Beyond the Black Box: Do More Complex Models Provide Superior XAI Explanations?

Assessing Deep Learning Model Complexity in the Evaluation and Interpretability of Medical X-ray Images

Mateusz Cedro, Marcin Chlebus
University of Warsaw, Faculty of Economic Sciences
mw.cedro@student.uw.edu.pl, m.chlebus@uw.edu.pl

Abstract

The increasing complexity of Artificial Intelligence (AI) models poses challenges to interpretability, particularly in domains like healthcare. This study investigates the impact of the complexity of deep learning models and Explainable AI (XAI) efficacy, utilizing four ResNet architectures (ResNet-18, 34, 50, 101). Through methodical experimentation on 4,369 lung X-ray images of COVID-19-infected and healthy patients, the research evaluates the models' classification performance and the relevance of corresponding XAI explanations with respect to the ground-truth disease masks. Results indicate that the increase in model complexity is associated with the decrease in classification accuracy and AUC-ROC scores (ResNet-18: 98.4%, 0.997, ResNet-101: 95.9%, 0.988). Notably, no statistically significant differences were observed in XAI quantitative metrics — Relevance Rank Accuracy and proposed Positive Attribution Ratio — across trained models in this study. These results suggest that increased model complexity does not consistently lead to higher performance or the relevance of explanations of models' decision-making processes.

Introduction

Deep Neural Networks (DNNs) have garnered substantial success across diverse domains of Artificial intelligence (AI) applications. Nonetheless, the opacity of their decision-making processes presents considerable challenges, particularly in critical sectors such as healthcare where transparency is essential (Morch et al. 1995; Baehrens et al. 2010; Simonyan et al. 2013). Efforts to reconcile the trade-off between model's accuracy and interpretability have led to the development of methods to trace predictions back to input features, enhancing model transparency (Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017; Arras, Osman, and Samek 2022; Molnar 2022).

In the medical domain, the demand for interpretable models is heightened due to the life-or-death implications of decisions made (Rajkomar, Dean, and Kohane 2019; Holzinger et al. 2019). Despite the intrinsic complexity of accurate machine learning models, healthcare professionals require comprehensible insights into how specific features influence predictions (Che et al. 2017; Topol 2019).

Artificial neural networks, particularly DNNs, are designed to emulate the complexity of biological systems, resulting in architectures that are not inherently transparent, thus casting these models as "black-box" devices (Guidotti et al. 2019; Ribeiro, Singh, and Guestrin 2016; Apley and Zhu 2020). Consequently, enhancing model explainability is a critical factor influencing the adoption of machine learning models in sensitive applications (Holzinger et al. 2017).

Interpretable and Explainable AI

The field of Interpretable Machine Learning (IML) and Explainable Artificial Intelligence (XAI) has risen in response to the need for transparency in deep learning models, attracting significant attention in machine learning research (Ribeiro, Singh, and Guestrin 2016; Samek et al. 2019; Molnar, Casalicchio, and Bischl 2020). Interpretability is defined by the ability of a machine learning system to make its processes and decisions understandable to human users. The quality of these explanations is crucial for model evaluation, validation, and debugging, and it's evaluated on the clarity of the model's decision-making process, not merely on prediction accuracy (Adebayo et al. 2020).

Interpretability is examined from two perspectives: global and local. Global interpretability provides an overarching understanding of the model's functioning and decision patterns, while local interpretability focuses on the specific rationale behind individual predictions. Both levels of interpretability are critical, serving varied purposes from enhancing scientific understanding to identifying biases and substantiating individual decisions (Doshi-Velez and Kim 2017; Molnar 2022).

The challenge of explaining how neural networks arrive at their predictions is central to XAI, where the goal is to map and quantify the influence of each input feature on the final decision. This is particularly valuable in medical settings where healthcare professionals benefit from understanding the model's reasoning (Samek, Wiegand, and Müller 2017; Adebayo et al. 2020). Despite their capabilities, many machine learning models remain opaque, acting as "black boxes" without revealing the rationale behind their decisions.

The attribution of a deep network's predictions to its input features has been identified as a central problem (Sundararajan, Taly, and Yan 2017). This attribution is represented as a vector, quantifying each input feature's contribution to the network's prediction, thereby clarifying the decision-making process, particularly beneficial for health-

care professionals in understanding the strengths and limitations of the model (Samek, Wiegand, and Müller 2017; Adebayo et al. 2020).

Techniques such as DeepLIFT (Shrikumar, Greenside, and Kundaje 2019), Layerwise Relevance Propagation (Bach et al. 2015), LIME (Ribeiro, Singh, and Guestrin 2016), and Integrated Gradients (Sundararajan, Taly, and Yan 2017) have been developed to unravel these decisions, breaking down the contributions of individual neurons to input features and thus advancing model interpretability. Moreover, in the GradientSHAP methodology, SHAP values (SHapley Additive exPlanations), have been adopted to attribute importance to each feature in a prediction, grounded in cooperative game theory, enhancing the interpretability of model features (Lundberg and Lee 2017; Shapley 1953).

Further refinements in interpretability methods, such as sensitivity and saliency maps, highlight influential image regions. Innovations like SmoothGrad and NoiseGrad have improved these techniques, reducing visual noise and integrating stochastic elements into models, which enhances both local and global interpretive clarity (Smilkov et al. 2017; Bykov et al. 2022).

The assessment of feature significance within Explainable AI is critical for model refinement and establishing trust in model predictions (Hooker et al. 2019). Challenges arise from the lack of universally accepted interpretability standards and the complexity involved in selecting and configuring appropriate interpretability methods (Kindermans et al. 2017; Arras, Osman, and Samek 2022).

Evaluating feature significance often involves analyzing the effects of feature removal on model performance. This method, while effective, can alter the evaluation data distribution and thus potentially compromise the assessment's validity (Bach et al. 2015; Sundararajan, Taly, and Yan 2017). With the growing necessity for transparent AI, objective and reproducible evaluation metrics are increasingly important (Ribeiro, Singh, and Guestrin 2016). One of the comprehensive frameworks for assessing the quality of the model explanations is Quantus (Hedström et al. 2023). This framework provides a comprehensive set of tools for accurate assessment of explanations and follows a transparent and impartial validation process for various XAI methodologies.

AI and XAI in Medicine

The integration of AI into healthcare is a strategic initiative aimed at personalizing patient treatment by harnessing the analytical prowess of AI to process and interpret large-scale clinical datasets (LeCun, Bengio, and Hinton 2015; Holzinger et al. 2019). Deep learning architectures, capable of sifting through extensive data such as hundreds of thousands of labelled X-ray images, are particularly instrumental in this shift from traditional rule-based diagnostics to a more nuanced, data-driven approach. This transition necessitates a framework within which the complex outputs of these models can be understood and trusted by medical professionals, a need met by the emerging field of Explainable AI (Cabitza, Rasoini, and Gensini 2017; Rajkomar, Dean, and Kohane 2019).

XAI in medicine not only seeks to illuminate the opaque decision-making processes of deep learning models but also strives to validate the reliability of AI-generated recommendations. The overarching goal is to establish a symbiotic relationship where AI systems are not merely tools for data extrapolation but partners in clinical decision-making, providing transparent and interpretable explanations that foster trust and facilitate informed medical judgments (Katuwal and Chen 2016; Che et al. 2017; Hinton 2018).

AI carries transformative economic implications, necessitating a balance between peak performance and operational efficiency (LeCun, Bengio, and Hinton 2015; Hinton 2018; Holzinger et al. 2019). The deepening of neural networks, while advancing capabilities, approaches a threshold beyond which additional layers yield minimal performance gains, as identified by Wu, Shen, and Hengel (2016) and Zhao et al. (2016). Contemporary advancements in complex architectural designs, such as Generative Pre-trained Transformers (GPT), have brought to the fore the significant financial and environmental costs inherent in the training processes. This development necessitates a judicious equilibrium between the advantages conferred by AI and the consumption of resources it entails (Brown et al. 2020; Menghani 2023).

In healthcare, the role of AI is especially critical as it offers the dual benefits of cost reduction and enhanced patient care. However, the adoption of AI must consider not just technological prowess but also the practicalities of application (Davenport and Kalakota 2019; Secinaro et al. 2021). This balance is crucial in ensuring that AI's integration into healthcare remains both efficient and beneficial, providing clear, interpretable outcomes that align with the overarching goals of medical practice.

The COVID-19 pandemic has accelerated the application of deep learning in medical diagnostics, with researchers leveraging transfer learning to address data limitations such as size and quality (Chowdhury et al. 2020; Degerli et al. 2021; Rahman et al. 2021). Studies utilizing ResNet architectures on COVID-19 datasets have yielded promising results, underscoring the potential of deep learning in aiding pandemic response (Showkat and Qureshi 2022). Furthermore, the use of saliency maps in medical image segmentation has provided visual explanations that enhance the interpretability of model predictions, essential for medical diagnostics (Saporta et al. 2022).

Influence of Model Scale on Performance and XAI Evaluations

In the field of machine learning, there is a common hypothesis that an increase in model capacity should correlate with enhanced training efficacy (Eigen et al. 2014). Nonetheless, this correlation is not absolute, as studies have shown variable performance benefits with the scaling of model complexity, particularly in ResNet architectures (Dauphin and Bengio 2013; Wu, Shen, and Hengel 2016). While deeper networks such as ResNet-50 have demonstrated improvements in specific tasks, they do not universally outperform across all scenarios, with instances where less complex models like ResNet-18 match or exceed the accuracy of their larger counterparts (Khan et al. 2018; Sarwinda et al. 2021).

The concept of diminishing returns becomes evident as network complexity increases beyond a certain threshold, resulting in marginal performance enhancements that do not justify the additional complexity (Eigen et al. 2014; Wu, Shen, and Hengel 2016). In particular, ResNet-18 has been noted for its competitive performance against more elaborate models in certain classification tasks, prompting a reevaluation of the efficacy of scaling up network depth (Guo and Du 2019). These observations underscore the imperative for a strategic approach in model selection that weighs computational efficiency against the specific performance requirements of the given task, thereby optimizing the balance between model architecture size and functional output.

To the best of our knowledge, no previous research has explored the relationship between the complexity of deep learning model architectures and the quality of XAI explanations. Our study is the first to address the problem in the literature by conducting experiments to investigate this relationship. In sectors where transparency is paramount, such as healthcare, understanding how architectural complexities affect both model performance and the interpretability of XAI explanations becomes crucial. By conducting methodical experiments, this study aims to gain in-depth insight into the relationship between the complexity of deep learning models and the greatest possible interpretability, ultimately aiming to increase the accuracy and reliability of XAI explanations. Therefore, this study proposed two hypotheses.

Hypothesis 1: As the model's complexity increases, characterized by a greater number of trainable parameters, it exhibits better classification performance.

Hypothesis 2: As the model's complexity increases, characterized by a greater number of trainable parameters, XAI assessment indicators are anticipated to yield inferior results, indicating an increased challenge in elucidating the underlying decision-making process.

Methodology

To answer the underlying question, of whether more complex architectures provide better explainability in image classification tasks, in the conducted research the same workflow was employed for all of the trained ResNet models (ResNet-18, ResNet-34, ResNet-50, and ResNet-101). Initially, each ResNet model was trained from scratch, utilizing a consistent subset of randomly assigned images and model hyper-parameters to ensure equitable training conditions across all architectures.

After the training phase, a focused exploration into model explainability was undertaken by generating XAI explanations for each trained model, employing the Quantus library (Hedström et al. 2023). Three XAI methodologies were leveraged: Saliency Maps (Morch et al. 1995; Simonyan et al. 2013), GradientShap (Lundberg and Lee 2017), and Integrated Gradients (Sundararajan, Taly, and Yan 2017), each providing distinct perspectives into model decision-making processes.

The derived explanations were then subjected to a quantitative evaluation utilizing two pertinent metrics: Relevance Rank Accuracy (Arras, Osman, and Samek 2022) and proposed in this paper Positive Attribution Ratio, providing insightful revelations regarding the reliability and interpretability of the explanations propagated by each model. Having this approach, the following experiment provides a clear evaluation of the models' behaviour in the conducted image classification task.

Data

The dataset used in the following experiment was the COVID-QU-Ex dataset formulated by researchers from Qatar University and the University of Dhaka, which is a collection of the X-rays lung images obtained from various resources (Tahir et al. 2021). The dataset contains three groups of X-rays: COVID-19 pneumonia, other diseases (non-covid), and healthy patients' lungs. For the X-rays from COVID-QU-Ex, corresponding ground-truth masks from the QaTa-COV19 dataset were used. QaTa-COV19 dataset was developed by Qatar University and Tampere University which provides binary segmentation masks of COVID-19 pneumonia (Chowdhury et al. 2022).

For the following experiment, 4369 X-ray lung images of different patients and corresponding ground truth masks were used. 2,913 images were labeled as *COVID-19* infected and 1,456 as *Healthy*, non-infected patients. For training, validation and testing, X-ray images were randomly split on 70%, 20%, and 10% dataset fractions respectively.

Before training, all the images were resized to the size of 224x224 pixels, turned into grayscale, transformed into Tensors, and normalized. Transformations were done with the use of PyTorch's Torchvision library (TorchVision maintainers and contributors 2016).

Models

In the experiment four Residual Network (ResNet) architectures were explored, each distinguished by its depth: ResNet-18, ResNet-34, ResNet-50, and ResNet-101 (He et al. 2015). Recognized for effectively addressing challenges in training deep networks for image classification tasks, these architectures were selected to probe the relationship between network depth and performance. Figure 1 shows a building block containing the residual connection that provides an *identity* input to every other layer, which becomes a state-of-the-art building block of deep learning architectures. Figure 3 presents a ResNet-34 architecture in comparison with plain 34-layer deep learning architecture (He et al. 2015).

Baseline performance was established using ResNet-18 and ResNet-34, which were chosen for their balance of predictive power and computational efficiency. In contrast, ResNet-50 and ResNet-101 were scrutinized for potential accuracy improvements, despite their increased computational demands and complexity. A uniform training and testing process was applied to all models to ensure a fair comparison, and the trade-offs between model size, computational demand, and predictive accuracy were elucidated in the context of our research.

In Table 1, the number of trainable parameters for various ResNet models is presented. Each consequent ResNet

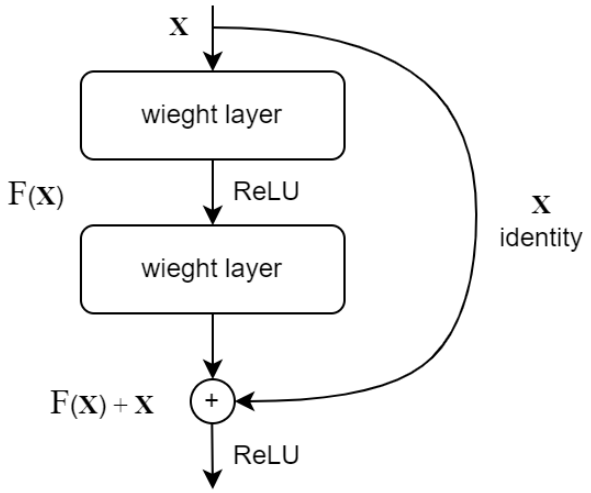

Figure 1: Residual learning: a building block.
Source: Own preparation based on He et al. (2015).

Model	Number of Trainable Parameters
ResNet-18	6,139,842
ResNet-34	12,329,218
ResNet-50	23,532,418
ResNet-101	42,550,658

Table 1: Number of Trainable Parameters in ResNet Models.
Source: Own calculations.

model has approximately double the number of the trainable parameters of the former model.

Model Training Setup

In the research all ResNet architectures (ResNet-18, ResNet-34, ResNet-50, and ResNet-101) were trained from scratch for the image classification tasks. From the original dataset, the X-rays labelled as other diseases (non-covid) were excluded, leaving a dataset categorized into two label groups: *COVID-19* and *healthy*. Under the aforementioned approach, all models conducted binary classification tasks.

A concerted approach was employed to ensure the coherent training, validation, and testing of all models, with the images being randomly partitioned into respective subgroups comprising 70%, 20%, and 10% of the data that contained in total 4,369 images subdivided into 2,913 and 1,456 images of *COVID-19* and *healthy* groups respectively.

The models were developed using the PyTorch library (Paszke et al. 2019) and utilized a Cross-Entropy Loss criterion. This criterion computes the cross-entropy loss between predicted and target class labels, facilitating the models' learning from the logits.

The optimization of the model parameters was undertaken using Stochastic Gradient Descent (Zinkevich et al. 2010) with a learning rate and momentum of 0.001 and 0.9, respectively. All models were subjected to the training for 50 epochs, with a batch size of 64, to gauge their efficacy in distinguishing between the defined label groups under consistent hyper-parameter settings (Bertrand 2019).

Ensuring experimental reproducibility and consistency across all training sessions, the random seeds for PyTorch and NumPy were fixed at a value of 42 (Chen et al. 2022).

All model training sessions and subsequent Explainable Artificial Intelligence analyses were conducted utilizing the Nvidia A100 GPU with 40 GB of RAM capacity. This computational environment was facilitated through the virtualized infrastructure provided by Google Colab Pro+.

Gradient-base techniques

In the field of machine and deep learning, gradients are defined as the rate of change of the output with respect to the input and are acknowledged for their pivotal roles beyond mere optimization. Historically, the product of model coefficients with feature values has been examined by practitioners to interpret simpler, usually linear models. In deep neural networks, gradients are perceived as intrinsic coefficients, signifying the intricate connection between input and output (Baehrens et al. 2010; Simonyan et al. 2013). With advancements in research, gradient-based techniques have been introduced in the realm of explainable artificial intelligence, enabling a more profound interpretation of model behaviour and given prediction.

In this section, three gradient-based methods are outlined, specifically Saliency Maps (Morch et al. 1995; Baehrens et al. 2010; Simonyan et al. 2013), GradientShap (Lundberg and Lee 2017), and Integrated Gradients (Sundararajan, Taly, and Yan 2017). In the Saliency Maps method, the derivative of the class score with respect to the input image is calculated, identifying pixels that, when slightly altered, are found to have the most significant influence on the class score. Subsequently, the GradientShap method synthesizes Shapley values and gradients, further enriching our understanding of model predictions. Lastly, the Integrated Gradients method is presented, wherein the path integration between input and output is detailed, providing a comprehensive attribution explanation.

A thorough examination of these gradient-based methodologies is undertaken in this chapter, highlighting their roles in augmenting the interpretability and transparency of deep neural architectures.

X-rays of both healthy and COVID-19-infected lungs, along with their respective ground-truth masks and pixel attribution maps, are presented in Figure 4 and Figure 5.

Saliency Maps One of the pioneering methodologies in the realm of explainable artificial intelligence is denoted as *saliency maps* (Morch et al. 1995; Simonyan et al. 2013), which delineates the significance of specific components, such as pixels on the image, with respect to the observed empirical relationships.

Given the inherent nonlinearity of the models with complex architecture, straightforward interpretations become elusive (Simonyan et al. 2013). In this context, the saliency maps serve as an instrumental visualization mechanism, highlighting regions within the image that exhibit strong

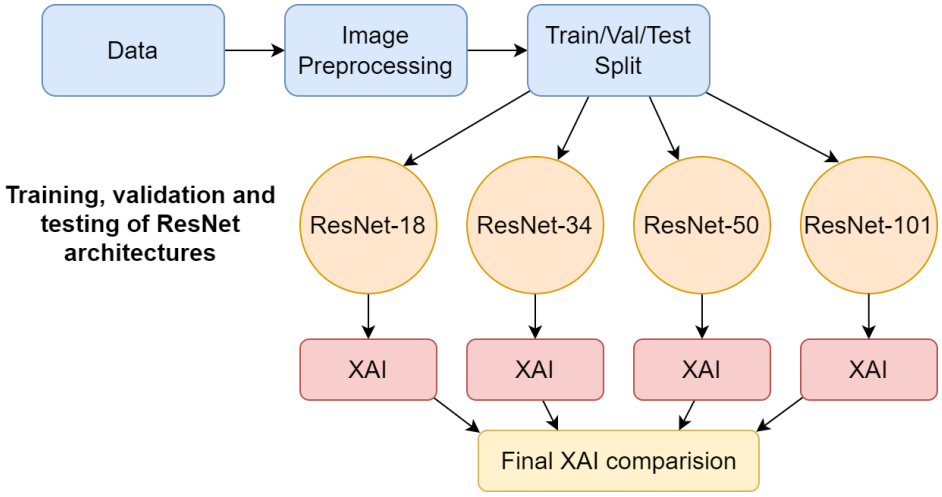

Figure 2: Comprehensive Workflow Schema of the Research Methodology.
Source: Own preparation.

correlations to distinct tasks. By employing this technique, a transition from the high-dimensional input data space to a substantially reduced vector of projections is facilitated. This process inherently engenders profound weight sharing, underlined by associations amongst weights interfacing the input and hidden layers of the feed-forward neural architecture, as Convolutional Neural Networks are. The saliency attributed to an input channel (for instance, the pixel i of an image vector) is quantified by the noticeable alteration in the cost function upon its exclusion.

In the research presented in *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps* (Simonyan et al. 2013), a gradient-based technique was introduced to compute an image-specific class saliency map, tailored to a distinct image and class combination. This method was harnessed using classification ConvNets and was designed to identify and highlight the spatial significance of a specific class within a given image. Essentially, for a given image I_0 and its associated class c, the pixels of I_0 were ranked based on their impact on the class score $S_c(I_0)$. Given the intricate non-linearity of deep ConvNets, the class score $S_c(I)$ was approximated linearly in the vicinity of I_0 using the first-order Taylor expansion. Through this approach, pixels that could be altered minimally to most influence the class score were elucidated.

The procedure involved first determining the derivative w via back-propagation. Subsequently, the saliency map was extracted by reorganizing the components of vector w. For grey-scale images, the dimensions of w were found to align with the pixel count of I_0, allowing for the map's computation as $M_{ij} = |w_{h(i,j)}|$, where $h(i, j)$ denoted the index of w that corresponded to the pixel situated in the i-th row and j-th column. Notably, this saliency map derivation utilized a classification ConvNet, trained exclusively on image labels, thereby eliminating the need for supplemental annotations, such as bounding boxes or segmentation masks.

These saliency maps, inherently weakly supervised, were found to encode the object's location of the assigned class within the image. As a result, they showed notable potential for object localization tasks, despite their training being restricted solely to image labels. To summarize, the framework proposed by Simonyan et al. (2013) explains a technique for formulating image-specific saliency maps, emphasizing segments within the image characteristic of the pertinent class.

GradientShap The problem with the evolving trend of developing more and more complex deep learning architectures to achieve higher accuracy and model precision, models become hardly interpretable forcing unwanted tradeoffs between the accuracy and the model interpretability for end-users. To address that problem Lundberg and Lee (2017) proposed an explanation framework named SHAP - SHapley Additive exPlanations. SHAP bases its feature interpretability on the concept from cooperative game theory (Shapley 1953) by allocating an importance value to each feature for a specific prediction.

In research on model interpretability, it is commonly addressed that a simple model acts naturally as its own best explanation, eliminating the need for additional clarifications (Lundberg and Lee 2017). However, for complex models like deep neural network architectures, the original model is not inherently interpretable. Thus, a more straightforward, interpretable model approximation or the *explanation model* is needed. Consider denoting the original model as f and the *explanation model* as g. Explanation models typically employ simplified inputs, x_0, which correlate to the original inputs via a transformation function, $x = h_x(x_0)$. The objective of local methods is to ensure that $g(z_0)$ closely mirrors $f(h_x(z_0))$ whenever z_0 is akin to x_0.

Shapley values. In the domain of cooperative game theory, the Shapley value stands out as a pivotal mechanism designed to equitably allocate gains and costs among various participants within a coalition (Shapley 1953). This concept, originally formulated by Lloyd Shapley, becomes indispensable in scenarios where distinct actors contribute

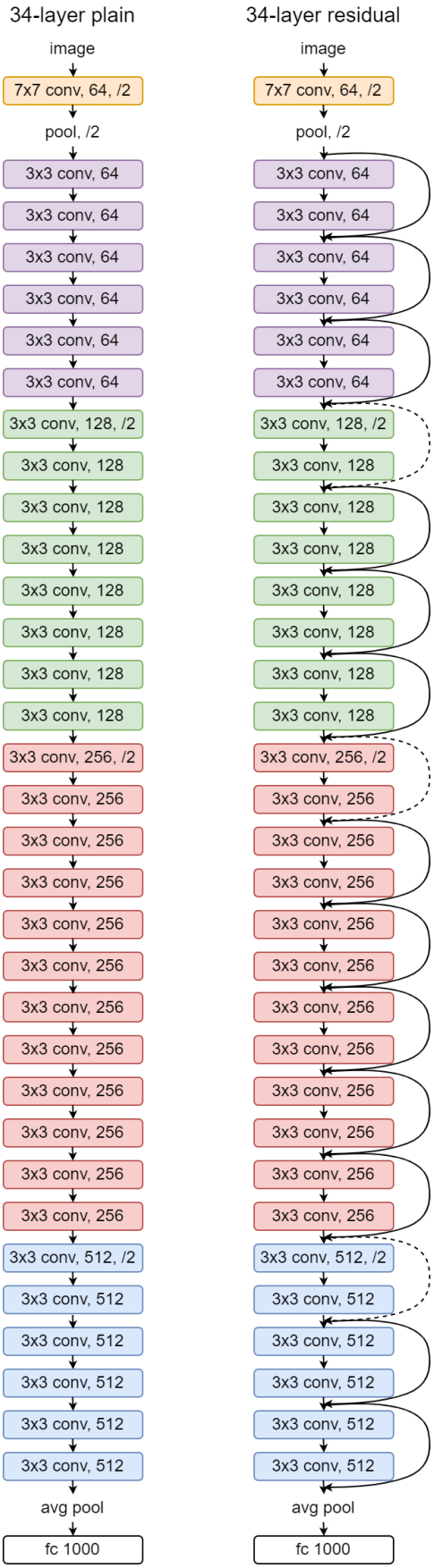

Figure 3: Computer Vision network architectures. Left: 34-layer plain network. Right: 34-layer residual network. Dotted lines represent dimension-expanding connections. Source: Own preparation based on He et al. (2015).

unequally yet collaborate towards a shared objective. The central premise of the Shapley value is to guarantee that each participant receives a payoff commensurate with their contribution, ensuring it is not less than what they would achieve independently. To elucidate, within a strategic game involving multiple players aiming for a specific outcome, the Shapley value quantifies the average marginal contribution of each player, after considering all feasible combinations.

In a machine learning framework, the traditional players of the cooperative game are analogously represented by the features inherent to the machine learning model, with the model's output serving as a corollary to the game's payoff (Merrick and Taly 2019). Shapley values offer a perspective on feature importance within linear models, particularly when multicollinearity is present. The application of this method necessitates the retraining of the model for all feature subsets $S \subseteq F$, where F denotes the complete set of features. Each feature has assigned an importance value, representing its impact on the model prediction when included. To determine this impact, one model, $f_{S\cup\{i\}}$, incorporates the particular feature, while the other, f_S, excludes it. The predictions of these two models are subsequently contrasted based on the current input: $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, wherein x_S symbolizes the values of the input features contained within set S. Given that the ramifications of omitting a feature are influenced by the model's other features, the aforementioned differences are evaluated across all feasible subsets $S \subseteq F \setminus \{i\}$. Subsequent calculations yield the Shapley values, then formally, the contribution ϕ of model feature i is defined as:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{1}$$

Conceptually, the Shapley value quantifies the average contribution of a specific feature i, by evaluating the incremental payoff introduced by i across all possible coalitions that exclude player i.

SHAP values. SHAP values are proposed as a unified measure of feature importance, representing the Shapley values of a conditional expectation function of the original model (Lundberg and Lee 2017). These values are attributed to each feature, reflecting the change in the expected model prediction upon conditioning on that particular feature. The transition from the base value $E[f(z)]$ — which would have been predicted in the absence of any known features — to the current output $f(x)$ is elucidated by SHAP values.

The unique additive feature importance measure that adheres to several properties is provided by SHAP values. These properties encompass:

Local accuracy — ensuring that the explanation model $g(x_0)$ corresponds with the original model $f(x)$ when $x = h_x(x_0)$;

Missingness — where features with $x_{0i} = 0$ are constrained to have no attributed impact;

Consistency — which mandates that if a model's alteration causes a simplified input's contribution to either increase or remain unchanged irrespective of other inputs, the

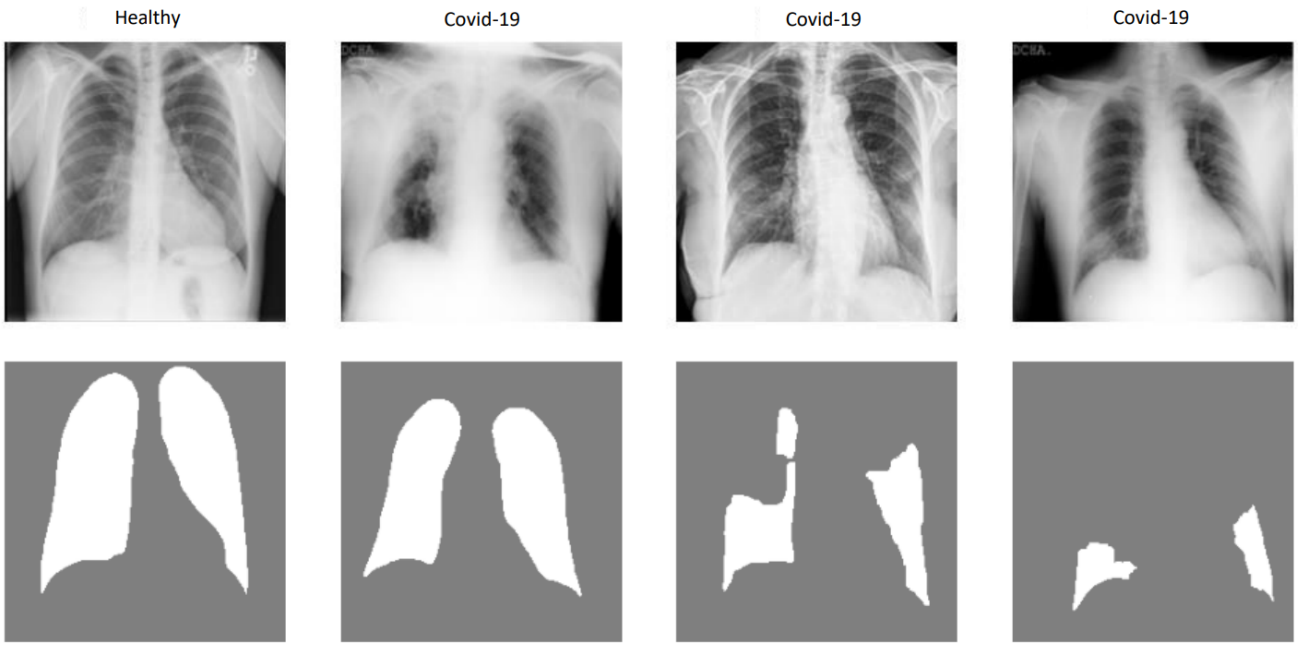

Figure 4: X-rays of *healthy* and *COVID-19* infected lungs, with corresponding ground-truth masks. Masks for healthy individuals encompass the entire lungs. In contrast, for COVID-19 patients, masks delineate areas identified by radiologists as diseased, potentially covering specific regions or the entirety of the lungs.
Source: Own preparation.

attribution of that input should not diminish.

Conditional expectations are utilized to define simplified inputs within these values. Inherent in the SHAP value definition is a simplified input mapping, denoted as $h_x(z_0) = z_S$, where z_S contains missing values for features absent in set S. Owing to most models' inability to process arbitrary patterns of missing input values, $f(z_S)$ is approximated with $E[f(z)|z_S]$. This definition of SHAP values is structured to closely resonate with the foundational Shapley values (Shapley 1953; Lundberg and Lee 2017).

GradientShap. GradientShap method estimates SHAP values by evaluating the gradient expectations, achieved by random sampling from a baseline distribution. By introducing white noise to input samples multiple times, it randomly selects a baseline and an intermediate point between the baseline and input, then calculates the gradient with respect to these random points. The resulting SHAP values mirror the expected values of these gradients multiplied by the difference between inputs and baselines.

The underlying assumption with GradientShap presumes that input features are independent and the explanation model is linear, indicating that the interpretations are modelled using the additive composition of feature effects. However, if the model exhibits non-linearity or the input features lack independence, the sequence in which features are incorporated into the expectation becomes significant. Under these circumstances, SHAP values are derived by averaging the Shapley values across all conceivable sequences. Given these conditions, the SHAP value can be approximated by the expected gradients computed for randomly generated samples, after Gaussian noise has been added to each input across various baselines.

Integrated Gradients The problem addressed in *Axiomatic Attribution for Deep Networks* publication (Sundararajan, Taly, and Yan 2017) concerned the issue that many previous gradient base methods broke at least one of the two axioms that should always be satisfied in feature attribution methods, namely *sensitivity* and *implementation invariance* axioms. To address this problem, the *Integrated Gradients* method was presented.

The Integrated Gradients approach has emerged as a notable solution in the realm of deep neural network interpretation. Rooted from an axiomatic framework inspired by economics literature, Integrated Gradients seeks to fulfil both *sensitivity* and *implementation invariance* axioms. This ensures that the computed attributions are not just artefacts of the method but genuinely reflect the network's behaviour (Sundararajan, Taly, and Yan 2017).

An attribution technique adheres to the *sensitivity* criterion when, for any input and baseline differing in just one feature with different predictions, the differing feature receives a non-zero attribution. consequently, if the deep network's function exhibits no mathematical dependence on a particular variable, that variable's attribution is always zero. In practical terms, the absence of sensitivity can lead to gradients predominantly concentrating on irrelevant features.

Within the context of neural networks, two architectures are deemed functionally equivalent when they produce con-

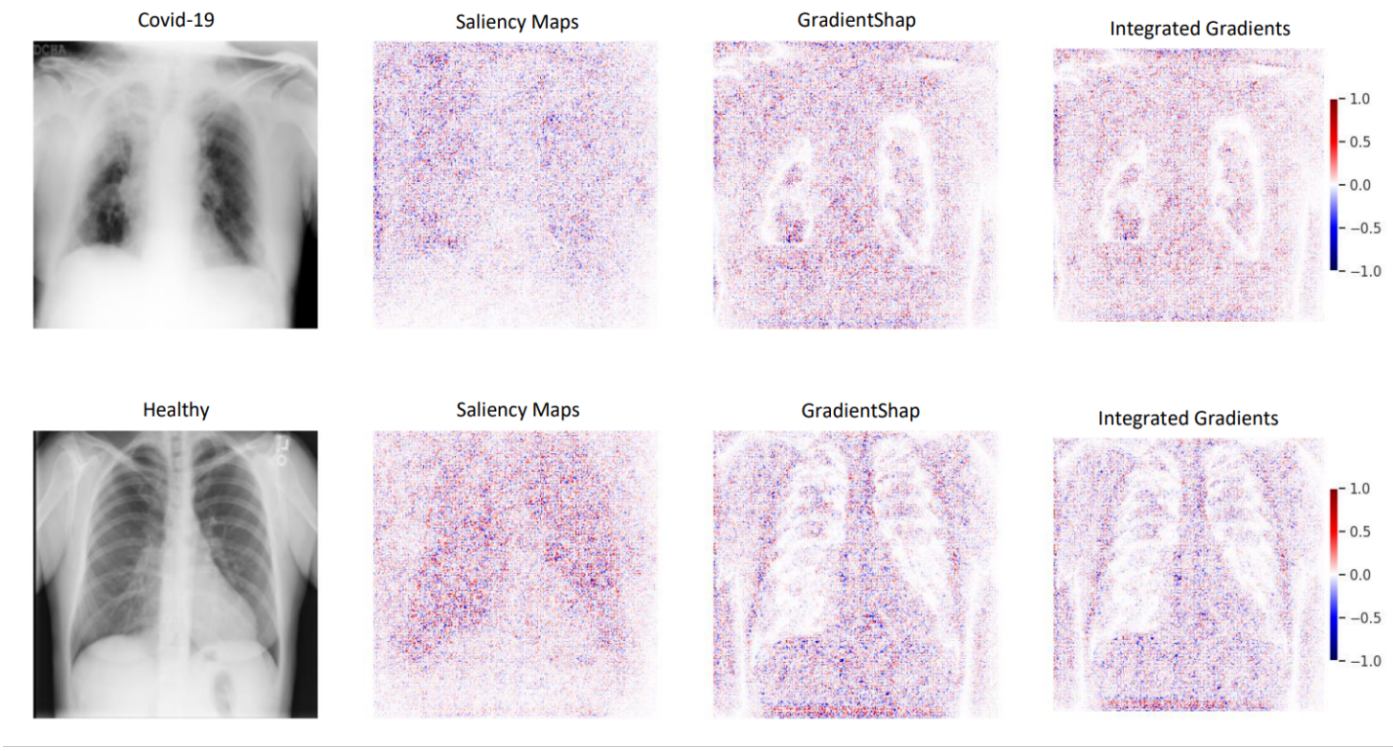

Figure 5: X-rays of *healthy* and *COVID-19*-infected lungs, accompanied by gradient-based attributions from Saliency Maps, GradientShap, and Integrated Gradients methodologies.
Source: Own preparation with the use of Quantus library.

sistent outputs across all given inputs, even if their internal implementations differ considerably. For attribution techniques, it is essential to adhere to the principle of *implementation invariance*. This principle ensures that the attributions remain consistent for networks that are functionally equivalent, regardless of their distinct structures.

In the Integrated Gradients method, gradients are systematically integrated between a designated baseline, usually a black image, and the actual input image. This technique identifies the presence or absence of distinct features, thereby highlighting the significance of specific pixels or features within the contextual framework. It is commonly called a path-attribution technique (Molnar 2022). Critically, Integrated Gradients is deemed a *complete* path-attribution approach. This implies that the cumulative relevance scores across all input features equate to the disparity between the prediction derived from the actual image and that of the reference image. In computer vision applications, pixel-wise attributions are presented, highlighting the areas of an image that resulted in the model's decision-making process.

Evaluation Metrics

In the context of machine learning interpretability, ensuring rigorous evaluation of explanatory heatmaps is crucial for computer vision models, especially when discerning model relevance. Therefore, to evaluate all the XAI approaches that have been used in this research study, *Relevance Rank Accuracy* (Arras, Osman, and Samek 2022) and proposed in this paper *Positive Attribution Ratio* metrics were used. Since both metrics are calculating the ratios, their values fall within the [0-1] range, with a higher score signifying a more precise relevance heatmap.

Relevance Rank Accuracy The Relevance Rank Accuracy is defined to gauge the degree to which the most pronounced relevance points are aligned with the ground truth. First, K is determined, representing the size of the ground truth mask. Then, the top K relevance values are extracted. Afterwards, the number of these values that correspond to locations within the ground truth is counted. This count is subsequently normalized by the dimension of the ground truth mask. Formally, this procedure can be expressed as:

$$P_{\text{top}K} = \{p_1, p_2, \ldots, p_K \mid R_{p_1} > R_{p_2} > \ldots > R_{p_K}\} \quad (2)$$

where $P_{\text{top}K}$ represents the set of pixels, each associated with relevance values $R_{p_1}, R_{p_2}, \ldots, R_{p_K}$, arranged in descending order up to the K-th pixel. Subsequently, the rank accuracy is determined as:

$$\text{Rank Accuracy} = \frac{|P_{\text{top}K} \cap GT|}{|GT|} \quad (3)$$

where GT represents the set of pixel positions contained within the ground truth mask, and $|GT|$ denotes the total count of pixels within this mask.

Positive Attribution Ratio The Positive Attribution Ratio derives its foundation from the Relevance Mass Accuracy as outlined by (Arras, Osman, and Samek 2022). Nevertheless, a pivotal distinction exists, since it solely operates on pixels that possess positive attribution. We believe that for the future end-user, it is more important to be informed about the ratio of the number of pixels that have positive attribution localized inside the ground truth mask with respect to all positively attributed pixels on the investigated image.

The Relevance Mass Accuracy is calculated by dividing the aggregated sum of relevance values located within the ground truth mask by the total relevance values across the entire image. Essentially, this metric evaluates the proportion of the explanation method's "mass" attributed to the pixels within the ground truth. Conversely, the Positive Attribution Ratio operates in a similar manner but focuses solely on pixels with positive attributions. As such, the Positive Attribution Ratio gauges the portion of positive attributions within the ground truth mask R_{within} in relation to the positive attributions across the entire image R_{total}. This might be formally represented as:

$$\text{Positive Attribution Ratio} = \frac{R_{\text{within}}}{R_{\text{total}}} \tag{4}$$

where

$$R_{\text{within}} = \sum_{\substack{k=1 \\ \text{s.t. } p_k \in \text{GT}}}^{|\text{GT}|} R_{p_k}, \forall R_{p_k} > 0 \tag{5}$$

and

$$R_{\text{total}} = \sum_{k=1}^{N} R_{p_k}, \forall R_{p_k} > 0 \tag{6}$$

where R_{p_k} denotes the relevance value corresponding to pixel p_k which has positive relevance attribution, GT encompasses pixel locations present within the ground truth mask, $|\text{GT}|$ signifies the count of pixels within this mask, and N stands for the overall pixel count in the image.

Experiments and Results

Models' Performance

Performance of the each ResNet model in terms of accuracy, AUC-ROC and Cross-Entropy Loss metrics on the separated test set is presented in Table 2. The ResNet-18 architecture achieved the highest accuracy of 98.4% and an AUC-ROC of 0.997, alongside maintaining the lowest cross-entropy loss of 0.066, misclassifying only 7 out of 437 X-ray images in the hold-out test set. Although all models demonstrated high accuracies and AUC-ROC values exceeding 95.9% and 0.988 respectively, an inverse relationship was noted between model complexity and performance metrics, with ResNet-101 registering the lowest accuracy and AUC-ROC scores in the series. These findings are consistent with the results reported by Guo and Du (2019).

This evaluation underscores the criticality of considering the trade-off between model complexity and predictive performance in the selection of suitable deep learning architectures for image classification.

Model	Accuracy (%)	AUC-ROC	Cross-Entropy Loss
ResNet-18	98.4	0.997	0.066
ResNet-34	97.3	0.996	0.097
ResNet-50	96.1	0.995	0.168
ResNet-101	95.9	0.988	0.153

Table 2: Performance Metrics of ResNet Models on Test Set.
Source: Own calculations.

Class	Model	Saliency Maps Mean (SD)	Gradient SHAP Mean (SD)	Integrated Gradients Mean (SD)
COVID-19	ResNet-18	**0.199** *(0.10)*	0.116 *(0.10)*	0.118 *(0.10*
	ResNet-34	0.198 *(0.10)*	0.117 *(0.10)*	0.115 *(0.10)*
	ResNet-50	0.191 *(0.11)*	0.115 *(0.10)*	0.116 *(0.10)*
	ResNet-101	0.175 *(0.12)*	**0.118** *(0.09)*	**0.119** *(0.09)*
Healthy	ResNet-18	0.303 *(0.05)*	0.248 *(0.07)*	**0.251** *(0.07)*
	ResNet-34	**0.305** *(0.05)*	**0.249** *(0.07)*	0.248 *(0.07)*
	ResNet-50	0.301 *(0.05)*	0.243 *(0.07)*	0.248 *(0.07)*
	ResNet-101	0.290 *(0.06)*	0.248 *(0.08)*	0.249 *(0.07)*

Table 3: Mean and Standard Deviation Scores for Relevance Rank Accuracy.
Source: Own calculations.

Results

The quantitative evaluations of all ResNet architectures, utilizing both the Relevance Rank Accuracy and Positive Attribution Ratio metrics, are meticulously detailed in Table 3 and Table 4 respectively. These evaluations incorporated the aforementioned XAI methodologies: Saliency Maps, GradientShap, and Integrated Gradients. The same interpretative methodologies were uniformly implemented across four ResNet models and evaluated independently on a test set consisting of 292 X-ray images labelled as *COVID-19* class and 145 X-ray images *Healthy* class.

In the context of the class *COVID-19*, clear fluctuations in performance indicators were evident. For the Relevance Rank Accuracy metric, ResNet-18 registered the highest mean score of 0.199 (*SD=0.1*) when analyzed through the Saliency Maps approach. Conversely, the application of GradientSHAP and Integrated Gradients methodologies resulted in the highest scores of 0.118 (*SD=0.09*) and 0.119 (*SD=0.09*), respectively, which were attributed to the ResNet-101 architecture.

In the evaluation of the *Healthy* class, the Relevance Rank

Class	Model	Saliency Maps Mean (SD)	Gradient SHAP Mean (SD)	Integrated Gradients Mean (SD)
COVID-19	ResNet-18	**0.186** *(0.12)*	0.117 *(0.10)*	**0.120** *(0.10)*
	ResNet-34	0.185 *(0.12)*	0.118 *(0.10)*	0.118 *(0.10)*
	ResNet-50	0.182 *(0.12)*	0.116 *(0.10)*	0.118 *(0.10)*
	ResNet-101	0.169 *(0.13)*	**0.120** *(0.10)*	0.119 *(0.10)*
Healthy	ResNet-18	0.308 *(0.07)*	0.250 *(0.08)*	**0.253** *(0.08)*
	ResNet-34	**0.315** *(0.07)*	0.255 *(0.08)*	0.250 *(0.08)*
	ResNet-50	0.304 *(0.07)*	0.242 *(0.07)*	0.245 *(0.07)*
	ResNet-101	0.292 *(0.08)*	**0.263** *(0.09)*	0.252 *(0.08)*

Table 4: Mean and Standard Deviation Scores for Positive Attribution Ratio.
Source: Own calculations.

Accuracy metric exposed varying performance paths. The ResNet-34 architecture, when interfaced with the Saliency Maps methodology, achieved an exemplary mean score of 0.305 (*SD=0.05*). Contrarily, when subjected to the GradientSHAP and Integrated Gradients methodologies, mean scores of 0.249 (*SD=0.07*) and 0.251 (*SD=0.07*) were predominantly associated with ResNet-34 and ResNet-18 architectures, respectively.

Referring to the Positive Attribution Ratio Scores within the *COVID-19* group, ResNet-18 emerged preeminent with a mean score of 0.186 (*SD=0.12*) and 0.120 (*SD=0.1*), attributable to the Saliency Maps and Integrated Gradients methodologies. Concurrently, the ResNet-101 configuration demonstrated its ascendancy in the Gradient SHAP, registering the highest mean score of 0.120 (*SD=0.1*).

For the *Healthy* class, under the Positive Attribution Ratio metric, ResNet-34's proficiency was salient with the highest mean score of 0.315 (*SD=0.07*) using the Saliency Maps approach. In a contrastive scenario, the GradientSHAP methodology witnessed ResNet-101 achieving a paramount mean score of 0.263 (*SD=0.09*), while ResNet-18 achieved a mean score of 0.253 (*SD=0.08*) within the Integrated Gradients approach.

To assess the statistical significance of the differences between the scores obtained from each of the ResNet architectures (18, 34, 50, and 101) for both metrics (Relevance Rank Accuracy and Positive Attribution Ratio), a procedure of statistical analyses was executed. These evaluations spanned across each of the three XAI methodologies (Saliency Maps, GradientShap, and Integrated Gradients) and were further bifurcated based on two distinct subgroups: *COVID-19* and *Healthy*. The analyses for the *COVID-19* subgroup were conducted utilizing a set of 292 X-ray images, whereas the *Healthy* class was assessed based on 145 X-rays. Both subgroups utilized images from the test set.

The analysis sequence was initialized with a one-way ANOVA, which provided a preliminary insight into the variance between group means. Given the potential non-parametric distribution of the data, the Kruskal-Wallis test was subsequently employed as a non-parametric alternative to the one-way ANOVA. This test facilitated the detection of differences in the medians among the ResNet models. To clarify these differences, pairwise comparisons among the four ResNet models were performed using the Mann-Whitney U test. Considering the risk of type I errors due to multiple comparisons, the p-values obtained were adjusted using the Bonferroni correction method. Through this thorough statistical approach, a clear understanding of performance disparities across distinct model architectures with specific XAI techniques and image classes was achieved.

Upon analyzing the undertaken dataset, several observations emerge. Notably, ResNet-50 did not attain the top performance in either Relevance Rank Accuracy or Positive Attribution Ratio metric. Meanwhile, ResNet-18 secured the highest scores in five out of the twelve evaluated instances. ResNet-101 achieved the highest score in four out of the twelve instances, and lastly, ResNet-34 secured the top scores in three of the twelve evaluations.

From the derived observations, it becomes clear that there is no direct correlation between the size or complexity of the ResNet model architecture and the resultant performance metrics like accuracy or AUC-ROC, aligning with the findings reported by Khan et al. (2018); Guo and Du (2019); Sarwinda et al. (2021). In terms of XAI quantitative metrics results, while ResNet-18 often displayed superior results, ResNet-50 did not necessarily follow suit, despite its increased complexity. Conversely, in certain scenarios, both ResNet-101 and ResNet-34 demonstrated superior performances, surpassing the results achieved by ResNet-50. Hence, it is imperative to understand that the choice of model architecture should not be solely based on its size or complexity. The results emphasize the importance of context-specific evaluations and suggest that in the domain of explainable AI for medical imaging, no one-size-fits-all approach is suitable.

The statistical analysis reveals that the sole statistically significant divergence in means across the ResNet architectures, at a threshold of $p < 0.05$, is discerned within the Relevance Rank Accuracy metric for the *COVID-19* category, only for the Saliency Maps methodology ($p = 0.03$). Subsequent application of the non-parametric Kruskal-Wallis test, assessing differences in medians, corroborated this significance, yielding a p-value of 0.02 for the same category. Extended analysis utilizing the Mann-Whitney U Test elucidated a statistically significant distinction between the outcomes for the ResNet-18 and ResNet-101 architectures, with $p = 0.03$. Furthermore, a marginal approach to significance within the same group and approach was observed between the ResNet-34 and ResNet-101 models, registering a p-value of 0.053.

In contrast, the remaining comparisons failed to evince

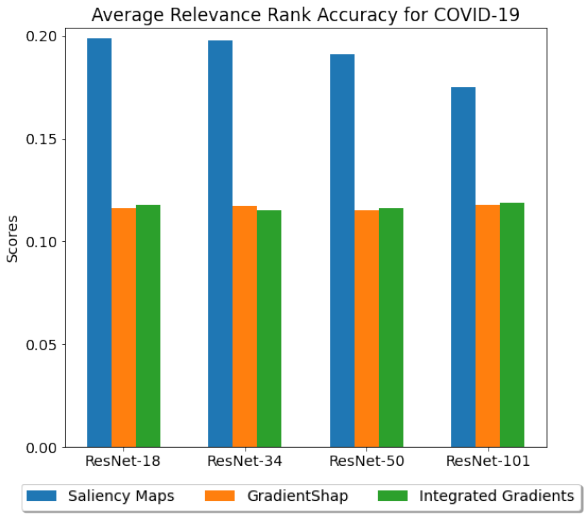

Figure 6: Average results for relevance rank accuracy within *COVID-19* class.

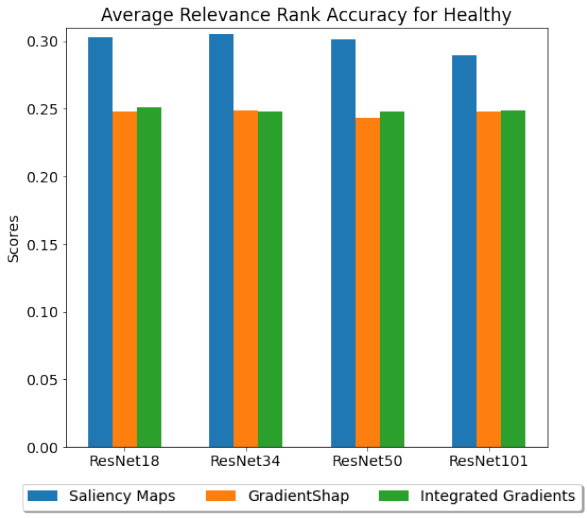

Figure 7: Average results for relevance rank accuracy within *healthy* class.

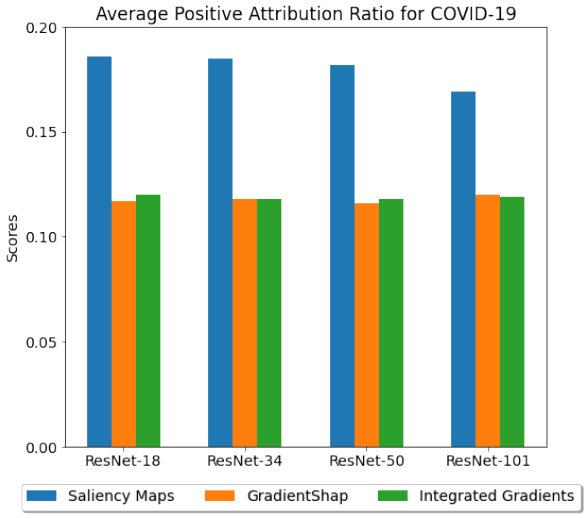

Figure 8: Average results for positive attribution ratio within *COVID-19* class.

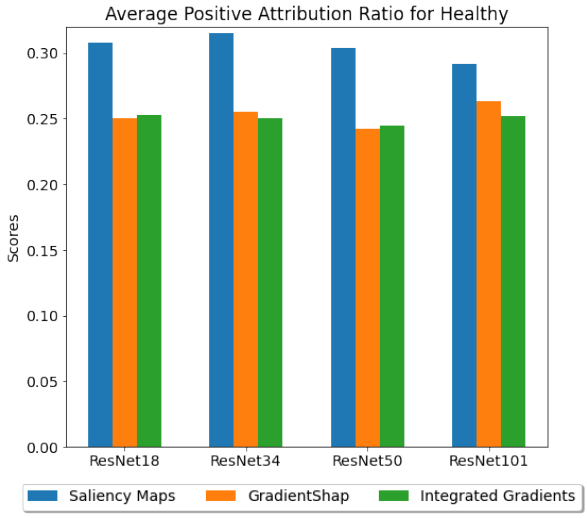

Figure 9: Average results for positive attribution ratio within *healthy* class.

any statistical discrepancies across the quartet of models, irrespective of the metric, category, or XAI technique in question. Notably, within the *Healthy* category utilizing the Saliency Maps, there was a borderline approach to statistical significance in the variance of means across ResNet models during the one-way ANOVA assessment. This culminated in p-values of 0.06 and 0.07 for the Relevance Rank Accuracy and Positive Attribution Ratio metrics, respectively.

Discussion

It is pertinent to note that the efficacy of interpretative methods in XAI hinges on their proper configuration (Montavon et al. 2017; Sundararajan, Taly, and Yan 2017). Incorrect settings can substantially diminish their effectiveness, as evidenced by past research (Kindermans et al. 2017). Therefore, constructing an empirical framework is crucial for validating the effectiveness and reliability of these methods (Hooker et al. 2019).

In healthcare and finance, users may mistakenly view predictive model outputs as causal, for instance, interpreting high saliency metrics as confirmation of specific health conditions. The capability of adversarial attacks to subtly alter inputs and shift focus from relevant to irrelevant features poses a significant challenge; such manipulations often go undetected as they do not change the diagnostic labels (Ghorbani, Abid, and Zou 2018). The vulnerability of deep neural networks to these adversarial attacks is a documented concern, casting doubt on the trustworthiness of their predictive labels (Goodfellow, Shlens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 2017; Papernot et al. 2017).

The research by Ghorbani, Abid, and Zou (2018) delves

Metric	Class	ANOVA	Methodology	ANOVA F-statistic	ANOVA p-value	Kruskal-Wallis H-statistics	Kruskal-Wallis p-value
Relevance Rank Accuracy	*COVID-19*	F(3, 1164)	Saliency Maps	3.06	**0.03**	9.79	**0.02**
			GradientShap	0.07	0.98	0.62	0.89
			Integrated Gradients	0.10	0.96	1.19	0.76
	Healthy	F(3, 576)	Saliency Maps	2.48	0.06	5.83	0.12
			GradientShap	0.25	0.86	0.58	0.90
			Integrated Gradients	0.06	0.98	0.20	0.98
Positive Attribution Ratio	*COVID-19*	F(3, 1164)	Saliency Maps	1.18	0.32	4.87	0.18
			GradientShap	0.10	0.96	0.05	1
			Integrated Gradients	0.04	0.99	0.30	0.96
	Healthy	F(3, 576)	Saliency Maps	2.41	0.07	7.37	0.06
			GradientShap	1.59	0.19	3.90	0.27
			Integrated Gradients	0.34	0.80	0.82	0.84

Table 5: Statistical Test for Differences in Scores Across ResNet-18, ResNet-34, ResNet-50, and ResNet-101 Models. Source: Own calculations.

into the impact of adversarial perturbations on the interpretations provided by neural networks. The interpretation of a neural network is considered vulnerable if there is a possibility to manipulate an image without a perceptual difference, maintaining the initial classification label, while significantly altering the network's interpretation of that image (Ghorbani, Abid, and Zou 2018).

Conclusions

The influence of architectural complexity on the performance and explainability of ResNet models in medical image classification was investigated in this study. It was found that models with reduced complexity could deliver performance and interpretability comparable to or surpassing that of their more intricate counterparts. Specifically, architectures such as ResNet-18 were shown to provide effective accuracy and interpretability, challenging the prevailing belief that increased complexity ensures enhanced efficacy of the model. This provides grounds for **rejecting Hypothesis 1**.

Statistical analysis conducted on interpretability metrics on four ResNet models highlighted a lack of consistent correlation between architectural complexity and interpretability. The outcomes of this study necessitate a sensible approach to the selection of deep learning models, especially for applications that demand high precision and transparent explanations, such as those prevalent in healthcare. The results suggest that the additional resources required for more complex architectures, e.g. increased memory usage, higher financial costs, greater environmental impact, and longer training times, may not be justified, given that less complex architectures could achieve similar or superior levels of interpretability. This provides a rationale for **rejecting Hypothesis 2**.

The study highlights the importance of properly configuring XAI methods to prevent misinterpretation of model predictions and calls for the development of an empirical framework to establish the reliability of these interpretive approaches. The conducted research reinforces the principle of a context-specific selection of neural network architectures underscoring the importance of both performance and interpretability, especially in applications within sensitive domains.

Future Work

In consideration of future explorations within the domain of Explainable Artificial Intelligence and image classification, it is imperative to address the burgeoning interest in the Vision Transformer (ViT) architecture which exceeds the traditional CNN models in a variety of deep learning tasks (Vaswani et al. 2017; Dosovitskiy et al. 2020). The inherent capacity of Transformers to facilitate complex, sequential data processing through self-attention mechanisms posits them as a prime candidate for augmenting the interpretability of deep learning models (Vaswani et al. 2017).

Future investigations should strive to establish methodological approaches that quantify the effect of Vision Transformer complexity on explanation quality. This research should also extend to examining the capability of Transformers to preserve explainability when processing image datasets.

Additionally, it is crucial to extend the assessment of the explainability of Vision Transformers by leveraging the CheXpert dataset, a comprehensive repository of 224,316 chest X-rays across 65,240 patients (Irvin et al. 2019). The dataset encompasses 14 diverse radiological observations, each accompanied by annotations that mark uncertain diagnoses, providing a robust framework for appraising the interpretability of AI in the intricate realm of medical image analysis.

Such research endeavours are expected to contribute significantly to the development of AI systems that are both advanced in their operational capabilities and transparent in their reasoning processes. This balance is essential for fostering trust and facilitating effective human-AI interaction, propelling the field of XAI forward.

Reproducibility

The code utilized for replicating the experimental results is accessible at https://github.com/mateuszcedro/xai-and-model-size/blob/main/notebook/XAI-ResNet50-notebook.ipynb.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2020. Sanity Checks for Saliency Maps. ArXiv:1810.03292 [cs, stat].

Apley, D. W.; and Zhu, J. 2020. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4): 1059–1086.

Arras, L.; Osman, A.; and Samek, W. 2022. Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI. *Information Fusion*, 81: 14–40. ArXiv:2003.07258 [cs, eess].

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7): e0130140.

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; and Hansen, K. 2010. How to Explain Individual Classification Decisions.

Bertrand, H. 2019. Hyper-parameter optimization in deep learning and transfer learning: applications to medical imaging.

Brown, T. B.; et al. 2020. Language Models are Few-Shot Learners. *CoRR*, abs/2005.14165.

Bykov, K.; Hedström, A.; Nakajima, S.; and Höhne, M. M.-C. 2022. NoiseGrad: Enhancing Explanations by Introducing Stochasticity to Model Weights. ArXiv:2106.10185 [cs].

Cabitza, F.; Rasoini, R.; and Gensini, G. F. 2017. Unintended Consequences of Machine Learning in Medicine.

Che, Z.; Purushotham, S.; Khemani, R.; and Liu, Y. 2017. Interpretable Deep Models for ICU Outcome Prediction.

Chen, B.; Wen, M.; Shi, Y.; Lin, D.; Rajbahadur, G. K.; and Jiang, Z. M. 2022. Towards Training Reproducible Deep Learning Models. *CoRR*, abs/2202.02326.

Chowdhury, M. E. H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M. A.; Mahbub, Z. B.; Islam, K. R.; Khan, M. S.; Iqbal, A.; Emadi, N. A.; Reaz, M. B. I.; and Islam, M. T. 2020. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access*, 8: 132665–132676.

Chowdhury, M. E. H.; et al. 2022. QaTa-COV19 Database. https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset. Accessed: 2022-12-01.

Dauphin, Y. N.; and Bengio, Y. 2013. Big Neural Networks Waste Capacity. ArXiv:1301.3583 [cs].

Davenport, T.; and Kalakota, R. 2019. The potential for artificial intelligence in healthcare.

Degerli, A.; Ahishali, M.; Yamac, M.; Kiranyaz, S.; Chowdhury, M. E. H.; Hameed, K.; Hamid, T.; Mazhar, R.; and Gabbouj, M. 2021. COVID-19 infection map generation and detection from chest X-ray images. *Health Information Science and Systems*, 9(1): 15.

Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv:1702.08608 [cs, stat].

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv:2010.11929 [cs].

Eigen, D.; Rolfe, J.; Fergus, R.; and LeCun, Y. 2014. Understanding Deep Architectures using a Recursive Convolutional Network. ArXiv:1312.1847 [cs].

Ghorbani, A.; Abid, A.; and Zou, J. 2018. Interpretation of Neural Networks is Fragile. ArXiv:1710.10547 [cs, stat].

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. ArXiv:1412.6572 [cs, stat].

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5): 1–42.

Guo, M.; and Du, Y. 2019. Classification of Thyroid Ultrasound Standard Plane Images using ResNet-18 Networks. In *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, 324–328. Xiamen, China: IEEE. ISBN 978-1-72812-458-2.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [cs].

Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; and Höhne, M. M. M. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24(34): 1–11.

Hinton, G. 2018. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA*, 320(11): 1101.

Holzinger, A.; Biemann, C.; Pattichis, C. S.; and Kell, D. B. 2017. What do we need to build explainable AI systems for the medical domain? ArXiv:1712.09923 [cs, stat].

Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; and Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4): e1312.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. ArXiv:1806.10758 [cs, stat].

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R. L.; Shpanskaya, K. S.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; and Ng, A. Y. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *CoRR*, abs/1901.07031.

Katuwal, G. J.; and Chen, R. 2016. Machine Learning Model Interpretability for Precision Medicine.

Khan, R. U.; Zhang, X.; Kumar, R.; and Aboagye, E. O. 2018. Evaluating the Performance of ResNet Model Based on Image Recognition. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, 86–90. Chengdu China: ACM. ISBN 978-1-4503-6419-5.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2017. The (Un)reliability of saliency methods. ArXiv:1711.00867 [cs, stat].

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial examples in the physical world. ArXiv:1607.02533 [cs, stat].

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444.

Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. ArXiv:1705.07874 [cs, stat].

Menghani, G. 2023. Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ACM Computing Surveys*, 55(12): 1–37.

Merrick, L.; and Taly, A. 2019. The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory. *CoRR*, abs/1909.08128.

Molnar, C. 2022. *Interpretable Machine Learning*. 2 edition.

Molnar, C.; Casalicchio, G.; and Bischl, B. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. volume 1323, 417–431. ArXiv:2010.09337 [cs, stat].

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65: 211–222.

Morch, N.; Kjems, U.; Hansen, L.; Svarer, C.; Law, I.; Lautrup, B.; Strother, S.; and Rehm, K. 1995. Visualization of neural networks using saliency maps. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, 2085–2090. Perth, WA, Australia: IEEE. ISBN 978-0-7803-2768-9.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical Black-Box Attacks against Machine Learning. ArXiv:1602.02697 [cs].

Paszke, A.; et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Rahman, T.; Khandakar, A.; Qiblawey, Y.; Tahir, A.; Kiranyaz, S.; Abul Kashem, S. B.; Islam, M. T.; Al Maadeed, S.; Zughaier, S. M.; Khan, M. S.; and Chowdhury, M. E. 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, 132: 104319.

Rajkomar, A.; Dean, J.; and Kohane, I. 2019. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14): 1347–1358.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv:1602.04938 [cs, stat].

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K.-R., eds. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Cham: Springer International Publishing. ISBN 978-3-030-28953-9 978-3-030-28954-6.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ArXiv:1708.08296 [cs, stat].

Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S. Q. H.; Nguyen, C. D. T.; Ngo, V.-D.; Seekins, J.; Blankenberg, F. G.; Ng, A. Y.; Lungren, M. P.; and Rajpurkar, P. 2022. Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, 4(10): 867–878.

Sarwinda, D.; Paradisa, R. H.; Bustamam, A.; and Anggia, P. 2021. Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. *Procedia Computer Science*, 179: 423–431.

Secinaro, S.; Calandra, D.; Secinaro, A.; Muthurangu, V.; and Biancone, P. 2021. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, 21(1): 125.

Shapley, L. S. 1953. *Contributions to the Theory of Games*. "A value for n-person games", pp. 307–317.

Showkat, S.; and Qureshi, S. 2022. Efficacy of Transfer Learning-based ResNet models in Chest X-ray image classification for detecting COVID-19 Pneumonia. *Chemometrics and Intelligent Laboratory Systems*, 224: 104534.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2019. Learning Important Features Through Propagating Activation Differences. ArXiv:1704.02685 [cs].

Simonyan; et al. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ArXiv:1312.6034 [cs].

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. ArXiv:1706.03825 [cs, stat].

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks.

Tahir, A.; et al. 2021. COVID-QU-Ex Dataset. https://www.kaggle.com/datasets/anasmohammedtahir/covidqu/. Accessed: 2022-12-01.

Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1): 44–56.

TorchVision maintainers and contributors. 2016. TorchVision: PyTorch's Computer Vision library. https://github.com/pytorch/vision.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. ArXiv:1706.03762 [cs].

Wu, Z.; Shen, C.; and Hengel, A. v. d. 2016. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. ArXiv:1611.10080 [cs].

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Xiao, T.; and Jia, J. 2016. Understanding scene in the wild.

Zinkevich, M.; Weimer, M.; Li, L.; and Smola, A. J. 2010. Parallelized Stochastic Gradient Descent.