Towards Adaptive Attacks on Constrained Tabular Machine Learning

Thibault Simonetto [1] **Salah Ghamizi** [2] **Maxime Cordy** [1]

Abstract

State-of-the-art deep learning models for tabular data have recently achieved acceptable performance to be deployed in industrial settings. Contrary to computer vision, there is to date no efficient constrained whitebox attack to evaluate the adversarial robustness of deep tabular models due to intrinsic properties of tabular data such as categorical features, immutability, and feature relationship constraints. To fill this gap, we propose CAPGD, the first efficient evasion attack for constrained tabular deep learning models. CAPGD is an iterative parameter-free attack to generate adversarial examples under constraints. We evaluate CAPGD across four critical use cases: credit scoring, phishing, botnet attacks, and ICU survival prediction. Our empirical study covers 5 modern tabular deep learning architectures and demonstrates the effectiveness of our attack which improves over the most effective constrained attack by 81% points.

1. Introduction

Evasion attack is the process of slightly altering an original input into an *adversarial example* designed to force a machine learning (ML) model to output a wrong decision. Robustness to adversarial examples is a problem of growing concern among the secure ML community, with over 10,000 publications on the subject since 2014 (Carlini et al., 2019). Recent studies also report real-world occurrences of evasion attacks, which demonstrate the importance of studying and defending against this phenomenon (Grosse et al., 2024).

While research has studied the robustness of deep learning models in Computer Vision (CV) and Natural Language Processing (NLP) tasks, many real-world applications instead deal with tabular data, including in critical fields like finance, energy, and healthcare. If classical "shallow" models (e.g. random forests) have been the go-to solution to learn from tabular data (Hancock & Khoshgoftaar, 2020), deep learning models are becoming competitive (Borisov et al., 2022). This raises anew the need to study the robustness of these models.

However, robustness assessment for tabular deep learning models brings a number of new challenges that previous solutions — because they were originally designed for CV or NLP tasks — do not consider. One such challenge is the fact that tabular data exhibit complex relationships and constraints across features. The satisfaction of these feature constraints can be a non-convex or even non-differentiable problem; this implies that established evasion attack algorithms relying on gradient computation do not create valid adversarial examples (i.e., constraint satisfying) (Ghamizi et al., 2020). Meanwhile, attacks designed for tabular data also ignore feature type constraints (Ballet et al., 2019) or, in the best case, consider categorical features without feature relationships (Wang et al., 2020; Xu et al., 2023; Bao et al., 2023) and are evaluated on datasets that exclusively contain such features. This restricts their application to other domains that present heterogeneous feature types.

The only published evasion attacks that support feature constraints are *Constrained Projected Gradient Descent* (CPGD) and *Multi-Objective Evolutionary Adversarial Attack* (MOEVA) (Simonetto et al., 2021). CPGD is an extension of the classical gradient-based PGD attack with a new loss function that encodes how far the generated examples are from satisfying the constraints. Although theoretically elegant and practically efficient, this attack suffers from a low success rate due to its difficulty to converge toward both model classification and constraint satisfaction (Simonetto et al., 2021). Conversely, MOEVA is based on genetic algorithms. It offers an outstanding success rate compared to CPGD and works on shallow and deep learning models. However, it is computationally expensive and requires numerous hyper-parameters to be tuned (population size, mutation rate, generations, etc.). This prevents this attack from scaling to larger models and datasets.

Overall, research on adversarial robustness for tabular machine learning in general (and tabular deep learning in particular) is still in its infancy. This is in stark contrast to the

[1] University of Luxembourg, Luxembourg, Luxembourg [2] LIST/RIKEN AIP, Esch-sur-Alzette, Luxembourg. Correspondence to: Thibault Simonetto <thibault.simonetto@uni.lu>, Salah Ghamizi <salah.ghamizi@gmail.com>, Maxime Cordy <maxime.cordy@uni.lu>.

abundant literature on adversarial robustness in CV (Long et al., 2022) and NLP tasks (Dyrmishi et al., 2023). Given this limited state of knowledge, the **objective** of this paper is to propose novel and effective attack methods for tabular models subject to feature constraints.

Our first hypothesis is that gradient-based algorithms have been insufficiently explored in (Simonetto et al., 2021) and that the introduction of dedicated adaptive mechanisms can outperform CPGD. To verify this, we design a new adaptive attack, named *Constrained Adaptive PGD* (CAPGD), whose only free parameter is the number of iterations and that does not require additional parameter tuning (Section 4). We demonstrate that the different mechanisms we introduced in CAGPD contribute to improving the success rate of this attack compared to CPGD, by 81% points. Across all our datasets, the set of adversarial examples that CAPGD generates subsumes all of the examples generated by any other gradient-based method.

Our contributions can be summarized as follows:

1. We design a new parameter-free attack, CAPGD that introduces momentum and adaptive steps to effectively evade DL models while enforcing the domain constraints.
2. We evaluate CAPGD in a large-scale evaluation over four datasets, five architectures. Our results show that CAPGD outperforms the other gradient-based attacks in terms of capability to generate valid (constraint-satisfying) adversarial examples. CAPGD improves over the most effective constrained attack by up to 81% points.

2. Related work

2.1. Tabular Deep Learning

Tabular data remains the most commonly used form of data (Shwartz-Ziv & Armon, 2021), especially in critical applications such as medical diagnosis (Ulmer et al., 2020; Somani et al., 2021), financial applications (Ghamizi et al., 2020; Clements et al., 2020; Cartella et al., 2021), user recommendation systems (Zhang et al., 2019), cybersecurity (Chernikova & Oprea, 2019; Aghakhani et al., 2020), and more. Improving the performance and robustness of tabular machine learning models for these applications is becoming critical as more ML-based solutions are cleared to be deployed in critical settings.

Borisov et al. (2022) showed that traditional deep neural networks tend to yield less favorable results in handling tabular data when compared to more shallow machine learning methods, such as XGBoost. However, recent approaches like RLN (Shavitt & Segal, 2018) and TabNet (Arik & Pfister, 2021) are catching up and even outperforming shallow models in some settings. We argue that DNNs for Tabular Data are sufficiently mature and competitive with shallow models and require therefore a thorough investigation of their safety and robustness. Our work is the first exhaustive study of these critical properties.

2.2. Realistic Adversarial Examples

Initially applied to computer vision, adversarial examples have also been adapted and evaluated on tabular data. Ballet et al. (2019) considered feature importance to craft the attacks, Mathov et al. (2022) considered mutability, type, boundary, and data distribution constraints, Kireev et al. (2022) suggested considering both the cost and benefit of perturbing each feature, and Simonetto et al. (2021) introduced domain-constraints (relations between features) as a critical element of the attack. This last approach is closest to the trend in adversarial machine learning in critical scenarios such as malware and finance (Pierazzi et al., 2020; Dyrmishi et al., 2022).

Our work follows this last hypothesis and focuses on constrained feature-space attacks to realistically assess the robustness of deep tabular learning models.

While domain constraints satisfaction is essential for successful attacks, research on robustness for industrial settings (eg Ghamizi et al. (2020) with a major bank) also demonstrated that imperceptibility remains important for critical systems with human-in-the-loop mechanisms, which could deflect attacks with manual checks from human operators. Imperceptibility is domain-specific, and multiple approaches have been suggested (Ballet et al., 2019; Kireev et al., 2022; Dyrmishi et al., 2022). None of these approaches was confronted with human assessments or compared with each other, and in our study we decided to use the most established L_2 norm.

Overall, except the work from Simonetto et al. (2021), none of the existing attacks for tabular machine learning supports the feature relationships inherent to realistic tabular datasets, as summarized in Table 1. Nevertheless, in our empirical study we evaluate all the approaches that support continuous values and where a public implementation is available to confirm our claims: LowProFool, BF*, CPGD, and MOEVA.

3. Problem formulation

We formulate in the following the problem of evasion attacks under constraints. We assume the attack to be untargeted (i.e. it aims to force misclassification in any incorrect class); the formulation for targeted attacks is similar.

We denote by $x \in \mathbb{R}^d$ an input example and by $y \in$

Table 1. Recent literature on evasion attacks for tabular machine learning models. In bold the attacks where a public implementation is disclosed.

Attack	Supported features	Supported constraints		
		Categorical	Discrete	Relations
LowProFool (LPF) (Ballet et al., 2019)	Continuous	No	No	No
Cartella et al. (2021)	Continous, Discrete, Categorical	Yes	Yes	No
Gressel et al. (2021)	Continous, Discrete, Categorical	Yes	Yes	No
Xu et al. (2023)	Categorical	Yes	No	No
Wang et al. (2020)	Categorical	Yes	No	No
Bao et al. (2023)	Categorical	Yes	No	No
BF*/BFS (Kulynych et al., 2018; Kireev et al., 2023)	Continous, Discrete, Categorical	Yes	Yes	No
Mathov et al. (2022)	Continous, Discrete, Categorical	Yes	Yes	No
CPGD, MOEVA (Simonetto et al., 2021)	Continous, Discrete, Categorical	Yes	Yes	Yes
CAPGD, CAA (OURS)	Continous, Discrete, Categorical	Yes	Yes	Yes

$\{1, \ldots, C\}$ its correct label. Let $h : \mathbb{R}^d \to \mathbb{R}^C$ be a classifier and $h_{c_k}(x)$ the classification score that h outputs for input x to be in class c_k. Let $\Delta \subseteq \mathbb{R}^d$ be the space of allowed perturbations. Then, the objective of an evasion attack is to find a $\delta \in \Delta$ such that

$$argmax_{c \in \{1,\ldots,C\}} h_c(x + \delta) \neq y. \tag{1}$$

In image classification, the set Δ is typically chosen as the perturbations within some l_p-ball around x, i.e. $\Delta_p = \{\delta \in \mathbb{R}^d, ||\delta||_p \leq \epsilon\}$ for a maximum perturbation threshold ϵ. This restriction aims at preserving the semantics of the original input by assuming that small enough perturbations will yield images that humans perceive the same as the original images and would therefore classify the perturbed input into the same class (while the classifier predicts another class). This also guarantees that the example remains meaningful, that is, $x + \delta$ is not an image with random noise.

Tabular data are by nature different from images. They typically represent objects of the considered application domain (e.g. botnet traffic (Chernikova & Oprea, 2022), financial transaction (Ghamizi et al., 2020)). We denote by $\varphi : Z \to \mathbb{R}^d$ the feature mapping function that maps objects of the problem space Z to a d-dimensional feature space defined by the feature set $F = \{f_1, f_2, ...f_d\}$. Each object $z \in Z$ must inherently respect some natural condition to be valid (to be able to exist in reality). In the feature space, these conditions translate into a set of constraints on the feature values, which we denote by Ω. By construction, any input example x obtained from a real-world object z satisfies Ω, noted $x \models \Omega$.

Thus, in the case of tabular data, we additionally require the perturbation δ applied to x to yield a valid example $x + \delta$ satisfying Ω, that is, $\Delta_p(x) = \{\delta \in \mathbb{R}^d : ||\delta||_p \leq \epsilon \wedge x + \delta \models \Omega\}$.

To define the constraint language expressing Ω, we consider the four types of constraint introduced by Simonetto et al. (Simonetto et al., 2021), which we found to be sufficient for the constraints related to the datasets we used in our experiments. Hence, *immutability* defines what features cannot be changed by an attacker; *boundaries* define upper / lower bounds for feature values; *type* specifies a feature to take continuous, discrete, or categorical values; and *feature relationships* capture numerical relations between features. These four types of constraints can be encoded using the following grammar:

$$\omega \coloneqq \omega_1 \wedge \omega_2 \mid \omega_1 \vee \omega_2 \mid \psi_1 \succeq \psi_2 \mid f \in \{\psi_1 \ldots \psi_k\} \tag{2}$$

$$\psi \coloneqq c \mid f \mid \psi_1 \oplus \psi_2 \mid x_i \tag{3}$$

where $f \in F$ is a feature, c is a constant, $\omega, \omega_1, \omega_2$ are constraint formulae, $\succeq \in \{<, \leq, =, \neq, \geq, >\}$ is a comparison operator, $\psi, \psi_1, \ldots, \psi_k$ are numeric expressions, $\oplus \in \{+, -, *, /\}$ is a numerical operator, and x_i is the value of the i-th feature of the original input x.

3.1. Constrained Projected Gradient Descent

Constrained Projected Gradient Descent (CPGD) Simonetto et al. (2021) is an extension of the well-established PGD attack (Madry et al., 2017) to generate adversarial examples satisfying constraints in tabular machine learning. Its key principle is to integrate constraint satisfaction into the loss function that PGD optimizes. This is achieved by translating each constraint ω into a differentiable function $penalty(x, \omega)$ that values to zero if $x \models \omega$; otherwise, the (positive) value of the function for x represents how far x is from satisfying ω. Table 2 shows how each construct of the constraint grammar translates into a penalty function.

Based on this, CPGD produces adversarial examples from an initial sample x_{orig} classified as y by iteratively computing:

$$x^{(k+1)} = R_\Omega(P_\mathcal{S}(x^{(k)} + \eta^{(k)} \nabla \mathcal{L}(x^{(k)}, y, h, \Omega)))) \tag{4}$$

where $x^0 = x_{orig}$ (the original input), $P_\mathcal{S}$ is a projection

Table 2. Translation from constraint formulae to penalty functions. τ is an infinitesimal value. ω a constraint and ψ a numerical value.

ID	Constraints formulae	Penalty function
$\wedge$	$\omega_1 \wedge \omega_2$	$\omega_1 + \omega_2$
$\vee$	$\omega_1 \vee \omega_2$	$\min(\omega_1, \omega_2)$
$\in$	$\psi \in \Psi = \{\psi_1, \dots \psi_k\}$	$\min(\{\psi_i \in \Psi : \mid \psi - \psi_i \mid\})$
$\leq$	$\psi_1 \leq \psi_2$	$max(0, \psi_1 - \psi_2)$
$<$	$\psi_1 < \psi_2$	$max(0, \psi_1 - \psi_2 + \tau)$
$=$	$\psi_1 = \psi_2$	$\mid \psi_1 - \psi_2 \mid$

onto $\mathcal{S} = \{x \in \mathbb{R}^d, ||x - x_{orig}||_p \leq \epsilon\}$, $\nabla \mathcal{L}$ is the gradient of loss function $\mathcal{L}$, defined as

$$\mathcal{L}(x, y, h, \Omega) = l(h(x), y) - \sum_{\omega_i \in \Omega} penalty(x, \omega_i). \quad (5)$$

In the original CPGD implementation (Simonetto et al., 2021), the step size $\eta^{(k)}$ follows a predefined decay schedule, $\eta^{(k)} = \epsilon \times 10^{-(1+\lfloor k/\lfloor K/M \rfloor \rfloor)}$, with $M = 7$, and $K = max(k)$. $\mathcal{L}'(x)$ abbreviates $\mathcal{L}(x, y, h, \Omega)$.

3.2. Experimental settings

Our experiments are driven by the following datasets, models, and attack parameters.

Datasets. To conduct our study, we selected tabular datasets that present feature relations based on domain constraints. **URL** (Hannousse & Yahiouche, 2021) is a dataset of legitimate and phishing URLs. With only 14 linear domain constraints and 63 features, it is the simplest of our benchmark. **LCLD** (George, 2018) is a credit-scoring dataset with non-linear constraints. The **WiDS** (Lee et al., 2020) dataset contains medical data on the survival of patients admitted to the ICU. It has only 30 linear domain constraints. The **CTU** (Chernikova & Oprea, 2022) dataset reports legitimate and botnet traffic from CTU University. The challenge of this dataset lies in its large number of linear domain constraints (360).

Architectures. We evaluate five top-performing architectures from a recent survey on tabular ML (Borisov et al., 2022): **TabTransformer** (Huang et al., 2020) and **TabNet** (Arik & Pfister, 2021) are transformer-based models. **RLN** (Shavitt & Segal, 2018) uses a regularization coefficient to minimize a counterfactual loss. **STG** (Yamada et al., 2020) optimizes feature selection with stochastic gates, and **VIME** (Yoon et al., 2020) relies on self-supervised learning. These deep learning architectures achieve equivalent performance to XGBoost, the best shallow machine learning model for our use cases.

Perturbation parameters. We use L2-norm to measure distance between original and perturbed input, because this norm is suitable for both numerical and categorical features. We set ϵ to 0.5 for all datasets. Each of these datasets has a critical (negative) class, respectively phishing URLs, rejected loans, botnets, and not surviving patients. Hence, we only attack clean examples from the critical class that are not already misclassified by the model. In these settings, the relevant success metric is robust accuracy, which enables cross-model comparisons.

4. Our Constrained *Adaptive* PGD

The relative lack of effectiveness of CPGD as reported in its original publication Simonetto et al. (2021) leads us to investigate the cause of these weaknesses. We investigate four factors that may affect the success rate of the attack: (1) because the choice of the step size is known to largely impact the effectiveness of gradient-based attacks (Mosbach et al., 2018), we conjecture that the fixed step size and predefined decay in CPGD might be suboptimal; (2) the algorithm is unaware of the trend, i.e. it does not consider whether the optimization is evolving successfully and is not able to react to it; (3) CPGD does not check constraint satisfaction between the iterations, which could "lock" the algorithm into a part of the invalid data space; (4) CPGD starts with the original example, whereas classical gradient-based attacks often benefit from random initialization.

4.1. CAPGD components

We propose Constrained Adaptive PGD (CAPGD), a new constraint-aware gradient-based attack that aims to overcome the limitations of CPGD and improve its effectiveness.

Step size selection We introduce a step-size adaptation. We follow the exploration-exploitation principle by gradually reducing the gradient step (Croce & Hein, 2020). However, unlike CPGD, this reduction does not follow a fixed schedule but is determined by the optimization trend. If the value of the loss function grows, we keep the same step size; otherwise, we halve it. That is, we start with a step $\eta^{(0)} = 2\epsilon$, and we identify checkpoints $w_0 = 0, w_1, ..., w_n$ at which we decide whether it is necessary to halve the size of the current step. We halve the step size if any of the following two conditions holds. First, we count how many cases since the last checkpoint w_{j-1} the update step has successfully increased $\mathcal{L}'$. The condition holds if the loss has increased for at least a fraction of ρ steps (we set $\rho = 0.75$)

$$\sum_{i=w_{j-1}}^{w_j - 1} \mathbf{1}_{\mathcal{L}'(x^{(i+1)}) > \mathcal{L}'(x^{(i)})} < \rho \cdot (w_j - w_{j-1})$$

Second, the step has not been reduced at the last checkpoint and the loss is less or equal to the loss of the last checkpoint:

$$\eta^{(w_{j-1})} \equiv \eta^{(w_j)} \wedge \mathcal{L}_{\text{max}}^{(w_{j-1})} \equiv \mathcal{L}_{\text{max}}^{(w_j)}$$

where $\mathcal{L}_{\text{max}}^{(w_j)}$ is the highest objective value in the first $j+1$ iterations.

Repair operator We also introduce a new "repair" operator denoted R_Ω that projects back the example produced at each iteration into the valid data space. The idea is to force the value of any feature f that occurs in constraints of the form $f = \psi$ (see Equation 3) to be ψ valued based on all other feature values in the example.

Initial state As for initialization, we apply the attack from two initial states: the original example x_{orig} and a random example sampled from $\mathcal{S}$ (the Lp-ball around x_{orig}). The goal behind this second initialization is to reduce the risk of being immediately locked into local optima that encompass only invalid examples.

Gradient step Finally, we introduce in CAPGD a momentum (Dong et al., 2018)

. Let $\eta^{(k)}$ be the step size at iteration k, then we first compute $z^{(k+1)}$ before the updated example $x^{(k+1)}$.

$$z^{(k+1)} = P_{\mathcal{S}}(x^{(k)} + \eta^{(k)}(\nabla\mathcal{L}'(x^{(k)}))) \tag{6}$$

$$\begin{aligned} x^{(k+1)} = R_\Omega(P_{\mathcal{S}}(x^{(k)} + \alpha \cdot (z^{(k+1)} - x^{(k)}) \\ + (1-\alpha)\cdot(x^{(k)} - x^{(k+1)})) \end{aligned} \tag{7}$$

where $\alpha \in [0,1]$ (we use $\alpha = 0.75$ following (Croce & Hein, 2020)) regulates the influence of the previous update on the current, and $P_{\mathcal{S}}$ is the projection onto $\mathcal{S} = \{x \in \mathbb{R}^d, ||x - x_{orig}||_p \leq \epsilon\}$.

4.2. Comparison of CAPGD to gradient-based attacks

To evaluate the benefits of CAPGD, we compare it with CPGD as well as LowProFool, the only other public gradient attack for tabular models that can be extended to support all feature types.

CAPGD is more successful than existing gradient attacks. We compare the robust accuracy across our five datasets and five architectures against CPGD, LowProFool, and CAPGD, and report the results in Table 3. CAPGD significantly outperforms CPGD and LowProFool. It decreases the robust accuracy on URL, LCLD, and WIDS datasets to as low as 10.9%, 0.2%, and 10.2% respectively.

CAPGD subsumes all gradient attacks. We analyze in detail the original examples from which attacks could generate valid and successful adversarial examples. For each attack, we take the union of the sets of clean examples across 5 seeds. We generate the Venn diagram for CPGD, LowProFool, and CAPGD, for all datasets and model architectures. We sum the partition values in Figure 1. CAPGD generates adversarial examples for 6597 original examples from which none of the other gradient attacks could produce adversarial examples. In contrast, all successful adversarial examples by CPGD (132) and LowProFool (3) are also generated by CAPGD.

Table 3. Robust accuracy for CAPGD and SOTA gradient attacks. A lower robust accuracy means a more effective attack. The lowest robust accuracy is in bold.

Dataset	Model	Clean	LPF	CPGD	CAPGD
URL	TabTr.	93.6	93.6	91.9	**10.9**
	RLN	94.4	94.4	92.8	**12.6**
	VIME	92.5	92.5	90.7	**56.3**
	STG	93.3	93.3	93.3	**72.6**
	TabNet	93.4	93.4	88.5	**19.3**
LCLD	TabTr.	69.5	69.2	69.5	**27.1**
	RLN	68.3	68.3	68.3	**0.2**
	VIME	67.0	67.0	67.0	**2.6**
	STG	66.4	66.4	66.4	**55.5**
	TabNet	67.4	67.4	67.4	**6.3**
CTU	TabTr.	**95.3**	**95.3**	**95.3**	**95.3**
	RLN	**97.8**	**97.8**	**97.8**	**97.8**
	VIME	**95.1**	**95.1**	**95.1**	**95.1**
	STG	**95.3**	**95.3**	**95.3**	**95.3**
	TabNet	**96.1**	**96.1**	**96.1**	**96.1**
WIDS	TabTr.	75.5	75.5	75.2	**48.0**
	RLN	77.5	77.5	77.3	**61.8**
	VIME	72.3	72.3	71.5	**51.4**
	STG	77.6	77.6	77.5	**65.1**
	TabNet	79.7	79.7	76.0	**10.2**

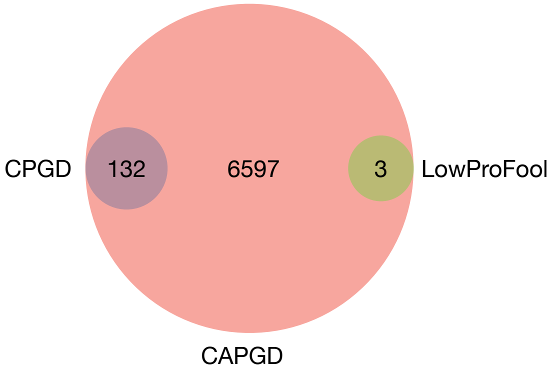

Figure 1. Complementarity of CAPGD, CPGD and LowProFool with the number of successful adversarial examples.

Conclusion

In this work, we first propose CAPGD, a new parameter-free gradient attack for constrained tabular machine learning. We evaluate our attack over four datasets and five architectures and demonstrated that our new attack outperforms all previous attacks in terms of effectiveness and efficiency.

References

Aghakhani, H., Gritti, F., Mecca, F., Lindorfer, M., Ortolani, S., Balzarotti, D., Vigna, G., and Kruegel, C. When malware is packin'heat; limits of machine learning classifiers based on static analysis features. In *Network and Distributed Systems Security (NDSS) Symposium 2020*, 2020.

Arik, S. Ö. and Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021.

Ballet, V., Aigrain, J., Laugel, T., Frossard, P., Detyniecki, M., et al. Imperceptible adversarial attacks on tabular data. In *NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness and Privacy (Robust AI in FS 2019)*, 2019.

Bao, H., Han, Y., Zhou, Y., Gao, X., and Zhang, X. Towards efficient and domain-agnostic evasion attack with high-dimensional categorical inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6753–6761, 2023.

Borisov, V., Leemann, T., Sessler, K., Haug, J., Pawelczyk, M., and Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2022. doi: 10.1109/tnnls.2022.3229161. URL https://doi.org/10.1109%2Ftnnls.2022.3229161.

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

Cartella, F., Anunciaçao, O., Funabiki, Y., Yamaguchi, D., Akishita, T., and Elshocht, O. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. 2021.

Chernikova, A. and Oprea, A. Fence: Feasible evasion attacks on neural networks in constrained environments. *arXiv preprint arXiv:1909.10480*, 2019.

Chernikova, A. and Oprea, A. Fence: Feasible evasion attacks on neural networks in constrained environments. *ACM Transactions on Privacy and Security*, 25(4):1–34, 2022.

Clements, J. M., Xu, D., Yousefi, N., and Efimov, D. Sequential deep learning for credit risk monitoring with tabular financial data. *arXiv preprint arXiv:2012.15330*, 2020.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Dyrmishi, S., Ghamizi, S., Simonetto, T., Traon, Y. L., and Cordy, M. On the empirical effectiveness of unrealistic adversarial hardening against realistic adversarial attacks. *arXiv preprint arXiv:2202.03277*, 2022.

Dyrmishi, S., Ghamizi, S., and Cordy, M. How do humans perceive adversarial text? a reality check on the validity and naturalness of word-based adversarial attacks, 2023.

George, N. Lending club loan data. https://www.kaggle.com/datasets/wordsforthewise/lending-club, 2018.

Ghamizi, S., Cordy, M., Gubri, M., Papadakis, M., Boystov, A., Le Traon, Y., and Goujon, A. Search-based adversarial testing and improvement of constrained credit scoring systems. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1089–1100, 2020.

Gressel, G., Hegde, N., Sreekumar, A., Radhakrishnan, R., Harikumar, K., Achuthan, K., et al. Feature importance guided attack: a model agnostic adversarial attack. *arXiv preprint arXiv:2106.14815*, 2021.

Grosse, K., Bieringer, L., Besold, T. R., Biggio, B., and Alahi, A. When your ai becomes a target: Ai security incidents and best practices. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 23041–23046, Mar. 2024. doi: 10.1609/aaai.v38i21.30347. URL https://ojs.aaai.org/index.php/AAAI/article/view/30347.

Hancock, J. T. and Khoshgoftaar, T. M. Survey on categorical data for neural networks. *Journal of Big Data*, 7:1–41, 2020.

Hannousse, A. and Yahiouche, S. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Engineering Applications of Artificial Intelligence*, 104:104347, 2021.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Kireev, K., Kulynych, B., and Troncoso, C. Adversarial robustness for tabular data through cost and utility awareness. *arXiv preprint arXiv:2208.13058*, 2022.

Kireev, K., Kulynych, B., and Troncoso, C. Adversarial robustness for tabular data through cost and utility awareness. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Kulynych, B., Hayes, J., Samarin, N., and Troncoso, C. Evading classifiers in discrete domains with provable optimality guarantees. *arXiv preprint arXiv:1810.10939*, 2018.

Lee, M., Raffa, J., Ghassemi, M., Pollard, T., Kalanidhi, S., Badawi, O., Matthys, K., and Celi, L. A. Wids (women in data science) datathon 2020: Icu mortality prediction. PhysioNet, 2020.

Long, T., Gao, Q., Xu, L., and Zhou, Z. A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions. *Computers & Security*, 121:102847, 2022. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2022.102847. URL https://www.sciencedirect.com/science/article/pii/S0167404822002413.

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

Mathov, Y., Levy, E., Katzir, Z., Shabtai, A., and Elovici, Y. Not all datasets are born equal: On heterogeneous tabular data and adversarial examples. *Knowledge-Based Systems*, 242:108377, 2022.

Mosbach, M., Andriushchenko, M., Trost, T., Hein, M., and Klakow, D. Logit pairing methods can fool gradient-based attacks. *NeurIPS 2018 Workshop on Security in Machine Learning*, 2018.

Pierazzi, F., Pendlebury, F., Cortellazzi, J., and Cavallaro, L. Intriguing properties of adversarial ml attacks in the problem space, 2020.

Shavitt, I. and Segal, E. Regularization learning networks: deep learning for tabular datasets. *Advances in Neural Information Processing Systems*, 31, 2018.

Shwartz-Ziv, R. and Armon, A. Tabular Data: Deep Learning is Not All You Need. *arXiv preprint arXiv:2106.03253*, 2021.

Simonetto, T., Dyrmishi, S., Ghamizi, S., Cordy, M., and Traon, Y. L. A unified framework for adversarial attack and defense in constrained feature space. *arXiv preprint arXiv:2112.01156*, 2021.

Somani, S., Russak, A. J., Richter, F., Zhao, S., Vaid, A., Chaudhry, F., De Freitas, J. K., Naik, N., Miotto, R., Nadkarni, G. N., et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace*, 2021.

Ulmer, D., Meijerink, L., and Cinà, G. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *Machine Learning for Health*, pp. 341–354. PMLR, 2020.

Wang, Y., Han, Y., Bao, H., Shen, Y., Ma, F., Li, J., and Zhang, X. Attackability characterization of adversarial evasion attack on discrete data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1415–1425, 2020.

Xu, H., He, P., Ren, J., Wan, Y., Liu, Z., Liu, H., and Tang, J. Probabilistic categorical adversarial attack and adversarial training. In *International Conference on Machine Learning*, pp. 38428–38442. PMLR, 2023.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates. In *Proceedings of Machine Learning and Systems 2020*, pp. 8952–8963. 2020.

Yoon, J., Zhang, Y., Jordon, J., and van der Schaar, M. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

Zhang, S., Yao, L., Sun, A., and Tay, Y. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.