Machine Learning for Cyber security
Lab 4 Report

Name: Karthvik Sarvade **NetId**: ks6807

Introduction:

This report presents the development and evaluation of GoodNet models aimed at detecting backdoor attacks in neural networks. The focus was on countering a specific "sunglasses backdoor" attack in a BadNet model trained on the YouTube Face dataset.

Methodology:

1. **Dataset Overview**:
- The project utilized the YouTube Face dataset, which comprises facial images from YouTube videos. This dataset is significant for its diversity in terms of facial expressions, poses, and lighting conditions.
- Two subsets of data were employed: a clean dataset and a poisoned dataset. The poisoned dataset was crafted by introducing a sunglasses trigger, emblematic of the backdoor attack.

2. **Pruning Defense Implementation**:
- The defense strategy involved pruning the last pooling layer of the compromised BadNet. This was executed by sequentially removing channels based on their activation values.
- Pruning was controlled by thresholds of 2%, 4%, and 10%, representing the allowable drop in validation accuracy.

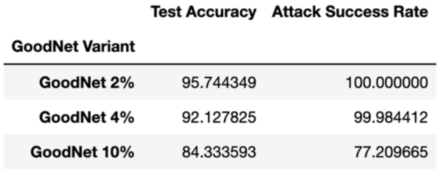
3. **GoodNet Model Construction**:
- The GoodNet models integrated the original BadNet with its pruned counterparts.
- These models were designed to flag inputs as potentially backdoored if there was a disagreement in classification between the original and pruned networks.

4. **Model Evaluation**:
- Evaluations were conducted on both clean and poisoned datasets to measure clean classification accuracy and attack success rate (ASR).

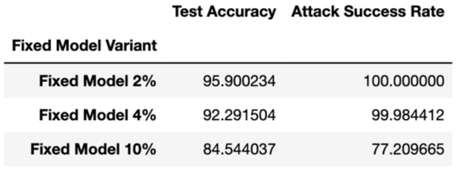
Results:

1. Data Handling and Analysis:
- The handling of both clean and poisoned datasets was crucial in assessing the effectiveness of the GoodNet models. Accurate processing and labeling of these datasets ensured the reliability of the evaluation metrics.
- The poisoned dataset, in particular, was instrumental in determining the models' robustness against the backdoor attack.

2. Performance Metrics:
- The models showed varied performance based on the pruning thresholds. Lower thresholds maintained higher accuracy on clean data but were less effective in lowering ASR. The 10% threshold, while reducing clean data accuracy, significantly lowered the ASR.



	Test Accuracy	Attack Success Rate
GoodNet Variant		
GoodNet 2%	95.744349	100.000000
GoodNet 4%	92.127825	99.984412
GoodNet 10%	84.333593	77.209665



	Test Accuracy	Attack Success Rate
Fixed Model Variant		
Fixed Model 2%	95.900234	100.000000
Fixed Model 4%	92.291504	99.984412
Fixed Model 10%	84.544037	77.209665

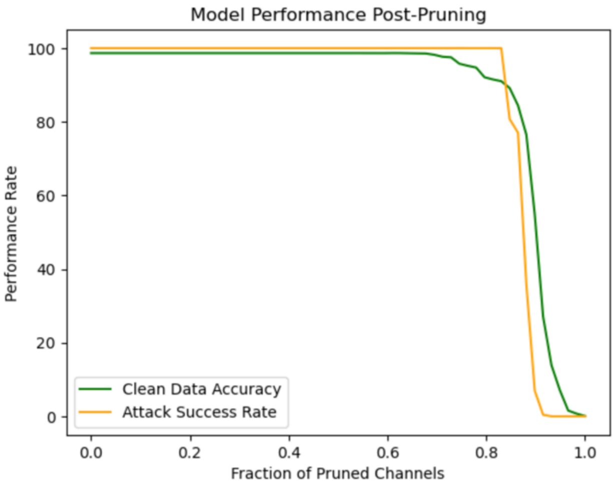


Figure 1: The graph displays the model's performance after pruning. The clean data accuracy (green line) remains relatively stable until a certain point, beyond which it decreases sharply. Similarly, the attack success rate (orange line) maintains a high percentage until faced with substantial pruning, where it then drops significantly. This visualization clearly shows the effectiveness of pruning in reducing the attack success rate, though it also highlights the corresponding decrease in clean data accuracy, demonstrating the inherent trade-offs in model performance.

3. Visualization and Interpretation:

- Graphs visually represented the impact of different pruning thresholds on model performance. These visualizations were key in illustrating the trade-offs inherent in the pruning process.

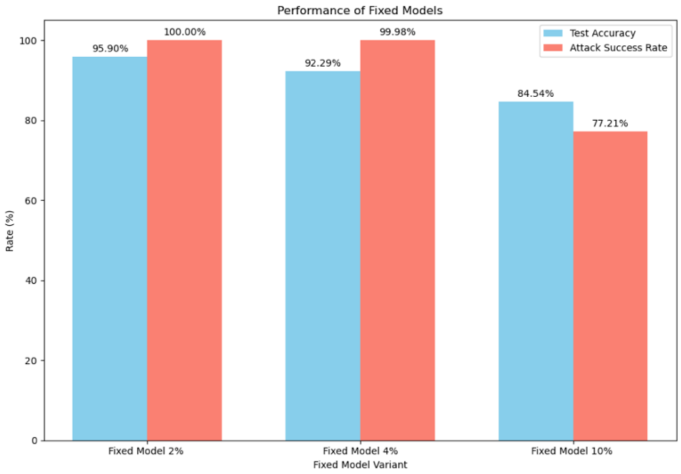


figure 2

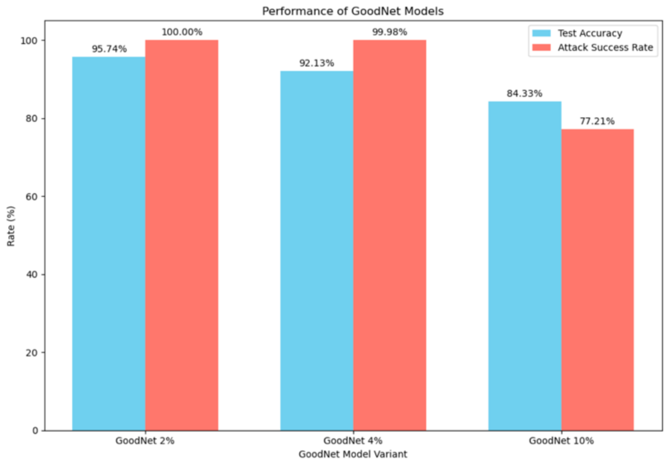


figure 3

figure 2 illustrates the test accuracy and attack success rate for the fixed models.
figure3 illustrates the test accuracy and attack success rate for the GoodNet models.

Conclusion:

The project demonstrated that channel pruning is a viable defense against backdoor attacks in neural networks. The GoodNet models balanced clean data accuracy with backdoor attack mitigation, with the 10% pruning threshold model showing the most promise in terms of security, albeit at a cost to clean data accuracy. This study highlights the challenges in achieving an optimal balance between security and performance in machine learning models and underscores the importance of dataset integrity and handling in model evaluation.