A Novel Approach to Explainable AI using Formal Concept Lattice

Bhaskaran Venkatsubramaniam, Pallav Kumar Baruah

Abstract*: Current approaches in explainable AI use an interpretable model to approximate a black box model or use gradient techniques to determine the salient parts of the input. While it is true that such approaches provide intuition about the black box model, the primary purpose of an explanation is to be exact at an individual instance and also from a global perspective, which is difficult to obtain using such model based approximations or from salient parts. On the other hand, traditional, deterministic approaches satisfy this primary purpose of explainability of being exact at an individual instance and globally, while posing a challenge to scale for large amounts of data. In this work, we propose a deterministic, novel approach to explainability using a formal concept lattice for classification problems, that reveal accurate explanations both globally and locally, including generation of similar and contrastive examples around an instance. This technique consists of preliminary lattice construction, synthetic data generation using implications from the preliminary lattice followed by actual lattice construction which is used to generate local, global, similar and contrastive explanations. Using sanity tests like Implementation Invariance, Input transformation Invariance, Model parameter randomization sensitivity and model-outcome relationship randomization sensitivity, its credibility is proven. Explanations from the lattice are compared to a white box model in order to prove its trustworthiness.

***Keywords*: Explainable AI, Deterministic methods for XAI, Concept Lattice, Formal Concept Analysis, Lattice explanation for black box models, Lattice for XAI, XAI.**

I. INTRODUCTION

A Machine Learning or a Deep Learning model learns patterns from training data and predicts an outcome for an instance or maps an instance to a class. In a black box model, the learnt patterns of data are not evident and the reasons why a model decided an outcome is not clear. In order to make these models trustworthy and therefore acceptable, it is necessary to augment the model with explanations of its decisions. Current approaches in explainable AI either approximate the black box model with an inherently interpretable model [15] or use an unified approach [3], or produce saliency visualizations using class activation maps [11][12][13].

Inherent to these approaches, there are several drawbacks that challenge their usability, specifically in sensitive domains like medicine. While approximating a black box model by a linear model, it can be locally faithful, an aggregate of the linear model's weights need not necessarily represent the black box model [5]. While linear models consider each feature independent from the other, it may not be inherently true in the dataset. Visualizations are good at generating intuition about an instance but there are no rigorous validations to extrapolate these intuitions to the entire model [2]. While heat map visuals depict where the model is looking, most often human intuition fills up what the model is looking at. Many techniques indicate the reasons why an instance belongs to a specific class, while not presenting reasons for the instance not being in another class. Presenting a heat map or weights as explanations needs user expertise to understand the explanations themselves [7]. Some of these techniques are not invariant to functionally equivalent implementation or non-intrusive input transformations [14], while some are not sensitive to model parameter or model-outcome relationship randomization [1]. In this work, we propose a deterministic, data driven, model agnostic technique to extract an explanation by building a formal concept lattice of the dataset and the model outcome. It derives several combinations of features and their values responsible for an outcome retention or change. It not only extracts a set of feature implications to different classes acting as global explanation rightfully representing the black box model, but also produces a hierarchy of minimal feature combinations of an instance that lead to the model's decision of a specific class and not other classes acting as local explanation. Apart from global and local explanations, this technique can also provide similar and contrastive explanations around an instance. Presenting textual, detailed, features-to-class intuitive implications does not weigh heavy on user expertise to understand. Feature combinations that distinguish an instance away from other classes are considered salient while feature combinations that do not contribute to this distinction do not appear in the explanation. Implications gathered from the dataset and the model outcome act as rigorous validations enabling extrapolation of the global explanation to the entire model. Meaningful relationships in the dataset are utilized to build the lattice on realistic synthetically generated data instead of unrealistic random data. In similar works that use a lattice for explanation, [10] use the lattice to guide samples chosen by LIME, while [9] build a lattice based model independent of the black box model and compare it with the classification model to produce only individual features responsible for the decision.

Manuscript received on 26 May 2022.
Revised Manuscript received on 02 June 2022.
Manuscript published on 30 June 2022.
* Correspondence Author

Bhaskaran Venkatsubramaniam*, Department of Math and Computer Science, Sri Sathya Sai Institute of Higher Learning, Muddenahalli (Karnataka), India. Email: vbhaskaran@sssihl.edu.in

Prof. Pallav Kumar Baruah, Department of Math and Computer Science, Sri Sathya Sai Institute of Higher Learning, Puttaparthi (Andhra Pradesh), India. Email: pkbaruah@sssihl.edu.in

The proposed technique is different from the above as it constructs the lattice on the dataset and the model outcome producing combinations of features responsible for the model's decision in the form of multiple explanations.

The structure of the paper is in the following order. Section II introduces terms in Formal Concept Lattice, while Section III describes our novel technique to extract explanations using a lattice, illustrating it with a simple dataset. Section IV proves that this technique passes sanity tests for a known non-linear model, while Section V evaluates lattice based explanations by comparing it to a white box model on a popular dataset. Section VI contains conclusions and future work.

II. 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 $A^{/} = \{ m \epsilon M \mid gIm \; \forall g \epsilon A\}$ [$A^{/}$ is the set of attributes common to all the objects in A]

For a set $B \subseteq M$, define $B^{/} = \{ g \epsilon G \mid gIm \; \forall m \epsilon B\}$ [$B^{/}$ 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 (A_1,B_1) and (A_2,B_2) are concepts of a context (G,M,I), then (A_1,B_1) is a subconcept of (A_2,B_2) (or (A_1,B_1) is a superconcept of (A_2,B_2)), denoted by $(A_1,B_1) \leq (A_2,B_2)$ (or $(A_2,B_2) \leq (A_1,B_1)$) if $A_1 \subseteq A_2$, equivalently $B_2 \subseteq B_1$ (or $A_2 \subseteq A_1$, equivalently $B_1 \subseteq B_2$). The relation $\leq$ 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 [8].

Table-I illustrates a formal context and Fig 1 its concept lattice.

III. THE NOVEL APPROACH OF GENERATING EXPLANATION FROM THE FORMAL CONCEPT LATTICE

In this work, we propose a novel approach to extract different kinds of explanations of a trained black box model using the formal concept lattice of the dataset in Table-I [6]. At this point, this work can be applied to a dataset with categorical features and on a model performing a classification task. The proposed technique involves five steps.

A. Lattice construction and feature implication extraction

Standard concept generation algorithms work on binary featured datasets to generate concepts. Generally, categorical features are converted to binary valued features using standard techniques like one-hot encoding before concept generation [6][8]. To keep programming simple and not increase features, we modified the concept generation algorithm to match a feature along with its value to directly work with multi-valued features instead of converting to binary features. Implications are extracted from the lattice using an implication finding algorithm [6] along with the number of instances that support the implication and redundant implications are eliminated [4]. Few verifiable feature-to-feature implications with their instance support from the constructed lattice for the formal context from Table-I are as follows (⇒ stands for "implies"):

((Lives in water 1)) ⇒ ((Can fly 0)(Has hands 0)(Has skeleton 1)(Is viviparous 0)) (4)

((Breathes in water 1)) ⇒ ((Can fly 0)(Has hands 0)(Has skeleton 1)(Has wings 0)(Lives in water 1)(Is viviparous 0)) (3)

((Has beak 0)(Lives in water 1)) ⇒ ((Breathes in water 1)(Can fly 0)(Has hands 0)(Has skeleton 1)(Has wings 0)(Is viviparous 0)) (2)

((Has hands 1)) ⇒ ((Breathes in water 0)(Can fly 0)(Has beak 0)(Has skeleton 1)(Has wings 0)(Lives in water 0)(Is viviparous 1)(Produces light 0)) (1)

((Can fly 1)) ⇒ ((Breathes in water 0)(Has hands 0)(Has skeleton 1)(Has wings 1)(Lives in water 0)(Produces light 0)) (2)

((Has beak 1)(Has wings 1)) ⇒ ((Breathes in water 0)(Has hands 0)(Has skeleton 1)(Is viviparous 0)(Produces light 0)) (2)

These implications are easily verifiable from the formal context.

B. Synthetic Data Generation

Each implication is accompanied by the number of instances that support it. While some of these may be realistic, some of them with lower support may not be true to the actual event described by the dataset. With these implications and an implication cutoff (> 0, defined by the user at runtime), we generate synthetic data for each column from its range of values respecting all implications whose instance support is greater than or equal to the implication cutoff. In Table-II, we present the number of instances in the generated synthetic dataset for different implication cutoff values. For example, for an implication cutoff value 6, since the only implication with support 6 for the formal context from Table-I, is every species 'Has skeleton', different possibilities of feature values are generated for the remaining 8 features except 'Has skeleton', leading to a count of 256. It is to be noted that the relation between the implication cutoff and synthetic dataset instance count depends on the implications from the dataset and hence case specific.

C. Lattice construction with synthetic dataset and feature implication extraction

Step A is repeated with the generated synthetic dataset. We illustrate this with an implication cutoff value 1, leading to a synthetic dataset same as the original. With the modification that considers a feature with each of its values, the lattice in Fig 2 is more informative than the one shown in Fig 1. The generated implications are applied at each node to obtain the reduced set of features at each node. The lattice with the reduced set of features is shown in Fig 3.

Table-I: Formal context of a few species with their attributes

	Breathes in water (a)	Can fly (b)	Has beak (c)	Has hands (d)	Has skeleton (e)	Has wings (f)	Lives in water (g)	Is viviparous (h)	Produces light (i)
Bat		X			X	X		X	
Eagle		X	X		X	X			
Monkey				X	X			X	
Parrot Fish	X		X		X		X		
Penguin			X		X	X	X		
Shark	X				X		X		
Lantern Fish	X				X		X		X

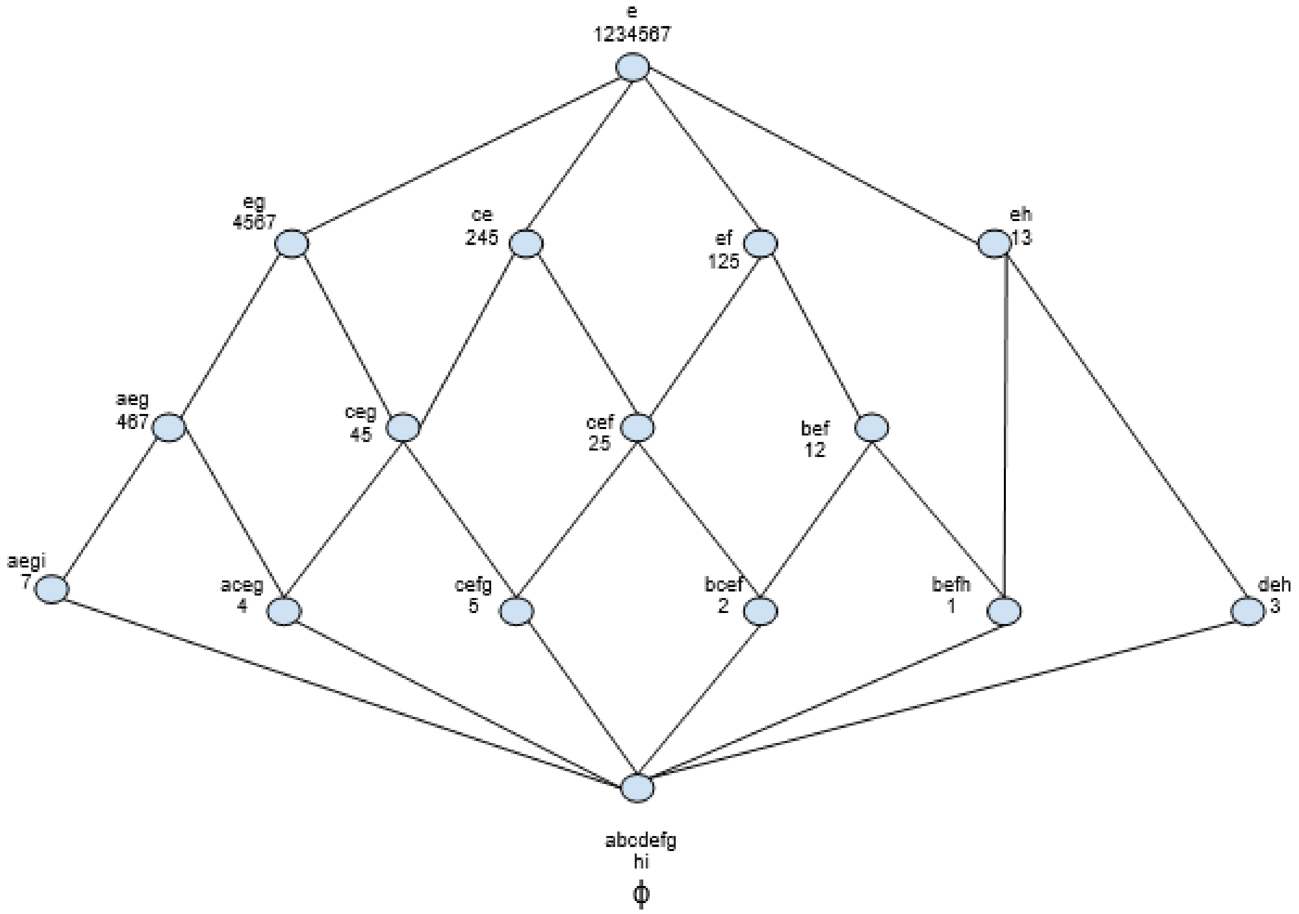


Fig. 1. Concept Lattice of the formal context in Table-I

Table-II: Synthetic Data Instance Count for different implication cutoff values

Implication Cutoff	1	2	3	4	5	6
Synthetic Dataset Instance Count	7	7	8	42	192	256

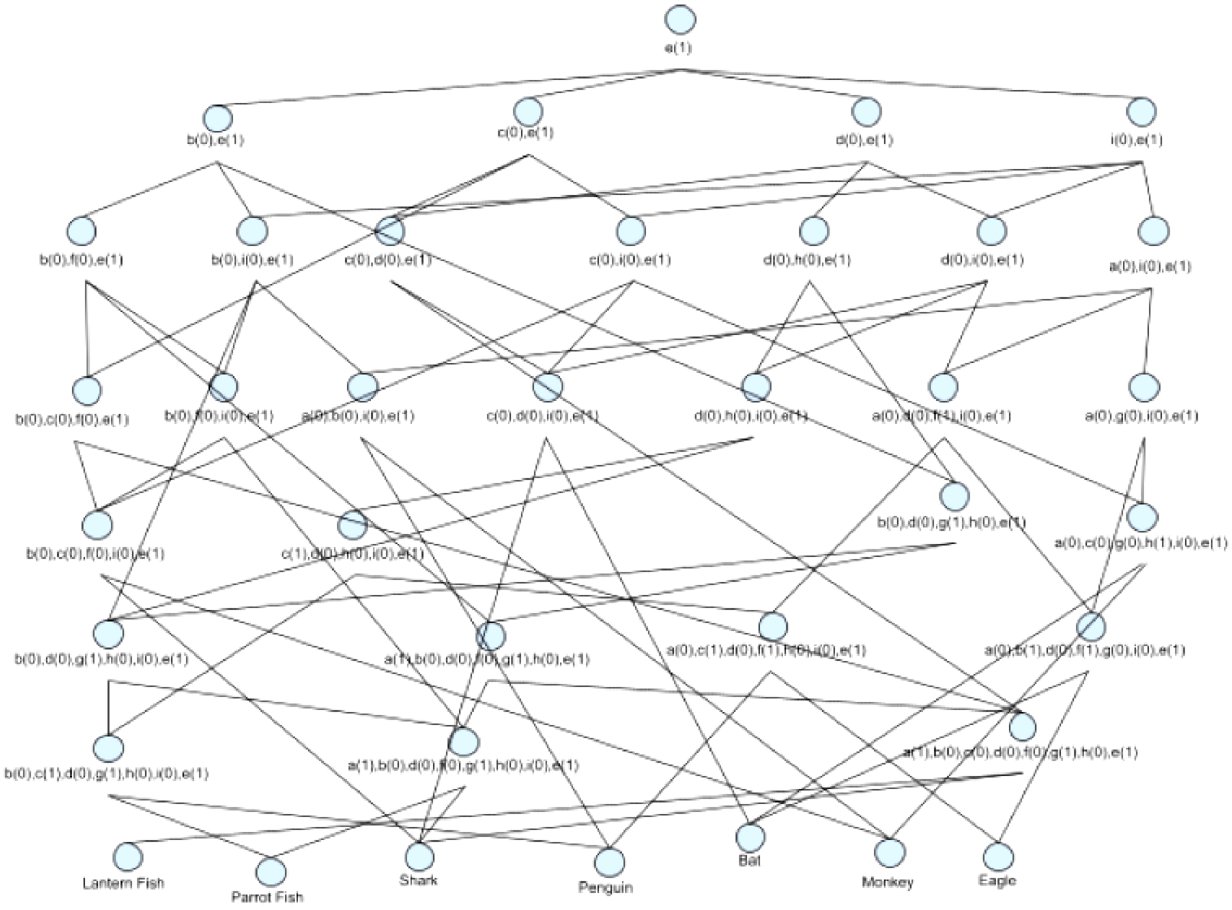


Fig. 2. Concept lattice considering every value of a feature (the last level with an empty object set can be assumed)

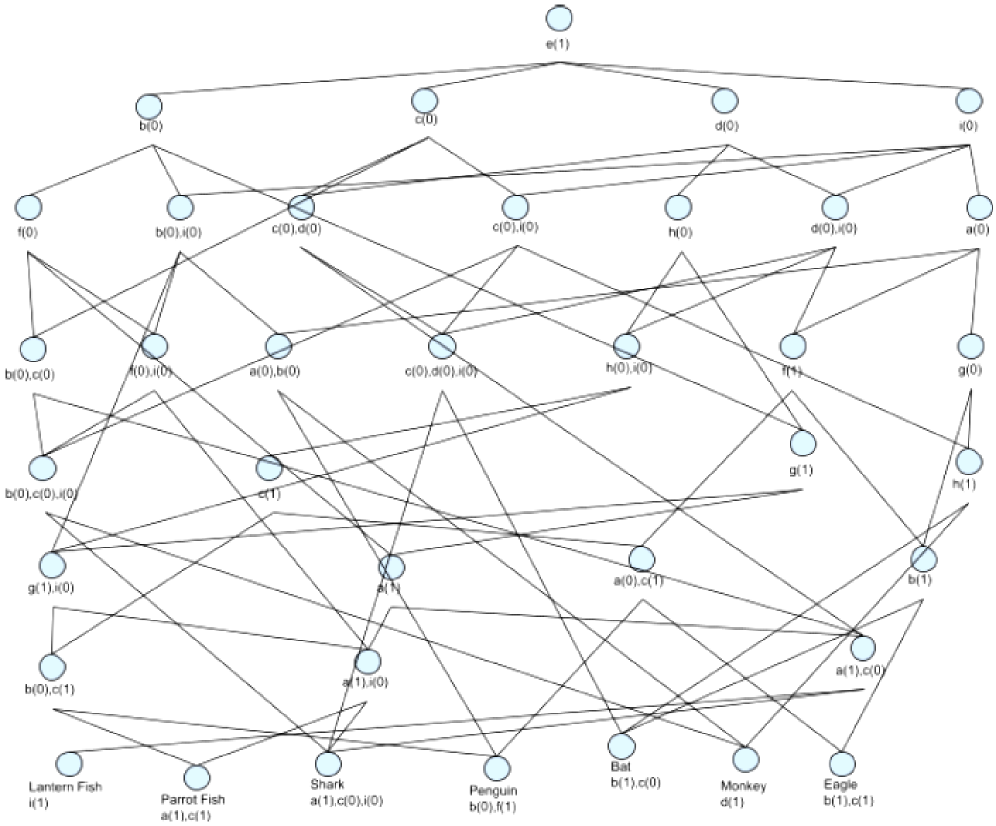


Fig. 3. Lattice with reduced set of features (the last level with an empty object set can be assumed)

D. Communicating outcomes from dataset instances

The term dataset refers to the synthetic dataset henceforth. Each instance of the dataset will be a node in the lattice. The model outcome is obtained for these instances and communicated to all its superconcepts. At each node, the union of outcomes of its subconcepts is computed. This process is continued recursively until all the nodes in the lattice gather their set of outcomes. With this information, the minimum set of features that imply a specific outcome or deny a set of outcomes is derived. These are presented as a global explanation of the model. To illustrate this, let us assume a black box model trained on the dataset in Table-III. The model outcome of each data instance in the lattice is sent to all its superconcepts. The union of these are gathered and shown in Fig. 4. Global explanation of the black box model is derived from each node of the lattice. If there is only one outcome at a node, that set of features positively leads to that outcome, while if an outcome is not present at a node, that set of features cannot lead to that outcome. These are presented as feature-to-class implications (⇒ stands for "implies" and ~ for "not"):

1. *(Has beak,0) ⇒ ~Bird*
2. *(Has wings,0) ⇒ ~Bird*
3. *(Is viviparous,0) ⇒ ~Mammal*
4. *(Breathes in water,0) ⇒ ~Fish*
5. *(Breathes in water,1) ⇒ ~Bird*
6. *(Breathes in water,1) ⇒ ~Mammal*
7. *(Breathes in water,1) ⇒ Fish*
8. *(Is viviparous,1) ⇒ ~Fish*
9. *(Is viviparous,1) ⇒ ~Bird*
10. *(Is viviparous,1) ⇒ Mammal*
11. *(Breathes in water,0)(Has beak,1) ⇒ ~Mammal*
12. *(Breathes in water,0)(Has beak,1) ⇒ ~Fish*
13. *(Breathes in water,0)(Has beak,1) ⇒ Bird*

These implications are easily verifiable from the dataset in Table-III.

E. Lattice traversal for an instance explanation

In order to generate an explanation for a specific data instance, the nodes of the lattice are traversed in the sorted order of their number of reduced features.

Table-III: Dataset with the black box model outcome

	Breathes in water (a)	Can fly (b)	Has beak (c)	Has hands (d)	Has skeleton (e)	Has wings (f)	Lives in water (g)	Is viviparous (h)	Produces light (i)	Model Outcome
Bat		X			X	X		X		Mammal
Eagle		X	X		X	X				Bird
Monkey				X	X			X		Mammal
Parrot Fish	X		X		X		X			Fish
Penguin			X		X	X	X			Bird
Shark	X				X		X			Fish
Lantern Fish	X				X		X		X	Fish

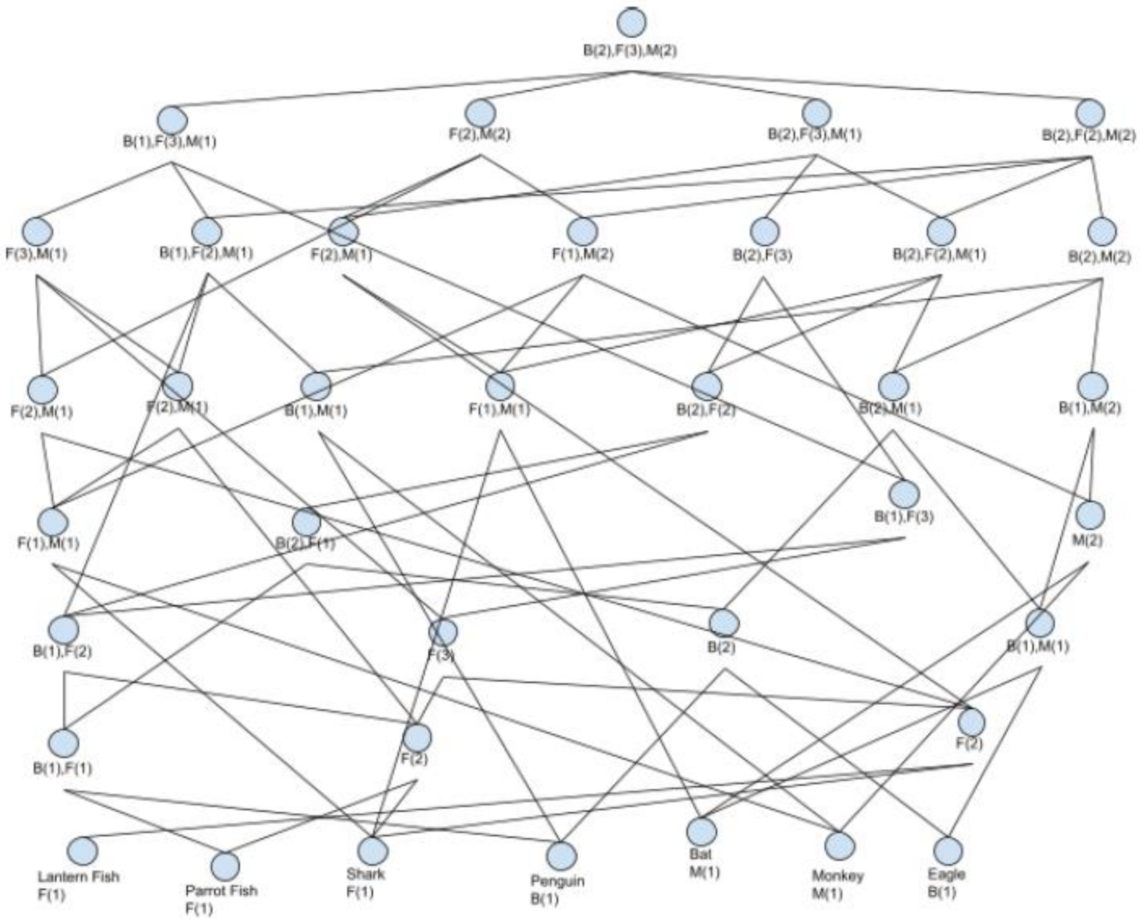


Fig. 4. Lattice nodes with gathered model outcomes and number of instances of each (the last level with an empty object set and empty set of classes can be assumed)

If an instance satisfies feature values at a node with no implication violations, the set of outcomes at the node are used to derive the minimum feature combinations that deny non-outcome classes or those that imply the outcome. They are derived along with a confidence value, a ratio of the number of instances the node satisfies towards a specific outcome to the number of instances it satisfies for all outcomes. For similar and contrastive examples, the nearest instance neighbours in the lattice are found and based on its outcome, the change in feature values leading to a change in the outcome or not are derived. Since there can be many such examples, these are derived incrementally on user request. For illustration, the examples from the first iteration alone are presented henceforth. To illustrate this, we use 4 instances from Table 3 and present local explanations of each, with similar and contrastive examples.

BAT: 0,1,0,0,1,1,0,1,0

Features: (Has beak,0) deny class(es): Bird
Remaining class(es) are: Fish Mammal
Features: (Breathes in water,0) deny class(es): Fish
Remaining class(es) are: Mammal
Lattice traversal has denied all class(es) except "Mammal"
Features: (Is viviparous,1) lead to class "Mammal" with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Has beak (0 to 1) Is viviparous (1 to 0) changes the class to Bird.

EAGLE: 0,1,1,0,1,1,0,0,0

Features: (Is viviparous,0) deny class(es): Mammal
Remaining class(es) are: Bird Fish
Features: (Breathes in water,0) deny class(es): Fish
Remaining class(es) are: Bird
Lattice traversal has denied all class(es) except "Bird"
Features: (Has wings,1) lead to class "Bird" with a confidence of 0.666667
Features: (Breathes in water,0) (Has beak,1) lead to class "Bird" with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Has beak (1 to 0) Is viviparous (0 to 1) changes the class to Mammal.
Changing features: Can fly (1 to 0) Lives in water (0 to 1) does not change the class.

PARROT FISH: 1,0,1,0,1,0,1,0,0

Features: (Lives in water,1) deny class(es): Mammal
Remaining class(es) are: Bird Fish
Features: (Has wings,0) deny class(es): Bird
Remaining class(es) are: Fish
Lattice traversal has denied all class(es) except "Fish"
Features: (Breathes in water,1) lead to class "Fish" with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Has beak (1 to 0) does not change the class.

PENGUIN: 0,0,1,0,1,1,1,0,0

Features: (Lives in water,1) deny class(es): Mammal
Remaining class(es) are: Bird Fish
Features: (Breathes in water,0) deny class(es): Fish
Remaining class(es) are: Bird
Lattice traversal has denied all class(es) except "Bird"
Features: (Has wings,1) lead to class "Bird" with a confidence of 0.666667
Features: (Can fly,0) (Has wings,1) lead to class "Bird" with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Breathes in water (0 to 1) Has wings (1 to 0) changes the class to Fish.
Changing features: Breathes in water (0 to 1) Has beak (1 to 0) Has wings (1 to 0) changes the class to Fish.

IV. SANITY TEST EVALUATION

To prove that the proposed technique passes sanity tests for explanations, we evaluate it for a binary classification problem on a simple dataset with a known non-linear model in order to be intuitive to human understanding. We use Implementation Invariance, Input transformation Invariance, Model parameter randomization sensitivity and model-outcome relationship randomization sensitivity as the Sanity tests. Implementation invariance checks if an explanation model provides the same or equivalent explanation to black box models that are functionally equivalent. Input transformation invariance detects if an explanation model provides the same or equivalent explanation when data is modified in a way that does not affect the outcome [14]. Model parameter sensitivity tests if an explanation model is sensitive or changes the explanation when the model's learnt parameters are modified or randomized. Model-Outcome relationship sensitivity tests if an explanation model is sensitive or changes the explanation when the model and outcome relationship is randomized [1]. In general, for explanation models, these test results are presented statistically using measures of difference between the new explanation (say, a heat map) and the original. But such statistical measures are not applicable here as the proposed technique is deterministic and not model driven. Hence we use local explanation around an instance to prove its credibility. Sample Dataset: Four features F_0, F_1, F_2 and F_3, integer valued ranging from -5 to 5.

Non-linear Model:

If $-1 * (F_0) + 2 * (F_1)^2 + -3 * (F_2)^3 + 4 * (F_3)^4 > 0$
    Outcome is Class 1
Else
    Outcome is Class 0

We assume no implications leading to a synthetic dataset with 11^4 instances on which the lattice is constructed. Few of the feature-to-class implications (part of global explanation) is as follows:

1. (F2,-2) ⇒ 1
2. (F2,-3) ⇒ 1
3. (F2,-4) ⇒ 1
4. (F2,-5) ⇒ 1
5. (F3,-4) ⇒ 1
6. (F3,-5) ⇒ 1
7. (F3,4) ⇒ 1
8. (F3,5) ⇒ 1
9. (F2,3)(F3,-1) ⇒ 0
10. (F2,3)(F3,0) ⇒ 0
11. (F2,3)(F3,1) ⇒ 0
12. (F2,4)(F3,-1) ⇒ 0
13. (F2,4)(F3,-2) ⇒ 0
14. (F2,4)(F3,0) ⇒ 0
15. (F2,4)(F3,1) ⇒ 0
16. (F2,4)(F3,2) ⇒ 0

All the implications are intuitive to human understanding. For example, we can understand implication 1 as follows:

For $F_2 < -1$, $-3 * (F_2)^3 \geq 24$, and the overall value can be pulled down maximum by F_0 with value 5, retaining it positive, hence rightfully classified to be in class 1.

Using implications 5 to 8, for F_3, such a decision of class 1 can be made only for values $F_3 < -3$ or $F_3 > 3$. On the contrary, single feature values are not enough to decide class 0, pointed out in implications 10 to 16.

We use the local explanation for the instance -5,-5,4,-2 in order to evaluate sanity tests.

Local explanation from the original model:
Features: (F2,4) (F3,-2) deny class(es): 1
Remaining class(es) are: 0
Lattice traversal has denied all class(es) except 0
Features: (F2,4) lead to class(es) 0 with a confidence of 0.454545
Features: (F2,4) (F3,-2) lead to class(es) 0 with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: F3 (-2 to -1) does not change the class.
Changing features: F0 (-5 to -1) does not change the class.
Changing features: F0 (-5 to -2) does not change the class.
Changing features: F0 (-5 to -3) does not change the class.
Changing features: F0 (-5 to -4) does not change the class.
Changing features: F1 (-5 to -1) does not change the class.
Changing features: F1 (-5 to -2) does not change the class.
Changing features: F1 (-5 to -3) does not change the class.
Changing features: F1 (-5 to -4) does not change the class.
Changing features: F2 (4 to -1) changes the class to 1.
Changing features: F2 (4 to -2) changes the class to 1.
Changing features: F2 (4 to -3) changes the class to 1.
Changing features: F2 (4 to -4) changes the class to 1.
Changing features: F2 (4 to -5) changes the class to 1.
Changing features: F2 (4 to 0) changes the class to 1.
Changing features: F2 (4 to 1) changes the class to 1.
Changing features: F2 (4 to 2) changes the class to 1.
Changing features: F2 (4 to 3) changes the class to 1.

In order to prove implementation invariance, we change the non-linear model as follows, that does not affect the model outcome from the original model.

If $(-1 * (F_0) + 2 * (F_1)^2 + -3 * (F_2)^3 + 4 * (F_3)^4) / 10 > 0$

Outcome is Class 1

Else

Outcome is Class 0

The lattice is constructed with the model outcome of the modified version and the local explanation generated for the instance -5,-5,4,-2 is exactly the same as the original.

Features: (F2,4) (F3,-2) deny class(es): 1
Remaining class(es) are: 0
Lattice traversal has denied all class(es) except 0
Features: (F2,4) lead to class(es) 0 with a confidence of 0.454545
Features: (F2,4) (F3,-2) lead to class(es) 0 with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: F3 (-2 to -1) does not change the class.
Changing features: F0 (-5 to -1) does not change the class.
Changing features: F0 (-5 to -2) does not change the class.
Changing features: F0 (-5 to -3) does not change the class.
Changing features: F0 (-5 to -4) does not change the class.
Changing features: F1 (-5 to -1) does not change the class.
Changing features: F1 (-5 to -2) does not change the class.
Changing features: F1 (-5 to -3) does not change the class.
Changing features: F1 (-5 to -4) does not change the class.
Changing features: F2 (4 to -1) changes the class to 1.
Changing features: F2 (4 to -2) changes the class to 1.
Changing features: F2 (4 to -3) changes the class to 1.
Changing features: F2 (4 to -4) changes the class to 1.
Changing features: F2 (4 to -5) changes the class to 1.
Changing features: F2 (4 to 0) changes the class to 1.
Changing features: F2 (4 to 1) changes the class to 1.
Changing features: F2 (4 to 2) changes the class to 1.
Changing features: F2 (4 to 3) changes the class to 1.

Non-intrusive input transformation invariance is naturally proven from similar and contrastive examples generated in the explanation.

In order to prove model parameter sensitivity, we change the non-linear model as follows, that affects the model outcome from the original model.

If $F_0 + -2 * (F_1)^2 + 3 * (F_2)^3 + -4 * (F_3)^4 > 0$

Outcome is Class 1

Else

Outcome is Class 0

The lattice is constructed with the model outcome of the modified version and the local explanation generated for the instance -5,-5,4,-2 is clearly different from the original.

Features: (F2,4) (F3,-2) deny class(es): 0
Remaining class(es) are: 1
Lattice traversal has denied all class(es) except 1
Features: (F2,4) lead to class(es) 1 with a confidence of 0.454545
Features: (F2,4) (F3,-2) lead to class(es) 1 with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: F3 (-2 to -1) does not change the class.
Changing features: F0 (-5 to -1) does not change the class.
Changing features: F0 (-5 to -2) does not change the class.
Changing features: F0 (-5 to -3) does not change the class.
Changing features: F0 (-5 to -4) does not change the class.
Changing features: F1 (-5 to -1) does not change the class.
Changing features: F1 (-5 to -2) does not change the class.
Changing features: F1 (-5 to -3) does not change the class.
Changing features: F1 (-5 to -4) does not change the class.
Changing features: F2 (4 to -1) changes the class to 0.
Changing features: F2 (4 to -2) changes the class to 0.
Changing features: F2 (4 to -3) changes the class to 0.
Changing features: F2 (4 to -4) changes the class to 0.
Changing features: F2 (4 to -5) changes the class to 0.
Changing features: F2 (4 to 0) changes the class to 0.
Changing features: F2 (4 to 1) changes the class to 0.
Changing features: F2 (4 to 2) changes the class to 0.
Changing features: F2 (4 to 3) changes the class to 0.

Model-Outcome relationship randomization sensitivity is naturally proven from the difference in the explanation for model parameter sensitivity considering it as a model-outcome randomization and the fact that the technique is a deterministic one depending only on the input parameters and the model outcome.

V. EVALUATION OF EXPLANATION BY COMPARISON

In order to evaluate the explanation from the lattice, it is compared with a decision tree on multi-class classification dataset (Zoo) from the UCI Machine Learning repository. The attributes and their value ranges for the Zoo dataset are:

1. hair: Boolean
2. feathers: Boolean
3. eggs: Boolean
4. milk: Boolean
5. airborne: Boolean
6. aquatic: Boolean
7. predator: Boolean
8. toothed: Boolean
9. backbone: Boolean
10. breathes: Boolean
11. venomous: Boolean
12. fins: Boolean
13. legs: Numeric (set of values: {0,2,4,5,6,8})
14. tail: Boolean
15. domestic: Boolean
16. catsize: Boolean

Class: Numeric (integer values in range [1 - 7]) (which is converted to the names [1-Mammal, 2-Bird, 3-Reptile , 4-Fish, 5-Amphibian, 6-Insect, 7-Invertebrate] for better readability)

A. Global Explanation comparison

A decision tree model was trained on the Zoo dataset with 90.9% accuracy. The decision tree used by the model after training is depicted in Fig. 5.

The dataset and the model outcome were used to construct the lattice (with implication cutoff as 1 leading to synthetic dataset same as the original) and the explanations were derived. A part of the global explanation is as follows (⇒ stands for "implies"):

1. (Feathers,0) ⇒ Amphibian or Fish or Insect or Invertebrate or Mammal or Reptile
2. (Milk,0) ⇒ Amphibian or Bird or Fish or Insect or Invertebrate or Reptile
3. (Aquatic,1) ⇒ Amphibian or Bird or Fish or Invertebrate or Mammal
4. (Predator,0) ⇒ Amphibian or Bird or Fish or Insect or Mammal or Reptile
5. (Predator,1) ⇒ Amphibian or Bird or Fish or Invertebrate or Mammal or Reptile
6. (Backbone,1) ⇒ Amphibian or Bird or Fish or Mammal or Reptile
7. (Breathes,1) ⇒ Amphibian or Bird or Insect or Invertebrate or Mammal or Reptile
8. (Fins,0) ⇒ Amphibian or Bird or Insect or Invertebrate or Mammal or Reptile
9. (tail,1) ⇒ Amphibian or Bird or Fish or Invertebrate or Mammal or Reptile
10. (domestic,1) ⇒ Bird or Fish or Insect or Mammal
11. (catsize,1) ⇒ Bird or Fish or Invertebrate or Mammal or Reptile
12. (Eggs,0) ⇒ Amphibian or Invertebrate or Mammal

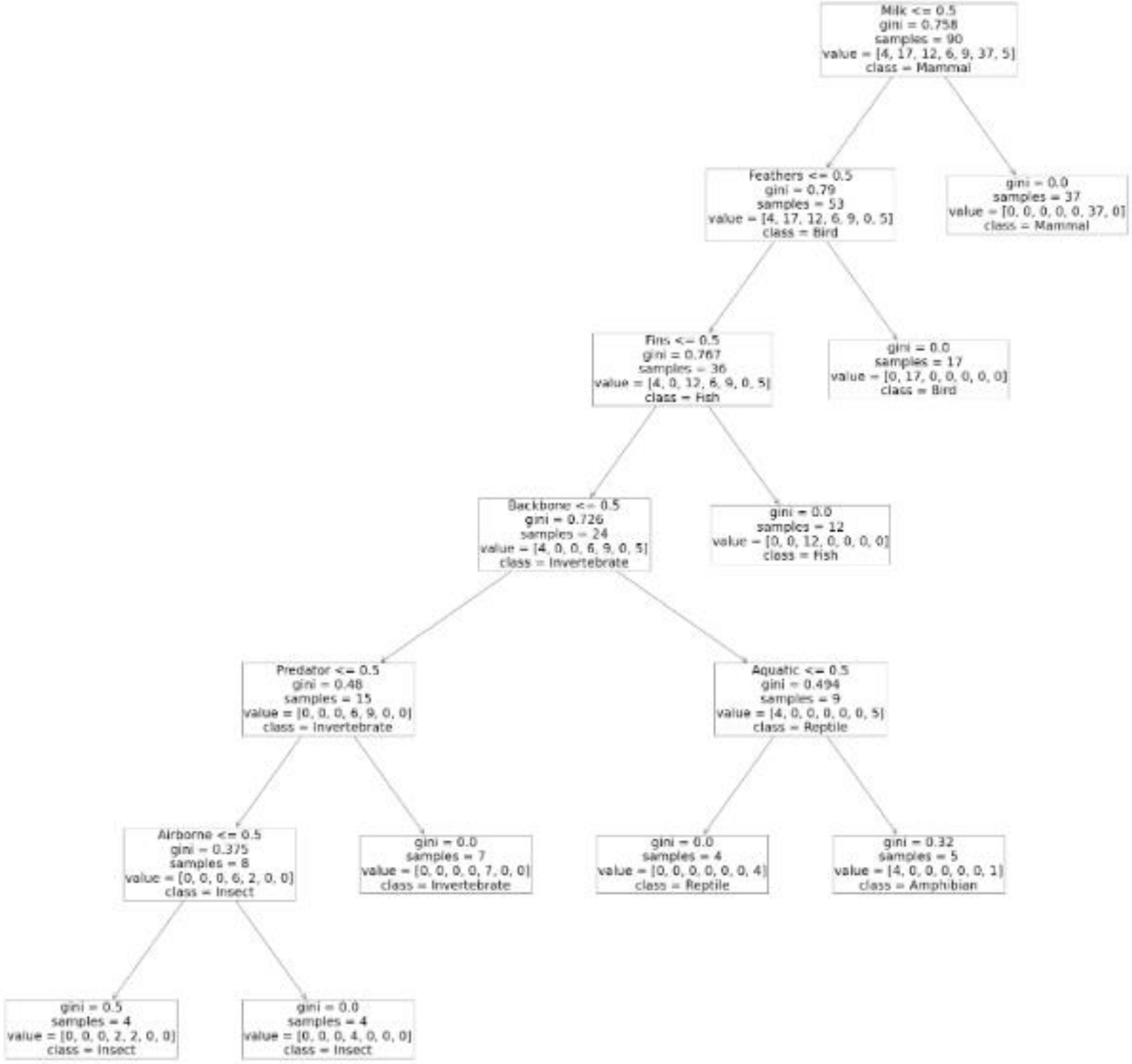
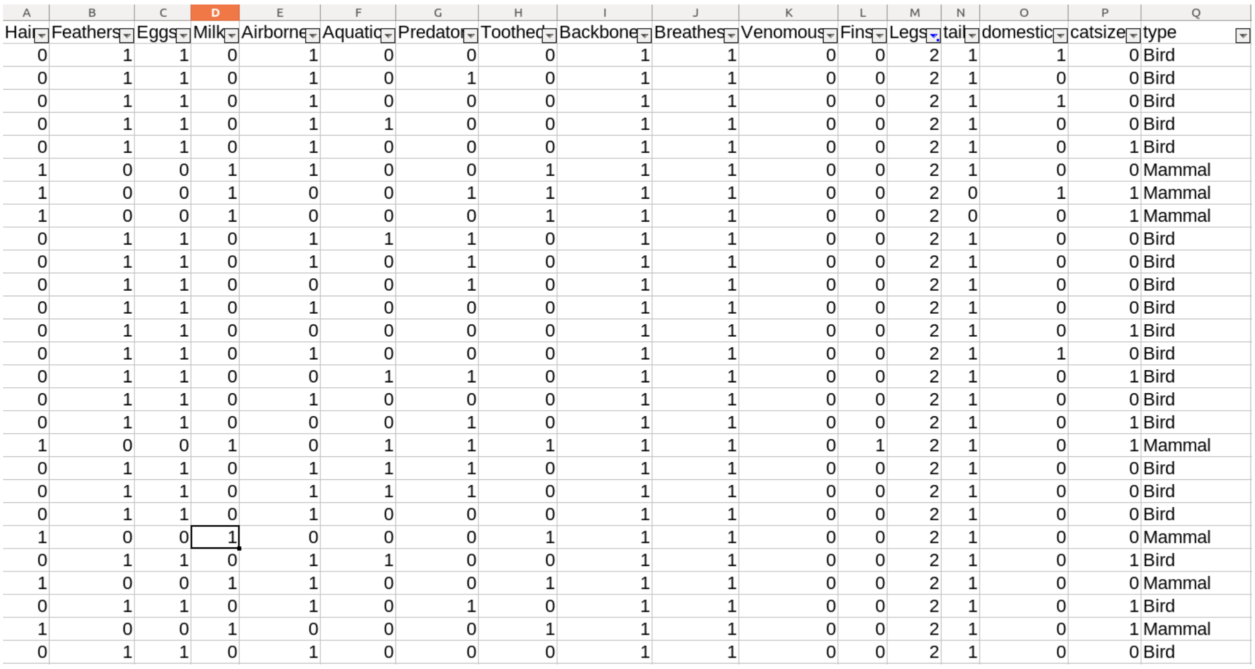


Fig. 5. Decision Tree of the trained model

13. (tail,0) ⇒ Amphibian or Insect or Invertebrate or Mammal
14. (Aquatic,0) ⇒ Bird or Insect or Invertebrate or Mammal or Reptile
15. (Toothed,0) ⇒ Bird or Insect or Invertebrate or Mammal or Reptile
16. (Hair,1) ⇒ Insect or Mammal
17. (Airborne,1) ⇒ Bird or Insect or Invertebrate or Mammal

18. (Legs,4) ⇒ Amphibian or Invertebrate or Mammal or Reptile
19. (Legs,2) ⇒ Bird or Mammal
20. (Breathes,0) ⇒ Amphibian or Fish or Invertebrate
21. (Backbone,0) ⇒ Insect or Invertebrate
22. (Milk,1) ⇒ Mammal
23. (Fins,1) ⇒ Fish or Mammal
24. (Feathers,1) ⇒ Bird
25. (Feathers,0)(Milk,0) ⇒ Amphibian or Fish or Insect or Invertebrate or Reptile
26. (Feathers,0)(Predator,0) ⇒ Amphibian or Fish or Insect or Mammal or Reptile
27. (Feathers,0)(Predator,1) ⇒ Amphibian or Fish or Invertebrate or Mammal or Reptile
28. (Feathers,0)(Backbone,1) ⇒ Amphibian or Fish or Mammal or Reptile
29. (Feathers,0)(Breathes,1) ⇒ Amphibian or Insect or Invertebrate or Mammal or Reptile
30. (Feathers,0)(Fins,0) ⇒ Amphibian or Insect or Invertebrate or Mammal or Reptile

The global explanation from the lattice is verifiable and reveals all the decision rules used by the decision tree. In some cases, explanation from the lattice uses feature values that seem not to directly correlate with the decision tree. But it is verifiable from the dataset that the features used in the explanation inherently imply the ones used in the decision tree for the dataset on which the lattice was constructed. For example, for a species with feature "Legs", having a value "2", rule 19 indicates that it must be a Bird or a Mammal. Apparently, this feature is not used in the decision tree at all. But if the values in the Zoo dataset with the trained decision tree outcome is explored, it is clear that all instances where "Legs" has value "2" are either a Bird or a Mammal as shown in Fig. 6. Among these instances, those that have feature "Milk" with value "1" are Mammals, while those that have feature "Feathers" with value "1" are Birds, which is part of the decision tree. There are many more rules present in the lattice's explanation compared to the decision tree as these are derived from feature implications in the dataset.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Hai	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Breathes	Venomous	Fins	Legs	tail	domestic	catsize	type
0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	Bird
0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	Bird
0	1	1	0	1	1	0	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	1	Bird
1	0	0	1	1	0	0	1	1	1	0	0	2	1	0	0	Mammal
1	0	0	1	0	0	1	1	1	1	0	0	2	0	1	1	Mammal
1	0	0	1	0	0	0	1	1	1	0	0	2	0	0	1	Mammal
0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	0	0	1	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	0	0	0	0	1	1	0	0	2	1	0	1	Bird
0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	Bird
0	1	1	0	0	1	1	0	1	1	0	0	2	1	0	1	Bird
0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	0	0	1	0	1	1	0	0	2	1	0	1	Bird
1	0	0	1	0	1	1	1	1	1	0	1	2	1	0	1	Mammal
0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0	Bird
0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	Bird
1	0	0	1	0	0	0	1	1	1	0	0	2	1	0	0	Mammal
0	1	1	0	1	1	0	0	1	1	0	0	2	1	0	1	Bird
1	0	0	1	1	0	0	1	1	1	0	0	2	1	0	0	Mammal
0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	1	Bird
1	0	0	1	0	0	0	1	1	1	0	0	2	1	0	1	Mammal
0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	Bird

Fig 6: Zoo Dataset and Decision Tree outcome with Filter for "Legs" = "2"

Table-IV: Instances used for local explanation comparison

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	Class
E1	1	0	0	1	0	1	1	1	1	1	0	1	2	1	0	1	1
E2	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	2
E3	0	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0	3
E4	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0	4
E5	0	0	1	0	0	1	1	1	1	1	1	0	4	0	0	0	5
E6	1	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0	6
E7	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0	0	7

B. Local Explanation comparison

To compare the local explanation of the lattice to the decision path from the Decision tree, the instances (one instance from each class - E1 to E7 in classes 1 - 7) in Table-IV are used.

Decision Path of E1(Mammal) from Decision tree:
Milk 1 > 0.5
Class: ['Mammal']

Explanation of E1(Mammal) from Lattice:
Features: (Feathers,0) deny class(es): Bird
Remaining class(es) are: Amphibian Fish Insect Invertebrate Mammal Reptile
Features: (Aquatic,1) deny class(es): Insect Reptile
Remaining class(es) are: Amphibian Fish Invertebrate Mammal
Features: (Backbone,1) deny class(es): Invertebrate
Remaining class(es) are: Amphibian Fish Mammal
Features: (Breathes,1) deny class(es): Fish
Remaining class(es) are: Amphibian Mammal
Features: (catsize,1) deny class(es): Amphibian
Remaining class(es) are: Mammal
Lattice traversal has denied all class(es) except Mammal
Features: (Milk,1) lead to class(es) Mammal with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Fins (1 to 0) Legs (2 to 4) does not change the class.
Changing features: Legs (2 to 0) tail (1 to 0) does not change the class.
Changing features: Hair (1 to 0) Legs (2 to 0) does not change the class.

Decision Path of E2(Bird) from Decision tree:
Milk 0 <= 0.5
Feathers 1 > 0.5
Class: ['Bird']

Explanation of E2(Bird) from Lattice:
Features: (Milk,0) deny class(es): Mammal
Remaining class(es) are: Amphibian Bird Fish Insect Invertebrate Reptile
Features: (Predator,0) deny class(es): Invertebrate
Remaining class(es) are: Amphibian Bird Fish Insect Reptile
Features: (Backbone,1) deny class(es): Insect
Remaining class(es) are: Amphibian Bird Fish Reptile
Features: (Breathes,1) deny class(es): Fish
Remaining class(es) are: Amphibian Bird Reptile
Features: (domestic,1) deny class(es): Amphibian Reptile
Remaining class(es) are: Bird
Lattice traversal has denied all class(es) except Bird
Features: (Feathers,1) lead to class(es) Bird with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: domestic (1 to 0) does not change the class.

Decision Path of E3(Reptile) from Decision Tree:
Milk 0 <= 0.5
Feathers 0 <= 0.5
Fins 0 <= 0.5
Backbone 1 > 0.5
Aquatic 0 <= 0.5

Explanation of E3(Reptile) from Lattice:
Features: (Feathers,0) deny class(es): Bird
Remaining class(es) are: Amphibian Fish Insect Invertebrate Mammal Reptile
Features: (Milk,0) deny class(es): Mammal
Remaining class(es) are: Amphibian Fish Insect Invertebrate Reptile
Features: (Predator,1) deny class(es): Insect
Remaining class(es) are: Amphibian Fish Invertebrate Reptile
Features: (Backbone,1) deny class(es): Invertebrate
Remaining class(es) are: Amphibian Fish Reptile
Features: (Breathes,1) deny class(es): Fish
Remaining class(es) are: Amphibian Reptile
Features: (Aquatic,0) deny class(es): Amphibian
Remaining class(es) are: Reptile
Lattice traversal has denied all class(es) except Reptile
Features: (Legs,0) lead to class(es) Reptile with a confidence of 0.153846
Features: (Aquatic,0) (Legs,0) lead to class(es) Reptile with a confidence of 0.500000
Features: (Aquatic,0) (Toothed,1) (Legs,0) lead to class(es) Reptile with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Venomous (0 to 1) does not change the class.
Changing features: Legs (0 to 4) does not change the class.

Decision Path of E4(Fish) from Decision Tree:
Milk 0 <= 0.5
Feathers 0 <= 0.5
Fins 1 > 0.5
Class: ['Fish']

Explanation of E4(Fish) from Lattice:
Features: (Feathers,0) deny class(es): Bird
Remaining class(es) are: Amphibian Fish Insect Invertebrate Mammal Reptile
Features: (Milk,0) deny class(es): Mammal
Remaining class(es) are: Amphibian Fish Insect Invertebrate Reptile
Features: (Aquatic,1) deny class(es): Insect Reptile
Remaining class(es) are: Amphibian Fish Invertebrate
Features: (Predator,0) deny class(es): Invertebrate
Remaining class(es) are: Amphibian Fish
Features: (Fins,1) deny class(es): Amphibian
Remaining class(es) are: Fish
Lattice traversal has denied all class(es) except Fish
Features: (Fins,1) lead to class(es) Fish with a confidence of 0.625000
Features: (Eggs,1) (Fins,1) lead to class(es) Fish with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: domestic (0 to 1) does not change the class.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Decision Path of E5(Amphibian) from Decision tree:
Milk 0 <= 0.5
Feathers 0 <= 0.5
Fins 0 <= 0.5
Backbone 1 > 0.5
Aquatic 1 > 0.5
Class: ['Amphibian']

Explanation of E5(Amphibian) from Lattice:
Features: (Feathers,0) deny class(es): Bird
Remaining class(es) are: Amphibian Fish Insect Invertebrate Mammal Reptile
Features: (Milk,0) deny class(es): Mammal
Remaining class(es) are: Amphibian Fish Insect Invertebrate Reptile
Features: (Aquatic,1) deny class(es): Insect Reptile
Remaining class(es) are: Amphibian Fish Invertebrate
Features: (Backbone,1) deny class(es): Invertebrate
Remaining class(es) are: Amphibian Fish
Features: (Breathes,1) deny class(es): Fish
Remaining class(es) are: Amphibian
Lattice traversal has denied all class(es) except Amphibian
Features: (Venomous,1) lead to class(es) Amphibian with a confidence of 0.250000
Features: (Venomous,1) (Legs,4) lead to class(es) Amphibian with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Venomous (1 to 0) does not change the class.

Decision Path of E6(Insect) from Decision tree:
Milk 0 <= 0.5
Feathers 0 <= 0.5
Fins 0 <= 0.5
Backbone 0 <= 0.5
Predator 0 <= 0.5
Airborne 1 > 0.5
Class: ['Insect']

Explanation of E6(Insect) from Lattice:
Features: (Feathers,0) deny class(es): Bird
Remaining class(es) are: Amphibian Fish Insect Invertebrate Mammal Reptile
Features: (Milk,0) deny class(es): Mammal
Remaining class(es) are: Amphibian Fish Insect Invertebrate Reptile
Features: (Predator,0) deny class(es): Invertebrate
Remaining class(es) are: Amphibian Fish Insect Reptile
Features: (Breathes,1) deny class(es): Fish
Remaining class(es) are: Amphibian Insect Reptile
Features: (tail,0) deny class(es): Reptile
Remaining class(es) are: Amphibian Insect
Features: (Aquatic,0) deny class(es): Amphibian
Remaining class(es) are: Insect
Lattice traversal has denied all class(es) except Insect
Features: (Legs,6) lead to class(es) Insect with a confidence of 0.714286
Features: (Predator,0) (Backbone,0) lead to class(es) Insect with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Hair (1 to 0) does not change the class.

Decision Path of E7(Invertebrate) from Decision tree:
Milk 0 <= 0.5
Feathers 0 <= 0.5
Fins 0 <= 0.5
Backbone 0 <= 0.5
Predator 1 > 0.5
Class: ['Invertebrate']

Explanation of E7(Invertebrate) from Lattice:
Features: (Feathers,0) deny class(es): Bird
Remaining class(es) are: Amphibian Fish Insect Invertebrate Mammal Reptile
Features: (Milk,0) deny class(es): Mammal
Remaining class(es) are: Amphibian Fish Insect Invertebrate Reptile
Features: (Aquatic,1) deny class(es): Insect Reptile
Remaining class(es) are: Amphibian Fish Invertebrate
Features: (Fins,0) deny class(es): Fish
Remaining class(es) are: Amphibian Invertebrate
Features: (Toothed,0) deny class(es): Amphibian
Remaining class(es) are: Invertebrate
Lattice traversal has denied all class(es) except Invertebrate
Features: (Backbone,0) lead to class(es) Invertebrate with a confidence of 0.571429
Features: (Predator,1) (Backbone,0) lead to class(es) Invertebrate with a confidence of 1.000000
Generating similar & contrastive explanations:
Changing features: Legs (4 to 5) does not change the class.
All the local explanations generated from the lattice for the Zoo dataset are verifiable, minimal feature combinations that deny or lead to the outcome. Using specific filters on the dataset, similar to Fig. 6., the explanations are exact to the decision tree path.

VI. Conclusion and Future Work

From the sanity test evaluations and explanation comparisons and evaluations, it is evident that our novel, deterministic lattice based approach to explainability is exact, integrated and intuitive to human understanding. Its feasibility to larger datasets is worth studying using parallel and distributed approaches. This novel technique can be developed further to extend to text and images. A comparative study of explanations from the proposed lattice based technique to explanations from existing state-of-the-art techniques can further prove its credibility and utility.

ACKNOWLEDGMENT

The authors offer this work to the Founder Chancellor of Sri Sathya Sai Institute of Higher Learning, Bhagawan Sri Sathya Sai Baba. We are grateful to Prof. P. Sadayappan, Prof. Ashok Srinivasan and Dr. Sudheer Chunduri for sharing their valuable suggestions.

REFERENCES

1. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I. J.; Hardt, M. & Kim, B., “Sanity Checks for Saliency Maps”., in Samy Bengio; Hanna M. Wallach; Hugo Larochelle; Kristen Grauman; Nicolò Cesa-Bianchi & Roman Garnett, ed., 'NeurIPS' , 2018, pp. 9525-9536 .
2. Leavitt, M. L. and Morcos, A., “Towards falsifiable interpretability research”, Neural Information Processing Systems Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA), Online, 2020.
3. Lundberg, S. M. & Lee, S.-I., “A Unified Approach to Interpreting Model Predictions”, in I. Guyon; U. V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan & R. Garnett, ed., 'Advances in Neural Information Processing Systems 30' , Curran Associates, Inc., 2017, pp. 4765-4774 .
4. Maier, D., “The Theory of Relational Databases (Book)”, Computer Science Press, 1983 .
5. Marzyeh Ghassemi, Luke Oakden-Rayner, Andrew L Beam, “The false hope of current approaches to explainable artificial intelligence in healthcare”, The Lancet Digital Health, Volume 3, Issue 11, 2021, Pages e745-e750, ISSN 2589-7500. [CrossRef]
6. Carpineto, C., “Concept Data Analysis Theory and Applications (Book)”, Wiley & Sons, 2004 . [CrossRef]
7. Rudin, C., 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nature Machine Intelligence 1 (5), 2019, 206--215. [CrossRef]
8. Wille, R., 'Concept lattices and conceptual knowledge systems', Computers and Mathematics with Applications 23 , 1992, 493-515. [CrossRef]
9. Sangroya, A.; Anantaram, C.; Rawat, M. & Rastogi, M., “Using Formal Concept Analysis to Explain Black Box Deep Learning Classification Models”., in Sergei O. Kuznetsov; Amedeo Napoli & Sebastian Rudolph, ed., 'FCA4AI@IJCAI' , CEUR-WS.org, 2019, pp. 19-26.
10. Sangroya, A.; Rastogi, M.; Anantaram, C. & Vig, L., “Guided-LIME: Structured Sampling based Hybrid Approach towards Explaining Blackbox Machine Learning Models”., in Stefan Conrad & Ilaria Tiddi, ed., 'CIKM (Workshops)', 2020, CEUR-WS.org.
11. Selvaraju, R. R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D. & Batra, D., “Grad-CAM: Why did you say that?”, 2016, CoRR abs/1611.07450.
12. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. B. & Wattenberg, M., “SmoothGrad: removing noise by adding noise.”, 2017, CoRR abs/1706.03825.
13. Springenberg, J. T.; Dosovitskiy, A.; Brox, T. & Riedmiller, M. A., “Striving for Simplicity: The All Convolutional Net.”, in Yoshua Bengio & Yann LeCun, ed., 2015, ICLR (Workshop) .
14. Sundararajan, M.; Taly, A. & Yan, Q., “Axiomatic attribution for deep networks”, in 'International conference on machine learning', 2017, pp. 3319-3328.
15. Ribeiro, M. T.; Singh, S. & Guestrin, C., "Why Should I Trust You?: Explaining the Predictions of Any Classifier”., in Balaji Krishnapuram; Mohak Shah; Alexander J. Smola; Charu C. Aggarwal; Dou Shen & Rajeev Rastogi, ed., 'KDD' , ACM, 2016, pp. 1135-1144 . [CrossRef]

AUTHORS PROFILE

Bhaskaran Venkatsubramaniam, B.Sc (Math, Hons.), M.Sc (Math), M.Tech (Computer Science), Assistant Professor, Faculty Research Scholar, Sri Sathya Sai Institute of Higher Learning, Muddenahalli Campus. https://www.sssihl.edu.in/faculty/v-bhaskaran/ Assistant Professor, registered for Ph.D, Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, Muddenahalli Campus. After an industry experience of 8 years. research interests include High Performance Computing, Explainable AI and software development for society. Current research focuses on using the formal concept lattice as an integrated model to derive different types of explanations for a black box model.

Prof. Pallav Kumar Baruah, B.Sc (Math), M.Sc (Math), Ph.D Professor, Department of Math and Computer Science, Sri Sathya Sai Institute of Higher Learning, Prasanthinilayam Campus. https://www.sssihl.edu.in/faculty/pallav-kumar-baruah/ Professor, Department of Mathematics and Computer Science, Sri Sathya Sai Institute of Higher Learning, Prasanthinilayam Campus. Research interest includes High Performance Computing and its applications to various domains in Computational Science, Bioinformatics, Actuarial Data Science etc. Current research interest is in developing XAI models using concept lattice. Also working on Scalability and Interoperability of Blockchain. Collaborations with Actuarial practitioners to develop solutions for problems in the domain of Actuarial Science using AI, ML, DL, Blockchain & HPC that has resulted in some work on Fraud Detection and prevention in Motor vehicle and Health-Insurance, ERM dashboard with KPIs for various lines of business with Predictive capability, AI, ML models for Pricing, Reserving, claims count estimation etc. Few other works in recent times are E-Voting Protocol using Blockchain, Platform for sharing Credential and background using Blockchain, Blended Learning: Blockchain based process for authenticated data assimilation into Deep Learning models. Recent work includes data marketplace and Blockchain based model to offer AI as a service (AIaaS).

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)