Securing On-Device AI From Model Probing Attacks Using Autoencoder

Overview:

This paper summarizes the main findings of the project, including evaluation on CIFAR-10 and ImageNet datasets, the role of synthetic data, and robustness testing against various adversarial probing attacks. Results highlight the detection accuracy, false positive rates, and performance under adversarial injection scenarios.

Variational Autoencoder Results:

The Variational Autoencoder (VAE) successfully generated synthetic benign data that closely matched the statistical properties of real queries. As can be seen in Figure 1, the feature distributions for the first logit, second logit, and margin loss showed stable alignment with their real counterparts. Redundancy analysis further confirmed the diversity of the generated data. As shown in Table 1, the synthetic set contained more than 72,000 unique sequences compared to about 12,000 real ones, with no exact duplication. This expansion provided broader coverage while avoiding replication of training data, ultimately strengthening generalization in the detection framework.

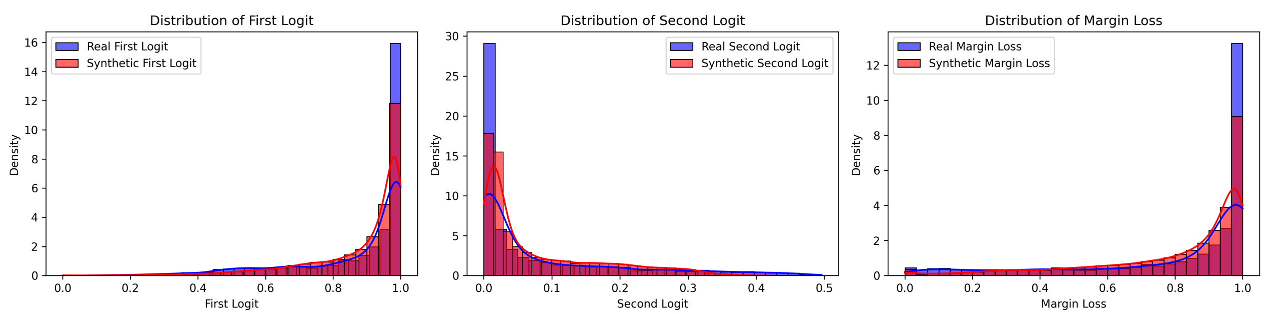

Figure 1. Distribution Alignment of First Logit, Second Logit, and Margin Loss for Real and Synthetic Queries

Table 1. VAE Redundancy Analysis

Metric/Feature	Synthetic Benign Data	Real Benign Data
Unique First Logit Values	24,663	4,074
Unique Second Logit Values	23,350	4,074
Unique Margin Loss Values	24,828	4,044
Exact Match Samples	0	0
Total Unique Sequences	72,841	12,192

Evaluation on CIFAR-10 Datasets:

The Autoencoder (AE) trained exclusively on real benign CIFAR-10 queries achieved consistently high detection rates. As shown in Table 2, the model maintained perfect classification for benign inputs with no false positives and achieved near-perfect detection across all six black-box probing attacks. The confusion matrix in Figure 2 illustrates this performance, where Square, QEBA, HSJA, Boundary, and SurFree attacks were detected with 100% accuracy, while NES remained highly effective with a recall of 0.98. These findings confirm the AE's ability to generalize strongly when exposed only to benign data during training.

Table 2. CIFAR-10 Classification Report

Class	Precision	Recall	F1-Score	Support
Benign	1	1	1	23
SurFree	1	1	1	224
QEBA	1	1	1	224
HSJA	1	1	1	224
Boundary	1	1	1	224
NES	1	0.98	0.99	224
Square	1	1	1	224

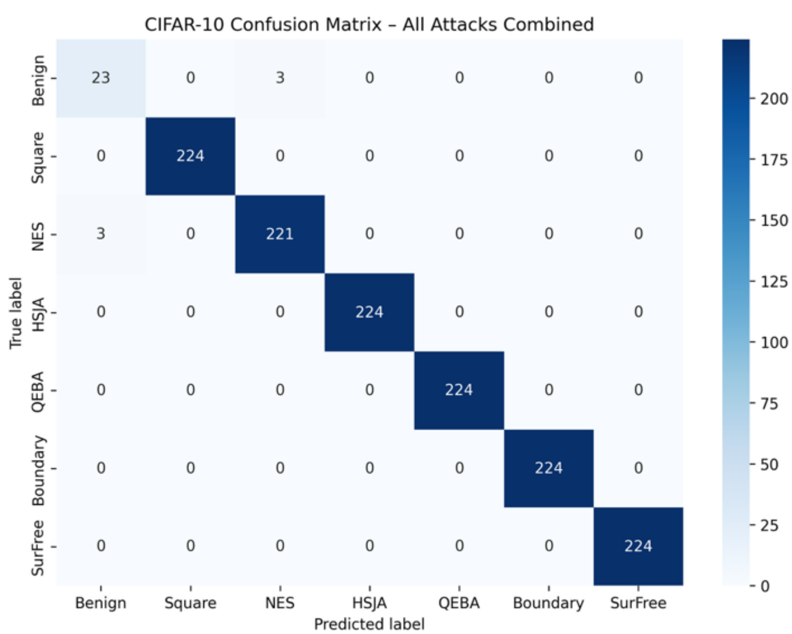

Figure 2. CIFAR-10 Confusion Matrix Results

Evaluation on ImageNet Datasets:

When evaluated on ImageNet, the AE was trained on synthetic benign data generated by the VAE but tested using real benign sequences. As can be seen in Table 3, the model maintained high accuracy and zero false positives, while successfully detecting nearly all adversarial attacks. SurFree, QEBA, HSJA, and Boundary reached perfect detection scores, while NES and Square remained slightly lower but still highly effective, with recalls of 0.96 and 0.98, respectively. The confusion matrix shown in Figure 3 highlights the balanced performance across both benign and adversarial classes, underscoring the robustness of the approach even under more complex dataset conditions.

Table 3. ImageNet Classification Report

Class	Precision	Recall	F1-Score	Support
Benign	1	1	1	58
SurFree	1	1	1	224
QEBA	1	1	1	224
HSJA	1	1	1	224
Boundary	1	1	1	224
NES	1	0.96	0.98	224
Square	1	0.98	0.99	224

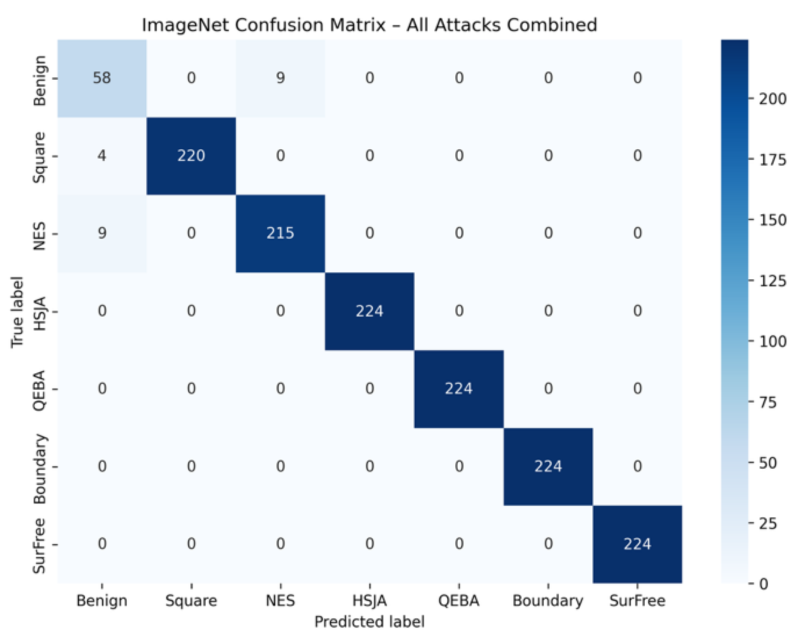

Figure 3. ImageNet Confusion Matrix Results

Adversarial Injection Test:

To simulate realistic conditions, adversarial sequences were injected into continuous benign query streams and transformed using strategies such as spike, stretched, ramped, hybrid, and fragmented modifications. As demonstrated in Table 4 and Figure 4, the AE consistently detected nearly all injected adversarial groups in CIFAR-10, including stretched and ramped variants with accuracy exceeding 99%. Similarly, on ImageNet, shown in Table 5 and Figure 5, the AE preserved strong detection performance across diverse variants. While fragmented attacks posed the greatest challenge due to their sparse structure, the model still generated meaningful alerts, confirming its sensitivity to incomplete or dispersed adversarial patterns.

Table 4. Autoencoder Detection Results for Injected Adversarial Sequences on CIFAR-10

Index	Attack	Variant	Range	Flagged (%)
1	QEBA	Stretched	222 – 670	100%
2	Square	Spike	581 – 607	100%
3	SurFree	Stretched	883 – 907	95.8%
4	NES	Ramped	1096 – 1320	99.6%
5	Boundary	Ramped	1523 – 1747	99.6%
6	HSJA	Ramped	1922 – 2146	99.6%
7	HSJA	Fragmented	2400 – 2404	75.0%
8	NES	Stretched	2688 – 2766	98.7%
9	Boundary	Spike	3155 - 3255	99.0%
10	QEBA	Spike	3592 – 3815	99.6%

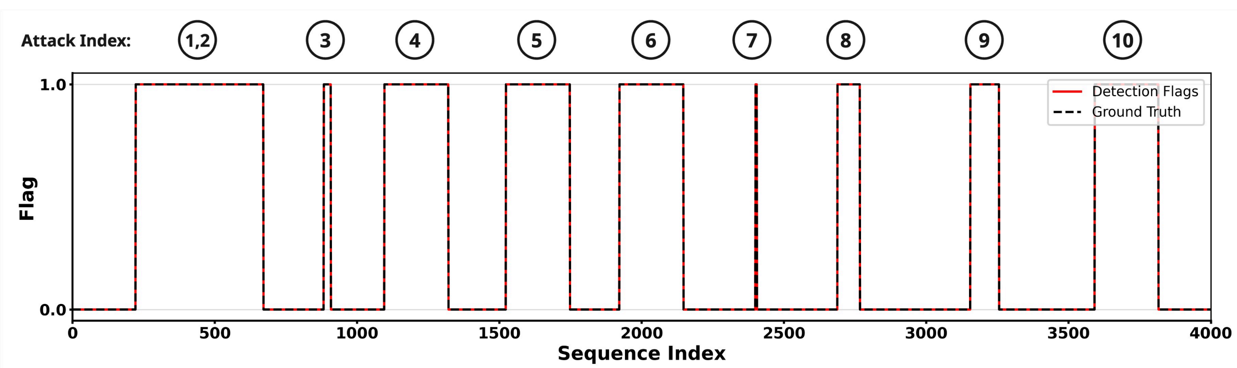

Figure 4. Visualization Results for Injected Adversarial Sequences on CIFAR-10

Table 5. Autoencoder Detection Results for Injected Adversarial Sequences on ImageNet

Index	Attack	Variant	Range	Flagged (%)
1	HSJA	Ramped	229 - 292	98.4%
2	NES	Ramped	490 – 515	96.0%
3	QEBA	Ramped	722 – 767	97.9%
4	SurFree	Spike	1054 – 1102	97.9%
5	Square	Hybrid	1328 – 1426	99.0%
6	Boundary	Fragmented	1559 – 1634	98.7%
7	HSJA	Ramped	1936 – 1944	87.5%
8	Boundary	Stretched	2161 – 2359	99.5%
9	QEBA	Stretched	2632 – 2932	99.7%
10	SurFree	Fragmented	3123 - 3198	98.7%

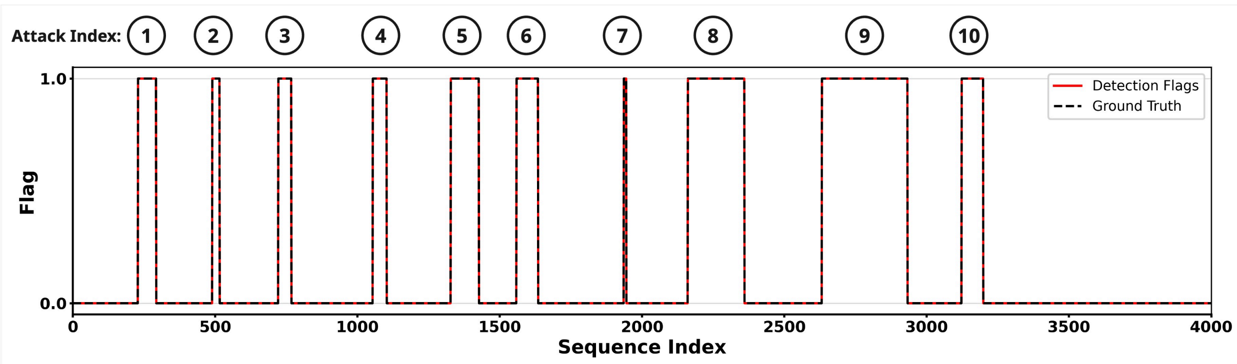

Figure 13. Visualization Results for Injected Adversarial Sequences on ImageNet

Model Efficiency:

The AE maintained a lightweight architecture with approximately 44,000 parameters, of which 43,719 are trainable. The final model occupied only about 650 KB in memory. As tested during evaluation, inference averaged 0.0026 seconds per query and achieved throughput of roughly 23,000 queries per second. These results indicate the system's suitability for deployment in resource-constrained environments without sacrificing detection quality.

Review:

Overall, the results demonstrate that the proposed Autoencoder framework effectively distinguishes between benign and adversarial queries across both CIFAR-10 and ImageNet datasets. Detection remained reliable even when synthetic benign data was used for training, validating the generalization capability of the model. Adversarial injection tests further confirmed its resilience under adaptive and concealed attack strategies, while maintaining efficiency with minimal memory overhead. These findings establish the framework as a practical and scalable defense mechanism for securing on-device machine learning systems against probing-based threats.