AIShield
Powered by Bosch

Vulnerability Analysis

Image Classification

Attack Type
Evasion

Date
2024-03-20

Author
AIShield

Job Id
gAAAAABl...XcNsvvnQ==

Executive Summary:

Our analysis unveils substantial vulnerabilities, highlighting a heightened risk of attackers exploiting the model through Evasion with 72.0% efficacy. These vulnerabilities could lead to significant financial, legal, and reputational repercussions. Employing AIShield's sample attack vectors for adversarial hardening or implementing our Threat Informed Defense Engine for real-time protection are essential. Proactive measures will act as enablers of regulatory preparedness and minimize potential risks.

This assesment aligns with the following OWASP Machine Learning Security Top-10 Vulnerabilities(v0.3 Draft):

- ML01:2023 Input Manipulation Attack

The proposed measures align with these CVE/CWE Entries:

- CWE-20: Improper Input Validation
- CWE-707: Improper Neutralization
- CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
- CWE-1288: Improper Validation of Consistency within Input

For more details about OWASP Vulnerabilities, please refer to Section 4.4 of the report.
For more details about CVE/CWE Entries, please refer to Section 4.5 of the report

1. Security:

1.1 Strength Vs Efficacy on Extracted Model:

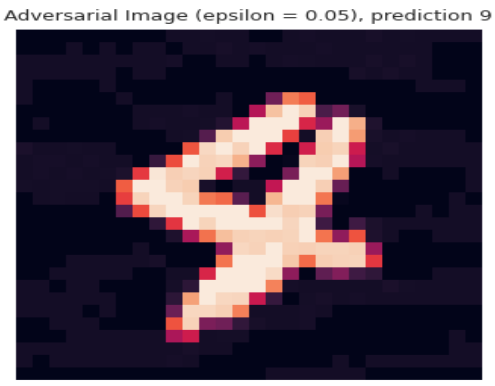
Below table shows the change in epsilon vs attack efficacy

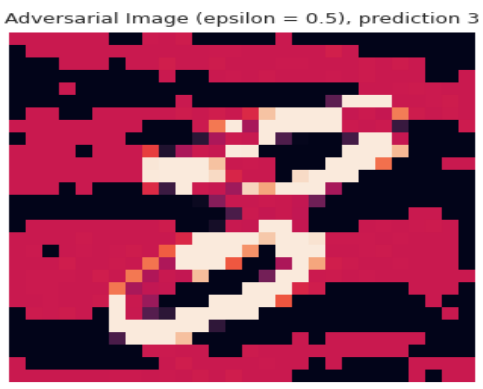
Epsilon	Attack Efficacy %
0.05	2.0
0.1	4.0
0.2	17.0
0.3	44.0
0.4	63.0
0.5	72.0

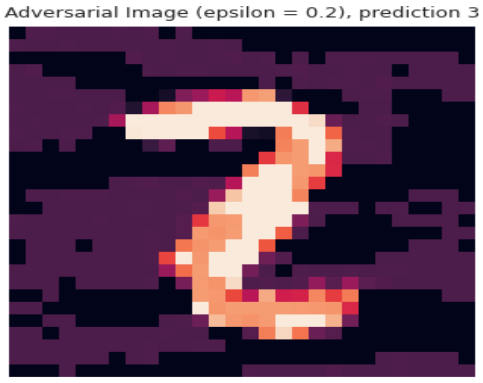
For more details on the metrics, refer to the Appendix Section of the report

1.2 Adversarial samples:

Below images represent original images and their repsective adversarial images







2. Performance:

2.1 Inference Time of models:

Inference time of a model refers to the amount of time it takes for the model to generate predictions on new input data. It is the time taken by the model to process input data and produce output. inference time is influenced by factors such as the model's complexity, the input data's size, and the computational resources available for inference. It is an important metric to consider when deploying machine learning models in production environments where fast and efficient processing is necessary.

In the context of this scenario, Original Model Inference Time in ms is 55.85.

More details on the hardware infrastructure, and inference time distributions will be populated in the upcoming reports.

2.2 Accuracy and F1 score of Models (Model):

Accuracy and F1 score on original data is necessary to understand the model performance even under imbalanced data distribution. The below table represents the performance comparison of two models - the Original model and the Extracted model - based on the number of samples, their relative accuracy, and the relative F1 score performed on original data as ground truth.

Model	Number of Samples	Model Accuracy	F1 Score
Original Model	6000	99.0	0.99

2.3 Adversarial Accuracy on Extracted Model:

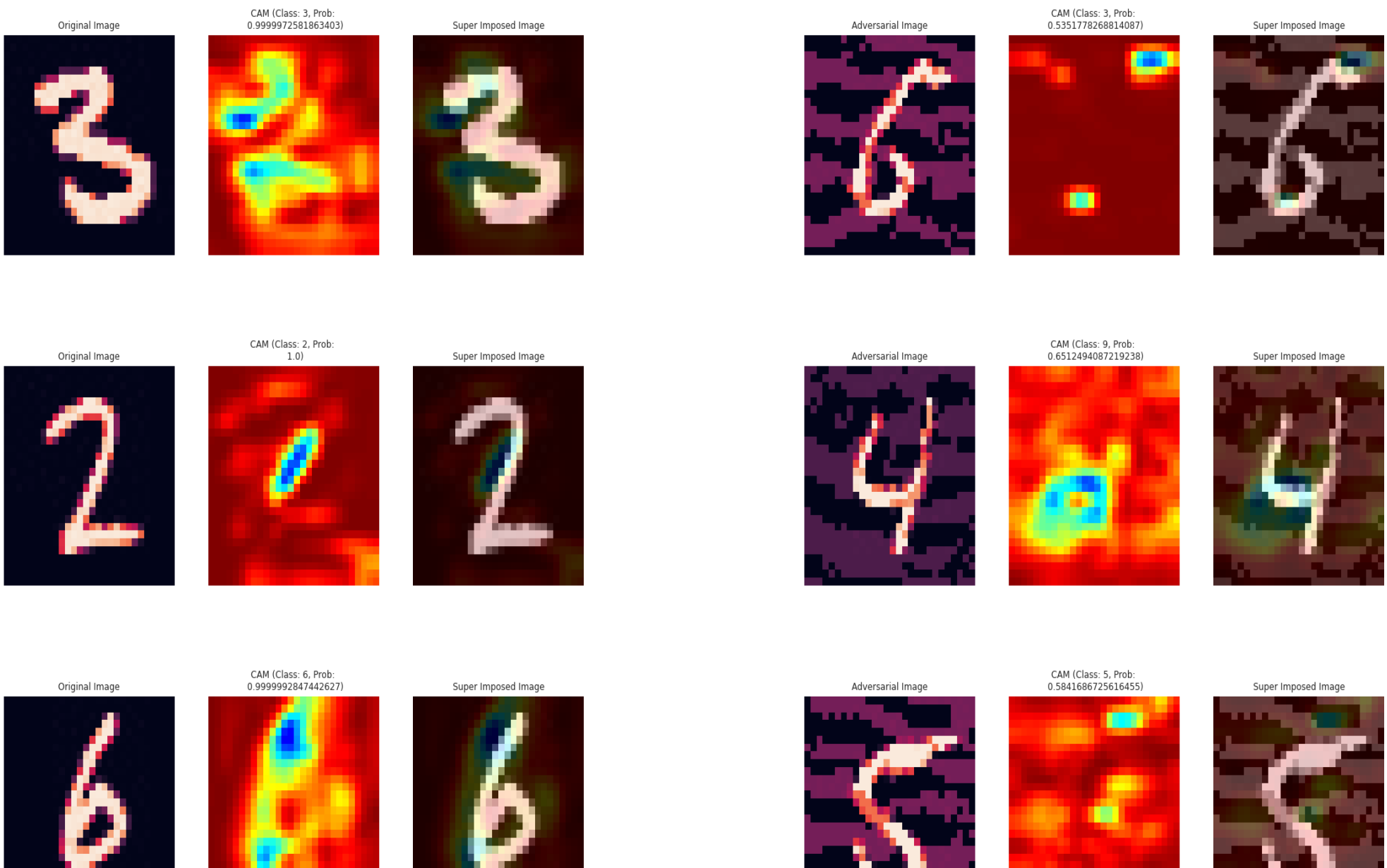
Below table shows the concentration of adversarial samples vs accuracy (adversarial accuracy)

Original samples (6000)	Adversarial samples (6000)	Model Accuracy %
100 %	0 %	99.18
80 %	20 %	79.47
60 %	40 %	59.77
40 %	60 %	40.1
20 %	80 %	20.42
0 %	100 %	0.52

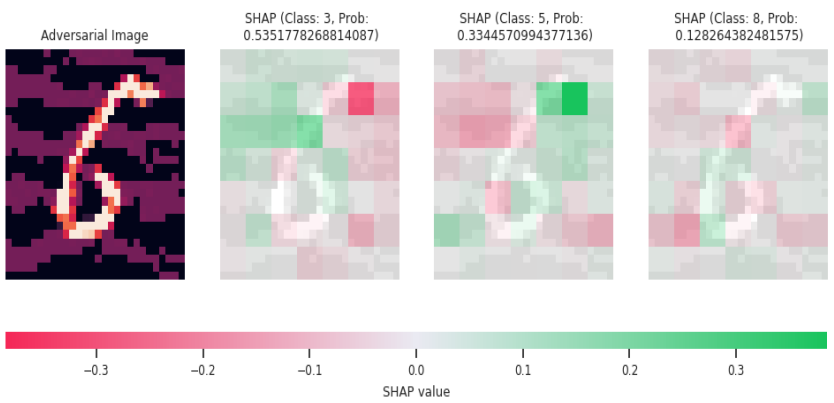
3. Explainability Report:

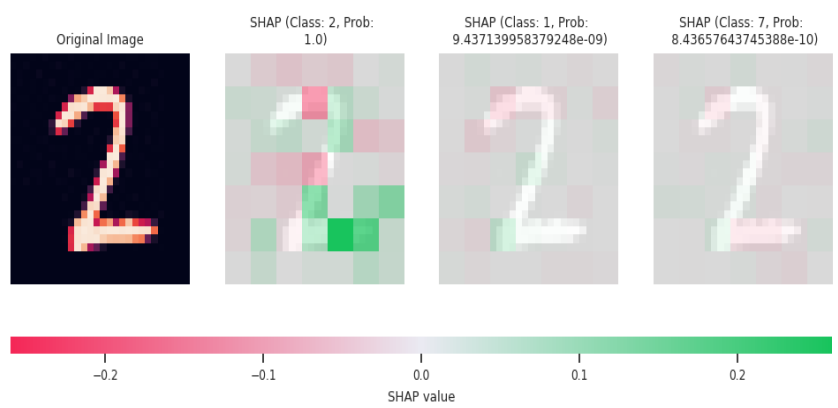
Explainability report provides a clear and concise explanation of the factors that contribute to a model's behavior or decisions, helping stakeholders to better understand and trust the model's outputs. The below report typically includes information on the data used to train the model, the model's architecture and performance, and the techniques or tools used to extract insights from the model.

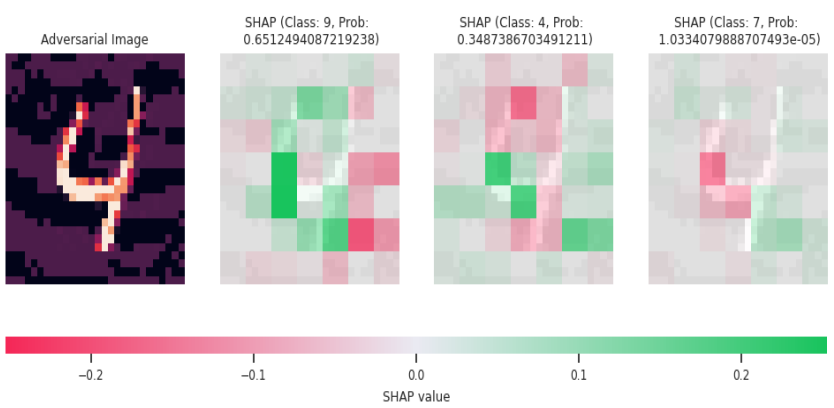
3.1 Class Activation Map (GradCAM++):

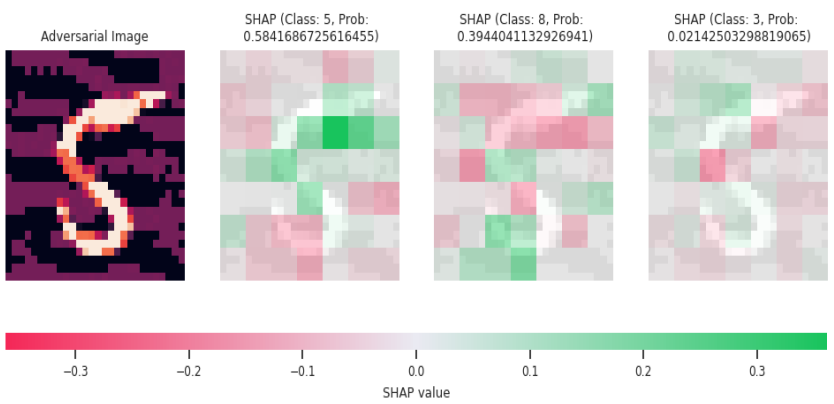


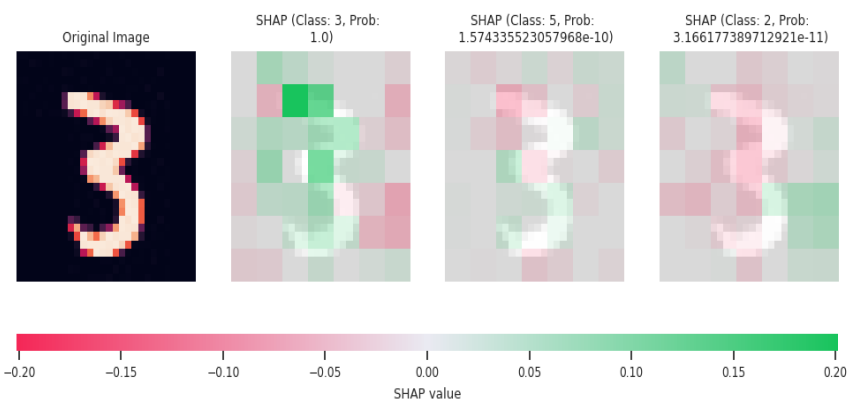
3.2 SHapley Additive exPlanations (SHAP):

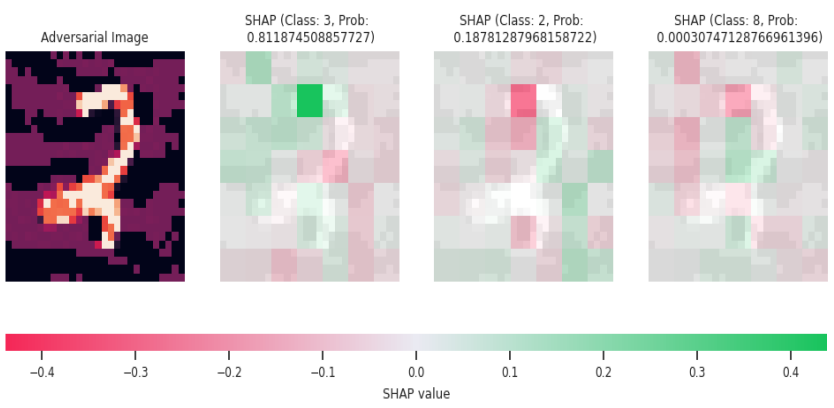


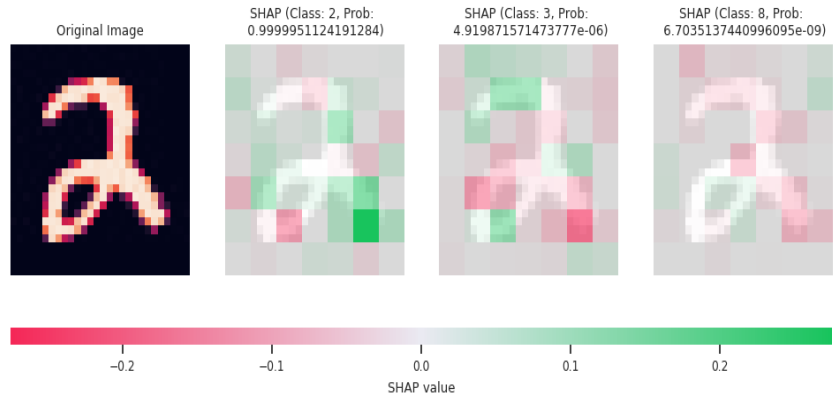












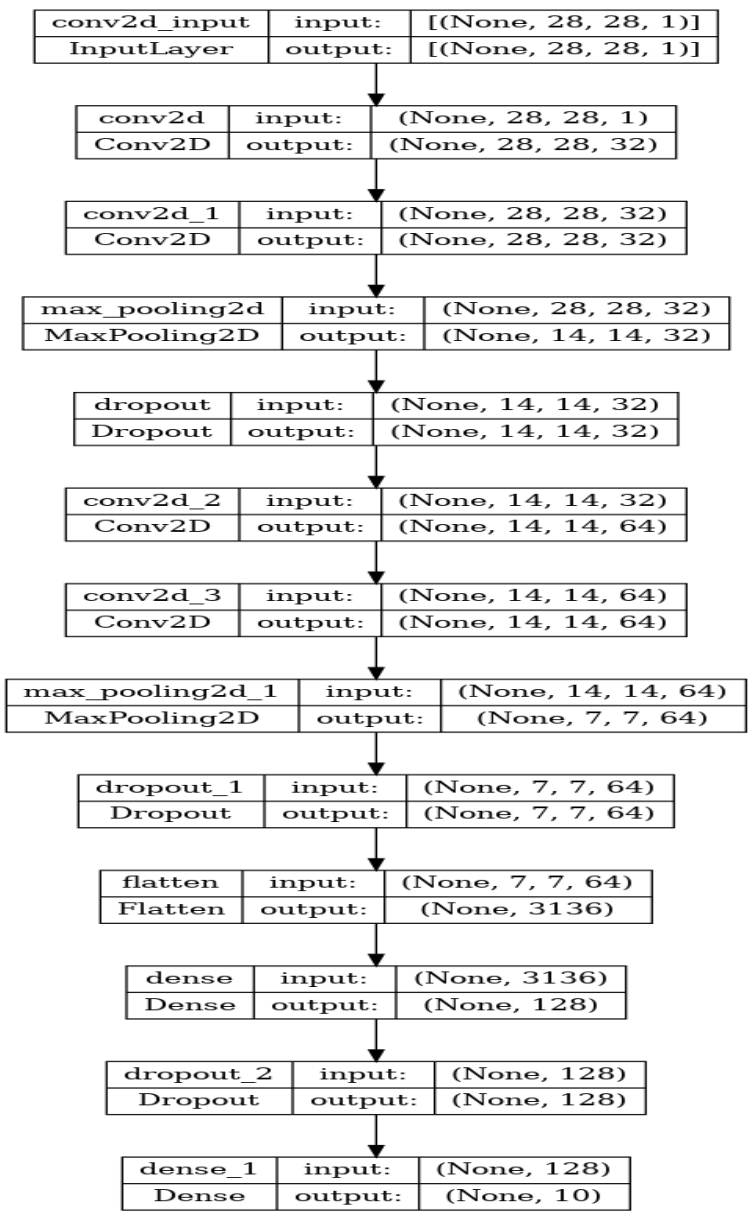
4. Appendix:

This section offers supplementary data on our methodology, the architectures of both the Original and Extracted Models, and other pertinent data like the Confusion matrix and Classification reports.

4.1 Model Architecture:

Below image represent the model architecture.

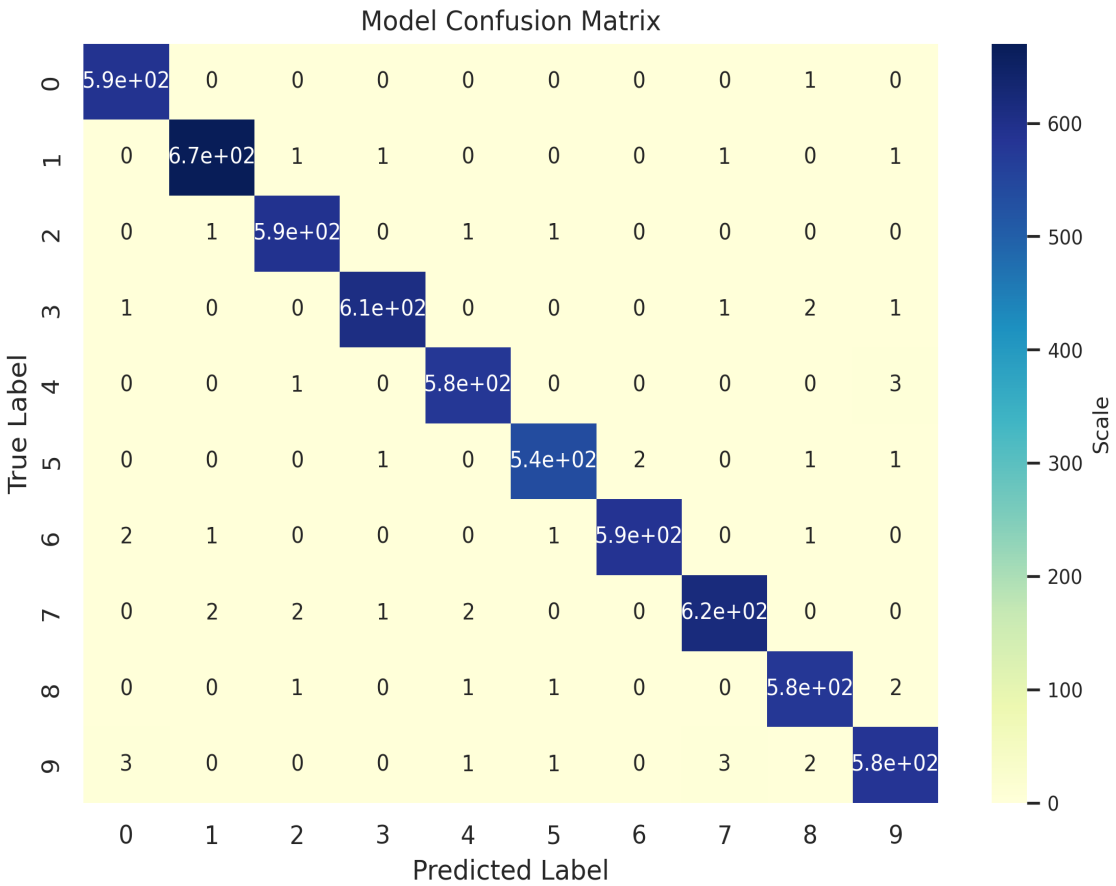
Model Architecture



4.2 Confusion Matrix:

Confusion matrix provides a comprehensive overview of the classification model's performance. The rows of the matrix represent the actual class labels, while the columns represent the predicted class labels. Each cell of the matrix shows the number of samples that belong to a particular combination of actual and predicted class labels.

Below images represent the Confusion Matrix.



4.3 Classification Report:

Classification report helps to evaluate the performance of your model for each class and make decisions about which classes need further improvement. It can also be used to compare the performance of different models or different versions of the same model.

Below images represent the Classification Report.

Model Classification Report

	precision	recall	f1-score	support
0	0.99	1.00	0.99	592
1	0.99	0.99	0.99	674
2	0.99	0.99	0.99	596
3	1.00	0.99	0.99	613
4	0.99	0.99	0.99	584
5	0.99	0.99	0.99	542
6	1.00	0.99	0.99	592
7	0.99	0.99	0.99	627
8	0.99	0.99	0.99	585
9	0.99	0.98	0.98	595
accuracy			0.99	6000
macro avg	0.99	0.99	0.99	6000
weighted avg	0.99	0.99	0.99	6000

4.4 OWASP Vulnerabilities(v0.3 Draft):

ML01:2023 Input Manipulation Attack
This weakness directly relates to adversarial attacks where an automated mechanism, such as machine learning, doesn't properly detect or handle inputs that have been modified to mislead the model.

For an expanded understanding of the OWASP Top - 10 ML Vulnerabilities, please visit https://owasp.org/www-project-machine-learning-security-top-10/

4.5 CVE(Common Vulnerabilities and Exposures)/CWE(Common Weakness Enumeration):

CWE-20: Improper Input Validation
A weakness where the product receives input or data but does not validate or incorrectly validates that the input has the properties required to process the data safely and correctly. This can lead to various security issues, such as altered control flow, arbitrary control of a resource, or arbitrary code execution.

CWE-707: Improper Neutralization
The product does not ensure or incorrectly ensures that structured messages or data are well-formed and that certain security properties are met before being read from an upstream component or sent to a downstream component.

CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
This weakness occurs when a product uses an automated mechanism, to recognize complex data inputs, but it does not properly detect or handle inputs that have been modified or constructed such that it causes the mechanism to detect an incorrect concept.

CWE-1288: Improper Validation of Consistency within Input
The product receives a complex input with multiple elements or fields that must be consistent with each other, but it does not validate or incorrectly validates that the input is actually consistent.This can allow attackers to trigger unexpected errors, cause incorrect actions to take place, or exploit latent vulnerabilities.

For comprehensive data related to CVE/CWE list,please visit https://cwe.mitre.org/

4.6 Utilized Attack Methods in Computer Vision Analysis:

In the comprehensive vulnerability analysis of the computer vision model, a variety of sophisticated evasion attack methods are employed.

AIShield integrates its proprietary evasion attack methodologies with an array of additional evasion attack techniques, ensuring a thorough and multi-dimensional assessment. The following section outlines the specific evasion attack techniques utilized in the vulnerability assessment by AIShield attack engine

List of Blackbox Attacks:

AIShield Propreitary methods; Auto Attack; Decision-based/Boundary Attack; Geometric Decision-based Attack (GeoDA); HopSkipJump Attack; No-Score/Label-Only attacks; Pixel Attack; Query-efficient Black-box; Simple Black-box Adversarial (SimBA); Spatial Transformation; Square Attack; Threshold Attack; Transfer Learning Attacks; Zeroth Order Optimisation (ZOO), and Hybrid(combination of above)

List of Greybox Attacks:

AIShield Propreitary methods; Adversarial Patch ; Auto-Attack ; Auto Projected Gradient Descent (Auto-PGD); Basic Iterative Method (BIM); Brendel & Bethge Attack; Carlini & Wagner (C&W) L_2 and L_inf attack; DeepFool; DPatch; Elastic Net; Fast Gradient Method; Feature Adversaries; Jacobian Saliency Map; NewtonFool; Optimisation attack (L-BFGS); Projected Gradient Descent (PGD); Robust DPatch; Shadow Attack; ShapeShifter; Targeted Universal Adversarial Perturbations; Universal Perturbation; Virtual Adversarial Method; Wasserstein Attack, and Hybrid(combination of above)

Report can be verified for its integrity with SHA-256 checksum:
2af8f68e53207538cfdc4265be6faf8495efb00652ab918205bdd8de48e2878e