Are Neural Networks Robust to Gradient-Based Adversaries Also More Explainable? Evidence from Counterfactuals

Rithik Appachi Senthilkumar[1]
Supervisors: Patrick Altmeyer[1]**, Dr. Cynthia Liem**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Rithik Appachi Senthilkumar
Final project course: CSE3000 Research Project
Thesis committee: Dr. Cynthia Liem, Patrick Altmeyer, Dr. Bernd Dudzik

Abstract

Adversarial Training has emerged as the most reliable technique to make neural networks robust to gradient-based adversarial perturbations on input data. Besides improving model robustness, preliminary evidence presents an interesting consequence of adversarial training – increased explainability of model behaviour. Prior work has explored the effects of adversarial training on gradient stability and interpretability, as well as visual explainability of counterfactuals. Our work presents the first quantitative, empirical analysis of the impact of model robustness on model explainability by comparing the plausibility of faithful counterfactuals for both robust and standard networks. We seek to determine whether robust networks learn more plausible decision boundaries and representations of the data than regular models, and whether the strength of the adversary used to train robust models affects their explainability. Our finidngs indicate that robust networks for image data learn more explainable decision boundaries and representations of data than regular models, with more robust models producing more plausible counterfactuals. Robust models for tabular data, however, only conclusively exhibit this phenomenon along decision boundaries and not for its overall data representations, possibly due to its high robustness-accuracy trade-off and the difficulties associated with traditional adversarial training due to its innate properties. We believe our work can help guide future research towards improving the robustness of machine learning models keeping their explainability in mind.

1 Introduction

Neural networks are vulnerable to *adversarial attacks*, small and often imperceptible perturbations to inputs engineered to elicit misclassifications by the network [1]. Such a vulnerability can lead to security threats [2, 3] when neural networks are deployed in critical scenarios, which necessitates developing defense mechanisms for neural networks. Adversarial training – augmenting the training set with adversarial examples – has been shown to improve the robustness of neural networks to adversarial attacks [4, 5].

Counterfactual Explanations (CEs) allow us to intuitively explain the decision-making of machine learning models by exploring how inputs to a model have to change for it to predict a different outcome [6]. A variety of desiderata guide the counterfactual search process and are used to measure the suitability of counterfactuals generated. Existing research has emphasized the *plausibility* [7] of counterfactuals, i.e. consistency with the underlying data distribution, and *faithfulness*, i.e. consistency with the model's *learned representation* of the data [8].

Adversarial training forces the network to generalize to both regular and perturbed points, implicitly smoothing the loss gradients, steering the network towards differentiating between classes based on robust features. Adversarially trained neural networks therefore produce more visually interpretable loss gradients compared to standard networks, since adversarial training restricts gradients closer to the data manifold [9]. Interestingly, counterfactual generators [7, 6] function similarly to gradient-based adversaries: strategically perturbing inputs by navigating along their loss gradients. It therefore intuitively follows that adversarial training on gradient-based attacks can *steer the counterfactual search along relevant and robust features, thereby making networks more explainable.* Augustin et al. [10] demonstrate this qualitatively, showing that counterfactuals produced from adversarially robust models exhibited class-specific features more evidently than standard models.

However, to the best of our knowledge, the systematic and quantitative exploration of how adversarial robustness of neural networks impacts their explainability via counterfac-

tuals is an under-explored research direction. We perform a series of experiments on both the *MNIST* [11] and *California Housing* [12] datasets to investigate this link, generating faithful counterfactuals from both robust and regular neural networks. Using the plausibility of faithful counterfactuals generated as the explainability measure, we seek to determine **whether robust neural networks are more explainable than standard networks**, and **among robust models, whether the strength of adversary used during training affects their explainabilities**. We observed that:

- Faithful counterfactuals for robust neural networks generated along the decision boundary are more plausible than those generated for standard neural networks for both datasets, demonstrating that **robust networks differentiate between classes more plausibly than regular models**.
- Robust *MNIST* models generate starkly more plausible counterfactuals beyond the decision boundary than standard models, showing that **robust networks learn more plausible representations of classes than regular networks**. However, a similar trend is not observed for the tabular *California Housing* dataset, likely because **for tabular data, adversarial training moves the region of maximum likelihood away from the data manifold to accommodate for adversarial examples**, further evidenced by its high robustness-accuracy tradeoff.
- Among robust *MNIST* networks, **those trained on stronger adversaries produced more plausible counterfactuals both along the decision boundaries and beyond**, indicating a positive relation between model robustness and model explainability.

We hope our work can help encourage future research on developing robust neural networks that consider both the desirable objectives of robustness against gradient-based adversarial perturbations, and explainability of model decisions.

Our paper is organized as follows: Section 2 provides the relevant background on adversarial machine learning and counterfactual explanations, Section 3 details our research objectives and methodology, followed by a thorough description of our experimental setup in Section 4. Section 5 describes our findings, assessments and key takeaways, and Section 6 highlights the steps we took to ensure our research was carried out *responsibly*. A transparent discussion of our limitations and our suggestions for future work can be found in Section 7, and Section 8 concludes our paper.

2 Background

In this section, we present an overview of relevant research, detailing prior work in adversarial machine learning and the concept of robustness in background:adversarialrobustness, followed by the counterfactual approach to model explainability in background:counterfactual.

2.1 Adversarial Machine Learning

Given the vulnerability of neural networks to adversarial attacks [1], an illustration of which can be found in Figure 1, a large body of work has been dedicated towards developing mechanisms to *defend* neural networks against adversarial examples [4, 5, 13], and to evaluate the *adversarial robustness* of neural networks [14, 15].

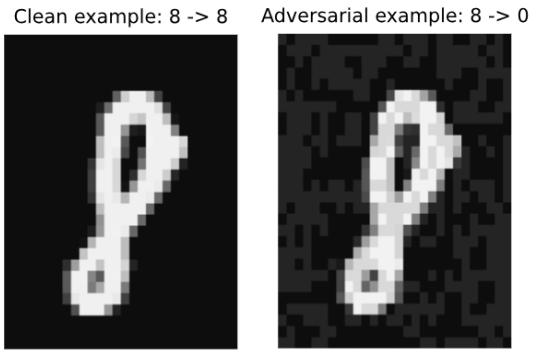

Figure 1: Left: A clean *MNIST* [11] datapoint correctly classified by a neural network as an '8'. Right: A Fast Gradient Sign Method (FGSM) attack [4] applied to it, causing the network to misclassify it as a '0'

Adversarial training, augmentation of the training data with adversarial examples, has emerged as the most reliable and empirically validated defence mechanism for neural networks against adversarial attacks. Given a network f_θ parameterized by θ, a dataset (x_i, y_i), a loss function ℓ, and a threat model Δ representing the type of ℓ-norm to bound the adversarial perturbations, adversarial training is typically formulated as the following robust optimization problem:

$$\min_{\theta} \sum_{i} \max_{\delta \in \Delta} \ell(f_\theta(x_i + \delta), y_i) \tag{1}$$

where the threat model is expressed as $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$ for some $\epsilon > 0$ and p representing the norm.

Prior work has explored the usage of single-step adversaries such as the Fast Gradient Sign Method (FGSM) [4] to generate adversarial examples during training, which computes an adversarial example based on the sign of the loss gradient with respect to the input:

$$x + \epsilon \, sgn(\nabla_x \mathcal{L}(\theta, x, y)) \tag{2}$$

On the other hand, Madry et al. [5] introduced the multi-step Projected Gradient Descent (PGD) attack:

$$x_{t+1} = \Pi_{x+S} \left(x_t + \alpha \, sgn(\nabla_x \mathcal{L}(\theta, x, y)) \right) \tag{3}$$

repeated across steps to determine the final adversarial example. The PGD attack is essentially FGSM carried out across multiple iterations with a smaller step size α.

Madry et al. [5] empirically observed that the variability of the loss landscape for adversarially trained models trained with the PGD attack was significantly lower than that of standard models. This smoothening of the loss landscape implies the loss gradients of robust models are more stable, uniform and consistent than in standard models, leading to predictable and stable model behavior despite slight perturbations to inputs.

Evaluation of a model's adversarial robustness entails measuring the degree to which it resists misclassifying adversarial examples. Robustness can be measured by performing

heuristic evaluations of model performance on adversarial examples [5, 15], or by identifying and formally proving robustness up to a lower bound of input perturbation for a specific set of examples [16]. Carlini et al. provide a comprehensive set of principles that should guide defense evaluations, and a checklist to avoid common evaluation pitfalls [14].

2.2 Counterfactual Explanations

Counterfactual explanations (CEs) for classification models explore how the inputs to the model have to change, for the model to classify it differently. Modifying *factual* input instances intohypothetical *counterfactuals* allows us to better understand how the underlying model makes its predictions, by observing the feature modifications necessary to elicit a different prediction by the model [17]. Counterfactuals offer the possibility of algorithmic recourse [18], empowering people to carry out changes to alter unfavorable outcomes delivered by automated decision-making systems, such as those involved with loan approvals and hiring.

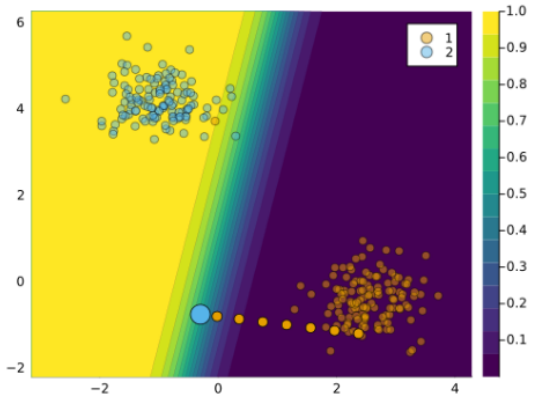

Figure 2: Generating a counterfactual of class '2' from a factual of class '1' for a conventional binary classifier, using a generic gradient-based counterfactual generator, from [19]

Most counterfactual generators perform gradient descent with respect to the original input to navigate it from the original class to the target class by minimising a loss function, modifying the input at every iteration. Figure 2 demonstrates a gradient-based counterfactual generator in action. While this describes the simplest gradient-based generator, researchers have proposed variations to the counterfactual search objective to guide the counterfactual search to satisfy certain desiderata. Wachter et al. [6], for instance, added a distance-based penalty term to the search objective as a regularizer, emphasizing *closeness* to the original factual as a desideratum. Other desiderata explored in research include sparsity [20] and plausibility [7].

Plausibility describes how well a counterfactual adheres to the underlying distribution of the data. While the proximity of plausible counterfactuals to target class data may incentivise a *plausible-first approach* to counterfactual search, Altmeyer et al. [8] argue that doing so can produce counterfactuals that seem plausible but need not describe the opaque model's behavior faithfully. They introduce the notion of *faithfulness* as a desideratum, which describes how well the counterfactual adheres to the *model's learned representation* of the data. They developed the *Energy-Constrained Conformal Counterfactuals (ECCCo)*

generator which leverages advances in energy-based modelling to produce faithful counterfactuals. Its search objective is as follows:

$$\min_{Z' \in Z^L} \left\{ \ell(f(Z'); M_\theta, y^+) + \lambda_1 \cdot cost(f(Z')) + \lambda_2 \cdot E_\theta(f(Z') \mid y^+) + \lambda_3 \cdot \Omega(C_\theta(f(Z'); \alpha)) \right\} \quad (4)$$

Here, ℓ represents any standard classification loss, $f(\cdot)$ maps from the counterfactual state space to the feature space, y^+ represents the target class, M_θ the underlying model parametrized by θ and Z' denotes a counterfactual from an L-dimensional array of counterfactual states. The penalty terms involving λ_1, λ_2 and λ_3 induce closeness [6] by constraining distance from the factual, faithfulness by constraining the counterfactual's energy [21], and low predictive uncertainty via conformal predictions [22] respectively.

3 The Effect of Robustness on Model Explainability

Prior research has explored the effect of adversarial training on the interpretability of loss gradients [9] and the plausibility of counterfactual explanations generated [10]. However, these works investigated the aforementioned effects only for image classification tasks, and the latter work solely relied on a qualitative analysis of counterfactuals by visually inspecting them to identify class-specific features.

Through this paper, we seek to expand upon past research to answer the following research questions:

> ***Research Question 1:*** *Are neural networks made robust to gradient-based adversaries through adversarial training more explainable than a regularly trained neural network?*

> ***Research Question 2:*** *Among adversarially robust neural networks, are those trained with a stronger gradient-based adversary more explainable than those trained with a weaker gradient-based adversary?*

Our work, to the best of our knowledge, is the first attempt at quantitatively determining the effect of adversarial robustness on model explainability via counterfactuals. We experiment across both tabular and image data, training models with varying strengths of adversaries to systematically and comprehensively determine the robustness-explainability connection. Moreover, we leverage a counterfactual generator [8] that produces counterfactuals in line with the underlying model's learned representation of the data, ensuring that they faithfully describe model behavior.

The following subsections will further describe our methods of adversarial training and robustness evaluation, as well as our notion of model explainability and our counterfactual generation process.

3.1 Adversarial Training

Adversarial training can be formulated as the robust min-max optimization problem defined in Equation 1, with adversarial attacks approximating the inner maximization problem. Gradient-based adversarial attacks can either take a single step along the loss gradient, such as the Fast Gradient Sign Method (FGSM) [4] described in Equation 2 that perturbs the input based on the sign of the loss gradient w.r.t the input, or multiple steps like

the Projected Gradient Descent (PGD) attack [5], where multiple smaller FGSM steps are carried out iteratively.

While adversarial training based on FGSM or its variations provides robustness against single-step adversaries, they were shown to be vulnerable to more sophisticated multi-step attacks [23]. We choose to carry out PGD-based adversarial training as it has been shown to learn models resistant to both single- and multi-step adversaries [5]. Moreover, we limit our scope to the ℓ_∞ **threat model**, which means the ℓ_∞-norm (Chebyshev distance) is used as the distance measure to bound the perturbations.

To answer our second research question, we configure three PGD adversaries with increasing step sizes, number of iterations and perturbation bounds. The adversary with the highest measure for these three attributes is our strongest and leads to the best robustness, while the opposite is true for the adversary with the lowest measures. The standard (non-robust) model is subjected to the regular training process.

To quantitatively determine the robustness of our neural networks, we measure the accuracy score on adversarial examples generated with FGSM [4] and PGD [5] to cover both single- and multi-step robustness. An alternate strategy to measure robustness of a model is neural network verification, in which robustness within a certain bound of perturbations for a set of examples is formally proven [24, 25]. However, formal verification is computationally intractable for even modestly sized neural networks and only provides guarantees under a narrow subset of examples, hence we do not consider it.

3.2 Measuring Model Explainability

Once we train our standard and robust models, the next step involves generating many *faithful* counterfactual explanations for each model, measuring their *plausibility* to assess model explainability.

We use an adapted version of the *Energy-Constrained Conformal Counterfactuals (ECCCo)* generator [8] whose counterfactual search objective is described in Equation 4. We modify it by removing the conformal set penalty (the λ_3 penalty in the equation). Our generator (hereinafter referred to as *ECCo*) still retains its closeness [6] (λ_1) and faithfulness [8] (λ_2) inducing constraints. The faithfulness constraint is especially important in our case. Faithful counterfactuals are those in line with the model's learned representation of the underlying data [8]. Modelling faithfulness as a counterfactual search objective ensures that those we obtain truthfully represent the underlying model's behavior and can be assessed to ascertain the model's properties.

In this work, we define the *explainability* of a model as the degree to which faithful counterfactuals generated for the model are also plausible. The plausibility of a counterfactual refers to its consistency with the underlying (target class) data [7, 8]. To quantitatively determine the extent of plausibility, we use the implausibility metric initially proposed by Guidotti [26] and later adapted by Altmeyer et al. [8]:

$$impl\left(\mathbf{x}', \mathbf{X}_{\mathbf{y}^+}\right) = \frac{1}{\left|\mathbf{X}_{\mathbf{y}^+}\right|} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbf{y}^+}} dist\left(\mathbf{x}', \mathbf{x}\right) \tag{5}$$

It measures the average distance between the counterfactual $\mathbf{x}'$ and its nearest neighbors $\mathbf{X}_{\mathbf{y}^+}$ in the target class $\mathbf{y}^+$. The further a point is from its nearest neighbors in the target class, the less consistent it is with the underlying data, therefore it is more implausible.

3.3 Effect of Adversarial Training on Gradients and Counterfactual Search

Madry et al. [5], who framed the robust optimization approach to adversarial training (Equation 1), performed an elaborate empirical comparison of the loss landscape of adversarially robust models with those of standard, non-robust models. Their key takeaway was that the inner maximization problem of the adversarial training objective was tractable, with the loss values increasing *consistently* with an increase in adversary strength (iterations). This indicates that adversarial training stabilizes the gradient landscape, making the underlying model less susceptible to perturbations along random, irrelevant directions. Kim et al. [9] further demonstrated that adversarial training leads to more interpretable loss gradients than standard training both qualitatively and quantitatively.

Most counterfactual generators (including *ECCo*) are gradient-based, which means they rely on variations of gradient descent to modify a factual input to a counterfactual belonging to a different class. Since adversarial training stabilizes the gradient landscape of models, one can intutively reason that it guides the gradient descent process towards meaningful modifications, i.e. along more *robust* features, rather than along irrelevant directions. When the counterfactual search is steered towards more relevant features, it naturally follows that we obtain counterfactuals that are more *plausible* representations of their intended target classes.

This overall line of reasoning leads us to believe that adversarial training has a positive effect on model explainability, and that explainability is directly proportional to robustness. The observations of Augustin et al. [10] demonstrate this phenomenon qualitatively, observing that counterfactuals generated for robust models consisted of class-specific features to a greater extent than regular models. Our goal with this paper is to subject this hypothesis to a rigorous and systematic analysis, quantitatively exploring the robustness-explainability link.

4 Experimental Setup

This section of the paper will comprehensively describe our experimental process, elaborating on the datasets and models used, configurations of adversaries for training and robustness evaluations, and the counterfactual search and evaluation process. Finally, we describe the software we utilized to carry out our experiments. The source code for our experiment can be found online [1]

4.1 Datasets

We experiment with *MNIST* [11] as our vision dataset, and *California Housing* [12] as our tabular dataset. In addition to the quantitative analysis we conduct for both datasets, using MNIST allows us to qualitatively assess the counterfactual examples generated for both the standard and adversarially trained models, through visual inspection of class-specific features.

[1] https://github.com/JuliaTrustworthyAI/what-makes-models-explainable/tree/rithik-adv-ml

4.2 Models and Adversaries

For both our datasets, we initialize and train multi-layer perceptrons (MLPs). While convolutional neural networks (CNNs) like LeNet-5 [11] exhibit better performance on *MNIST* than regular MLPs, adversarially training them was computationally prohibitive since generating a single Projected Gradient Descent (PGD) adversarial attack necessitates multiple gradient computations [5].

For *MNIST*, our neural networks consisted of a 784 neuron input layer, a 128 neuron hidden layer and 10 output neurons. For *California Housing*, our networks consisted of an 8 neuron input layer, a 10 neuron hidden layer and 2 output neurons. Rectified Linear Unit (ReLU) [27] was used as the activation function for the hidden layer in all modes. The models were trained using the cross-entropy loss, and Xavier initialization (using a normal distribution) [28] was used to initialize weights.

We use three different PGD [5] adversaries varying in strength to train our robust models:

- **Strong Adversary:** PGD with $\epsilon = 0.3$, 40 iterations and 0.01 step-size.
- **Medium Adversary:** PGD with $\epsilon = 0.1$, 13 iterations and 0.01 step-size.
- **Weak Adversary:** PGD with $\epsilon = 0.05$, 7 iterations and 0.01 step-size.

This altogether gives us four neural networks for each dataset: one regularly trained and three adversarially trained.

We measure the robustness of models as accuracy on an unseen test set perturbed by the Fast Gradient Sign Method (FGSM) adversary [4], with $\epsilon = 0.2$, and a PGD adversary with $\epsilon = 0.2$, 26 iterations and 0.01 step size.

4.3 Counterfactual Search and Analysis

Our counterfactual search leverages the *ECCo* generator [8] with a distance penalty (λ_1) and energy constraint (λ_2) of 0.01 and 0.5 respectively for *MNIST*, and 0.1 and 0.5 respectively for *California Housing*. Across both datasets, we prioritized the energy constraint over closeness to encourage generating counterfactuals that faithfully represent the underlying model's behavior. The lower closeness penalty for *MNIST* is based on the understanding that unlike for tabular data, minor distortions in images do not significantly affect interpretability, allowing for a broader range of counterfactuals that retain the digit's essence.

An experimental run consists of randomly selecting 100 factual datapoints from the unseen test set and generating one counterfactual for each factual with pre-determined target classes kept the same across all models for each dataset. We then measure their implausibilities as per Equation 5 with 100 nearest neighbors, relying on Euclidean distance for tabular data and the Structural Similarity Index (SSIM) [29] for image data. Finally, we average their implausibilities. This run is repeated five times, each with a different selection of factuals. We report the mean and standard deviation of implausibilities across the five experimental runs.

To add another dimension to our analysis, we generate *two distinct types of counterfactuals*: those along the decision boundaries and those in the underlying model's region of maximum likelihood for the classes. The former kind is generated by setting a *Decision Threshold Convergence* criterion, concluding search once the counterfactual is predicted to belong to the target class with probability over 0.5. The latter kind is generated by setting a *Generator Conditions Convergence* criterion, concluding search once the loss of the gradient

Table 1: Clean and robust accuracies, and average implausibilities ± one standard deviation for both decision-boundary counterfactuals (DTC - Decision Threshold Convergence) and maximum likelihood counterfactuals (GCC - Generative Conditions Convergence) for standard and robust models trained on *MNIST* [11] and *California Housing* [12] datasets. Lowest average implausibility for each dataset and type of counterfactual marked in bold.

Dataset	Training	Accuracies			Model Implausibilities	
		Clean	FGSM	PGD	Impl. (DTC)	Impl. (GCC)
MNIST	Standard	0.981	0.032	0.002	0.437 ± 0.002	0.390 ± 0.004
	Strong-AT	0.969	0.714	0.653	0.412 ± 0.002	**0.221 ± 0.005**
	Medium-AT	0.984	0.379	0.298	**0.411 ± 0.002**	0.236 ± 0.005
	Weak-AT	0.983	0.248	0.062	0.416 ± 0.003	0.262 ± 0.006
California Housing	Standard	0.857	0.211	0.217	1.673 ± 0.072	**2.358 ± 0.124**
	Strong-AT	0.771	0.644	0.647	**1.196 ± 0.072**	2.762 ± 0.149
	Medium-AT	0.810	0.563	0.572	1.224 ± 0.070	2.884 ± 0.188
	Weak-AT	0.840	0.337	0.350	1.390 ± 0.074	2.632 ± 0.122

w.r.t the counterfactual (for the target class), reduces below a threshold of 0.01 or undergoes a maximum of 1000 descent steps. Generating decision boundary counterfactuals can help us determine **whether the model learned plausible ways to differentiate between classes**, while generating gradient minimisation counterfactuals helps us determine **whether the model learned plausible representations of the classes themselves**.

4.4 Software

Our experiments are carried out in Julia, with our datasets obtained from the TaijaData.jl[2] package, neural networks initialized and trained using the Flux.jl[3] package and counterfactuals generated via the CounterfactualExplanations.jl[4] package, all of which are open source.

5 Results and Discussion

In this section, we present and discuss the results of our experiments, outlining how they answer the research questions we investigated and providing explanations. Our aim was to explore whether robustness to gradient-based adversaries makes neural network decisions more explainable, and whether the strength of network robustness plays a role. Table 1 demonstrates the quantitative results we obtained, consisting of both robustness and explainability evaluations of regular and adversarially trained models. The robust accuracies indicate that adversarial training worked as intended, with the robustness of networks directly proportional to the strength of adversary they were trained on.

[2]https://github.com/JuliaTrustworthyAI/TaijaData.jl/tree/main
[3]https://fluxml.ai/Flux.jl/stable/
[4]https://juliatrustworthyai.github.io/CounterfactualExplanations.jl/stable/

We generated counterfactuals both along the model's learned inter-class decision boundaries and per-class maximum likelihood regions to assess how robust training impacts the network's ability to (1) differentiate between classes, and (2) learn plausible representations of the classes.

5.1 Explainability of Decision Boundaries

To generate counterfactuals in the vicinity of the inter-class decision boundaries, we defined a *Decision Threshold* convergence criteria for the search. It returns the first counterfactual it generates that the underlying network predicts as being in the target class with a confidence of at least 0.5, which represents a certain crossing of the decision boundary.

The results demonstrate that across both *MNIST* and *California Housing* datasets, robust neural networks generated decision boundary counterfactuals closer to its nearest neighbors, i.e. more *plausible*, than the regularly trained model. We observe that the difference in decision-boundary explainability between robust models is not significant (as indicated by their standard deviations), but all three robust models clearly generated more plausible counterfactuals than the regularly trained model. This shows that even weak adversarial training on neural networks can enable it to differentiate between classes more robustly than regular training, but there are diminishing returns with an increase in training adversary strength.

Our findings are corroborated by the observations of Ilyas et al. [30], who argue that models trained in the standard manner tend to establish decision boundaries by learning non-robust, non-interpretable features, while adversarial training causes networks to leverage robust, human-interpretable features to differentiate between classes. Our empirical assessment using decision boundary counterfactuals demonstrates exactly this point.

5.2 Explainability of Class Representations

Faithful counterfactuals generated beyond the decision boundary give us insights into the model's learned representation of the data distribution. By specifying the *Generator Conditions* convergence criteria based on minimizing the gradients below a threshold of 0.01 (or stopping after 1000 iterations), we produced counterfactuals within or close to the network's learned representation of the target class. Unlike decision boundary counterfactuals, here we observe stark differences in explainability trends between *MNIST* and *California Housing* networks.

MNIST results demonstrate that adversarially trained networks produced counterfactuals significantly more plausible than the standard network. This is an encouraging result, indicating that adversarial training causes models to learn more plausible representations of the classes. Figure 3 demonstrates this observation for a random factual from the test set. Even among robust networks, we observe that as we increase the strength of adversary during training, the counterfactuals they produce tend to be closer to its nearest neighbors in the target class, implying that more robust networks are also more explainable.

For neural networks trained on the *California Housing* dataset, however, we do not notice the same results. In fact, we observe that the standard model generates more plausible counterfactuals for the opposite class to its factual than any of the robust models, and among robust models there does not seem to be a noticeable proportionality between robustness and explainability.

A potential explanation for this discrepancy is that adversarial training with the *California Housing* data significantly moved the region of maximum likelihood further away from

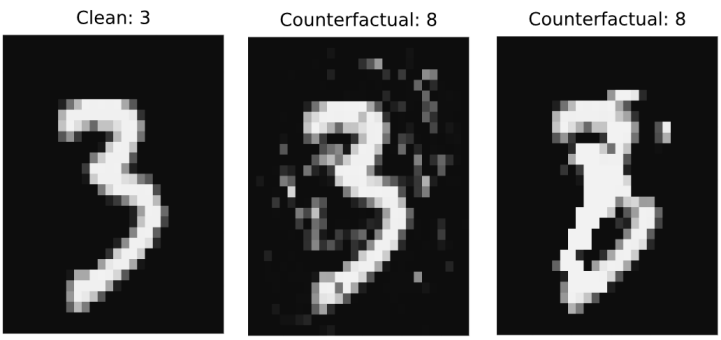

Figure 3: A factual datapoint of class '3' (left) and counterfactual '8's produced by the standard model (middle) and most robust model (right), using the *Generator Conditions* convergence criterion. The robust model produced a counterfactual visually far more plausible than the standard model.

the data manifold to satisfy the demands of adversarial robustness. We can also observe this in the trade-off between clean and robust accuracies in *California Housing* models, which was significantly more drastic than in *MNIST* models.

More importantly, these results also underscore **drawbacks with traditional deep-learning techniques and adversarial training with tabular data** compared to image data, due to the innate characteristics of both these data types and the differences in their structures. Tabular data consists of heterogenous features – dense numerical data and sparse categorical features, and the correlations between individual features in tabular data are generally not as well-defined as in image data with its spatial relationships. Effectively dealing with tabular data remains a challenge to deep learning models [31, 32]. While a small perturbation does not affect an image's essence and human-interpretability, the same cannot be said as easily for tabular data, for which a perturbation can produce a datapoint that can be interpreted completely differently. Therefore, the benefits conferred onto image classification neural networks by adversarial training [5, 4], which aims to minimize susceptibility to perturbations, need not translate into similar effects in tabular data.

6 Responsible Research

Given the focus of our research on the robustness and explainability of machine learning models, domains which have a variety of ethical considerations [18, 2, 3],we made concerted efforts to ensure our research adheres to accessibility and reproducibility standards.

With regards to accessibility, our data is obtained from the open-source TaijaData.jl package which is under the MIT free license and properly anonymized. The Julia packages we use to instantiate and train neural networks (Flux.jl), and generate counterfactuals (CounterfactualExplanations.jl) are both open-source, and so is the language of Julia which we use to conduct our experiments. The source code of our experiments can also be found online under the JuliaTrustworthyAI organization on GitHub.

Concerning reproducibility, we take special care to perform multiple runs for each experiment across multiple initializations of models, finally averaging results to ensure mitigation of biases. We make sure to report the hyperparameter values and exact training and ex-

periment procedures to ensure reproducibility. Our work adheres to the standards of transparency associated with experimental and dataset details as specified in the Netherlands Code of Conduct for Research Integrity [33].

7 Limitations and Future Work

Despite considerable efforts to ensure the reliability of our methodology and research, we can identify certain limitations.

Firstly, our notion of *adversarial robustness* for neural networks was limited to robustness towards adversaries exploiting access to the network's gradients to craft perturbations [5, 4]. These adversaries are *white-box* in nature, as the attack algorithm assumes access to model parameters. However, there exist *black-box* adversaries [34, 35], which learn about the model by repeatedly querying it. Moreover, robustness evaluation metrics explored in research have relied on *ensemble* adversarial attacks, where multiple attacks are attempted for each datapoint to prevent an overestimation of adversarial robustness. A popular robustness metric that leverages this technique is RobustBench [15], which uses AutoAttack [36], an ensemble of four attacks consisting of both white- and black-box attacks. We hope future work exploring the robustness-explainability connection considers both black-box attacks and attack ensembling techniques to provide a more realistic and accurate robustness estimate avoiding overestimation.

For both our datasets, we explored the robustness-explainability connection for one neural network architecture each. Moreover, for the *MNIST* image dataset, we relied on a regular multi-layered perceptron architecture, although convolutional neural networks (CNNs) exhibit state-of-the-art performance on image data. An interesting direction for future research can be measuring the impact of robustness on explainability across a richer spectrum of models to paint a complete picture.

In terms of our counterfactual generation process, our notion of *plausibility* was based on the distance between a counterfactual and its nearest neighbors in the target class. Although it is intuitive to use distance-based metrics for image data, it may not be applicable to tabular data, especially due to the heterogeneity associated with its features. Identifying universally applicable plausibility measures can help future work.

Finally, we identified limitations associated with traditional deep learning techniques and adversarial training with tabular data. Due to the heterogeneity of features and potentially weak correlations between features in tabular data, learning shallow neural networks to tabular data representations is a greater challenge than with image or speech data, for which spatial or semantic similarities can be exploited. Unlike for image data, slight perturbations can severely impact the interpretability of tabular datapoints. An interesting direction for future research to address these concerns is by leveraging tree-based models such as Deep Neural Decision Trees [37] that combines the benefits of neural networks and decision trees and are known to perform better on tabular data than regular shallow neural networks.

8 Conclusions

In this work, we investigate whether neural networks robust to gradient-based adversaries learn more plausible representations of data than standard networks. Training neural networks on adversarial attacks has been shown to stabilize gradients [5], causing networks to learn more robust features [30]. We perform the first quantitative study across both

image and tabular data, empirically determining whether faithful counterfactuals produced for robust networks are more plausible than those for standard networks, and whether the strength of adversary used to train robust networks impacts the plausibility of their faithful counterfactuals. Our results indicate robust models for image data are more explainable than regular models, learning more plausible inter-class decision boundaries and class representations. For tabular data, robustness did not lead to better explainability of class representations, possibly due to its high robustness-accuracy trade-off and innate properties which make it difficult to apply traditional adversarial training and deep learning principles on it. We believe our findings can encourage future work on improving model robustness in ways that promote model explainability.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 2154–2156.

[3] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Simple physical adversarial examples against end-to-end autonomous driving models," in *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*. IEEE, 2019, pp. 1–7.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[6] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[7] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[8] P. Altmeyer, M. Farmanbar, A. van Deursen, and C. C. Liem, "Faithful model explanations through energy-constrained conformal counterfactuals," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 10 829–10 837.

[9] B. Kim, J. Seo, and T. Jeon, "Bridging adversarial robustness and gradient interpretability," *arXiv preprint arXiv:1903.11626*, 2019.

[10] M. Augustin, A. Meinke, and M. Hein, "Adversarial robustness on in-and out-distribution improves explainability," in *European Conference on Computer Vision*. Springer, 2020, pp. 228–245.

[11] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[12] R. K. Pace and R. Barry, "Sparse spatial autoregressions," *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997.

[13] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *Advances in neural information processing systems*, vol. 32, 2019.

[14] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[15] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," *arXiv preprint arXiv:2010.09670*, 2020.

[16] M. H. Meng, G. Bai, S. G. Teo, Z. Hou, Y. Xiao, Y. Lin, and J. S. Dong, "Adversarial robustness of deep neural networks: A survey from a formal verification perspective," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2022.

[17] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[18] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects," *arXiv preprint arXiv:2010.04050*, 2020.

[19] P. Altmeyer, A. van Deursen, and C. Liem, "Explaining black-box models through counterfactuals," *arXiv preprint arXiv:2308.07198*, 2023.

[20] L. Schut, O. Key, R. Mc Grath, L. Costabello, B. Sacaleanu, Y. Gal *et al.*, "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1756–1764.

[21] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 2011, pp. 681–688.

[22] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[23] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[24] L. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, L. Daniel, D. Boning, and I. Dhillon, "Towards fast computation of certified robustness for relu networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5276–5285.

[25] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*. Springer, 2017, pp. 97–117.

[26] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[27] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402 Vol.2, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:60600316

[30] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf

[31] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[32] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[33] KNAW, NFU, NWO, TO2-federatie, Vereniging Hogescholen, and VSNU, "Nederlandse gedragscode wetenschappelijke integriteit," 2018. [Online]. Available: https://doi.org/10.17026/dans-2cj-nvwu

[34] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[35] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248*, 2017.

[36] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.

[37] Y. Yang, I. G. Morillo, and T. M. Hospedales, "Deep neural decision trees," *arXiv preprint arXiv:1806.06988*, 2018.