Explainable Artificial Intelligence (XAI) in Chest X-ray Classification Using an Interaction-based Approach

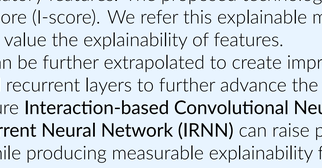
Yiqiao Yin [1]

[1]Columbia University

Abstract and Motivation

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 view of the above needs, this study proposes an interaction-based methodology – Influence Score (I-score) – to screen out the noisy and non-informative variables in the dataset hence it nourishes an environment with explainable and interpretable features that are directly associated to feature predictivity. The contribution of this paper opens a novel angle that moves the community closer to the future pipelines of XAI problems.

Figure 1. This diagram is a recreation DARPA document (DARPA-BAA-16-53) [2, 8]. The diagram presents the relationship between learning performance (usually measured by prediction performance) and effectiveness of explanations (also known as explainability).



Proposed Conditions for Explainable and Interpretable Methodology

To shed light to these questions, we define the following three necessary conditions ($\mathcal{C}1$, $\mathcal{C}2$, and $\mathcal{C}3$) for any feature selection methodology to be explainable and interpretable.

- $\mathcal{C}1$. The first condition states that the feature selection methodology do not require the knowledge of the underlying model of how explanatory variables affects outcome variable.
- $\mathcal{C}2$. An explainable and interpretable feature selection method must clearly state to what degree a combination of explanatory variables influence the response variable. Moreover, it is beneficial if a statistician can directly compute a score for a set of variables in order to make reasonable comparisons.
- $\mathcal{C}3$. In order for a feature assessment and selection technique to be interpretable and explainable, it must directly associate with the predictivity of the explanatory variables (for definition of predictivity, please see [3, 4]).

Table 1. **Explainability Satisfaction Table.** The table summarizes whether famous XAI methods and proposed I-score satisfy the definition of Explainability of a set of variables according to definition in Innovation 2.

Definition of Explainability	CAM	LIME	RISE	I-score
$\mathcal{C}1$ Non-parametric	No	No	No	Yes
$\mathcal{C}2$ Quantifiable Measure	No	Yes	Yes	Yes
$\mathcal{C}3$ Predictivity	No	No	No	Yes

Highlight of Our Work

- We propose **a novel interaction-based study** to examin the explainability and interpretability of explanatory features. The proposed technology is called Influence Measure or Influence Score (I-score). We refer this explainable measure I-score and the final assessed numerical value the explainability of features.
- The proposed **I-score** can be further extrapolated to create improved and explainable convolutional layers and recurrent layers to further advance the field of deep learning.
- The proposed architecture **Interaction-based Convolutional Neural Network (ICNN) and Interaction-based Recurrent Neural Network (IRNN)** can raise prediction performance to state-of-the-art level while producing measurable explainability for end-users.

Proposed I-score

The Influence Score (I-score) is a statistic derived from the partition retention method [1]. Consider a set of m binary features X (each feature is binary, so total of 2^m partitions) with target outcome Y. Based on these 2^m partitions created by X, we can compute I-score using the following formula.

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

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.

The concept of interaction-based Feature is initially proposed in Lo and Yin (2021) [5, 6]. In their work, the I-score is accompanied with a greedy search algorithm called the Backward Dropping Algorithm (call this algorithm $\mathcal{B}$). The features selected using I-score and BDA would be explainable according to the definition. Then the information (represented by a combination of features) can be combined using **Interaction-based Feature Engineer**.

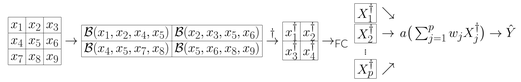
Suppose we have a supervised learning problem and we are given explanatory variables X (with k partitions) and response variable Y. We can create a novel non-parametric feature using the following formula

$$X^{\dagger} := \bar{Y}_j, \text{ while } j \in \{1, 2, ..., k\} \tag{2}$$

where k is the size of the total partitions formed by X.

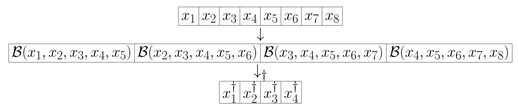
Interaction-based Convolutional Neural Network (ICNN)

Figure 2. **Interaction-based Convolutional Neural Network.** The diagram outlines the procedure of constructing convolutional layer using I-score and BDA.



Interaction-based Recurrent Neural Network (IRNN)

Figure 3. **Interaction-based Recurrent Neural Network.** The diagram outlines the procedure of constructing recurrent layer using I-score and BDA.



Application

Figure 4. This executive diagram summarizes the key components of the method: Interaction Convolutional Neural Network, proposed in this paper. This design heavily relies on the I-score and has an architecture that is interpretable at each location of the image at each convolutional layer. More importantly, the proposed design satisfies all three dimensions ($\mathcal{C}1$, $\mathcal{C}2$, and $\mathcal{C}3$ in the Introduction) of the definition of interpretability and explainability.

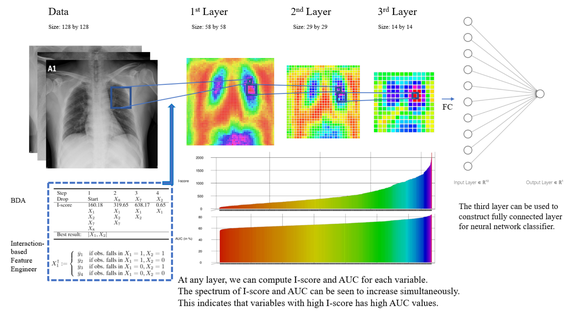


Figure 5. **Visualization for Multi-class Classification Using I-score Enhanced Deep Learning.** This figure presents 4 samples (each row is a sample with different label). There are 4 classes (0: Healthy, 1: COVID, 2: Other Pneumonia, 3: Tuberculosis).

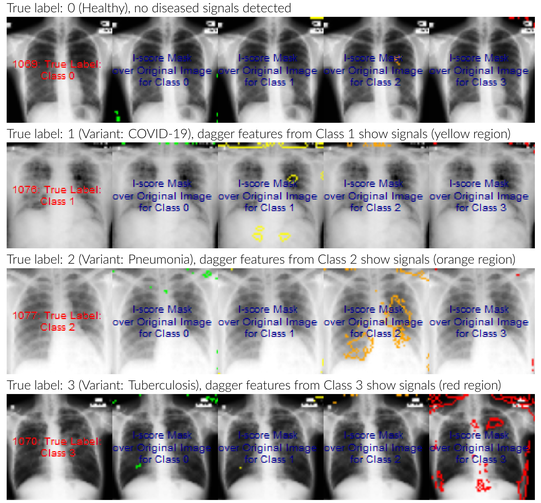


Table 2. **Multi-class Lung Cancer Variants Diagnosis.** This table presents experiment results for multi-class lung cancer variants classification. In total, there are 4 classes (0: Healthy, 1: COVID-19, 2: Pneumonia, 3: Tuberculosis).

Model	AUC (Test Set)	No. of Parameters
Benchmarks:		
ResNet [7]	0.82 - 0.90	11-25 million
Inception [7]	0.89 - 0.91	23-56 million
DenseNet [7]	0.93 - 0.94	0.8-40 million
Average	0.89	26 million
Proposed:		
ICNN (Θ_1: {starting point: 6, window size: 2 by 2, stride: 2})	0.97	12,000
ICNN (Θ_2: {starting point: 4, window size: 3 by 3, stride: 3})	0.98	13,000
ICNN (use Θ_1, Θ_2, then concatenate)	0.98	20,000
IRNN (for details, see [5] and [6])	0.99	15,000
Average	0.98	15,000

References

[1] Herman Chernoff, Shaw-Hwa Lo, and Tian Zheng. Discovering influential variables: A method of partitions. *The Annals of Applied Statistics*, 3(4):1335 – 1369, 2009.

[2] DARPA. Broad agency announcement, explainable artificial intelligence (xai). *DARPA*, 2016.

[3] Adeline Lo, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, 112(45):13892–13897, 2015.

[4] Adeline Lo, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. Framework for making better predictions by directly estimating variables' predictivity. *Proceedings of the National Academy of Sciences*, 113(50):14277–14282, 2016.

[5] Shaw-Hwa Lo and Yiqiao Yin. An interaction-based convolutional neural network (icnn) toward a better understanding of covid-19 x-ray images. *Algorithms*, 14(11), 2021.

[6] Shaw-Hwa Lo and Yiqiao Yin. Language semantics interpretation with an interaction-based recurrent neural network. *Machine Learning and Knowledge Extraction*, 3(4):922–945, 2021.

[7] Narinder Singh Punn and Sonali Agarwal. Automated diagnosis of covid-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *Applied Intelligence*, 51(5):2689–2702, 2021.

[8] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.