The Pros and Cons of Adversarial Robustness

Yacine Izza[1] and Joao Marques-Silva[2]
[1]*CREATE, NUS, Singapore*
[2]*ICREA, University of Lleida, Spain*

Keywords: Local & Global Robustness, Certified Robustness, Adversarial Examples, Explainable AI.

Abstract: Robustness is widely regarded as a fundamental problem in the analysis of machine learning (ML) models. Most often robustness equates with deciding the non-existence of adversarial examples, where adversarial examples denote situations where small changes on some inputs cause a change in the prediction. The perceived importance of ML model robustness explains the continued progress observed for most of the last decade. Whereas robustness is often assessed locally, i.e. given some target point in feature space, robustness can also be defined globally, i.e. where any point in feature space can be considered. The importance of ML model robustness is illustrated for example by the existing competition on neural network (NN) verification (VNN-COMP), which assesses the progress of robustness tools for NNs, but also by efforts towards robustness certification. More recently, robustness tools have also been used for computing rigorous explanations of ML models. Despite the continued advances in robustness, this paper uncovers some limitations with existing definitions of robustness, both global and local, but also with efforts towards robustness certification. The paper also investigates uses of adversarial examples besides those related with robustness.

1 INTRODUCTION

For more than a decade, Machine Learning (ML) has been the subject of remarkable advances. However, such advances have also been marred by a number of persistent challenges. One of the best known of these challenges is the *brittleness* of ML models (Hein and Andriushchenko, 2017). An ML model is brittle if it exhibits *adversarial examples* (AExs), i.e. small changes to the inputs of the ML model can cause (unexpected and unwanted) changes to the prediction (Biggio et al., 2013; Szegedy et al., 2014; Goodfellow et al., 2015). Intuitively, an ML model is robust if it exhibits no AExs. ML model robustness has been extensively studied over the last decade (Goodfellow et al., 2015; Zhang et al., 2022b). The importance of deciding the robustness of ML models motivated an outpouring of competing approaches, ranging for rather informal solutions, to those based on automated reasoners, and even those based on domain-specific reasoners. Furthermore, the significance of assessing and asserting robustness is further highlighted by VNN-COMP (Verification of Neural Networks Competition (Brix et al., 2023b; Brix et al., 2023a)), a competition for assessing robustness tools for NNs, that has been running since 2020. In addition, there are also efforts targeting the *robustness certification* of ML models (Cohen et al., 2019; Rosenfeld et al., 2020; Dvijotham et al., 2020; Huang et al., 2021; Vorácek and Hein, 2023; Carlini et al., 2023). Furthermore, there have been proposals towards the verification and validation of systems based on AI (Seshia et al., 2022), covering not only robustness, but also explainability and fairness. The uses of systems of AI in high-risk and safety-critical domains had motivated calls for the use of so-called *interpretable models* (Rudin, 2019; Rudin et al., 2022), with the purpose of enabling human-decision makers to explain the decisions taken by such systems. Unfortunately, such calls have not deterred proposal for the use of complex systems of AI in high-risk and safety-critical domains (Huang et al., 2020).

Robustness is often defined with respect to a concrete input to the ML model and its associated prediction. In this case, one is referring to what is called *local robustness*. An alternative view is *global robustness*, where the goal is to assess robustness for *any* input of the ML model. Nevertheless, there exist tools that target both local and global robustness (Katz et al., 2017). Furthermore, to understand whether an ML is locally/globally robust, it is also fundamental to outline an adequate experimental setup.

Izza, Y. and Marques-Silva, J.
The Pros and Cons of Adversarial Robustness.
DOI: 10.5220/0013166300003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 386-397
ISBN: 978-989-758-737-5; ISSN: 2184-433X

More recently, robustness has been linked with formal approaches to explainable artificial intelligence (XAI) (Wu et al., 2023; Huang and Marques-Silva, 2023; Izza and Marques-Silva, 2024; Izza et al., 2024). For example, it is the case that so-called abductive explanations (Marques-Silva and Ignatiev, 2022; Marques-Silva, 2024) are such that no adversarial examples can be identified.

This paper is in part motivated by a number of negative results in explainability (Letoffe et al., 2024; Marques-Silva and Huang, 2024; Huang and Marques-Silva, 2024; Izza et al., 2022; Marques-Silva and Ignatiev, 2023; Marques-Silva, 2023; Ignatiev, 2020; Marques-Silva and Ignatiev, 2022), but targets instead robustness. Moreover, the paper argues that existing definitions of local and global robustness are problematic. Concretely the paper shows that the experimental setup used for assessing robustness is invariably inconclusive with respect to deciding the robustness of the given ML model. In a similar fashion, the paper argues that efforts to deliver *robustness certification* can be ineffective. Motivated by these results, the paper hypothesizes that there is no simple solution to the basic shortcomings of existing approaches for deciding the robustness of ML models. Nevertheless, the paper also underlines the importance of robustness tools, not specifically for deciding robustness, but instead as a key building block for the computation of formal explanations by iteratively deciding the existence of adversarial examples.

A longer version of this paper is published on arXiv (Izza and Marques-Silva, 2023), which includes all proofs of the propositions and detailed results of the experiments conducted in this work .

2 PRELIMINARIES

Measures of Distance. We consider the well-known l_p measure of distance,

$$||\mathbf{x}-\mathbf{y}||_p = \left(\sum_{i=1}^{m} |x_i - y_i|^p\right)^{1/p} \quad (1)$$

Which is referred as the Minkowski distance. Special cases include the Manhattan distance l_1, the Euclidean distance l_2 and the Chebyshev distance l_∞, which is defined by,

$$\lim_{p\to\infty} ||\mathbf{x}-\mathbf{y}||_p = \max_{1\leq i\leq m} \{|x_i - y_i|\} \quad (2)$$

Moreover, l_0 denotes the Hamming distance, defined by:

$$||\mathbf{x}-\mathbf{y}||_0 = \sum_{i=1}^{m} \text{ITE}(x_i = y_i, 0, 1) \quad (3)$$

Classification Problems. A classification problem is defined on a set of features $\mathcal{F} = \{1,\ldots,m\}$ and a set of classes $\mathcal{K} = \{c_1, c_2, \ldots, c_K\}$. Each feature $i \in \mathcal{F}$ takes values from a domain $\mathcal{D}_i$. Domains can be categorical or ordinal. If ordinal, domains can be discrete or real-valued. Throughout the paper, and unless otherwise stated, domains will be assumed to be real-valued. Feature space is defined by $\mathbb{F} = \mathcal{D}_1 \times \mathcal{D}_2 \times \ldots \times \mathcal{D}_m$. The notation $\mathbf{x} = (x_1, \ldots, x_m)$ denotes an arbitrary point in feature space, where each x_i is a variable taking values from $\mathcal{D}_i$. Moreover, the notation $\mathbf{v} = (v_1, \ldots, v_m)$ represents a specific point in feature space, where each v_i is a constant representing one concrete value from $\mathcal{D}_i$. An *instance* denotes a pair $(\mathbf{v}, c)$, where $\mathbf{v} \in \mathbb{F}$ and $c \in \mathcal{K}$. An ML classifier $\mathcal{M}$ is characterized by a non-constant *classification function* κ that maps feature space $\mathbb{F}$ into the set of classes $\mathcal{K}$, i.e. $\kappa : \mathbb{F} \to \mathcal{K}$. Given the above, we associate with a classifier $\mathcal{M}$, a tuple $(\mathcal{F}, \mathbb{F}, \mathcal{K}, \kappa)$. Since we assume that κ is non-constant, then the ML classifier $\mathcal{M}$ is declared *non-trivial*, i.e. $\exists(\mathbf{a}, \mathbf{b} \in \mathbb{F}).(\kappa(\mathbf{a}) \neq \kappa(\mathbf{b}))$.

When reasoning about systems with formal methods, it is often assumed that all inputs are possible, or alternatively, that there exist explicit constraints that disallow some inputs. In contrast, reasoning methods rooted in machine learning usually assume some input distribution, which most often needs to be inferred (or approximated) from training data or be user-specified. We say that a point $\mathbf{x}$ is *viable* if, for reasoning purposes, $\mathbf{x}$ must be accounted for. As a result, when making statements about ML models, we consider three possible scenarios on the inputs:

1. *Unconstrained inputs*, i.e. any point $\mathbf{x}$ in feature space is viable.
2. *Constrained inputs*, i.e. any point $\mathbf{x}$ in feature space is viable iff some set of constraints $\mathcal{C}$ is satisfied by $\mathbf{x}$. This is represented by the predicate $\xi(\mathbf{x})$.
3. *Distribution-restricted* inputs, i.e. any point $\mathbf{x}$ in feature space is viable iff it respects some distribution $\mathcal{D}$ i.e. $\mathbf{x} \sim \mathcal{D}$, which is either user-specified or it is inferred from training data.

Most work on adversarial robustness implicitly assumes distribution-restricted inputs. In contrast, throughout the paper, we assume the case of *unconstrained inputs* for the following main reasons. First, the most rigorous approaches for robustness make that implicit assumption. Second, assuming that inputs are distribution-restricted does not account for data drift. Nevertheless, in cases where the distinction matters, the paper also accounts for the other possible cases on the inputs.

2.1 Adversarial Robustness

Local Robustness. Given a classifier $\mathcal{M}$ with a classification function κ, and an instance $(\mathbf{v}, c)$, the classifier is *locally robust* (or just robust) for $\mathbf{v}$ if,

$$\forall(\mathbf{x} \in \mathbb{F}).\, [||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \quad (4)$$

If the classifier is not robust, then any point $\mathbf{x} \in \mathbb{F}$ satisfying the condition,

$$[||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon \land (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v}))] \quad (5)$$

is referred to as an *adversarial example* for distance ε (εAEx). (Observe that (5) consists of selecting one of the counterexamples to (4).)

The definitions of local robustness in (4) and of adversarial example in (5) assume unconstrained inputs. For completeness, we include the definitions of local robustness for the other cases regarding assumptions on the inputs.

For constrained inputs we have,

$$\forall(\mathbf{x} \in \mathbb{F}).\, [\xi(\mathbf{x}) \land (||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon)] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \quad (6)$$

For distribution-restricted inputs we have,

$$\mathcal{Q}_{\varsigma}(\mathbf{x} \sim \mathcal{D}).\, [||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \quad (7)$$

where $\mathcal{Q}_{\varsigma}$ captures the sampling according to some distribution and target confidence (ς), and where $\sim$ is interpreted as a predicate, with two arguments $\mathbf{x}$ and $\mathcal{D}$, that holds iff $\mathbf{x}$ respects the distribution $\mathcal{D}$.

There exist a multitude of proposed robustness tools dedicated to local robustness, many of which are regularly assessed in the VNN-COMP (Brix et al., 2023b). Section 3 briefly overviews existing work on robustness.

One additional observation is that tools that exploit automated reasoners assume unconstrained inputs, and so the definition of local robustness considered is (4). If information about the context in which the ML models is to be deployed, then $\xi(\mathbf{x})$ may be available, and so it is to be expected that (6) would be used instead. Finally, incomplete methods, often based on the sampling of feature space, assume the somewhat different definition (7), which can offer probabilistic guarantees, but not absolute guarantees.

Certified Robustness. Earlier research have also proposed *certified robustness*, which has been defined as follows:

Definition 1 (From (Cohen et al., 2019))**.** "A classifier is said to be certifiably robust if **for any input x**, one can easily obtain a guarantee that the classifier's prediction is constant within some set around **x**, often an l_2 or l_∞ ball".

(We underscore that Definition 1 is taken verbatim from (Cohen et al., 2019), although we highlight the universal quantification on the inputs.) We will refer to this definition throughout the paper.

Global Robustness. Given the definition of local robustness, a possible definition of *global robustness* is,

$$\forall(\mathbf{v}, \mathbf{x} \in \mathbb{F}).\, (||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon) \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \quad (8)$$

(Observe that (8) is just a formalization of Definition 1, by allowing the norm l_p to be any.) Similar definitions have been studied in the literature (Seshia et al., 2018; Narodytska, 2018; Cohen et al., 2019; Rosenfeld et al., 2020; Dvijotham et al., 2020; Huang et al., 2021; Chen et al., 2021; Carlini et al., 2023). For example, Reluplex (Katz et al., 2017) defines global robustness by allowing small differences in predictions. This alternative definition raises concerns in classifications problems, because the differences between predicted values may be small, but the predicted classes may be different. Moreover, there are other variants of this definition, which the paper also studies. However, by default we will assume this definition throughout the paper. (As shown in the paper, this apparently sensible definition actually raises a number of critical issues.)

As shown in Section 3, other definitions of global robustness can be related with the one proposed above. Section 3 also briefly overviews approaches for deciding global robustness, local robustness and certified robustness.

Running Example. To motivate the claims in the paper, the following very simple classifiers are used as the running examples throughout the paper.

Example 1. *A first classifier $\mathcal{M}_1$ is defined on a single feature $\mathcal{F}_1 = \{1\}$, with $\mathbb{D}_{11} = \mathbb{R}$. The set of classes is $\mathcal{K}_1 = \{0,1\}$, and the training data is given by: $\{(0.0,0),(0.3,0),(0.4,0),(0.7,1),(1.0,1)\}$. Furthermore, we use an off-the-shelf ML toolkit, e.g. scikit-learn (Pedregosa et al., 2011), to learn the classifier's function κ_1 as a* linear classifier *$\kappa_1 : \mathbb{D}_{11} \rightarrow \mathcal{K}_1$. Accordingly, the model learned by scikit-learn is,*

$$\kappa(x_1) = \text{ITE}(0.93198992 \times x_1 - 0.64735516 \geq 0, 1, 0)$$

As can be observed, the accuracy of the learned classifier over training data is 100%. Moreover, the question we seek to answer is: is the classifier (locally or globally) robust?

Example 2. *A second classifier $\mathcal{M}_2$ is obtained from the first one above (see Example 1), but defined as*

follows:

$$\kappa_2(x_1,x_2) = \begin{cases} \kappa_1(x_1) & \text{if } x_1 \leq 1 \\ \text{ITE}(x_1 > x_2, 1, 0) & \text{otherwise} \end{cases}$$

In this case, $\mathcal{F}_2 = \{1,2\}$, $\mathbb{D}_{21} = \mathbb{D}_{22} = \mathbb{R}$, $\mathbb{F} = \mathbb{R} \times \mathbb{R}$, *and* $\mathcal{K}_2 = \{0,1\}$.

Example 3 exemplifies the definition of adversarial examples.

Example 3. *For the classifier of Example 1, and for the instance* $(0.7,1)$*, from training data (and assuming 100% accuracy on training data), it is apparent that an AEx exists by setting* $\varepsilon = 0.3$*. By manual inspection of the learned model, we conclude that we obtain an AEx with a smaller* ε*, e.g.* $\varepsilon = 0.1$ *suffices.*

2.2 Symbolic Explainability

As mentioned in Section 1, the concepts of adversarial examples and explainability are tightly related. As a result, we include a brief introduction to symbolic (or logic) explainability. More detailed accounts are available (Marques-Silva and Ignatiev, 2022; Darwiche, 2023; Marques-Silva, 2024).

An explanation problem $\mathcal{E}$ is a tuple $(\mathcal{M},(\mathbf{v},c))$, where $\mathcal{M}$ is a classifier, and $(\mathbf{v},c)$ is an instance. When describing concepts in explainability, it is to be understood an underlying explanation problem $\mathcal{E}$. Prime implicant (PI) explanations (Shih et al., 2018) denote a minimal set of literals (relating a feature value x_i and a constant v_i from its domain $\mathcal{D}_i$) that are sufficient for the prediction. PI-explanations can be formulated as a problem of logic-based abduction, and so are also referred to as *abductive explanations* (AXp) (Ignatiev et al., 2019). Formally, given $\mathbf{v} = (v_1,\ldots,v_m) \in \mathbb{F}$ with $\kappa(\mathbf{v}) = c$, an AXp is any minimal subset $\mathcal{X} \subseteq \mathcal{F}$ such that,

$$\forall(\mathbf{x} \in \mathbb{F}).\left[\bigwedge\nolimits_{i\in\mathcal{X}}(x_i = v_i)\right] \rightarrow (\kappa(\mathbf{x}) = c) \quad (9)$$

AXps can be viewed as answering a 'Why?' question, i.e. why is some prediction made given some point in feature space. A different view of explanations is a contrastive explanation (Miller, 2019), which answers a 'Why Not?' question, i.e. which features can be changed to change the prediction. A formal definition of contrastive explanation is proposed in recent work (Ignatiev et al., 2020). Given $\mathcal{E} = (\mathcal{M},(\mathbf{v},c))$, a CXp is any minimal subset $\mathcal{Y} \subseteq \mathcal{F}$ such that,

$$\exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{i\in\mathcal{F}\setminus\mathcal{Y}}(x_i = v_i) \land (\kappa(\mathbf{x}) \neq c) \quad (10)$$

Note that any set $\mathcal{Y} \subseteq \mathcal{F}$ for which (10) holds but not necessarily minimal is referred to as *weak* CXp. Similarly, any $\mathcal{X} \subseteq \mathcal{F}$ for which (9) is a *weak* AXp.

The relationship between explanations and adversarial examples can be further clarified (Izza et al., 2024), as follows.

Proposition 1. *Given an explanation problem* $\mathcal{E} = (\mathcal{M},(\mathbf{v},c))$*,* $\mathcal{Y} \subseteq \mathcal{F}$ *is a weak CXp iff* $\mathcal{M}$ *has an* εAEx*, with* l_0 *distance* $\varepsilon = |\mathcal{Y}|$*.*

Building on the results of R. Reiter in model-based diagnosis (Reiter, 1987), (Ignatiev et al., 2020) proves a minimal hitting set (MHS) duality relation between AXps and CXps, i.e. AXps are MHSes of CXps and vice-versa. Thus, as long as one can devise logic encodings for an ML classifier (and this is possible for most ML classifiers) and have access to a suitable reasoner, then (9) and (10) offer a solution for computing one AXp/CXp.

Distance-Restricted Explanations. With the purpose of relating explanations with robustness, recent work (Huang and Marques-Silva, 2023; Wu et al., 2023) introduced the concept of distance-restricted explanation.

Given an explanation problem $\mathcal{E} = (\mathcal{M},(\mathbf{v},c))$ and $\varepsilon > 0$, a distance-restricted AXp (εAXp) is a subset-minimal set of features $\mathcal{X} \subseteq \mathcal{F}$ such that,

$$\forall(\mathbf{x} \in \mathbb{F}).\Big[\bigwedge\nolimits_{i\in\mathcal{X}}(x_i = v_i) \land ||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon] \rightarrow (\kappa(\mathbf{x}) = c) \quad (11)$$

We define distance-restricted CXps (εCXps) accordingly. A εCXp is a subset-minimal set of features $\mathcal{Y} \subseteq \mathcal{F}$, such that,

$$\exists(\mathbf{x} \in \mathbb{F}).\Big[\bigwedge\nolimits_{i\in\mathcal{F}\setminus\mathcal{Y}}(x_i = v_i) \land ||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon \land (\kappa(\mathbf{x}) \neq c)] \quad (12)$$

MHS duality between distance-restricted AXps and CXps has been proved, which enables the development of algorithms for navigating the space of εAXps and εCXps (Izza et al., 2024).

Finally, there is a simple relationship between (distance-unrestricted or plain) AXps/CXps and εAXps/εCXps. By picking l_0, i.e. Hamming distance, and letting $\varepsilon = m$, i.e. the number of features, then εAXps/εCXps represent *exactly* the (plain) AXps/CXps. Observe that, by setting $\varepsilon = m$, we allow any subset of the features to be included/excluded from AXps/CXps. Hence, we will be computing distance-unrestricted AXps/CXps using algorithms developed for εAXps/εCXps.

3 RELATED WORK

The realization that ML models are most often brittle (Biggio et al., 2013; Szegedy et al., 2014; Goodfellow et al., 2015), i.e. that ML models can exhibit adversarial examples, motivated a massive body of research over the last decade on deciding the robustness of ML models. The goal of this section is to briefly overview works that are of special interest to the paper's topics, especially ML model robustness and the identification of AExs. Moreover, a growing number of surveys (Wiyatno et al., 2019; Zhang and Li, 2020; Chen et al., 2020; Chakraborty et al., 2021; Rosenberg et al., 2022; Liang et al., 2022; Zhang et al., 2022b; Zhou et al., 2023; Han et al., 2023) illustrate the importance of robustness and AExs for the practical deployment of ML models.

Local Robustness. The definition of local robustness proposed in most of past work matches the one used in this paper (see (7)). Examples of tools that decide local robustness are those evaluated in VNN-COMP (Brix et al., 2023b).

Global Robustness. Past work considers global robustness as proposed in Eq. (8), which formalizes Definition 1. This is the case with (Seshia et al., 2018; Chen et al., 2021), but also (Cohen et al., 2019; Rosenfeld et al., 2020; Dvijotham et al., 2020; Huang et al., 2021; Carlini et al., 2023). Some works propose a slightly modified definition of global robustness (Katz et al., 2017):

$$\forall(\mathbf{v},\mathbf{x}\in\mathbb{F}).\,(||\mathbf{x}-\mathbf{v}||_p\leq\varepsilon)\rightarrow(|\kappa(\mathbf{x})-\kappa(\mathbf{v})|\leq\delta) \tag{13}$$

where $\delta > 0$. When compared with (8), the modified definition targets neural networks, especially when these compute real-valued outputs. A number of works adopt this definition of global robustness (Wang et al., 2022a; Wang et al., 2022b; Fu et al., 2022), but (Fu et al., 2022) imposes no constraint on $\mathbf{x}$ and $\mathbf{v}$. It should be noted that this definition is not without problems. For ML classifiers, e.g. image classification, conditions on the values of the outputs are uninteresting, and so a definition similar to (8) must be considered.

A different approach is adopted in (Ruan et al., 2019) where global robustness is defined with respect to a *finite set* of points in feature space, and not *all* points in feature space. Yet another take on global robustness is to reject inputs that are classified as AEx (Leino et al., 2021; Baharlouei et al., 2023). Thus a model can return class *abstain* on a given input $\mathbf{x}$. Finally, another line of research is represented by DeepSafe, which finds safe regions where robustness is guaranteed (Gopinath et al., 2018; Dimitrov et al., 2022).

Robustness Certification. Work on certifying robustness can be traced to (Cohen et al., 2019; Weng et al., 2019; Gehr et al., 2018; Singh et al., 2018) and more recently (Rosenfeld et al., 2020; Dvijotham et al., 2020; Huang et al., 2021; Vorácek and Hein, 2023; Carlini et al., 2023) that leverage on local robustness property to provide certification and/or quantify robustness of models against AExs. As illustration, empirical evaluation reported in (Carlini et al., 2023) (resp. (Vorácek and Hein, 2023)) considers a collection of 100,000 and 10,000 (resp. 2000 and 500) samples, respectively, drawn from CIFAR10 and ImageNet datasets, that serve to certify *robustness accuracy*, i.e. percentage of samples failed/succeeded in the local robustness test. Besides, works reported in (Liu et al., 2020; Wang et al., 2022b; Wang et al., 2022a) adopt global robustness property to certify whether or not the analyzed model is robust. Furthermore, some works (Fu et al., 2022; Wang et al., 2022a) use local and global robustness techniques to measure lower and upper bounds for robustness. Another recent work (Dimitrov et al., 2022) computes regions with robustness certification on all possible points in these regions. In these earlier works, the implications of global robustness on local robustness are not discussed.

4 THE CONS OF ROBUSTNESS

This section proves a number of negative results regarding global and local robustness, but also regarding robustness certification. More importantly, those negative results impact the conclusions drawn in earlier work on robustness. Nevertheless, this section also discusses ways to cope with these negative results.

4.1 Basic Negative Results

4.1.1 Continuous Domains

There Is no Global Robustness. A straightforward observation is that, given the proposed definition of global robustness (see (8)), then there exist no non-trivial globally robust classifiers.

Proposition 2. *Any non-trivial classifier defined on continuous (real-valued) features is not globally robust, independently of the value of* ε *chosen. (We as-*

sume that the measure of distance considered is l_p, with $p \geq 1$.)

An observation that mimics Proposition 2 is made in recent work (Leino et al., 2021). However, the consequences of such observation were not investigated further. Instead, a proposed solution to the problem of global robustness was to change the training of the classifier to return an indication of inputs that cannot be guaranteed to be robust. Hence, the solution proposed is to move from *a posteriori* deciding robustness to training for robustness.

Example 4. *With respect to Example 1, recall that the model learned by scikit-learn is,*

$$\kappa(x_1) = \text{ITE}(0.93198992 \times x_1 - 0.64735516 \geq 0, 1, 0)$$

Clearly, the value of x_1 for which the predicted class transitions from 0 to 1 is 0.69459459. Hence, exhibiting point $x_1 = 0.69459459$ is a proof that the model is not *globally robust.*

A key conclusion of the results above is that the definition of (certified) robustness used in earlier works (Cohen et al., 2019; Narodytska, 2018) (among others, see also Section 2) can only be achieved with an ML model that predicts a constant value; and this of course unsatisfactory. Furthermore, the solution proposed in other works (Lécuyer et al., 2019), i.e. to restricted robustness certification to a fixed test set, is also unsatisfactory because validating robustness on such a fixed test set defeats the purpose of machine learning.

No Global Robustness Implies no Local Robustness. One observation that stems from the proof of Proposition 2 is that deciding local robustness is also problematic. Concretely, if we are allowed to select a suitable point in feature space, then it is the case that,

Proposition 3. *For any non-trivial classifier defined on real-valued features, there exists a point for which the classifier is not locally robust, independently of the value of ε chosen.*

Example 5. *With respect to Example 1, if we sample training data, then the model will be declared locally robust for $\varepsilon < 0.00540541$, due to point $x_1 = 0.7$. Clearly, if we allow complete freedom on which values of x_1 to sample, we will then conclude that, for $x_1 = 0.69459459$, robustness is non-existing for any value of $\varepsilon > 0$, and so the classifier is not robust.*

Observe that what Proposition 3 claims is that one can always find points in feature space for which local robustness does not hold. Thus, claiming local robustness based on successful proving local robustness for some selected points in feature space, does not equate with global robustness holding in all points in feature space.

The Robustness of Non-Trivial Classifiers Cannot Be Certified. In recent years, a number of works have studied robustness certification (Cohen et al., 2019; Dvijotham et al., 2020; Rosenfeld et al., 2020; Huang et al., 2021; Carlini et al., 2023).[1] Using the definition of robustness certification proposed in Section 2 (see Definition 1), which is taken verbatim from (Cohen et al., 2019), then we can claim that,

Proposition 4. *For any non-trivial classifier defined on real-valued features, robustness cannot be certified, independently of the value of ε chosen.*

Example 6. *The analysis of Example 5 also demonstrates that robustness certification will fail for the example classifier, for any $\varepsilon > 0$, as long as the point $x_1 = 0.69459459$ is analyzed. In contrast, if the training data were to be sampled for local robustness, then one would declare robustness to be certified for $\varepsilon < 0.00540541$. Evidently, such conclusion would be in error.*

Counterexamples to Local Robustness. Under the assumption of unconstrained inputs, we proved earlier in this section that no ML model is (locally) robust. As a result, and given some $\varepsilon > 0$, counterexamples to (unconstrained) local robustness are guaranteed to exist, and can be obtained from any pair of points $\mathbf{v}, \mathbf{x}$ in feature space such that,

$$\exists(\mathbf{x}, \mathbf{v} \in \mathbb{F}).\, (||\mathbf{x} - \mathbf{v}||_p \leq \varepsilon) \land (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v})) \quad (14)$$

Observe that requiring $(\mathbf{x} \neq \mathbf{v})$ is unnecessary. Also, we note that (14) is just the negation of (8), i.e. the condition for global robustness. In addition, it is easy to accommodate the cases where a classifier can flag points as not being robust (Leino et al., 2021). (Earlier work proposes that such points be flagged with a different class $\perp$, but in this paper we use $\oslash$ instead.) Finally, it is also plain to formulate the search for counterexamples as a decision problem. Given some logic formula φ and a target logic theory $\mathcal{T}$, then $[\![\varphi]\!]_{\mathcal{T}}$ represents the encoding of φ in the target logic theory $\mathcal{T}$. Also, let $\mathsf{SAT}_{\mathcal{T}}$ represent a reasoner for theory $\mathcal{T}$. Then, we use (14) to formulate the decision problem

[1] By being based on restricted forms of sample, the approaches used for *robustness certification* do not capture the dictionary meaning of *certification*, i.e. *to state in a formal way that something is correct*, and are fundamentally different from the meaning ascribed to certification in formal methods.

in theory $\mathcal{T}$ as follows:

$$\mathsf{SAT}_{\mathcal{T}}\left([\![(||\mathbf{x}-\mathbf{v}||_p \leq \varepsilon) \land (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v}))]\!]_{\mathcal{T}}\right) \quad (15)$$

4.1.2 Discrete Domains

The negative results in Section 4.1.1 also hold in the case of classifiers with discrete features. However, the values of ε considered are further constrained. First, we consider a classifier with categorical features, an l_0 norm and unconstrained inputs. In this case, it is guaranteed that one cannot have εAExs if $\varepsilon < 1$, i.e. if one prevents any feature from changing value. If we impose the constraint $\varepsilon \geq 1$, then the results of Section 4.1.1 hold, i) no non-trivial classifier is globally robust; ii) for any non-trivial classifier there are points in feature space that are not locally robust; and iii) robustness cannot be certified.

There are also settings where features are discrete and result from discretizing real-valued features. This is the case for example when using binarized neural networks (BNNs) (Hubara et al., 2016). In these cases we assume a discretization step δ. As a result, the comments made for classifiers with categorical features also apply in this case, with the constraint that $\varepsilon \geq \delta$, i.e. the smallest distance to consider is no less than the discretization step.

For example, for the experiments of Section 6, when the l_0 norm is used, it is the case that $\delta = 1$. As a result, in the experiments we used $\varepsilon = 1$, thus targeting the *smallest* distance that could possibly be considered. As the results show, and as it should be expected, one can find counterexamples to local/global robustness for all the experiments.

4.2 Practical Consequences

The results in Section 4.1.1 have important practical consequences. First, the experimental setup most often used in the assessment of local robustness exhibits critical shortcomings. In a significant body of earlier work, local robustness is assessed by randomly sampling feature space. This is the case with evaluations of local robustness in (Brix et al., 2023b). This means that, either sampling does not pick the right points in feature space, and so one is allowed to (incorrectly) decide for local robustness, or sampling picks one of the points that must exist, and so (the expected) non local robustness is decided. One paradigmatic example is VNN-COMP(Brix et al., 2023b), where uniform random sampling has been employed in all the previous competitions. The bottom line is that any classifier declared local robust is guaranteed *not* to be so. The same remarks apply in the case of global robustness, but in this case the number of existing works is a fraction of the number of works on deciding local robustness.

Examples of Ineffective Robustness Assessment. One example of the limitations of randomly sampling for assessing robustness is VNN-COMP (Brix et al., 2023b; Brix et al., 2023a). The description of the most recent competitions (Müller et al., 2022; Brix et al., 2023c; Brix et al., 2023a) confirms that robustness is assessed by randomly sampling existing datasets. In these cases, *any* results indicating that NNs are robust are necessarily *incorrect*. Furthermore, past works claiming robustness certification are inaccurate. Methods based on automated reasoners, which assume unconstrained inputs, will be in error if ML models are declared robust. Distribution-restricted methods can only provide probabilistic guarantees. In such cases, one must trust that the inferred input distribution faithfully captures possible inputs to the ML model. More importantly, one must also trust that the sampling methods used will offer enough rigor.

Besides the shortcomings of local robustness and certification, the limitations of existing methods for attaining global robustness are also clear.

Practical Assessment of Shortcomings. As noted earlier (15), in practice it is conceptually simple to demonstrate the limitations of local robustness using a dedicated automated reasoner. It suffices to decide the existence of a point $\mathbf{z}$ in feature space which, for arbitrarily small ε, it is the case that there exists a point in the ε ball corresponding to a prediction other than $\kappa(\mathbf{z})$. Section 6 summarizes results illustrating not only the existence of such points for complex classifiers, but also the practical scalability of this approach.

Existing Solutions. Some approaches avoid the problems reported in this section by curbing the claims about (certified) robustness. For example, the definition of global robustness in some works (Ruan et al., 2019) considers a finite set of points where local robustness is assessed. As long as the inputs respect such a finite set of points, then the non-existence of AExs for robust ML models is guaranteed. Unfortunately, if the inputs were to be known, then the need for ML models would be non-existing. In a similar vein, other tools (Gopinath et al., 2018) ensure robustness in specific regions of feature space. Although more general that (Ruan et al., 2019), similar limitations apply.

Some works exploit sampling of the inputs according to some inferred input distribution (Cohen

et al., 2019; Rosenfeld et al., 2020; Dvijotham et al., 2020; Huang et al., 2021; Carlini et al., 2023). Such works offer no formal guarantees of rigor, and so claims of *certification* hold probabilistically assuming that all possible inputs respect the assumed input distribution.

4.3 Threats to Validity & Discussion

One possible criticism to the results in this section is that some works on robustness do not assume unconstrained inputs, but instead sample the inputs according to the distribution inferred from training data or specific imposed distribution (Weng et al., 2019; Cohen et al., 2019; Yan et al., 2024). In such a situation, one can argue that both local and global robustness can still be safely decided. Clearly, the claims of rigor differ in the two cases. As the paper shows, approaches based on automated reasoners that consider all possible inputs in feature space cannot provide guarantees of robustness. However, the same applies to any other approach when one seeks the strongest guarantees of robustness. In addition, there is recent work showing the importance of sampling out of the distribution (Yin et al., 2019; Hendrycks et al., 2020; Hendrycks et al., 2021; Zhang et al., 2022a), but this again exhibits the shortcomings of local/global robustness identified earlier in this section.

Furthermore, the sampling of inputs according to some inferred input distribution is not without clear limitations. First, inferring an input distribution from a negligible fraction of feature space can be a source of error. Second, an input distribution declares inputs more or less likely, but does not prevent the possibility, no matter how negligible, of inputs not in the distribution. Third, robustness tools based on automated reasoners assume that *all* inputs are possible, i.e. unconstrained inputs are assumed, when deciding local robustness for a concrete point in feature space (Katz et al., 2017). If input distributions were to be taken into account for such tools when deciding the points in feature space to analyze, then even for deciding local robustness for a single point in feature space, input distributions would have to be accounted for.

To the best of our knowledge, there is no clear way on how to instrument automated reasoners to account for input distributions. The key point is that sampling according to the input distribution could be a source of error, and so the integration with tools based on automated reasoners is unclear.

A different approach is to explicitly assume that some inputs are not accepted. For example, allowing an elderly person to be of a fairly young age. To the best of our knowledge, part work on robustness does not account for disallowed inputs. In contrast, the topic has been researched in formal explainability (Gorji and Rubin, 2022; Yu et al., 2023), by considering constraints on the inputs. Future work may re-analyze robustness in light of such constraints.

5 SOME PROS OF ROBUSTNESS

Under a scenario in which all inputs are possible, Section 4 raises concerns about the usefulness of attempting to prove robustness, be it local or global. This might seem to suggest the ongoing efforts towards devising efficient and rigorous robustness tools are misguided. Despite the negative results of the previous section, robustness should be expected to find critical applications in the near future. This section briefly overviews one such application.

From εAExs to εAXps/εCXps. Recent work revealed a tight relationship between formal explanations and the non-existence of (constrained) adversarial examples (Huang and Marques-Silva, 2023), enabled by the concept of distance-restricted AXps/CXps (see definition in Page 4). Furthermore, MHS duality between distance-restricted AXps and CXps enable the navigation of the space of AXps and CXps. More importantly, the computation of εAXps/εCXps can be instrumented using tools for finding εAExs.

Throughout this section, we assume that the existence of εAExs is decided by calls to a suitable oracle. The procedure invoking the oracle is represented by a predicate $\mathsf{FindAEx}(\varepsilon, Q; \mathcal{E}, p)$, parameterized by the explanation problem $\mathcal{E} = (\mathcal{M}, (\mathbf{v}, x))$, where $\mathcal{M}$ is the classifier, and by the norm l_p, and with arguments $\varepsilon > 0$ and $Q \subseteq \mathcal{F}$. Furthermore, the predicate $\mathsf{FindAEx}(\varepsilon, Q; \mathcal{E}, p)$ holds true if the ML classifier $\mathcal{M}$ exhibits an AEx within distance ε, with the features in Q fixed to the values dictated by $\mathbf{v}$.

To illustrate the relationships between explanations and adversarial examples, we propose a simple linear search algorithm for computing one εCXp, depicted in Algorithm 1. (In the same vein, the linear search can be adapted for computing one εAXp, as shown in (Wu et al., 2023; Izza et al., 2024)). One argument is the value of ε, whereas the other argument is a set of features that must be kept free (for CXps) or fixed (for AXps) . The loop invariant is that $\mathcal{S}$ is a weak εCXp (resp. εAXp). In the case of an εCXp, the features in $\mathcal{S}$, if freed, cause that an AEx can be identified (given the distance ε). In the case of an εAXp, the features in $\mathcal{S}$, if fixed, cause that no AEx can be identified (given the distance ε). Features are freed/fixed

Algorithm 1: Linear algorithm to find CXp using AEx.

Input: Arguments: $\mathcal{F}$, ε; Parameters: $\mathcal{E}$, p
Output: One CXp $\mathcal{S}$

function FindCXpDel($\mathcal{F}, \varepsilon; \mathcal{E}, p$) ▷ Inv: FindAEx($\varepsilon, \mathcal{F}$)
  $\mathcal{S} \leftarrow \mathcal{F}$ ▷ Initially, all features are free
  for $i \in \mathcal{F}$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$ ▷ Fix feature i
    outc = FindAEx($\varepsilon, \mathcal{F} \setminus \mathcal{S}; \mathcal{E}, p$)
    if not outc **then** ▷ no AEx
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{i\}$ ▷ Free again feature i
  return $\mathcal{S}$

as long as the loop invariant is preserved. An oracle for deciding the existence of an AEx is used to decide whether or not the loop invariant is preserved when another feature is freed/fixed. It can be shown that the final set $\mathcal{S}$ is a εCXp/εAXp.

Navigating the Space of εAXps/εCXps. Besides computing one distance-restricted explanation, one may be interested in navigating the sets of distance-restricted explanations. For example, we may be interested in deciding whether a sensitive feature can occur in some explanation, or we may just be interested in finding some other explanation when the reported one is uninteresting.

New Insights into Explanations. For non-trivial classifiers defined on real-values features, the fact that local robustness does not hold on all points of feature space reveals not only the guaranteed existence of adversarial examples, but it also reveals new properties about explanations. We discuss one such example.

Proposition 5. *For a non-trivial classifier defined on real-valued features, and for a measure l_p, there exist instances for which, for any $\varepsilon > 0$, there exist* εAXps *and* εCXps.

6 EXPERIMENTAL EVIDENCE

This section presents a summary of practical evidence of our results on global robustness (and so indirectly on the impossibility of global local robustness) for the case study of dense NNs trained with image datasets.

The assessment is performed on a selection of 5 NNs examples, publicly available, used in robustness formal verification. (Additional results on discrete data (Binarized NNs) are included in the appendix of (Izza and Marques-Silva, 2023).) Furthermore, the experiments are conducted on a MacBook Pro with a Dual-Core Intel Core i5 2.3GHz CPU with 8GByte RAM running macOS Ventura. The time limit is set 3600 s and the memory limit is set to 16 Gb.

We implemented a formal global robustness verifier for NNs in Python. Concretely, we generate two copies of the neural network in ONNX format, one replica that represents $\kappa(\mathbf{x})$ and another one for $\kappa(\mathbf{y})$ and then encode the constraints on the input layer to enforce $||\mathbf{x}-\mathbf{y}||_p \leq \varepsilon$ and output layer to enforce them to pick different classes, i.e. $\kappa(\mathbf{x}) \neq \kappa(\mathbf{y})$. Moreover, Marabou oracle[2] (Katz et al., 2019) is instrumented to solve the robustness targeted problem.

Table 1: Assessment of global robustness verification for deep NNs. The table shows results for 4 image datasets.

Model	ε	AEx	Time
KJ_TinyTaxiNet	0.1	✓	0.069
KJ_TinyTaxiNet	0.05	✓	0.070
KJ_TinyTaxiNet	0.001	✓	0.113
MNIST-dense	0.1	✓	0.897
MNIST-dense	0.05	✓	0.899
MNIST-dense	0.001	✓	1.805
MNIST-conv	0.1	—	*TO*
cifar-convSmall	0.1	—	*TO*
gtsrb-dense	0.1	✓	42.535
gtsrb-dense	0.05	✓	28.677
gtsrb-dense	0.001	✓	50.556

Table 2: Detailed performance evaluation of computing εCXp for DNNs. Columns **Avg** and **nCalls** in column **AEx** report, resp. the average time and total number of instrumented oracle (AEx robustness) calls. Column **Avg** (resp. **Mn** and **Mx**) in column **εCXp** reports the average time (resp. min and max) to deliver a εCXp and **Len** reports the average length of the εCXps.

Model	AEx			εCXp			
	ε	Avg	nCalls	Len	Mn	Mx	Avg
gtsrb	0.03	0.08	1023	218	74.2	87.4	77.1
mnist	0.08	0.41	464	360	131.0	355.6	188.3

Table 1 summarizes the results on deep neural networks, on different l_∞ Chebyshev distance ε value ranging from 0.001 to 0.1 for each considered classifier. As can be observed from the results, global robustness formulation is able to identify adversarial examples for all tests, with a few exception when the neural network reasoner (i.e. Marabou) exceeds the time limit. Clearly, our results confirm the theoretical findings presented earlier that global robustness query will always report an adversarial example for

[2]Marabou is a complete neural network verifier powered with an SMT solver CVC4 (Barrett et al., 2011).

any non-trivial classifier.

Besides, we assess the performance-wise our basic algorithm for computing εCXps using AEx search. Results of the evaluation on *mnist* and *gtsrb* benchmarks are reported in Table 2. As can be seen from the table, the average number of pixels in generated εCXps is relatively small w.r.t. the image size of considered data, e.g. ∼21% for *gtsrb* and ∼46% for *mnist*, which illustrates the succinctness of the explanations and subsequently more interpretable. Moreover, the average runtimes are less than 77.1 and 188.3 seconds (maximum 87.4 and 355.6 seconds), resp., for *gtsrb* and *mnist* DNN, which demonstrates the practical effectiveness of our method of image classification benchmarks. The focus of our future works will be on devising efficient algorithms to compute smallest (cardinality) contrastive explanations.

7 CONCLUSIONS

This paper presents simple arguments that reveal the shortcomings of deciding robustness, be it global or local. Similarly, the paper uncovers related pitfalls of attempts to certifying robustness, namely when inputs are unconstrained. In addition, the paper also argues that possible attempts at solving the identified shortcomings are not entirely satisfactory.

In contrast to the negative results presented in the paper, the paper also details recently proposed uses of robustness tools, building on the connections between adversarial examples and explainability. Furthermore, the negative results on robustness are used to shed light on the properties of distance-restricted explanations of ML models. Future work will further investigate the links between adversarial examples and symbolic explanations, e.g. smallest size contrastive explanations will be investigated.

ACKNOWLEDGEMENTS

This work was supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, by the Spanish government under grant PID2023-152814OB-100, and by ICREA starting funds.

REFERENCES

Baharlouei, S., Sheikholeslami, F., Razaviyayn, M., and Kolter, Z. (2023). Improving adversarial robustness via joint classification and multiple explicit detection classes. In *AISTATS*, pages 11059–11078.

Barrett, C. W., Conway, C. L., Deters, M., Hadarean, L., Jovanovic, D., King, T., Reynolds, A., and Tinelli, C. (2011). CVC4. In *CAV*, pages 171–177.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *ECML*, pages 387–402.

Brix, C., Bak, S., Liu, C., and Johnson, T. T. (2023a). The fourth international verification of neural networks competition (VNN-COMP 2023): Summary and results. *CoRR*, abs/2312.16760.

Brix, C., Müller, M. N., Bak, S., Johnson, T. T., and Liu, C. (2023b). First three years of the international verification of neural networks competition (VNN-COMP). *Int. J. Softw. Tools Technol. Transf.*, 25(3):329–339.

Brix, C., Müller, M. N., Bak, S., Johnson, T. T., and Liu, C. (2023c). First three years of the international verification of neural networks competition (VNN-COMP). *CoRR*, abs/2301.05815.

Carlini, N., Tramèr, F., Dvijotham, K. D., Rice, L., Sun, M., and Kolter, J. Z. (2023). (certified!!) adversarial robustness for free! In *ICLR*.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45.

Chen, K., Zhu, H., Yan, L., and Wang, J. (2020). A survey on adversarial examples in deep learning. *Journal on Big Data*, 2(2):71.

Chen, Y., Wang, S., Qin, Y., Liao, X., Jana, S., and Wagner, D. A. (2021). Learning security classifiers with verified global robustness properties. In *CCS*, pages 477–494.

Cohen, J., Rosenfeld, E., and Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In *ICML*, pages 1310–1320.

Darwiche, A. (2023). Logic for explainable AI. In *LICS*, pages 1–11.

Dimitrov, D. I., Singh, G., Gehr, T., and Vechev, M. T. (2022). Provably robust adversarial examples. In *ICLR*.

Dvijotham, K. D., Hayes, J., Balle, B., Kolter, J. Z., Qin, C., György, A., Xiao, K., Gowal, S., and Kohli, P. (2020). A framework for robustness certification of smoothed classifiers using F-divergences. In *ICLR*.

Fu, F., Wang, Z., Fan, J., Wang, Y., Huang, C., Chen, X., Zhu, Q., and Li, W. (2022). REGLO: Provable neural network repair for global robustness properties. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS*.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. T. (2018). AI2: safety and robustness certification of neural networks with abstract interpretation. In *IEEE S&P*, pages 3–18.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*.

Gopinath, D., Katz, G., Pasareanu, C. S., and Barrett, C. W. (2018). DeepSafe: A data-driven approach for assessing robustness of neural networks. In *ATVA*, pages 3–19.

Gorji, N. and Rubin, S. (2022). Sufficient reasons for classifier decisions in the presence of domain constraints. In *AAAI*, pages 5660–5667.

Han, S., Lin, C., Shen, C., Wang, Q., and Guan, X. (2023). Interpreting adversarial examples in deep learning: A review. *ACM Computing Surveys*.

Hein, M. and Andriushchenko, M. (2017). Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NeurIPS*, pages 2266–2276.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8320–8329.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. In *ACL*, pages 2744–2751.

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., and Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.*, 37:100270.

Huang, X. and Marques-Silva, J. (2023). From robustness to explainability and back again. *CoRR*, abs/2306.03048.

Huang, X. and Marques-Silva, J. (2024). On the failings of shapley values for explainability. *Int. J. Approx. Reason.*, page 109112.

Huang, Y., Zhang, H., Shi, Y., Kolter, J. Z., and Anandkumar, A. (2021). Training certifiably robust neural networks with efficient local lipschitz bounds. In *NeurIPS*, pages 22745–22757.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. (2016). Binarized neural networks. In *NeurIPS*, pages 4107–4115.

Ignatiev, A. (2020). Towards trustable explainable AI. In *IJCAI*, pages 5154–5158.

Ignatiev, A., Narodytska, N., Asher, N., and Marques-Silva, J. (2020). From contrastive to abductive explanations and back again. In *AIxIA*, pages 335–355.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019). Abduction-based explanations for machine learning models. In *AAAI*, pages 1511–1519.

Izza, Y., Huang, X., Morgado, A., Planes, J., Ignatiev, A., and Marques-Silva, J. (2024). Distance-Restricted Explanations: Theoretical Underpinnings & Efficient Implementation. In *KR*, pages 475–486.

Izza, Y., Ignatiev, A., and Marques-Silva, J. (2022). On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75:261–321.

Izza, Y. and Marques-Silva, J. (2023). The pros and cons of adversarial robustness. *CoRR*, abs/2312.10911.

Izza, Y. and Marques-Silva, J. (2024). Efficient contrastive explanations on demand. *CoRR*.

Katz, G., Barrett, C. W., Dill, D. L., Julian, K., and Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. In *CAV*, pages 97–117.

Katz, G., Huang, D. A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljic, A., Dill, D. L., Kochenderfer, M. J., and Barrett, C. W. (2019). The marabou framework for verification and analysis of deep neural networks. In *CAV*, pages 443–452.

Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2019). Certified robustness to adversarial examples with differential privacy. In *IEEE S&P*, pages 656–672.

Leino, K., Wang, Z., and Fredrikson, M. (2021). Globally-robust neural networks. In *ICML*, pages 6212–6222.

Letoffe, O., Huang, X., and Marques-Silva, J. (2024). On correcting SHAP scores. In *AAAI*.

Liang, H., He, E., Zhao, Y., Jia, Z., and Li, H. (2022). Adversarial attack and defense: A survey. *Electronics*, 11(8):1283.

Liu, X., Han, X., Zhang, N., and Liu, Q. (2020). Certified monotonic neural networks. In *NeurIPS*.

Marques-Silva, J. (2023). Disproving XAI myths with formal methods - initial results. In *ICECCS*, pages 12–21.

Marques-Silva, J. (2024). Logic-based explainability: Past, present and future. In *ISoLA*, pages 181–204.

Marques-Silva, J. and Huang, X. (2024). Explainability is not a game. *Commun. ACM*, pages 66–75.

Marques-Silva, J. and Ignatiev, A. (2022). Delivering trustworthy AI through formal XAI. In *AAAI*, pages 12342–12350.

Marques-Silva, J. and Ignatiev, A. (2023). No silver bullet: interpretable ML models must be explained. *Frontiers in Artificial Intelligence*, 6:1128212.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Müller, M. N., Brix, C., Bak, S., Liu, C., and Johnson, T. T. (2022). The third international verification of neural networks competition (VNN-COMP 2022): Summary and results. *CoRR*, abs/2212.10376.

Narodytska, N. (2018). Formal analysis of deep binarized neural networks. In Lang, J., editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5692–5696. ijcai.org.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artif. Intell.*, 32(1):57–95.

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L. (2022). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Comput. Surv.*, 54(5):108:1–108:36.

Rosenfeld, E., Winston, E., Ravikumar, P., and Kolter, J. Z. (2020). Certified robustness to label-flipping attacks via randomized smoothing. In *ICML*, pages 8230–8241.

Ruan, W., Wu, M., Sun, Y., Huang, X., Kroening, D., and Kwiatkowska, M. (2019). Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *IJCAI*, pages 5944–5952.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85.

Seshia, S. A., Desai, A., Dreossi, T., Fremont, D. J., Ghosh, S., Kim, E., Shivakumar, S., Vazquez-Chanlatte, M., and Yue, X. (2018). Formal specification for deep neural networks. In *ATVA*, pages 20–34.

Seshia, S. A., Sadigh, D., and Sastry, S. S. (2022). Toward verified artificial intelligence. *Commun. ACM*, 65(7):46–55.

Shih, A., Choi, A., and Darwiche, A. (2018). A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, pages 5103–5111.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. T. (2018). Fast and effective robustness certification. In *NeurIPS*, pages 10825–10836.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). Intriguing properties of neural networks. In *ICLR*.

Vorácek, V. and Hein, M. (2023). Improving l1-certified robustness via randomized smoothing by leveraging box constraints. In *ICML*, pages 35198–35222.

Wang, Z., Huang, C., and Zhu, Q. (2022a). Efficient global robustness certification of neural networks via interleaving twin-network encoding. In *DATE*, pages 1087–1092.

Wang, Z., Wang, Y., Fu, F., Jiao, R., Huang, C., Li, W., and Zhu, Q. (2022b). A tool for neural network global robustness certification and training. *CoRR*, abs/2208.07289.

Weng, L., Chen, P., Nguyen, L. M., Squillante, M. S., Boopathy, A., Oseledets, I. V., and Daniel, L. (2019). PROVEN: verifying robustness of neural networks with a probabilistic approach. In *ICML*, pages 6727–6736.

Wiyatno, R. R., Xu, A., Dia, O., and de Berker, A. (2019). Adversarial examples in modern machine learning: A review. *CoRR*, abs/1911.05268.

Wu, M., Wu, H., and Barrett, C. W. (2023). Verix: Towards verified explainability of deep neural networks. In *NeurIPS*.

Yan, G., Romano, Y., and Weng, T. (2024). Provably robust conformal prediction with improved efficiency. In *ICLR*.

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. (2019). A fourier perspective on model robustness in computer vision. In *NeurIPS*, pages 13255–13265.

Yu, J., Ignatiev, A., Stuckey, P. J., Narodytska, N., and Marques-Silva, J. (2023). Eliminating the impossible, whatever remains must be true: On extracting and applying background knowledge in the context of formal explanations. In *AAAI*, pages 4123–4131.

Zhang, J. and Li, C. (2020). Adversarial examples: Opportunities and challenges. *IEEE Trans. Neural Networks Learn. Syst.*, 31(7):2578–2593.

Zhang, M., Levine, S., and Finn, C. (2022a). MEMO: test time robustness via adaptation and augmentation. In *NeurIPS*.

Zhang, X., Zheng, X., and Mao, W. (2022b). Adversarial perturbation defense on deep neural networks. *ACM Comput. Surv.*, 54(8):159:1–159:36.

Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., and Yu, P. S. (2023). Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Comput. Surv.*, 55(8):163:1–163:39.