DOI: 10.21917/ijsc.2023.0445

EVALUATION OF LATTICE BASED XAI

Bhaskaran Venkatsubramaniam and Pallav Kumar Baruah
Department of Math and Computer Science, Sri Sathya Sai Institute of Higher Learning, India

Abstract

With multiple methods to extract explanations from a black box model, it becomes significant to evaluate the correctness of these Explainable AI (XAI) techniques themselves. While there are many XAI evaluation methods that need manual intervention, in order to be objective, we use computable XAI evaluation methods to test the basic nature and sanity of an XAI technique. We pick four basic axioms and three sanity tests from existing literature that the XAI techniques are expected to satisfy. Axioms like Feature Sensitivity, Implementation Invariance, Symmetry preservation and sanity tests like Model parameter randomization, Model-Outcome relationship, Input transformation invariance are used. After reviewing the axioms and sanity tests, we apply it on existing XAI techniques to check if they satisfy them or not. Thereafter, we evaluate our lattice based XAI technique with these axioms and sanity tests using a mathematical approach. This work proves these axioms and sanity tests to establish the correctness of explanations extracted from our Lattice based XAI technique.

Keywords:

Explainable AI, XAI, Formal Concept Analysis, Lattice for XAI, XAI Evaluation

1. INTRODUCTION

Artificial Intelligence systems are quite popular and it is everyone's dream to perform routine tasks using trained AI models. As AI continues to penetrate wider areas of tasks and is allowed to make autonomous decisions, justifying these decisions also becomes quite critical. Most often, a Machine Learning or a Deep Learning model is trained on a dataset and later deployed in production as the model improves in its accuracy. It is the property of the model to expose the function that is fit to the curve or not. Especially in Deep Learning, these functions are not explicitly available. This led to the growth of Explainable AI techniques that bring out different kinds of explanations. Some of them excel at global explanation of the model, while some bring out the best explanation around an instance [1]. There are a large number of techniques that are well tuned to images and produce heat maps as explanations [2-3], [20]. In our work, we propose a novel XAI technique using a Formal Concept Lattice [5]. This technique can generate global, local, similar and contrastive explanations and has been tested on tabular data and images [6]. It has also been compared to standard techniques to prove its credibility [7]. Lack of such explainability can put users away from adopting AI into their domain of work [8] and rather adopt standard white box models that may not suit all domains. In order to avoid such extreme opinions, it is necessary to encourage XAI techniques [9]. There is little doubt that XAI techniques will continue to rise in prominence and produce much needed work for the future [10].

As these XAI techniques continue to grow, it would be worthwhile to evaluate these techniques themselves. We need to evaluate whether the explanations are good, whether they satisfy the user and if it led to the improvement of trust in the model. There can be multiple ways an XAI technique can produce explanations in order to build trust in the model. It can evaluate if the model is performing its primary task or not. If it performs the primary task, it can bring out the mental model behind the model which can be compared by the user to his/her own mental model to understand how the model works. It can explain why a particular instance was mapped by the model to a specific outcome and why not other outcomes. It may answer questions on what happens if some of the features changed values or what change in feature values would force the model to change its decision. In effect, an XAI technique needs to have many aspects - understandability, sufficiently detailed, complete, correct, useful and trustworthy [11].

In order to measure these aspects, most often domain experts are needed to study these aspects and verify it. It needs time and resources to conduct these studies, analyze feedback forms and measure scales of these aspects. There are also computational methods to measure some of these aspects in XAI. Fidelity of an XAI technique measures the correctness of the technique in generating true explanations for model predictions [12]. One such method to measure fidelity is by comparative evaluation. It compares the XAI technique under question to an existing, well accepted, XAI technique that is already proven to do well. In [13], a set of empirical evaluations are designed to compare their technique's consistency with an existing technique like LIME. Another example is [14], where it is compared with LIME and DeepLIFT [15]. Another method to measure fidelity is to compare the explanations against inherently interpretable models or white box models. LIME [1] itself compares its explanations to inherently interpretable models like linear or logistic regression. Specific to Deep Learning models on images and saliency based XAI techniques, Congruence and Annotation classification are two metrics to measure the correctness of an XAI technique. Congruence measures the proportion of model attention within expert annotated regions, while Annotation classification measures how much of the expert annotation the model pays attention to [16]. These are similar to precision and recall specific to images. Even in these two methods it is evident that a domain expert was involved in creating the annotation apriori

There are also computational methods to measure the basic sanity of XAI techniques without involvement of a domain expert. In this work, we restrict ourselves to evaluating the XAI technique using these sanity tests and basic axioms. Section 2 introduces these tests and axioms. Section 3 reviews these metrics on existing XAI techniques. Section 4 introduces the formal concept lattice and our lattice based XAI technique. Section 5 proves the set of axioms and sanity tests for our lattice based XAI technique. Section 6 contains conclusions and future work.

2. AXIOMS AND SANITY TESTS FOR XAI

We consider three axioms and four sanity tests to evaluate an XAI technique: Feature Sensitivity-a axiom, Feature Sensitivity-

b axiom, Implementation Invariance axiom, Symmetry preserving axiom [17], Model parameter randomization sensitivity test, Input transformation invariance test and the Model-Outcome relationship sensitivity test [18].

- *Feature Sensitivity-a*: If two instances differ in one feature but have different predictions from a trained model, then the differing feature should be highlighted in the model explanation brought out from the XAI technique.
- *Feature Sensitivity-b*: If the trained model does not depend on some variable, then the explanation brought from the XAI technique should not highlight this variable.
- *Implementation Invariance*: Two trained models are functionally equivalent if their outputs are equal for all inputs, despite having different implementations. A good XAI technique should provide identical explanations for two functionally equivalent networks.
- *Symmetry preserving*: Two input variables are symmetric with respect to a function if swapping them does not change the function. An XAI technique is Symmetry preserving, if for all inputs that have identical values for symmetric variables, the symmetric variables receive identical attributions.
- *Model Parameter Randomization Sensitivity*: Explanation of a model from an XAI technique is compared to the explanation from the copy of the model with randomly initialized parameters. A good XAI technique should be able to differ substantially in its explanation output for the two cases.
- *Model-Outcome Relationship Sensitivity*: Explanation of a model from an XAI technique is compared to the explanation on the model with the same architecture but trained with the copy of the data set with permuted labels. A good XAI technique should depend on the relation between the instances and the labels.
- *Input transformation sensitivity*: If a data instance is modified such that it does not affect the model outcome, explanations for the original and the modified instance must be equivalent. A good XAI technique must demonstrate such input transformation invariance.

3. AXIOMS AND SANITY TEST EVALUATION FOR EXISTING XAI TECHNIQUES

We review the techniques presented in [17] and [18] together with the results of the three axioms and four sanity tests as stated in the previous section to evaluate an XAI technique: Feature Sensitivity-a axiom, Feature Sensitivity-b axiom, Implementation Invariance axiom, Symmetry preservation axiom, Model Parameter Randomization sensitivity test, Model-Outcome Relationship sensitivity test and Input transformation sensitivity test.

For a deep network that classifies a given input, the gradient of the output with respect to each input for each class, captures the importance of each input feature for a specific output class. The product of this gradient and the input feature values is a good starting point as attribution of a feature towards a class. But [17] states that all gradient based methods break the sensitivity-a axiom and proves it with a simple one variable network. They break this axiom as the function flattens out despite having a change of value from the one at the baseline. Hence it leads gradients to focus on irrelevant features as captured in the fireboat picture and its gradients, as shown in Fig 1 from [17].

Fig.1. Fireboat picture and its gradients

Deconvolutional Networks [19] visualize concepts learnt by neurons in higher layers of a convolutional neural network by inverting the data flow, moving from neuron activations in a specific layer back to the image. The resulting reconstructed image shows the part of the input image that is most responsible to activate this neuron. A schematic illustration from [19] of this technique is shown in Fig 2.

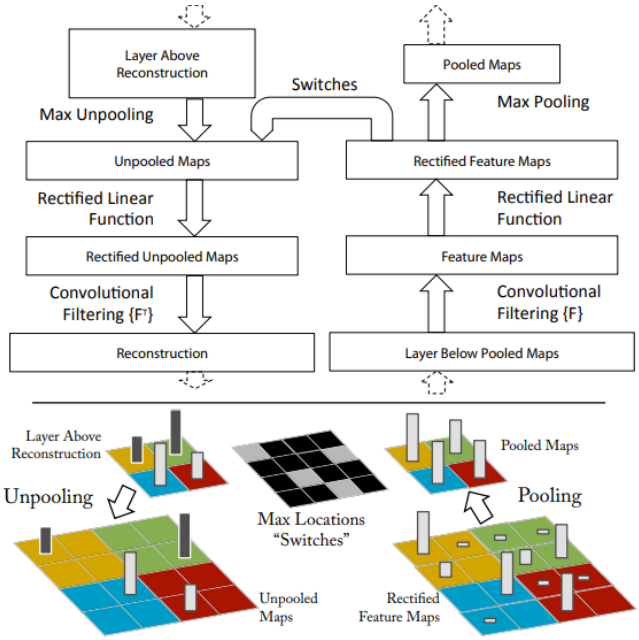

Fig.2. Schematic representation of Deconvolution from [19]

To examine a convolutional neural network, a deconvolutional network is attached to each of its layers providing a continuous path back to the image. All other activations in a layer are set to zero except the specific neuron and pass the feature maps as input to the deconvolutional layer.

Guided Backpropagation [20] modifies deconvolutional networks to make image reconstructions more accurate. It combines the techniques from backpropagation and deconvolution to reduce the attribution signal to zero at a ReLU when the gradient is negative or if the input to ReLU at the time of forward pass was negative. This idea is represented in Fig.3, as presented in [20].

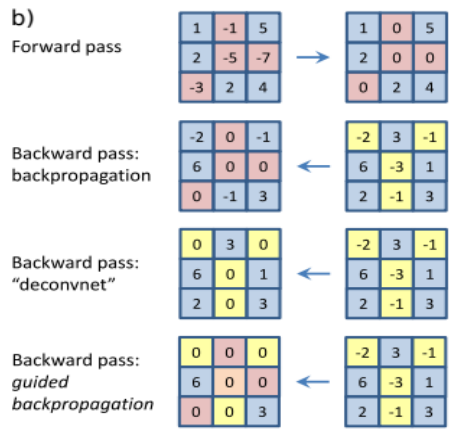

Fig.3. Combined technique of Guided Backpropagation from [20]

In effect, this combination stops negative gradients from flowing backwards, thereby not allowing those corresponding neurons to decrease the activation of the unit that is being visualized.

But [17] states that Deconvolution and Backpropagation fail sensitivity-a axiom, since both these methods back propagate through a ReLU node only if it is turned on at the input, which makes it similar to gradients. Hence the attribution can be zero for features with zero gradient at the input even though there is a non-zero gradient at the baseline.

DeepLift [15] is another technique that uses backpropagation to calculate its attribute scores but uses difference with respect to a reference state rather than instantaneous gradients. Hence it circumvents the saturation problem that the prior two techniques fall into. This technique also gives separate consideration to positive and negative contributions at non-linear units, hence revealing dependencies that may be missed out by the other techniques. The choice of reference input is critical for gaining insightful results from this technique. In [6], we compare our lattice based XAI technique to DeepLift.

Layer-wise relevance propagation [21] propagates the prediction backward in the network with a conservation property. Relevance scores at the given layer are used to calculate the relevance scores at the previous layer by using the extent to which the previous layer neuron contributed in activating the neuron in the given layer. A simple schematic diagram representing this idea is shown in Fig.4 from [21].

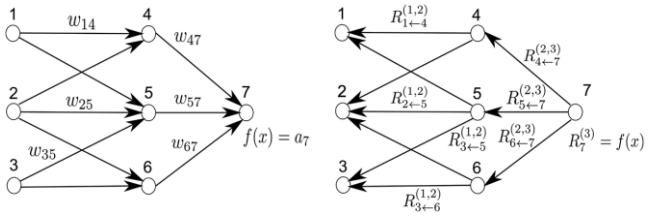

Fig.4. Layer wise relevance propagation from [21]

Moving from layer to its previous layers, the relevance scores are scaled using the weights learned in the forward pass leading to noise reduction.

But [17] states that both DeepLift and LRP fail the implementation invariance axiom. Both techniques replace instantaneous gradients by discrete gradients and still use backpropagation to compute attribution scores. Since the chain rule does not hold for discrete gradients, it is very much possible that there are multiple sets of values with which the same function can be achieved. If the network converges to one set instead of the other, it leads to different attribution scores for an equivalent network, thereby failing to satisfy implementation invariance.

Integrated Gradients [17] is a technique that combines gradients with Layer wise relevance propagation or DeepLift without their weaknesses of loss of sensitivity and implementation variance. It computes the gradients at all the points along the straight line from the baseline to the input and accumulates them all. In other words it calculates the path integral of the gradients along the straight line path from the baseline to the input. This technique satisfies sensitivity and implementation invariance. It also provides another desirable property called completeness, where the attributions add up to the difference between the network output between the baseline and the input.

Gradient-weighted class activation mapping (Grad-CAM) [2] is a technique that builds over class activation mapping without needing any architectural changes or the necessity for re-training. It enhances the class discriminative features compared to guided backpropagation. While Grad-CAM produces class discriminative low resolution feature maps, it is fused with point wise multiplication to produce Guided Grad-CAM that is both class discriminative and high resolution. In class activation mapping, the convolutional feature maps from the penultimate convolutional layer are global average pooled and linearly transformed to produce a score for each class. It computes the linear combination of the final feature maps using the learned weights of the final layer. But it needs architectural changes when there are multiple fully connected layers before the output layer. These are replaced by convolutional layers and the network is re-trained. But in Grad-CAM, the weights are directly produced by computing a global average pool of the gradients of the class score function with respect to the feature maps of a convolutional layer.

SmoothGrad [3] is another simple technique to visually sharpen gradient based heat maps. To an image of interest, it generates more samples by adding noise to the image. It then averages the heat maps for each of the sampled images. Such a technique of adding noise at training time has a de-noising effect on the heat maps.

In [18], the Inception V3 architecture trained on the ImageNet dataset was used with random weights (model parameter randomization test). Random weights were applied independently to each layer and also in a cascading pattern, i.e. applied from the beginning till the specific layer of the network. This modified network was used on different XAI techniques to test if they produced different explanations or not. The Fig 5 shows the results of this experiment with independent randomization and Fig 6 shows the results of this experiment with cascading randomization from [18].

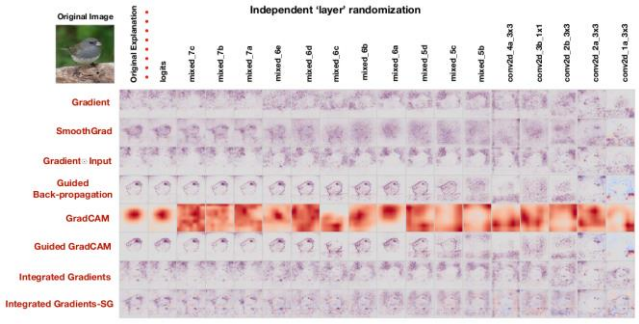

Fig.5. XAI techniques on independently random weighted Inception V3 network from [18]

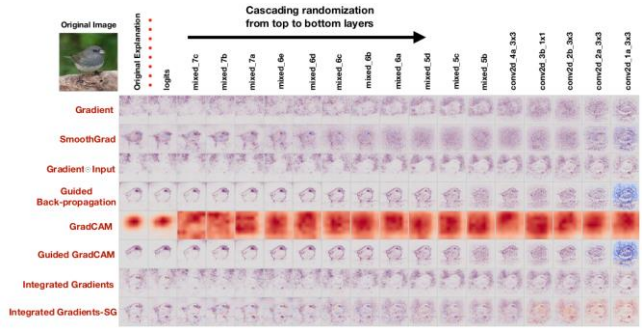

Fig.6. XAI techniques on cascading random weighted Inception V3 network from [18]

While most of the techniques pass this test on a bird image for independent randomization, it is clear that Guided backpropagation and Guided GradCAM pass the test only for the lower layers. It also shows that the gradient and GradCAM techniques are sensitive to both independent and cascading parameter randomization which is expected from a good XAI technique. Again, Guided backpropagation and Guided Grad-CAM are insensitive to the cascading model parameter randomization and are sensitive only in the last column where all the weights of the network are completely randomized. While Integrated gradients shows some promise during independent randomization, it clearly reveals part of the bird during cascading weight randomization which is not expected from a good XAI technique.

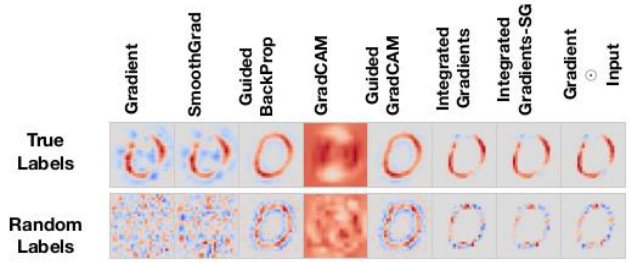

Fig.7. XAI techniques of randomly labeled MNIST dataset from [18]

Another test result from [18] is the model-outcome relationship sensitivity, when the model is trained on a dataset with permuted labels. A model achieving high training accuracy on such a permuted label dataset can only memorize the labels instead of learning the structure. If such a model is asked to be explained by an XAI technique, it should not indicate the pattern or structure behind an artifact. If it does then it does not pass the model-outcome relationship sensitivity test. The Fig.7 from [18] shows the output from different XAI techniques on the MNIST dataset.

Yet again, both Guided Backpropagation and Guided Grad-CAM techniques reveal the structure of the digit as the reason behind model classification, thereby not passing the model-outcome relationship sensitivity test. While gradient and its SmoothGrad variants show random pixels, Grad-CAM shows disconnected patches, convincingly passing the test. Integrated Gradients and its SmoothGrad variant also show change in the sign of the attributions, but yet reveal the structure. These two techniques cannot be considered to have passed the test convincingly.

We summarize our review in Table.1 for the Feature Sensitivity-a axiom, Feature Sensitivity-b axiom, Implementation Invariance axiom, Symmetry preservation axiom, Model Parameter Randomization sensitivity test and Model-Outcome Relationship sensitivity test.

Table.1. Summary of the review

	I	II	III	IV	V	VI
Gradient	X	✓	✓	✓	✓	✓
Guided Back Propagation	✓	✓	✓	✓	X	X
Deconvolution	X	✓	✓	✓	-	-
GradCAM	✓	✓	✓	✓	✓	✓
Guided GradCAM	✓	✓	✓	✓	X	X
Integrated Gradients	✓	✓	✓	✓	X	X
DeepLIFT	✓	✓	X	✓	-	-
Layer wise Relevance Propagation	✓	✓	X	✓	-	-

The Table.1 makes it clear that mere visual inspection of the heatmaps does not prove its credibility. Though GradCAM satisfies all the axioms and passes the tests, it can produce only low resolution heatmaps. But its associated guided GradCAM technique, which can produce high resolution heatmaps, does not pass all the tests.

3.1 XAI USING THE FORMAL CONCEPT LATTICE

A context is a triple (G,M,I), where G is a set of objects, M is a set of attributes and I the relation between them. The notation gIm means that the object g has the attribute m.

For a set $A \subseteq G$, define $A' = \{m \in M \mid gIm \ \forall \ g \in A\}$ [A' is the set of attributes common to all the objects in A]

For a set $B \subseteq M$, define $B' = \{g \in G \mid gIm \ \forall \ m \in B\}$ [B' is the set of objects which have all attributes in B]

A concept of the context (G,M,I) is a pair (A,B) such that, $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. A is called the extent and B the intent of the concept (A,B).

If $(A1,B1)$ and $(A2,B2)$ are concepts of a context (G,M,I), then $(A1,B1)$ is a subconcept of $(A2,B2)$ (or $(A1,B1)$ is a superconcept

of (*A*2,*B*2)), denoted by (*A*1,*B*1) ≤ (*A*2,*B*2) (or (*A*2,*B*2) ≤ (*A*1,*B*1)) if A1 ⊆ A2, equivalently *B*2 ⊆ *B*1 (or *A*2 ⊆ *A*1, equivalently *B*1 ⊆ B2). The relation ≤ is called the hierarchical order of the concepts. The ordered set of concepts is called the concept lattice of the context (*G*,*M*,*I*). Concept lattices are represented using a hasse line diagram [4]. The Table.2 contains a simple formal context and Fig 8, its concept lattice.

Table.2. Formal context of a few species with their attributes (columns split into two parts)

	Breathes in water (a)	Can fly (b)	Has beak (c)	Has hands (d)	Has skeleton (e)
Bat		X			X
Eagle		X	X		X
Monkey				X	X
Parrot Fish	X		X		X
Penguin			X		X
Shark	X				X
Lantern Fish	X				X

	Has wings (f)	Lives in water (g)	Is viviparous (h)	Produces light (i)
Bat	X		X	
Eagle	X			
Monkey			X	
Parrot Fish		X		
Penguin	X	X		
Shark		X		
Lantern Fish		X		X

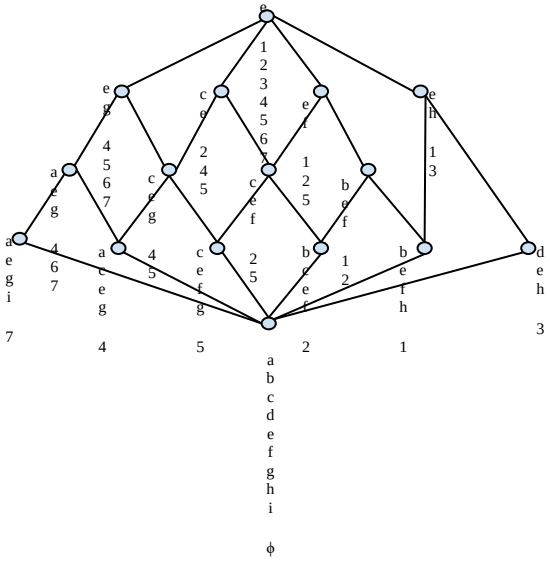

Fig.8. Concept Lattice of the formal context in Table 2

In [5], we first build a formal concept lattice from the given dataset and extract implications. With the instance support of each implication and a user provided implication cutoff, we generate a synthetic dataset that respects all the implications whose support is greater than or equal to the implication cutoff. We then build a formal concept lattice with this newly generated synthetic dataset. The crux behind this step is to cover all kinds of data points without generating unnecessary points that were not implied from the original dataset. We use this formal concept lattice to extract global, local, similar and contrastive explanations of a black box model around an instance of interest. Model outcome of the data instances in the synthetic dataset formal concept lattice is communicated to all its super concepts. At each node the union of all the sub concept outcomes is computed and communicated recursively. For local, similar and contrastive explanation, this lattice is traversed to find minimum feature combinations that lead to a specific outcome. Given a data instance, we find the path from the root of the lattice to the specific data instance recording the change in the set of outcomes. Multiple paths from the root to the instance indicate equivalent change in outcomes. Using this invariant/varying set of outcomes, as the lattice is traversed, we generate the contrastive and similar explanations that explain the set of features that impacted a change or not.

4. AXIOMS AND SANITY TEST EVALUATION FOR LATTICE BASED XAI TECHNIQUE

We prove the correctness for all the chosen axioms and sanity tests for our lattice based XAI technique.

4.1 FEATURE SENSITIVITY - A

This axiom states that if two instances differ in a specific feature and have different predictions from the model, then this feature should be brought out in the explanation presented by the XAI technique under consideration.

A context is a triple (*G*,*M*,*I*), where G is a set of objects, M is a set of attributes and *I* the relation between them. A concept of the context (*G*,*M*,*I*) is a pair (*A*,*B*) such that, $A \subseteq G$, $B \subseteq M$, $A^{\prime}=B$ and $B^{\prime}=A$, where *A* is the extent and B the intent of the concept (*A*,*B*). In our work, we consider a feature and its specific value together as an attribute. Hence *M* consists of a feature F_i with its value V_i.

Let us consider two concepts (A_1,B_1) and (A_2,B_2) that are part of the formal concept lattice, with outcomes C_1 and C_2 respectively, such that, $|B_1|=|B_2|=|M|$ and $C_1 \neq C_2$. Let us also assume (F_i,V_1) ∈ B_1 and (F_i,V_2) ∈ B_2, such that, $V_1 \neq V_2$ and same values for rest of the features F_j, for $j \neq i$. From the definition of the formal concept lattice, it implies that there must be a concept (A_3,B_3), where $A_3=A_1 \cup A_2$ and $B_3=B_1 \cap B_2$, with $|B_3|=|M|-1$. B_3 consists of all attributes except the feature F_i. After lattice construction, the outcomes are communicated from subconcepts to superconcepts. Each superconcept maintains a union of all the outcomes it receives and passes it below to its superconcepts recursively. This implies that the outcomes computed at the concept (A_3,B_3) must contain {C_1,C_2}. In drawing contrastive explanation from the lattice, the traversal records changes in outcome in the lattice, specifically, where nodes contain outcomes that differ from the outcomes of their subconcepts. The concept (A_3,B_3) would be found in this process as its outcomes contain

$\{C_1,C_2\}$, whereas its subconcept (A_1,B_1) has outcome C_1 and its subconcept (A_2,B_2) has outcome C_2. This implies that the contrastive explanation brings out feature F_i with value V_1 leading to outcome C_1 and changing F_1 to value V_2 would lead to outcome C_2, thereby highlighting the differing feature, proving the feature sensitivity - a axiom.

4.2 FEATURE SENSITIVITY - B

This axiom states that if two instances differ in a specific feature but have the same predictions from the model, then this feature should not be brought out in the explanation presented by the XAI technique under consideration.

Let us consider two concepts (A_1,B_1) and (A_2,B_2) that are part of the formal concept lattice, with outcomes C_1 for both, such that, $|B_1|=|B_2|=|M|$. Let us also assume $(F_i,V_1) \in B_1$ and $(F_i,V_2) \in B_2$, such that, $V_1 \neq V_2$ and same values for rest of the features F_j, for $j \neq i$. From the definition of the formal concept lattice, it implies that there must be a concept (A_3,B_3), where $A_3=A_1 \cup A_2$ and $B_3=B_1 \cap B_2$, with $|B_3|=|M|$-1. B_3 consists of all attributes except the feature F_i. After lattice construction, the outcomes are communicated from subconcepts to superconcepts. Each superconcept maintains a union of all the outcomes it receives and passes it below to its superconcepts recursively. This implies that the outcomes computed at the concept (A_3,B_3) must contain $\{C_1\}$. In drawing contrastive explanation from the lattice, the traversal seeks to find concepts in the lattice where nodes contain outcomes that differ from the outcomes of their subconcepts. The concept (A_3,B_3) would not be found in this process as its outcome is $\{C_1\}$, the same outcome as both its subconcepts (A_1,B_1) and (A_2,B_2). In fact, these nodes would be found while drawing out similar explanation, stating that changing feature F_i from value V_1 to V_2 does not change the outcome. This implies that the contrastive explanation will not bring out feature F_i, thereby not highlighting the differing feature, proving the feature sensitivity - b axiom.

4.3 IMPLEMENTATION INVARIANCE

The axiom states that if two trained models are functionally equivalent then the XAI technique should provide identical explanations for the two.

Let F_1 and F_2 be two different models that predict the same outcome for all the data instances of any dataset. In our lattice based XAI technique, the formal concept lattice is constructed on the generated synthetic dataset respecting implications based on user provided implication cutoff. In this lattice, we communicate the outcomes of the dataset instances to their superconcepts recursively. Since the lattice is constructed on the generated synthetic dataset, which is common for both the models, the constructed lattices will be identical for both F_1 and F_2. Since the outcomes match for both models F_1 and F_2 for any data instance, the set of outcomes that is communicated from each data instance to the superconcepts would also be identical throughout the lattice. Effectively, this means that both the lattice and the outcomes are identical for F_1 and F_2. Explanations are drawn out from the lattice and hence it would be identical for a data instance across F_1 and F_2 as their lattices are identical. Let us assume that the outcome of F_1 and F_2 differ for only one data instance I. If so, then the set of collected outcomes would also differ in the two lattices and hence the path traversed in order to generate the explanation would also differ, hence modifying the explanation for F_1 and F_2. But when there are no differing outcomes, explanations are identical. This proves implementation invariance of the lattice based XAI technique.

4.4 SYMMETRY PRESERVING

This axiom states that two input variables are symmetric with respect to a function if swapping them does not change the function outcome. An XAI technique is Symmetry preserving, if for all inputs that have identical values for symmetric variables, the symmetric variables receive identical attributions.

This is trivially proven for the lattice based XAI technique as explanations are drawn from a formal concept lattice, where a concept consists of a pair (A,B) such that, $A \subseteq G$, $B \subseteq M$, $A'=B$ and $B'=A$, of the context (G,M,I). In a subset of features and their values, the order of their presence is not considered, proving the symmetry preserving nature.

4.5 MODEL PARAMETER RANDOMIZATION SENSITIVITY

In this sanity test, explanation of a model from an XAI technique is compared to the explanation from the copy of the model with randomly initialized parameters. A good XAI technique should differ substantially in their explanation of the two cases.

This test assumes without stating that on random initialization of a deep learning model, the outcome of the model differs substantially compared to the outcomes from the properly trained model. Let us assume a model F_1 which has been properly trained and let its outcome be O for the data instance I. Let F_2 be the model that has the same design as F_1, but has been initialized randomly and let its outcome be O' for the data instance I. In the lattice constructed for F_1, the data instance I would have outcome O and it follows that all its superconcepts would also have O in their set of outcomes. In the lattice constructed for F_2, the data instance I would have outcome O' and it follows that all its superconcepts would also have O' in their set of outcomes. While the concept lattice would remain identical, the set of outcomes gathered at each node would differ substantially between the two lattices. In deriving an explanation for the data instance I, the paths traversed will be similar in both the lattices, but the explanation produced will differ as the set of outcomes in the nodes differ. This proves that the lattice based XAI technique is sensitive to model parameter randomization.

4.6 MODEL-OUTCOME RELATIONSHIP SENSITIVITY

In this sanity test, explanation of a model from an XAI technique is compared to the explanation on the model with the same architecture but trained with the copy of the data set with permuted labels. A good XAI technique should depend on the relation between the instances and the labels and if there is a change in the label, the explanation should also differ appropriately. Let F_1 be the model trained on the original dataset and its labels. Let F_2 be the model trained on the dataset with randomly permuted labels. Let us consider a data instance I which has label L_1 in F_1 and with label L_2 in F_2. Since the set of data instances are the same there will not be any difference between the lattices of F_1 and F_2. But there will be differences in the set of

collected outcomes as the labels differ between F_1 and F_2. In the data instance I and its path to the root of the lattice, the collected labels would have L_1 for the lattice of F_1 and L_2 for the lattice of F_2. When considering the explanation for the data instance I, the lattice based technique will traverse identical paths for F_1 and F_2 but since the labels across these paths differ, there will be a clear difference in the explanations, thereby proving model-outcome relationship sensitivity.

4.7 INPUT TRANSFORMATION SENSITIVITY

In this sanity test, if a data instance is modified such that it does not affect the model outcome, explanations for the original and the modified instance must be equivalent. A good XAI technique must demonstrate such input transformation invariance. Let us assume a model F trained on a dataset with n features. Let I be a data instance with n features and its specific values, say $\{(f_1,v_1), (f_2, v_2), \ldots, (f_n,v_n)\}$. Let us consider an index set J, such that, there is another instance I' with I' having same values as I on all $i \notin J$ and different values on all $i \in J$. For example, if $J=\{2\}$, then $I'=\{(f_1,v_1), (f_2,v_2'), \ldots, (f_n,v_n)\}$. Let us assume that the model F produces the outcome O for both I and I'. In the lattice constructed from the dataset, there will be a node K with features $\{(f_i,v_i)\}$, for $i \notin J$ with outcome O as part of its set of outcomes. In the traversal from the root of the lattice to the instance I or I', the rest of the path must be the same except the path from node K to the specific instance. Hence the explanation will be the same for I and I' from the root till node K. From node K to the specific instance, the explanation would state different values on the feature index set J without any change in the outcome set. Thus while one can observe a change in values of features in the index set J, there will not be any change in the outcome, which proves Input transformation sensitivity.

5. CONCLUSION AND FUTURE WORK

It is clear that our Lattice based XAI technique satisfies all the axioms and passes all sanity tests. Our earlier approach [5] proved some of these empirically, while in this work, we have used a mathematical approach to prove these. This clearly proves that our Lattice based approach to generate explanations is accurate, correct and hence reliable. Evaluation of XAI techniques is an emerging area and more such fundamental axioms and sanity tests need to be standardized to evaluate any XAI technique. Such standardization will clearly bring out the quality of an XAI technique. We also intend to apply many more XAI evaluation techniques, including methods that need human involvement, in order to further prove the credibility of our Lattice based XAI technique.

REFERENCES

[1] M.T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.

[2] R. Selvaraju, R. Ramprasaath, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra, "Grad-Cam: Visual Explanations from Deep Networks via Gradient-based Localization", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618-626, 2017.

[3] D. Smilkov and M. Wattenberg, "SmoothGrad: Removing Noise by Adding Noise", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 18-26, 2017.

[4] R. Wille, "Concept Lattices and Conceptual Knowledge Systems", *Computers and Mathematics with Applications*, Vol. 23, pp. 493-515, 1992.

[5] Bhaskaran Venkatsubramaniam and Pallav Kumar Baruah, "A Novel Approach to Explainable AI using Formal Concept Lattice", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 11, No. 7, pp. 1-13, 2022.

[6] Bhaskaran Venkatsubramaniam and Pallav Kumar Baruah, "XAI using Formal Concept Lattice for Image Data", *ICTACT Journal on Image and Video Processing*, Vol 13, No. 3, pp. 2904-2913, 2023.

[7] Bhaskaran Venkatsubramaniam and Pallav Kumar Baruah, "Comparative Study OF XAI using Formal Concept Lattice and LIME", *ICTACT Journal on Soft Computing*, Vol 13, No. 1, pp. 2782-2791, 2022.

[8] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead", *Nature Machine Intelligence*, Vol. 1, No. 5, pp. 206-215, 2019.

[9] Alejandro Barredo Arrieta, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila and Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI", *Information Fusion*, Vol. 58, pp. 82-115, 2020.

[10] O. Biran and C. Cotton, "Explanation and Justification in Machine Learning: A Survey", *Proceedings of Workshop on Explainable Artificial Intelligence*, pp. 1-6, 2017.

[11] R.R. Hoffman and Jordan Litman, "Metrics for Explainable AI: Challenges and Prospects", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1-14, 2018.

[12] Sina Mohseni, Niloofar Zarei and Eric D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems", *ACM Transactions on Interactive Intelligent Systems*, Vol. 11, No. 3-4, pp. 1-45, 2021.

[13] Andrew Slavin Ross, Michael C. Hughes and Finale Doshi-Velez, "Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations", *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 2662-2670, 2017.

[14] S.M. Lundberg and S.I. Lee, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765-4774, 2017.

[15] Avanti Shrikumar, Peyton Greenside and Anshul Kundaje, "Learning Important Features through Propagating Activation Differences", *Proceedings of International Joint Conference on Machine Learning*, pp. 3145-3153, 2017.

[16] Oliver Zhang, Randall J. Lee, Yiran Chen and Xiao Hu, "Explainability Metrics of Deep Convolutional Networks

for Photoplethysmography Quality Assessment", *IEEE Access*, Vol. 9, pp. 29736-29745, 2021.

[17] M. Sundararajan and Q. Yan, "Axiomatic Attribution for Deep Networks", *Proceedings of International Conference on Machine Learning*, pp. 3319-3328, 2017.

[18] J. Adebayo and B. Kim, "Sanity Checks for Saliency Maps", *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 9525-9536, 2018.

[19] M.D. Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks", *Proceedings of International Joint Conference on Computer Vision*, pp. 818-833, 2014.

[20] Springenberg, Jost Tobias, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, "Striving for Simplicity: The all Convolutional Net.", *Proceedings of International Joint Conference on Computer Vision*, pp. 1-8, 2014.

[21] Sebastian Bach, Frederick Klauschen, Klaus-Robert Muller and Wojciech Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation", *PloS One*, Vol. 10, No. 7, pp. 1-12, 2015.