Hybrid Adversarial Training with CausalMixFT and SMOTE on TabRobust Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 2 peer-reviewed papers addressing the following research question: To what extent do hybrid adversarial training methods (combining CausalMixFT with traditional data augmentation like SMOTE) improve robustness on TabRobust benchmarks compared to standalone methods,. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Unlearning: Reducing Confidence Along Adversarial Directions. Research question: To what extent do hybrid adversarial training methods (combining CausalMixFT with traditional data augmentation like SMOTE) improve robustness on TabRobust benchmarks compared to standalone methods, as measured by adversarial accuracy and generalization across domains?.

2 Methodology

Systematic literature search across multiple databases yielded 2 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

2 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The experiments study the effect of RCAD on test accuracy in comparison to and in addition to existing regularization te	×	0.10
Six image classification benchmarks are used: CIFAR-10, CIFAR-100, SVHN, Tiny Imagenet, CIFAR-100-2k, and CIFAR-100-10k.	×	0.03
CIFAR-100-2k and CIFAR-100-10k are created by randomly sub-sampling 2,000 and 10,000 training examples from the original	×	0.03
If the validation split is not provided by the benchmark, 10% of the training examples are held out for validation.	×	0.04
All methods are trained using the ResNet-18 backbone unless specified otherwise.	×	0.03
Gradient clipping in the l2 norm (at 1.0) is used to accelerate training loss convergence.	×	0.04
All models are trained for 200 epochs using SGD with an initial learning rate of 0.1 and Nesterov momentum of 0.9.	×	0.02
The learning rate is decayed by a factor of 0.1 at epochs 100, 150, and 180.	×	0.01
The model checkpoint corresponding to the epoch with the best accuracy on validation samples is selected as the final mo	×	0.03
For all datasets except CIFAR-100-2k and CIFAR-100-10k, the methods were trained with a batch size of 128.	×	0.03
The primary aim of the experiments is to study whether entropy maximization along the adversarial direction shrinks the	×	0.02
Baselines include methods that directly constrain the model's predictive distribution on observed samples (label smoothi	×	0.11

References

- http://arxiv.org/abs/2206.01367v1
- http://arxiv.org/abs/2408.02710v1