

# CodeT5 Robustness Under FGSM and PGD Attacks at Varying Epsilon Levels

Assignee Research

June 9, 2026

## Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the comparative degradation in CodeT5 code generation accuracy when subjected to FGSM versus PGD attacks with varying epsilon values. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Brain Tumor Classifiers Under Attack: Robustness of ResNet Variants Against Transferable FGSM and PGD Attacks. Research question: What is the comparative degradation in CodeT5 code generation accuracy when subjected to FGSM versus PGD attacks with varying epsilon values?.

## 2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

## 3 Results

14 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.2/10.

## 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
BrainNeXt152 demonstrates strong resilience against black-box attacks and produces the least transferable adversarial sa	✓	0.15
The lowest black-box accuracy for the Dilation3 model is 61.09% under attacks crafted from the Dilation4 model using the	×	0.09
The lowest black-box accuracy for the BrainNet model is 64.31% under attacks crafted from the Dilation4 model using the	×	0.10
Excluding attacks from BrainNeXt152, both BrainNet and Dilation3 experience an average drop of approximately 40% in blac	×	0.05
Under PGD attacks with alpha=0.0075, the Dilation3 model accuracy drops to 60.35% when attacked with adversarial samples	×	0.13
Under PGD attacks with alpha=0.0075, the BrainNet model reaches a lowest accuracy of 58.36% when targeted by samples fro	×	0.13
Dilated CNN variations with dilation rates of 2, 3, and 4 each have 42,626,560 parameters.	×	0.02
Dilated CNN variations with dilation rates of 2, 3, and 4 took 15 minutes to train.	×	0.02
Dilated CNN variations with dilation rates of 2, 3, and 4 have an inference time of 3 seconds.	×	0.02
The study assumes a black-box adversarial setting where the attacker lacks access to the target model’s architecture, we	×	0.03
Experiments were conducted on three MRI dataset variations: full-sized augmented, shrunk augmented, and shrunk non-augme	×	0.14
Attacks were tested with epsilon values of 0.02, 0.03, 0.04, and 0.05.	×	0.02
For normalized images in the [0,1] range, an epsilon value of 0.04 corresponds to a $\pm 10$ pixel intensity change on a 0–25	×	0.02
Figure 10 evaluates model accuracy under FGSM attacks at epsilon = 0.05 using the shrunk and augmented MRI dataset.	×	0.08
Figure 12 evaluates model accuracy under PGD attacks at epsilon = 0.03 and alpha = 0.003 for 10 iterations using the ful	×	0.08

## References

- <http://arxiv.org/abs/2408.13274v1>
- <http://arxiv.org/abs/2307.02055v1>
- <http://arxiv.org/abs/2602.11646v1>