UNIVERSITI TEKNOLOGI MARA

A FORENSIC FRAMEWORK FOR POST-DEPLOYMENT ANALYSIS OF DATA POISONING ATTACKS IN IMAGE DEEP NEURAL NETWORKS

ABDULRUHMAN I AHMED ABOMAKHLEB

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Information Technology)

Faculty of Computer and Mathematical Sciences

August 2026

ABSTRACT

Deep neural network (DNN) models are increasingly deployed in security-critical applications, raising concerns about their vulnerability to data poisoning attacks, particularly dirty-label and clean-label attacks. These attacks embed malicious behaviour while preserving accuracy, making detection difficult. Existing approaches focus on prevention, with limited support for post-deployment analysis, trace-back, and forensic attribution, leaving a critical gap in the investigation of compromised models. This thesis addresses these issues by proposing a digital forensic trace-back framework for post-deployment analysis of data poisoning attacks. The framework integrates three complementary algorithms designed to enable systematic detection, attribution, and reconstruction of adversarial manipulation in DNN systems. First, the Cosine Similarity Forensics for Deep Neural Networks (CSF-DNN) algorithm enables the detection and trace-back of poisoned samples in dirty-label backdoor attacks by analysing weight-update signatures using cosine similarity. Evaluated on the CIFAR-10 dataset (80% training and 20% testing samples) with the WideResNet architecture, CSF-DNN achieves 98.4% precision and 96.8% recall, outperforming the poison forensics benchmark approach. Secondly, the Autoencoder-Based Trigger Identification for Backdoor Forensic Attack (ATI-BFA) approach improves upon existing methods, which require reverse-engineering and assumptions about trigger patterns. Previous studies often fail when trigger characteristics differ from the patterns identified through reverse-engineering. To address this issue, ATI-BFA introduces a novel unsupervised autoencoder-based approach for trigger detection, enabling the identification and localisation of hidden trigger patterns directly from model behaviour without prior knowledge, achieving up to 91.5% accuracy while preserving clean-data performance. Finally, the Poison Image Forensics via Feature Clustering (PIFC) algorithm addresses a limitation of existing defences, which can be evaded by transferable clean-label attacks such as the Bullseye Polytope method. To tackle this issue, PIFC introduces a forensic approach leveraging feature-space clustering and Euclidean distance to detect and trace poisoned samples responsible for misclassification, demonstrating robust performance across varying attack intensities. PIFC-CLD achieves strong detection performance across varying attack intensities while minimising false positives. Collectively, these contributions establish an integrated forensic framework for tracing the sources of poisoning and identifying triggers. This framework advances AI security by adopting an evidence-driven forensic methodology.

ACKNOWLEDGEMENT

Firstly, I wish to thank God for giving me the opportunity to embark on my PhD and for completing this long and challenging journey successfully.

My sincere appreciation goes to my main supervisor, Prof. Dr. Kamarularifin Abd Jalil, and my co-supervisor, Assoc. Prof. Ts. Dr. Alya Geogiana Buja, for their invaluable guidance, encouragement, and continuous support throughout the course of this research.

Finally, this thesis is dedicated to the loving memory of my very dear late father and mother for their vision and determination to educate me. This piece of victory is dedicated to both of you. Alhamdulillah.

TABLE OF CONTENTS

		Page
CONFIRMATION BY PANEL OF EXAMINERS		**ii**
AUTHOR'S DECLARATION		**iii**
ABSTRACT		**iv**
ACKNOWLEDGEMENT		**v**
TABLE OF CONTENTS		**vi**
LIST OF TABLES		**xiii**
LIST OF FIGURES		**xv**
LIST OF SYMBOLS		**xviii**
LIST OF ABBREVIATIONS		**xix**
CHAPTER 1 INTRODUCTION		**1**
1.1	Introduction	1
1.2	Research Background	3
1.3	Research Motivation	8
1.4	Problem Statement	9
1.5	Research Objectives	12
1.6	Research Question	12
1.7	Significance of the Study	13
1.8	Research Contributions	13
1.9	Thesis Scope	15
1.10	Thesis Outline	16
CHAPTER 2 LITERATURE REVIEW		**19**
2.1	Introduction	19
2.2	Security Challenges in Deep Neural Networks	19
	2.2.1 Adversarial ML	20
	2.2.2 Adversarial ML in Practice	21
	2.2.3 Adversarial Life Cycle	21
	2.2.4 Threat Model	23

CHAPTER 1
INTRODUCTION

1.1 Introduction

Artificial intelligence (AI) has emerged as a transformative technological paradigm, with widespread adoption across critical domains such as healthcare, autonomous systems, and cybersecurity. These advancements have significantly improved efficiency and decision-making across critical domains. However, their security and resilience have become paramount as AI systems become integral to critical infrastructure and decision-making processes.

The growing reliance on AI has also introduced new risks and vulnerabilities, particularly in the context of adversarial manipulation and tampering. Cyber attackers are increasingly targeting AI models to exploit their weaknesses through techniques such as adversarial attacks, data poisoning attacks, and backdoor implantation. These attacks can compromise the integrity, confidentiality, and reliability of AI systems, leading to severe consequences, including financial loss, operational disruption, and even risks to human safety.

To address these challenges, researchers have been actively developing strategies to secure AI systems and detect malicious tampering or hacking attempts. These efforts include designing robust defence mechanisms against adversarial attacks, methods for detecting poisoned data, and strategies for mitigating backdoor vulnerabilities in deployed models. By securing AI systems, researchers aim to protect individual users and broader societal interests from the potentially catastrophic effects of compromised intelligent systems.

Furthermore, forensic analysis in AI has become increasingly important. Machine learning (ML) forensics refers to the post-hoc analysis of model behaviour to identify, trace, and attribute adversarial manipulations within trained systems. This area of research enhances defensive capabilities while improving accountability through traceability and auditability.

The study of AI security is vital to ensuring these powerful technologies can be leveraged safely and effectively. As AI continues to permeate every aspect of modern life, securing these systems against threats is no longer optional but an essential