AIShield
Powered by Bosch

Defense Analysis

Tabular Classification

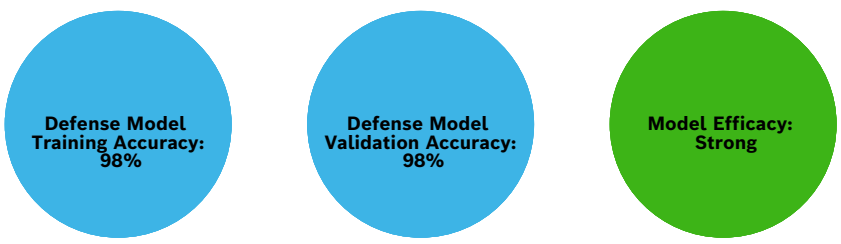
Attack Type
Extraction

Date
2024-08-05

Author
AIShield

Job Id
gAAAAABm...ETFc7xoQ==

Executive Summary:



A defense model with Strong efficacy is built and the defense model can detect an attack on the original model with 98% accuracy. This can help prevent malicious attacks on data and help avoid adverse consequences.

1. Performance:

1.1 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 complexity of the model, the size of the input data, 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, Inference Time is 52.9 ms

1.2 Simulation Report:

The table show efficacy of Defense Model in discriminating Original data and Attack data.

Input	Size	% attack detected
Original Data	2058	3.4
Attack Data	2058	99.03

The first row in the table represents false positives for attacks, indicating instances that are incorrectly identified as attacks by the model. The second row represents true positives for attacks, indicating instances that are correctly identified as attacks by the model.

2. Defense Model Efficacy:

2.1 Defense Model Accuracy:

Defense Model Accuracy is a measure of how accurate the defense model's predictions arefor the dataset which is combination of original and attack data [low-suspicious-attack, high-suspicious-attack] i.e. how well the model is able to detect a element of the dataset as original or attack. It is calculated by taking the total number of correct predictions and dividing by the total number of predictions made.

In the context of this scenario, Accuracy of the Defense Model is 98%.

2.2 Defense Model F1 Score:

Defense Model F1 score measures of how well the model is performing. It is calculate by considering both precision (how many of the model's positive predictions were correct) and recall (how many of the actual positives the model identified) to calculate a single score. An F1 score of 1.0 represents perfect precision and recall, while a score of 0.0 represents no better than chance performance.

In the context of this scenario, F1 score of the Defense Model is 0.98.

3. Appendix:

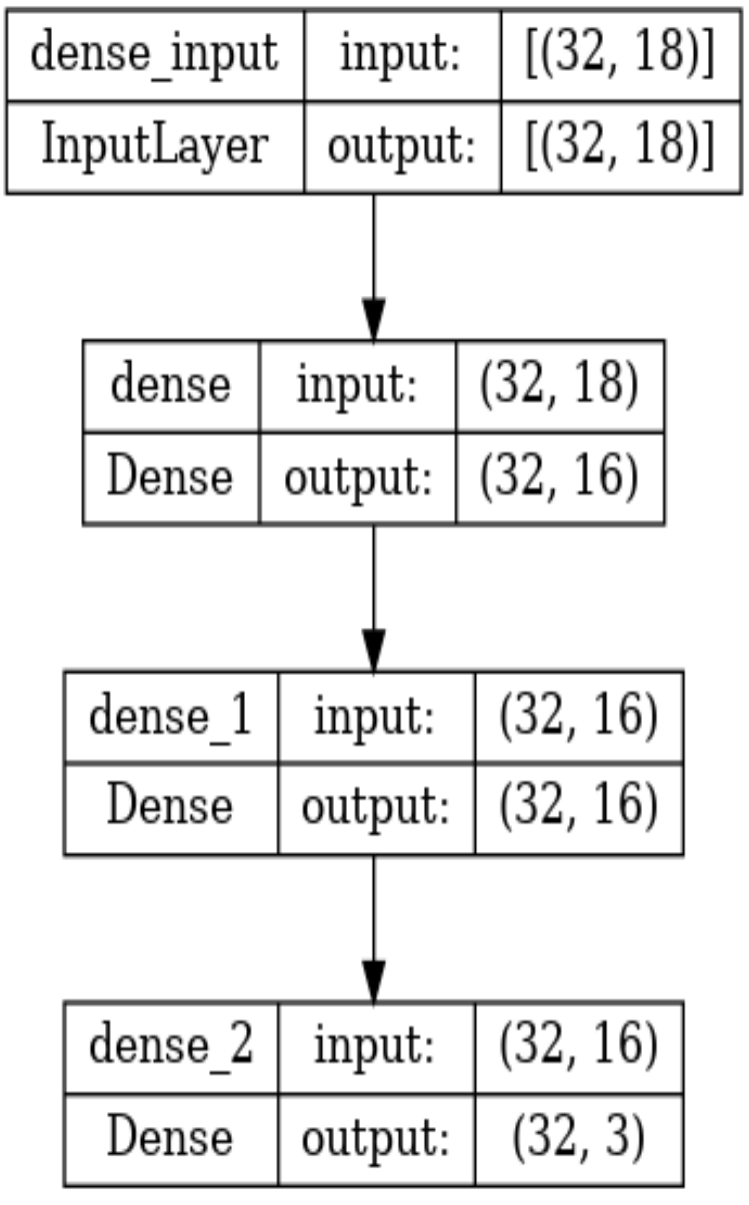
Appendix section provides additional information on our methodology, Defense Model architecture, and other relevant details like Confusion matrix , Classification reports of the model.

In Extraction right now , low-suspicious-attack is a dormant class.

3.1 Model Architecture:

Below image represent the model Architecture of the Defense Model.

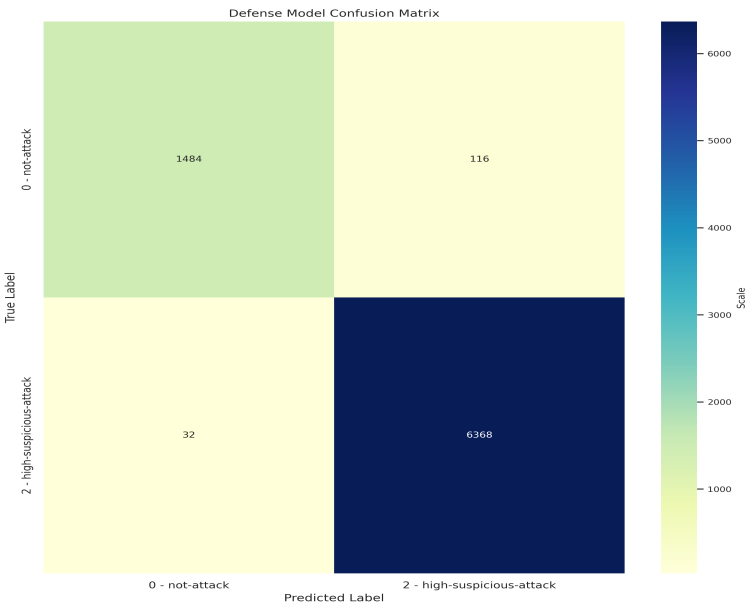
Defense Model Architecture



3.2 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.



3.3 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.

Defense Model Classification Report

	precision	recall	f1-score	support
0	0.98	0.93	0.95	1600
2	0.98	0.99	0.99	6400
accuracy			0.98	8000
macro avg	0.98	0.96	0.97	8000
weighted avg	0.98	0.98	0.98	8000

The digest value of the pdf file is c2cef932de4866d45f355a94a34d40641cacc83b598442ce3a3eb4ca2947e4aa