TabAttackBench: A Benchmark for Adversarial Attacks on Tabular Data

Zhipeng He[a,b,*], Chun Ouyang[a,b], Lijie Wen[c], Cong Liu[d], Catarina Moreira[e,b,f]

[a]*School of Information Systems, Queensland University of Technology, Brisbane, Australia*
[b]*Center for Data Science, Queensland University of Technology, Brisbane, Australia*
[c]*School of Software, Tsinghua University, Beijing, China*
[d]*NOVA Information Management School, NOVA University of Lisbon, Lisboa, Portugal*
[e]*Data Science Institute, University of Technology, Sydney, Australia*
[f]*INESC-ID/Instituto Superior Técnico, University of Lisboa, Lisboa, Portugal*

Abstract

Adversarial attacks pose a significant threat to machine learning models by inducing incorrect predictions through imperceptible perturbations to input data. While these attacks have been extensively studied in unstructured data like images, their application to tabular data presents new challenges. These challenges arise from the inherent heterogeneity and complex feature interdependencies in tabular data, which differ significantly from those in image data. To address these differences, it is crucial to consider imperceptibility as a key criterion specific to tabular data. Most current research focuses primarily on achieving effective adversarial attacks, often overlooking the importance of maintaining imperceptibility. To address this gap, we propose a new benchmark for adversarial attacks on tabular data that evaluates both effectiveness and imperceptibility. In this study, we assess the effectiveness and imperceptibility of five adversarial attacks across four models using eleven tabular datasets, including both mixed and numerical-only datasets. Our analysis explores how these factors interact and influence the overall performance of the attacks. We also compare the results across different dataset types to understand the broader implications of these findings. The findings from this benchmark provide valuable insights for improving the design of adversarial attack algorithms, thereby advancing the field of adversarial machine learning on tabular data.

Keywords: Adversarial attack, Tabular data, Benchmark, Machine learning, Robustness

1. Introduction

In recent years, the field of machine learning has seen substantial advancements, leading to the deployment of models across a wide range of applications. However, with these advancements comes increasing concern about the robustness and security of models, particularly in the context of adversarial attacks. Adversarial attacks involve the intentional manipulation of input data to deceive machine learning models, causing incorrect or misleading outputs [1]. This area of research has drawn significant attention as researchers strive to understand and mitigate the vulnerabilities in various types of data and models. For instance, adversarial attacks on image data can cause misclassification of objects, which is concerning for applications like autonomous driving, surveillance, and facial recognition systems [2]. Similarly, Natural Language Processing (NLP) models are susceptible to attacks that can alter the meaning of sentences or generate misleading summaries, impacting applications in sentiment analysis, machine translation, and chatbots [3]. Additionally, speech recognition systems can be tricked by adversarial audio inputs, leading to incorrect transcriptions or commands, which has serious implications for virtual assistants and voice-controlled devices [4]. By addressing the vulnerabilities in these types of data, researchers aim to develop more robust and secure machine learning systems across various domains.

1.1. Challenges in Adversarial Attacks on Tabular Data

Tabular data, structured yet rich in semantics, heterogeneity, and interdependencies, is prevalent in domains such as finance, healthcare, and e-commerce. These datasets often contain vital information used for decision-making processes, predictive modelling, and anomaly detection. Despite their significance, machine learning models trained on tabular data (which can be referred to as tabular data models) remain underexplored regarding the vulnerabilities to adversarial attacks.

*Corresponding author

Email addresses: zhipeng.he@hdr.qut.edu.au (Zhipeng He), c.ouyang@qut.edu.au (Chun Ouyang), wenlj@tsinghua.edu.cn (Lijie Wen), cliu@novaims.unl.pt (Cong Liu), catarina.pintomoreira@uts.edu.au (Catarina Moreira)

The potential impact of adversarial attacks on tabular data models is profound. Such attacks can compromise the integrity and reliability of machine learning models, resulting in misclassification and potentially severe consequences for applications relying on precise prediction and data-driven decisions. The vulnerabilities of tabular data models to adversarial attacks are particularly notable due to the unique characteristics of tabular data. Unlike image or text data, where each data point is typically represented as pixels or words, tabular data presents a different challenge due to its varied nature. For example, consider a dataset containing customer information for a bank loan application. It includes categorical variables like *marital status* and *employment type*, numerical variables such as *income*, and possibly missing values in fields like *previous loans*. Additionally, these features often exhibit diverse distributions; for instance, income might follow a skewed distribution, while employment type is categorical. These complexities make applying adversarial attacks to tabular data more intricate compared to image or text data.

1.2. Benchmarking Adversarial Attacks on Tabular Data

An important aspect of advancing adversarial attack research is the establishment of benchmarks. These benchmarks function as standardised tests that evaluate the robustness of machine learning models against adversarial attacks. They provide a common ground for comparing different approaches and methodologies, thereby facilitating the development of more robust models. While considerable progress has been made in understanding adversarial attacks on image [5] and text [6], there remains a relatively underexplored area: adversarial attacks on tabular data. Our work addresses this gap by introducing a new benchmark specifically designed for attacks on tabular data.

Existing benchmarks, as summarised in Table 1, primarily focus on evaluating attacks on image, graph, and time-series data, covering a range of adversarial techniques such as black-box attacks [9], patch-based attacks [10], and transferability-based attacks [7]. These benchmarks typically evaluate adversarial robustness using metrics like attack success rate, adversarial accuracy, and norm-based metrics (e.g., ℓ_∞, ℓ_2) to quantify the strength of the adversarial perturbations. While these metrics are well-suited for image data, where

Table 1: Overview of existing benchmarks on adversarial attacks across different data types, attack types, and evaluation metrics; our work introduces a new benchmark for attacks on tabular data.

Benchmark	**Data Type**	**Attack Type**	**Evaluation Metric**
Jin et al. [7]	Image	Transferable Attacks	Attack Transferability Score
Dong et al. [8]	Image	White-box, Black-box Attacks	Adversarial Accuracy, ℓ_∞ Norm
Zheng et al. [9]	Image	Black-box Attacks	Attack Success Rate, Query Count
Croce et al. [5]	Image	ℓ_∞, ℓ_2 Norm-based Attacks	ℓ_∞, ℓ_2 Norm, Corruption Robustness
Hingun et al. [10]	Image	Patch-based Attacks	Patch Success Rate, Realism Score
Cinà et al. [11]	Image	Gradient-based Attacks	Adversarial Success Rate
Zheng et al. [12]	Graph	Attacks on Graphs	Adversarial Robustness on Graphs
Li et al. [13]	VQA	Attacks on VQA	VQA Accuracy under Adversarial Conditions
Siddiqui et al. [14]	Time-series	Attacks on Time-Series	Time-Series Attack Success Rate
Our paper	*Tabular*	*Attacks on Tabular Data*	*Attack Success Rate, Imperceptibility Metrics*

imperceptibility is measured by slight pixel changes that remain "indistinguishable to the human eye" [15], they do not translate directly to tabular data. Adversarial examples are created by strategically perturbing these pixel values to cause misclassification by machine learning models, while still preserving the visual similarity to the original image.

In contrast, tabular data presents unique challenges for adversarial attacks. Any changes to feature values can be easily spotted, making imperceptibility a more complex issue. For example, in image-based attacks, imperceptibility is concerned with altering pixels without compromising the visual integrity of the image. However, in tabular data, the definition of imperceptibility must account for human detectability across various feature dimensions.

Our benchmark includes the concept of imperceptibility for tabular data by focusing on four key quantitative properties: **Proximity**, **Sparsity**, **Deviation**, and **Sensitivity** [16]. These properties ensure that adversarial examples closely resemble original data, minimise feature alterations, and respect the statistical distribution of the data.

Additionally, while existing benchmarks are primarily focused on effectiveness (attack success rates), our benchmark uniquely combines this with an evaluation of imperceptibility, offering a more comprehensive assessment of adversarial robustness for tabular data. The incorporation of these imperceptibility metrics differentiates our work from prior benchmarks, which mostly focus on images, graphs, or time-series data and lack a detailed assessment of imperceptibility, particularly in the context of tabular datasets.

1.3. Contribution

This paper aims to address three main research questions. First, *how effective are the evaluated adversarial attack algorithms on tabular data?* Second, *how imperceptible are these adversarial attack algorithms on tabular data?* Finally, *whether and how the evaluated algorithms can achieve a balance between both imperceptibility and effectiveness*, striving for optimal results in both aspects. To address these three questions, the paper evaluate the effectiveness of various adversarial attack algorithms on tabular data, examines the imperceptibility of these attacks, and explores the potential for achieving a balance between imperceptibility and effectiveness.

In the remainder of this paper, we explore the current state of adversarial attack research, with a particular focus on tabular data (Section 2). We propose a benchmark evaluation of the effectiveness and imperceptibility of adversarial attacks on tabular data (Section 3). By examining the trade-offs between attack success rates and imperceptibility, our evaluation framework provides valuable insights for developing both effective and imperceptible adversarial attacks (Section 4). Our analysis further illuminates how different attack strategies prioritise either maximising attack imperceptibility or attack effectiveness, enabling researchers to strategically balance these competing objectives when designing novel adversarial techniques for tabular data (Section 5).

2. Background and Related Work

2.1. Adversarial Attacks

Adversarial attacks aim to mislead a machine learning model into making incorrect classifications by generating deliberately perturbed input data, known as *adversarial examples*. Consider a dataset where each input data point, represented by a vector $\boldsymbol{x} \in \mathbb{X}$, is associated with a class label $y \in \mathbb{Y}$. We define a machine learning classifier $f(\cdot)$. An adversarial example $\boldsymbol{x}^{adv}$ is a perturbed variant of $\boldsymbol{x}$ that remains similar to $\boldsymbol{x}$ but is specifically designed to cause the classifier to incorrectly predict the label of $\boldsymbol{x}$. This can be mathematically defined as:

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \boldsymbol{\delta} \quad \text{subject to } f(\boldsymbol{x}^{adv}) \neq y$$

where $\boldsymbol{\delta}$ denotes input perturbation.

A comprehensive taxonomy of adversarial attacks can be categorised based on various dimensions, considering factors such as the adversary's goals, capabilities, and knowledge [17].

2.1.1. Adversary's Goals

These can be categorised into four types based on how the adversarial perturbation affects the model's classification output [17, 18]: (1) *Confidence reduction* undermines prediction certainty without changing class labels, (2) *Nontargeted misclassification* seeks any form of classification error regardless of outcome specificity, (3) *Targeted misclassification* forces specific wrong outputs, and (4) *Source/target misclassification* requires dual specification of input and output classes.

Our focus lies on **nontargeted misclassification**, where the attacker's goal is to cause the model to produce any incorrect class prediction, regardless of specificity. For example, in a credit scoring model that predicts the likelihood of loan repayment based on tabular data such as income, credit history, and debt, this could mean changing the classification of a "low-risk" borrower to any other incorrect class, such as "medium" or "high-risk", without the attacker caring which specific misclassification occurs. Unlike targeted attacks, this approach does not require a specific incorrect outcome — its success depends only on causing a mismatch with the ground-truth label.

This category is particularly consequential for tabular data systems, where structured feature interdependencies amplify the potential effects of even arbitrary misclassifications. Its computational efficiency—since it does not require targeting specific classes—and its broad relevance—since any error can disrupt real-world systems—make it a priority in our benchmark.

2.1.2. Adversary's Capabilities

Adversarial attacks can also be classified based on the capabilities of the adversary, specifically in terms of how much control they have over the perturbations applied to the input data. These capabilities are generally divided into two categories: unbounded attacks

and bounded attacks. Each of these attack types represents different levels of access and constraints that an adversary may face when attempting to compromise a model.

- **Unbounded Attacks:** In an unbounded attack, the adversary has no restrictions on the magnitude or extent of the perturbations they can apply to the input data. Unbounded attacks attempt to minimise the distance between input $\boldsymbol{x}$ and adversarial example $\boldsymbol{x}^{adv}$ to obtain the minimal perturbation $\boldsymbol{\delta}$ without constraints on the magnitude of attack perturbation.

$$\min \|\boldsymbol{\delta}\| \quad \text{subject to } f(\boldsymbol{x}^{adv}) \neq y \tag{1}$$

- **Bounded Attacks:** Bounded attacks are more constrained, as the adversary is limited in how much they can perturb the input data. These attacks are defined by a upper bound ϵ on the amount of changes, such as keeping perturbation within a certain range. The goal is to find an adversarial example $\boldsymbol{x}^{adv}$, which has perturbation $\boldsymbol{\delta}$, within the budget ϵ, to an input $\boldsymbol{x}$ that misleads the prediction by maximising the loss function $\mathcal{L}$ of the machine learning model being attacked.

$$\max \mathcal{L}(f(\boldsymbol{x}^{adv}), y) \quad \text{subject to } \|\boldsymbol{\delta}\| \leq \epsilon \tag{2}$$

Our benchmark will encompass both unbounded and bounded attacks to provide a comprehensive evaluation of adversarial attacks on tabular data.

2.1.3. Adversary's Knowledge

Adversarial attacks are categorised into three primary threat models based on the adversary's knowledge of the target system: *white-box*, *gray-box*, and *black-box* attacks [17, 18]. These classifications reflect the attacker's access to knowledge about the model's architecture, parameters, and training data, ranging from full transparency (white-box) to complete opacity (black-box), with gray-box scenarios representing partial knowledge. Among these, white-box attacks pose the most severe threat, as they leverage comprehensive model insights

to craft precise adversarial perturbations.

A **white-box attack** represents the most severe threat scenario, where the adversary possesses full knowledge of the target model, including its architecture (e.g., layer types, activation functions), trained parameters (weights and biases), and the distribution of training data. This complete access allows the attacker to compute exact gradients of the loss function with respect to the input data, enabling highly optimised adversarial example generation through techniques like gradient ascent. By exploiting the model's mathematical structure, adversaries can systematically identify input perturbations that maximally degrade performance while remaining imperceptible or semantically valid.

2.2. White-box Adversarial Attacks

In this work, we prioritise white-box attacks for benchmarking adversarial attacks in tabular data systems for two reasons. First, white-box attacks provide an upper-bound evaluation of vulnerability by assuming worst-case adversarial scenarios, thereby revealing fundamental weaknesses in model design or training. Second, most state-of-the-art attack methods are developed under white-box assumptions, enabling systematic comparisons with existing literature. While gray-box and black-box attacks have practical relevance, white-box analysis offers a rigorous baseline for assessing inherent model robustness before considering real-world constraints on adversarial knowledge. Here are some popular methods widely used for white-box attacks:

2.2.1. Fast Gradient Sign Method

The fast gradient sign method (FGSM) attack [15] works by calculating the gradient of the neural network with respect to the input data and using the sign of this gradient to determine the direction of the perturbation. Given an original datapoint $\boldsymbol{x}$ and the corresponding label y, the problem can be formalised as:

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \mathbf{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)) \tag{3}$$

where ϵ is an initialised hyperparameter for determining the size of perturbation, $\nabla_{\boldsymbol{x}} J(\cdot)$

represents the gradient of with respect to datapoint $\boldsymbol{x}$. The term 'fast' in the context of the FGSM attack refers to the fact that it can be executed quickly with just a single forward pass through the neural network to generate the gradient, and a single backward pass to update the input data with the calculated perturbation. This simplicity and efficiency make it a popular choice for generating adversarial examples in machine learning applications.

2.2.2. Iterative Method

Iterative methods are a family of techniques used to produce adversarial examples by introducing perturbations into the input data in a series of small steps rather than a single large step. Two famous examples are the *Basic Iterative Method (BIM)* and *Projected Gradient Descent (PGD)*, both derived from FGSM.

BIM. The Basic Iterative Method [19] extends FGSM by repeatedly applying gradient-guided perturbations. Starting with the original input $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$, it iteratively updates adversarial examples by ascending the loss gradient while constraining perturbations within a predefined ϵ-ball. For the ith iteration, the update formulations are following:

$$\begin{aligned} \boldsymbol{x}_0^{adv} &= \boldsymbol{x}, \\ \boldsymbol{x}_{i+1}^{adv} &= \mathbf{Clip}\{\boldsymbol{x}_i^{adv} + \alpha \cdot \mathbf{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_i^{adv}, y))\} \end{aligned} \tag{4}$$

where α controls the magnitude of perturbation. Also, the perturbation is clipped to prevent it from becoming too large and potentially distorting the input beyond recognition.

PGD. Projected Gradient Descent [20] generalises BIM by incorporating two key modifications: (1) initialising $\boldsymbol{x}_0^{adv}$ as a random point within the ϵ-neighborhood of $\boldsymbol{x}$, and (2) projecting perturbed samples back to the feasible region after each update:

$$\begin{aligned} \boldsymbol{x}_0^{adv} &= \mathbf{Random}(\boldsymbol{x}), \\ \boldsymbol{x}_{i+1}^{adv} &= \mathbf{Proj}\{\boldsymbol{x}_i^{adv} + \alpha \cdot \mathbf{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_i^{adv}, y))\} \end{aligned} \tag{5}$$

Here, **Proj** ensures $\boldsymbol{x}^{adv}$ remains within the the ϵ-bound via L_∞-norm projection. Unlike BIM's deterministic initialisation, PGD's stochastic start helps escape local optima, making

it a more robust attack framework. BIM can be viewed as a special case of PGD without random initialisation or explicit projection steps.

2.2.3. Carlini and Wagner Attack

The Carlini and Wagner attack [21], also known as the C&W attack, is a state-of-the-art adversarial attack that aims to find the minimum perturbation that can cause a misclassification in a targeted deep neural network. Unlike some other attacks, the C&W attack is designed to optimise a loss function that combines both the perturbation magnitude with distance metrics and the prediction confidence with objective function, which is:

$$\arg\min_{\boldsymbol{x}^{adv}} \|\boldsymbol{x} - \boldsymbol{x}^{adv}\|_p + c \cdot z(\boldsymbol{x}^{adv}) \tag{6}$$

where $z(\boldsymbol{x}^{adv})$ is the objectve function. This approach enables the C&W attack to be effective even against models that are robust to other types of attacks. Moreover, the C&W attack can also be adapted to perform targeted attacks, where the attacker aims to cause the model to predict a specific target class.

2.2.4. DeepFool

DeepFool [22] constitutes an attack method based on the ℓ_2-norm. Its underlying assumption is that the predictive model is linear and that a hyperplane $\mathcal{F} = \{\boldsymbol{x} : \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$ exists that separates one class from another. Consequently, the search for adversarial examples can be framed as an optimization problem, expressed as follows:

$$\begin{gathered} r_*(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}^{adv}} \|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_2 = -\frac{f(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2^2 \boldsymbol{w}} \\ \text{subject to } f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{adv}), \end{gathered} \tag{7}$$

where $r_*(\boldsymbol{x})$ is the minimum perturbation to change the class of $\boldsymbol{x}$. It measures the closest distance from datapoint to the decision boundary hyperplane (shown in Figure 1).

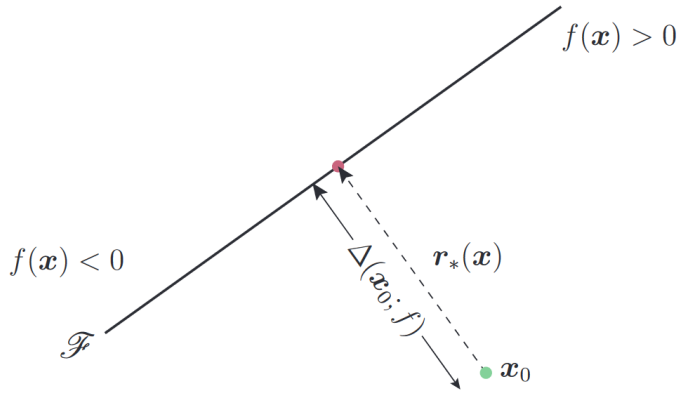


Figure 1: DeepFool attacks in a linear binary classifier [22].

2.2.5. LowProFool

LowProFool is proposed by Ballet et al. [23], an updated version of DeepFool for tabular data, which uses a weighted ℓ_p norm to determine the set of features to perturb. This attack utilizes the absolute value of the Pearson's correlation coefficient for each numerical feature as feature importance $\boldsymbol{v}$. Specifically, this metric is utilized to identify which features are comparatively inconspicuous and more challenging for human observers to detect.

$$\begin{gathered} r_*(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}^{adv}} d(\boldsymbol{x}^{adv} - \boldsymbol{x}) = \|(\boldsymbol{x}^{adv} - \boldsymbol{x}) \odot \boldsymbol{v}\|_p^2 \\ \text{subject to } f(\boldsymbol{x}) \neq f(\boldsymbol{x}^{adv}) \end{gathered} \tag{8}$$

2.3. Adversarial Machine Learning on Tabular Data

Adversarial attacks on tabular data aim to manipulate inputs such that machine learning models produce incorrect predictions while keeping modifications imperceptible. These attacks can be categorised into white-box (requiring full model access) and black-box (requiring only query-based access) methods. This review groups adversarial attack methods accordingly and presents their key strengths and weaknesses.

White-box attacks leverage full access to the model's internal mechanisms, making them highly effective in generating optimised adversarial perturbations. One of the most fundamental gradient-based white-box methods is Projected Gradient Descent (PGD), introduced by Madry et al. [20], which iteratively perturbs input features in a direction that maximizes

classification error. Building upon PGD, Simonetto et al. [24] developed Constrained Adaptive PGD (CAPGD), which modifies traditional gradient attacks to account for feature constraints, ensuring that adversarial perturbations remain realistic within a tabular dataset's domain constraints. Another prominent gradient-based attack is LowProFool attack [23]. This method selectively perturbs low-importance features, ensuring that adversarial modifications remain imperceptible to human scrutiny while still deceiving the model.

Beyond gradient-based attacks, some white-box approaches incorporate additional constraints to align adversarial perturbations with real-world feasibility. Mathov et al. [25] proposed a method that preserves tabular feature interdependencies by ensuring that modifications remain consistent with the underlying data structure. Similarly, FENCE [26] introduced a framework for crafting adversarial examples in security-related datasets, where feature dependencies must be respected to maintain feasibility. Other constraint-aware attacks focus on financial constraints, such as cost-aware adversarial framework [27], which generates adversarial examples that adhere to a given budget, ensuring that attacks are feasible from an economic standpoint.

Generative model-based attacks also play a role in white-box strategies. Zhou et al. [28] leveraged generative adversarial networks (GANs) to create adversarial examples that preserve statistical properties of the original dataset. This technique enhances attack stealth by ensuring that adversarial modifications follow the distributional characteristics of clean data, making detection more difficult. Despite their high attack success rates, white-box methods have notable limitations. They assume full knowledge of the target model, which is often unrealistic in real-world applications where machine learning models are deployed in black-box environments. Additionally, white-box attacks can be countered through adversarial training, wherein models are retrained with adversarial examples to improve their robustness.

Unlike white-box attacks, black-box attacks do not require access to model parameters or gradients. Instead, they rely on query-based techniques to infer decision boundaries and generate adversarial perturbations. One of the earliest and most well-known black-box methods is the Zero-Order Optimisation (ZOO) attack [29], which approximates gradients

using finite-difference methods. However, ZOO is computationally expensive due to the large number of queries required to estimate the gradient with high precision. Other decision-based attacks, such as Boundary Attack [30] take an adversarial approach by starting with a large perturbation and iteratively refining it while maintaining misclassification. A more refined version of this approach is the HopSkipJump Attack (HSJA) [31], which adapts decision-based attacks using dynamic step sizes to minimise queries while still achieving high attack success rates.

In addition to query-based attacks, some black-box methods adopt model-agnostic approaches that do not require direct access to model gradients. Feature Importance Guided Attack (FIGA) [32] perturbs the most influential features of a dataset without relying on internal model parameters. By focusing on high-importance features, FIGA maximises the likelihood of misclassification while minimising the number of modified features. Cartella et al. [33] extended black-box adversarial attacks into real-world fraud detection systems by adapting boundary-based methods to bypass anomaly detection algorithms. These adaptations enable adversarial perturbations to remain undetected while still achieving model evasion.

Despite their practical advantages, black-box attacks have inherent limitations. They often require a large number of queries to approximate gradients, making them computationally expensive and slow. Moreover, query-efficient black-box attacks typically struggle to generate minimal perturbations, resulting in adversarial examples that are more perceptible compared to their white-box counterparts. Nevertheless, black-box attacks are more applicable in real-world scenarios, as attackers typically do not have access to model internals.

3. Methodology

Integrating adversarial attacks as an additional phase in the standard machine learning pipeline enables a systematic assessment of model robustness against adversarial perturbations. Inspired by the established machine learning benchmark guidelines [34], we design our benchmark to evaluate adversarial attacks on tabular data. Figure 2 provides an overview of the proposed evaluation framework for this benchmark.

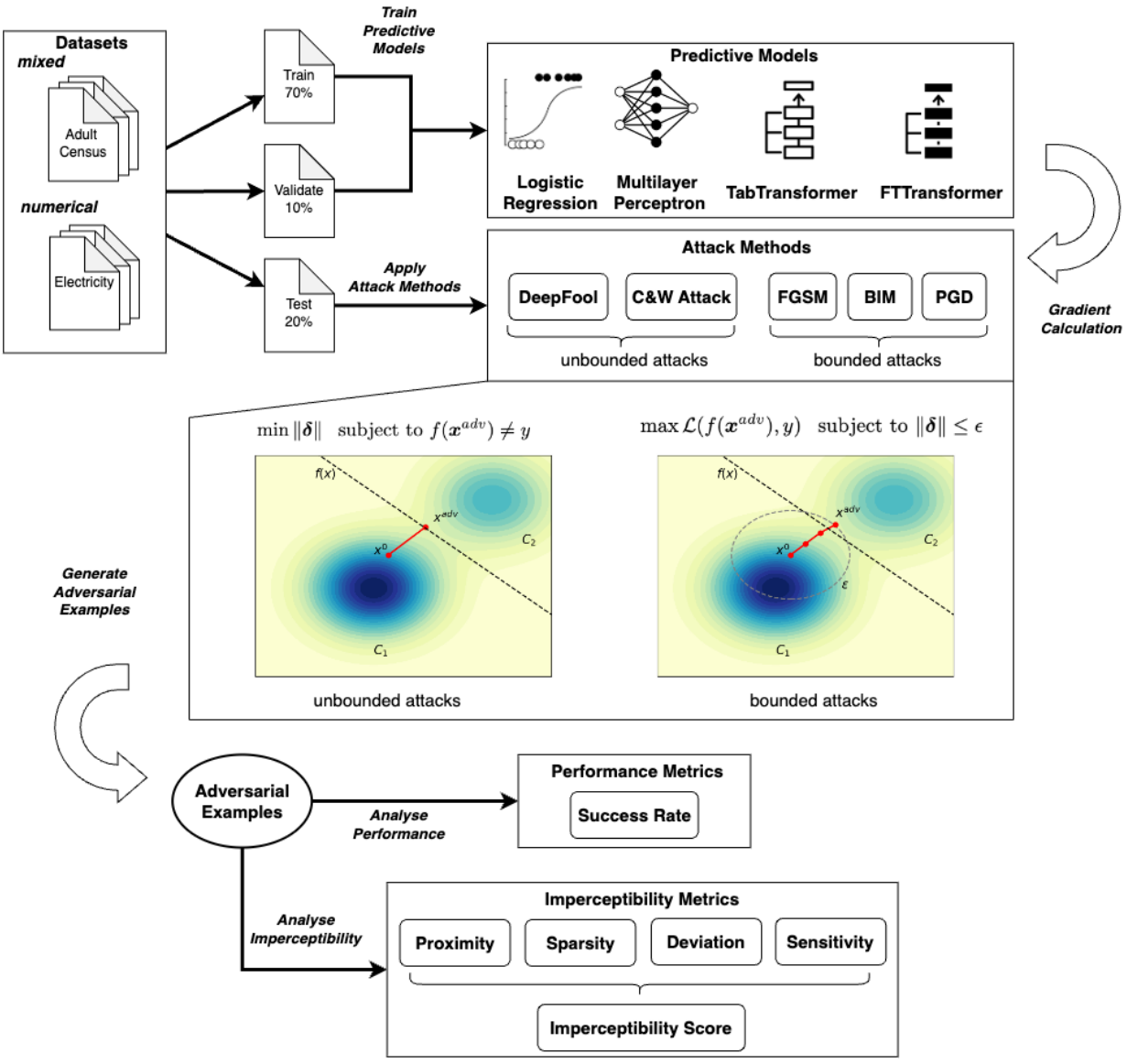


Figure 2: Overview of the evaluation framework for benchmarking adversarial attacks on tabular data.

3.1. Datasets

When selecting a dataset for the benchmark, several key criteria must be considered to ensure its suitability and effectiveness. Firstly, the dataset should be appropriate for classification tasks and include at least two classes to support the evaluation of binary classification. Secondly, to maintain computational feasibility and prevent excessive dimensionality, the total number of features, including those generated by one-hot encoding of categorical features, should not exceed 5000. Moreover, the dataset should be representative of real-world scenarios, with diverse instances to effectively assess model robustness and generalisation. Table 2 presents the data profiles of all 11 datasets used in the benchmark, summarising their key characteristics.[1]

3.2. Adversarial Attacks

We frame our adversarial attack problem based on the threat model taxonomy proposed in [17]. Specifically, we focus on the white-box attack scenario, where the attack algorithms

[1]WineQuality (White) and WineQuality (Red) originate from the same dataset but are treated as two separate datasets in this benchmark.

Table 2: Data profiles of the 11 datasets used in the benchmark, including the total number of instances (N_{total}), instances for training (N_{train}), validation ($N_{validate}$), and adversarial perturbation (N_{test}), as well as the number of numerical features (x_{num}), categorical features (x_{cat}), one-hot encoded (categorical) features ($x_{encoded}$), and the total number of features (x_{total}).

Dataset	N_{total}	N_{train}	$N_{validate}$	N_{test}	x_{num}	x_{cat}	$x_{encoded}$	x_{total}
Adult	32561	22792	3256	6513	6	8	99	105
Electricity	45312	31717	4532	9063	7	1	7	14
COMPAS	16644	11650	1665	3329	8	8	50	58
Higgs	1000000	700000	100000	200000	28	0	0	28
house_16H	22784	15948	2279	4557	16	0	0	16
jm1	10885	7619	1089	2177	21	0	0	21
BreastCancer	569	398	57	114	30	0	0	30
WineQuality-White	4898	3428	490	980	11	0	0	11
WineQuality-Red	1599	1119	160	320	11	0	0	11
phoneme	5404	3782	541	1081	5	0	0	5
MiniBooNE	130064	91044	13007	26013	50	0	0	50

have access to both the dataset and the predictive model's configuration. The objective of the adversarial attack is to deceive the predictive model's predictions. Notably, our benchmark does not enforce targeted misclassification, and all experiments generate untargeted attacks.

In practice, the choice of attack methods for structured data is based on various considerations such as their efficacy, efficiency, and complexity. However, it should be noted that a systematic review or benchmark on adversarial attacks on structured data is currently lacking. Consequently, the approach taken has been to explore existing attack benchmarks on images and then screen the attacks to identify those that can be extended to tabular data. Furthermore, most existing attack methods are designed for white-box settings, which allow the attacker to generate highly effective and efficient adversarial examples. Taking into account these factors, our selection of attack methods is guided by the following criteria:

1. The selected attack methods should be applicable to tabular data.
2. The selected attack methods should be designed for white-box attack.

Based on these criteria, we have identified five attack methods that are vary in complexity and approach and have demonstrated high effectiveness and efficiency in the field of computer vision. These include three unbounded attacks—FGSM, BIM, and PGD, and two bounded attacks—DeepFool and C&W. In addition, we include Gaussian noise as a baseline to evaluate the impact of random noise on model performance, providing a reference point to assess the effectiveness of intentional perturbation techniques in the selected attack methods compared to simple noise injection.

3.3. Predictive Models

When selecting machine learning models for adversarial attack benchmarking on tabular datasets, three crucial criteria guide our choices: *diversity* (spanning classical and modern architectures), *interpretability* (balancing transparency with complexity), and *performance* (ensuring competitive accuracy). Based on these principles, we evaluate four representative models:

1. **Logistic Regression (LR):** A simple, interpretable linear baseline that establishes performance lower bounds and vulnerability benchmarks.

2. **Multilayer Perceptron (MLP):** A foundational neural network adept at capturing nonlinear patterns, offering a mid-complexity comparison point.
3. **TabTransformer [35]:** An attention-based model that processes tabular features via transformer layers, leveraging contextual relationships among features.
4. **FTTransformer [36]:** A transformer-based architecture that tokenizes numerical and categorical features, enabling unified processing through self-attention mechanisms.

3.4. Evaluation Metrics

In the context of machine learning, especially in adversarial settings, a successful attack occurs when the model's predictions are manipulated or altered to produce incorrect or unintended results. From common practice, the attack success rate (Eq. 9) is used to measure the effectiveness of an adversarial attack. It represents the percentage of instances $\boldsymbol{x}_i$ in a dataset for which the attack is successful in misleading the predictive model or causing misclassifications.

$$\text{Attack Success Rate} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x}_i^{adv} \neq y_i) \tag{9}$$

Considering the characteristics of tabular data, besides the traditional effectiveness metrics of adversarial attacks, the metrics of imperceptibility are also be included in the benchmark. According to our previous research, four quantitative metrics of imperceptibility can be employed, including proximity, sparsity, sensitivity and deviation.

3.4.1. Proximity

- *Definition:* Average distance between inputs and generated adversarial examples.
- *Purpose:* Measures how close the adversarial examples are to the original inputs in terms of Euclidean distance.
- *Considerations:* Lower proximity values indicate better imperceptibility.

$$\ell_p(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_p = \begin{cases} \left(\sum_{i=1}^{n}(x_i^{adv} - x_i)^p\right)^{1/p}, & p \in \{1, 2\} \\ \sup_n |x_n^{adv} - x_n|, & p \to \infty \end{cases} \tag{10}$$

3.4.2. Sparsity

- *Definition:* Uses average ℓ_0 distance to calculate the average number of perturbed features.
- *Purpose:* Quantifies how many features are modified on average in the adversarial examples.
- *Considerations:* Lower sparsity values indicate better imperceptibility.

$$Spa(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \ell_0(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \sum_{i=1}^{n} \mathbb{1}(x_i^{adv} - x_i) \tag{11}$$

3.4.3. Deviation

- *Definition:* Uses Mahalanobis Distance (MD) to calculate the distance between generated adversarial examples and the original datasets. Thresholds are determined using statistical methods to identify outliers, and the total outlier rate is calculated.
- *Purpose:* Captures how much the adversarial examples deviate from the normal data distribution.
- *Considerations:* Lower MD values and Moutlier rates indicate better imperceptibility.

$$\text{MD}(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \sqrt{(\boldsymbol{x}^{adv} - \boldsymbol{x})V^{-1}(\boldsymbol{x}^{adv} - \boldsymbol{x})^T} \tag{12}$$

MD is a measure of the distance between a point and a distribution, taking into account the covariance structure of the data. In this context, a threshold is required to determine which data points are considered outliers based on their MD. We calculates the critical value for the MD using the chi-squared distribution. Mathematically, the

critical value for the MD is determined by first calculating the chi-squared critical value $\mathcal{X}^2_{\alpha,d}$ corresponding to the desired significance level α and the degrees of freedom d, where d is the number of dimensions or features in the data. This critical value represents the boundary beyond which a certain proportion α of the chi-squared distribution lies. The Mahalanobis distance threshold t is then computed by taking the square root of the chi-squared critical value. If MD value of a data point exceeds the threshold, it suggests that the point is farther away from the center of the distribution than expected and we consider it as outlier.

$$\text{Outlier Rate} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\text{MD}(\boldsymbol{x}^{adv}, \boldsymbol{x}) > t), \text{where } t = \sqrt{\mathcal{X}^2_{\alpha,d}} \tag{13}$$

3.4.4. Sensitivity

- *Definition:* A metric to check if sensitive features are changed, based on distance metrics.
- *Purpose:* Focuses on the impact of the attack on sensitive features, which may be critical for certain applications.
- *Considerations:* The metric should be sensitive to changes in important features.

$$\begin{aligned} \text{SDV}(x_i) &= \sqrt{\frac{\sum^{m}(x_i - \bar{x}_i)^2}{m}} \\ \text{SEN}(\boldsymbol{x}, \boldsymbol{x}^{adv}) &= \sum_{i=1}^{n} \frac{|x_i^{adv} - x_i|}{\text{SDV}(x_i)} \end{aligned} \tag{14}$$

4. Evaluation

Our evaluation methodology aims to address three primary research questions related to adversarial attacks on tabular data as follows.

- **RQ1.** *How effective are the evaluated adversarial attack algorithms on tabular data?* In our benchmark, this is measured by the success rates of individual adversarial attack methods in deceiving target model's classification.

- **RQ2.** *How imperceptible are these adversarial attack algorithms on tabular data?* Our benchmark evaluates the imperceptibility of individual adversarial attack methods on tabular data in terms of the four key quantitative properties specified in Section 3.4.
  - **RQ2.1 (Sparsity)**: How many features are modified in the adversarial examples?
  - **RQ2.2 (Proximity)**: How close are the adversarial examples to the original samples in the feature space?
  - **RQ2.3 (Deviation)**: How significantly do the modified features differ from their original values?
  - **RQ2.4 (Sensitivity)**: How much do perturbations respect narrow-guard feature perturbation?
- **RQ3.** *Whether and how can the evaluated algorithms achieve a balance between both imperceptibility and effectiveness?* Based on the benchmark evaluation, we conduct a trade-off analysis of different adversarial attack methods to identify those that strike the best balance between both effectiveness and imperceptibility.

4.1. Experiment Setup

4.1.1. Datasets and Preprocessing

Following the dataset selection criteria outlined in Section 3.1, we implement a standardised preprocessing pipeline for all benchmark datasets. Each dataset is first partitioned using stratified sampling to maintain class distributions, allocating 70% for training, 10% for validation, and 20% for testing and adversarial evaluation. We fix the random seed (42) throughout this process to ensure reproducibility.

For feature engineering, we remove constant and duplicate features, then address missing values through median imputation for numerical features and mode imputation for categorical variables. Categorical features are transformed via one-hot encoding, while all numerical features are normalised to the $[0, 1]$ range using min-max scaling. This consistent preprocessing approach ensures fair comparison across models while preserving each dataset's inherent characteristics documented in Table 2.

4.1.2. Models

Based on the methods in previous section, we select four predictive models in our research. LR is an representation of transparent model. Other three deep learning models are MLP, TabTransformer and FT-Transformer, which are hard for human beings to explain them. All models were trained for 20 epochs with a batch size of 512, optimised via Adam (learning rate=1e-3) using cross-entropy loss. For regularisation, we applied dropout (p=0.2) to MLP and both transformer models. The MLP uses ReLU activations, while the transformers employ ReLU in their feed-forward components. Transformer-specific configurations include 8 attention heads per layer and 6 stacked layers, with all embeddings dimensioned to match the MLP's hidden layer widths (64 units) for fair comparison.

4.1.3. Adversarial Attack Configuration

To thoroughly evaluate model robustness, employs five white-box attack methods, including the foundational FGSM attack along with its iterative variants (BIM and PGD), plus optimisation-based approaches (DeepFool and C&W). We test these across a carefully designed spectrum of perturbation budgets ($\epsilon \in \{0, 0.01, 0.03, 0.05, 0.1, 0.3, 0.5, 1\}$), where $\epsilon = 0$) serves as our natural accuracy baseline. This graduated approach allows us to precisely characterise how different models degrade under increasing adversarial pressure.

To provide a thorough evaluation, each attack algorithm is configured with specific hyperparameters tailored to its design. The FGSM attack represents our simplest case, applying the full ϵ perturbation in a single gradient step. This provides a fundamental benchmark against which we compare more sophisticated methods. For iterative attacks (BIM/PGD), we implement a relative step size strategy where each of the $T = 10$ iterations applies a perturbation of magnitude ϵ/T, ensuring controlled approach to the total budget.

The DeepFool attack employs an iterative boundary-crossing approach with carefully calibrated parameters: a maximum of 50 iterations ensures convergence while maintaining computational efficiency, and a 2% overshoot factor (0.02) guarantees reliable crossing of decision boundaries. We configure it to evaluate 10 candidate classes per iteration and operate directly on model logits, providing precise gradient information for minimal adversarial

perturbations.

For the C&W attack, we implement a rigorous optimisation process controlled by three key parameters: (1) 10 binary search steps to optimally scale the penalty constant, (2) an initial constant value of 0.001 for gradual constraint adjustment, and (3) zero-confidence ($\kappa = 0$) attacks to produce minimally-perturbed adversarial examples. Each binary search phase executes 10 optimisation steps with a learning rate of 0.1, balancing attack success rate with computational cost.

4.2. RQ1: How effective are the evaluated adversarial attack algorithms on tabular data?

We first evaluate model accuracy to understand the performance of predictive models before exposing them to adversarial attacks. If the model's accuracy is too low, such as below 60%, it may be easily deceived even without attacks, making further evaluation unnecessary. Table 3 presents the accuracy results for the four selected models across 11 datasets. As shown, three deep learning models generally outperform Logistic Regression (LR) on most datasets; whereas, for certain datasets, including Adult, jm1, and COMPAS, LR achieves similar performance. Overall, the models demonstrate sufficient accuracy, with all exceeding 63%, making them suitable for further adversarial testing.

Table 3: Model accuracy of four predictive models across 11 datasets.

Dataset	LR	MLP	TabTransformer	FTTransformer
Adult	0.834	0.8337	0.8328	0.799
BreastCancer	0.9386	0.9737	0.9035	0.9737
Compas	0.6654	0.6738	0.7053	0.6858
Electricity	0.6607	0.7635	0.762	0.7712
Higgs	0.6366	0.7234	0.6951	0.7296
MiniBooNE	0.7724	0.8372	0.8497	0.8402
WineQuality-Red	0.7219	0.7344	0.7281	0.7344
WineQuality-White	0.6745	0.7469	0.7316	0.752
house_16H	0.7029	0.8578	0.8251	0.8466
jm1	0.8075	0.8098	0.8066	0.8107
phoneme	0.7095	0.7872	0.7882	0.8002

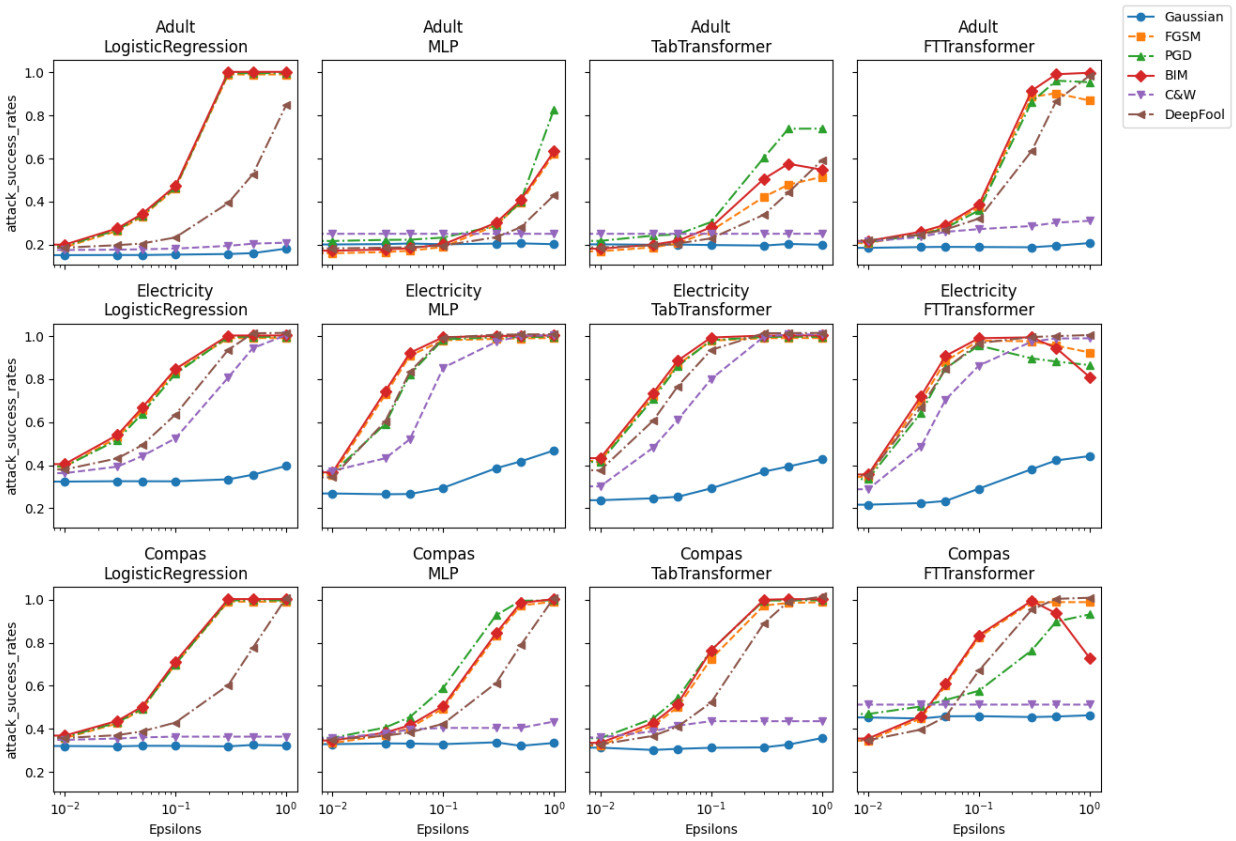


Figure 3: Attack success rate of evaluated attack methods on all three *mixed* datasets and four ML models.

Mixed Datasets. Figure 3 indicates the Electricity dataset demonstrates unique characteristics compared to other mixed datasets. Here, all attack methods achieve notably high success rates across model architectures, with even the typically underperforming C&W attack reaching nearly 100% success on LR, MLP, and TabTransformer. This suggests that the feature distribution or model decision boundaries for this dataset may be particularly susceptible to adversarial manipulation. Interestingly, while other datasets show clear differentiation between attack types, Electricity exhibits more uniform patterns across attack methods.

In contrast, the Adult and Compas datasets show clearer distinctions between attack effectiveness. The ℓ_∞-based attacks (FGSM, PGD, and BIM) consistently outperform ℓ_w-based attacks (C&W and DeepFool) on these datasets. This performance gap suggests that the bounded perturbation approach of ℓ_∞ attacks is particularly effective for these mixed-

type datasets, possibly due to their ability to make targeted changes to critical features without being constrained by the ℓ_2 norm's emphasis on overall perturbation magnitude.

From a model architecture perspective, Transformer-based models demonstrate greater robustness against adversarial attacks compared to traditional approaches. Both TabTransformer and FTTransformer require higher epsilon values to achieve the same attack success rates as seen with LR and MLP, particularly on the Adult dataset. This suggests that the attention mechanisms and deeper architectural features of transformers may provide some inherent robustness to adversarial perturbations when handling mixed tabular data.

Numerical Datasets. As shown in Figure 4, 5 and 6, our analysis of eight numerical datasets reveals more consistent patterns compared to mixed datasets, though with several dataset-specific characteristics worth noting.

The Higgs and house_16H datasets (Figure 4) exhibit remarkably uniform vulnerability to ℓ_∞-based attacks across model architectures. On these datasets, FGSM, PGD, and BIM produce nearly identical attack success curves, suggesting that the simpler FGSM approach may be sufficient for compromising models trained on these data distributions. The jm1

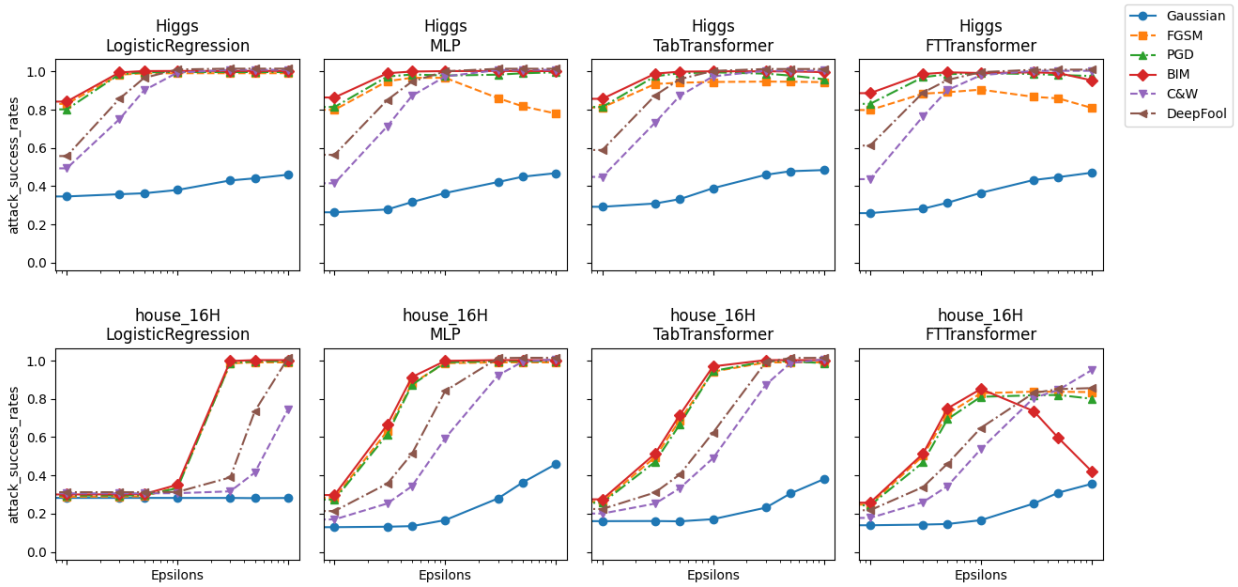


Figure 4: Attack success rate of evaluated attack methods on two (out of eight) *numerical* datasets and four ML models.

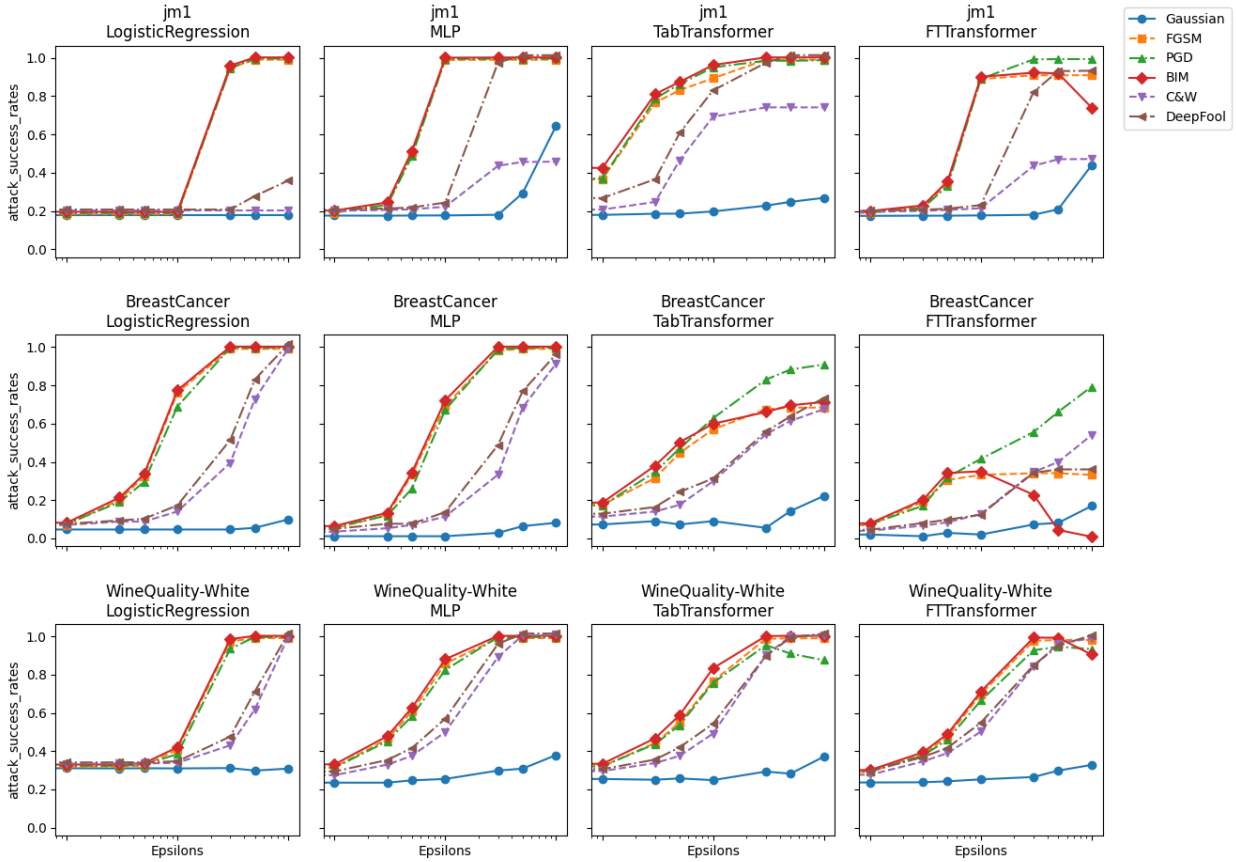


Figure 5: (Cont.) Attack success rate of evaluated attack methods on another three (out of eight) *numerical* datasets and four ML models.

dataset demonstrates a fascinating threshold phenomenon where attack success suddenly jumps from baseline to nearly perfect across multiple attack methods and models. This sharp transition suggests a critical vulnerability point in the feature space where slight perturbations beyond a specific threshold completely undermine model performance.

The BreastCancer dataset (Figure 5) provides perhaps the most diverse response to different attack methods. Here, PGD shows superior performance when targeting transformer-based models, while the FTTransformer exhibits unusual non-monotonic vulnerability patterns where attack success sometimes decreases at higher epsilon values. This counter-intuitive behaviour suggests potential overfitting of attack algorithms to specific decision boundary regions or gradient masking effects in the transformer architecture.

Both WineQuality datasets (Red and White, Figure 5 and 6) show that ℓ_∞-based at-

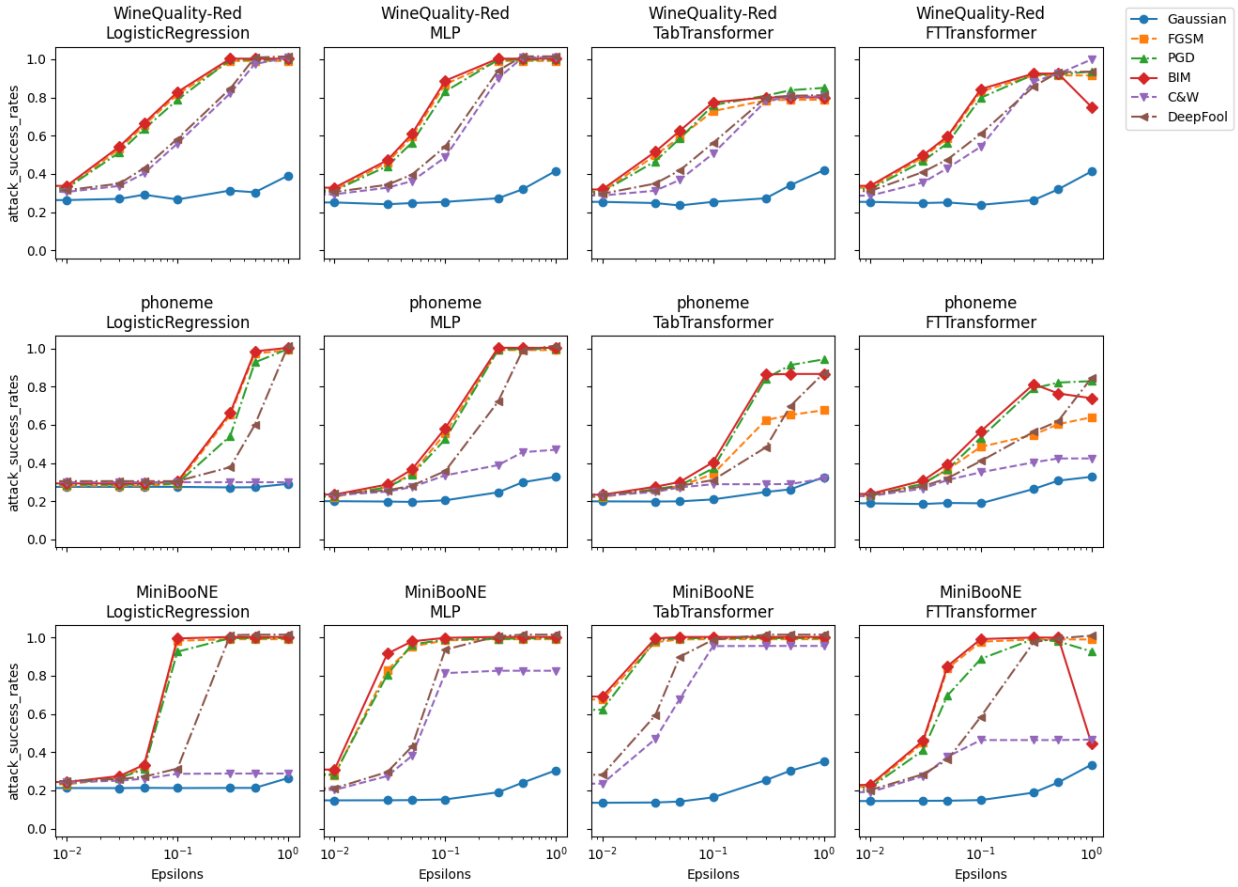


Figure 6: (Cont.) Attack success rate of evaluated attack methods on the remaining three (out of eight) *numerical* datasets and four ML models.

tacks require substantially lower epsilon values to achieve high success rates compared to ℓ_2-based approaches. DeepFool eventually reaches comparable performance but demands significantly higher perturbation budgets, making it less efficient from an adversarial perspective. The phoneme and MiniBooNE datasets (Figure 6) further confirm the superiority of ℓ_∞-based attacks, with all three methods (FGSM, PGD, BIM) demonstrating nearly identical performance trajectories.

From an architectural standpoint, LR models consistently demonstrate the highest vulnerability across numerical datasets, often exhibiting sharp threshold effects where attack success rates increase dramatically at specific epsilon values. This suggests that the linear decision boundaries of logistic models may be easier to exploit with minimal perturbations. In contrast, MLP and transformer models typically require larger epsilon values before showing

significant compromise, though the specific patterns vary by dataset.

The C&W attack shows highly inconsistent performance across numerical datasets, sometimes barely outperforming the random noise baseline while occasionally achieving competitive results on specific model-dataset combinations. This variability suggests that C&W's effectiveness is highly dependent on the specific characteristics of the dataset and model architecture, making it a less reliable general-purpose attack for numerical tabular data.

Overall, our findings on numerical datasets indicate that ℓ_∞-based attacks provide the most consistent and efficient approach for compromising tabular models, with FGSM often performing similarly to the more computationally intensive PGD and BIM methods.

4.3. RQ2: How imperceptible are these adversarial attack algorithms on tabular data?

Based on our analysis of attack success rates across varying ϵ values, we establish a systematic approach for selecting optimal attack budgets. For each experimental setting, we identify the value at which attack success rates first reach a plateau (the stationary point of the curve), beyond which further increases in ϵ yield negligible performance improvements. The specific attack budgets selected through this methodology are detailed in Appendix Appendix A. However, direct comparisons across different models and datasets would be methodologically unsound, as optimal ϵ values vary significantly between these contexts. To ensure fair and consistent benchmarking, we address this variation by identifying the most frequently occurring ϵ value for each attack method across all tested models and datasets, as presented in Table 4. These representative ϵ values serve as our standardised benchmark parameters for subsequent comparative analyses in **RQ2** and **RQ3**.

Table 4: Standardised ϵ value settings for each attack method used in the analysis of **RQ2** and **RQ3**.

Attacks	Gaussian	FGSM	BIM	PGD	C&W	DeepFool
ϵ	1	0.3	0.3	0.3	1	1

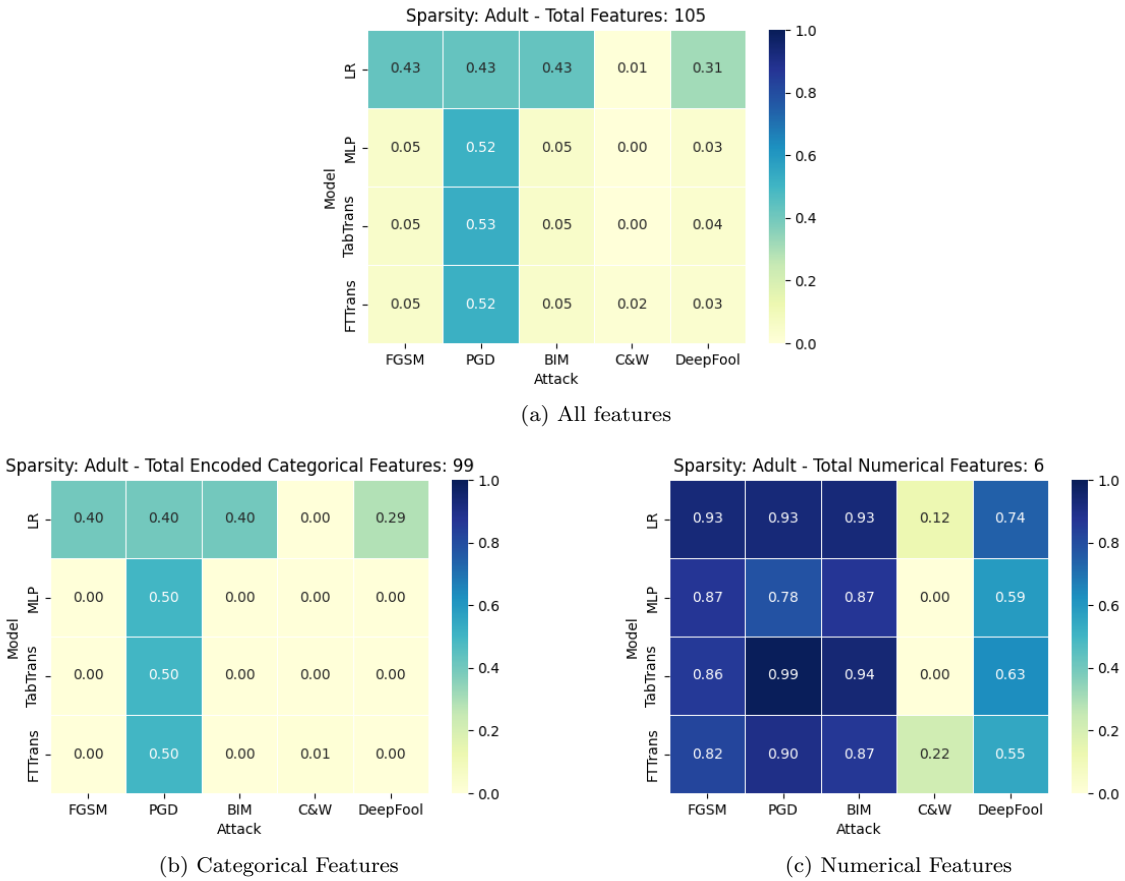
4.3.1. RQ2.1: How many features are modified in the adversarial examples?

Our analysis of sparsity patterns reveals distinct behavioural characteristics among adversarial attack methods while highlighting the influence of dataset dimensionality and model

architecture. The sparsity measure quantifies the proportion of features modified in adversarial examples, with higher values indicating more features being perturbed.

Sparsity – Mixed Datasets. Our analysis of sparsity patterns in mixed datasets (Figures 7, 8 and 9) reveals complex interaction effects between feature types, attack algorithms, and model architectures, providing critical insights into the selective vulnerability of different feature categories to adversarial perturbation.

In datasets with categorical feature dominance like Adult (105 total features, 99 categorical, Figure 7) and Compas (58 total features, 50 categorical, Figure 8), an articulate divide in treatment between categorical and numerical features is evident across attack methods. FGSM, BIM, C&W, and DeepFool demonstrate a strong numerical feature bias when attacking neural network models (MLP, TabTransformer, FTTransformer), with sparsity rates

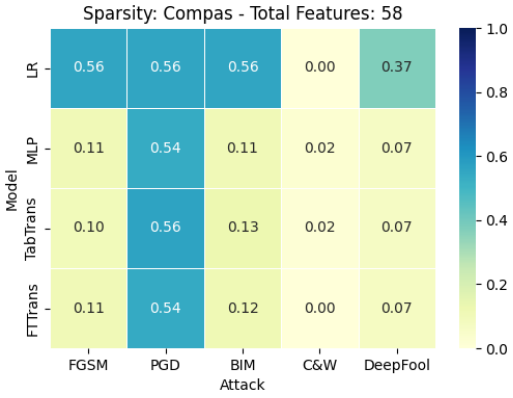


(a) All features

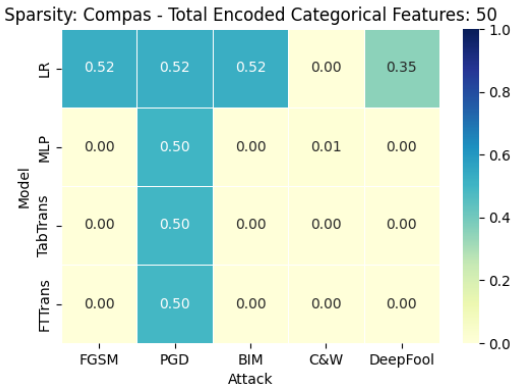
(b) Categorical Features

(c) Numerical Features

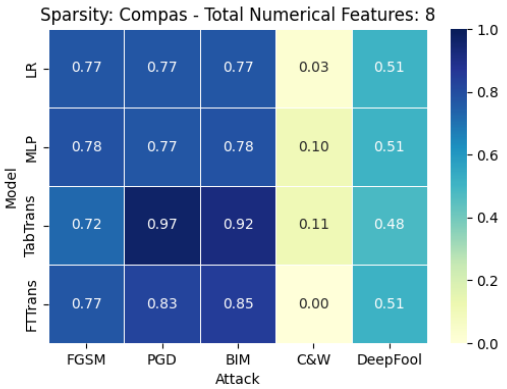
Figure 7: Sparsity results of five evaluated attack methods and four ML models on the Adult dataset.



(a) All features



(b) Categorical Features



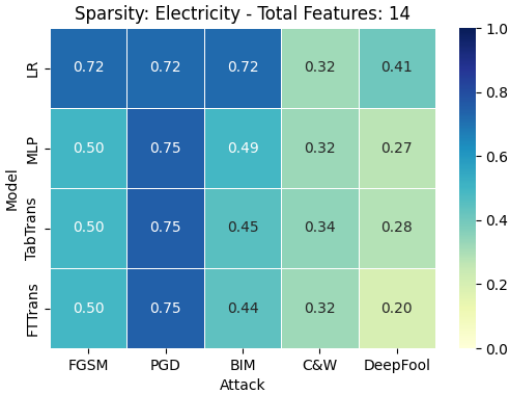
(c) Numerical Features

Figure 8: Sparsity results of evaluated attack methods and four ML models on the COMPAS dataset.

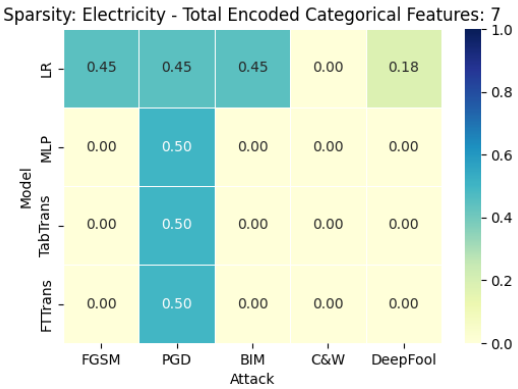
for categorical features near zero (0-1%) while maintaining high sparsity rates for numerical features (72-99%). This pronounced selectivity indicates these attacks algorithmically prioritise numerical features, effectively ignoring categorical dimensions despite their prevalence in the feature space.

PGD stands apart as the only attack capable of consistently modifying categorical features across all models and datasets, achieving approximately 50% sparsity on categorical features regardless of dataset composition. This unique capability suggests PGD's perturbation mechanism operates fundamentally differently from other ℓ_∞-based approaches, likely due to its projection mechanism that allows effective navigation of the discrete space represented by one-hot encoded categorical features.

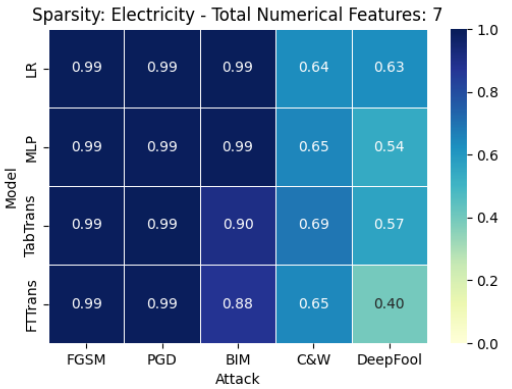
The LR model exhibits markedly different vulnerability patterns from neural architec-



(a) All features



(b) Categorical Features



(c) Numerical Features

Figure 9: Sparsity results of evaluated attack methods and four ML models on the Electricity dataset.

tures. When attacked by FGSM, PGD, and BIM, LR models show moderate categorical feature sparsity (40-52%) across all mixed datasets, suggesting these models encode information differently, making categorical features more susceptible to perturbation. This architectural effect is particularly evident in the Compas dataset, where categorical feature sparsity rates for LR (52%) significantly exceed those of neural networks (0%) when attacked by FGSM and BIM.

The Electricity dataset (14 total features with equal distribution of 7 categorical and 7 numerical features, Figure 9) offers a unique perspective on attack behaviour in balanced feature spaces. Here, ℓ_∞-based attacks maintain their numerical feature bias despite equal feature distribution, with FGSM and BIM achieving 99% sparsity on numerical features while leaving categorical features unmodified (0% sparsity) when attacking neural networks.

This persistent selectivity in a balanced feature environment further confirms the algorithmic preference of these attacks for continuous variables over discrete ones.

C&W attacks demonstrate consistently low overall sparsity rates across mixed datasets (0-2% on most models), but exhibit moderate performance on numerical features in the Electricity dataset (64-69%). This selective numerical targeting despite general low effectiveness highlights how ℓ_2-based attacks struggle with the mixed feature landscape of tabular data, particularly with one-hot encoded categorical variables.

DeepFool shows minimal categorical feature modification (0-0.4%) across all neural network models while achieving moderate sparsity on numerical features (40-74%), positioning it as even more numerically-focused than other attacks. This extreme preference for numerical features suggests DeepFool's gradient-based optimisation may be fundamentally incompatible with the discrete nature of categorical variables in tabular data.

These findings collectively demonstrate that with the exception of PGD, current adversarial attacks on tabular data exhibit a strong inherent bias toward perturbing numerical features while largely ignoring categorical dimensions, regardless of their prevalence in the feature space.

Sparsity – Numerical Datasets. For numerical datasets, as illustrated in Figure 10 and 11, we observe distinct patterns of feature perturbation across different attack methods. FGSM, PGD, and BIM consistently demonstrate high sparsity rates, modifying nearly all features of the original inputs (approximately 80-100% sparsity) across most models and datasets. This comprehensive modification approach persists regardless of feature dimensionality—from the low-dimensional phoneme dataset (5 features, Figure 11a) to the high-dimensional MiniBooNE dataset (50 features, Figure 11b). PGD exhibits particularly aggressive feature modification, achieving near-perfect sparsity (99-100%) in many configurations, especially with TabTransformer models.

In contrast, C&W attacks display highly selective and context-dependent behaviour. C&W sparsity rates range dramatically from 0% (no features modified) in extreme cases—such as the LR model on phoneme and jm1 (Figure 10c) datasets—to moderate rates (50-

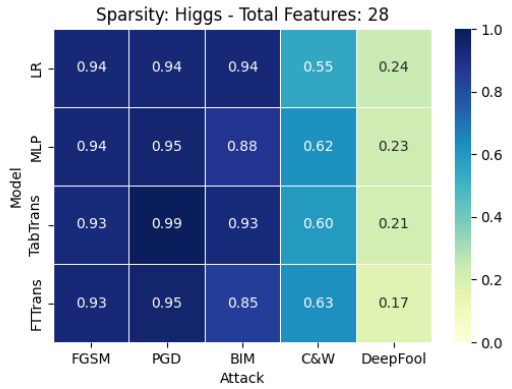
74%) in most other configurations. This variability suggests that C&W's optimisation approach identifies and targets only the most influential features for classification disruption. The C&W attack shows particularly low sparsity rates on the MiniBooNE dataset (5-28%) compared to smaller datasets, suggesting increased selectivity in higher-dimensional feature spaces.

DeepFool consistently occupies an intermediate position in feature modification strategy, with sparsity rates typically ranging from 17-80%. This attack shows its most selective behaviour on the Higgs dataset (17-24%, Figure 11b) and more moderate selectivity on other datasets. Interestingly, DeepFool's sparsity rates appear least affected by model architecture differences, maintaining relatively consistent modification patterns across different models for the same dataset.

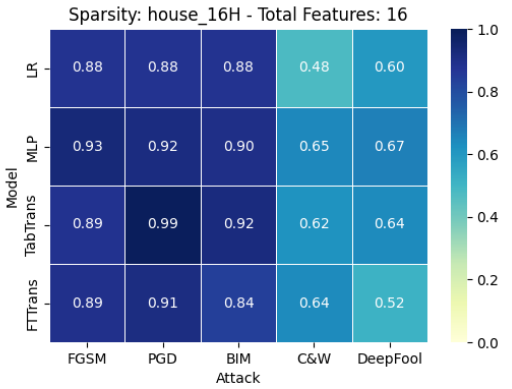
Model architecture significantly influences adversarial sparsity patterns. The LR model experiences the most extreme variations in feature modification, particularly with C&W and DeepFool attacks. TabTransformer shows notable variability in response to different attacks, while MLP models generally exhibit more consistent sparsity rates. The Breast-Cancer dataset (Figure 10d) uniquely demonstrates high sparsity rates (¿80%) for all attack methods and model combinations, suggesting that all features in this dataset are relevant to the classification task.

Feature dimensionality appears inversely correlated with sparsity rates for ℓ_2-based attacks, with C&W and DeepFool showing increased selectivity (lower sparsity) on larger datasets. This dimensional effect is particularly pronounced for C&W attacks, which modify only 5-28% of features on the 50-feature MiniBooNE dataset compared to 50-74% on smaller datasets like phoneme (5 features

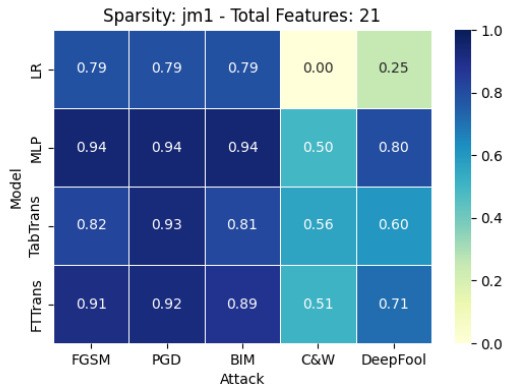
These findings reveal fundamental algorithmic differences in adversarial feature selection: ℓ_∞-based methods distribute perturbations broadly across the feature space, while ℓ_2-based methods strategically modify subsets of features.



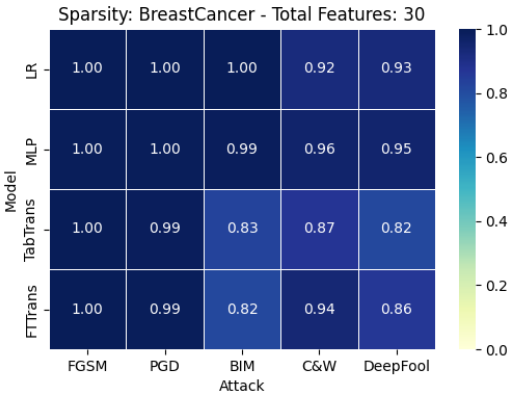
(a) Higgs



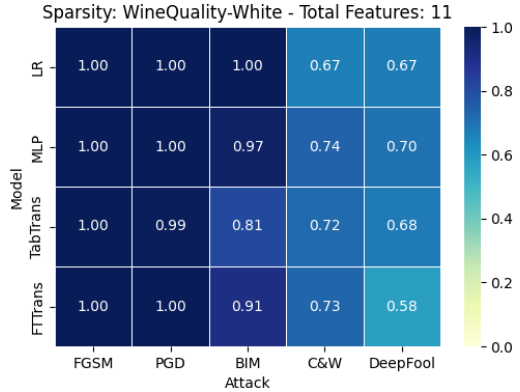
(b) house_16H



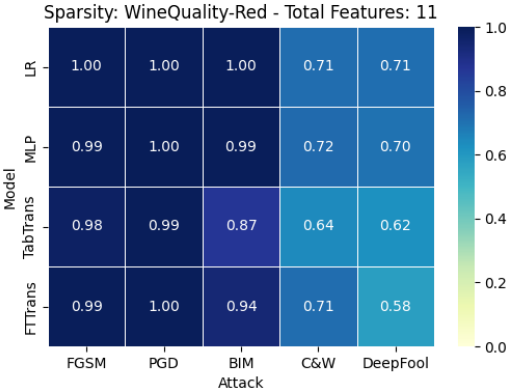
(c) jm1



(d) BreastCancer

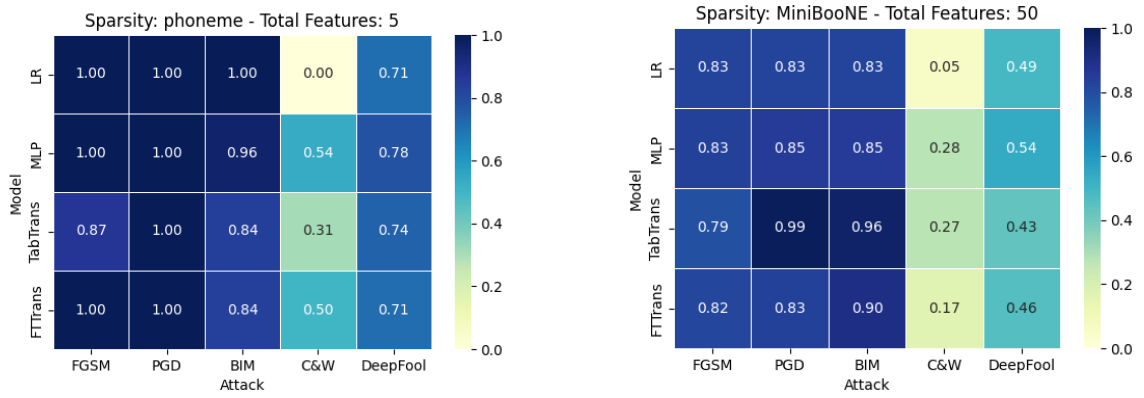


(e) WineQuality-White



(f) WineQuality-Red

Figure 10: Sparsity results of evaluated attack methods and four ML models on six (out of eight) *numerical* datasets.



(a) phoneme (b) MiniBooNE

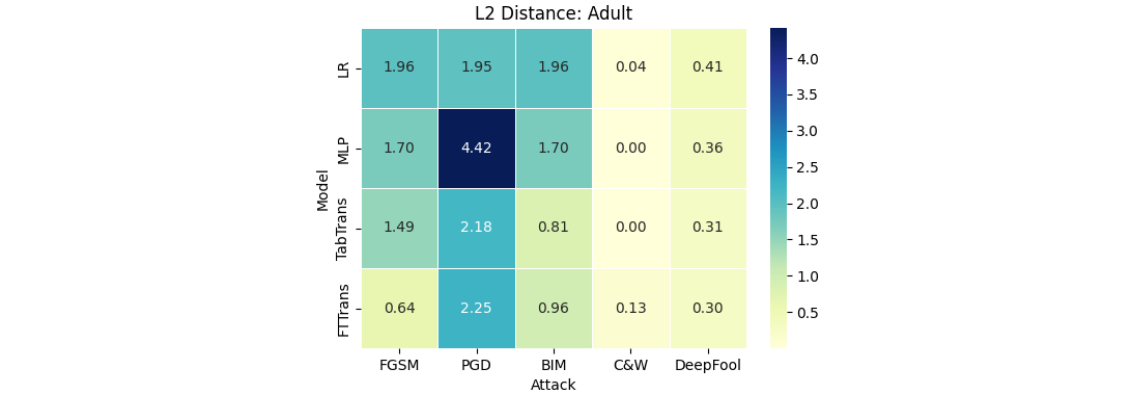
Figure 11: (Cont.) Sparsity results of evaluated attack methods and four ML models on the remaining two (out of eight) *numerical* datasets.

4.3.2. RQ2.2: How close are the adversarial examples to the original samples in the feature space?

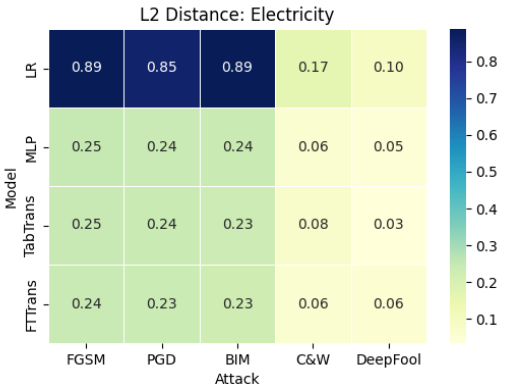
Our proximity analysis measures how close adversarial examples remain to their original samples in the feature space using ℓ_2 distance metrics. The heatmaps in Figures 12, 13 and 14 reveal distinct patterns across attack types, model architectures, and datasets that provide important insights into the imperceptibility of different adversarial approaches.

Proximity – Mixed dataset. The proximity results for mixed datasets (Adult, Electricity, and Compas) demonstrate clear differences between ℓ_2-based and ℓ_∞-based attack algorithms.

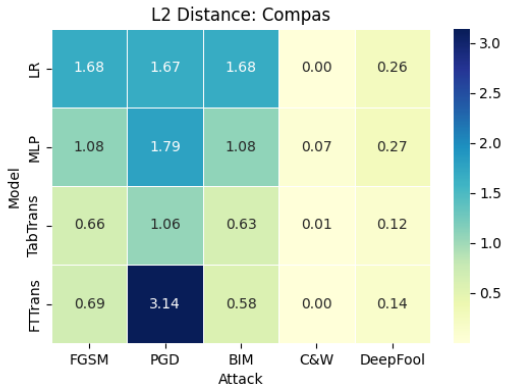
In the Adult dataset (Figure 12a), we observe a substantial proximity advantage for ℓ_2-based attacks across all model architectures. C&W consistently produces the closest adversarial examples to original samples, with remarkably low ℓ_2 distances ranging from 0.00 to 0.13, followed by DeepFool with distances between 0.30 and 0.41. In contrast, ℓ_∞-based attacks generate examples significantly further from originals, with distances typically ranging from 0.64 to 1.96. PGD is particularly notable for creating the most distant adversarial examples, reaching an exceptional ℓ_2 distance of 4.42 in the MLP model and 2.25 in the FTTransformer. This extreme difference suggests that PGD's optimisation approach, while effective at finding adversarial examples, sacrifices proximity considerably compared to other methods in this dataset.



(a) Adult



(b) Electricity



(c) Compas

Figure 12: Proximity results of evaluated attack methods and four ML models on all three *mixed* datasets.

The Electricity dataset (Figure 12b) exhibits lower ℓ_2 distances overall compared to Adult, but maintains the same pattern of ℓ_2-based attacks preserving significantly better proximity. The distance gap between attack types is most pronounced in LR models, where ℓ_∞-based attacks produce distances around 0.89, while C&W and DeepFool achieve distances of only 0.17 and 0.10 respectively. Interestingly, all neural network architectures demonstrate similar proximity values within each attack type, suggesting that model complexity has minimal impact on proximity in this dataset. This could indicate that the Electricity dataset's feature space permits finding closer adversarial examples regardless of model architecture.

For the Compas dataset (Figure 12c), we again observe the ℓ_2-based attacks' superior proximity performance, but with interesting model-specific variations. While C&W achieves remarkable proximity (distances between 0.00 and 0.07), PGD demonstrates extremely poor

proximity in certain model architectures, particularly with FTTransformer where it reaches 3.14 – the highest ℓ_2 distance in all mixed datasets. This suggests that certain combinations of dataset characteristics, model architectures, and attack algorithms can produce significantly outlying proximity behaviours.

Proximity – Numerical dataset. Our analysis of numerical datasets reveals both consistent patterns and intriguing variations in proximity metrics across the eight datasets examined.

The Higgs dataset (Figure 13a) demonstrates uniformly small ℓ_2 distances across all attack methods and models, with distances ranging from 0.01 to 0.25. While ℓ_2-based attacks still maintain better proximity (0.01-0.07) than ℓ_∞-based attacks (0.19-0.25), the difference is less pronounced than in mixed datasets. This suggests that the Higgs feature space may be structured in a way that adversarial examples can be found relatively close to original samples regardless of attack methodology.

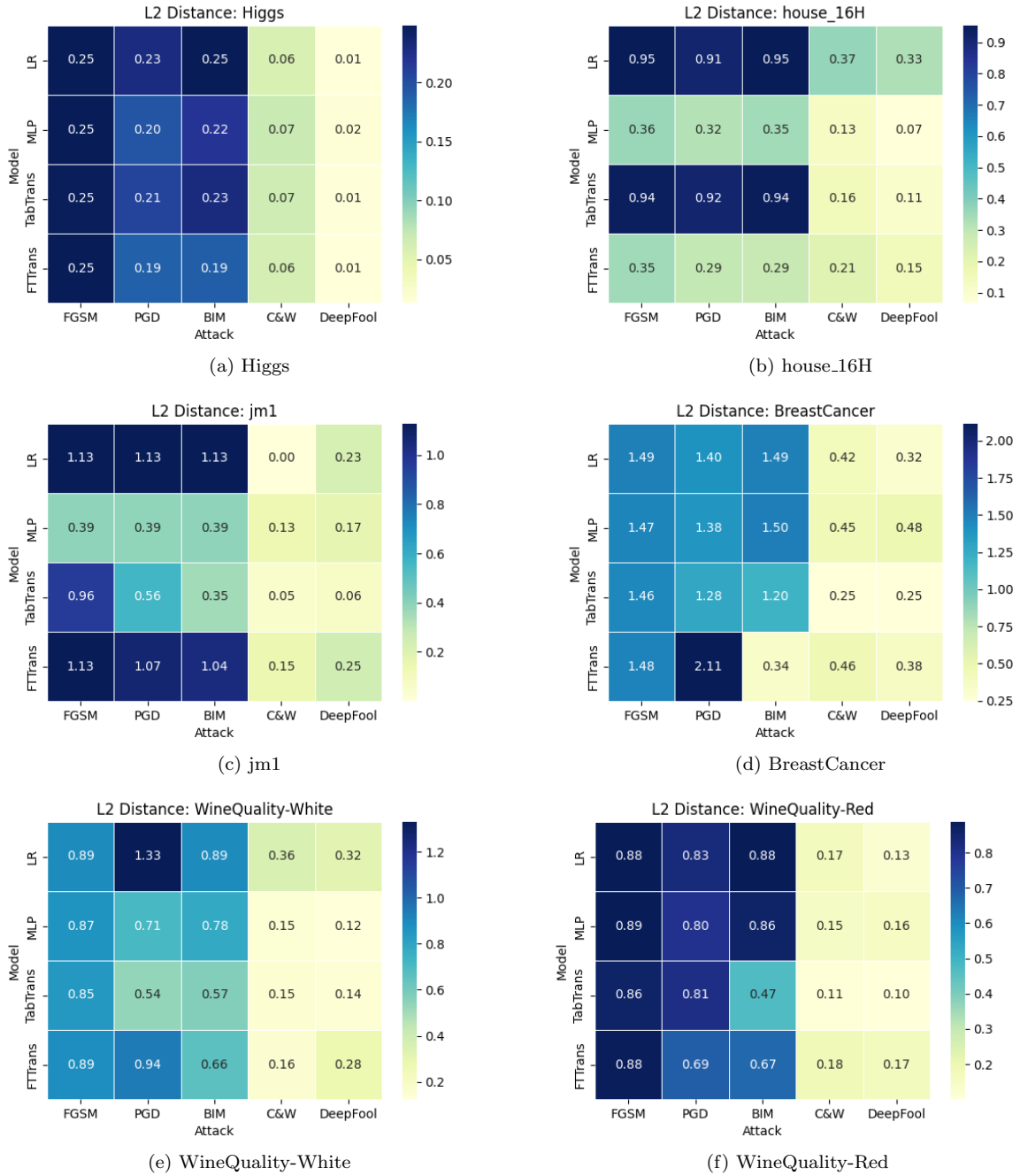
In contrast, the house_16H dataset (Figure 13b) shows much greater variability in proximity across model architectures. LR and TabTransformer models exhibit substantially higher ℓ_2 distances for ℓ_∞-based attacks (0.91-0.95) compared to MLP and FTTransformer models (0.29-0.36). This pattern suggests that model architecture plays a significant role in determining proximity characteristics for this dataset, potentially due to differences in decision boundary complexity.

The jm1 dataset (Figure 13c) reveals an interesting interaction between attack algorithms and model architectures. While ℓ_∞-based attacks maintain consistent ℓ_2 distances (1.13) for LR model, their proximity varies significantly for TabTransformer, with distances ranging from 0.35 for BIM to 0.96 for FGSM. This three-fold difference in proximity despite all attacks using the same ℓ_∞ constraint highlights how attack optimisation approaches interact differently with various model architectures.

The BreastCancer dataset (Figure 13d) demonstrates some of the highest ℓ_2 distances overall among numerical datasets, particularly for ℓ_∞-based attacks against LR around 1.49). PGD again produces exceptional distances in certain cases, reaching 2.11 with FTTransformer—more than six times the distance of BIM (0.34) on the same model. This

extreme variation suggests that PGD's optimisation approach can sometimes explode the ℓ_2 distance while pursuing adversarial examples under an ℓ_∞ constraint.

The WineQuality datasets (White and Red, Figure 13e and 13f) display moderate ℓ_2



(a) Higgs

(b) house_16H

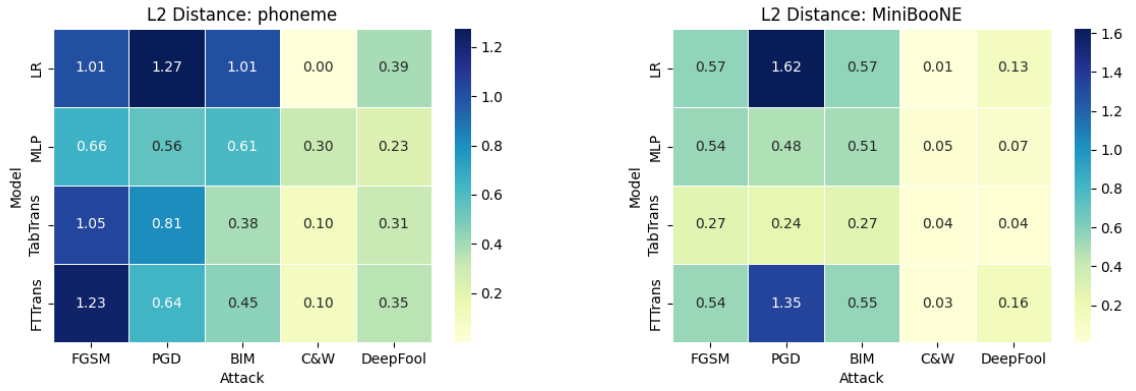
(c) jm1

(d) BreastCancer

(e) WineQuality-White

(f) WineQuality-Red

Figure 13: Proximity results of evaluated attack methods and four ML models on six (out of eight) *numerical* datasets.



(a) phoneme

(b) MiniBooNE

Figure 14: Proximity results of evaluated attack methods and four ML models on the remaining two (out of eight) *numerical* datasets.

distances for ℓ_∞-based attacks, but with notable attack-specific patterns. In WineQuality-White, PGD produces consistently higher ℓ_2 distances compared to FGSM and BIM across all model architectures, most pronounced in LR model (1.33 versus 0.89). However, this pattern is less evident in WineQuality-Red, suggesting dataset-specific interactions with attack algorithms.

The phoneme dataset (Figure 14a) exhibits significant model-dependent proximity variations. For LR, PGD again produces the highest ℓ_2 distance (1.27), while for TabTransformer, FGSM generates the most distant examples (1.05). C&W achieves exceptional proximity across all models, reaching zero for LR and 0.10 for transformer models. This highlights C&W's effectiveness at finding minimal-distance adversarial examples due to its direct ℓ_2 optimisation objective.

The MiniBooNE dataset (Figure 14b) continues the trend of PGD generating the most distant adversarial examples, with PGD producing ℓ_2 distances of 1.62 for LR and 1.35 for FTTransformer, significantly higher than other attack methods on the same models. These extreme values for PGD across multiple datasets suggest a fundamental characteristic of its optimisation approach that consistently sacrifices proximity for adversarial effectiveness.

From a model architecture perspective, LR generally exhibits the highest ℓ_2 distances across datasets, particularly for ℓ_∞-based attacks. This suggests that the linear decision

boundaries of logistic models may require larger perturbations to cross, resulting in less proximate adversarial examples. Transformer-based models show more variable proximity patterns across datasets, sometimes exhibiting better proximity than simpler models (as in jm1) and sometimes worse (as in BreastCancer with PGD).

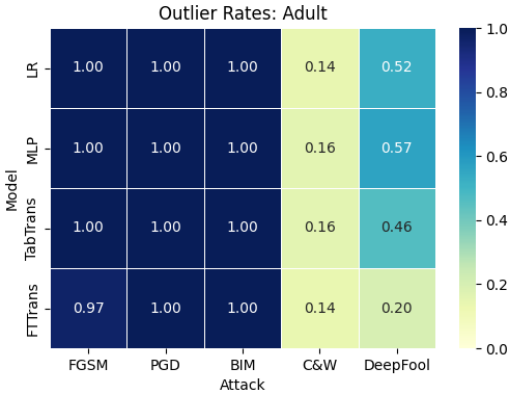
Overall, our proximity analysis confirms that ℓ_2-based attacks consistently generate adversarial examples that remain closer to original samples compared to ℓ_∞-based attacks. This aligns with their objective functions since L2 attacks directly optimise for minimal distance, while ℓ_∞ attacks focus on limiting the maximum change to any individual feature. Among ℓ_∞-based attacks, PGD frequently produces the most distant examples, suggesting its iterative process and strong adversarial optimisation may come at a significant cost to proximity.

4.3.3. RQ2.3: How significantly do the modified features differ from their original values?

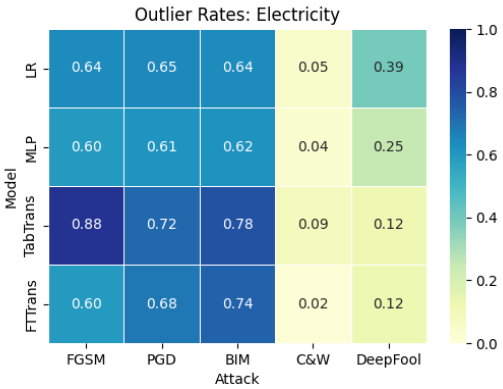
Our deviation analysis examines how significantly the adversarial examples differ from the original data distribution. The heatmaps presented in Figure 15 - 17 reveal clear patterns in the outlier rates produced by different attack algorithms across model architectures and datasets.

Deviation – Mixed dataset. On the Adult dataset (Figure 15a), we observe a striking dichotomy: all three ℓ_∞-based attacks (FGSM, PGD, and BIM) generate adversarial examples with outlier rates consistently at or near 100% across all model architectures. This indicates that these attacks produce perturbations that push samples substantially outside their original feature distributions. In contrast, C&W and DeepFool exhibit significantly lower outlier rates, with C&W ranging from 0.14 to 0.34 and DeepFool from 0.20 to 0.46, depending on the model architecture. This pattern suggests that ℓ_2-based attacks tend to preserve the original data distribution more effectively, potentially making them more difficult to detect through distribution-based defences.

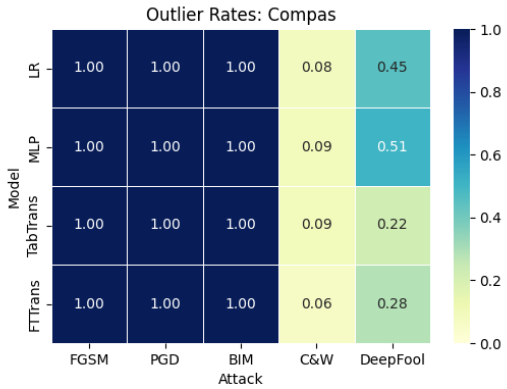
The Electricity dataset (Figure 15b) presents an interesting deviation from this pattern. While ℓ_∞-based attacks still generally produce higher outlier rates than ℓ_2-based approaches, the overall rates are lower compared to other mixed datasets. FGSM, PGD, and BIM



(a) Adult



(b) Electricity



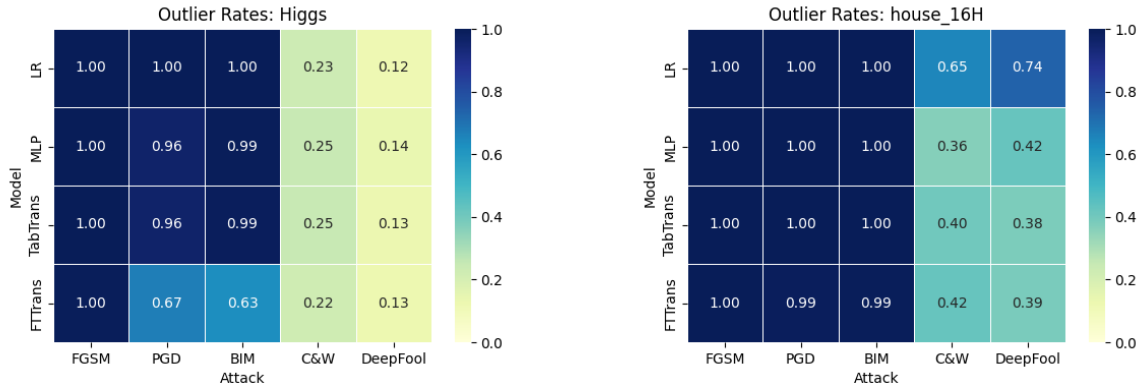
(c) Compas

Figure 15: Deviation results of evaluated attack methods and four ML models on all three *mixed* dataset.

generate outlier rates ranging from 0.60 to 0.88, with the TabTransformer model showing particular vulnerability to distribution shifts with outlier rates reaching 0.88 for FGSM. Both C&W and DeepFool maintain substantially lower outlier rates across all models (0.02-0.09 for C&W and 0.12-0.39 for DeepFool), reinforcing the trend that ℓ_2-based attacks tend to remain closer to the original data distribution.

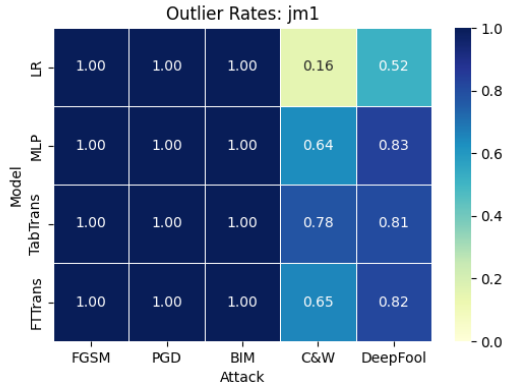
For the Compas dataset, we see the clearest demarcation between attack types. All ℓ_∞-based attacks generate perfect 1.00 outlier rates across all model architectures, indicating complete departure from the original feature distributions. Meanwhile, C&W consistently produces the lowest outlier rates (0.06-0.09), and DeepFool generates moderate outlier rates (0.22-0.51) that vary by model architecture. This stark contrast highlights the fundamentally different approaches to perturbation optimisation between ℓ_∞ and ℓ_2 norm constraints.

Deviation – Numerical dataset. The Higgs dataset (Figure 16a) displays the expected pattern where ℓ_∞-based attacks predominantly produce outlier rates of 1.00, with some exceptions for FTTransformer models where PGD and BIM show reduced rates of 0.67 and

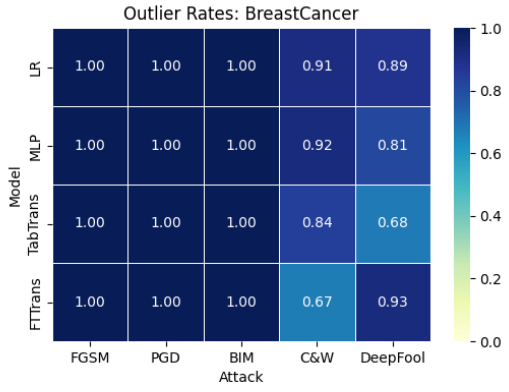


(a) Higgs

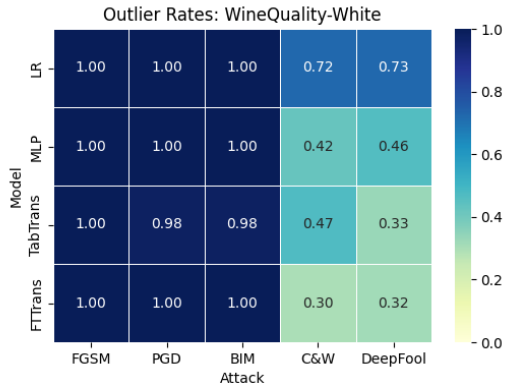
(b) house_16H



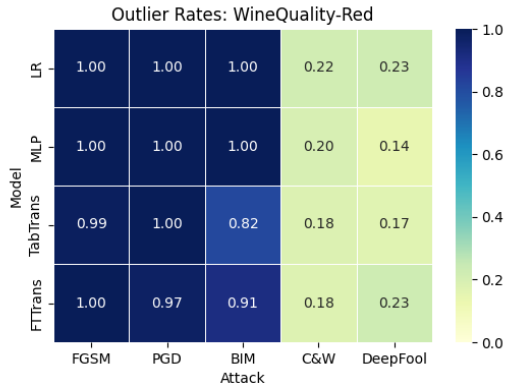
(c) jm1



(d) BreastCancer

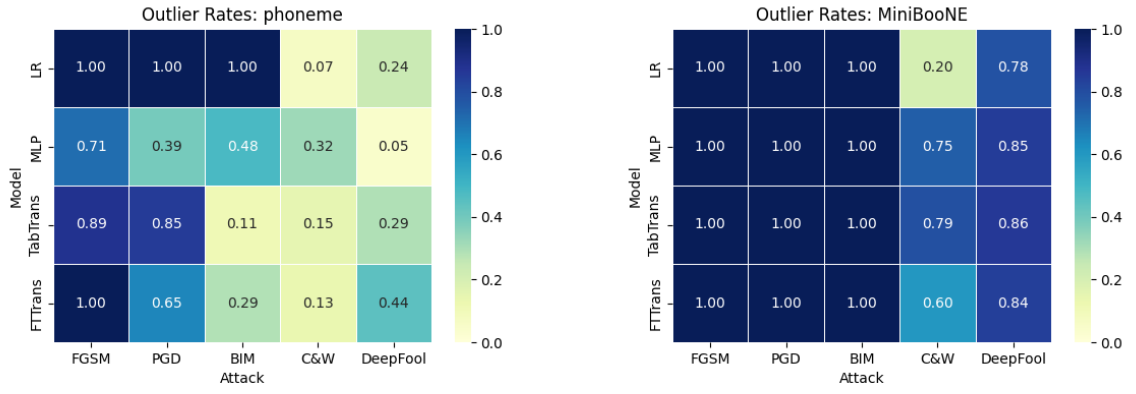


(e) WineQuality-White



(f) WineQuality-Red

Figure 16: Deviation results of evaluated attack methods and four ML models on six (out of eight) *numerical* datasets.



(a) phoneme (b) MiniBooNE

Figure 17: (Cont.) Deviation results of evaluated attack methods and four ML models on the remaining two (out of eight) *numerical* datasets.

0.63, respectively. ℓ_2-based attacks maintain significantly lower outlier rates (0.12-0.25), consistent with their tendency to preserve data distributions.

The house_16H dataset (Figure 16b) presents an interesting case where ℓ_∞-based attacks maintain near-perfect outlier rates across all models, but ℓ_2-based attacks also demonstrate relatively high outlier rates compared to other datasets. C&W produces outlier rates ranging from 0.36 to 0.65, while DeepFool generates rates from 0.38 to 0.74. This suggests that the feature distribution of house_16H may be particularly sensitive to perturbations, causing even ℓ_2-constrained modifications to push samples outside their original distributions.

The jm1 dataset (Figure 16c) exhibits uniformly high outlier rates for ℓ_∞-based attacks (1.00 across all models) and surprisingly high rates for ℓ_2-based attacks as well, with C&W reaching 0.16-0.78 and DeepFool achieving 0.52-0.83. This indicates that jm1's feature space may be particularly conducive to generating out-of-distribution samples regardless of the attacks employed.

The BreastCancer dataset (Figure 16d) stands out as an anomaly among numerical datasets, with all attack methods producing remarkably high outlier rates. Even C&W and DeepFool, which typically generate in-distribution samples, produce outlier rates ranging from 0.67 to 0.93. This suggests that the BreastCancer dataset may have a particularly compact or tightly clustered feature distribution where even small perturbations can push

samples beyond distribution boundaries.

The WineQuality datasets (White and Red, Figure 16e and 16f) display contrasting behaviors. WineQuality-White shows high outlier rates for ℓ_∞-based attacks (0.98-1.00) and moderate rates for ℓ_2-based attacks (0.30-0.73). In contrast, WineQuality-Red exhibits a wider range of outlier rates even for ℓ_∞-based attacks, with BIM producing rates as low as 0.82 on TabTransformer. This suggests that the Red variant may have a more dispersed feature distribution that can accommodate certain perturbations while remaining in-distribution.

The phoneme dataset (Figure 17a) reveals the most variable behaviour across models and attacks. While LR model remains highly susceptible to distribution shifts from ℓ_∞-based attacks (1.00 outlier rates), other models show surprising resistance. BIM produces outlier rates as low as 0.11 on TabTransformer, and even FGSM shows reduced effectiveness on MLP with a 0.71 outlier rate. ℓ_2-based attacks maintain their typical low outlier pattern, with rates as low as 0.05 for DeepFool on MLP.

Finally, the MiniBooNE dataset (Figure 17b) demonstrates consistently high outlier rates across all attack types and models. Even C&W, which typically preserves distribution characteristics, produces outlier rates from 0.20 to 0.79. This suggests that MiniBooNE may have a feature space where adversarial perturbations, regardless of norm constraints, readily push samples outside their original distributions.

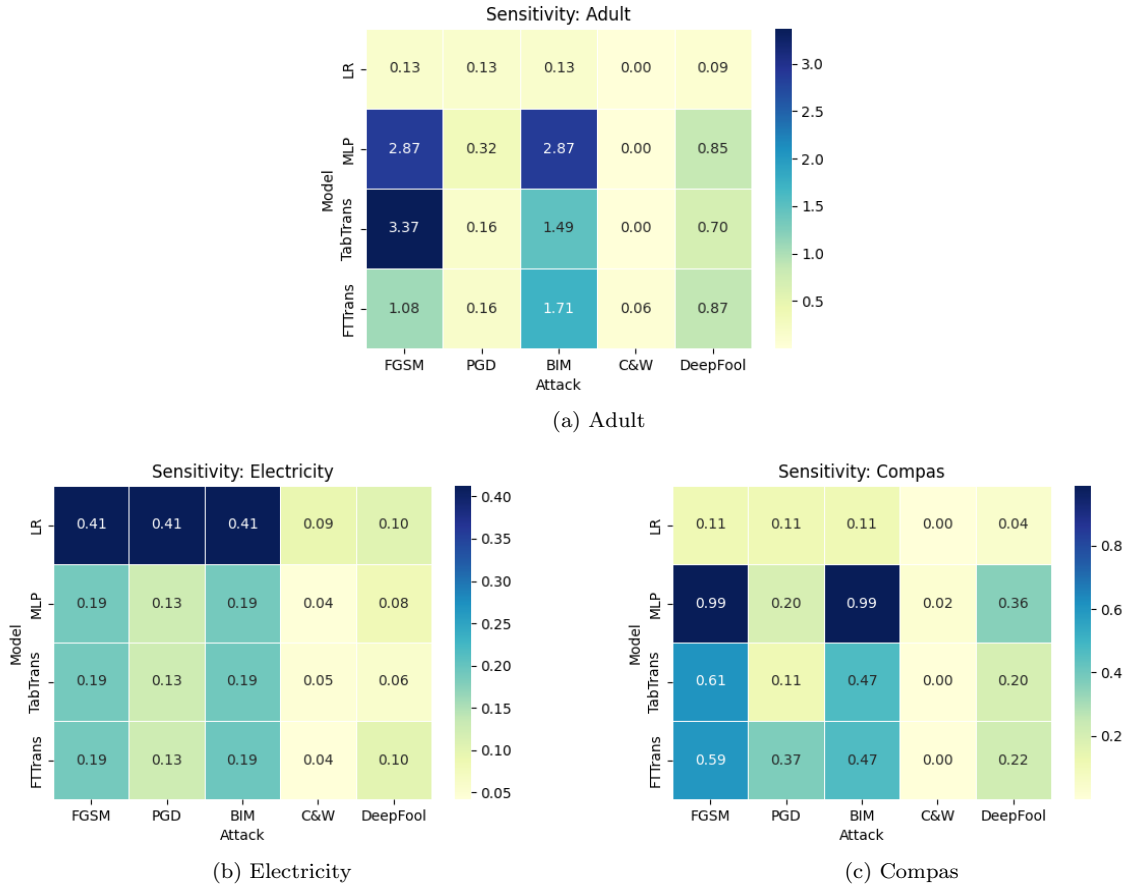
From a model architecture perspective, there are notable variations in how different models respond to distribution-shifting attacks. LR models generally exhibit the highest vulnerability to distribution shifts across datasets, particularly for ℓ_∞-based attacks. In contrast, transformer-based models occasionally demonstrate some resilience to certain attacks, as seen in the phoneme dataset where TabTransformer shows a remarkably low outlier rate (0.11) for BIM. This suggests that the more complex decision boundaries of transformer architectures may sometimes accommodate certain perturbations while maintaining in-distribution characteristics.

Overall, our deviation analysis confirms that ℓ_∞-based attacks consistently generate adversarial examples that significantly deviate from original data distributions, while ℓ_2-based

attacks tend to produce more in-distribution perturbations. However, the specific patterns vary notably by dataset characteristics and model architecture, highlighting the complex interplay between attack methods and the underlying data structures they attempt to exploit.

4.3.4. RQ2.4: How much do perturbations respect narrow-guard feature perturbation?

Our sensitivity analysis examines how adversarial attacks handle narrow-guard feature perturbation, particularly for features with narrow distributions in tabular data. The heatmaps in Figure 18, 19 and 20 reveal complex patterns that vary significantly across datasets, attack algorithms, and model architectures. Rather than showing consistent behaviours, the sensitivity metrics highlight the contextual nature of how perturbations interact with narrowly distributed features.



(a) Adult

(b) Electricity

(c) Compas

Figure 18: Sensitivity results of evaluated attack methods and four ML models on all three *mixed* dataset.

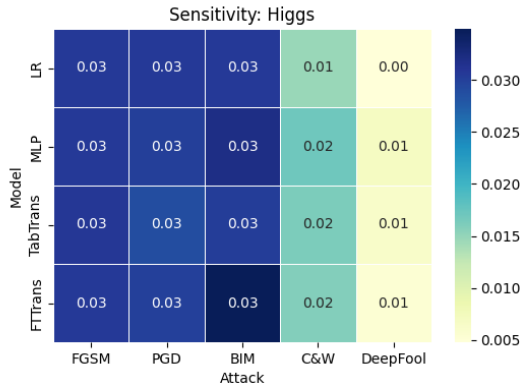
Sensitivity – Mixed Dataset. In the Adult dataset (Figure 18a), we observe striking differences across model architectures. While LR shows relatively low sensitivity scores (0.13) uniformly across ℓ_∞-based attacks, more complex models exhibit much higher sensitivity values. TabTransformer, when targeted by FGSM, produces the highest sensitivity score (3.37) among all mixed datasets, indicating that this combination significantly perturbs narrow-distribution features. Most notably, C&W consistently demonstrates the lowest sensitivity scores (0.00-0.06) across all models, suggesting its superior ability to preserve the integrity of narrow-guard features. This aligns with its ℓ_2-norm objective function that naturally penalises large deviations in any individual feature.

The Electricity dataset (Figure 18b) presents a more uniform sensitivity pattern across model architectures for the same attack method. LR model consistently shows higher sensitivity scores (0.41) for ℓ_∞-based attacks compared to other models (0.13-0.19). This suggests that simpler model architectures may induce attackers to make more substantial modifications to narrowly distributed features in this dataset. The three ℓ_∞-based attacks (FGSM, PGD, and BIM) produce identical sensitivity scores within each model architecture, indicating that these attacks, despite their algorithmic differences, alter narrow-distribution features similarly when applied to electricity data.

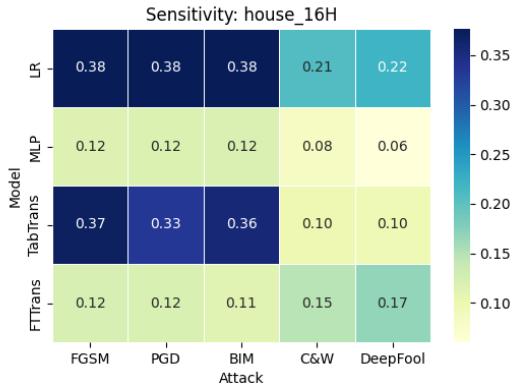
For the Compas dataset (Figure 18c), we observe moderate sensitivity scores overall but with notable variations across model architectures. MLP shows significantly higher sensitivity scores (0.99) for FGSM and BIM compared to transformer-based models (0.47-0.61), suggesting that the decision boundaries of MLPs may encourage more aggressive perturbations to narrow-distribution features. PGD consistently demonstrates lower sensitivity scores compared to other ℓ_∞-based attacks across all models, indicating its potentially more controlled approach to perturbing features with narrow distributions.

Sensitivity – Numerical Dataset. The Higgs dataset (Figure 19a) stands out for its remarkably low sensitivity scores across all attack methods and model architectures (0.00-0.03). This suggests that either the Higgs dataset lacks features with sufficiently narrow distributions or that its feature space allows effective adversarial examples without significantly

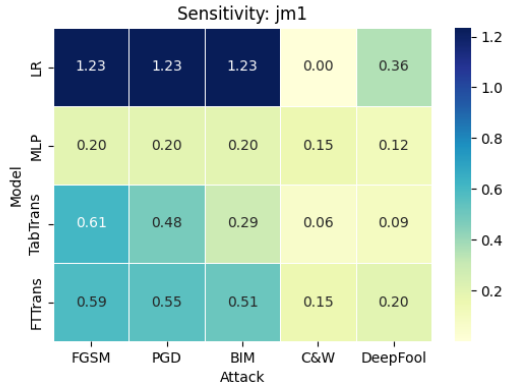
altering narrow-distribution features. The uniformity of scores across different attack methods suggests that the dataset characteristics, rather than attack algorithms, primarily drive the sensitivity outcomes in this case.



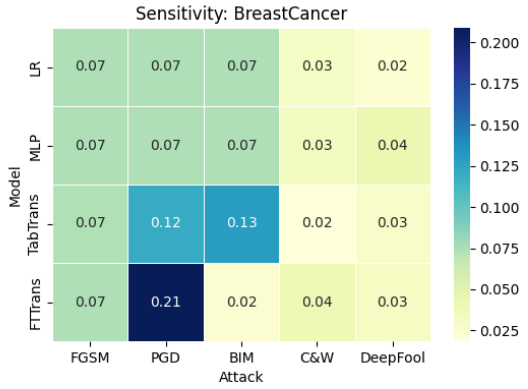
(a) Higgs



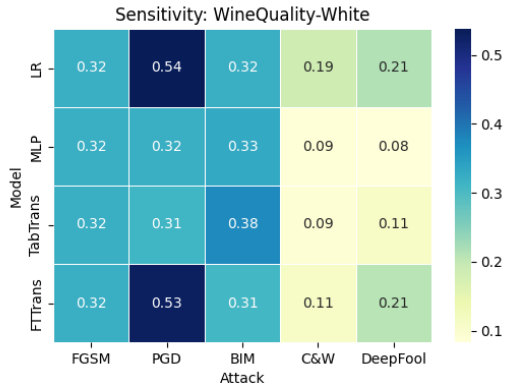
(b) house_16H



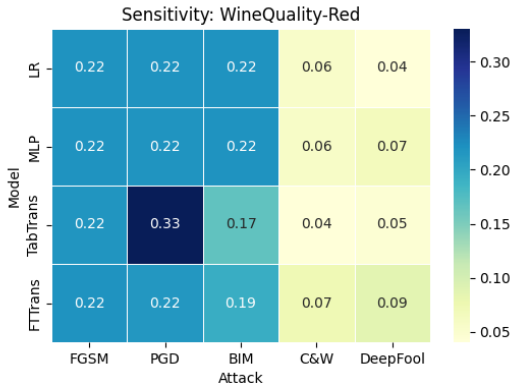
(c) jm1



(d) BreastCancer

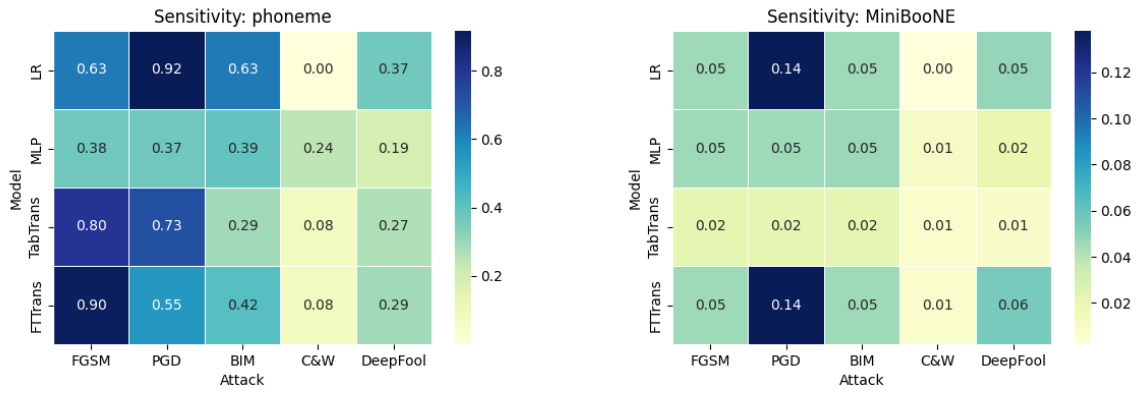


(e) WineQuality-White



(f) WineQuality-Red

Figure 19: Sensitivity results of evaluated attack methods and four ML models on six (out of eight) *numerical* datasets.



(a) phoneme (b) MiniBooNE

Figure 20: (Cont.) Sensitivity results of evaluated attack methods and four ML models on the remaining two (out of eight) *numerical* datasets.

In stark contrast, the jm1 dataset (Figure 19c) exhibits substantially higher sensitivity scores, particularly for LR models targeted by ℓ_∞-based attacks (1.23). This dramatic difference from other numerical datasets indicates that jm1 likely contains several features with narrow distributions that significantly influence model predictions. The sensitivity scores decrease markedly for more complex models, with TabTransformer and FTTransformer showing progressively lower scores (0.29-0.61) for the same attacks, suggesting that more sophisticated architectures may rely less on narrowly distributed features for their predictions.

The WineQuality datasets (White and Red, Figure 19e and 19f) demonstrate moderate sensitivity scores with interesting attack-specific patterns. For WineQuality-White, PGD produces notably higher sensitivity scores (0.31-0.54) compared to other attacks, particularly with LR and FTTransformer models. This suggests that PGD's iterative approach may target narrow-distribution features more aggressively in this dataset. The WineQuality-Red dataset shows more uniform sensitivity scores across ℓ_∞-based attacks but consistently lower values for ℓ_2-based attacks.

The phoneme dataset (Figure 20a) exhibits some of the highest sensitivity scores among numerical datasets, particularly for transformer models targeted by FGSM (0.80-0.90) and PGD (0.73). This suggests that phoneme's feature space contains influential narrow-distribution

features that transformer models heavily rely on for predictions. Conversely, the MiniBooNE dataset (Figure 20b) shows consistently low sensitivity scores across most attack-model combinations (0.01-0.06), with only PGD occasionally producing slightly higher values (0.14).

From a model architecture perspective, we observe that LR models often show either the highest or the lowest sensitivity scores depending on the dataset, suggesting that the interaction between model simplicity and dataset characteristics strongly influences how narrow-distribution features are perturbed. Transformer-based models show more variable sensitivity patterns across datasets, sometimes exhibiting high sensitivity (as in phoneme) and sometimes low (as in MiniBooNE).

Among attack methods, C&W consistently demonstrates the lowest sensitivity scores across nearly all datasets and models, confirming its tendency to preserve the characteristics of narrow-distribution features. PGD shows the most variable behaviour, sometimes producing the highest sensitivity scores (as in WineQuality-White and phoneme) and sometimes moderate values, suggesting its perturbation strategy may be more adaptive to dataset-specific characteristics.

The lack of consistent patterns across datasets underscores that sensitivity to narrow-guard feature perturbation is highly contextual, depending on the specific combination of dataset characteristics, model architecture, and attack algorithm. This variability highlights the importance of dataset-specific evaluation when assessing the imperceptibility of adversarial attacks from a sensitivity perspective, rather than attempting to draw universal conclusions about attack or model behaviours.

4.4. RQ3: Whether and how can the evaluated algorithms achieve a balance between both imperceptibility and effectiveness?

Evaluating the relationship between effectiveness and imperceptibility in adversarial attacks is crucial for understanding how well an attack balances both aspects. Effectiveness is measured by the attack success rate (RQ1) and imperceptibility is assessed using metrics including sparsity, proximity, sensitivity, and deviation (RQ2). Rather than comparing effectiveness against each imperceptibility metric individually, our approach evaluates the

overall imperceptibility, offering a more comprehensive understanding of its influence on attack success. This method supports the development of more robust adversarial attacks.

We propose a weighted harmonic mean to assess comprehensively imperceptibility, and refer to this overall metric as the Imperceptibility Score (IS). This assessment encompasses four metrics: proximity, deviation, sparsity, and sensitivity. This approach enables a nuanced and balanced evaluation, considering multiple facets of imperceptibility in the overall analysis.

The construction of the Imperceptibility Score follows a systematic approach, consisting of the following steps:

1. **Metric Definition:** Define and establish the four metrics, including proximity, deviation, sparsity, and sensitivity, each representing distinct aspects of imperceptibility in the context of our evaluation. All four metrics are already defined and employed for evaluation in the prior sections.
2. **Weight Assignment:** Assign appropriate weights to each metric based on their relative importance in the imperceptibility assessment. Considering the significance of each imperceptibility metric, we set the equal weight for each metric as:

$$proximity : sparsity : deviation : sensitivity = 0.25 : 0.25 : 0.25 : 0.25$$

3. **Score Normalisation:** Normalise the scores obtained for each metric to a common scale, such as 0-1, ensuring uniformity and comparability. This step is crucial to prevent biases arising from differences in the measurement scales of individual metrics. Sparsity and deviation can easily be converted into 0-1 scaling by using sparsity rate and outlier rate. For the other two metrics, an auxiliary normalisation function is required. Considering that the possible range of both ℓ_2 distance and sensitivity is $[0, +\infty)$, common normalisation method (Eq. 15) is not suitable since it is hard to seek the max value.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{15}$$

Practically, we select $x_{norm} = ln(x+1)$ as normalisation function to normalise all four metrics into the same scale.

4. **Imperceptibility Score Calculation**: Calculate the harmonic mean for the normalised scores of the four metrics. The harmonic mean, being more sensitive to lower values, ensures that deficiencies in any individual metric have a noticeable impact on the overall evaluation.

$$\text{IS} = \frac{n}{\sum_{i=1}^{n} \frac{w_i}{x_i}} \tag{16}$$

Analysing the relationship between attack success rate (ASR) and imperceptibility score (IS) provides critical insights into the relationship between effectiveness and imperceptibility of adversarial attacks on tabular data. By visualising this relationship through a 2D density plot in Figure 21, we can discern patterns that illuminate the interplay between these two crucial factors. The graphs were divided into four distinct sections based on specific thresholds, enabling us to categorise different scenarios and gain a clearer understanding of their impact. These thresholds were determined using our Gaussian noise method, which selects the maximum ASR value (0.659) and the minimum IS value (0.181) from all adversarial examples generated by Gaussian noise.

Effective and Imperceptible *(High ASR, Low IS).* The most desirable outcome for adversarial attacks occurs when examples successfully fool models while remaining nearly indistinguishable from original data. The density plot reveals that DeepFool consistently achieves this balance, with its highest density region falling in this quadrant. DeepFool's iterative approach of finding minimal perturbations to cross decision boundaries clearly excels at preserving tabular data characteristics while maintaining high effectiveness.

C&W also demonstrates strong performance in this quadrant for a portion of its examples, though it shows a bimodal distribution across both imperceptible regions. This suggests that C&W can achieve the ideal balance in many cases but may sometimes sacrifice effectiveness to maintain imperceptibility.

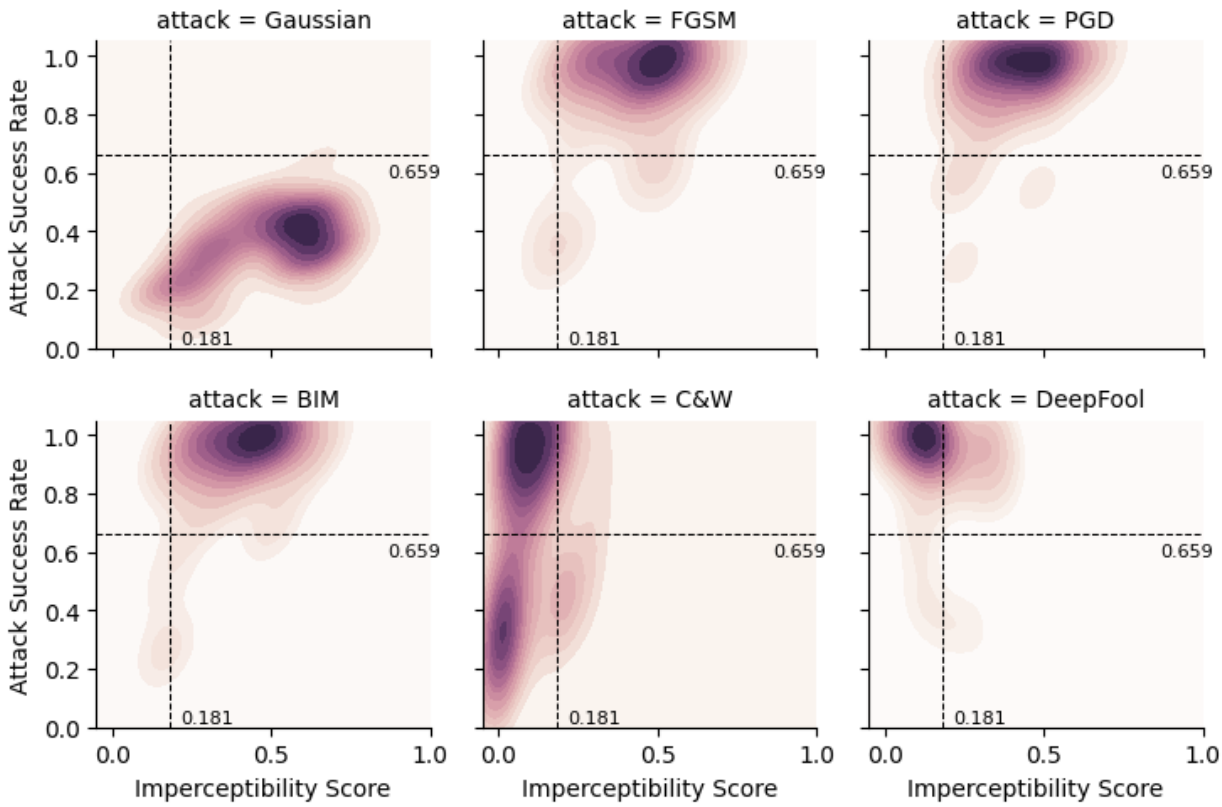


Figure 21: The 2D density plot shows the attack success rate (ASR) and imperceptibility score (IS) across Gaussian noise and five different attack methods. The plot is divided into four sectors based on the maximum ASR value (0.659) and the minimum IS value (0.181) observed among all adversarial examples generated by Gaussian noise. Gaussian noise is considered an ineffective and perceptible method for generating adversarial examples for tabular data. FGSM, PGD, and BIM are categorised as effective but perceptible methods. C&W attack has two high-density regions: one that is Effective and Imperceptible, and another that is Ineffective but Imperceptible. Most of DeepFool attack's high-density regions fall into the Effective and Imperceptible sector.

Effective but Perceptible *(High ASR, High IS).* This quadrant contains attacks that successfully mislead models but make noticeable modifications to the data. The density plots show that FGSM, PGD, and BIM consistently fall into this category, achieving high attack success rates at the cost of more significant data alterations. These ℓ_∞-based attacks effectively fool models but often modify features in ways that could compromise data integrity or be detected in quality control processes.

Ineffective but Imperceptible *(Low ASR, Low IS).* Attacks in this quadrant make subtle changes that preserve data characteristics but fail to successfully mislead models. C&W

shows a significant density in this region, indicating that it sometimes generates examples that maintain excellent imperceptibility but cannot effectively fool the model. This highlights C&W's explicit optimisation for minimal perturbations, which can sometimes come at the expense of attack effectiveness.

Ineffective and Perceptible *(Low ASR, High IS).* The least desirable outcome occurs when attacks make noticeable changes yet fail to mislead the model. Gaussian noise predominantly falls in this category, confirming its poor performance as a baseline comparison. Its high-density region centres on moderate ASR values with high imperceptibility scores, demonstrating why random noise is considered both ineffective and easily perceptible.

Overall Performance Comparison. The density plots provide clear evidence for ranking the overall performance of different attack methods:

1. DeepFool emerges as the most balanced approach, consistently generating examples that are both highly effective and imperceptible. Its iterative linearisation of decision boundaries enables precise identification of minimal perturbations needed to cross classification boundaries, resulting in subtle modifications that maintain data integrity while achieving high success rates.
2. C&W shows mixed results with two distinct behaviour patterns - one group achieving the ideal balance and another maintaining imperceptibility at the cost of effectiveness.
3. The ℓ_∞-based attacks (FGSM, PGD, and BIM) prioritise effectiveness over imperceptibility, making them suitable for scenarios where attack success is more important than maintaining data characteristics.
4. Gaussian noise serves as an appropriate baseline, demonstrating poor performance in both dimensions as expected.

This analysis provides valuable guidance for selecting appropriate attack methods based on specific requirements for tabular data scenarios, highlighting the fundamental trade-off between effectiveness and imperceptibility in adversarial machine learning.

5. Discussion

5.1. Investigating the Inverse Relationship Between BIM Attack Budget and Success Rate

As presented in Section 4.2, our **RQ1** evaluation results across both mixed and numerical datasets reveal an intriguing and counterintuitive phenomenon. While increasing epsilon ϵ values generally leads to improved success rates for most attack methods, the BIM attack on the FTTransformer model shows a notable decline in success rates at higher perturbation budgets. This inverse relationship between attack budget and effectiveness contradicts conventional adversarial attack theory, where larger perturbation budgets typically enable more successful attacks.

The plots in Figures 3 to 6 clearly demonstrate this unexpected pattern across multiple datasets, including Electricity, Compas, house_16H, BreastCancer, and MiniBooNE. In these cases, BIM attack success rates initially increase with epsilon values but then significantly decline at higher epsilon values, sometimes dropping dramatically. For example, on the BreastCancer dataset, the success rate drops from approximately 35% to nearly 0% at the highest epsilon value, while on MiniBooNE, it plummets from 100% to about 40%.

Two primary factors may explain this counterintuitive behavior:

1. **Gradient Saturation Effects:** In BIM's iterative approach, the step size (α) and number of iterations (T) play critical roles. When using default parameters (α=0.2, T=10), the relatively large step size may cause overshooting at higher epsilon values. This occurs because BIM computes gradients with respect to the input and takes steps in that direction. As epsilon increases, these steps can become too large, causing the attack to miss optimal adversarial regions and produce less effective perturbations.
2. **Decision Boundary Characteristics:** FTTransformer models may have complex decision boundaries with unique topological properties. At higher epsilon values, adversarial examples might cross these boundaries multiple times, potentially returning to regions where the model correctly classifies inputs. This threshold effect suggests that beyond certain perturbation magnitudes, the adversarial examples become perceptually more similar to their correct classes.

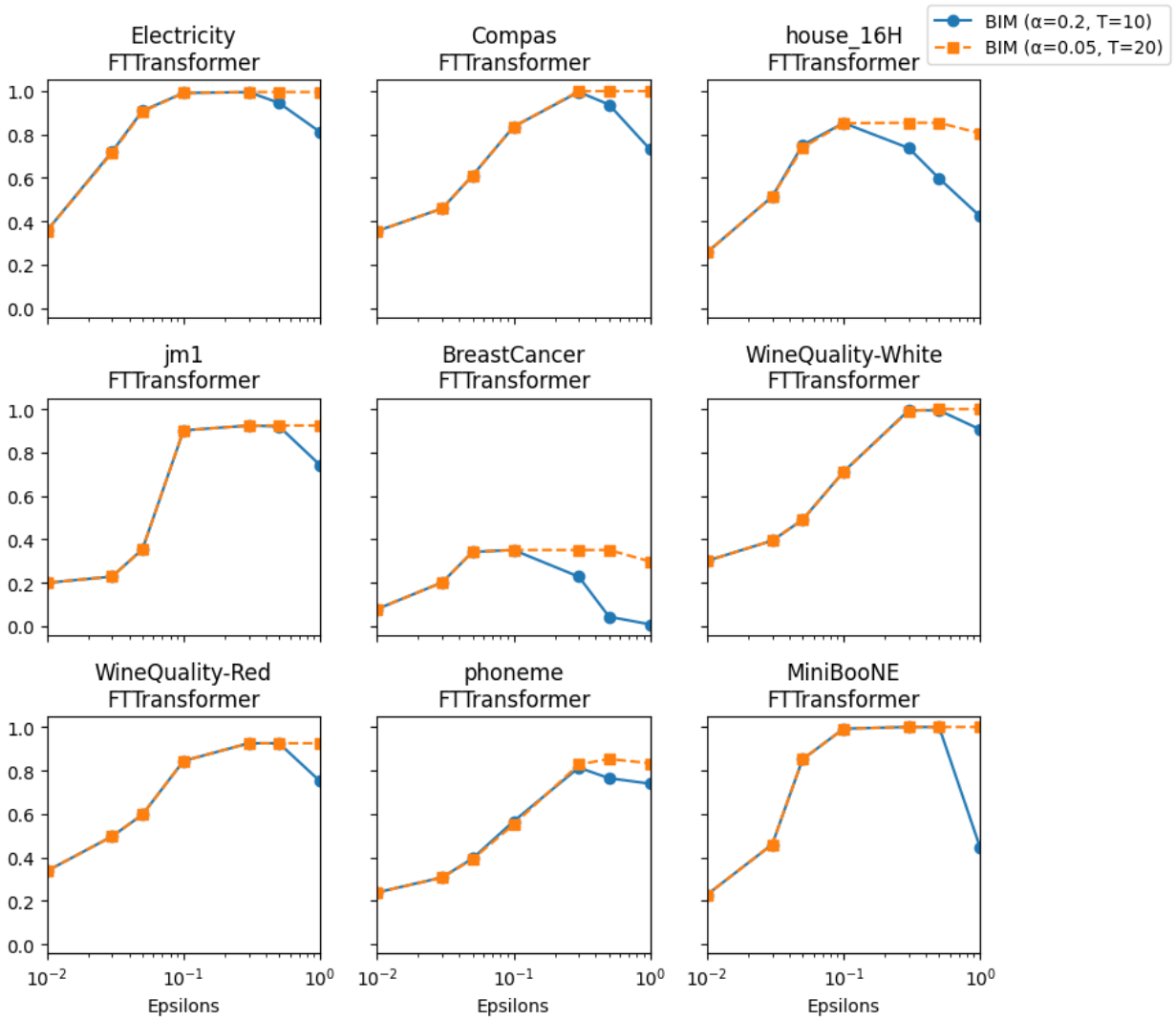


Figure 22: BIM attacks on FTTransformer stop dropping in attack success rate after adjusting step size (α) hyperparameters. The orange lines, representing the adjusted BIM implementation, consistently maintain high attack success rates across all epsilon values, eliminating the dramatic drops observed with the default parameters.

Our follow-up experiment demonstrates that adjusting BIM's hyperparameters can mitigate this issue. By reducing the step size (α) from 0.2 to 0.05 and increasing iterations (T) from 10 to 20, in Figure 22, we observe that the modified BIM attack maintains high success rates even at larger epsilon values across all datasets. This confirms that the original decline was primarily due to optimisation challenges rather than fundamental limitations of the attack method.

This finding has important implications for adversarial attack research on tabular data:

- FTTransformer models possess unique adversarial robustness characteristics that differ from other model architectures.
- Attack hyperparameters require careful tuning based on both the model architecture and dataset characteristics.
- When evaluating adversarial robustness, researchers should consider a range of attack configurations beyond default parameters to ensure comprehensive assessment.

This investigation highlights the complex interplay between attack algorithms, model architectures, and dataset characteristics in the tabular domain. It also demonstrates the importance of parameter optimisation when deploying adversarial attacks, particularly for transformer-based models that may have more complex decision boundary topologies than traditional neural networks.

5.2. Exploring Design Strategies for Effective and Imperceptible Adversarial Attacks on Tabular Data

In light of the results from analysing the relationship between attack success rate (ASR) and imperceptibility score (IS), achieving an optimal balance between effectiveness and imperceptibility is crucial in designing adversarial attack algorithms for tabular data. One notable observation is that ℓ_∞ attacks tend to generate highly effective adversarial examples, whereas ℓ_2 attacks are more adept at producing imperceptible examples. The key challenge lies in finding the equilibrium between these two aspects.

To design effective and imperceptible adversarial attack algorithms for tabular data, several strategies can be explored:

- **Optimisation Techniques**: Employing advanced optimisation techniques can enhance the efficiency of adversarial attack algorithms. Techniques such as evolutionary algorithms, genetic algorithms, or gradient-based optimisation methods can be tailored to optimise both ASR and IS simultaneously, thereby facilitating the creation of more effective and imperceptible adversarial examples.

- **Feature Engineering**: Leveraging domain-specific knowledge and feature engineering techniques can enhance the robustness and imperceptibility of adversarial attacks. By identifying and manipulating key features within the tabular data that are most susceptible to manipulation, attackers can craft adversarial examples that achieve their objectives while minimising perceptible changes to the data.

5.3. Evaluating the Suitability of One-Hot Encoding for Adversarial Attacks on Tabular Data

Adversarial attacks in machine learning have predominantly focused on image data, which are continuous and typically measured in the [0, 255] range. These attacks often involve adding small perturbations to the original samples, and the perturbations are evaluated using distance metrics like the ℓ_p-norm. However, when it comes to tabular data, the challenge becomes more complex due to the presence of both numerical and categorical features. Categorical data can further be divided into nominal data, which are used for naming variables, and ordinal data, which possess an intrinsic order. Encoding these categorical features into numerical values is crucial for applying adversarial attack algorithms effectively.

Several existing studies have explored different strategies for handling categorical features for adversarial attacks on tabular data in Table 5. Ballet et al. [23] proposed dropping all categorical features and using the ℓ_p-norm and Weighted ℓ_p-norm as distance metrics. Mathov et al. [25] suggested using label encoding for categorical data, though they did not specify the distance metric employed. Chernikova and Oprea [26] and Cartella et al. [33] both used one-hot encoding and applied the ℓ_2-norm as their distance metric. On the other

Table 5: The encoding methods employed in recent papers on adversarial attacks targeting tabular data.

Paper	Year	Encoding method	Distance metric
Ballet et al. [23]	2019	Drop all categorical features	ℓ_p-norm & Weighted ℓ_p-norm
Mathov et al. [25]	2021	Label encoding	ℓ_2-norm
Chernikova and Oprea [26]	2022	One-hot encoding	ℓ_2-norm
Cartella et al. [33]	2021	One-hot encoding	ℓ_2-norm
Kireev et al. [27]	2022	Discrete continuous features	Cost function
Zhou et al. [37]	2022	Discrete continuous features	ℓ_1-norm

hand, Kireev et al. [27] and Zhou et al. [37] opted for discretising continuous features and used a cost function as a distance measure, without specifying a particular norm.

In our work, we also adopted one-hot encoding for handling categorical features. However, our evaluation revealed that one-hot encoding significantly impacts the sparsity of adversarial attacks on tabular data. One-hot encoding transforms categorical variables into a high-dimensional binary vector space, which can lead to an increase in the dimensionality of the dataset and consequently affect the efficiency and effectiveness of adversarial attacks. The high-dimensional space created by one-hot encoding may dilute the perturbations, making it harder to generate effective adversarial examples while maintaining the integrity of the data.

This observation underscores the importance of choosing an appropriate encoding method for categorical features in adversarial attacks on tabular data. While one-hot encoding preserves the categorical nature of the data, its impact on sparsity and dimensionality needs careful consideration. Alternative encoding methods, such as label encoding for ordinal data or exploring more advanced techniques like embedding-based methods, might offer better trade-offs between preserving the data structure and maintaining the attack's effectiveness. Future research should focus on evaluating these methods to develop robust adversarial attack strategies for tabular data that balance the trade-offs between sparsity, dimensionality, and attack performance.

Moreover, exploring alternative distance metrics presents a promising direction for future research. Traditional metrics like the Lp-norm may not be well-suited for the mixed data types often found in tabular datasets. Metrics such as Gower's distance [38], which can handle mixed types of data (continuous, ordinal, and categorical), could provide a more accurate measure of similarity for tabular data. Additionally, other categorical feature similarity measures, such as those proposed by Cost and Salzberg [39] and Le and Ho [40], offer potential improvements by considering the unique characteristics of categorical data. By integrating these distance metrics into the design of adversarial attack algorithms, researchers can develop more effective and nuanced methods that are better tailored to the complexities of tabular data.

6. Conclusion

In this paper, we conducted a comprehensive benchmark analysis of adversarial attacks on tabular data, focusing on both their effectiveness and imperceptibility. Using a diverse set of 11 datasets, encompassing both mixed and numerical data types, we evaluated the performance of five different adversarial attacks across four predictive models. Our findings reveal significant variations in attack effectiveness depending on the dataset and model combination. Furthermore, the study highlights the challenge of maintaining attack imperceptibility, particularly in the context of tabular data, where subtle modifications can become perceptually noticeable.

The results of our benchmark provide valuable insights into the strengths and limitations of existing adversarial attack methods when applied to tabular data. By analysing the trade-offs between attack success rates and their imperceptibility, we offer an understanding of how different attacks perform under varying conditions. This study lays the groundwork for future research aimed at developing more robust adversarial defences that can effectively counteract these attacks while preserving the integrity of tabular datasets.

This research assumes that all features contribute equally to the predictive models, similar to the notion that each pixel holds equal importance in images. However, real-world tabular datasets often exhibit complex inter-dependencies among features. This observation points to the need for future work to explore non-uniform adversarial attacks [41, 42]. Addressing these challenges will contribute to a more comprehensive understanding of the robustness and generalisation capabilities of predictive models in practical applications.

Code Availability

The implementation code, including data processing scripts and experimental pipelines, is openly available at https://github.com/ZhipengHe/TabAttackBench/

CRediT authorship contribution statement

Zhipeng He: Conceptualisation, Methodology, Software, Investigation, Writing – Original Draft, Visualisation. **Chun Ouyang:** Conceptualisation, Methodology, Investigation, Writing – Original Draft. **Lijie Wen:** Methodology, Investigation. **Cong Liu:** Supervision. **Catarina Moreira:** Methodology, Investigation, Supervision.

Appendix A. Selected Attack Budgets (ϵ) by ASR

Table A.6: Best attack budget (ϵ) settings for four models on different datasets from the evaluation.

Datasets	**Model**	**Guassian**	**FGSM**	**PGD**	**BIM**	**C&W**	**DeepFool**
Adult	LR	1	0.3	0.3	0.3	0.5	1
Adult	MLP	0.5	1	1	1	0.01	1
Adult	TabTrans	0.5	1	0.5	0.5	0.01	1
Adult	FTTrans	1	0.3	0.5	0.5	0.5	1
Electricity	LR	1	0.3	0.3	0.3	1	0.5
Electricity	MLP	1	0.1	0.1	0.1	0.3	0.3
Electricity	TabTrans	1	0.1	0.1	0.1	0.3	0.3
Electricity	FTTrans	1	0.1	0.1	0.1	0.3	0.3
Compas	LR	0.5	0.3	0.3	0.3	0.1	1
Compas	MLP	0.3	0.5	0.5	0.5	1	1
Compas	TabTrans	1	0.3	0.3	0.3	0.1	0.5
Compas	FTTrans	0.07	0.3	1	0.3	0.01	0.5
Higgs	LR	1	0.07	0.07	0.07	0.3	0.1
Higgs	MLP	1	0.07	0.07	0.07	0.3	0.1
Higgs	TabTrans	1	0.07	0.07	0.07	0.3	0.1
Higgs	FTTrans	1	0.07	0.07	0.07	0.3	0.1
house_16H	LR	0.01	0.3	0.3	0.3	1	1
house_16H	MLP	1	0.1	0.1	0.1	0.5	0.3

house_16H	TabTrans	1	0.3	0.3	0.3	0.5	0.5
house_16H	FTTrans	1	0.1	0.1	0.1	1	0.5
jm1	LR	0.01	0.5	0.5	0.5	0.01	1
jm1	MLP	1	0.1	0.1	0.1	0.5	0.5
jm1	TabTrans	1	0.3	0.3	0.3	0.3	0.5
jm1	FTTrans	1	0.3	0.3	0.3	0.5	0.5
BreastCancer	LR	1	0.3	0.3	0.3	1	1
BreastCancer	MLP	1	0.3	0.3	0.3	1	1
BreastCancer	TabTrans	1	0.3	0.5	0.5	1	1
BreastCancer	FTTrans	1	0.3	1	0.1	1	0.5
WineQuality-White	LR	1	0.3	0.5	0.3	1	1
WineQuality-White	MLP	1	0.3	0.3	0.3	0.5	0.5
WineQuality-White	TabTrans	1	0.3	0.3	0.3	0.5	0.5
WineQuality-White	FTTrans	1	0.3	0.5	0.3	1	1
WineQuality-Red	LR	1	0.3	0.3	0.3	1	0.5
WineQuality-Red	MLP	1	0.3	0.3	0.3	0.5	0.5
WineQuality-Red	TabTrans	1	0.3	0.5	0.3	0.5	0.5
WineQuality-Red	FTTrans	1	0.3	0.3	0.3	1	0.5
phoneme	LR	1	0.5	1	0.5	0.01	1
phoneme	MLP	1	0.3	0.3	0.3	1	1
phoneme	TabTrans	1	1	1	0.3	1	1
phoneme	FTTrans	1	1	0.5	0.3	0.5	1
MiniBooNE	LR	1	0.1	0.3	0.1	0.1	0.3
MiniBooNE	MLP	1	0.1	0.1	0.1	0.1	0.3
MiniBooNE	TabTrans	1	0.07	0.07	0.07	0.1	0.3
MiniBooNE	FTTrans	1	0.1	0.3	0.1	0.1	0.5

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: 2nd International Conference on Learning Representations, ICLR 2014, 2014.

[2] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, Ieee Access 6 (2018) 14410–14430.

[3] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020) 1–41.

[4] N. Carlini, D. Wagner, Audio adversarial examples: Targeted attacks on speech-to-text, in: 2018 IEEE security and privacy workshops (SPW), IEEE, 2018, pp. 1–7.

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

[6] S. Eger, Y. Benz, From hero to zéroe: A benchmark of low-level adversarial attacks, in: Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing, 2020, pp. 786–803.

[7] Z. Jin, J. Zhang, Z. Zhu, H. Chen, Short: Benchmarking transferable adversarial attacks, in: Network and Distributed System Security (NDSS) Symposium 2024, 2024.

[8] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, J. Zhu, Benchmarking adversarial robustness on image classification, in: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 321–331.

[9] M. Zheng, X. Yan, Z. Zhu, H. Chen, B. Wu, Blackboxbench: A comprehensive benchmark of black-box adversarial attacks, arXiv preprint arXiv:2312.16979 (2023).

[10] N. Hingun, C. Sitawarin, J. Li, D. Wagner, Reap: a large-scale realistic adversarial patch benchmark, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4640–4651.

[11] A. E. Cinà, J. Rony, M. Pintor, L. Demetrio, A. Demontis, B. Biggio, I. B. Ayed, F. Roli, Attackbench: Evaluating gradient-based attacks for adversarial examples, arXiv preprint arXiv:2404.19460 (2024).

[12] Q. Zheng, X. Zou, Y. Dong, Y. Cen, D. Yin, J. Xu, Y. Yang, J. Tang, Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning, arXiv preprint arXiv:2111.04314 (2021).

[13] L. Li, J. Lei, Z. Gan, J. Liu, Adversarial vqa: A new benchmark for evaluating the robustness of vqa models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2042–2051.

[14] S. A. Siddiqui, A. Dengel, S. Ahmed, Benchmarking adversarial attacks and defenses for time-series data, in: International Conference on Neural Information Processing, Springer, 2020, pp. 544–554.

[15] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: 3rd International Conference on Learning Representations, ICLR 2015, 2015.

[16] Z. He, C. Ouyang, L. Alzubaidi, A. Barros, C. Moreira, Investigating imperceptibility of adversarial attacks on tabular data: An empirical analysis, Intelligent Systems with Applications 25 (2025) 200461.

[17] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: IEEE European Symposium on Security and Privacy, EuroS&P 2016, IEEE, 2016, pp. 372–387.

[18] F. Assion, P. Schlicht, F. Greßner, W. Günther, F. Hüger, N. M. Schmidt, U. Rasheed, The attack generator: A systematic approach towards constructing adversarial attacks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 1370–1379.

[19] A. Kurakin, I. J. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: 5th International Conference on Learning Representations, ICLR 2017, 2017.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv:1706.06083 (2017).

[21] N. Carlini, D. A. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy, SP 2017, IEEE Computer Society, 2017, pp. 39–57.

[22] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, IEEE Computer Society, 2016, pp. 2574–2582.

[23] V. Ballet, X. Renard, J. Aigrain, T. Laugel, P. Frossard, M. Detyniecki, Imperceptible adversarial attacks on tabular data, arXiv 1911.03274 (2019).

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

[25] Y. Mathov, E. Levy, Z. Katzir, A. Shabtai, Y. Elovici, Not all datasets are born equal: On heterogeneous tabular data and adversarial examples, Knowl. Based Syst. 242 (2022) 108377.

[26] A. Chernikova, A. Oprea, FENCE: feasible evasion attacks on neural networks in constrained environments, ACM Trans. Priv. Secur. 25 (2022) 34:1–34:34.

[27] K. Kireev, B. Kulynych, C. Troncoso, Adversarial robustness for tabular data through cost and utility awareness, arXiv preprint arXiv:2208.13058 (2022).

[28] J. Zhou, N. Zaidi, Y. Zhang, P. Montague, J. Kim, G. Li, Leveraging generative models for combating adversarial attacks on tabular datasets, in: Pacific-Asia Conference on Knowledge Discovery and Data

Mining, Springer, 2023, pp. 147–158.

[29] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, C.-J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM workshop on artificial intelligence and security, 2017, pp. 15–26.

[30] W. Brendel, J. Rauber, M. Bethge, Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, arXiv preprint arXiv:1712.04248 (2017).

[31] J. Chen, M. I. Jordan, M. J. Wainwright, Hopskipjumpattack: A query-efficient decision-based attack, in: 2020 ieee symposium on security and privacy (sp), IEEE, 2020, pp. 1277–1294.

[32] G. Gressel, N. Hegde, A. Sreekumar, R. Radhakrishnan, K. Harikumar, K. Achuthan, et al., Feature importance guided attack: A model agnostic adversarial attack, arXiv preprint arXiv:2106.14815 (2021).

[33] F. Cartella, O. Anunciação, Y. Funabiki, D. Yamaguchi, T. Akishita, O. Elshocht, Adversarial attacks for tabular data: Application to fraud detection and imbalanced data, in: Workshop on Artificial Intelligence Safety 2021, 2021.

[34] J. Thiyagalingam, M. Shankar, G. Fox, T. Hey, Scientific machine learning benchmarks, Nature Reviews Physics 4 (2022) 413–420.

[35] X. Huang, A. Khetan, M. Cvitkovic, Z. Karnin, Tabtransformer: Tabular data modeling using contextual embeddings, arXiv preprint arXiv:2012.06678 (2020).

[36] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data, Advances in Neural Information Processing Systems 34 (2021) 18932–18943.

[37] J. Zhou, N. A. Zaidi, Y. Zhang, G. Li, Discretization inspired defence algorithm against adversarial attacks on tabular data, in: Advances in Knowledge Discovery and Data Mining - 26th Pacific-Asia Conference, PAKDD 2022, volume 13281 of *Lecture Notes in Computer Science*, 2022, pp. 367–379.

[38] J. C. Gower, A general coefficient of similarity and some of its properties, Biometrics (1971) 857–871.

[39] S. Cost, S. Salzberg, A weighted nearest neighbor algorithm for learning with symbolic features, Machine learning 10 (1993) 57–78.

[40] S. Q. Le, T. B. Ho, An association-based dissimilarity measure for categorical data, Pattern Recognition Letters 26 (2005) 2549–2557.

[41] E. Erdemir, J. Bickford, L. Melis, S. Aydore, Adversarial robustness with non-uniform perturbations, Advances in Neural Information Processing Systems 34 (2021) 19147–19159.

[42] J. Nandy, J. Chauhan, R. Saket, A. Raghuveer, Non-uniform adversarial perturbations for discrete tabular datasets, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 1887–1896.