Leveraging Verification to Enhance Formal Explainable AI for Neural Networks

19 Juin 2025

GDR GPL

Presented by:

Ryma Boumazouza

Work in collaboration with:

Mélanie Ducoffe

Raya Elsaleh

Shahaf Bassan

Guy Katz

Airbus AI Research

ANITI

DEEL Dependable, Explainable & Embeddable Learning France-Quebec

THE HEBREW UNIVERSITY OF JERUSALEM

Delivering trustworthy AI through XAI

Need for Trustworthy AI in High-Risk Settings

Ensure safety, compliance, and ethics in critical sectors (e.g., healthcare, aeronautics, finance, autonomous vehicles)

Guidelines and Regulations Driving Trustworthiness:

EU, OECD, and UNESCO guidelines emphasize the need for AI to be trustworthy, transparent, accountable, and ethically and legally sound

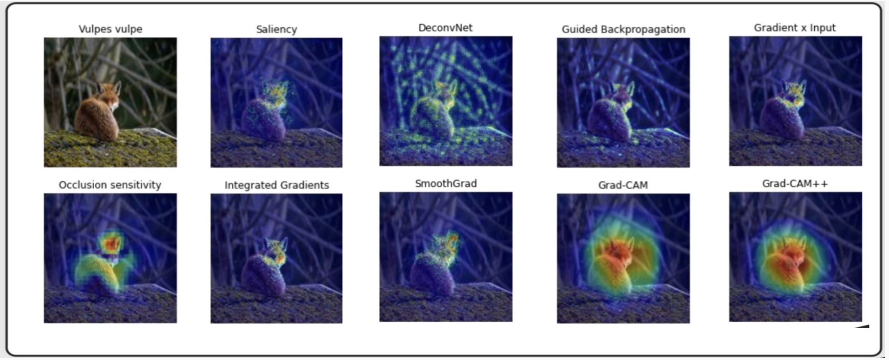
Challenges with Current eXplainable AI (XAI) Approaches

Scalability limits, high complexity, and difficulty integrating into existing AI systems

Can we Truly trust XAI Tools ?

XAI tools promise transparency but...

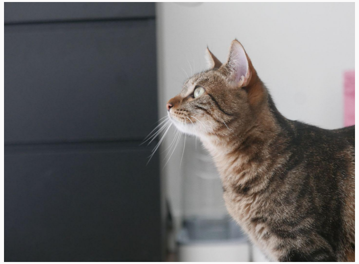
- are often heuristic
- do not provide guarantees



But,
If we can't trust the explainer, can we trust the model ?

→ **What about a Formal Explanation?**

What Does a Formal Explanation Look Like?

$N($  $) \stackrel{?}{=} \text{cat}$

We want an explanation to answer “why” the classifier predicted “cat”

A Sufficient reason/
Abductive explanation would be :

Formalizing the Concept of an Explanation

Formally, an abductive explanation is defined as:

$$\forall (x \in \mathbb{F}). \quad [\bigwedge_{i \in E} (x_i = v_i) \rightarrow (N(x) = c)]$$

Properties of an abductive explanation:

Minimality:

$N($  $) \neq \text{cat}$

Sufficiency:

$N($  $) = \text{cat}$

How to compute such abductive explanation ?

Algorithm 1 Deletion Algorithm to Find One Abductive Explanation

Input: Predictive model M and input $x = \langle \chi^1, \ldots, \chi^n \rangle$
Output: Explanation for class C of x
$Explanation \leftarrow \emptyset$ ▷ Set of relevant features
$I \leftarrow \emptyset$ ▷ Set of irrelevant features
$\pi \leftarrow TRAVERSALORDER(F)$ ▷ Sort F's features by ascending relevance
for each feature $x_i \in \pi$ **do**
  $\pi' \leftarrow \pi \setminus \{x_i\}$ ▷ Removing feature x_i
  if $\mathrm{CHECK}(M, \pi')$ **then** ▷ Check if prediction changes
    $Explanation \leftarrow Explanation \cup \{x_i\}$ ▷ Add feature x_i to relevant features
  else
    $I \leftarrow I \cup \{x_i\}$ ▷ Add feature x_i to irrelevant features
  end if
end for
Return *Explanation*

Challenges to address

- Challenge 1: The CHECK function is computationally expensive
- Challenge 2: Sequential traversal of the feature set (loop)
- Challenge 3: Impact of order on explanation interpretability (size)

About the CHECK method (Verification as a new paradigm for Abductive Explanation)

CHECK STEP

$\pi' \leftarrow \pi \setminus \{x_i\}$
if $\text{CHECK}(M, \pi')$ **then**
$\quad Explanation \leftarrow Explanation \cup \{x_i\}$

Property: Local Robustness

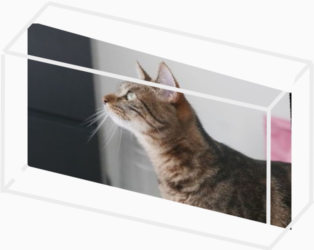
Under test
XAI set
Free set

$N\big($  $\big) \stackrel{?}{=} \text{cat}$

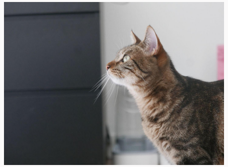
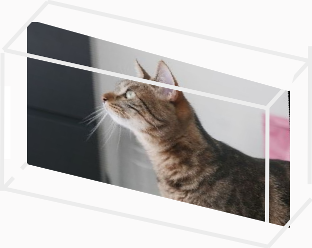
How it works?

1. *Fix XAI variables at their nominal values,*
2. *Allow the removed feature and all other inputs to vary within their valid domains,*
3. *Verify no property violations occur*

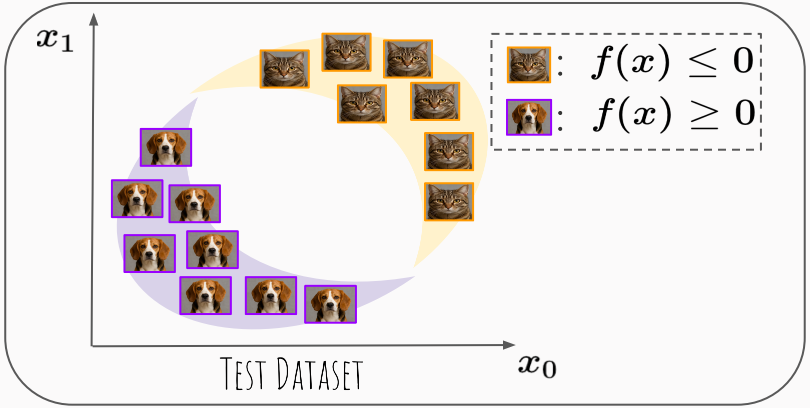
Verification property: Local Robustness

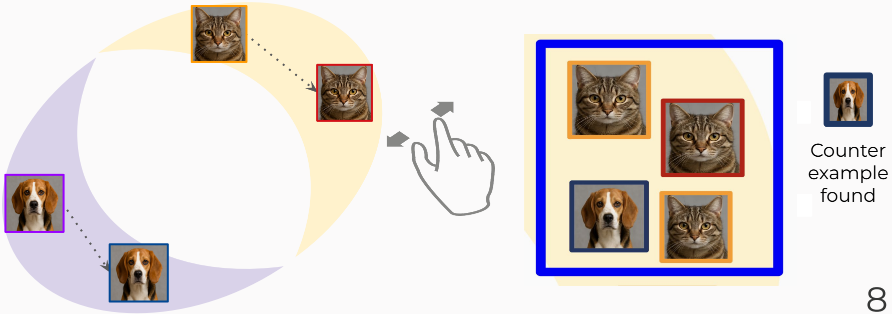
$N($  $) \stackrel{?}{=} \text{cat}$

An example of a perturbation with epsilon = +/- 1 pixel

 $+$  $\in$ 

Perturbation domain





Scaling up the performance of formal explainers

Algorithm 1 Deletion Algorithm to Find One Abductive Explanation

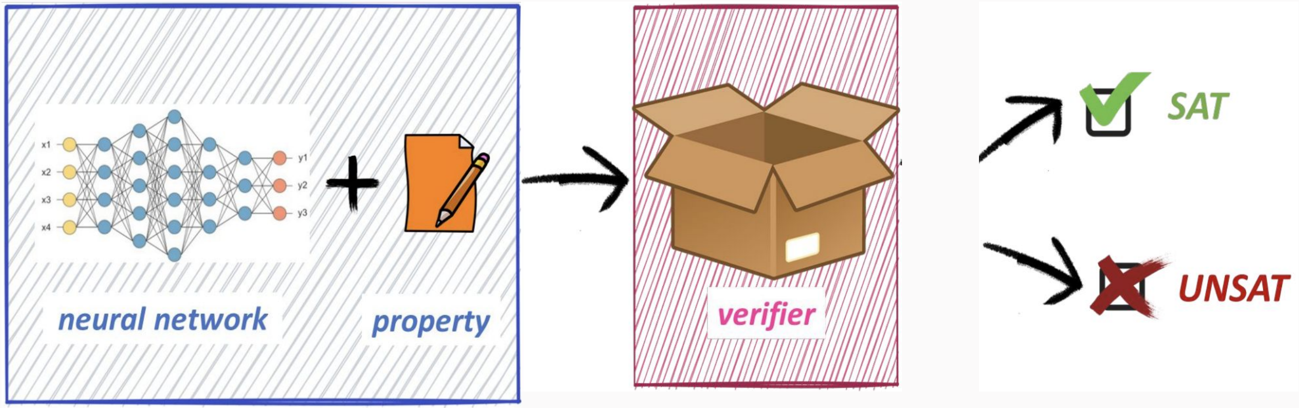
Input: Predictive model M and input $x = \langle \chi^1, \ldots, \chi^n \rangle$
Output: Explanation for class C of x
$\mathit{Explanation} \leftarrow \emptyset$ ▷ Set of relevant features
$I \leftarrow \emptyset$ ▷ Set of irrelevant features
$\pi \leftarrow \mathit{TRAVERSALORDER}(F)$ ▷ Sort F's features by ascending relevance
for each feature $x_i \in \pi$ **do**
    $\pi' \leftarrow \pi \setminus \{x_i\}$ ▷ Removing feature x_i
    if CHECK(M, π') **then** ▷ Check if prediction changes
        $\mathit{Explanation} \leftarrow \mathit{Explanation} \cup \{x_i\}$ ▷ Add feature x_i to relevant features
    else
        $I \leftarrow I \cup \{x_i\}$ ▷ Add feature x_i to irrelevant features
    end if
end for
Return *Explanation*

Challenges to address

Challenge 1: The CHECK function is computationally expensive

About the CHECK method (Verification step)

Challenge 1

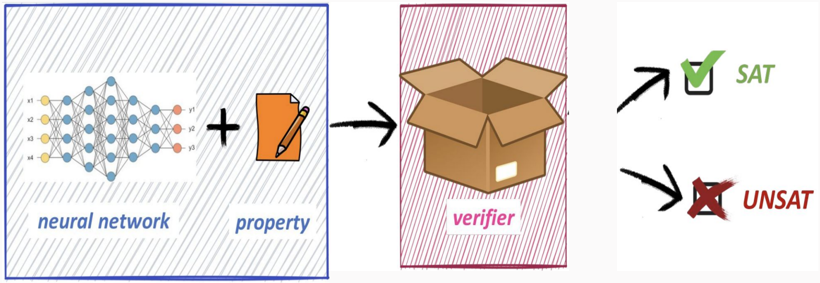


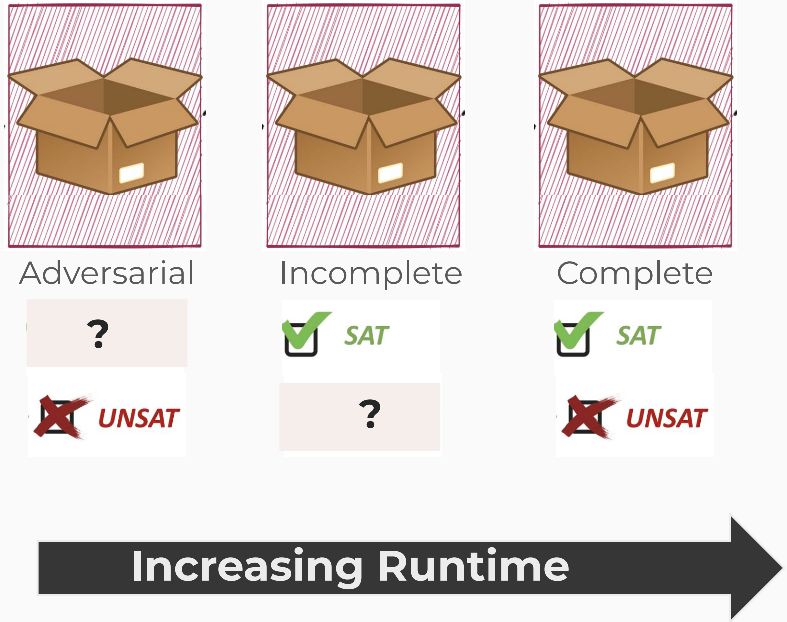
- ➔ SAT solvers (MiniSat, Glucose)
- ➔ SMT solvers (marabou, z3)
- ➔ CP solvers (OR-Tools CP-SAT)
- ➔ MILP solvers (Gurobi, CPLEX)

Complete

Link to figure [source]

Different techniques for NN verification





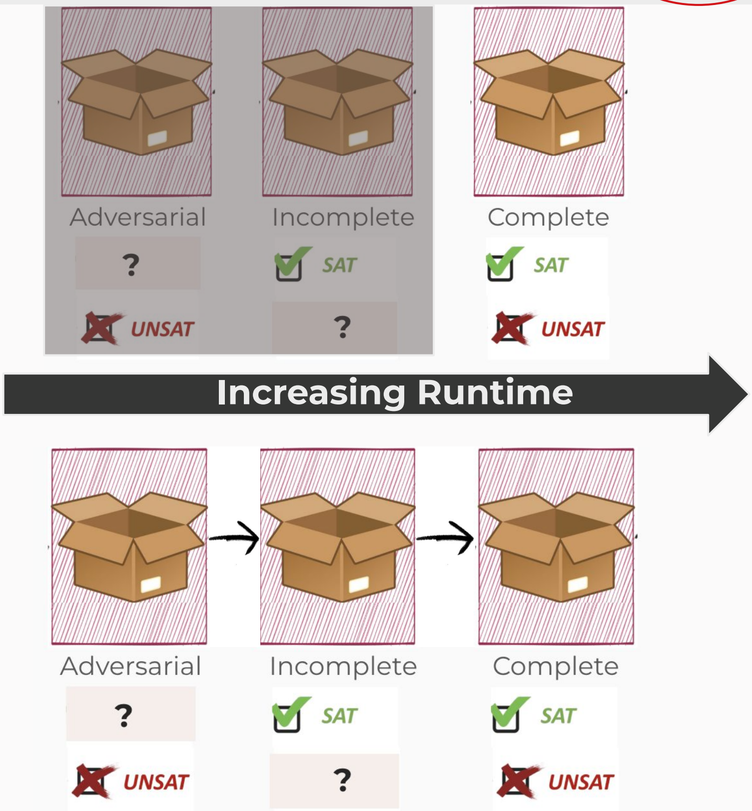
Combining methods in one verification 'pipeline'

Challenge 1

Algorithm 1 Deletion Algorithm to Find One Abductive Explanation

Input: Predictive model M and input $x = \langle \chi^1, \ldots, \chi^n \rangle$
Output: Explanation for class C of x
$Explanation \leftarrow \emptyset$ ▷ Set of relevant features
$I \leftarrow \emptyset$ ▷ Set of irrelevant features
$\pi \leftarrow TRAVERSALORDER(F)$ ▷ Sort F's features by ascending relevance
for each feature $x_i \in \pi$ **do**
  $\pi' \leftarrow \pi \setminus \{x_i\}$ ▷ Removing feature x_i
  if $\mathrm{CHECK}(M, \pi')$ **then** ▷ Check if prediction changes
    $Explanation \leftarrow Explanation \cup \{x_i\}$ ▷ Add feature x_i to relevant features
  else
    $I \leftarrow I \cup \{x_i\}$ ▷ Add feature x_i to irrelevant features
  end if
end for
Return $Explanation$

➛ Challenge 1: The CHECK function is computationally expensive



Parallelizing the computation of formal explanation

Challenge 2

Algorithm 1 Deletion Algorithm to Find One Abductive Explanation

Input: Predictive model M and input $x = \langle \chi^1, \ldots, \chi^n \rangle$
Output: Explanation for class C of x
$Explanation \leftarrow \emptyset$ ▷ Set of relevant features
$I \leftarrow \emptyset$ ▷ Set of irrelevant features
$\pi \leftarrow TRAVERSALORDER(F)$ ▷ Sort F's features by ascending relevance
for each feature $x_i \in \pi$ **do**
  $\pi' \leftarrow \pi \setminus \{x_i\}$ ▷ Removing feature x_i
  if CHECK(M, π') **then** ▷ Check if prediction changes
    $Explanation \leftarrow Explanation \cup \{x_i\}$ ▷ Add feature x_i to relevant features
  else
    $I \leftarrow I \cup \{x_i\}$ ▷ Add feature x_i to irrelevant features
  end if
end for
Return $Explanation$

Challenge 2: Sequential traversal of the feature set (loop)

Could we Add to explanation a batch of input Features ?

Challenge 2

Adversarial

?

Adversarial Attack: Add batch of input feature

Idea: Propose a new strategy that breaks the sequential query bottleneck in deletion-based formal XAI

How? Enable parallel removal of feature constraints and launch several adversarial attacks at once

Advantages? Batch processing & GPU implementation supported by existing adversarial methods, no extra development required!

Could we Free a batch of input Features ?

Challenge 2

Incomplete

SAT

?

Abstract Interpretation: free batch of input feature

Idea: Go beyond SAT/UNSAT verifier's decision! Leverage solver proofs to pinpoint and free multiple feature indices in a single iteration

How? Determine the largest subset of features that can be freed without compromising the property's & soundness

$$\mathbb{I}\ \text{s.t.}\ \forall\,\mathbb{I}',\ P(\mathbb{I}') \implies |\mathbb{I}'| \leq |\mathbb{I}|\,,$$
$$\text{where } P(\mathbb{I})\ :\ \exists\, E \in F \setminus \mathbb{I},\ \left(\bigwedge_{i \in E} x_i = v_i\right) \implies N(x) = c$$

Advantages? One call of abstract interpretation is enough !
Linear complexity + Soundness

Statistically-Guided Explanations

Challenge 3

Algorithm 1 Deletion Algorithm to Find One Abductive Explanation

Input: Predictive model M and input $x = \langle \chi^1, \ldots, \chi^n \rangle$
Output: Explanation for class C of x
$\mathit{Explanation} \leftarrow \emptyset$ ▷ Set of relevant features
$I \leftarrow \emptyset$ ▷ Set of irrelevant features
$\pi \leftarrow \mathit{TRAVERSALORDER}(F)$ ▷ Sort F's features by ascending relevance
for each feature $x_i \in \pi$ **do**
  $\pi' \leftarrow \pi \setminus \{x_i\}$ ▷ Removing feature x_i
  if CHECK(M, π') **then** ▷ Check if prediction changes
    $\mathit{Explanation} \leftarrow \mathit{Explanation} \cup \{x_i\}$ ▷ Add feature x_i to relevant features
  else
    $I \leftarrow I \cup \{x_i\}$ ▷ Add feature x_i to irrelevant features
  end if
end for
Return $\mathit{Explanation}$

Challenge 3: Impact of order on explanation interpretability (size)

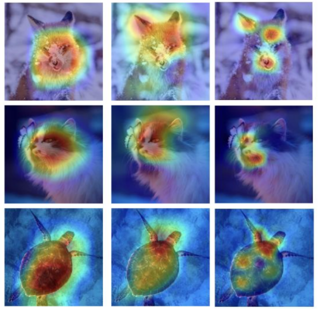
Statistically-Guided Explanations

Challenge 3

DEEL

XPLIQUE

Explainability Toolbox for Neural Networks



Feature attributions

Idea: Address the cardinality bottleneck in formal XAI by leveraging statistical explanation orderings

How? Synergies with statistical XAI to guide formal search, test different feature ordering and chose the best

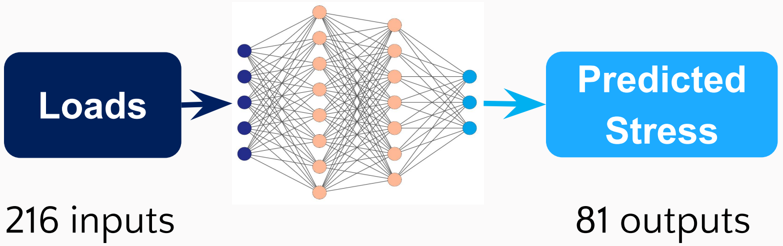
Advantages? Take advantage of the several XAI statistical-based techniques available and create synergies between the two communities!

Preliminary results

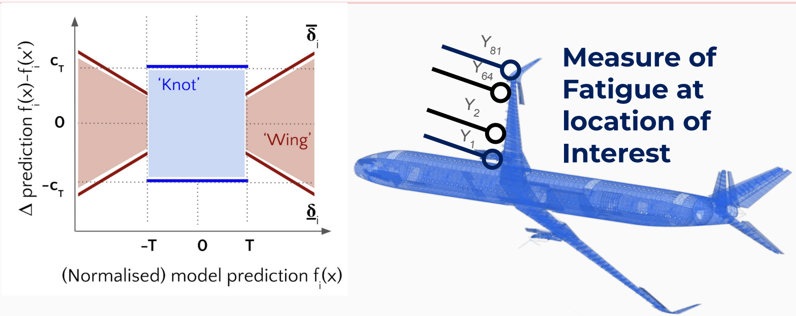
Uses cases

Local Stability

Industrial use case : Fatigue Digital Twin

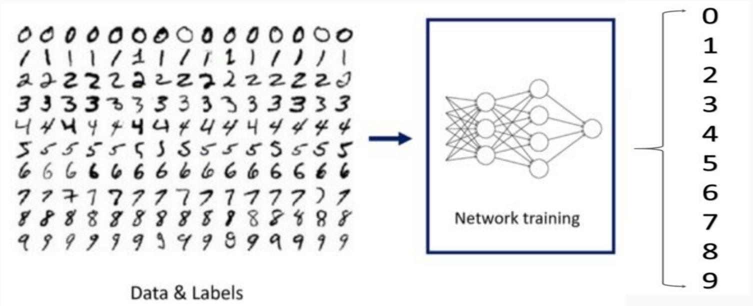


Property

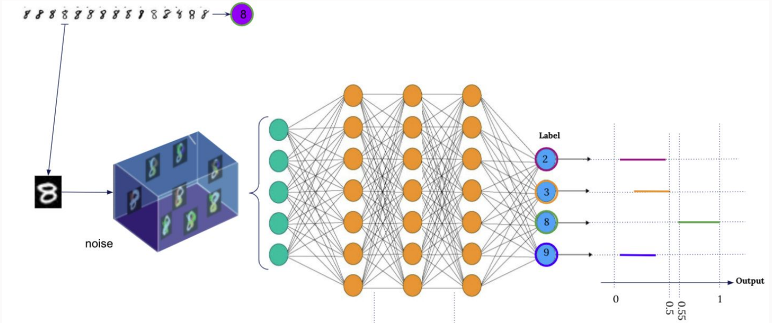


Local Robustness

Academic use case



Property



[Surrogate Neural Networks Local Stability for Aircraft Predictive Maintenance, FMICS 2024]()

Libraries & Tools

Explanation computation	Verification Pipeline	Abstract Interpretation
XAirobas	***Airobas***	***Decomon***
AIRBUS/ANITI	*AIRBUS*	*AIRBUS/ANITI*

Results on industrial use case (Fatigue Digital Twin)

Metrics \ CHECK config	Only complete **Baseline**	Verification pipeline **Challenge 1**	Pipeline & batch processing **Challenge 1+2**
Runtime per explanation	18mn48sec	**10mn79sec**	**2.14sec!**
Average explanation size	107	**72**	**36**
#Calls to adv attacks	0	104	34
#Calls to incomplete solver	0	99	2 + **1**
#Calls to complete solver	216	**13**	**4**

1 call to free in average
176 out of 216 at once !!!!

Results on academic use case (MNIST)

CHECK config / Metrics	Only complete **Baseline**	Verification pipeline **Challenge 1**	Pipeline & batch processing **Challenge 1+2**
Runtime per explanation	3876.61sec	**1783.28sec**	**495.5sec**
Average explanation size	111.85	111.85	**110.93**
#Calls to adv attacks	0	58	61
#Calls to incomplete solver	0	583	303 + **1**
#Calls to complete solver	784	**143**	134

1 call to free in average **286 out of 784 at once !!!!**

Takeaways

Algorithm 1 Deletion Algorithm to Find One Abductive Explanation

```
Input: Predictive model M and input x = ⟨χ^1, ..., χ^n⟩
Output: Explanation for class C of x
Explanation ← ∅                                    ▷ Set of relevant features
I ← ∅                                              ▷ Set of irrelevant features
π ← TRAVERSALORDER(F)                              ▷ Sort F's features by ascending relevance
for each feature x_i ∈ π do
    π′ ← π \ {x_i}                                 ▷ Removing feature x_i
    if CHECK(M, π′) then                           ▷ Check if prediction changes
        Explanation ← Explanation ∪ {x_i}          ▷ Add feature x_i to relevant features
    else
        I ← I ∪ {x_i}                              ▷ Add feature x_i to irrelevant features
    end if
end for
Return Explanation
```

XAirobas

Contributions

Contribution 1: Introduced a modern formal verification pipeline tailored to the scalability demands

Contribution 2: Propose a novel distributed strategy that breaks the sequential query bottleneck

Contribution 3: Statistically-Guided Explanations

Contacts:

Ryma Boumazouza :
ryma.boumazouza@airbus.com

Mélanie Ducoffe :
melanie.ducoffe@airbus.com

Thank you !