Discover Artificial Intelligence



Research

A novel interaction-based methodology towards explainable Al with better understanding of Pneumonia Chest X-ray Images

Shaw-Hwa Lo¹ · Yiqiao Yin¹

Received: 11 October 2021 / Accepted: 6 December 2021

Published online: 20 December 2021 © The Author(s) 2021 OPEN

Abstract

In the field of eXplainable AI (XAI), robust "blackbox" algorithms such as Convolutional Neural Networks (CNNs) are known for making high prediction performance. However, the ability to explain and interpret these algorithms still require innovation in the understanding of influential and, more importantly, explainable features that directly or indirectly impact the performance of predictivity. A number of methods existing in literature focus on visualization techniques but the concepts of explainability and interpretability still require rigorous definition. In view of the above needs, this paper proposes an interaction-based methodology-Influence score (I-score)—to screen out the noisy and non-informative variables in the images hence it nourishes an environment with explainable and interpretable features that are directly associated to feature predictivity. The selected features with high I-score values can be considered as a group of variables with interactive effect, hence the proposed name interaction-based methodology. We apply the proposed method on a real world application in Pneumonia Chest X-ray Image data set and produced state-of-the-art results. We demonstrate how to apply the proposed approach for more general big data problems by improving the explainability and interpretability without sacrificing the prediction performance. The contribution of this paper opens a novel angle that moves the community closer to the future pipelines of XAI problems. In investigation of Pneumonia Chest X-ray Image data, the proposed method achieves 99.7% Area-Under-Curve (AUC) using less than 20,000 parameters while its peers such as VGG16 and its upgraded versions require at least millions of parameters to achieve on-par performance. Using I-score selected explainable features allows reduction of over 98% of parameters while delivering same or even better prediction results.

1 Introduction

Many successful achievements in machine learning and deep learning have accelerated real-world implementations of Artificial Intelligence (AI). This issue has been greatly acknowledged by the Department of Defense (DoD) [7]. DARPA [7] initiated the eXplainable Artificial Intelligence (XAI) challenge and brought this new interest to the surface. In addressing the concepts of interpretability and explainability, these scholars and researchers have made attempts towards discussing a trade-off between learning performance (usually measured by prediction performance) and effectiveness of explanations (also known as explainability), which is presented in Fig. 1 [18, 19]. This trade-off often occurs in any supervised machine learning problems that aim to use explanatory variable to predict response variable (or outcome

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44163-021-00015-z.

Yiqiao Yin, yy2502@columbia.edu; Shaw-Hwa Lo, shl5@columbia.edu | ¹Columbia University, New York, USA.



Discover Artificial Intelligence

(2021) 1:16

| https://doi.org/10.1007/s44163-021-00015-z



Learning

Performance

(Prediction Performance)

Effectiveness of explanations (Explainability)

Fig. 1 This diagram is a recreation DARPA document (DARPA-BAA-16-53) [7, 26]. The diagram presents the relationship between learning performance (usually measured by prediction performance) and effectiveness of explanations (also known as explainability)

(2021) 1:16

variable) which happens between learning performance (also known as prediction performance) and effectiveness of explanations (also known as explainability). As illustrated in Fig. 1, the issue is that a learning algorithm such as linear regression modeling has a clear algorithmic structure and an explicitly written mathematical expression so that it can be understood with high effectiveness of explanations with yet relatively lower prediction performance. An algorithm such as linear regression can be positioned in the bottom right corner of the scale in this figure (in consensus, linear regression is regarded as an explainable learning algorithm). On the other hand, a learning algorithm such as a deep Convolutional Neural Network (CNN) with hundreds of millions of parameters would have much better prediction performance, yet it is extremely challenging to explicitly state the mathematical formulation of the architecture. A deep learning algorithm such as an ultra deep CNN with hundreds of millions of parameters would be positioned on the top left corner of the scale in the figure (which is generally considered inexplainable in consensus). In the field of transfer learning, it is a common practice to adopt a previously trained CNN model on a new data set. For example, one can adopt the VGG16 model and weights learned from ImageNet on a new data: Chest X-ray Images. The filters learned from ImageNet data may or may not be helpful on Chest X-ray. Due to large amount of filters used in VGG16, we can hope that some filters can capture important information on Chest X-ray scans. However, we will never truly know what features are important if we do not impose any feature assessment condition. This renders the adoption of a pretrained CNN model inexplicable. This calls for the need of a novel feature assessment and feature selection technique to shrink the dimension of the number of parameters while maintaining prediction performance. Hence, this paper focuses on feature and variable selection assessment to build explainability including trustworthy, fair, robust, and high performing models for real-world applications. A fruitful consequence of this delivery is to build learning algorithms with state-of-the-art performance while maintaining a small number of features and parameters, an algorithm that can be positioned on the top right corner of the scale in Fig. 1.

1.1 Organization of this paper

The paper is organized as follows. Section 2 proposes the definition for explainability of feature(s). This definition requires three major conditions for a method to be explainable and interpretable when it is used to measure the explainability and interpretability of features. The proposed definition leads us to discuss whether modern-day Convolutional Neural Networks (CNNs) satisfy the definition of explainable models. The discussion inspired us to create a novel statistical measure that satisfies the conditions required in the proposed definition. This leads to the proposed method in Sect. 3. Next, we apply the proposed method in a real-world application Pneumonia Chest X-ray Image Classification in Sect. 4. In this application, we show that the proposed method can achieve 99.7% Area-Under-Curve (AUC) in test set performance while using 98% less number of parameters in the neural network architecture and delivering explainable features.

2 Definition of feature explainability

A popular description of interpretability defines XAI as the ability to explain or to present in understandable terms to a human Doshi-Velez and Kim [8]. Another popular version states interpretability as the degree to which a human can understand the cause of a decision [23]. Though intuitive, these definitions lack mathematical formality and

¹ We use the terms response variable and outcome variable interchangeably throughout the article.



rigorousness [1]. Moreover, it is yet unclear why variables provide us the good prediction performance and, more importantly, how to yield a relatively unbiased estimate of a parameter that is not sensitive to noisy variables and is related to the parameter of interest.

To shed light to these questions, we define the following three necessary conditions (C1, C2, and C3) for any feature selection methodology to be explainable and interpretable. In other words, a variable and feature selection method can only be considered explainable and interpretable if all three conditions (C1, C2, and C3 defined below) are satisfied. Specifically, we regard the final assessment quantity of the importance evaluated for a set of features or variables to be the final score measured for feature assessment and selection method. More importantly, we define this importance score of a variable set from using only explainable feature assessment and selection methods to be the explainability of a combination of variables. There are three conditions defined below and we name these conditions C1, C2, and C3.

- C1. The first condition states that the feature selection methodology must be non-parametric and hence does not require any assumption of the true form of the real model.
- C2. An explainable and interpretable feature selection method must clearly state to what degree a combination of explanatory variables influences the response variable.
- C3. In order for a feature assessment and selection technique to be interpretable and explainable, it must related with the predictivity of the explanatory variables (for definition of predictivity, please see Lo et al. [21], Lo et al. [22]).

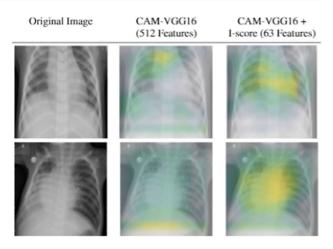
The first condition is required, because the assumption of the underlying model can affect end-users' ability to interpret the feature importance. This means the following. Suppose a prediction task has explanatory variables X and outcome variable Y. If we fit a model $f(\cdot)$ and we use f(X) to explain Y, the procedure requires us to explicitly understand the internal structure of $f(\cdot)$. This is extremely challenging to establish a well-written mathematical formulation when the fitted model is an ultra-deep CNN with hundreds of millions of parameters (sometimes it is combination of many deep CNNs). Worse yet, any attempted explanations would have mistakes of this model $f(\cdot)$ carried over. To avoid this bewilderment, scientists and statisticians desire explanation of how features affect the outcome variable. Hence, a feature selection methodology is required to understand how explanatory variables X impact the outcome variable Y without relying on the procedure of searching for a fitted model $f(\cdot)$. Instead of using a fitted model $f(\cdot)$ to explain feature importance, this condition proposes to use non-parametric procedure to directly explain X and the impact X has on the outcome variables without fitting model. Hence, the first condition, C1, requires the feature selection methodology to not depend on the model fitting procedure (to avoid the procedure of searching for $f(\cdot)$).

Next, the second condition requires an explainable and interpretable feature selection method to state quantitatively to what degree a combination of explanatory variables influence the response variable. It is beneficial if a statistician can directly compute a score for a set of variables in order to make reasonable comparisons. This means any additional influential variables should raise this score while any injection of noisy and non-informative variables should decrease the value of this score. Hence, this condition, C_2 , allows statisticians to pursue feature assessment and feature selection in a quantifiable and rigorous manner. Since we consider the explainability to be the final score and assessment of a variable set using only explainable feature assessment methodology, this second condition, C_2 , asserts that explainability of a combination of variables to be exactly the amount of assessment that explainable feature assessment and selection methodology evaluates and it is a concept states how important explanatory variables are at influencing outcome variable.

Thirdly, the last condition requires explainability to be defined using a measure of predictivity. This is because in deep learning era the variables are commonly explained in two ways. The first is to rely on the weights (also known as the parameters) found by backpropagation. For example, consider input variables (or explanatory variables) to be X_1 and X_2 . A very simple neural network architecture can be constructed using weights w_1 and w_2 . We can simply define estimated outcome variable Y to be $\hat{Y}:=\operatorname{sigmoid}(\sum_{j=1}^2 w_j X_j)=(1+\exp(-\sum_{j=1}^2 w_j X_j))^{-1}$. Though with little intuition, w_j is commonly used to illustrate how much X_j affects Y. The second is through visualization after model fitting procedure. For example, CAM (and its upgraded versions) [33] can be used to generate highlights of images that are important for making predictions. The first approach of using weight parameters can be quite challenging and even an impossible task when the neural network architecture has hundreds of hidden layers and hundreds of millions of parameters. The second approach, however, requires the statistician to have the access of internal structure of the model. This would be difficult if an ultra-deep CNN is used with millions of parameters.



Fig. 2 The graph presents two samples using CAM visualization technique before and after using proposed method. This marks the direction of explainability that this paper tackles. We will see more intriguing results in the application section



Only with all three conditions satisfied, a feature assessment and selection technique would be considered explainable and interpretable in this article. We regard these three conditions to be required in order for a feature assessment and selection methodology to be considered explainable and interpretable. In addition, we consider the explainability of a variable set to be exactly the outcome score of an explainable feature assessment and selection method. Only with this appropriate method that satisfies the definition of explainable feature assessment technique can we say how the explanatory variables explains the outcome variable. Hence, we propose a research agenda of feature selection methodology that evaluates the explainability of explanatory variables.

With these questions in front of us, research agenda towards searching for a criterion to locate highly predictive variables is imminent. Yang and Kim [32] raised the question of absolute feature importance (exactly how important each feature is), but it is yet unexplored how to search for important features by directly looking at the given data set before fitting a model which fails to check the first condition, C1. Amongst a variety of deep learning frameworks, the Convolutional Neural Networks (CNNs) [17] have been widely adopted by many scholarly work including computer vision, object detection [10, 12], image recognition [13, 16, 29], image retrieval [11, 12], and so on. Many famous network architectures that exist including VGG16 [28], VGG19 [28], DenseNet121 [15], DenseNet169 [15], DenseNet201 [15], ResNet [14], and Xception [5]. Despite their [5, 14–16, 28] remarkable prediction performance, some have proposed to avoid the use of fully-connected layers to minimize the number of parameters while maintaining high performance [33].

In the work of Zhou et al. [33], experiments show advantages of using global average pooling layer to retain localization ability until the final layer. However, these doctrines require the access of internal structure of a CNN and, hence, fail to meet the first condition, $\mathcal{C}1$. Zhou et al. [33] provided explanation of the important characteristics of regions in the image emphasized by CNNs for determining the classification of the entire image. However, weights of pooling layer, determined by backpropgation, may also take background information into consideration when the algorithm is making predictions. This may produce inconsistency with human beliefs since humans generally tend to focus on the foreground characteristics which do not meet the second condition, $\mathcal{C}2$. For example, the investigation in this paper used Class Activation Map (CAM) on the X-ray images of Pneumonia diseased patients. Human experts generally study the lung area and explores how disease manifests. However, an Al-trained model may use background area (such as liver, stomach, and intestines) instead of the foreground (features of lungs) to make predictions (see Fig. 2).

The above methods all focus on variable selection and feature importance ranking. However, no method has been able to provide sound approach to meet all three conditions: C1, C2, and C3 described above. Specifically, it is not yet fully discovered how to select variables without assuming any model formation from many noisy variables and, moreover, how to detect, out of many potential explanatory features, the most influential features that directly have impact on response variable Y.

Chernoff et al. [4] presents a general intensive approach, based on a method pioneered by Lo and Zheng [20] for detecting which, out of many potential explanatory variables, have an influence given a specific subset of explanatory variables on a dependent variable Y. This paper presents an interaction-based feature selection methodology incorporating the notion of influence score, l-score, as a major technique to detect the higher-order interactions in complex and large-scale data set. The proposed name "interaction" comes from the nature that the proposed l-score (see Sect. 3) is designed to select highly predictive variable set such that the set of variables, once selected by l-score, can form a higher-order interaction, i.e. multiple variables that have joint effect on predicting outcome variable. Our work also investigates the potential usage of l-score to explain and visualize the deep learning framework. It is not surprising that these new



tools (I-score and Backward Dropping Algorithm) are able to provide vast explanation directly associated to response variable and interpretation to visualize any given CNNs and other deep learning frameworks.

3 Proposed method

The proposed methodology comes with three stages. First, we investigate variables to identify those with high potential to form influential modules. Secondly, we generate highly influential variable modules from variables selected in the first stage, where variables of the same module interact with each other to produce a strong effect on Y. Lastly, we combine the variable modules to carry out the prediction process.

From prior simulation experience [21, 22], it is demonstrated that the two basic tools, I-score and Backward Dropping Algorithm, can extract influential variables from data set in regards of modules and interaction effect. However, questions remain on how to determine the input to Backward Dropping Algorithm and how to use the output from Backward Dropping Algorithm results to construct prediction estimates. Unless one can appropriately utilize input to and output from Backward Dropping Algorithm, the strength of Backward Dropping Algorithm cannot be fully excavated. In this sense, the innovation of the proposed method manifests itself in three ways. First, if one directly applies Backward Dropping Algorithm on high-dimensional data sets, one may still miss some useful information. We propose two-stage feature selection procedure: interaction-based variable screening and variable module generation via Backward Dropping Algorithm. Since the impurity of features is largely enhanced by the interaction-based variable selection algorithm in the first stage, we are able to construct variable modules that are higher order interactions with lots of information in the second stage. These variable modules will support as building blocks for us to form classification rules. This school of thoughts produce results significantly better than directly application of Backward Dropping Algorithm.

The proposed measure for feature explainability, I-score, is defined using discrete variables. If some explanatory variables are continuous, we first convert them into discrete ones for feature selection purpose (see Supplement for Artificial Example III for detailed discussion of discretization). After we have selected the important variables, we can use the original values to estimate their effects. We can rely on the influence score when we convert continuous variables into discrete variables. For example, if a random variable is drawn from normal distribution, then one optimal cutoff is to use the one that has the largest marginal I-score. There is a trade-off induced from this process: the information loss due to discretizing variables from continuous to discrete forms versus the information gain from robust detection of interactions by discretization. Wang et al. [31] demonstrated that the gain from robust detection of interactions is much more than enough to offset possible information loss due to discretization. Wang et al. [31] used the two-mean clustering algorithm to turn the gene expression level into a variable of two categories, high and low. As an additional piece of evidence supporting the proposed pre-processing step, the authors have tried more than two categories; e.g. three categories of high, medium and low. The empirical results show that the more categories used the worse classification error rates.

3.1 Influence score (I-score)

The Influence score (I-score) is a statistic derived from the partition retention method [4]. Consider a set of n observations of an outcome variable (or response variable) Y and a large number S of explanatory variables, $\mathbf{X} = \{X_1, X_2, ..., X_S\}$. Randomly select a small group, m, of the explanatory variables \mathbf{X} . We can denote this subset of variables to be $\{X_k, k=1,...,m\}$. We suppose X_k takes values of only 1 and 0 (though the variables are binary in this discussion, it can be generalized into continuous variables, see Supplement Artificial Example III). Hence, there are 2^m possible partitions of the set of observations for X's. The n observations are partitioned into 2^m cells according to the values of the m explanatory variables. We refer to this partition as Π_X . The proposed I-score (denoted by I_{Π_V}) is defined in the following

$$I_{\Pi_X} = \frac{1}{ns_n^2} \sum_{j=1}^{2^m} n_j^2 (\bar{Y}_j - \bar{Y})^2 \tag{1}$$

while $s_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$. We notice that the I-score is designed to capture the discrepancy between the conditional means of Y on $\{X_1, X_2, ..., X_m\}$ and the mean of Y. For the rest of the paper, we refer this formula as Influence Measure or Influence Score (I-score).

The statistics / is the summation of squared deviations of frequency of Y from what is expected under the null hypothesis. There are two properties associated with the statistics /. First, the measure / is non-parametric which requires no



Discover Artificial Intelligence

need to specify a model for the joint effect of $\{X_{b_1},...,X_{b_k}\}$ on Y. This measure I is created to describe the discrepancy between the conditional means of Y on $\{X_{b_1},...,X_{b_k}\}$ disregard the form of conditional distribution. With each variable to be dichotomous, the variable set $\{X_{b_1}, ..., X_{b_n}\}$ form a well-defined partition, \mathcal{P} [4]. Secondly, under the null hypothesis that the subset has no influence on Y, the expectation of I remains non-increasing when dropping variables from the subset. The second property makes I fundamentally different from the Pearson's χ^2 statistic whose expectation is dependent on the degrees of freedom and hence on the number of variables selected to define the partition. We can rewrite statistics I in its general form when Y is not necessarily discrete

$$I = \sum_{j \in \mathcal{P}} n_j^2 (\bar{Y}_j - \bar{Y})^2, \tag{2}$$

where \bar{Y}_i is the average of Y-observations over the jth partition element (local average) and \bar{Y} is the global average. It is shown (Chernoff et al. [4]) that the normalized I, $I/n\sigma^2$ (where σ^2 is the variance of Y), is asymptotically distributed as a weighted sum of independent χ^2 random variables of one degree of freedom each such that the total weight is less than one. Due to this reason, we typically recommend using the normalized I-score in equation 2. Empirically, the normalized I-score is shielded from exploding value under large amount of sample size, because n_i^2 in the equation 2 may lead to extremely large I-score values when the training sample size is large. It is precisely this property that serves the theoretical foundation for the following algorithm.

We further discuss in the Artificial Example II of the Supplement the comparison between AUC values and I-score in a simulated environment.

3.2 Backward Dropping Algorithm (BDA)

The Backward Dropping Algorithm is a greedy algorithm to search for the optimal subsets of variables that maximizes the I-score through step-wise elimination of variables from an initial subset sampled in some way from the variable space. The steps of the algorithm are presented in Algorithm 1.

```
Algorithm 1: Procedure of the Backward Dropping
Algorithm (BDA)
 Training Set: Suppose there is a training set
   \{(y_1, x_1), ..., (y_n, x_n)\} of n observations, where
   x_i = (x_{1i}, ..., x_{pi}) is a p-dimensional vector of
   explanatory variables. Typically p is very large. All
   explanatory variables are discrete.
 Sampling from Variable Space: Select an initial
   subset of k explanatory variables
   X_b = \{X_{b_1}, ..., X_{b_k}\}, b = 1, ..., B. The notation b
   indicates which rounds of BDA it is executing. In
   practice, we recommend \boldsymbol{B} to be a large number.
 Compute I-score: I(X_b) = \sum_{j \in \Pi_{X_b}} n_j^2 (Y_j - Y)^2.
 Initialize: set X_{\text{current}} to be X_b
 while the subset X_{current} has more than one variable
   do
     \begin{array}{l} 1. \ \ \text{tentatively drop each variable in } X_{\text{current}} \\ 2. \ \ \text{recalculate the I-score with one variable less} \\ 3. \ \ \text{then drop the one that gives the highest I-score and call} \end{array}
         this new subset X'_b, which has one variable less than
         X_{\text{current}}
     4. store this subset of variable and its corresponding I-score
         in a list for future reference
     5. reset X_{\text{current}} to be X_b'
 return The subset that yields the highest I-score in
   the whole dropping process.
```

The proposed BDA (see Algorithm 1) presents a systematic way of searching for important and explainable features. Suppose a data set has p explanatory variables. The BDA proposes to randomly sample a group of variables (usually we recommend a group of size 10). Each random sample (say 10 explanatory variables) can then be sent into the BDA to



iteratively drop variables until there is no variable left. In doing so, I-score value at each step of the variable dropping will increase and then decrease. This is because in each group of variables there is always a subset of that group to be the most predictive variables to explain the outcome variable. The goal is omitting the noisy variables until I-score reaches a top value. However, this value is not known unless we drop till the last variable. This is why the condition of the while loop is $l \ge 1$. Once the loop finishes, we examine the variables dropped and the path of I-score value. We select the subset where the I-score value is the highest. The following toy example demonstrates the procedure of BDA. Suppose there are $X_i \sim \text{Bernoulli}(1/2)$ while $i = \{1, 2, 3\}$. Assume there is a toy model $Y = X_1 + X_2 \pmod{2}$. In other words, the variable X_3 would be a noisy variable because it does not contribute to the definition of Y. In addition, suppose we generate 100 samples for X_i and Y. BDA can be used to screen out X_3 . We present the steps of BDA in Table 1.

The toy example only has three variables $\{X_1, X_2, X_3\}$, so we only executed BDA once (in Table 1). In practice, when the number of variables is large, we recommend to each round of BDA to sample a group of 8-10 variables to execute Algorithm 1. In addition, we recommend the total number of rounds of BDA, B, to be a large enough number so that we have good chance to stumble upon important high-order interactions in a round.

There are more detailed discussion of how to use the proposed I-score and BDA to conduct feature selection in Supplement. More details in the Artificial Example I of the Supplement and we illustrate the procedure of using the BDA in a simulated environment.

3.3 Executive diagram for the proposed method

The proposed method I-score and BDA discussed above can be used to construct a novel interaction-based classifier. The major steps of the proposed learning method is presented in Fig. 3.

4 Application

This section presents the results of the experiments.

4.1 Background of the Pneumonia disease

Pneumonia has been playing a major component of the children morality rate across the globe. According to statistics from World Health Organization (WHO), an estimated of 2 million deaths are reported every year for children under age 5. In the United States, pneumonia accounts for over 500,000 visits to emergency departments and 50,000 deaths in 2015, keeping the ailment on the list of the top 10 causes of death in the country. Chest X-ray (CXR) analysis is the most commonly performed radiographic examination for diagnosing and differentiating the types of pneumonia [3]. While common, accurately diagnosing pneumonia is possible with modern day technology, it is a requirement to review chest radiograph (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory examinations. Pneumonia usually manifests as an area or regions of increased opacity on CXR Franguet [9]. However, the diagnosis of pneumonia on CXR is complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post-radiation or surgical changes. Outside of the lungs, fluid in the pleural space (pleural effusion) also appears as increase opacity on CXR. Toğaçar et al. [30] used a deep CNN with 60 million parameters. Ayan and Ünver [2] tested an ultra deep convolutional neural network that has 138 million parameters. Rajaraman et al. [25] attempted a variety of architecture that ranges in 0.8–40 million parameters. While some of these architecture produced good prediction performance, the power of explainability is lost in these deep convoluted structures. The proposed method of using I-score and BDA to screen for important features, however, do not rely on deep convoluted network architecture. The features with high I-score values are directly responsible for the prediction of the diseased cases and these explainable features can be fed into a feedforward neural network with a single pass of propagation to make prediction (Fig. 4).



Table 1 Demonstration of RDA

l-score	13.10	25.90	1.05
Variables left	X_1, X_2, X_3	X_{1}, X_{2}	X_2
Investigate	Drop X_1 , I = 0.94	Drop X_1 , $I = 1.31$	
	Drop X_2 , I = 0.55	Drop X_2 , I = 1.05	
	Drop X_3 , I = 25.90		
Drop	X_3	X_1	
	3	1	

The subset of variables left with the highest l-score value is $\{X_1, X_2\}$

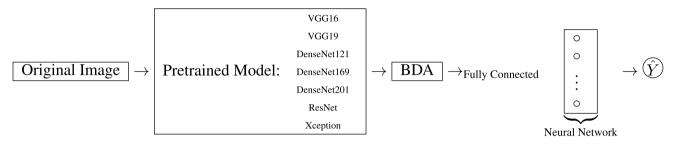


Fig. 3 Executive diagram for the proposed pipeline. The figure illustrates the major steps of the proposed pipeline. In the following flow-chart, the proposed algorithm starts with the training sample of the original image data. Training samples are fed into pretrained models such as VGG16/19, DenseNet121/169/201, ResNet, and Xception. Features of any layer of the above pre-trained CNNs can be extracted. From practice, it is often recommended to extract the last convolutional layer of a pre-trained CNN to be new features. The proposed I-score can then be applied to examine the explainability of each features by computing I-score directly. Alternatively, BDA can be applied and groups of features can be selected. Each group form a higher-order interaction and we call this group a variable module. Individual feature and groups of features can then be used to form classifier by feeding them into a feed-forward neural network. Because the features are finely selected, the feed-forward neural network requires very much less number of parameters then its peers to achieve similar or better results

4.2 Biological interpretation of the image data

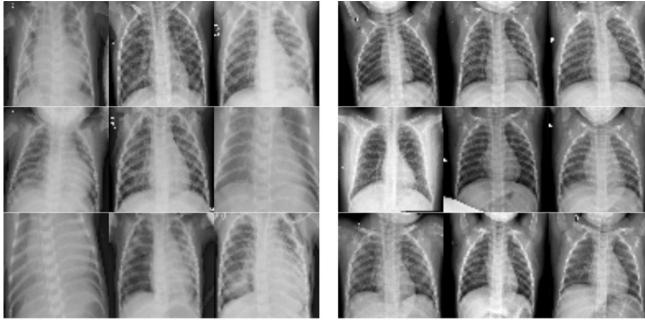
An important notion is opacity. Opacity refers to any area that preferentially attenuates the X-ray beam and therefore appears more opaque than the surrounding area. It is nonspecific term that does not indicate the size or pathological nature of the abnormality.²

From observation, lung opacity is not homogeneous and it does not have a clear center or clear boundaries. There is no universal methodology to properly segment opacity out of the entire picture. However, if one can segment the lungs properly and filter out the rest of the image, it is possible to create a clean image of the lungs for neural network to process.

There is a biological reason of why different healthy levels of lungs exhibit different level of opacity. To illustrate this idea, a diagram of lung anatomy and gas exchange is posed in the following (see Fig. 5). The structure of the human lungs consists of Trachea, Bronchi, Bronchioles, and Alveoli. The most crucial activity is the cycled called Gas Exchange. Trachea acts like a main air pipe allowing air to pass through from outside of the human body to inside human chest area. The Bronchi connects Trachea and are thinner pipes that allow air to move further into the lung area. The end of Bronchi has many tiny airbags called Alveoli. Alveioli is the center location for gas exchange and it has a thin membrane to separate air and bloodstream. As human beings conduct day-to-day activities such as running, walking, or even sleeping, bloodstream is constantly filled with Carbon Dioxide that is generated from these activities which then need to be passed out of the human body. The pass from bloodstream into Alveoli is the first step. The reverse direction also has an activity for Oxygen to pass into the bloodstream so human beings can continue to conduct normal day-to-day behavior. The in-and-out cycle with Carbon Dioxide and Oxygen is called the

² Felson's Principles of the Chest Roentgenology (4E), available from https://www.amazon.com/Felsons-Principles-Roentgenology-Programmed-Goodman/dp/1455774839.





Panel A Panel B

Fig. 4 The graph shows randomly sampled images from pediatric CXRs collected in Pneumonia data set. A Presents 9 images from Class Diseased. These CXR show diseased pictures with bacterial affected pneumonia disease. B Presents 9 images from Class Normal. These images show clear lungs with no abnormal opacity

Gas Exchange which is a normal microscopic activity occurs disregard whether human beings are conscious or not. For patients with diseased lungs, it is a natural reaction that the immune system is fighting against the germs and the inflammatory cells. This progress creates fluids inside Alveoli which generates grey area on image because not all CXR go through lungs. This creates the opaqueness in the images collected from CXR machine (Fig. 6).

Neuman et al. [24] proposed methods to evaluate radiography of children in a pediatric emergency department for suspicion of pneumonia. A team of six radiologists at two academic children's hospitals were formed to examine the image data of the chest X-ray pictures. Neuman's work suggested there was only a moderate level of agreement between radiologists about the presence of the opacity.

4.3 Pneumonia data set

The Pneumonia data set has images from different sizes (usually range from 300 to 400 pixels in width). For the experiment analyzing pixels, we reshape all images into 3-dimensional tensor that has size 224 by 224 by 3. In other words, each image has 150,528 pixels taking values from 0 to 255 before rescaling. There are 1341 images with label normal class and 3875 images with label pneumonia. In other words, this is a binary classification problem and we want to train a machine to learn the features of the images to be able to predict a probability that an image fall in class normal or pneumonia. We randomly select 300 images from normal cases and 300 images from diseased cases as test data. We then sample with replacement 3000 images from normal cases and 3000 images from diseased cases as training. This statistics can be summarized in the following table (see Table 2).

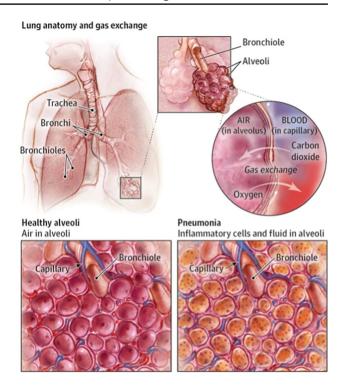
4.4 Transfer learning using VGG16

To borrow the strength of another deep Convolutional Neural Network, transfer learning is a common scheme in adopting many pre-trained models in a new data set. However, due to significantly large amount of filters, it is much less optimal if the prediction performance lack of explainability.

In our work, we adapt the architecture of VGG16. The key component is feature extraction using VGG16 feature maps. We feed every image in Chest X-ray data set into the architecture of VGG16. This means that each image goes through



Fig. 5 This image shows the lung anatomy and use graphical diagram to illustrate how gas exchange cycles in healthy human lungs versus diseased human lungs



the VGG16 procedure and we can produce many feature maps according to the parameters of the VGG16 architecture. The architecture of VGG16 model is presented in Fig. 7. A brief overview of the filter kernels used in the VGG16 model is presented in Fig. 8. A visualization of such feature map is produced in Fig. 9.

(2021) 1:16

From Fig. 8, we can see that each filter is designed with certain small patterns that aims to capture certain information in a local 3-by-3 window on an image. These designs are direct production of the work by Simonyan and Zisserman [28]. Though these filters produce robust performance in Simonyan and Zisserman [28], any new adaption of these filters on a new dataset may or may not produce intuition to human users. We regard the usage of many filters from a pre-trained model without any feature assessment methodology to be the major reason why transfer learning relying on a deep CNN such as VGG16 inexplainable.

Suppose we are given image data. This means one observation is a colored picture and it is a 3-dimensional array (a tensor). If it is a black and white picture, we treat is as a 2-dimensional array. However, in order to use VGG16, it is required that the input dimension to be 224 by 224 by 3. Any black and white picture is resized into 3-dimensional tensor by treating each color to have the same greyscale.

By using the filters (see Fig. 8), we are able to extract certain information from the original image. These new features are information that are essentially some transformation of the original pixels based on filters created to detect certain patterns in another dataset.

A deep CNN such as VGG16 typically has many convolutional layers. They are formed by standard techniques such as convolutional operation, pooling, and so on. There is not yet any explainable measure in the literature that can help us measure the explainability of these feature maps (Fig. 10).

4.5 Feature assessment and predictivity

The high I-score values (see Fig. 11) suggest that local information possess capability to have higher lower bounds of the predictivity. In Fig. 11, the 512 features from VGG16 are ranked using both AUC (on the x-axis) and I-score values (on the y-axis). The features with the high I-score values (values above 400 can be considered high I-score values) are also extremely predictive (with AUC above 73%). This is worth noticing because this information leads to not just high prediction performance but also explainable power. Specifically, the statistics of I-score does not rely on any assumption of the underlying model. This property eliminates the potential harm a fitted model can bring when we try to explain how features affect the prediction performance. Hence, the first condition $\mathcal{C}1$ satisfies, because the computation of I-score



Fig. 6 The lung anatomy and several annotated boxes doctors use to indicate important area in the image for diagnosis. Neuman et al. [24] suggested such annotation can be helpful for Pneumonia diagnosis

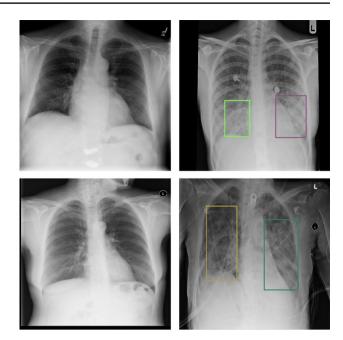


Table 2 The size of the training set, the validating set, and the test set for both approaches

Proposed approach Size: 224 by 224 by 3	Normal Obs. for	Pneumonia Obs. for
Training	2700	2700
Validating	300	300
Testing	300	300

The proposed approach is interaction-based feature learning which proposes to work with convolutional features instead of the original pixels

does not rely on any assumption of the underlying model. In addition, the magnitude of I-score allows statisticians to carry out feature assessment in order to make comparisons which subset of features are more explainable and influential at making predictions. This satisfies the second condition $\mathcal{C}2$ of an explainable measure. Third, the construction of the proposed I-score associates feature explainability with the predictivity of the features. This implies that I-score measure satisfies the third condition $\mathcal{C}3$. Hence, the proposed statistics I-score is an explainable measure and therefore the final score is the explainability of the variables.

Suppose we have true label Y and a predictor \hat{Y} . We can compute AUC of this predictor \hat{Y} . First, we use thresholds to convert \hat{Y} into two levels, i.e. "0" and "1", where the thresholds are formed by the unique levels of the real values in \hat{Y} . Then we compute a confusion table based on Y and converted \hat{Y} . The confusion table gives us sensitivity and specificity which allows us to plot ROC. The area under the ROC path is Area-Under-Curve (AUC). The notion of sensitivity is interchangeable with recall or true positive rate. In a simple two-class classification problem, the goal is to investigate covariate matrix X in order to produce an estimated value of Y. In our proposed neural network algorithm, we choose sigmoid activation function for the final output layer, which means that the predicted values are always between 0 and 1, which acts as a probabilistic statement to describe the chance an observation is class 1 or 0. Given a threshold between 0 and 1, we can compute sensitivity to be the following



Fig. 7 The VGG16 Architecture Simonyan and Zisserman [28]

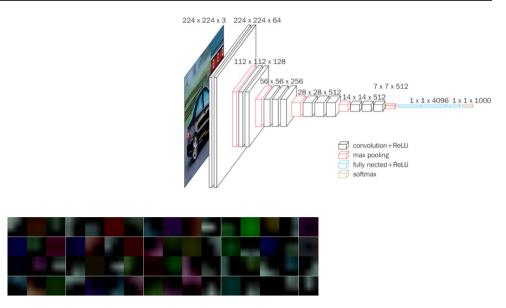
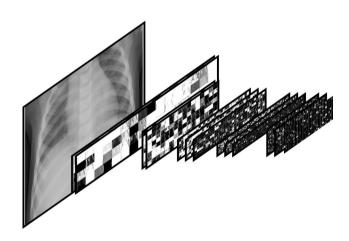


Fig. 8 Filters extracted from the architecture of VGG16 model. There are 64 filters in VGG16 model and they are presented above. Each filter has size 3 by 3 by 3. We code them into three colors by using RGB palette

Fig. 9 The resulting images from each layer of the VGG16 architecture Simonyan and Zisserman [28]. Given a new data set (i.e. the Pneumonia Chest X-ray dataset), we can feed each image into the VGG16 architecture to create feature maps from using the parameters and filter kernels in the original VGG16 model



On the other hand, specificity is also used to create ROC curve. Given a certain threshold between 0 and 1, we can compute specificity using the following

Specificity
$$= \frac{\text{True Negative}}{\text{Negative}}$$

$$= \frac{\text{#of Correctly Classified Non-Pneumonia Images}}{\text{#of Non-Pneumonia Images}}$$
(4)

Given different thresholds, a list of pair of sensitivity and specificity can be created. The Area-Under-Curve (AUC) is the area under the path plotted using pairs of sensitivity and specificity that is generated using different thresholds. For empirical results, please see Fig. 11.

4.6 Explainability and interpretation

Figure 12 presents ten samples from diseased classes. Each sample we present five plots: (1) the original picture sized 224 by 224, (2) visualization of the last convolutional layer of VGG16 (512 features) using heat-map generated from CAM, (3) the same visualization but using I-score to pick the top 53 out of 512 features from the last convolutional



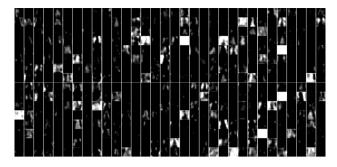


Fig. 10 These are the 512 feature maps extracted from the last convolutional layer of the VGG16 model. In transfer learning, we adopt a previously trained CNN model on current data set (i.e. pneumonia data). Due to lack of feature assessment procedure, we expect many of these feature maps to be inexplainable to detect the disease cases which are noisy information that do not contribute to a robust prediction performance. The motivation is to generate a robust understanding which features can explain the prediction performance.

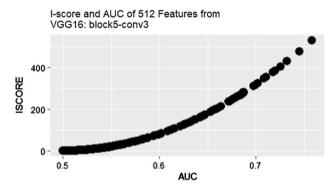


Fig. 11 Graph presents empirical relationship between proposed I-score and AUC. To plot this graph, we investigate the last convolutional layer (namely "block5-conv3") in the VGG16 architecture, i.e. there are 512 of these features extracted. We compute I-score and AUC value for each one of the 512 features. In this experiment, we directly compute the I-score of each feature and we do not need to use BDA. Since there are 512 features, all of their I-score values are computed and we use I-score value to measure the explainability and predictivity of each feature. To provide readers intuition how powerful each I-score value is, we also compute AUC for each feature alone (that is, by treating each feature itself as a predictor). The AUC values and I-score values are plotted for all 512 features in a scatter plot

layer of VGG16 (I-score greater than 200), (4) top 39 out of 512 features (I-score greater than 150), and (5) top 19 out of 512 features (I-score greater than 100).

In transfer learning, we adopt a previously trained model such as VGG16 and apply the convolutional features on chest X-ray data set. Due to significantly large number of filters in previous model and lack of feature selection method, it is extremely challenging to see the location in the image that is used to generate good performance using all 512 convolutional features which render the project inexplicable.

The proposed method I-score is capable of selecting and explaining, out of 512 convolutional features from a previously trained model, the important features (sometimes as little as less than 30 features) to create explainability and interpretability and the proposed procedure uses heat-map to suggest the exact location that the disease may occur in patients. The application of the proposed I-score and BDA are extremely flexible at tailoring into different scenarios. Under the scenario where the interest is the explainability of each feature, Fig. 11 can be created and we can directly compute I-score value of each feature. The ranking of these feature should directly correspond to AUC ranking as well for each interpretation, which is a novel discovery in the literature. BDA is not required in computation of explainability of each feature alone. Under scenarios where we need to extract higher-order interactions, we use BDA to search for explainability of a group of variables, which requires the usage of BDA. The groups of variables with high I-score values can be highlighted using heatmaps to create saliency mapping on the original picture, as shown in Fig. 12.



(2021) 1:16

Fig. 12 Ten samples from diseased classes. For each image, we develop 512 features using the VGG16 architecture. However, the significantly large number of filters from a pre-trained model renders the newly constructed features inexplainable. In regarding to this major pitfall, we propose to use I-score, an explainable measure, that produces explainability of variables and describes how a subset of variables affect the prediction performance. Among the 512 VGG16 features, the proposed statistics I-score can use the top 19 most explainable features to warn human users the location of chest X-ray that is directly associated with the pneumonia disease in patients.

To further confirm the improvement that the proposed method delivered, we conduct experiments by removing the influential features by I-score and we present AUC with the data (where these important features are not presented). For example, in Table 3 of the proposed approach for the top 2.5% modules, we delivered 99.3% AUC with approximately 200 features. The removal of these features drops AUC from 99.3% to 57.4% which is almost close to random guessing. For the top 5% of modules of 5,216 features from 7 different CNNs, we delivered 99.7% using approximately 400 features which are selected by I-score from the last convolutional layer of these ultra-deep CNNs. The removal of these features drop AUC to 54.2%.

5 Future scope

The proposed methods of using I-score and BDA to form explainable variable sets in order to build efficient and narrow neural network classifiers lay the foundation of using non-parametric approach to explain important features. In doing so, a major benefit is the immediately reduction of the size of the number of parameters and the number of hidden layers in the neural network architecture. However, this article has not yet investigated specific data structures and how the proposed methods may behave under new data structure. Hence, this paper also calls for future research opportunities of investigating the application of I-score and BDA under 2D (image data specifically), 3D (spatial data or MRI scans), and sequential data structure.

6 Conclusion

First, the paper provides a novel and rigorous definition for explainable and interpretable feature assessment and selection methodology (please see the boldface definition for three major conditions C1, C2, and C3 for fulfilling the necessary conditions of explainable and interpretable feature selection method). We define the explainability and interpretability



(2021) 1:16

Table 3 Test set performance of the Pneumonia Chest X-ray detection data set

Method	No. of Param.	Test AUC
Previous		
VGG16 [2]	approx. 138 million	87.0%
Saraiva et al. [27]		94.5%
AlexNet Toğaçar et al. [30]	approx. 60 million	96.8%
Cohen et al. [6]		99.0%
Rajaraman et al. [25]	approx. 0.8 - 40 million	99.0%
Proposed		
Interaction-based Feature Learning/Feature Generation & Extraction (use deep convolutional layers)	on 224 by 224 by 3	
512 Features from VGG16 + NeuralNet (Single-layer)	> 131,000	98.7%
512 Features from VGG19 + NeuralNet (Single-layer)	> 131,000	98.8%
512 Features use I-score		
~Top 0.5% Modules (approx. 70 features)	approx. 8,000	98.3%
~Top 1% Modules (approx. 100 features)	approx. 9,000	98.6%
~Top 2.5% Modules (approx. 200 features)	approx. 10,000	99.3%
Removal of the above 200 features		57.4%
5216 Features from 7 different CNNs + NeuralNet (Single-layer)	> 1.31 million	99.7%
5216 Features use I-score:		
~Top 1% Modules (approx. 100 features)	approx. 13,000	98.8%
~Top 2% Modules (approx. 150 features)	approx. 19,000	99.2%
\sim Top 5% Modules (approx. 400 features)	approx. 20,000	99.7%
Removal of the above 400 features		

The performance of held-out test set is measured by the Area-Under-Curve (AUC) from Receiver Operating Characteristic (ROC). The proposed approach works with images that are resized into 224 by 224 by 3 (i.e. 150,528 pixels) and the procedure uses transfer learning to take a combination of pre-trained models such as VGG16 to generate 512 features using its deep convolutional layers. Because models like VGG16 are trained using a different set of image data and with hundreds of filters tuned from prior learning experience, there is little hope that these features grant us passages explain exactly how the explanatory variables influence the response variable which raises ambiguity for explainability and interpretability. The proposal here is to use I-score. From using I-score and Backward Dropping Algorithm, we are able to select up to 400 features to construct a neural network with less than 20,000 parameters to produce state-of-the-art test set performance at 99.7% while originally the 512 features directly from VGG16 might need 1.31 million parameters to produce equal results. This is a 98% dimension reduction on the number of parameters required for building neural network architecture. Although in our experiments, we used seven well-known deep CNN models (VGG16/19, DenseNet121/169/201, ResNet, and Xception), the proposed method can be generalized and adapted to extract informative, explainable, and interpretable features (or variables) from the last convolutional layer generated by any combination deep CNN models

of a set of variables to be the final importance score measured and evaluated by only explainable and interpretable feature assessment and selection methodology. This allows us to establish rigorous discussion on the explainability and interpretability of features and variables specifically when it comes to explain what is the influence and impact a set of variables have on response variable.

Next, this paper delivers a novel interaction-based methodology to interpret and explain deep learning models while maintaining high prediction performance. In addition, we provide a way to contribute to the major issues about explainability, interpretability, and trustworthiness brought up by DARPA. We introduce and implement a non-parametric and interaction-based feature selection methodology. Under this paradigm, we propose *Interaction-based Feature Learning* that heavily relies on using an explainable measure, I-score, to evaluate and select the explainable features that are created from deep convolutional neural networks. This approach learns from the final convolutional layer of any deep CNN or any combination of deep learning frameworks. We conclude from both simulation and empirical application results that I-score shows unparalleled potential to explain informative and influential local information in a variety of large-scale data sets. The high I-score values suggest that local information possess capability to have higher lower bounds of the predictivity. This is worth noticing because this information leads to not just high prediction performance but also explainable power. The proposed methodology *Interaction-based feature learning* rely heavily on using I-score



to select, combine, and explain features for neural network classifiers. This establishes a domain of using I-score with neural network (as well as with Backward Dropping Algorithm) which we regard as the field of the undiscovered field of Interaction-based Neural Network. Although we only point out two approaches in this paper, we do believe in the future there will be many other approaches within the field of Interaction-based Neural Network. The methodologies proposed in this paper are novel. Though the experiment is successful in the Pneumonia Chest X-ray Image data set, the paper inspires future work to investigate explainability of subsets of variables in other forms of data sets to examine a more generalized procedure of using non-parametric approaches to explain features and predictivity of features rather than relying on "blackbox" algorithms. We encourage both the statistics and computer science community to further investigate this area to deliver more transparency and trustworthiness to deep learning era and to build a world with truly Responsible A.I..

Acknowledgements We would like to dedicate this to H. Chernoff, a well-known statistician and a mathematician worldwide, in honor of his 98th birthday and his contributions in Influence Score (I-score) and the Backward Dropping Algorithm (BDA). We are particularly fortunate in receiving many useful comments from him. Moreover, we are very grateful for his guidance on how I-score plays a fundamental role that measures the potential ability to use a small group of explanatory variables for classification which leads to much broader impact in fields of pattern recognition, computer vision, and representation learning.

Authors' contributions SL and YY wrote the main manuscript text. YY prepared the experiment and the code. All authors reviewed the manuscript. Both authors read and approved the final manuscript.

Funding This research is supported by National Science Foundation BIGDATAIIS 1741191.

Data availability The data sets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability Code for data cleaning and analysis is provided upon requests as part of the proposed statistical package. It will be accessible by request once the paper has been conditionally accepted.

Declarations

Competing interests The authors declare no competing interests

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit https://creativecommons.org/licenses/by/4.0/.

References

- 1. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access. 2018;6:52138–60.
- 2. Ayan E, Ünver HM. Diagnosis of pneumonia from chest x-ray images using deep learning. In 2019 Scientific Meeting on Electrical-Electronics Biomedical Engineering and Computer Science (EBBT), 2019;1–5.
- 3. Cherian T, Mulholland EK, Carlin JB, Ostensen H, Amin R, Campo M, Greenberg D, Lagos R, Lucero M, Madhi SA, et al. Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. Bull World Health Org. 2005;83:353–9.
- 4. Chernoff H, Lo S-H, Zheng T. Discovering influential variables: a method of partitions. Ann Appl Stat. 2009;3(4):1335–69.
- 5. Chollet F. Xception: Deep learning with depthwise separable convolutions. 2016 arXiv preprint arXiv:1610.02357.
- 6. Cohen J, Bertin P, Frappier V. A web delivered locally computed chest X-ray disease prediction system. In: Proceedings of machine learning research, under review 2020;1–12.
- 7. DARPA. Broad agency announcement, explainable artificial intelligence (xai). DARPA. 2016
- 8. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017 arXiv preprint arXiv:1702.08608.
- 9. Franquet T. Imaging of community-acquired pneumonia. J Thorac Imaging. 2018;33(5):282–94.
- 10. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014;580–587.
- 11. Gong Y, Wang L, Guo R, Lazebnik S. Multi-scale orderless pooling of deep convolutional activation features. In: European conference on computer vision, 2014;392–407. Springer.
- 12. Gordo A, Almazán J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. In: European conference on computer vision, 2016;241–257. Springer.
- 13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016a;770–778.



- 14. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016b;770–778.
- 15. Huang G, Liu Z, Maaten L, KQ W. Densely connected convolutional networks. Computer Vision and Pattern Recognition. 2017.
- Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
- 17. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Handwritten digit recognition with a back-propagation network. Advances in Neural Information Processing Systems (NIPS 1989) 2. 1990
- 18. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2021;23(1):18.
- 19. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16(3):31–57.
- 20. Lo S, Zheng T. Backward haplotype transmission association algorithm—a fast multiple-marker screening method. Hum Hered. 2002;53(4):197–215.
- 21. Lo A, Chernoff H, Zheng T, Lo S-H. Why significant variables aren't automatically good predictors. Proc Natl Acad Sci. 2015;112(45):13892-7.
- 22. Lo A, Chernoff H, Zheng T, Lo S-H. Framework for making better predictions by directly estimating variables' predictivity. Proc Natl Acad Sci. 2016;113(50):14277–82.
- 23. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019;267:1–38.
- 24. Neuman MI. Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. J Hosp Med. 2012;7(4):294-8.
- 25. Rajaraman S, Candemir S, Kim I, Thoma G, Antani S. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl Sci. 2018;8(10):1715.
- 26. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
- 27. Saraiva A, Carvalho da Costa N, Sousa J, Ferreira N, Valente A, Soares S. Models of learning to classify X-ray images for the detection of pneumonia using neural networks. In Bioimaging, 2019;76–83.
- 28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014 arXiv preprint arXiv:1409.1556.
- 29. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015;1–9.
- 30. Toğaçar M, Ergen B, Cömert Z, Özyurt F. A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models. IRBM. 2020;41(4):212–22.
- 31. Wang H, Lo S-H, Zheng T, Hu I. Interaction-based feature selection and classification for high-dimensional biological data. Bioinformatics. 2012;28(21):2834–42.
- 32. Yang M, Kim B. Benchmarking attribution methods with relative feature importance. 2019 arXiv preprint arXiv:1907.09701.
- 33. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;2921–2929.

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

