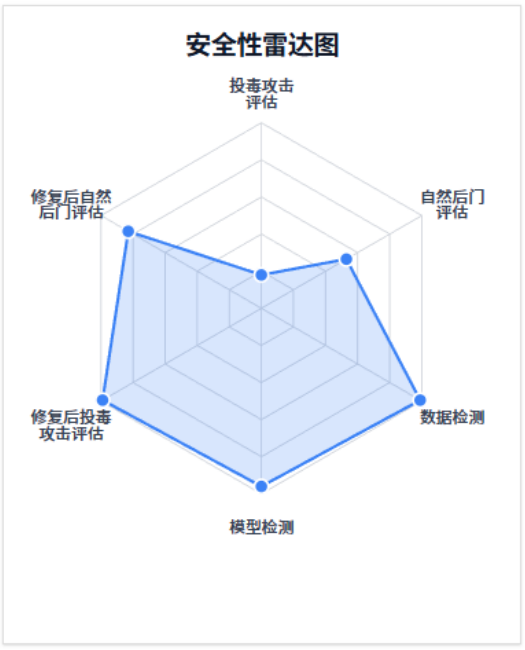
安全检测报告

Safety Inspection Report

任务配置 Configuration	模型结构 Structure	Simple Net	数据集 Dataset	MNIST
	优化器 Optimizer	SGD	学习率 Lr	0.01
防御设置 Defense	专家模式	抗攻击性模块	检测模块	缓解模块
测试项目 Project	BadNets、SSBAs、CASSOCK、通用触发器、样本专用触发器、SVD、STRIP、Neural Cleanse、DeBackdoor、Steps、数据清洗、数据增强、梯度加噪、样本对齐、剪枝、微调、机器遗忘			

一、总体分析



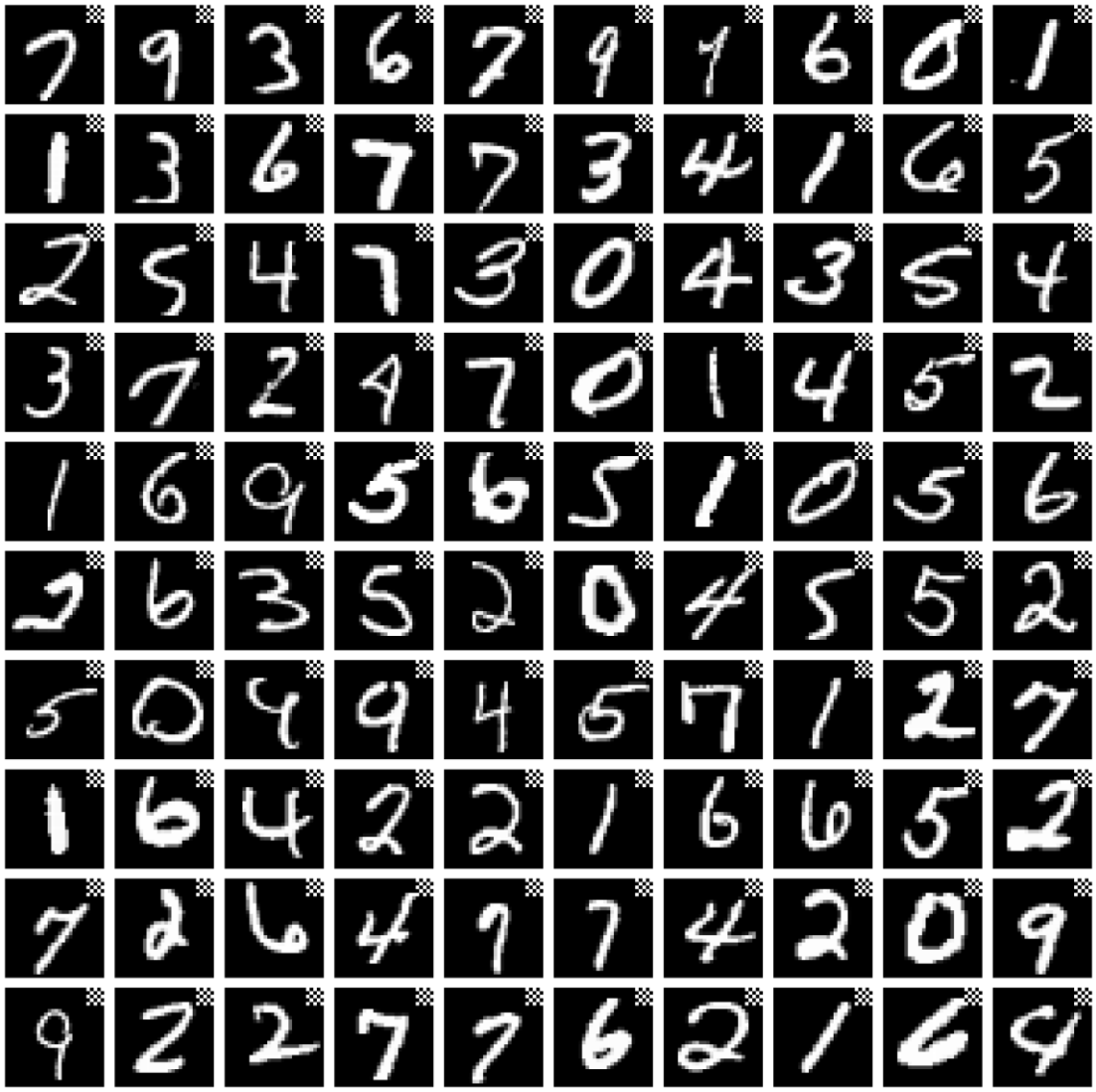
评估总结

本次安全评估覆盖了6个核心维度，综合得分为 **75分**，整体表现 较好。

二、投毒攻击评估

测试项目 Project	BadNets、SSBAs、CASSOCK					
数据集 Dataset	MNIST		模型结构 Structure	Simple Net		
批处理 BatchSize	64	优化率 Lr	0.01	轮次 Epoch	1	
结论 Conclusion	ASR(%)	99%	BACC(%)	98%	Time(s)	13

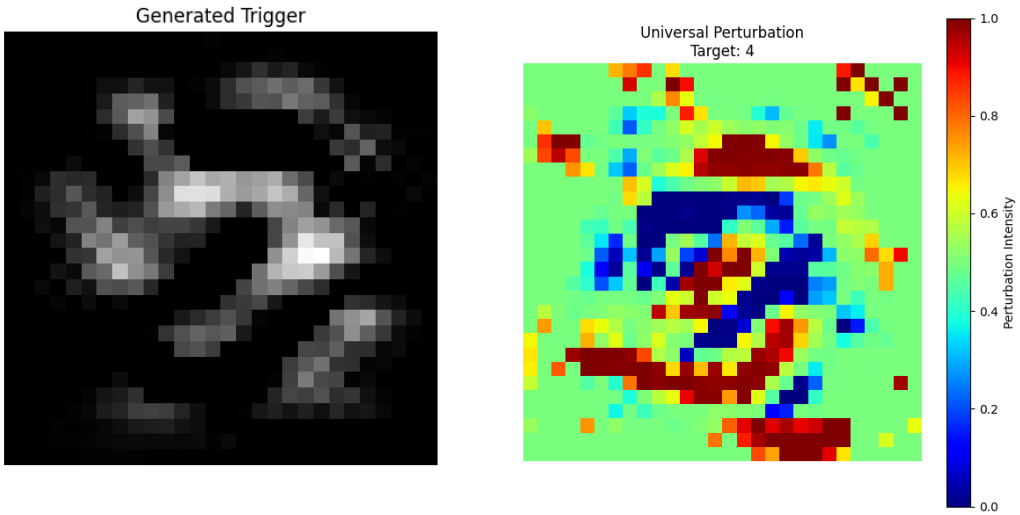
后门样本：



三、自然后门攻击评估

测试项目 Project	通用触发器				
数据集 Dataset	MNIST		模型结构 Structure	Simple Net	
批处理 BatchSize	1	优化率 Lr	0.0001	轮次 Epoch	1
结论 Conclusion	ASR(%)	9	Time(s)	12	

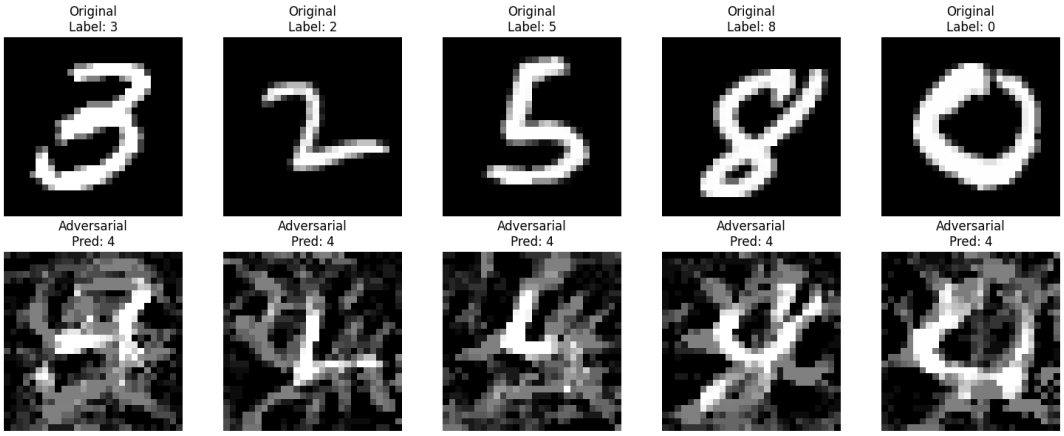
自然后门样本：



四、自然后门攻击评估

测试项目 Project	样本专用触发器				
数据集 Dataset	MNIST		模型结构 Structure	Simple Net	
批处理 BatchSize	1	优化率 Lr	0.0001	轮次 Epoch	1
结论 Conclusion	ASR(%)	99	Time(s)	8	

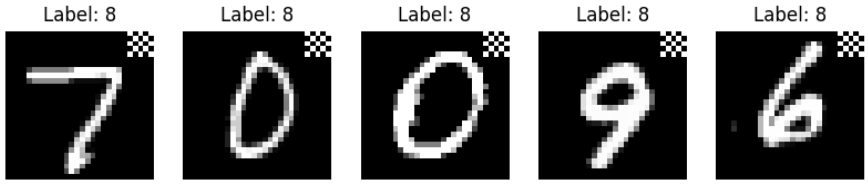
自然后门样本：

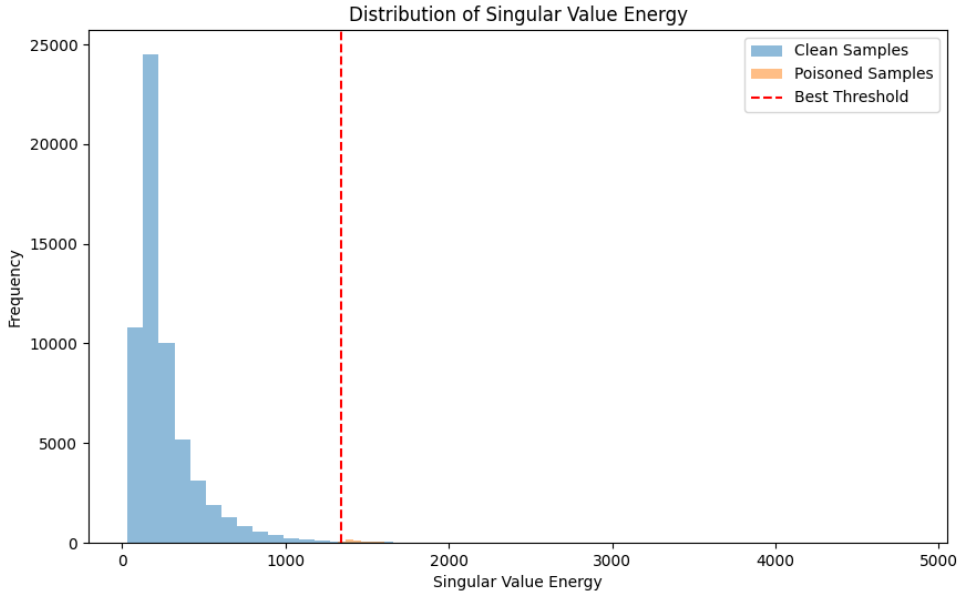


五、数据检测

测试项目 Project	SVD			
数据集 Dataset	MNIST	模型结构 Structure	Simple Net	
结论 Conclusion	异常分数(%)	99	Time(s)	18

异常样本：

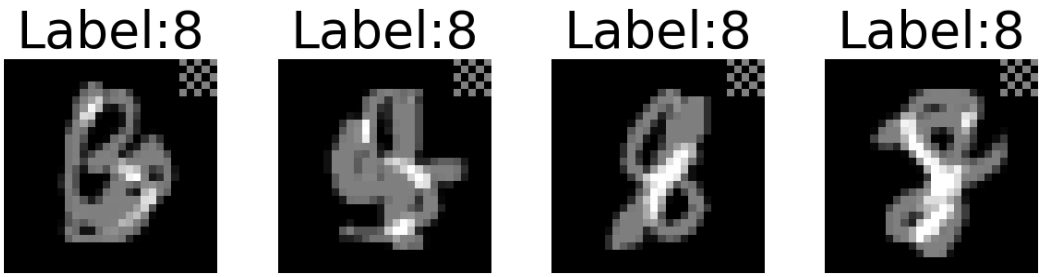




六、数据检测

测试项目 Project	STRIP			
数据集 Dataset	MNIST	模型结构 Structure	Simple Net	
结论 Conclusion	异常分数(%)	99	Time(s)	11

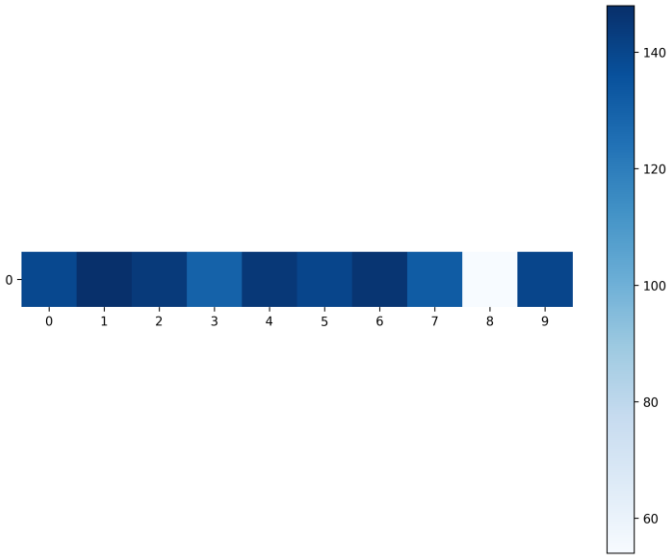
异常样本：



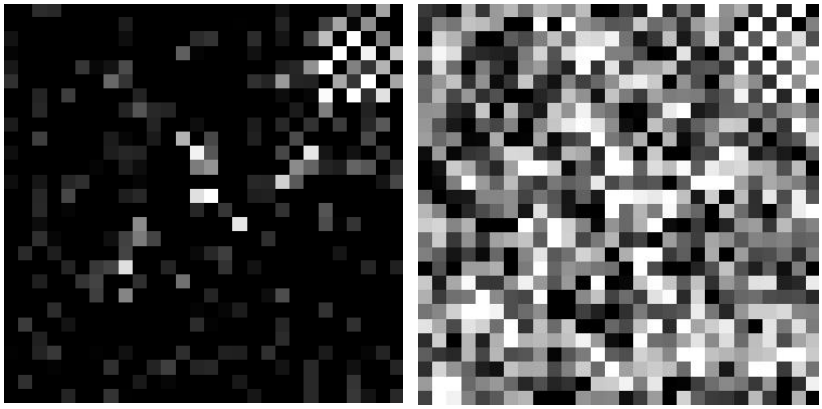
七、模型检测

测试项目 Project	Neural Cleanse			
数据集 Dataset	MNIST	模型结构 Structure	Simple Net	
结论 Conclusion	异常分数(%)	48	Time(s)	7

异常标签：



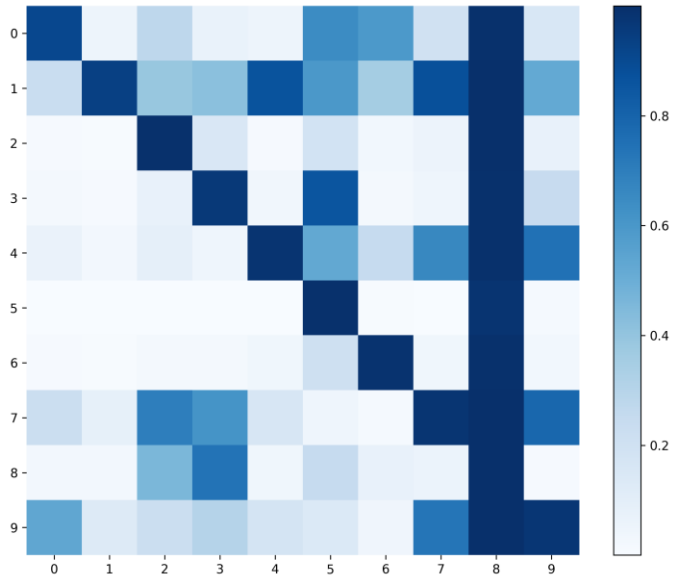
逆向触发器：



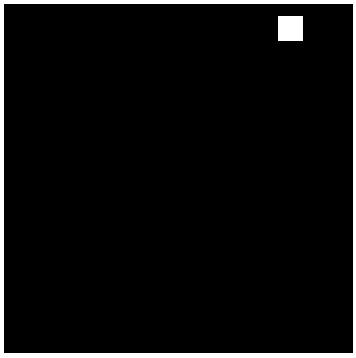
八、模型检测

测试项目 Project	DeBackdoor			
数据集 Dataset	MNIST	模型结构 Structure	Simple Net	
结论 Conclusion	异常分数(%)	32	Time(s)	5

异常标签：



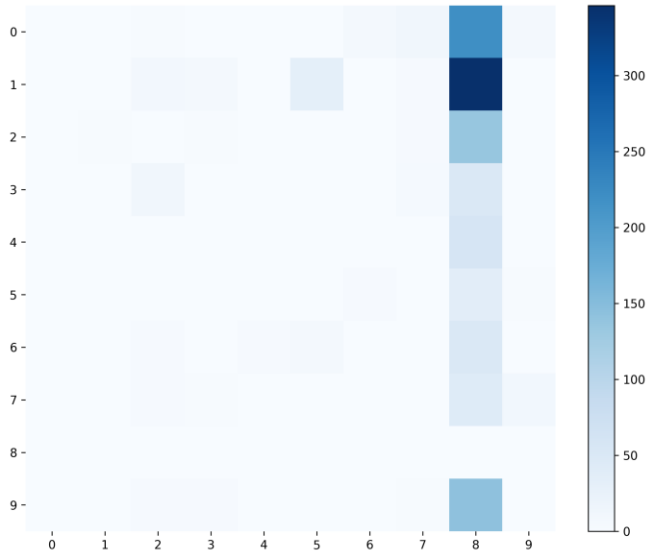
逆向触发器：



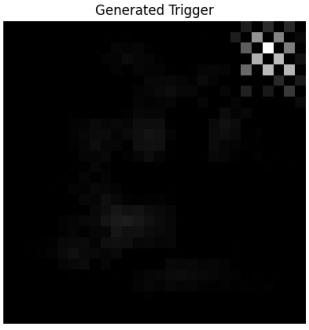
九、模型检测

测试项目 Project	Steps			
数据集 Dataset	MNIST	模型结构 Structure	Simple Net	
结论 Conclusion	异常分数(%)	84	Time(s)	2

异常标签：



逆向触发器：



十、修复后投毒攻击评估

修复项目 Project	数据清洗、数据增强、梯度加噪、样本对齐、剪枝、微调、机器遗忘				
数据集 Dataset	MNIST		模型结构 Structure	Simple Net	
评估 evaluate	投毒攻击 1	ASR(%)	1	BACC(%)	99
	自然后门 1	ASR(%)	4	BACC(%)	98
	自然后门 2	ASR(%)	3	BACC(%)	99