Securing Machine Learning: Privacy Preservation and Adversarial Resilience

Olivier BRUNET, Arij SAKHRAOUI

Université Paris Dauphine - PSL

November 25, 2023

Introduction

- **Differential Privacy applied to Healthcare:**
  - Exploring the implementation of differential privacy in diabetes prediction.
  - Evaluating the effectiveness of privacy-preserving techniques in protecting sensitive information.
- **Adversarial Attacks:**
  - Investigating the impact of adversarial attacks on model robustness, using the CIFAR dataset.
  - Understanding defensive strategies against such attacks to enhance model security.
- **Outcome:**
  - Insightful analysis of privacy-preserving techniques and adversarial robustness.
  - Enhancing awareness for ethical and responsible AI applications.

Dauphine | PSL

Agenda

1. Differential Privacy
   - Analysis with a Constant Epsilon
   - Accuracy Trends with Varying Epsilon
   - Conclusion
2. Adversarial Attacks
   - Convolutional Neural Network Architecture
   - Fast Gradient Sign Method
   - Projected Gradient Descent
   - Projected Gradient Descent with L_2 Norm
   - Adversarial Training
   - Local Intrinsic Dimensionality
   - No Free Lunch
3. Conclusion

Dauphine | PSL

Balancing Data Accessibility and Privacy in Healthcare

Key Challenges:

- Balancing the need for data publication (open data, surveys) with the imperative of data protection against leaks and attacks.
- Recognizing the risks of data re-identification, especially in easily accessible datasets.

Healthcare Case Study: Diabetes Prediction

- Using sensitive health data for predicting diabetes, highlighting privacy concerns even in "safe" data releases.

Advantage of Differential Privacy:

- Offers robust and quantifiable privacy protection, independent of potential attacks or auxiliary data.

Dauphine | PSL

Differential Privacy in Healthcare

Differential Privacy Definition:

- A randomized algorithm $\mathcal{A}$ is ϵ-differentially private if for any output set Z and for any two similar datasets X and X',

$$\mathbb{P}[\mathcal{A}(X) \in Z] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(X') \in Z].$$

- ϵ (epsilon) is a privacy budget that quantifies the allowable information leakage.

Applying to Diabetes Prediction:

- Utilizing randomized response technique to obscure individual diabetes status.
- Analysis of the trade-offs in predictive accuracy when varying ϵ in machine learning models.

Differential Privacy in Healthcare

Quantifying Privacy Loss:

- Privacy loss in differential privacy is represented by the value of ϵ. Lower ϵ means higher privacy (and usually less accuracy).
- Mathematical models to assess the impact of differential privacy on the statistical properties of healthcare data.

Noise Addition Techniques:

- Adding Laplace noise to numerical data to achieve differential privacy.

Ethical Implications:

- Understanding the balance between data privacy and the need for accurate medical predictions.
- Addressing concerns over data misuse and re-identification risks in healthcare datasets.

Data Anonymization Overview

Target (sensitive)

Identifiers (removed)

→ Quasi-identifiers (anonymized)

● The other columns are useless (removed)

	feature	type	# null	% null	# unique	% unique	sample
0	admitted_ts	object	0	0.0	1000	100.0	[2018-05-09 12:06:28, 2018-05-12 10:02:55, 201...
1	age	int64	0	0.0	73	7.3	[49, 82, 71, 87, 53]
2	ambulance_call	int64	0	0.0	2	0.2	[1, 0]
3	blood_sugar_reading	int64	0	0.0	80	8.0	[108, 70, 100, 113, 93]
	days_since_last_visit	int64	0	0.0	81	8.1	[99, 100, 78, 72, 80]
5	has_diabetes	int64	0	0.0	2	0.2	[1, 0]
6	hospital	object	0	0.0	4	0.4	[district, general, northern, central]
7	hours_hospitalized	int64	0	0.0	23	2.3	[15, 22, 1, 17, 13]
8	hydration_level	int64	0	0.0	10	1.0	[6, 1, 4, 8, 5]
9	id	int64	0	0.0	1000	100.0	[1000, 1001, 1002, 1003, 1004]
10	insulin	int64	0	0.0	2	0.2	[1, 0]
11	marital_status	object	0	0.0	4	0.4	[single, married, no_answer, divorced]
12	no_primary_dr	bool	0	0.0	2	0.2	[False, True]
13	patient_name	object	0	0.0	991	99.1	[Rachel Shelton, Barbara Medina, Kaitlyn Danie...
14	private_insurance	int64	0	0.0	2	0.2	[0, 1]
15	released_sameday	int64	0	0.0	2	0.2	[0, 1]
16	ssn	object	0	0.0	1000	100.0	[743-97-4081, 698-10-2230, 540-83-4297, 282-96...
17	symptom_code	int64	0	0.0	10	1.0	[4, 3, 2, 0, 8]

Dauphine | PSL

Target Distribution

- **Observation:** Target distribution shows imbalance.

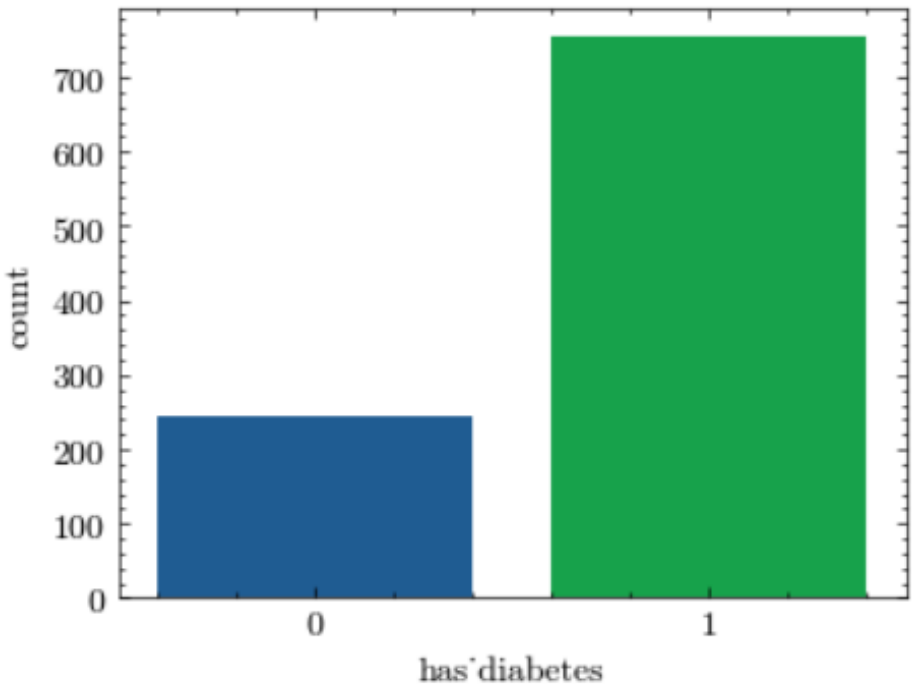


Figure: Visual representation of target distribution.

Analysis with a Constant Epsilon

- Impact of applying differential privacy with $\epsilon = 1.39$ on the dataset.

	age	blood_sugar_reading	has_diabetes	hydration_level	insulin	divorced	married	no_answer	single
0	49	108	1	6	1	0	0	0	1
1	82	70	1	1	1	0	1	0	0
2	71	100	1	4	1	0	0	1	0
3	87	113	1	4	1	0	0	1	0
4	53	93	1	8	0	0	0	1	0

For p = q = 0.5, epsilon = 1.39, preview of the epsilon DP private dataframe:

	age	blood_sugar_reading	has_diabetes	hydration_level	insulin	divorced	married	no_answer	single
0	49	107.983209	1	1.272616	1	0	0	0	1
1	82	69.990710	1	-2.687797	0	0	1	0	0
2	71	99.986916	1	1.261591	1	0	0	1	0
3	87	113.009946	0	0.191663	1	0	0	1	0
4	53	93.014373	0	0.614862	0	0	0	1	0

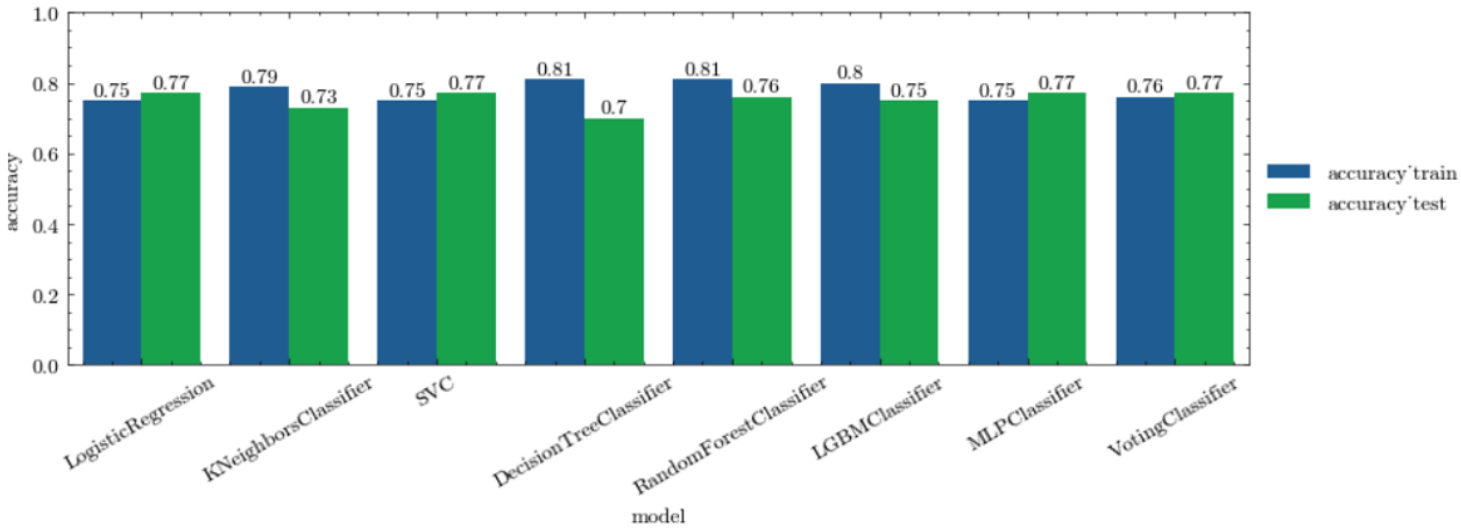
Model Initialization

- Initialization of a suite of classifier models for comparison.

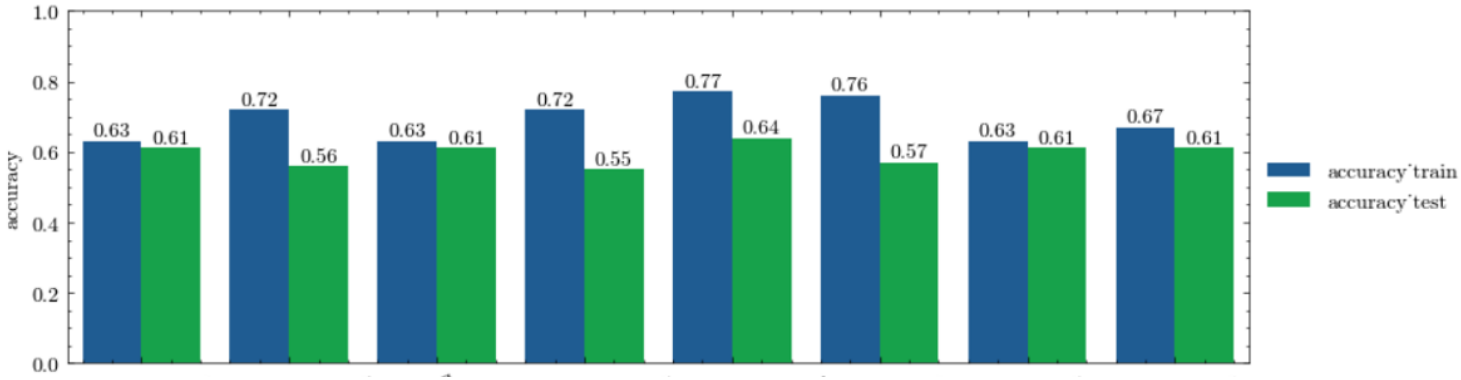
```
models = [
    LogisticRegression(solver='lbfgs', max_iter=1000, random_state=random_state),
    KNeighborsClassifier(),
    SVC(random_state=random_state),
    DecisionTreeClassifier(max_depth=7, random_state=random_state),
    RandomForestClassifier(n_estimators=30, max_depth=7, random_state=random_state),
    lgb.LGBMClassifier(n_estimators=30, max_depth=7, random_state=random_state),
    MLPClassifier(learning_rate = "adaptive", random_state=random_state)
]

voting_estimators = [(type(model).__name__, model) for model in models]

models.append(VotingClassifier(estimators=voting_estimators, voting='hard'))
```

Model Performance on Original & DP Enhanced Datasets

```
Comparison of accuracies for each models on the ORIGINAL train/test dataset
Acc. mean on train set: 0.78 / on test set: 0.75
```

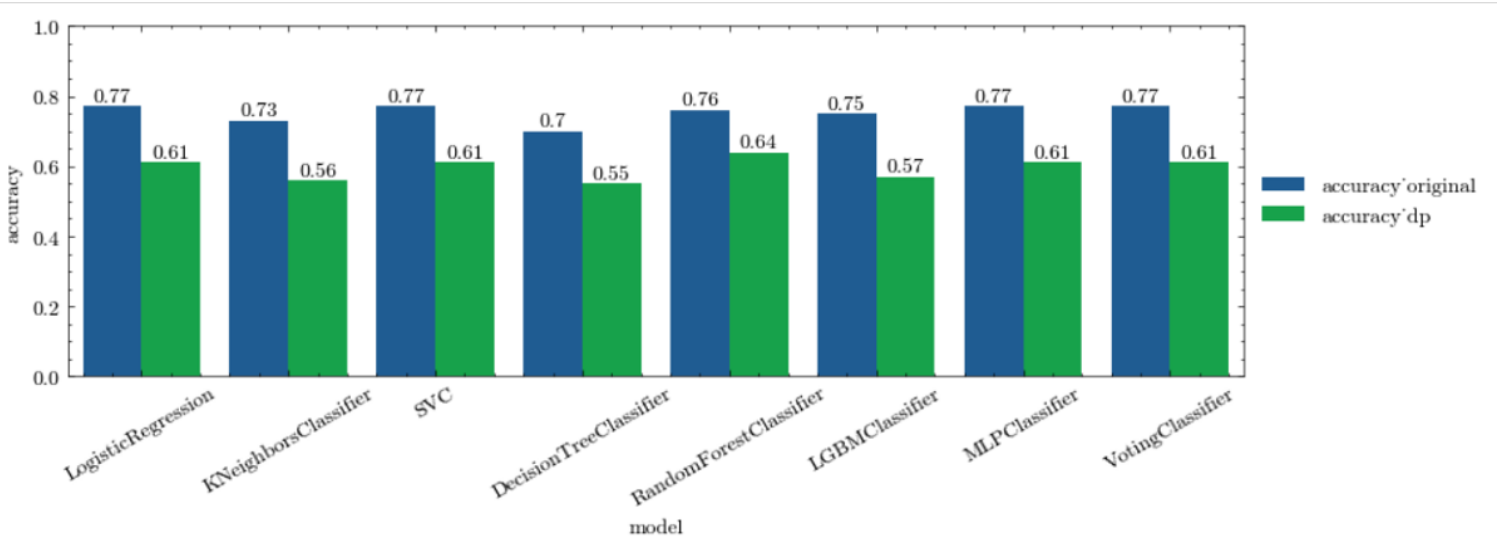



```
Comparison of accuracies for each models on the DP train/test dataset
Acc. mean on train set: 0.69 / on test set: 0.59
```



Model Accuracy on Original vs. DP Dataset

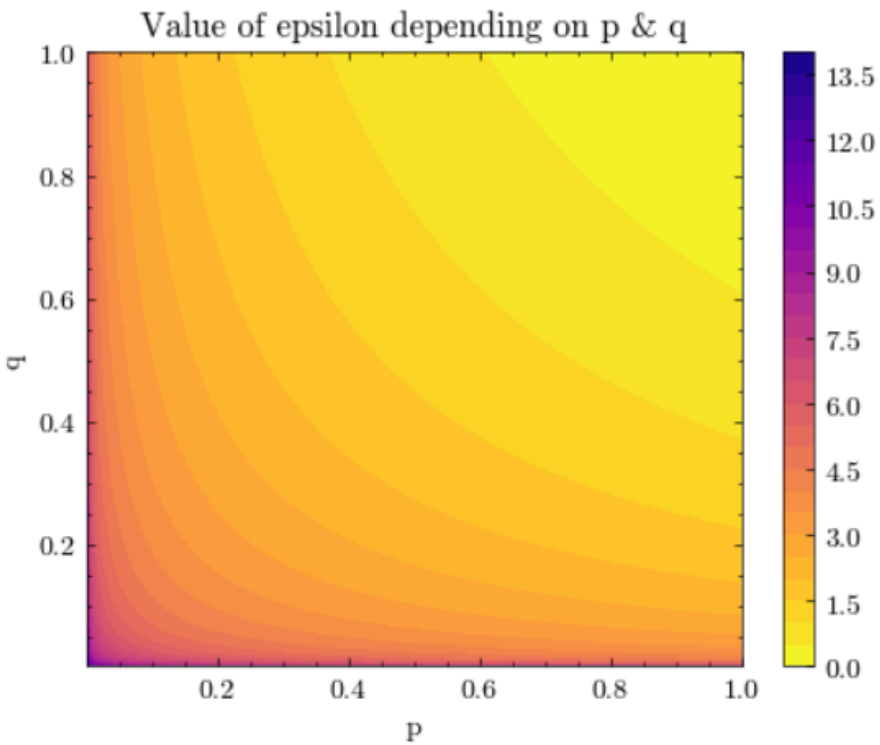
- Comparing model performance on the original and differentially private datasets.
- Highlights the trade-off between data privacy and model accuracy.

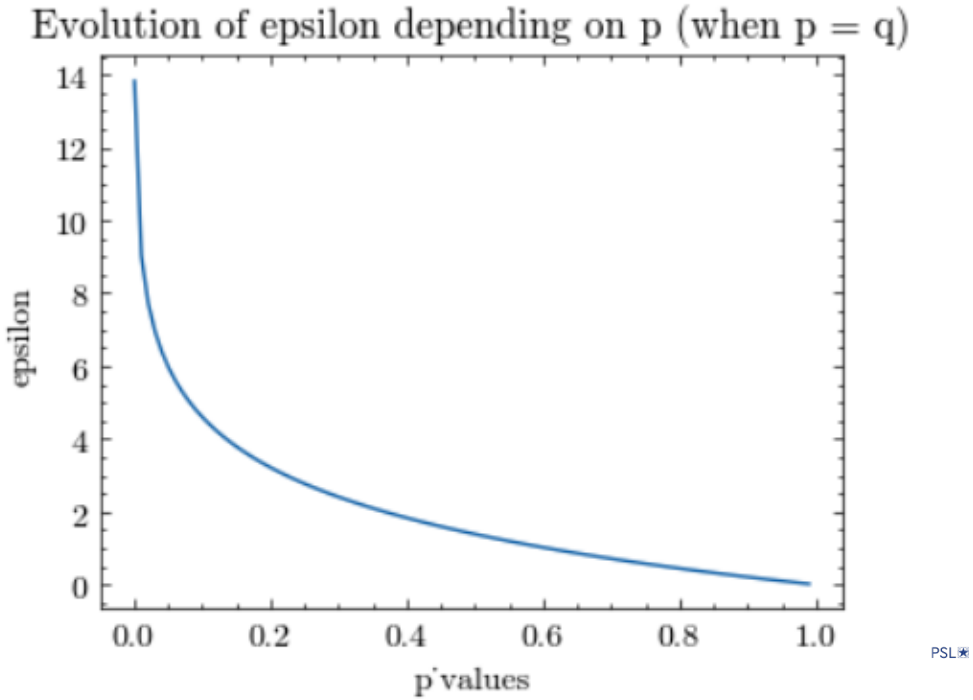


Accuracy Trends with Varying Epsilon

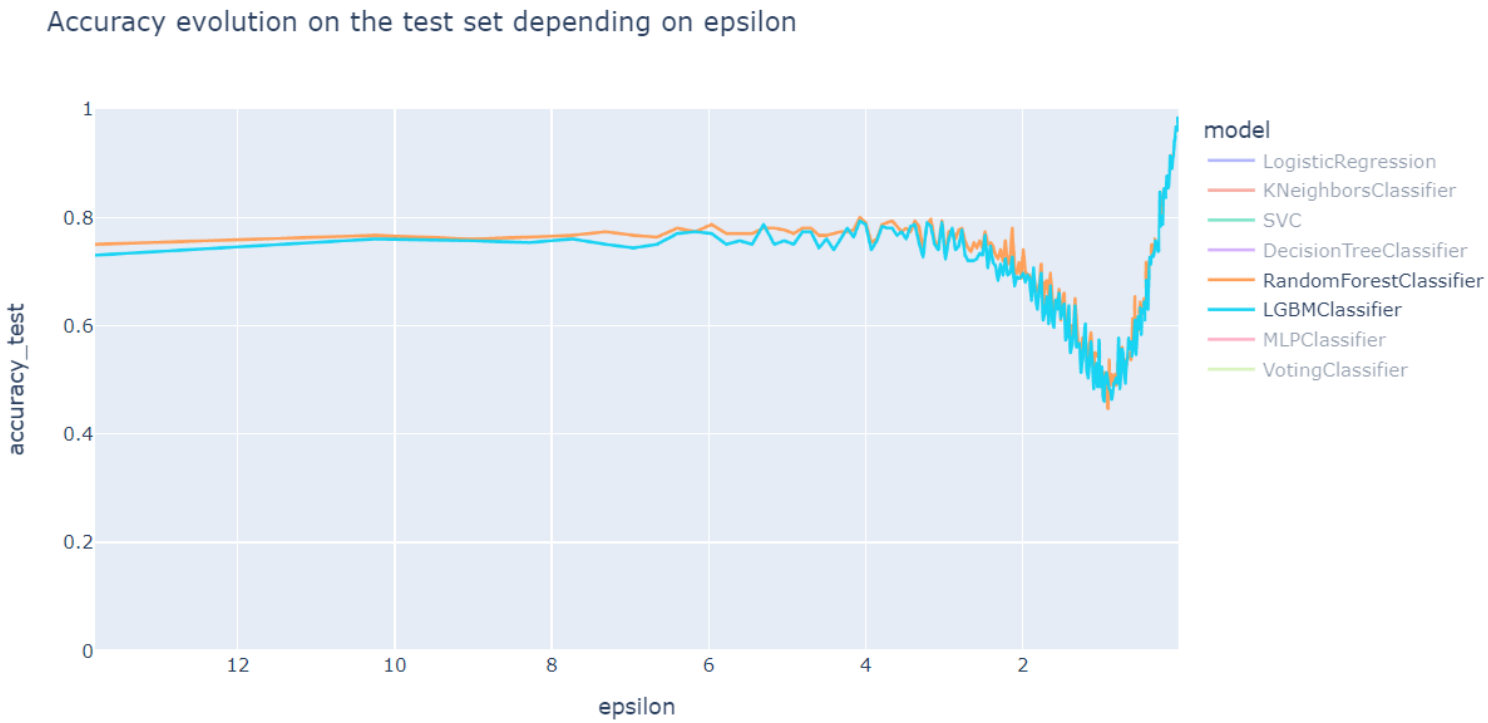
Determining ϵ:

- $\epsilon = -\ln(p \cdot q)$, where:
  - p is the probability of adding a random value instead of the true value.
  - q is the probability of assigning the random value as "has diabetes" (1) in the binary classification.

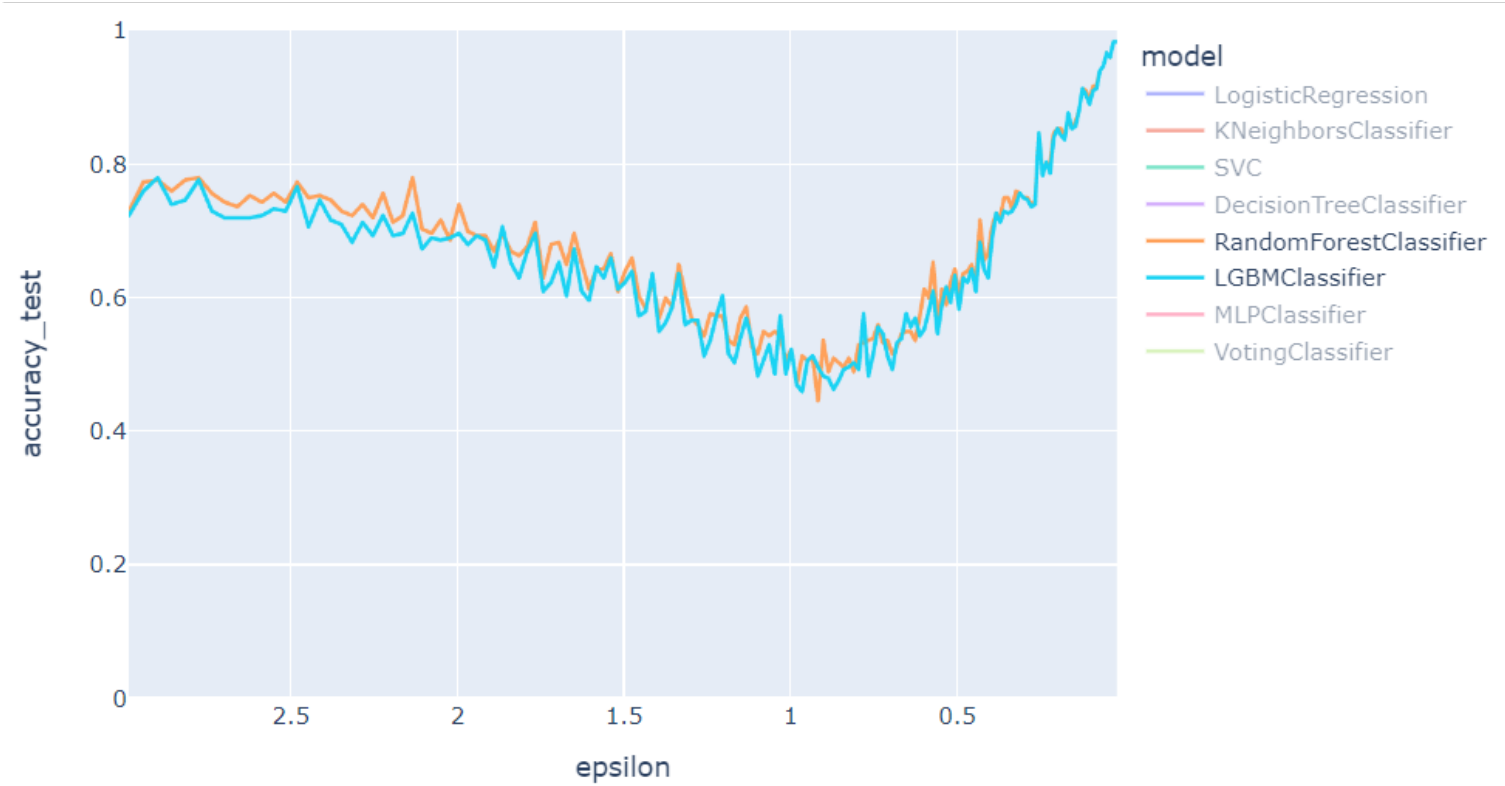




Accuracy Trends with Varying Epsilon

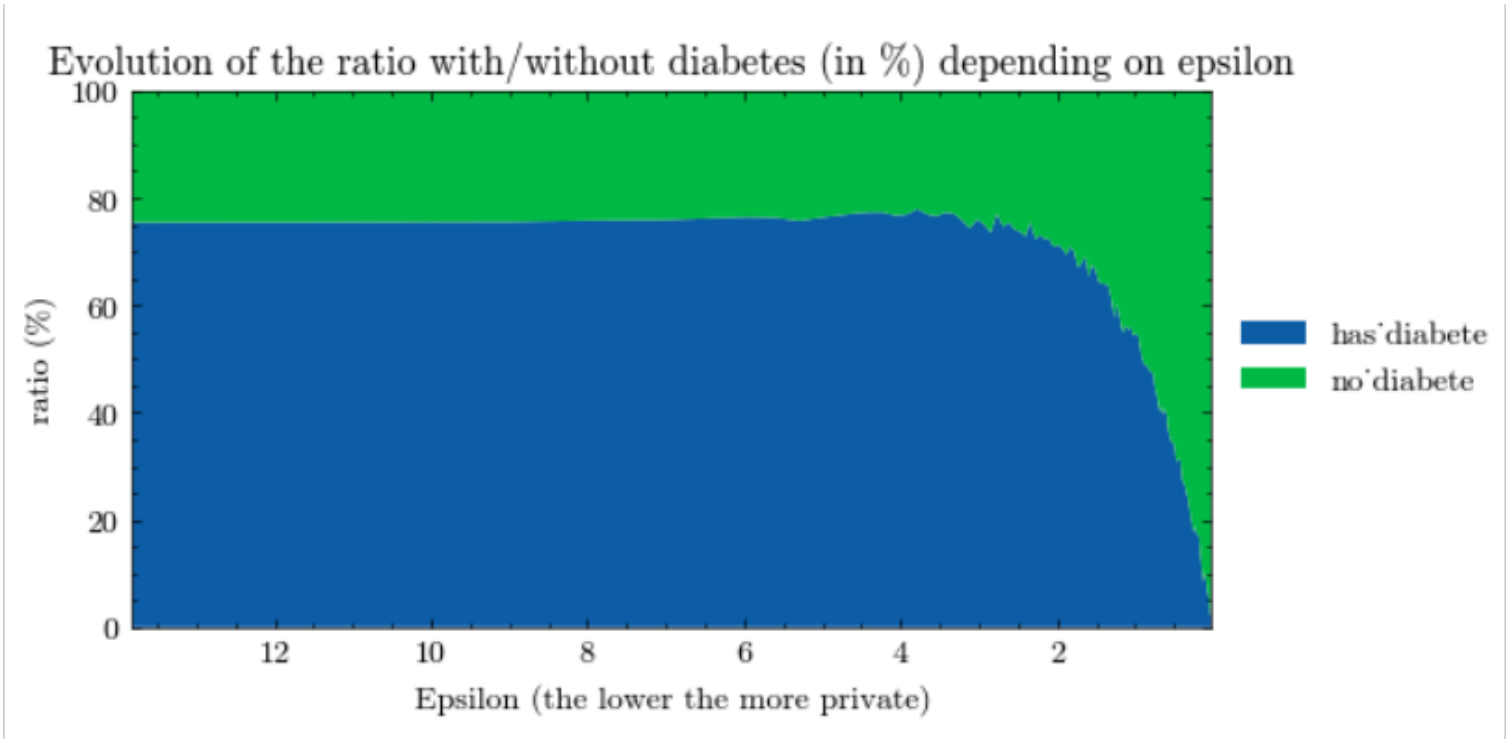


Accuracy Trends with Varying Epsilon



Accuracy Trends with Varying Epsilon

Accuracy Trends with Varying Epsilon



Concluding Remarks on Differential Privacy

Differential Privacy Strengths:

- Provides robust privacy without presuming the attacker's strategy or data access.
- Quantifiable privacy risk with mathematical rigor, through the epsilon parameter.

Trade-offs and Considerations:

- Performance impact: There's an inherent balance between privacy levels and model accuracy.
- Metrics re-calibration may be necessary for imbalanced datasets within this privacy framework.

Looking Ahead:

- Exploring alternatives like homomorphic encryption for secure data computation.
- Anticipating advancements to refine open data sharing while safeguarding against data breaches.

Dauphine | PSL

Adversarial Attacks

- Definition: Small, intentional perturbations that cause a model to make errors.
- Real-world implications: Security, reliability of AI systems.
- Loss function: $J(\theta, x, y)$
- Objective of the attacker: Maximize the loss.

Types of Adversarial Attacks

- White-box attacks: Attacker has complete knowledge of the model.
- Black-box attacks: Attacker has no knowledge of the model's internals.

$$x' = x + \delta$$

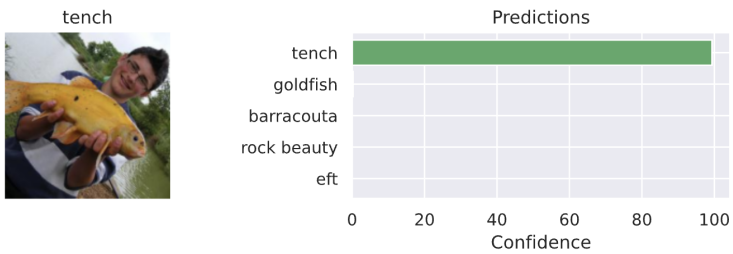


Figure: Original Prediction

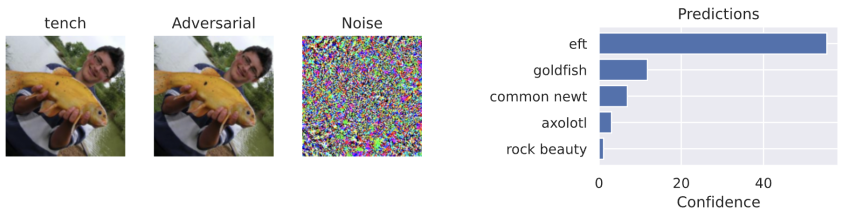


Figure: After FGSM Attack

Convolutional Neural Network Architecture

- **Convolutional Layers:**
  - 4 Convolutional layers with kernel size 3 and padding 1.
  - Increasing channels from $3 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256$.
- **Batch Normalization:**
  - Applied after each convolutional layer to stabilize learning.
- **Pooling and Dropout:**
  - Max pooling with a 2x2 window to reduce dimensions.
  - Dropout with a probability of 0.5 to prevent overfitting.
- **Fully Connected Layers:**
  - 3 Linear layers reducing dimensions from 256 to 10.
- **Activation:**
  - Uses ReLU (Rectified Linear Unit) activation functions.
- **Output:**
  - Log Softmax activation for classification.

Dauphine | PSL

Identifying Sensitive Pixels in Images

Objective:

- To determine which pixels significantly influence the model's predictions.
- Highlighting the pixels that contribute most to the classification decision.

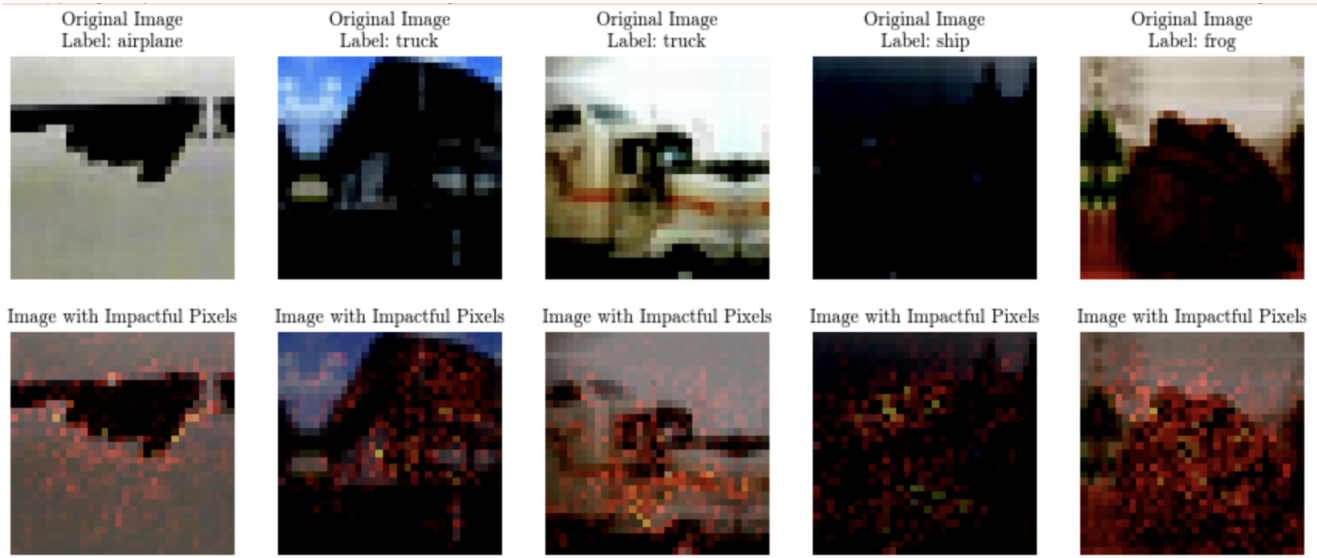
Methodology:

- The `most_impactful_pixels` function computes gradients of the model's loss with respect to the input image, identifying pixels that affect the output.
- The `overlay_impactful_pixels` function overlays these gradients onto the original image to visualize the sensitive areas.

Finding Sensitive Pixels

Results:

- Heatmaps indicate areas in the image that are most sensitive to changes.



Fast Gradient Sign Method (FGSM)

- Simple and powerful attack method.
- Generates adversarial examples as: $x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$.

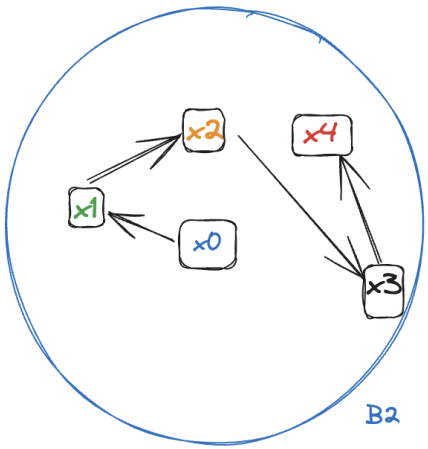
Projected Gradient Descent (PGD)

- Extends the FGSM approach by applying it multiple times with small step sizes.
- Formally, it solves the optimization problem: $x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^t, y)))$, where Π denotes the projection onto the set S of allowed perturbations.

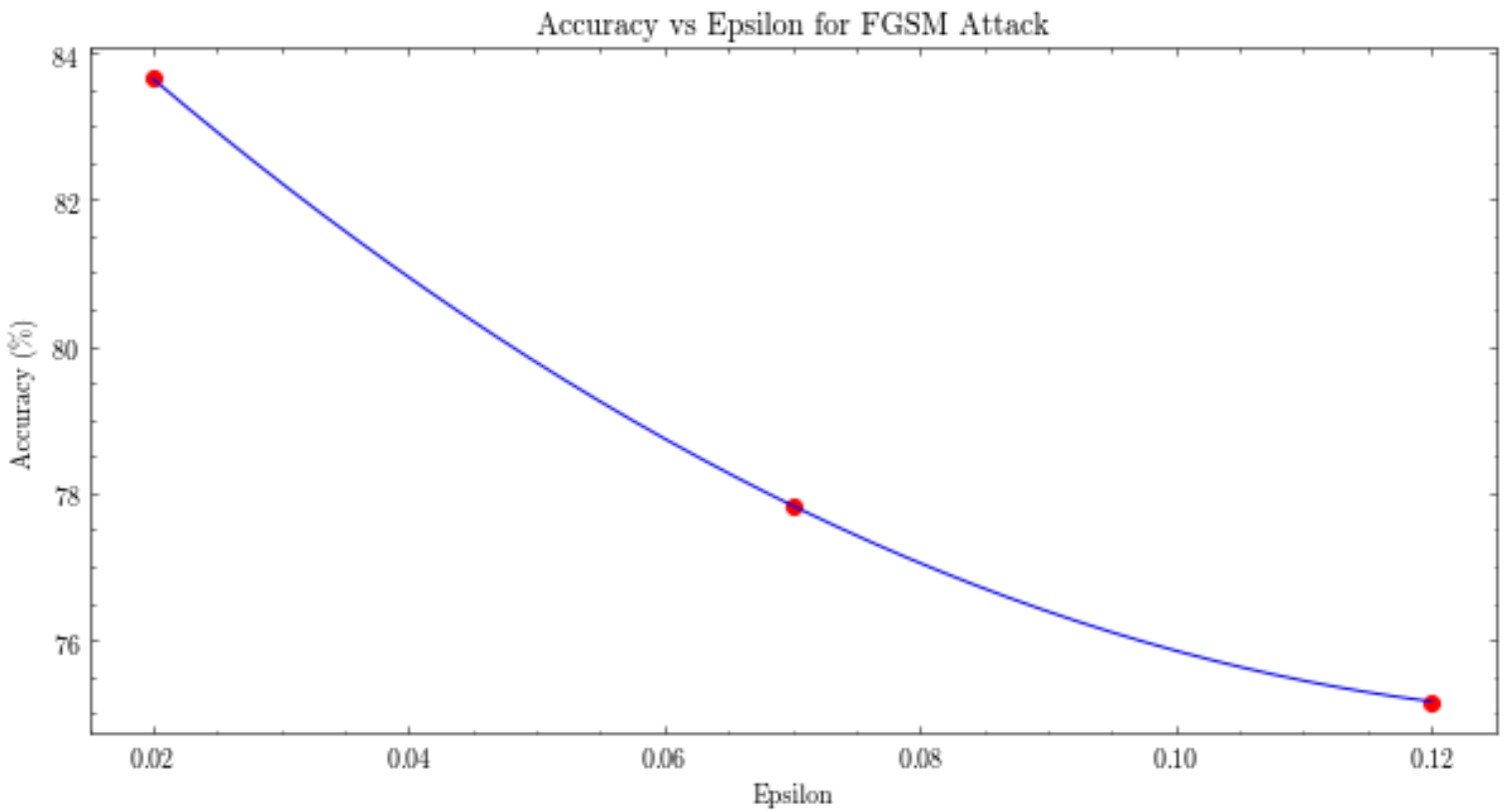
Projected Gradient Descent with L_2 Norm (PGD-L_2)

- An iterative attack method enforcing the L_2 norm constraint.
- Provides a balance between perturbation magnitude and attack efficacy.
- The L_2 norm measures the Euclidean distance of the perturbation vector.
- The PGD-L_2 update rule can be formulated as:

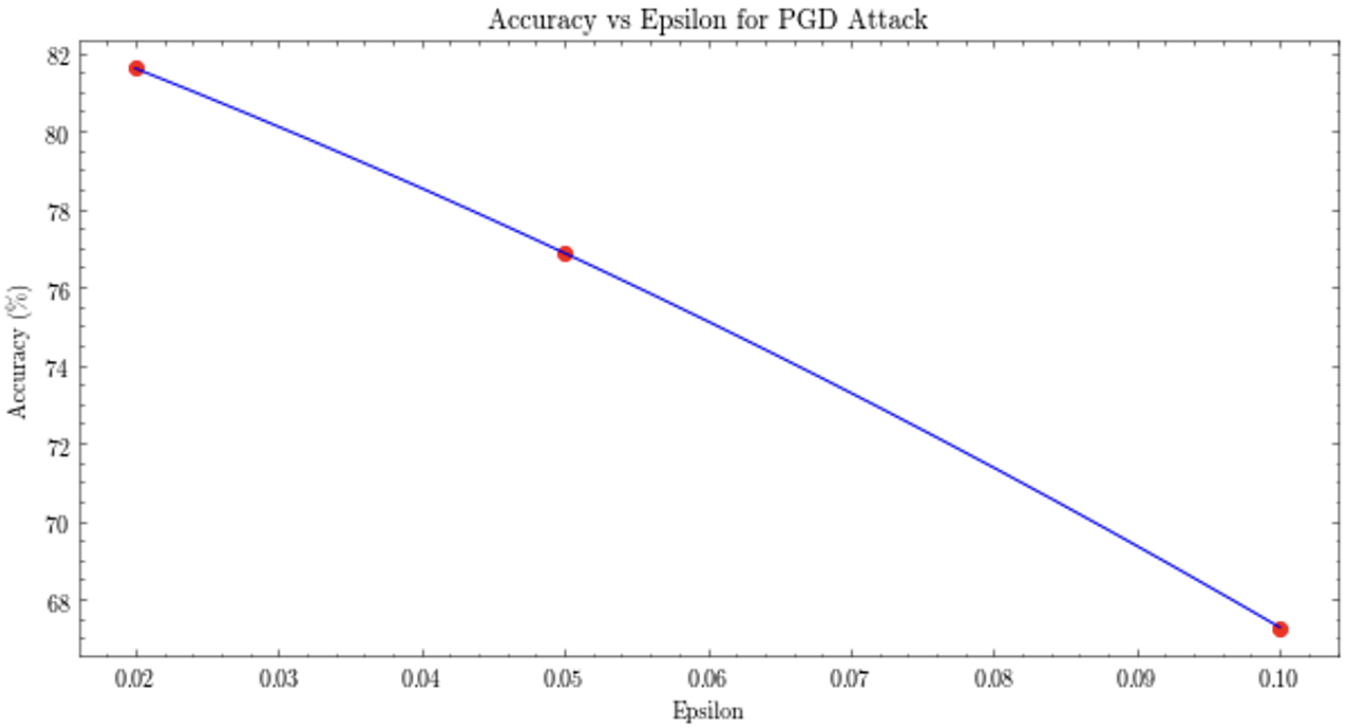
$$x^{t+1} = \Pi_{x+\epsilon \cdot B_2} \left(x^t + \alpha \cdot \frac{\nabla_x J(\theta, x^t, y)}{\|\nabla_x J(\theta, x^t, y)\|_2} \right)$$



Adversarial Training: FGSM



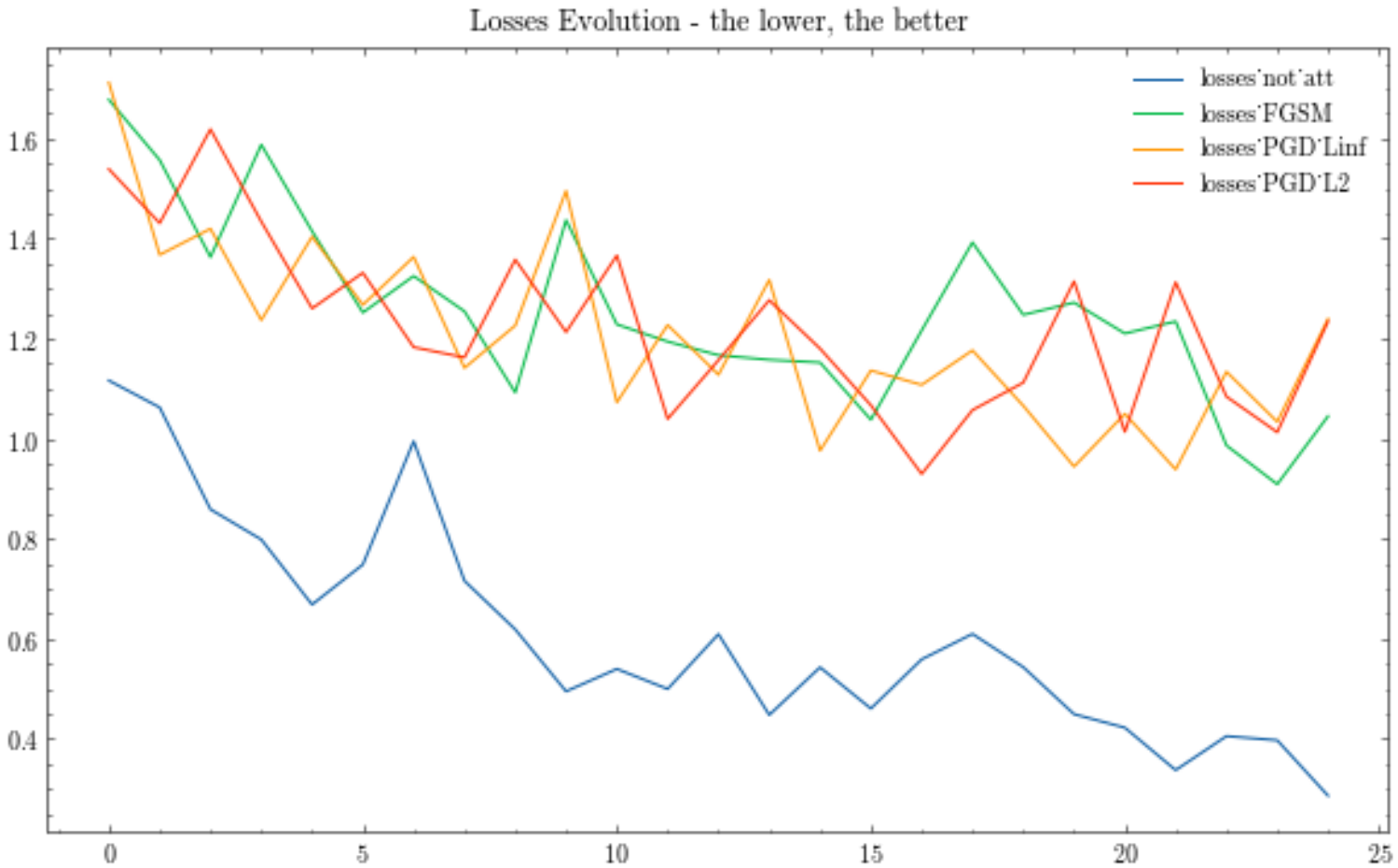
Adversarial Training: PGD L_∞



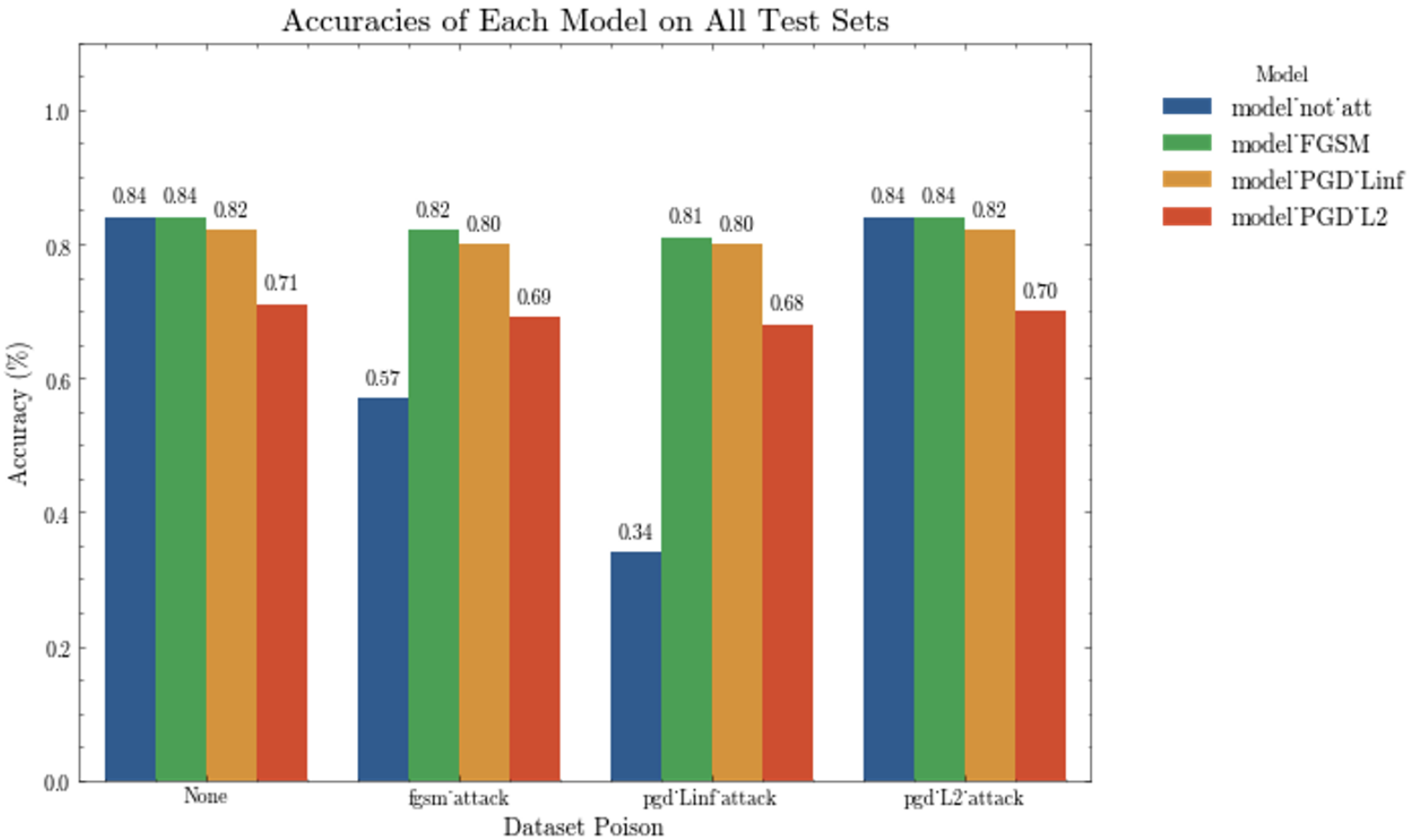
Adversarial Training: Variables

- **Model**:
  - Epochs Number: 25
  - Learning Rate: 1e-3
- **FGSM**:
  - Epsilon: 0.02
- **PGD Linf**:
  - Epsilon: 0.02
  - Alpha: 0.01
  - Number of Iterations: 40
- **PGD L2**:
  - Epsilon: 1.35
  - Alpha: 0.05
  - Number of Iterations: 40

Adversarial Training: Losses



Adversarial Training: Accuracy



Adversarial Training

Adversarial training:

- A simple but yet effective way to defend against attacks
- Very attack specific
- Not robust to black-box attacks
- Can be bypassed by 2-steps attacks

Local Intrinsic Dimensionality (Ma et al. 2018)

• Goal: estimate the density of clean samples around the current point using the theory of Local Intrinsic Dimensionality

- **Detection Mechanisms:** Utilizes LID metrics in intermediate data representations within neural networks.
- **Effectiveness:** Shows promise in high-dimensional spaces, typical in deep learning models.
- **Limitations:** Not foolproof; adversaries may craft attacks that evade LID-based detection.

Results:

- Estimated LID: 19.35 for the baseline model.
- Estimated LID: 18.87 for FGSM attack.
- Estimated LID: 18.85 for PGD (L-inf norm) attack.

No Free Lunch

The Delicate Balance:

- Active research underscores a fundamental challenge in machine learning: enhancing accuracy often increases susceptibility to adversarial attacks.

Insightful Findings:

- Studies, such as those by Dohmatob in 2018, have shown that across diverse data distributions, classifiers can be vulnerable if they focus solely on accuracy.

Implications for AI Security:

- There is a critical need for models that maintain both high predictive performance and resilience against deliberate manipulations.

Forward Path:

- Developers and researchers are encouraged to adopt a holistic view of model evaluation that equally weighs accuracy and adversarial robustness.

Concluding Reflections on Data Privacy and Security

Synthesis of Key Insights:

- Our analysis has highlighted the critical balance between data accessibility and the imperative for confidentiality in data privacy, particularly in the healthcare domain.
- Our study reveals a significant aspect of machine learning models: despite their high accuracy, they remain susceptible to adversarial attacks, a fact underscored by our experiments with the CIFAR dataset.

Closing Thought:

- A balanced approach is essential in machine learning to ensure both performance and ethical use.

Thank You

Thank you for your attention!