-weighted Class Activation Mapping (Grad-CAM) [38]. The proposed approach outperformed nine state-of-the-art techniques in the experiments. Wu et al. [45] developed a joint approach composed of segmentation and classification modules to perform real-time and explainable COVID-19 diagnosis using chest CT images. The idea is to perform segmentation only in the images that were classified as positive for COVID-19.

Brunese et al. [9] presented a three-step approach to distinguish between pneumonia and COVID-19: (i) first, given an X-ray image, the system detects whether it carries pneumonia or not; (ii) if so, the second step aims at differentiating between pneumonia and COVID-10; and (iii) the last step segments the regions that figure the disease indicators. The DeepCOVIDExplainer was proposed by Karim et al. [24] to provide explainable diagnosis using chest X-ray images. The authors used Grad-CAM, Grad-CAM++ [10], and Layer-wise Relevance Propagation (LRP) [23] tools to provide insights into the classification step, which used an ensemble of deep networks.

Alshazly et al. [5] also considered chest CT scans to diagnose COVID-19 infection automatically using transfer learning. Visualization techniques have been used to better understand the prediction step's outcomes. Hryniewska et al. [19] presented an interesting critic for proper usage of deep learning models and explainable tools in the context of COVID-19 diagnosis. The first concern is related to the quality of images available in the public datasets, for only a few use DICOM format for proper storage. They use 8-bit JPG or PNG format mostly. The second situation refers to the few images with low and medium severity cases. Usually, the works try to distinguish between healthy and infected individuals or between pneumonia and COVID-19. Other issues are imbalanced and mixed (CT with X-ray images) datasets. Data augmentation is used indiscriminately, for not all approaches are appropriate to the medical domain, e.g., rotation or flipping in CT and X-ray images since they are customarily taken using standardized protocols.

Bassi and Attux [7] used dense convolutional networks and transfer learning to classify X-ray images into three labels: COVID-19, pneumonia, and healthy individuals. LRP was further used to generate heat maps and analyze the outcomes. Fuhrman et al. [15] reviewed several explainable AI techniques to assist COVID-19 identification. The authors highlighted different aspects, advantages, and disadvantages of the techniques considered in their work. XAI can be embedded to bring interpretability but at the price of bringing the performance down. Therefore, its choice relies on the application itself.

Hu et al. [21] proposed a multi-input and fuzzy convolutional neural network to detect COVID-19 from torso X-ray images. Explainable approaches were used to investigate forecasts provided by the neural mode. The authors concluded that transfer learning and pre-trained models are helpful in such a context. Aviles-Ribeiro et al. [6] highlighted the problem of obtaining a suitable number of labeled samples for COVID-19 identification. The authors introduced a graph-based semi-supervised learning framework that used X-ray images to recognize COVID-19. Attention maps accommodate the radiologist's mental model.

3 Background theory

3.1 Taylor expansion

A Taylor series is an expansion of some infinitely differentiable function (in an open interval) into an infinite sum of terms, where each term has a larger exponent, as follows the example below:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{1}$$

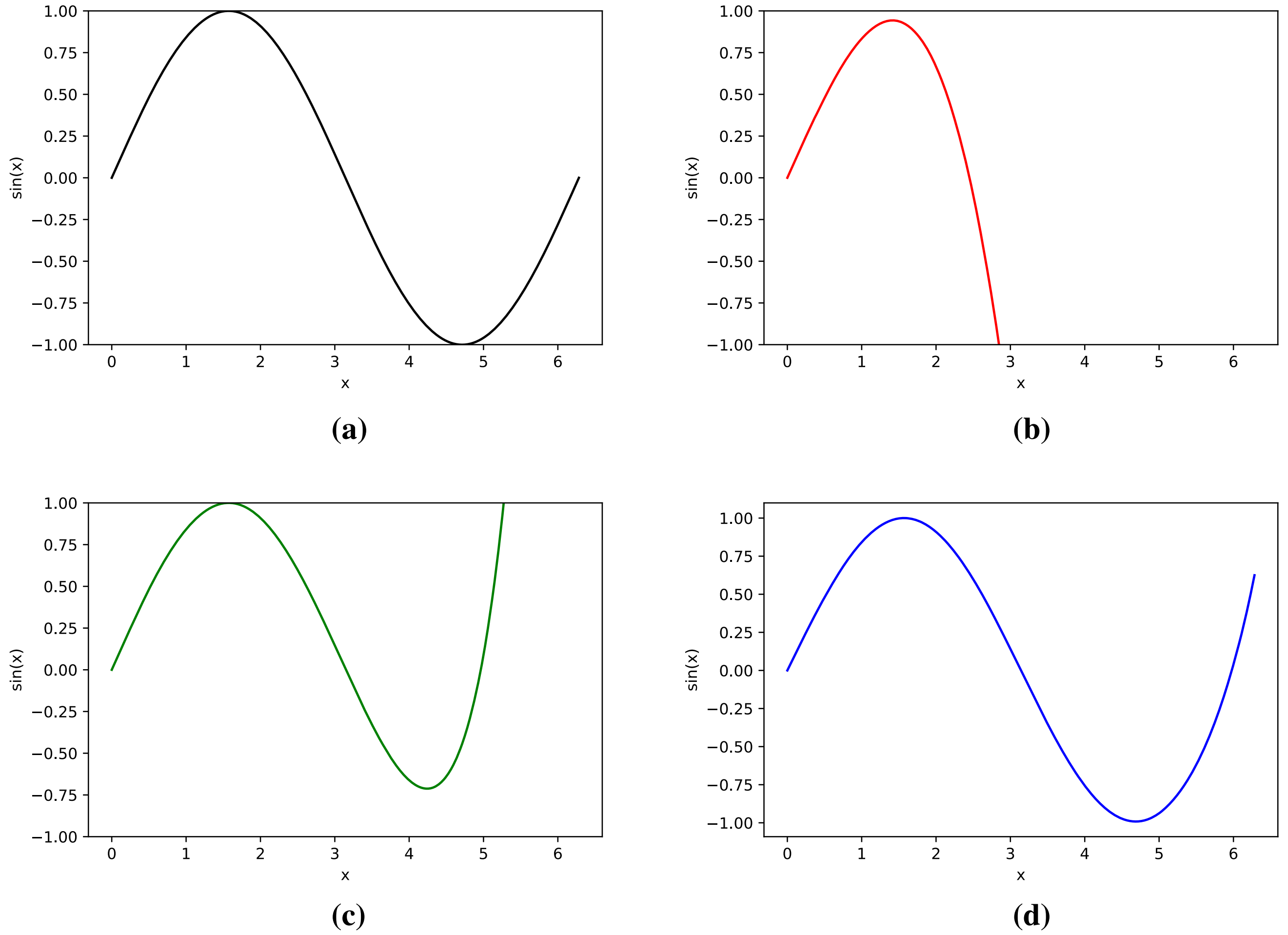
where n stands for the number of terms. The higher the number of terms, the better the approximation.

For the sake of explanation, consider approximating $\sin(x)$ function:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{(2n+1)}. \tag{2}$$

Figure 1 depicts the approximation of $\sin(x)$ function using different numbers of terms. One can observe that seven terms can approximate the function quite reasonably.

The key idea behind Taylor series concerns the fact that a function can be approximated using a summation of high-



(a)

(b)

(c)

(d)

Fig. 1 Approximating function $sin(x)$ using Taylor Series with different numbers of terms: **a** standard function, **b** approximated with two terms, i.e., $sin(x) = x - \frac{x^3}{3!}$, **c** approximated with five terms, i.e., $sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}$, and **d** approximated with five terms, i.e., $sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!}$

order polynomials around a neighborhood of some root point $\tilde{x} \in \Re$. In other words, one wants to evaluate/decompose $f(x)$ when it is close to $\tilde{x}$. The general formulation for such an assumption is given below:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\tilde{x})}{n!}(x - \tilde{x})^n, \quad (3)$$

where $f^{(n)}(\tilde{x})$ is the n-th derivative of f at the point $\tilde{x}$. A specific case of the above equation is termed as Maclaurin series when $\tilde{x} = 0$.

3.2 Deep Taylor expansion

Many machine learning models are complex and non-linear when considering them globally. On the other hand, they might be simpler and, sometimes, linear when taken locally. Now, let us assume that f is positive-valued and takes the form $f : \Re^d \rightarrow \Re$. Concerning image classification, the input $x \in \Re^d$ denotes an image with pixel values $x = \{x_p\}$, where p stands for a particular pixel.

Let us consider the first-order Taylor expansion of $f(x)$:

$$f(x) = f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) + \epsilon, \quad (4)$$

where ϵ denotes the higher-order terms of the expansion. In practice, Equation 4 is simply a different approach to represent the general formulation of the Taylor expansion (Equation 3).

According to Montavon et al. [27], Equation 4 can be reformulated as follows:

$$f(x) = f(\tilde{x}) + \sum_p f'(x_p)(x_p - \tilde{x}_p) + \epsilon, \quad (5)$$

where $\tilde{x}_p$ denotes the pixel values of the root point $\tilde{x}$, and $\sum_p$ runs over all pixels in the image.

Let $R_p(x)$ be a relevance score associated with each pixel p, i.e., it indicates to what extent pixel p contributes to explaining the classification decision $f(x)$. Besides, let $R(x) = \{R_p(x)\}$ be a heatmap that is composed of all pixel scores. According to Montavon et al. [27], a heatmapping $R(x)$ is *conservative* if it satisfies the condition below:

$$f(x) = \sum_p R_p(x), \quad \forall x. \tag{6}$$

Such a condition guarantees that the relevance scores correspond to the extent to which an object in the input image is detected by the model, i.e., $f(x)$. Also, a heatmapping is said to be *positive* if it obeys the following restriction:

$$f(x) = R_p(x) \geq 0, \quad \forall x, p. \tag{7}$$

The above constraint ensures the relevance scores are not contradictory regarding the presence or absence of the detected object in the image. We said that $f(x) = 0$ when the object is absent in the image, and $f(x) > 0$ quantifies the presence of this object. Last but not least, we say that a heatmapping $R(x)$ is *consistent* when it is conservative and positive.

According to Montavon et al. [27], the heatmapping $R(x)$ can be formulated as the element-wise product $\odot$ between the gradient of the function at the root point, i.e., $f'(\tilde{x})$, and the difference between the image and the root, as follows:

$$R(x) = f'(\tilde{x}) \odot (x - \tilde{x}). \tag{8}$$

Essentially, the formulation above says that the magnitude of the gradient at each pixel will tell us its relevance for classification purposes.

The Taylor decomposition concerning function $f(x)$ has one free variable, i.e., the choice of the root point $\tilde{x}$. Mathematically speaking, we want to observe the behavior of function $f(x)$ in the neighborhood of that root point. In general terms, we want to study how the function behaves when the object of interest is absent in the image, i.e., $f(\tilde{x}) = 0$, for we expect it to be in the image. Such a situation holds when the minimize the following objective function [42]:

$$\tilde{\zeta} = \underset{\zeta}{\operatorname{argmin}} \|\zeta - x\|^2 \quad \text{s.t.}\ f(\zeta) = 0 \ \text{and} \ \zeta \in \mathcal{X}, \tag{9}$$

where $\mathcal{X}$ stands for the image domain. Montavon et al. [27] stated that finding proper values for $\tilde{x}$ is time-consuming when $f(x)$ is computationally expensive. Moreover, for deep networks, nearest root points are usually not perceivable (visually speaking) from x.

We can rewrite the first-order Taylor expansion from Equation 5 as follows:

$$f(x) = f(\tilde{\zeta}) + \underbrace{f'(\tilde{\zeta})(x - \tilde{\zeta})}_{R(x)} + \epsilon \tag{10}$$

$$= f(\tilde{\zeta}) + R(x) + 0. \tag{11}$$

Since we are considering a first-order expansion, we can discard the higher-order terms, i.e., $\epsilon = 0$.

3.2.1 Extension to deep networks

Let us assume that $f(x)$ models a deep neural network. The idea of Deep Taylor Expansion is to understand that a complex and non-linear function learned by $f(x)$ can be decomposed into a set of simpler subfunctions [27]. Let us assume that $f(x)$ has been decomposed on the set of neurons at a given layer, and let x_j be such a neuron and R_j its assigned relevance. In a nutshell, we want to decompose R_j on the set of lower neurons $\{x_i\}$ to which x_j is connected.

Considering neuron x_j at the current layer to be analyzed, we define $\{\tilde{x}_i\}^j$ as the root point. Assuming that $\{x_i\}$ and R_j are related by a function $R_j(\{x_i\})$, the Taylor decomposition of R_j is computed as follows:

$$\begin{aligned} R_j &= R'_j(\{\tilde{x}_i\}^j)^T(\{x_i\} - \{\tilde{x}_i\}^j) + \epsilon_j \\ &= \sum_i \underbrace{R'_j(\{\tilde{x}_i\}^j)(x_i - \tilde{x}_i^j)}_{R_{ij}} + \epsilon_j, \end{aligned} \tag{12}$$

where ϵ_j denotes the Taylor residual at neuron x_j, and R_{ij} stands for the redistributed relevance from neuron x_j to neuron x_i.

In order to estimate the total relevance of neuron x_i, we need to consider all relevance values from neurons $\{x_j\}$ to which neuron x_i contributes:

$$R_i = \sum_j R_{ij}. \tag{13}$$

Combining Equations 12 and 13, we obtain:

$$R_i = \sum_j R'_j(\{\tilde{x}_i\}^j)(x_i - \tilde{x}_i). \tag{14}$$

Figure 2 illustrates the idea of layer-wise propagation on a deep network, where x_f denotes the output neuron of such deep network. One can observe that the neuron's activations are backpropagated to the input image to highlight relevant pixels.

4 Methodology

This section details the methodology used to evaluate some XAI approaches for the automatic COVID-19 identification.

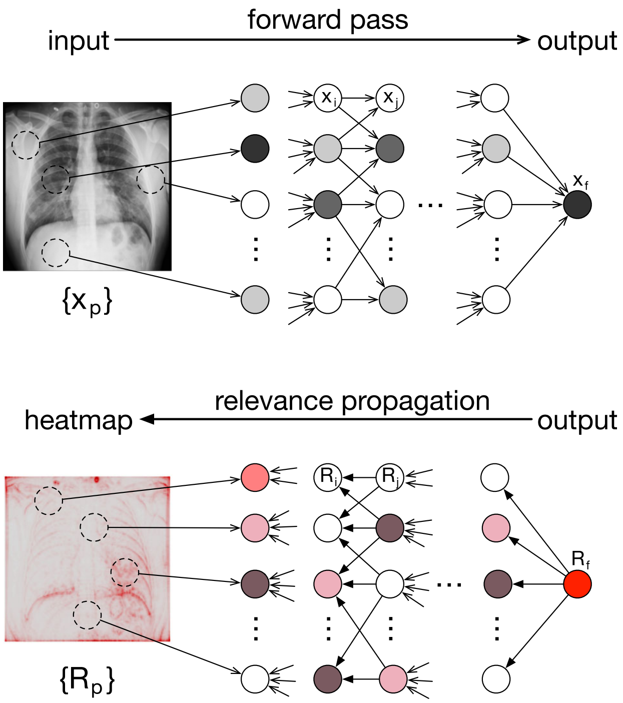


Fig. 2 The working mechanism of the deep Taylor decomposition. A prediction for the class “COVID-19” is estimated by forwarding the pixel values $\{x_p\}$ (input) to the deepest layers of the neural architecture. The output is encoded by neuron x_f. A relevance score $R_f = x_f$ is assigned to the output neuron and backpropagated to the input layer, where R_p denotes the pixel’s relevance scores, visualized as a heatmap. Adapted from [27]

4.1 Deep neural architectures

We considered three well-known deep neural architectures that employ different mechanisms for training purposes, for the primary idea of this work is to evaluate the behavior of some XAI tools in distinct scenarios:

- VGG11 [40]: it uses 224 × 224 RGB images forwarded through a stack of 3 × 3 convolutional filters with stride fixed to 1 pixel; five max-pooling layers carry out spatial pooling over 2 × 2-sized windows; the fully connected layers follow a stack of convolutional layers, where the first two have 4, 096 channels each, and the third contains 1, 000 outputs since it has been designed to address classification in the ImageNet dataset [34]; the last layer stands for a softmax layer.
- VGG16 [40]: it comprises a similar architecture to VGG11 but with extra convolutional layers. The output layer has been modified to accommodate three classes since it has been designed to address classification in the ImageNet dataset [34].

4.2 Dataset

We used the “COVID-19 Radiography Dataset”[1], which comprises 15, 153 chest X-ray images divided into three classes: (i) 1, 345 images positive to viral pneumonia, (ii) 3, 616 images positive to COVID-19, and (iii) 10, 192 images from healthy people. Figure 3 depicts some examples from the aforementioned dataset.

4.3 Experimental setting

Since we are using deep neural architectures, we perform data augmentation in the training dataset to double its size using horizontal flipping on every training image. Such transformation does not affect the natural appearance of the images, for we are dealing with chest X-ray data.

Out of the 15, 153 images, 15, 063 are employed to compose the training set (99.4%), and the remaining 90 images are used as the test set. Since we are interested in evaluating XAI techniques and not outperforming state-of-the-art approaches in terms of COVID-19 identification, we understand that large training sets will be further helpful in building consistent models so that their explainability can be clarified.

All deep architectures mentioned in Sect. 4.1 were first trained on ImageNet for further fine-tuning in the COVID-19 Radiography Dataset for one epoch only[2], which showed to be enough to reach recognition rates higher than 95%. We used mini-batches of size 6, cross-entropy as the loss function, Adam optimizer [25], and a learning rate of 3×10^{-5}. Concerning XAI tools[3], we compared Composite LRP [35], Single Taylor Decomposition [27], and Deep Taylor Decomposition [27].

4.4 Quantitative analysis

Explainable AI primarily refers to interpreting results using visual perspectives, i.e., qualitative understanding. However, one also can provide deeper insights using quantitative evaluation. In this paper, we consider three measures to accomplish this task: (i) input perturbation [36], (ii) selectivity [28], and (iii) continuity [28]. More details about their working mechanism are provided in the further section

[1] Available at https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[2] We used the official implementation provided by PyTorch [31].

[3] We used the code available at https://github.com/hans66hsu/nn_interpretability.

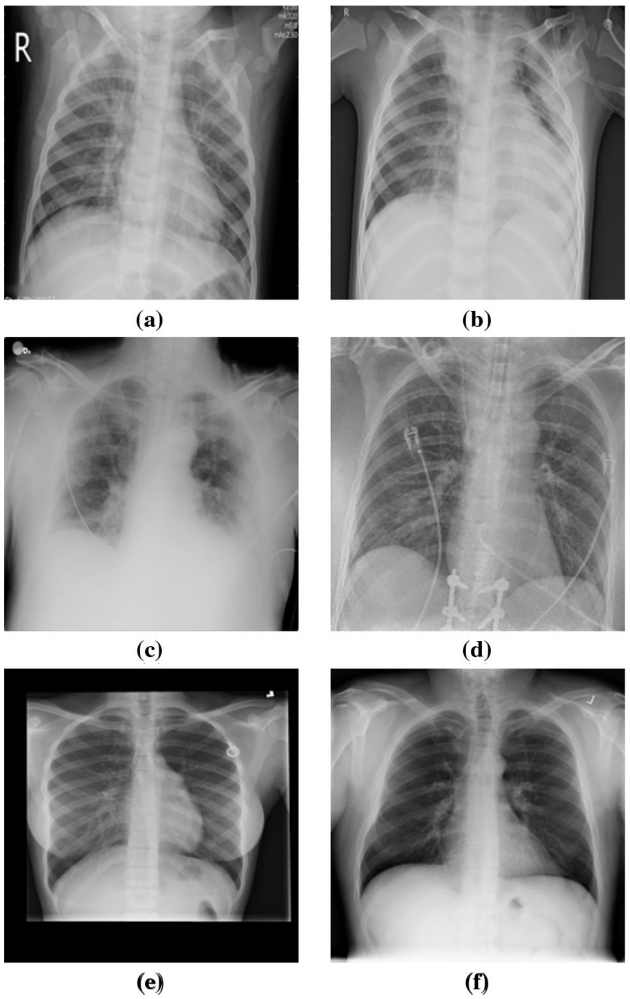


Fig. 3 Some samples from the dataset: **a**, **b** images positive to viral pneumonia, **c**, **d** images positive to COVID-19, and **e**, **f** images from healthy people

5 Experiments

In this section, we present the experimental results concerning the methodology described in the previous section.

5.1 Input perturbation

Input perturbation aims to evaluate to what extent regions of the input image identified as relevant by XAI tools take that role. The rationale is: Given a trained model, we use a test image as an input to obtain its heatmap (Fig. 2) according to some XAI technique designed for that specific deep network. Further, the most relevant regions[4], i.e., groups of pixels, have their values changed in the original (input) image by uniformly and randomly generated values. The modified image is then presented to the network once more for classification purposes. Such methodology is repeated a few more iterations to have a more significant number of patches changed. It is expected that the model effectiveness will be neglected as the number of modified patches increases. An effective good XAI technique will be more affected by these changes than the poorest one.

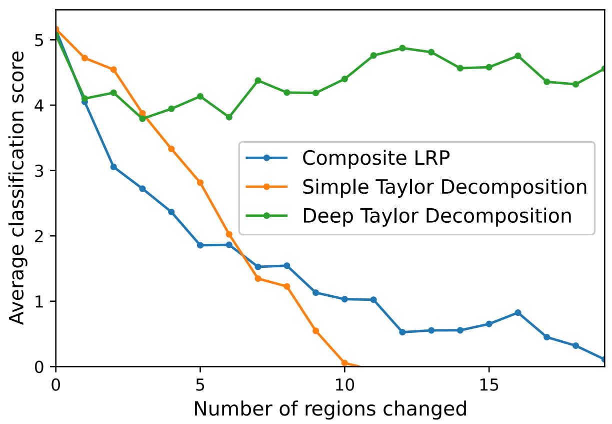


Fig. 4 VGG16 outcomes concerning COVID-19 label using input perturbation assessment

Figure 4 displays the VGG16 outcomes concerning the COVID-19 class[5]. We performed the above methodology for all test images so that prediction scores (y-axis) stand for the average values over the test set. Moreover, the y-axis stands for the so-called "classification score," which is the confidence value outputted by the neuron in charge of recognizing the COVID-19 label on the deep network's last layer. Therefore, the higher the classification score, the most accurate the model will be in identifying COVID-19 (true positive for that class).

One can observe that Composite LRP appears to be the most effective approach in identifying the relevant regions, for its classification score decreases faster than others. Deep Taylor Decomposition (DTD) figured as the worst since its behavior does not change that much when the number of modified regions increases. Such performance can be explained for DTD highlights high-frequency regions of the image, i.e., lung borders mainly. Figure 5 illustrates such scenario. It seems that such regions are not plausible to distinguish between COVID-19 and other classes (i.e., healthy and viral pneumonia), for internal parts of the lungs are the ones affected by the diseases. Since lung borders cover most of the image, only a very high number of modified patches will affect the classification score.

Concerning the explanations presented in Fig. 5, one can observe that both Composite LRP and Single Taylor Decomposition (STD) figure two distinct colors, i.e., red

[4] We used 32 × 32-sized patches.

[5] We present results only for COVID-19 patients, for we are interested in investigating the behavior of XAI tools for that scenario only.

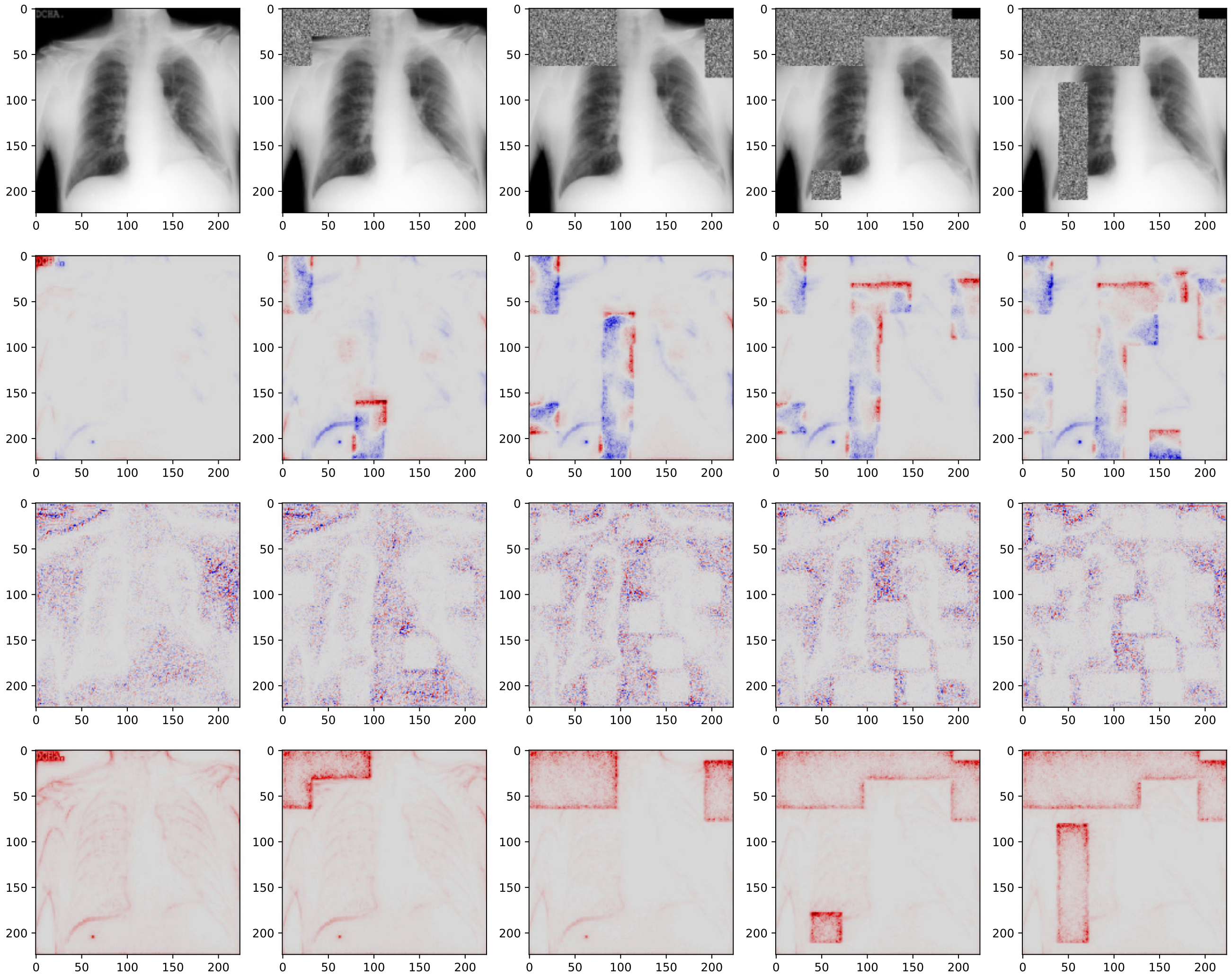


Fig. 5 From left to right: original image and its versions with three, eight, ten, and fourteen patches modified. From top to bottom: a random image from the test set positive to COVID-19, and heatmaps produced by Composite LRP, Single Taylor Decomposition, and Deep Taylor Decomposition, respectively (input perturbation analysis)

and blue. The former stands for the regions that are relevant for the identification of COVID-19, and the blue ones denote areas that do the opposite effect.

One can assess the performance of the XAI techniques by computing the area under the curve (AUC). The concept is that smaller AUC values stand for more precise approaches, i.e., the ones that can accurately highlight the most relevant parts of the image. Table 1 presents the AUC values concerning VGG16 deep network with respect to the results depicted in Fig. 4. The smallest AUC value is highlighted in bold.

Figure 6 illustrates the input perturbation analysis considering VGG11 deep network. In this case, Deep Taylor Decomposition obtained the best results (AUC = 223.54), followed by Single Taylor Decomposition (AUC = 261.33), and Composite LRP (AUC = 269.54). The difference now relies on the depth of the networks. Although VGG11 figures fewer convolutional layers than VGG16, its training loss (0.1271) was slightly smaller than VGG16 (0.1372). Such behavior might be due to the complexity of the VGG16 network, which may require more data for training purposes.

According to Montavon et al. [28], LPR tends to produce better explanations when the number of layers is kept

Table 1 AUC values concerning VGG16 deep network with respect to input perturbation analysis

Technique	AUC
Composite RLP	**28.68**
Single Taylor Decomposition	23.26
Deep Taylor Decomposition	82.96

The values in bold stand for the most effective outcomes

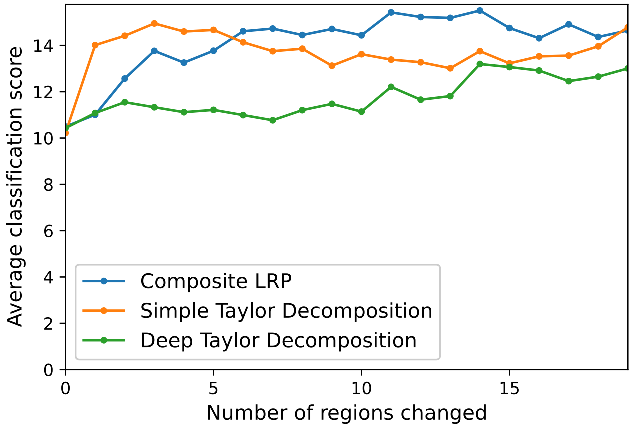


Fig. 6 VGG11 outcomes concerning COVID-19 label using input perturbation assessment

low, for the neurons' relevance is redistributed along with the network. Also, for LRP to best match DTD, average- or sum-pooling layers are preferred to max-pooling, and that does not happen in VGG-like models, which mainly use max-pooling layers. We, therefore, confirm the assumptions made by Montavon et al. [28].

5.2 Selectivity

Selectivity can be understood as a particular scenario of the input perturbation, for we "'remove" the most relevant areas of the input image instead of changing their values. In short, we zeroed the pixels' values inside that regions so that one can evaluate the robustness of the XAI technique. The rationale is the same as pixel perturbation, i.e., we expect that the prediction scores decrease as the number of zeroed (and relevant) regions increases.

Figure 7 depicts the selectivity results concerning VGG16 deep network. Once more, Composite LRP obtained the best results, for its classification score decreased faster than STD and DTD approaches as the number of removed regions increases. However, this experiment showed better effectiveness of Deep Taylor Decomposition when compared to its counterpart version, i.e., the Single Taylor Decomposition.

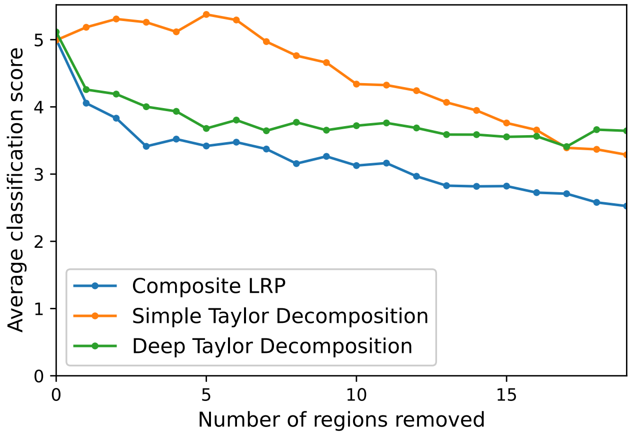


Fig. 7 VGG16 outcomes concerning COVID-19 label using selectivity assessment

Table 2 AUC values concerning VGG16 deep network with respect to selectivity analysis

Technique	AUC
Composite RLP	**59.79**
Single Taylor Decomposition	85.96
Deep Taylor Decomposition	72.06

The values in bold stand for the most effective outcomes

Deep Taylor Decomposition achieved results that are somehow close to the ones related to input perturbation, as one can observe in Table 2. However, both Composite LRP and Single Taylor Decompositions have their AUC values strongly affected by removing the most relevant regions. Although DTD is heavily based on border information and zeroing patches that fall in those regions induce discontinuities in the lung borders, the most relevant areas fall in the peripherical regions, thus affecting less the classification scores. Figure 8 illustrates such a scenario, where a considerable number of removed patches in the case of STD are spread over the entire image.

Figure 9 depicts the selectivity experiment concerning VGG11 model. One can observe that Composite LRP obtained results (AUC = 118.82) that are a bit better than Single Taylor Decomposition (AUC = 135.19), and Deep Taylor Decomposition (AUC = 135.41). Such results were different from those obtained in the input perturbation experiment, where DTD achieved the best results with the VGG11 model. It seems VGG16 leads to higher discrepancies among the XAI techniques than VGG11. We believe this scenario might be due to the better generalization capabilities demonstrated by VGG11, for it obtained a lower loss value during training. We understand that better-trained models lead to better explanations, regardless of the approach used for such a purpose.

5.3 Continuity

A suitable property of any explanation technique is to output continuous explanation functions, for it is often assumed that $f(x)$ is continuous either. Montavon et al. [28] stated that the following behavior should be ensured for a particular explanation technique: If two points are somehow equivalent, then the explanation of their predictions should also be comparable.

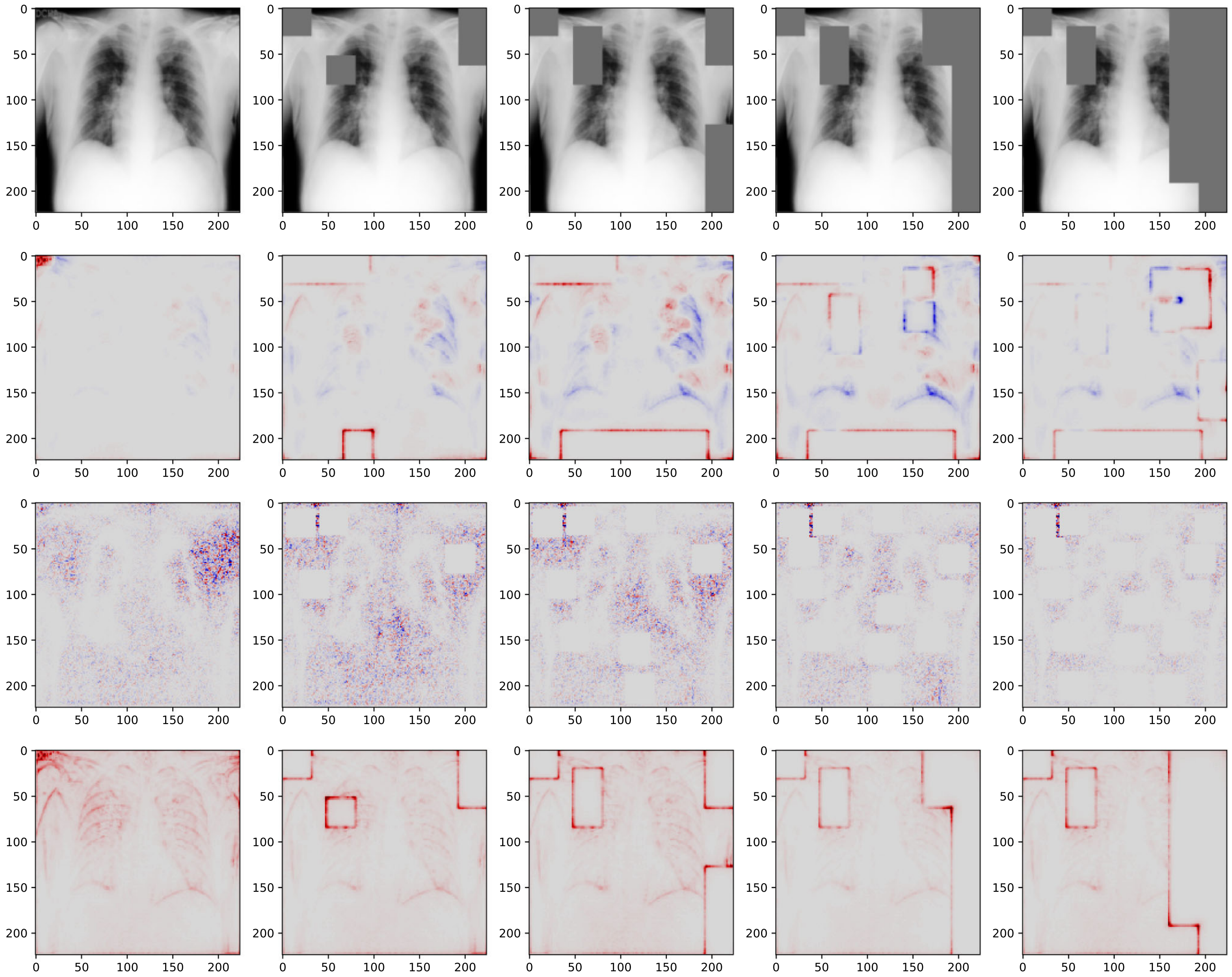


Fig. 8 From left to right: original image and its versions with three, eight, ten, and fourteen patches modified. From top to bottom: a random image from the test set positive to COVID-19, and heatmaps produced by Composite LRP, Single Taylor Decomposition, and Deep Taylor Decomposition, respectively (selectivity analysis)

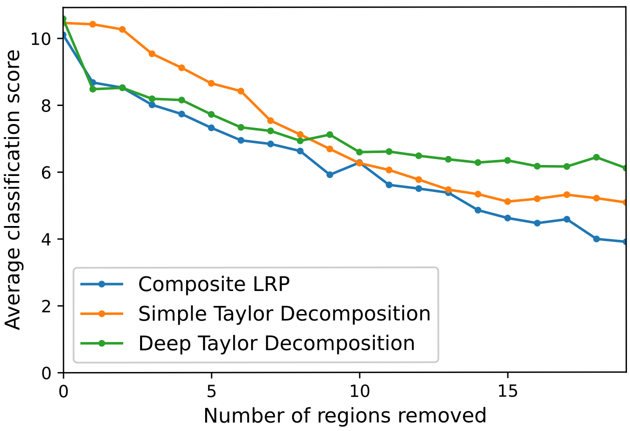


Fig. 9 VGG11 outcomes concerning COVID-19 label using selectivity assessment

The explanation continuity can be demonstrated by searching for the most substantial variation on the relevance maps. Montavon et al. [27] also stated that when $f(x)$ is a deep ReLU network, Simple Taylor Decomposition has sharp discontinuities in its explanation functions; on the other hand, Deep Taylor Decomposition produces continuous explanations.

In general, we can evaluate the robustness of the explanation approach by taking into account its “level of continuity” when we perform image translation (i.e., pixel shifting). The idea is to partition the image into quadrants (R_1 to R_4) so that explanation continuity will be assessed for each region, as depicted in Fig. 10.

Figure 11 illustrates the continuity assessment considering a random test image positive to COVID-19[6]. One can

[6] We used image COVID-82.png for such experiment.

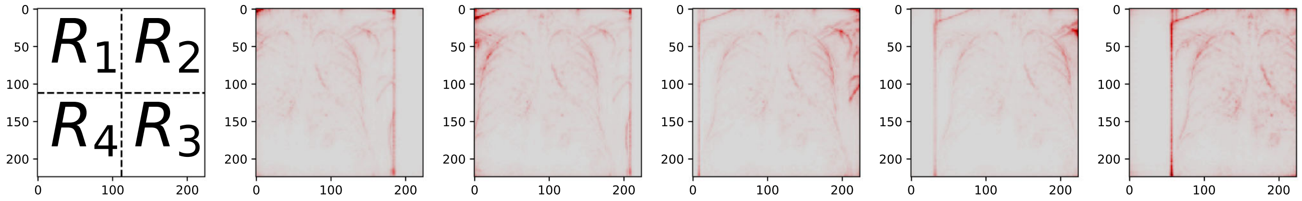
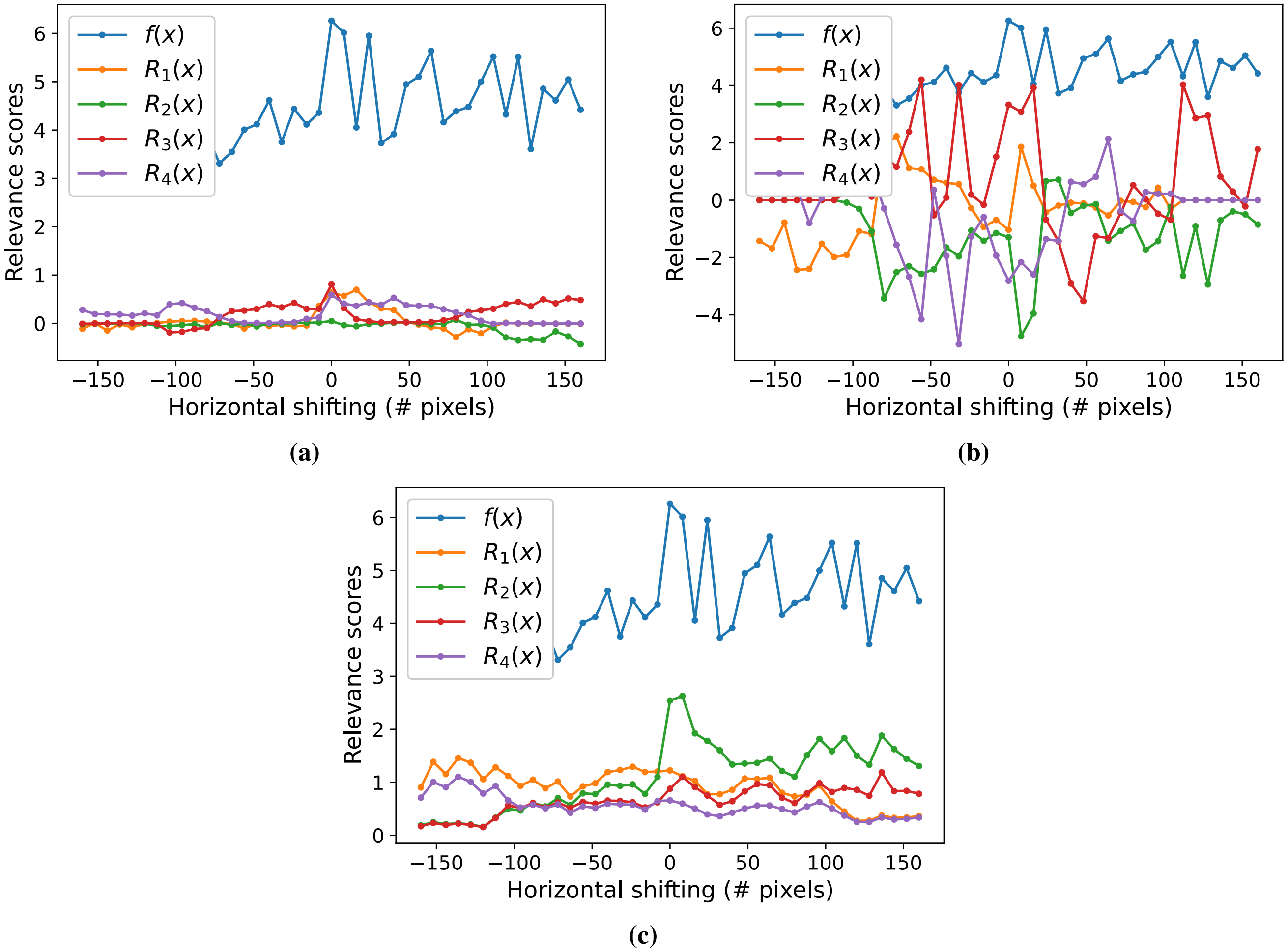


Fig. 10 Image is divided into quadrants for further explanation continuity assessment. Example of continuity analysis using Deep Taylor Decomposition



(a)

(b)

(c)

Fig. 11 Continuity analysis with VGG16 model: **a** Composite LRP, **b** Single Taylor Decomposition, and **c** Deep Taylor Decomposition

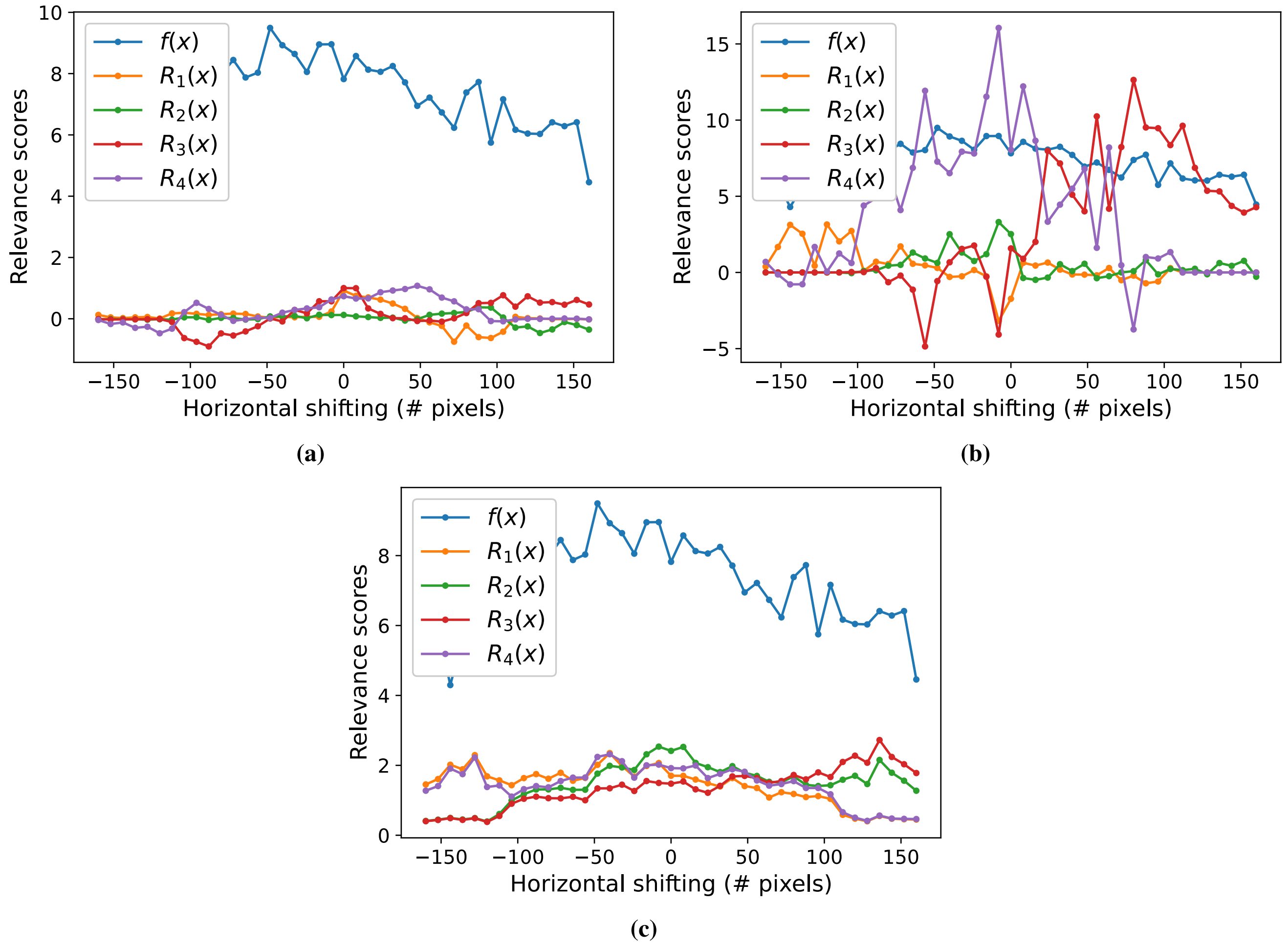
observe that Composite LRP and DTD produce “better behaved” functions, i.e., they tend to produce continuous functions as we shift the input image horizontally. On the other hand, STD figures sharp transitions, ending up in functions that are not continuous.

Figure 12 illustrated the continuity analysis under the same image used previously for VGG16. One can observe a similar behavior, i.e., Composite LRP and DTD tend to produce continuous functions. At the same time, STD has functions with sharp transitions when we increase the number of shifted pixels.

5.4 Discussion

The primary goal of the manuscript is to compare three XAI-based approaches to distinguish between viral pneumonia, COVID-19, and healthy individuals. The comparison considers three different aspects: (i) input perturbation, (ii) selectivity, and (iii) continuity. Besides, two neural backbones are employed: VGG-11 and VGG-16.

One can observe from Fig. 4 that Composite LRP can find the most relevant regions up to a certain extent, for the



(a)

(b)

(c)

Fig. 12 Continuity analysis with VGG1 model: **a** Composite LRP, **b** Single Taylor Decomposition, and **c** Deep Taylor Decomposition

classification accuracy drops rapidly until five regions are changed. For a number greater than that, STD takes the role, with all images classified incorrectly when ten regions are changed. However, with a lighter architecture (i.e., VGG-11), Deep Taylor Decomposition achieves better results (Fig. 6). The structure of the lungs is somehow well preserved when using DTD for explanation purposes (Fig. 5).

Selectivity plays a similar role (Figs. 7 and 9), except for the latter one, where Single Taylor Decomposition performed better than its deep counterpart. A possible explanation relies on the neural backbones, which might not be deep enough to benefit Deep Taylor Decomposition.

Figure 5 shows that all methods highlight high-frequency regions as important, which is usually expected. However, some of these regions comprise the patches that have been modified by either input perturbation or selectivity approaches. DTD appears to be less affected by the artificially changed patches, for it did not hallucinate about regions that have not been changed. Take the second column from left to right in Fig. 5. Composite LRP (second row from top to bottom) seemed to "see" patch patterns in the bottom-middle portion of the image, which does not happen to be. Single Taylor Decomposition also hallucinates about patches spread of the image (third row, second column).

6 Conclusions and future works

Explainable artificial intelligence has been a valuable asset to provide out-of-the-box explanations about the inner mechanisms of deep neural networks. Such a paradigm is a game-changer when dealing with automated decisions that must be further clarified.

In this manuscript, we coped with computer-assisted COVID-19 identification using chest X-ray images to assess three techniques' explanation quality further: Composite Layer-wise Relevance Propagation, Single Taylor Decomposition, and Deep Taylor Decomposition. We considered two well-known deep architectures for explanation: VGG11 and VGG16. Last but not least, three distinct quantitative measures were considered for comparison purposes: explanation continuity, explanation selectivity, and input perturbation.

We observe results that confirm some statements made by Montavon et al. [27], with VGG11 performing better than its counterpart with extra layers, i.e., VGG16. In general, Composite LRP achieved better results but was closely followed by Deep Taylor Decomposition. We understand that both approaches are suitable for explanation purposes if one takes into account the quantitative assessment. However, it seems that DTD highlights both lungs and the rib cage's boundaries, which does not seem to be a good choice. On the other hand, Composite LRP appears to highlight not only high-frequency regions, but others that seem to be relevant for COVID-19 automatic identification.

Concerning future works, we aim to evaluate other deep architectures such as ResNets, EfficientNets, and MobileNets. The latter models are pretty efficient and fast for training purposes, thus allowing us to retrain them whenever necessary.

Acknowledgments This work was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (5-21-01-001-0031).

Author's contribution MMH and JP were in charge of writing, coding, and paper reviewing. SAA and AA were in charge of paper reviewing, and research ideas.

Availability of data and material The data used in the manuscript are available publicly.

Code availability All source codes used in the manuscript were properly cited.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval It does not apply.

Informed consent It does not apply.

References

2. Abiodun KM, Awotunde JB, Aremu DR, Adeniyi EA (2022) Explainable ai for fighting covid-19 pandemic: opportunities, challenges, and future prospects. In: Computational intelligence for COVID-19 and future pandemics, pp 315–332. Springer
3. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L (2020) Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in china: a report of 1014 cases. Radiology, p 200642
4. Al-Waisy AS, Al-Fahdawi S, Mohammed MA, Abdulkareem KH, Mostafa SA, Maashi MS, Arif M, Garcia-Zapirain B (2020) COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images. Soft Computing
5. Alshazly H, Linse C, Barth E, Martinetz T (2021) Explainable COVID-19 detection using chest CT scans and deep learning. Sensors 21(2)
6. Aviles-Rivero AI, Sellars P, Schönlieb CB, Papadakis N (2022) Graphxcovid: explainable deep graph diffusion pseudo-labelling for identifying covid-19 on chest x-rays. Pattern Recogn 122:108274
7. Bassi PRAS, Attux R (2022) A deep convolutional neural network for covid-19 detection using chest x-rays. Res Biomed Eng 38(1):139–148
8. Bhattacharya S, Reddy Maddikunta PK, Pham QV, Gadekallu TR, Krishnan S, Chowdhary SR, Alazab CL, Jalil Piran M (2021) Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey. Sustain Cities Soc 65:102589
9. Brunese L, Mercaldo F, Reginelli A, Santone A (2020) Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. Comput Methods Prog Biomed 196:105608. https://doi.org/10.1016/j.cmpb.2020.105608. https://www.sciencedirect.com/science/article/pii/S0169260720314413
10. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp 839–847. IEEE
1. Clinical management of severe acute respiratory infection when novel coronavirus (2019-ncov) infection is suspected: interim guidance, 28 january 2020. Tech. rep., World Health Organization (2020). World Health Organization and others
11. Cohen JP, Morrison P, Dao L (2020) COVID-19 image data collection
12. Dansana D, Kumar R, Bhattacharjee A, Hemanth DJ, Gupta D, Khanna A, Castillo O (2020) Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm. Soft Comput. https://doi.org/10.1007/s00500-020-05275-y
13. DeGrave AJ, Janizek JD, Lee SI (2021) AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Mach Intell. https://doi.org/10.1038/s42256-021-00338-7
14. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W (2020) Sensitivity of chest ct for COVID-19: comparison to RT-PCR. Radiology, p 200432
15. Fuhrman JD, Gorre N, Hu Q, Li H, El Naqa I, Giger ML (2022) A review of explainable and interpretable ai with applications in covid-19 imaging. Med Phys 49(1):1–14
16. Gumaei A, Ismail WN, Hassan MR, Hassan MM, Mohamed E, Alelaiwi A, Fortino G (2022) A decision-level fusion method for covid-19 patient health prediction. Big Data Res 27:100287
17. Hassan MR, Hassan MM, Altaf M, Yeasar MS, Hossain MI, Fatema K, Shaharin R, Ahmed AF (2021) B5g-enabled distributed artificial intelligence on edges for covid-19 pandemic outbreak prediction. IEEE Netw 35(3):48–55
18. Hassan MR, Ismail WN, Chowdhury A, Hossain S, Huda S, Hassan MM (2022) A framework of genetic algorithm-based cnn on multi-access edge computing for automated detection of covid-19. J Supercomput 78(7):10250–10274
19. Hryniewska W, Bombiński P, Szatkowski P, Tomaszewska P, Przelaskowski A, Biecek P (2021) Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies
20. Hu Q, Gois FNB, Costa R, Zhang L, Yin L, Magaia N, de Albuquerque VHC (2022) Explainable artificial intelligence-based edge fuzzy images for covid-19 detection and identification. Appl Soft Comput 123:108966

21. Hu Q, Gois FNB, Costa R, Zhang L, Yin L, Magaia N, de Albuquerque VHC (2022) Explainable artificial intelligence-based edge fuzzy images for covid-19 detection and identification. Appl Soft Comput 123:108966
22. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506
23. Iwana BK, Kuroki R, Uchida S (2019) Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: 2019 IEEE/CVF international conference on computer vision workshop (ICCVW), pp 4176–4185. IEEE
24. Karim MR, Döhmen T, Cochez M, Beyan O, Rebholz-Schuhmann D, Decker S (2020) DeepCOVIDExplainer: Explainable COVID-19 diagnosis from chest x-ray images. In: 2020 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 1034–1037
25. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR
26. Liu T, Siegel E, Shen D (2022) Deep learning and medical image analysis for covid-19 diagnosis and prediction. Ann Rev Biomed Eng, 24
27. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn 65:211–222
28. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digital Signal Process 73:1–15
29. Ohata EF, Bezerra GM, das Chagas JVS, Neto AVL, Albuquerque AB, de Albuquerque VHC, Reboucas Filho PP (2020) Automatic detection of covid-19 infection using chest x-ray images through transfer learning. IEEE/CAA J Auto Sin 8(1):239–248
30. Parah SA, Kaw JA, Bellavista P, Loan NA, Bhat GM, Muhammad K, de Albuquerque VHC (2020) Efficient security and authentication for edge-based internet of medical things. IEEE Internet Things J 8(21):15652–15662
31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds.) Adv Neural Inf Process Syst 32:8024–8035. Curran Associates, Inc.
32. Pennisi M, Kavasidis I, Spampinato C, Schinina V, Palazzo S, Salanitri FP, Bellitto G, Rundo F, Aldinucci M, Cristofaro M, Campioni P, Pianura E, Di Stefano F, Petrone A, Albarello F, Ippolito G, Cuzzocrea S, Conoci S (2021) An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. Artif Intell Med, p 102114. https://doi.org/10.1016/j.artmed.2021.102114. https://www.sciencedirect.com/science/article/pii/S093336572100107X
33. Ribeiro MT, Singh S, Guestrin C (2016) "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16, pp 1135–1144. Association for Computing Machinery, New York, NY, USA
34. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
35. Samek W, Binder A, Lapuschkin S, Müller KR (2017) Understanding and comparing deep neural networks for age and gender classification. In: 2017 IEEE international conference on computer vision workshops, pp 1629–1638
36. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR (2017) Evaluating the visualization of what a deep neural network has learned. IEEE Trans Neural Netw Learn Syst 28(11):2660–2673
37. Santos CFG, Passos LA, Santana MC, Papa JP (2021) Normalizing images is good to improve computer-assisted covid-19 diagnosis. In: Kose U, Gupta D, de Albuquerque VHC, Khanna A (eds.) Data science for COVID-19, pp 51–62. Academic Press
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
39. Serte S, Demirel H (2021) Deep learning for diagnosis of covid-19 using 3d ct scans. Comput Biol Med 132:104306
40. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds.) 3rd International conference on learning representations, ICLR
41. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, Chen J, Wang R, Zhao H, Zha Y, Shen J, Chong Y, Yang Y (2021) Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with ct images. IEEE/ACM Trans Comput Biol Bioinf, pp 1–1. https://doi.org/10.1109/TCBB.2021.3065361
42. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks
43. Wang L, Lin ZQ, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 10(1):19549
44. Wang SH, Zhang Y, Cheng X, Zhang X, Zhang YD (2021) PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation. Comput Math Methods Med 2021:6633755 (**Publisher: Hindawi**)
45. Wu YH, Gao SH, Mei J, Xu J, Fan DP, Zhang RG, Cheng MM (2021) Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. IEEE Trans Image Process 30:3113–3126
46. Ye Q, Xia J, Yang G (2021) Explainable AI for COVID-19 CT classifiers: an initial comparison study
47. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 2921–2929

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.