Exploring Intrinsic Dimension Estimation for Enhanced Machine Learning Security

Seonghun Son, Debopriya Roy Dipta, Seyedmohammad Kashani, Grace Heron, Dr.Berk Gulmezoglu (Faculty)
Project Mentor: Dr.Bradford Kline, National Security Agency (NSA)

Project Statement

❖ **Motivation:**

- Machine learning models are actively utilized in security critical applications.
- The complexity of the dataset is on the rise and accommodates high dimensionality with a large number of features.
- There are several issues with representing and embedding data in wastefully large dimensions:

➢ ***Computing Resources:*** Requires more memory and computing power

➢ ***Accuracy:*** Dimension reduction generally improves classification results

➢ ***Security:*** An increased attack surface for adversarial attacks

❖ **Proposed Solution:**

1) Create a generalized Intrinsic Dimension Estimator (ID-E) tool to eliminate the insignificant dimensions from a dataset.

2) Leverage the ID-E tool to create a mitigation technique against adversarial attacks.

Methodology

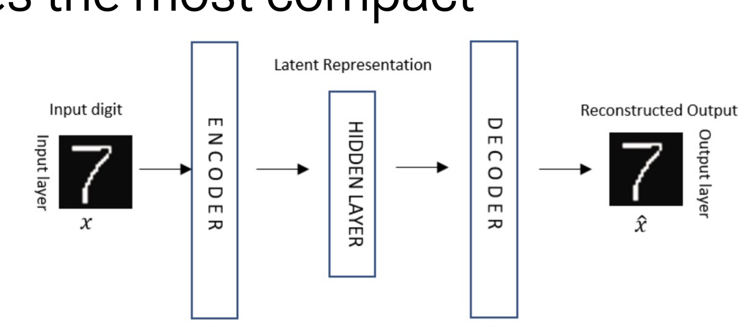
❖ ***Intrinsic Dimension Estimator (ID-E) Tool :***

1) Dataset:

- 8 different synthetic lab-generated datasets are used for the experiments (Created by our Project Manager, Dr. Bradford Kline)
- The datasets are diverse in terms of noise and complexity.
- Seven of the datasets are n-long feature vectors, while one is a collection of square grayscale images (m-by-m matrices).

2) Implementing Autoencoder (AE):

- Our purpose is to learn a compact input data representation, capturing its significant features in the latent space.
- The intrinsic dimension (ID) signifies the most compact representation of the input dataset.
- We gradually decrease the dimension of the latent space from the full dimension of the data.

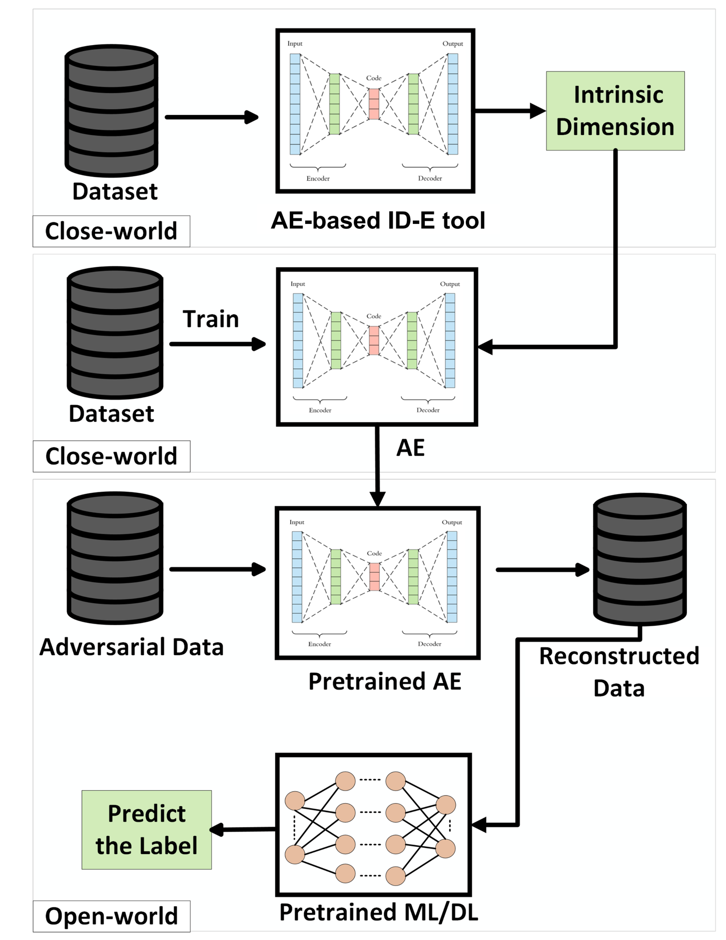


- The decoder reconstructs the image based on the features from the latent space during each iteration.
- The mean square error (MSE) of the reconstructed image is calculated during each iteration.
- The dimension at which the MSE function provides a knee-point corresponds to the ID of the dataset.

Observation 1: The "linear" activation function makes a clearer output than the conventional activation functions.

Observation 2: Vanilla Autoencoder estimates the intrinsic dimension (ID) value better than other Autoencoder types.

Other Autoencoder types tested in the ID-E tool:
- Regularized Autoencoder (RAE)
- Variational Autoencoder (VAE)
- Sparse Autoencoder (SAE)



❖ ***ID-E based mitigation tool***

1) Crafting adversarial examples

➢ Fast Gradient Sign Method (FGSM)

➢ Basic Iterative Method (BIM)

2) Building the Mitigation tool

- Finding the intrinsic dimension of a dataset using the AE-based ID-E tool
- Training an Autoencoder (AE) with the predetermined ID.
- Transforming adversarial data with pre-trained AE to filter out the induced perturbations through reconstruction.
- Feeding the reconstructed data into the pre-trained ML/DL model to decrease the success rate of the adversarial attack

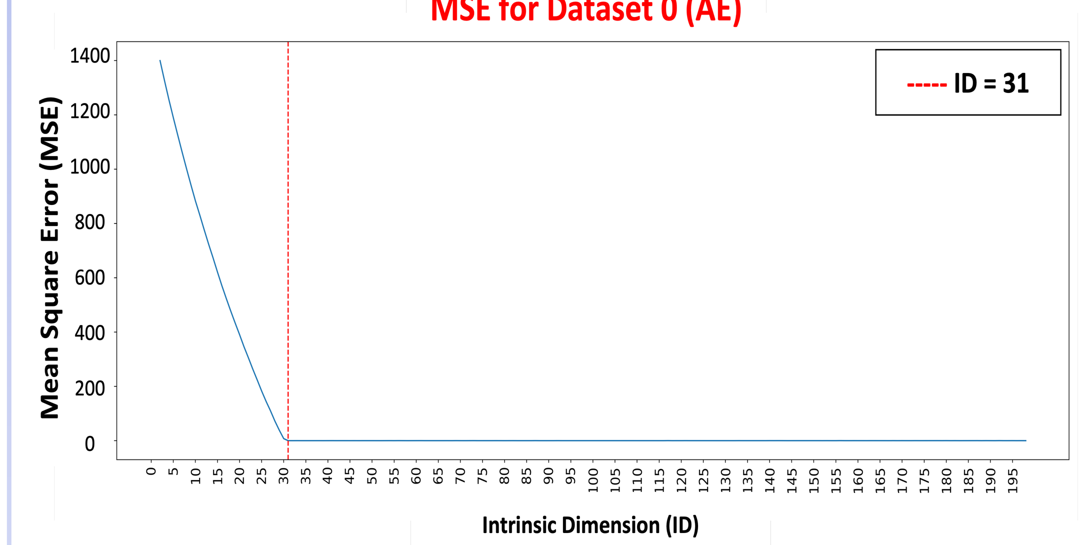
Results

❖ ***Intrinsic Dimension Estimator (ID-E) tool results***

a) One-dimensional lab-generated dataset

- Seven different datasets with different Intrinsic Dimension (ID) values
- Knee-point (red line) predicts the exact ID values of each given dataset

MSE for Dataset 0 (AE)

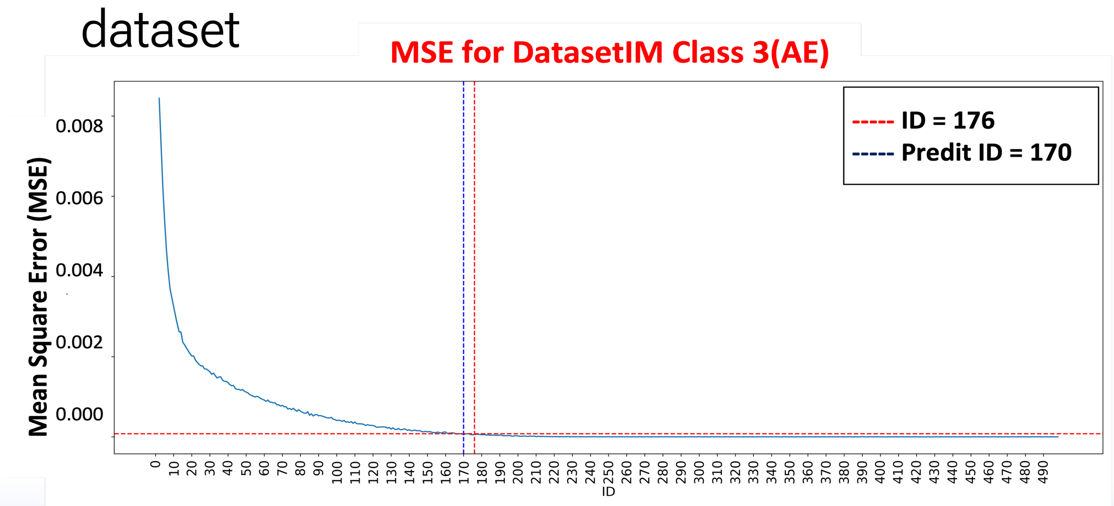


Dataset	Dimension	Answer ID	Predict ID
Dataset 0	200	30	31
Dataset 1	200	131	131
Dataset 2	150	64	65
Dataset 3	175	110	111
Dataset 4	200	37	38
Dataset 5	100	73	82
Dataset 6	28	20	20

b) Two-dimensional lab-generated dataset

- Six different classes with different Intrinsic Dimension (ID) values
- Knee-point (blue line) predicts the exact ID values (red line) of each given dataset

MSE for DatasetIM Class 3(AE)

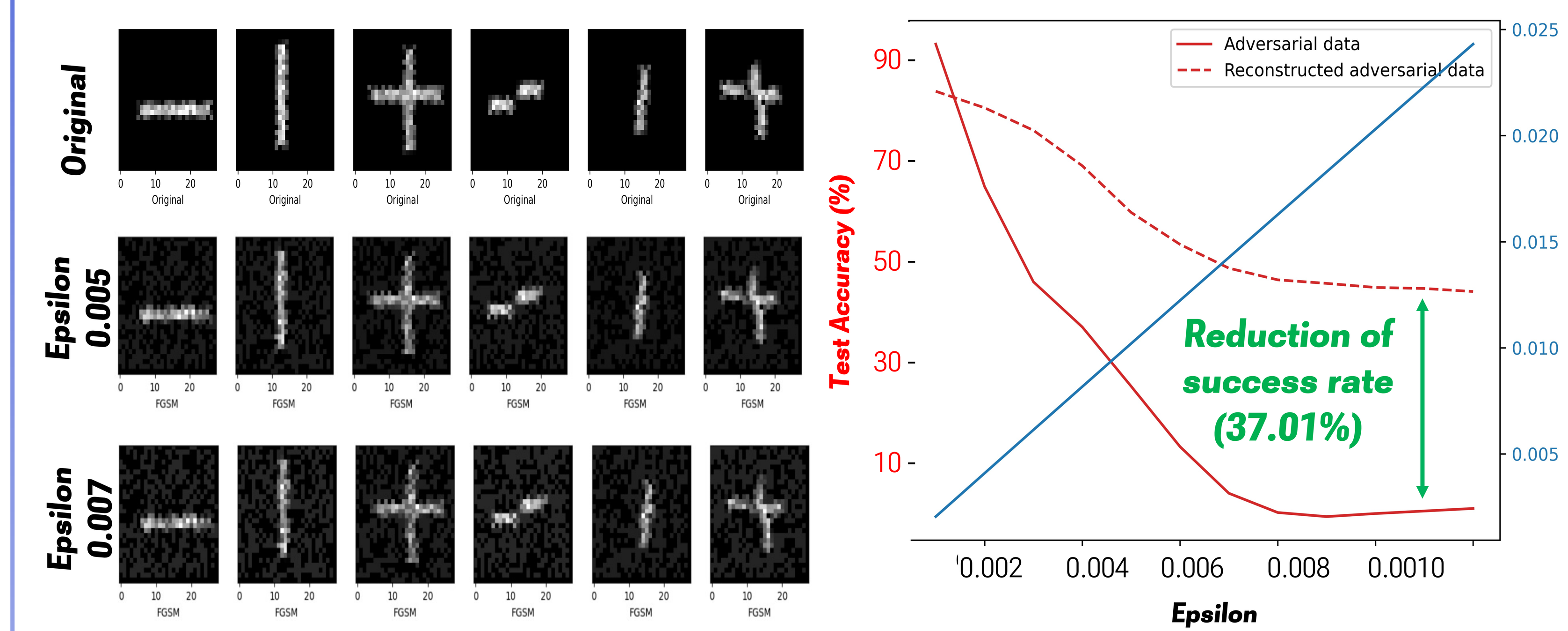


Class	Answer ID	Predict ID
Class 0	208	230
Class 1	208	230
Class 2	352	470
Class 3	176	170
Class 4	176	170
Class 5	288	320
Full Class	352	360

❖ ***ID-E based mitigation tool against Adversarial attacks:***

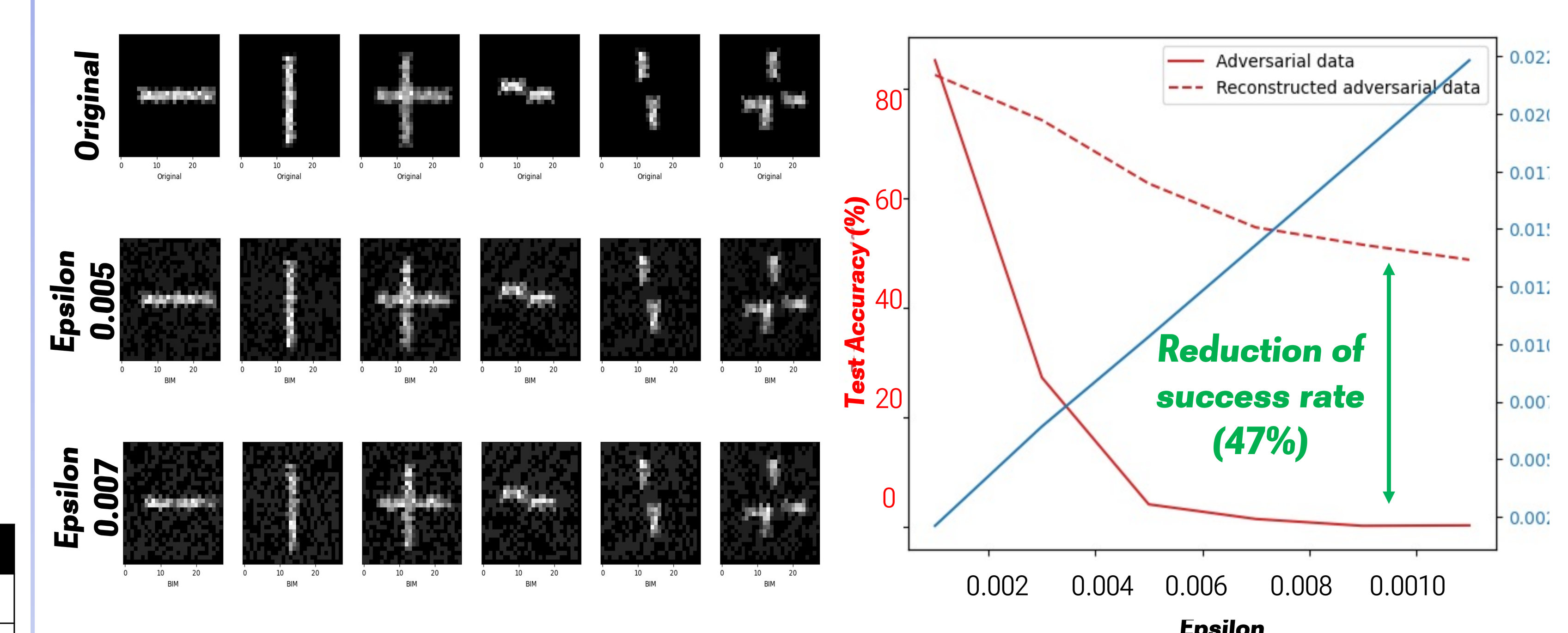
a) Fast Gradient Sign Method (FGSM)

- CNN classification accuracy drops below **20%** after applying the FGSM adversarial attack (epsilon=0.005).
- The ID-E tool restores the classification accuracy to over ***60%*** (epsilon=0.005).



b) Basic Iterative Method (BIM)

- CNN classification accuracy drops below ***10%*** after applying the BIM adversarial attack (epsilon=0.005).
- The ID-E tool restores the classification accuracy to over ***65%*** (epsilon=0.005).



Conclusion

- We created an Intrinsic Dimensional Estimation (ID-E) Tool using Autoencoder.
- The performance of our ID-E Tool is promising in finding the Intrinsic Dimension (ID) value.
- Created FGSM and BIM method Adversarial attacks on lab-generated datasets and achieved ***16%*** and ***20%***, respectively.
- We have successfully mitigated adversarial attacks on image datasets by achieving a classification accuracy of over ***65%***.

References

[1] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," 2016.
[2] B. Ghojogh, M. N. Samad, S. A. Mashhadi, T. Kapoor, W. Ali, F. Karray, and M. Crow- ley, "Feature selection and feature extraction in pattern analysis: A literature review," 2019.
[3] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in Proceedings 2018 Network and Distributed System Security Symposium, ser. NDSS 2018. Internet Society, 2018. [Online]. Available: http://dx.doi.org/10.14722/ndss.2018.23198
[4] H. Torabi, S. L. Mirtaheri, and S. Greco, "Practical autoencoder based anomaly detec- tion by using vector reconstruction error," Cybersecurity, vol. 6, no. 1, p. 1, 2023.