Investigating Imperceptibility of Adversarial Attacks on Tabular Data: An Empirical Analysis

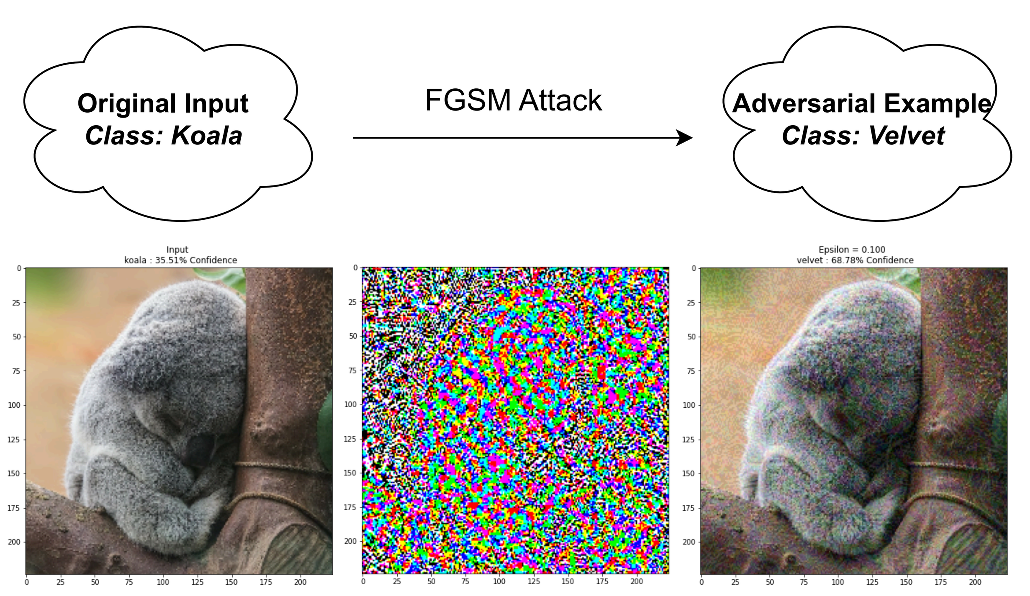
Zhipeng He[1], Chun Ouyang[1], Laith Alzubaidi[1], Alistair Barros[1], Catarina Moreira[2]

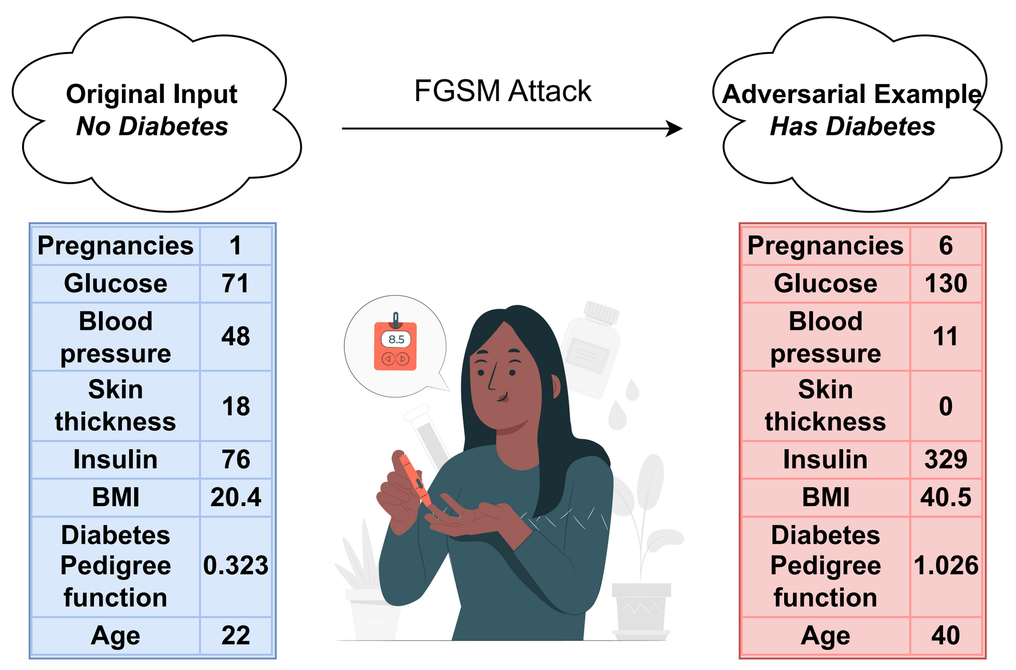
[1] Queensland University of Technology, [2] University of Technology Sydney

Problem Definition

Background: Adversarial attacks involve modifying input data to deceive machine learning models. These attacks have been studied for unstructured data (e.g., images), but structured tabular data poses unique challenges.

Gaps: 1) The imperceptibility of adversarial attacks on tabular data requires approaching different concepts compared to those for images. 2) Current adversarial attacks lack imperceptibility metrics tailored for tabular data.





Goal: Develop a set of standardised properties and metrics to evaluate the imperceptibility of adversarial attacks on tabular data.

Properties and Metrics of Imperceptibility

Proximity: A good adversarial example should introduce *minimal* changes, quantified by ensuring the smallest possible distance from the original feature vector. To measure the perturbation distance, we use the ℓ_2 and ℓ_∞ ***norms***.

Sparsity: An ideal adversarial example should misclassify the model's prediction by altering the fewest features possible. We use the ℓ_0 ***norm*** to count the number of perturbed features.

Deviation: To ensure imperceptibility, an adversarial example should closely resemble the majority of original inputs. We propose using the ***Mahalanobis distance*** to measure the deviation between an adversarial perturbation and the distribution of variations in the original data inputs.

Sensitivity: We adapt the concept of ***perturbation sensitivity*** as a metric to quantify the extent to which features with narrow distributions in tabular data are altered, which is defined as follows:

$$\mathrm{SDV}(x_i) = \sqrt{\frac{\sum_j^m (x_{i,j} - \bar{x}_i)^2}{m}}, \quad \mathrm{SEN}(\boldsymbol{x}, \boldsymbol{x}^{adv}) = \sum_{i=1}^{n} \frac{\|x_i^{adv} - x_i\|_2}{\mathrm{SDV}(x_i)}$$

where n is the number of numerical features, m is the number of all input vectors, and $\bar{x}_i$ represents the average of the ith features within all datapoints.

Immutability: ***Immutable features*** are fixed attributes in a dataset that are either inherently unchangeable or should remain unaltered due to ethical or practical constraints.

Feasibility: Adversarial attacks should avoid introducing perturbations that push feature values beyond ***feasible value ranges***, ensuring alignment with semantic correctness.

Feature Interdependency: Tabular data often contains features with non-linear and context-specific interactions or relationships. Altering a feature independently of its ***correlated features*** can create anomalies that are easily detectable.

Analysis of Imperceptibility using Qualitative Properties

Immutability: Feature *Race* should not be perturbed.

Feature Interdependency: If feature *Age*'s value is altered, feature *Age Category* should be correspondingly updated to reflect this change accurately.

Case	Attack	Age	Priors Count	Length of Stay	Age Cat.	Sex	Race	Class
#285	Original	80	0	0	Greater than 45	Male	*Caucasian*	*Medium-Low*
	DeepFool	18	38	799	Greater than 45	Male	*Native American*	*High*
#501	Original	83	0	0	Greater than 45	Male	*Hispanic*	*Medium-Low*
	DeepFool	18	38	799	Less than 25	Male	*African-American*	*High*

Fessibility: Features *Glucose* and *BMI* are prone to being perturbed into extreme values.

Case	Attack	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Age	Diabetes?
#19	Original	86.00	68.00	28.00	71.00	30.20	24.00	N
	DeepFool	159.50	62.21	30.24	51.32	49.00	32.69	Y
	C&W	135.98	63.93	29.18	69.29	43.22	24.32	Y
	LowProFool	199.00	56.17	32.57	30.80	67.10	41.74	Y
#57	Original	74.00	68.00	28.00	45.00	29.70	23.00	N
	DeepFool	191.86	58.71	31.59	13.43	59.85	36.93	Y
	C&W	136.88	62.84	29.65	43.81	45.97	23.24	Y
	LowProFool	199.00	55.53	32.81	2.65	67.10	41.69	Y

Contributions and Findings

Contributions: Our contributions to the field are twofold:

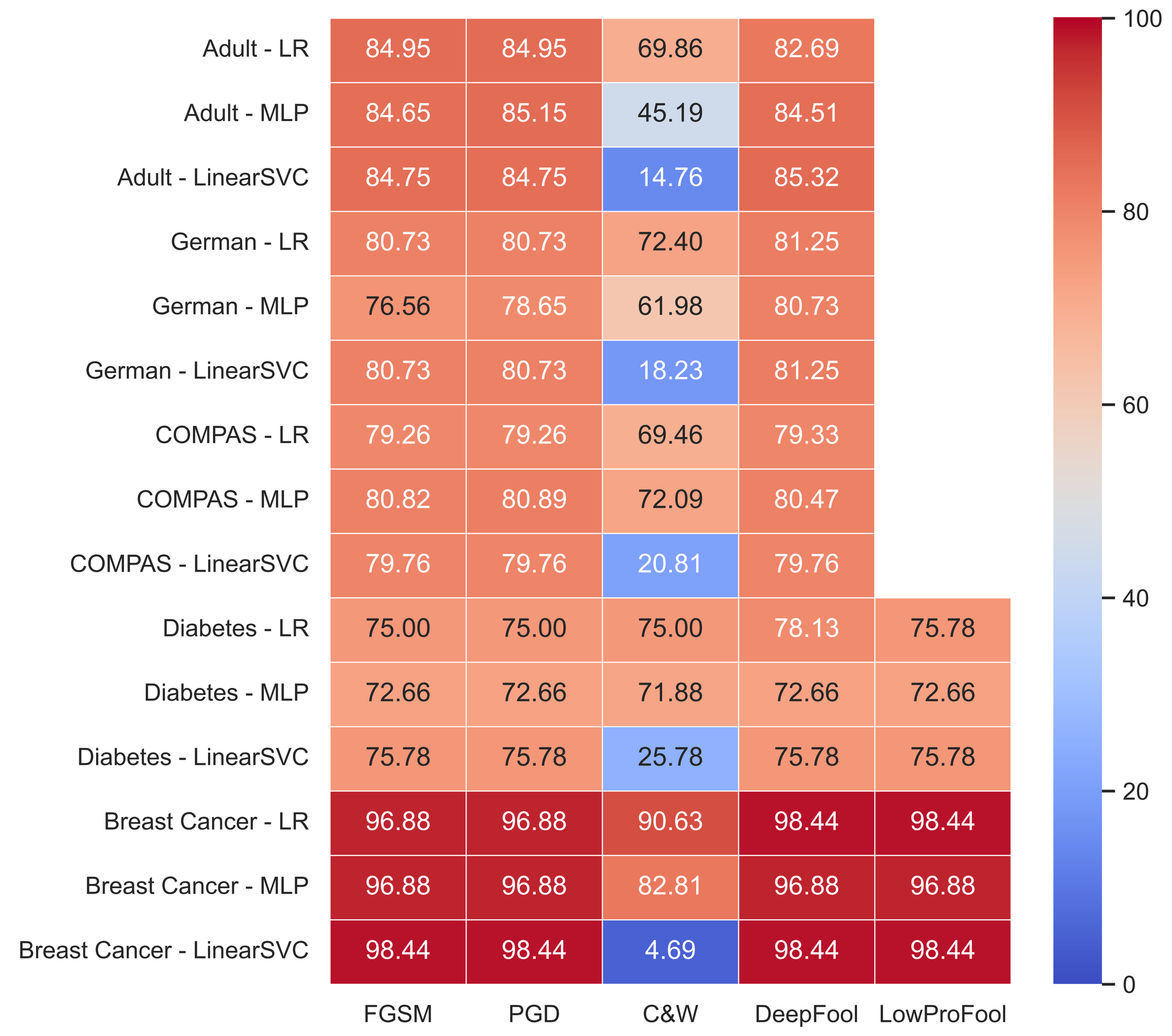
- We propose seven key properties that define imperceptible adversarial attacks for tabular data. These properties—***proximity***, ***sparsity***, ***deviation***, ***sensitivity***, ***immutability***, ***feasibility***, and ***feature interdependency***—are derived from the unique characteristics and challenges associated with tabular data.
- Using the proposed metrics, we empirically evaluated five adversarial attack methods, investigating all seven imperceptibility properties, analysing their relationship with attack effectiveness, and providing insights from the results.

Findings: The findings reveal that:

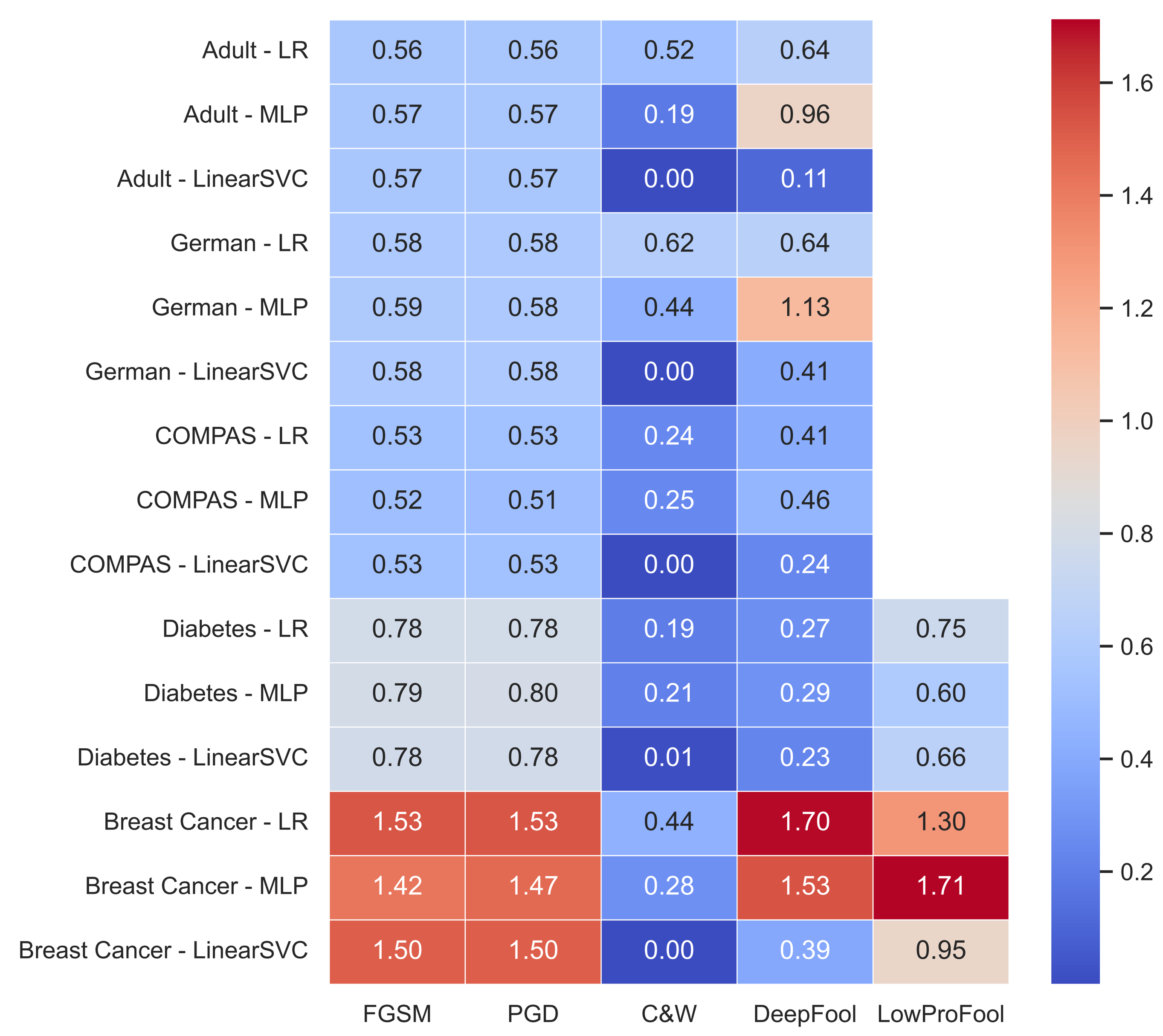
- With the exception of proximity, current adversarial attack methods often fail to consider the proposed imperceptibility properties in their algorithm designs.
- Additionally, our analysis highlights a trade-off between the effectiveness of adversarial attacks and their imperceptibility.

Experiments & Results

Results of Attack Effectiveness:



Results of Proxmity ℓ_2:



About Us

Project Webpage:

Code & Dataset & Model