Impact of Adversarial Training Strength on Tabular Foundation Model Generalization Under Clean Test Conditions

Assignee Research

June 11, 2026

Abstract

Deep Neural Networks (DNNs) have shown great promise in various domains. However, vulnerabilities associated with DNN training, such as backdoor attacks, are a significant concern. These attacks involve the subtle insertion of triggers during model training, allowing for manipulated predictions. More recently, DNNs used with tabular data have gained increasing attention due to the rise of transformer models. Our research presents a comprehensive analysis of backdoor attacks on tabular data using DNNs, mainly focusing on transformers. We propose a novel approach for trigger construction: in-bou

1 Introduction

This paper examines: Backdoor Attacks on Transformers for Tabular Data: An Empirical Study. Research question: What is the impact of varying the strength of adversarial training (e.g., perturbation magnitude, attack frequency) on the generalization performance of tabular foundation models, as measured by accuracy on the TabMNAR benchmark under clean test conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.0/10.

3 Results

12 papers retrieved. 12 claims extracted; 8 independently verified. Quality review score: 7.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Attack Success Rate (ASR) is defined as the proportion of poisoned test samples where the model predicts the target labe	×	0.10
Clean Data Accuracy (CDA) measures the accuracy of the poisoned model on clean test data compared to the baseline accura	✓	0.19
Out-of-Bounds Triggers are crafted by setting feature values to values outside the support of the original feature distr	×	0.11
Out-of-Bounds Triggers minimize false positives and potential drops in Clean Data Accuracy (CDA).	✓	0.19
Out-of-Bounds Triggers are not stealthy and can be easily spotted as outliers using dataset inspection methods.	✓	0.23
In-Bounds Triggers are crafted by setting feature values to values within the original data distribution support.	×	0.14
In-Bounds Triggers leverage common feature values such as the mean, median, or mode to reduce anomaly likelihood.	✓	0.28
Using the mode value for In-Bounds Triggers ensures the trigger value is prevalent in the dataset, making it less detect	✓	0.23
Trigger features are selected based on importance scores derived from surrogate models trained on clean data.	✓	0.22
Xie et al. demonstrated that using features with low importance scores led to a higher Attack Success Rate (ASR).	✓	0.24
The study deployed several models to determine feature importance scores, unlike the approach in reference [20] which us	✓	0.19
Deploying several models for feature importance determination led to backdoor performance improvement.	×	0.13

References

- `http://arxiv.org/abs/2311.07550v4`

- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2505.21027v2