Enhancing the Robustness of Credit Card Fraud Detection Systems Against Adversarial Attacks Using Machine Learning

Rathnayake R.M.C

219392J

MSc in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

Enhancing the Robustness of Credit Card Fraud Detection Systems Against Adversarial Attacks Using Machine Learning

Rathnayake R.M.C

219392J

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree MSc in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 05/06/2025

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. Sanath Jayasena

Signature of the Supervisor: Date: 05th June 2025

DEDICATION

This research is dedicated to the countless individuals who have fallen victim to credit card fraud, whose experiences have fueled the urgency for robust and resilient fraud detection systems. It is also dedicated to the tireless efforts of researchers, practitioners, and financial institutions working diligently to protect consumers from these malicious activities. May this research contribute to the collective knowledge and efforts aimed at safeguarding the integrity of digital financial transactions.

ACKNOWLEDGEMENT

The successful completion of this research would not have been possible without the invaluable contributions and support of numerous individuals and organizations, I extend my deepest appreciation to my advisor, Prof. Sanath Jayasena, for providing unwavering guidance, mentorship, and encouragement throughout this research journey. Your expertise, patience, and belief in my abilities have been instrumental in shaping this work. Thank you for your invaluable contributions to this research.

I am profoundly humbled and grateful for the opportunity to have embarked on this journey, and I earnestly hope that the findings of this work will contribute to the advancement of knowledge and the betterment of society.

ABSTRACT

Credit card fraud detection systems are essential for protecting money transactions and stopping illegal use. The usefulness of these systems is seriously threatened by the increasing prevalence of adversarial assaults, as attackers try to alter input data in order to trick machine learning algorithms and evade detection. The problem of making credit card fraud detection systems more resilient to such hostile attacks is addressed in this study. It explores current defense tactics, analyses the status of adversarial attacks, and evaluates how they affect model performance. The study offers a comprehensive strategy that involves implementing several defense strategies, modelling hostile settings, and assessing their effectiveness using important performance indicators. The results show that the proposed approach greatly enhances the robustness of fraud detection models, effectively reducing the influence of adversarial manipulations

Credit card fraud detection systems face a significant challenge as adversarial attacks grow more complex, enabling attackers to alter data and avoid detection. By methodically evaluating defense tactics against these threats, this study seeks to improve fraud detection models. Various adversarial attack types, including hybrid, white-box, and black-box attacks, are simulated to identify weaknesses in current systems. Four defense mechanisms, namely, Neural Cleanse, Random Noise, General Adversarial Training, and Defensive Distillation are implemented and assessed. The results demonstrate that none of these methods offer total protection, although they improve model resilience to some degree. It backs up the claim that financial institutions do not yet have infallible defenses against hostile attacks. The study provides insightful information about enhancing fraud prevention by outlining the advantages and disadvantages of each strategy.

This study supports the development of more effective credit card fraud detection systems by offering actionable insights and a structured approach to designing and assessing protective strategies against adversarial threats. The findings have implications for financial institutions seeking to fortify their security posture and protect customers from fraudulent activities, fostering trust and confidence in digital financial transactions.

Keywords: Credit card fraud; Adversarial attacks; Robustness; Defense mechanisms; Machine learning; Security; Financial transactions; Fraud detection

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACOFS	Ant Colony Optimization Feature Selection
AUC	Area Under the Curve
AUC-ROC	Area Under the ROC Curve
CNN	Convolutional Neural Network
DEFS	Differential Evolutionary Feature Selection
DNNs	Deep Neural Networks
ELM	Extreme Learning Method
FP	False Positive
FN	False Negative
GA	Genetic Algorithm
GAFS	Genetic Algorithm Feature Selection
GBM	Gradient Boosting Machine
GAT	Generative Adversarial Trainer
KNN	K-Nearest Neighbors
LR	Logistic Regression
LOF	Local Outlier Factor
LGBM	Light Gradient Boosting Machine
NB	Naive Bayes
PCA	Principal Component Analysis
PSOFS	Particle Swarm Optimization Feature Selection
RF	Random Forest
ROC	Receiver Operating Characteristic
RHSO	Rock Hyrax Swarm Optimization
SVM	Support Vector Machine
TN	True Negative
TP	True Positive