Chapter 19
Enhancing Machine Learning Security With Robust Discretization-Based Defenses Against Adversarial Attacks

Chetan Hiraman Patil
https://orcid.org/0009-0008-8220-9022
Madhyanchal Professional University, Bhopal, India

Mohd Zuber
https://orcid.org/0000-0002-1444-6535
Scope Global Skills University, Bhopal, India

Ankit Temurnikar
https://orcid.org/0000-0002-0416-5289
Madhyanchal Professional University, Bhopal, India

ABSTRACT

Machine learning (ML) models have security weaknesses, as small input perturbations cause misclassifications, risking healthcare, finance, and autonomous systems. A defense strategy using discrete representation transformation of continuous features enhances model resistance to adversarial attacks. A comparative analysis of discretization methods—quantization, binning, and entropy-based partitioning—applies to ML models like decision trees and deep neural networks. Our defense is tested on adversarial datasets (ImageNet-A, MNIST Adversarial, CIFAR-10 Adversarial) against FGSM, PGD, and CW attacks. Performance evaluation considers accuracy, robustness, adversarial transferability, and computational efficiency. Results show discretization reduces misclassification by 30% while maintaining strong prediction performance. Our research highlights its low cost compared to adversarial training, ensuring scalability. Future work explores adaptive and hybrid discretization to enhance ML security and optimize robustness-efficiency trade-offs.

DOI: 10.4018/979-8-3373-3241-3.ch019

Copyright ©2025, IGI Global Scientific Publishing. Copying or distributing in print or electronic forms without written permission of IGI Global Scientific Publishing is prohibited.

1. INTRODUCTION

Machine learning (ML) models have emerged as powerful tools in various domains, including healthcare, finance, and autonomous systems. Their ability to learn patterns from data and make real-time decisions has significantly transformed these industries. However, despite their widespread adoption (Zhang, Xu, Zhang, & Li, 2022), ML models suffer from profound security vulnerabilities. These vulnerabilities stem from the susceptibility of ML algorithms to adversarial attacks, where small, imperceptible perturbations in input data can lead to significant misclassifications. Such attacks pose severe risks to critical applications, potentially leading to misdiagnoses in healthcare, fraudulent transactions in financial systems, and failures in autonomous decision-making processes. Addressing these security challenges is imperative for ensuring the robustness and reliability of ML-driven applications (Jung, Woo, & Mukhopadhyay, 2024).

Adversarial attacks exploit the inherent weaknesses of ML models by introducing carefully crafted perturbations into input data. Popular attack strategies include Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner (CW) attacks, all of which manipulate model predictions while maintaining the appearance of normal input. These attacks have been successfully demonstrated against widely used ML models, including decision trees and deep neural networks, leading to erroneous classifications that can have serious real-world consequences (Huang & Li, 2023). For example, in healthcare, an adversarial perturbation in a medical image could mislead an AI-driven diagnostic system, resulting in incorrect treatment recommendations. Similarly, in financial institutions, adversarial attacks could manipulate fraud detection systems, allowing fraudulent transactions to bypass security mechanisms. The risks associated with adversarial threats necessitate the development of effective defense strategies to enhance ML model security (Jain, 2024).

Among various defensive strategies, discrete representation transformation of continuous features has emerged as a promising solution to mitigate adversarial attacks. The fundamental idea behind this approach is to discretize input data, thereby reducing the model's sensitivity to small perturbations introduced by adversarial attacks. Discretization techniques such as quantization, binning, and entropy-based partitioning offer a mechanism to transform continuous input data into discrete representations, making it more difficult for adversarial attacks to introduce imperceptible changes that significantly alter model predictions (Chen, Zhu, & He, 2024). Unlike adversarial training, which requires extensive computational resources and retraining on adversarial examples, discretization-based defenses operate efficiently at scale while maintaining strong adversarial robustness (Li, Zhou, Yuan, Li, & Leung, 2020).

This research presents a comprehensive comparative analysis of different discretization approaches and their impact on ML models, specifically decision trees and deep neural networks (Cao, Zhu, Wang, & Zhuang, 2022). We systematically evaluate various discretization methods, including uniform quantization, k-means clustering-based binning, and entropy-based partitioning, to determine their effectiveness in defending against adversarial attacks. Our defense approach is tested on General Adversarial Attack Datasets, including ImageNet-A, MNIST Adversarial, and CIFAR10 Adversarial, under adversarial attack scenarios involving FGSM, PGD, and CW attacks. The evaluation criteria include classification accuracy, robustness under attack, adversarial transferability resistance, and computational efficiency (Rahman, Pal, Habib, Pan, & Karmakar, 2025).

18 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/enhancing-machine-learning-security-with-robust-discretization-based-defenses-against-adversarial-attacks/379635

Related Content

The Satiation of Natural Curiosity
Felix Schoeller (2016). *International Journal of Signs and Semiotic Systems (pp. 27-34).*
www.irma-international.org/article/the-satiation-of-natural-curiosity/185499

Future Directions and Innovations in AI and Imaging
Induni Nayodhara Weerarathna, Matteo Hirushan Jayakody Arachchige, Austin Jojo Jallah, Mohammad Affan Kareemand Kidus Manaye Mulugeta (2026). *Radiodiagnosis in the Era of AI (pp. 225-264).*
www.irma-international.org/chapter/future-directions-and-innovations-in-ai-and-imaging/386078

Evaluation of Logistics Development Under the Visual Field of Low-Carbon Environmental Protection Based on Hierarchical Methods
Jinjuan Wang (2024). *International Journal of Ambient Computing and Intelligence (pp. 1-16).*
www.irma-international.org/article/evaluation-of-logistics-development-under-the-visual-field-of-low-carbon-environmental-protection-based-on-hierarchical-methods/360709

A Novel Sparse Representation Based Visual Tracking Method for Dynamic Overhead Cranes: Visual Tracking Method for Dynamic Overhead Cranes
Tianlei Wang, Nanlin Tan, Chi Zhang, Ye Liand Yikui Zhai (2019). *International Journal of Ambient Computing and Intelligence (pp. 45-59).*
www.irma-international.org/article/a-novel-sparse-representation-based-visual-tracking-method-for-dynamic-overhead-cranes/238053

Engineering Adaptive Multi-Agent Systems: The ADELFE Methodology
Carole Bernon, Valérie Camps, Marie-Pierre Gleizesand Gauthier Picard (2008). *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications (pp. 513-535).*
www.irma-international.org/chapter/engineering-adaptive-multi-agent-systems/24299