

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
YAHIA FARES UNIVERSITY OF MEDEA



Faculty of Sciences

Department of Mathematics and Computer Science

Final Master Project

Branch: Computer science

Specialty: Systems Engineering and Web Technologies

Deep Q-Learning based approach for DoS attacks detection

Prepared and presented by:
Haitham Attab
Amine Takdenti

Supervisor:
Prof. Brahim Sahmadi

To my family, for their endless love and support.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Brahim Sahmadi, for their invaluable guidance and unwavering support throughout this research.

I also wish to thank [mention other people, e.g., colleagues, friends, family] for their encouragement and insightful discussions.

This work was made possible by [mention any funding or support].

Abstract

With the growing frequency and sophistication of Denial-of-Service (DoS) attacks, robust and adaptive detection mechanisms have become critical to maintaining the availability and reliability of computer networks. These attacks can severely disrupt essential services and lead to substantial financial losses, highlighting the need for intelligent and resilient defense strategies.

This thesis introduces a novel intrusion detection system based on Reinforcement Learning (RL) to detect and mitigate DoS attacks. By framing network security as a sequential decision-making problem, the proposed approach enables an RL agent to learn optimal detection policies through continuous interaction with a simulated network environment. The agent incrementally refines its ability to differentiate between benign and malicious traffic by leveraging a reward-based feedback mechanism that reinforces accurate classifications and penalizes misclassifications.

Key traffic features are systematically extracted and provided as state inputs to the RL agent, facilitating effective and dynamic behavioral analysis. This adaptive learning paradigm enhances the system's capability to recognize evolving attack patterns, ultimately contributing to a more proactive and resilient network defense against DoS threats.

ملخص

مع تزايد وتيرة وتعقيد هجمات حجب الخدمة (DoS) أصبحت آليات الكشف المتقدمة والمتكيفة ضرورية لضمان توافر الشبكات وموثوقيتها. يمكن لهذه الهجمات أن تعطل الخدمات الحيوية وتسبب في خسائر مالية كبيرة، مما يُبرز الحاجة إلى استراتيجيات دفاع ذكية ومرنة.

تقدم هذه المذكرة نظام كشف تسلسل جديد يعتمد على التعلم المعزز (Reinforcement Learning) لاكتشاف هجمات DoS والتصدي لها. من خلال تصوير أمن الشبكات كمسألة اتخاذ قرارات متسلسلة، يتمكن العميل (Agent) في التعلم المعزز من تعلم سياسات الكشف المثلى من خلال التفاعل المستمر مع بيئة شبكة محاكاة. ويقوم بتحسين قدرته تدريجياً على التمييز بين حركة المرور السليمة والخبيثة باستخدام نظام تغذية راجعة قائم على المكافآت، يعزز التصنيفات الصحيحة ويعاقب التصنيفات الخاطئة.

يتم استخراج ميزات حركة المرور الأساسية بشكل منهجي وتقديمها كمدخلات حالة للعميل، مما يمكنه من تحليل السلوك بفعالية وديناميكية. تسهم هذه المنهجية التعليمية التكيفية في تعزيز قدرة النظام على التعرف على أنماط الهجمات المتغيرة، مما يؤدي في النهاية إلى دفاع شبكي أكثر مرونة واستباقية ضد تهديدات DoS.

Résumé

Face à la fréquence et à la complexité croissantes des attaques par déni de service (DoS), la mise en place de mécanismes de détection robustes et adaptatifs est devenue essentielle pour garantir la disponibilité et la fiabilité des réseaux informatiques. Ces attaques peuvent perturber gravement les services critiques et engendrer des pertes financières importantes, soulignant ainsi la nécessité de stratégies de défense intelligentes et résilientes.

Ce mémoire propose un système de détection d'intrusion innovant basé sur l'apprentissage par renforcement (Reinforcement Learning) pour détecter et atténuer les attaques DoS. En formulant la sécurité réseau comme un problème de prise de décision séquentielle, notre approche permet à un agent d'apprentissage de découvrir des politiques de détection optimales via une interaction continue avec un environnement réseau simulé. L'agent améliore progressivement sa capacité à distinguer le trafic bénin du trafic malveillant grâce à un système de rétroaction basé sur des récompenses, qui renforce les classifications correctes et pénalise les erreurs.

Les principales caractéristiques du trafic sont extraites de manière systématique et utilisées comme entrées d'état pour l'agent, facilitant une analyse comportementale efficace et dynamique. Ce paradigme d'apprentissage adaptatif améliore la capacité du système à reconnaître les schémas d'attaque en évolution, contribuant ainsi à une défense réseau plus proactive et résiliente contre les menaces DoS.

Contents

Acknowledgements	3
Abstract	4
5	Arabic in Abstract
Abstract in French	6
List of Figures	9
List of Tables	10
List of Acronyms	11
1 Intrusion Detection System (IDS)	16
1.1 Security Fundamentals	16
1.1.1 Core Security Concepts	16
1.2 Denial-of-Service (DoS) Attacks	17
1.2.1 Volumetric Attacks	17
1.2.2 Protocol Attacks	18
1.2.3 Application Layer Attacks	18
1.2.4 Distributed Denial-of-Service (DDoS)	19
1.2.5 Network Traffic Analysis	19
1.3 Intrusion Detection Systems (IDS)	19
1.3.1 IDS Architecture	20
1.3.2 IDS Types and Functionality	21
1.4 Classification in Security Contexts	23
1.4.1 Classification Paradigms	23
1.4.2 Machine Learning Techniques	23
2 Deep Q-Learning for Intrusion Detection	24
2.1 Introduction	24
2.2 Machine Learning Foundations	24
2.2.1 Introduction to Machine Learning	24
2.3 Machine Learning Paradigms	24
2.3.1 Supervised Learning	24
2.3.2 Unsupervised Learning	25
2.3.3 Reinforcement Learning	26
2.4 Markov Decision Process (MDP)	26
2.4.1 MDP Components	27
2.4.2 Markov Property	27
2.5 Reinforcement Learning Approaches	27
2.5.1 Value-Based Methods	27

2.5.2	Policy-Based Methods	27
2.5.3	Actor-Critic Methods	27
2.6	Deep Learning: Key Concepts	28
2.7	Reinforcement Learning	29
2.8	Deep Q-Learning (DQN)	30
2.9	IDS Implementation with Deep Learning	31
2.9.1	Introduction	31
2.9.2	Deep Learning Approaches in IDS	32
2.9.3	Implementation Strategies	32
2.9.4	Application Domains	33
2.9.5	Performance Advantages	33
2.9.6	Implementation Framework	33
2.9.7	Future Directions	34
2.10	Conclusion	34
3	Designing DQN Architectures for DoS/DDoS Attack Detection	35
3.1	Introduction	35
3.2	Problem Modeling for DoS/DDoS Detection using DQN	35
3.2.1	Environment Definition	35
3.2.2	Data Preprocessing Pipeline	36
3.2.3	Proposed DQN Architectures for Attack Detection	38
3.2.4	Conclusion	42
4	Experiments and Results	44
4.1	Introduction	44
4.2	Development Tools and Experimental Setup	44
4.2.1	Development Environment	44
4.2.2	Evaluation Metrics	45
4.3	Description of Datasets	45
4.3.1	Dataset Characteristics and Relevance	46
4.3.2	Preprocessing and Experimental Setup	46
4.4	Binary Classification Results	46
4.4.1	Results of MLP-DQN	46
4.4.2	Results of DDQN	48
4.4.3	Results of CNN-DQN	50
4.5	Multiclass Classification Results	52
4.5.1	Results of MLP-DQN	52
4.5.2	Results of CNN-DQN	53
4.6	Comparison and Discussion	55
4.7	Conclusion	57
	General Conclusion	58

List of Figures

1.1	Principles of information security	17
1.1	17
1.2	DoS and DDoS Attack Traffic Comparison	18
1.3	Network-based Intrusion Detection System (NIDS) Architecture	20
1.4	Network-based Intrusion Detection System (NIDS) Diagram	21
1.5	Host-based Intrusion Detection System (HIDS) Diagram	22
2.1	Fundamental structure of a deep neural network, showing the flow from input data through hidden layers to an output prediction.	28
2.2	The Agent-Environment Interaction Loop in Reinforcement Learning	30
4.1	DQN NN confusion matrix.	47
4.2	DQN learning rate and epsilon decay.	47
4.3	DQN NN training curves.	48
4.4	DDQN NN confusion matrix.	49
4.5	DDQN learning rate and epsilon decay.	49
4.6	DDQN NN training curves.	49
4.7	DDQN NN confusion matrix.	50
4.8	DDQN learning rate and epsilon decay.	50
4.9	DDQN NN training curves.	51
4.10	Confusion Matrix for the MLP-DQN Model on Multi-class Classification	52
4.11	Learning Rate and Epsilon Decay Schedule during MLP-DQN Training	53
4.12	Training Curve of MLP-DQN: Episode Rewards over Time	53
4.13	Confusion Matrix for the MLP-DQN Model on Multi-class Classification	54
4.14	Learning Rate and Epsilon Decay Schedule during MLP-DQN Training	54
4.15	Training Curve of MLP-DQN: Episode Rewards over Time	54
4.16	Accuracy Comparison Across All Models and Tasks	55
4.17	F1-Score Comparison Across All Models and Tasks	55
4.18	Model complexity Comparison Across All Models and Tasks	56
4.19	Performance metrics Across All Models and Tasks	56

List of Tables

4.1	Development Tools and Libraries	44
4.2	Binary Classification Results for MLP-DQN	46
4.3	Binary Classification Results for MLP-DDQN	48
4.4	Binary Classification Results for CNN-DQN	50
4.5	Multi-class Classification Results for MLP-DQN	52
4.6	Multi-class Classification Results for CNN-DQN	53

List of Acronyms

AI	Artificial Intelligence
CIA	Confidentiality, Integrity, and Availability
CIC	Canadian Institute for Cybersecurity
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSV	Comma-Separated Values
DDoS	Distributed Denial-of-Service
DNS	Domain Name System
DoS	Denial-of-Service
DQN	Deep Q-Network
FTP	File Transfer Protocol
HIDS	Host-based Intrusion Detection System
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection System
IP	Internet Protocol
IoT	Internet of Things
ML	Machine Learning
MLP	Multi-Layer Perceptron
NIDS	Network-based Intrusion Detection System
NTP	Network Time Protocol
PCAP	Packet Capture
RL	Reinforcement Learning
SIEM	Security Information and Event Management
SSDP	Simple Service Discovery Protocol

SVM Support Vector Machine

TCP Transmission Control Protocol

UDP User Datagram Protocol

General Introduction

Context and Motivation

In the contemporary digital era, the reliability and continuous availability of network services form the backbone of global commerce, communication, and critical infrastructure. A primary threat to this stability is the Denial-of-Service (DoS) attack, and its more potent variant, the Distributed Denial-of-Service (DDoS) attack. These malicious activities aim to render online services and network resources unavailable to legitimate users by overwhelming them with an insurmountable volume of traffic or by exploiting protocol vulnerabilities. As attackers devise increasingly sophisticated and large-scale attacks, the need for intelligent, adaptive, and robust defense mechanisms has become more critical than ever.

Traditional Intrusion Detection Systems (IDS) have served as the first line of defense. However, they often rely on static, signature-based methods that, while effective against known threats, are inherently incapable of identifying novel or zero-day attacks. Anomaly-based systems, which attempt to solve this by creating a baseline of "normal" behavior, frequently suffer from high false positive rates, leading to operational friction and alert fatigue. Furthermore, conventional machine learning classifiers, though an improvement, often operate as static models that require periodic retraining and struggle to adapt in real-time to the dynamic and adversarial nature of cybersecurity threats.

This clear gap in defensive capabilities motivates the exploration of more advanced paradigms. Reinforcement Learning (RL), a subfield of machine learning, offers a compelling alternative. Unlike supervised learning, RL focuses on training an autonomous agent to make optimal decisions through interaction with an environment, guided by a system of rewards and penalties. This approach is inherently adaptive, allowing the agent to learn and refine its strategies over time. The potential to create a self-improving, context-aware detection system that can respond dynamically to evolving threats provides the core motivation for this research.

Problem Statement

The central problem addressed by this thesis is the inadequacy of traditional and static machine learning-based Intrusion Detection Systems in providing adaptive, real-time defense against modern, sophisticated DDoS attacks. These existing systems exhibit several key limitations:

- **Lack of Adaptability:** Signature-based systems cannot detect new attack vectors, while static ML models fail to adapt to evolving attack patterns without manual retraining.
- **High False Positive Rates:** Anomaly detection systems often misclassify legitimate traffic spikes (flash crowds) as attacks, disrupting normal operations.
- **Reactive Nature:** Most systems are reactive, identifying attacks only after they have begun to impact the network, rather than learning to predict and preempt them.

This research seeks to answer the following question: **Can Reinforcement Learning, specifically a Deep Q-Network (DQN) approach, be effectively adapted to the supervised classification task of DDoS detection to create a more dynamic, accurate, and autonomous system?**

Objectives

To address the problem statement, this thesis sets out the following primary objectives:

1. To conduct a thorough review of Denial-of-Service attacks, traditional Intrusion Detection Systems, and the foundational principles of Reinforcement Learning to establish a strong theoretical basis.
2. To design and develop a novel detection framework that reframes the DDoS classification task as a sequential decision-making problem, solvable with a Deep Q-Network (DQN) agent.
3. To implement two distinct neural network architectures for the DQN’s Q-function approximator—a Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN)—to explore different feature representation capabilities.
4. To train and rigorously evaluate the proposed models using a modern, comprehensive benchmark dataset (CIC-DDoS2019), measuring performance based on key metrics such as accuracy, precision, recall, and F1-score.
5. To compare the performance of the MLP-based and CNN-based DQN agents to determine the most effective architecture for this task.

Overall Methodology

The research will be conducted through a structured, multi-stage process:

1. **Data Preprocessing:** The CIC-DDoS2019 dataset will be acquired and subjected to a comprehensive preprocessing pipeline. This includes data cleaning, binary label transformation, feature selection using a Random Forest classifier to identify the most salient features, Min-Max normalization to scale the data, and random undersampling to create a balanced dataset for training.
2. **Model Design:** A custom Reinforcement Learning environment will be simulated from the static dataset. A Deep Q-Network (DQN) agent will be designed, defining the state space (from network features), action space (classify as 'Benign' or 'Attack'), and a reward function that provides feedback based on classification correctness.
3. **Architectural Implementation:** Two Q-network architectures will be implemented in Python using deep learning libraries. The first will be a standard Multi-Layer Perceptron. The second will be a Convolutional Neural Network, which requires reshaping the one-dimensional feature vector into a two-dimensional "feature-image" to exploit spatial correlations between features.
4. **Training and Evaluation:** Both models will be trained using a novel loop that blends supervised signals (the true labels) with RL principles (value estimation of future states). The trained models will be evaluated on a held-out test set to assess their detection performance and generalization capabilities.

Thesis Structure

This thesis is organized into six chapters, each building upon the last to present a complete account of the research:

- **Chapter 1: Intrusion Detection Systems (IDS)** lays the theoretical groundwork by discussing fundamental security concepts, detailing the various types of DoS/DDoS attacks, and explaining the architecture and limitations of traditional IDS.
- **Chapter 2: Reinforcement Learning for Cybersecurity** introduces the core concepts of Reinforcement Learning, with a specific focus on Deep Q-Learning. It explores how the RL paradigm can be modeled for security tasks, covering state, action, and reward design.

- **Chapter 3: Architectural Blueprint and Implementation** presents the technical core of this work. It details the data preprocessing pipeline and the architectural design of the two proposed DQN agents (MLP and CNN-based).
- **Chapter 4: Experiments and Results** will describe the experimental setup, present the performance results of the trained models, and provide a comparative analysis and discussion of the findings.

Chapter 1

Intrusion Detection System (IDS)

1.1 Security Fundamentals

Introduction

Security fundamentals refer to the essential principles, concepts, and practices that form the foundation of information security. These fundamentals encompass a wide range of technical and organizational measures aimed at protecting sensitive information and systems from unauthorized access, theft, damage, or other forms of compromise.

Key security fundamentals include Confidentiality, Integrity, Availability, Authentication, Authorization, Encryption, Risk Management, Incident Response, and Disaster Recovery. Together, these principles establish the basis of a comprehensive information security program, enabling organizations to effectively safeguard their critical information assets and maintain the trust of stakeholders.[1]

1.1.1 Core Security Concepts

Network security encompasses strategies and technologies to protect systems from cyber threats, particularly Denial of Service (DoS) attacks that aim to disrupt service availability. Several core principles serve as the foundation for secure systems.

CIA Triad

Confidentiality Confidentiality ensures that sensitive information is only accessible to authorized users. Techniques such as encryption, user authentication, and access control policies prevent unauthorized data access. While DoS attacks do not typically aim to breach confidentiality directly, successful exploitation may lead to indirect confidentiality violations if attackers cause service misconfigurations or force failovers to insecure states.

Integrity Integrity guarantees that data is accurate and unaltered. It is maintained through cryptographic hash functions, checksums, and digital signatures that validate whether data has been tampered with. During a DoS attack, integrity may be compromised by interrupting legitimate updates or corrupting processes due to system overload.

Availability Availability ensures that services and systems remain accessible to legitimate users at all times. This principle is the primary target of DoS attacks, which flood networks or services with excessive requests, causing slowdowns or complete denial of access. Maintaining availability involves redundancy, load balancing, and proactive mitigation strategies like rate limiting and firewalls.



Figure 1.1: Principles of information security

Source: [2]

Extended Security Principles

Non-repudiation Non-repudiation guarantees that an entity cannot deny having performed a particular action, such as sending a message or initiating a transaction. This is enforced through digital signatures and secure logging mechanisms. In the context of DoS attacks, non-repudiation helps trace attack origins and supports legal accountability.

Authenticity Authenticity confirms that data, communications, or users are genuine and not forged. Authentication protocols, digital certificates, and cryptographic techniques are used to ensure that data comes from trusted sources. This is crucial for filtering legitimate traffic from spoofed attack traffic in DoS scenarios.

Accountability Accountability ensures that all actions within a system can be traced to responsible users or processes. It involves logging, auditing, and monitoring to track behavior. Accountability is vital for forensic analysis after a DoS attack and for strengthening defenses against future intrusions.

1.2 Denial-of-Service (DoS) Attacks

Introduction Denial-of-Service (DoS) attacks aim to make a system or network resource unavailable to its intended users by overwhelming it with excessive traffic or exploiting protocol-level vulnerabilities. These attacks can disrupt services, degrade performance, or completely shut down access. The most common types include:

1.2.1 Volumetric Attacks

Volumetric attacks aim to saturate a target's bandwidth by generating an overwhelming amount of traffic. This often involves amplification techniques or high-rate packet floods. Common examples include UDP floods and DNS amplification attacks, which leverage misconfigured servers to multiply traffic directed at the victim [3].

In a UDP flood, attackers send large numbers of spoofed UDP packets to random or specific ports, consuming bandwidth and processing power. DNS amplification uses small queries to open resolvers with the victim's IP address, causing them to return large responses to the victim.

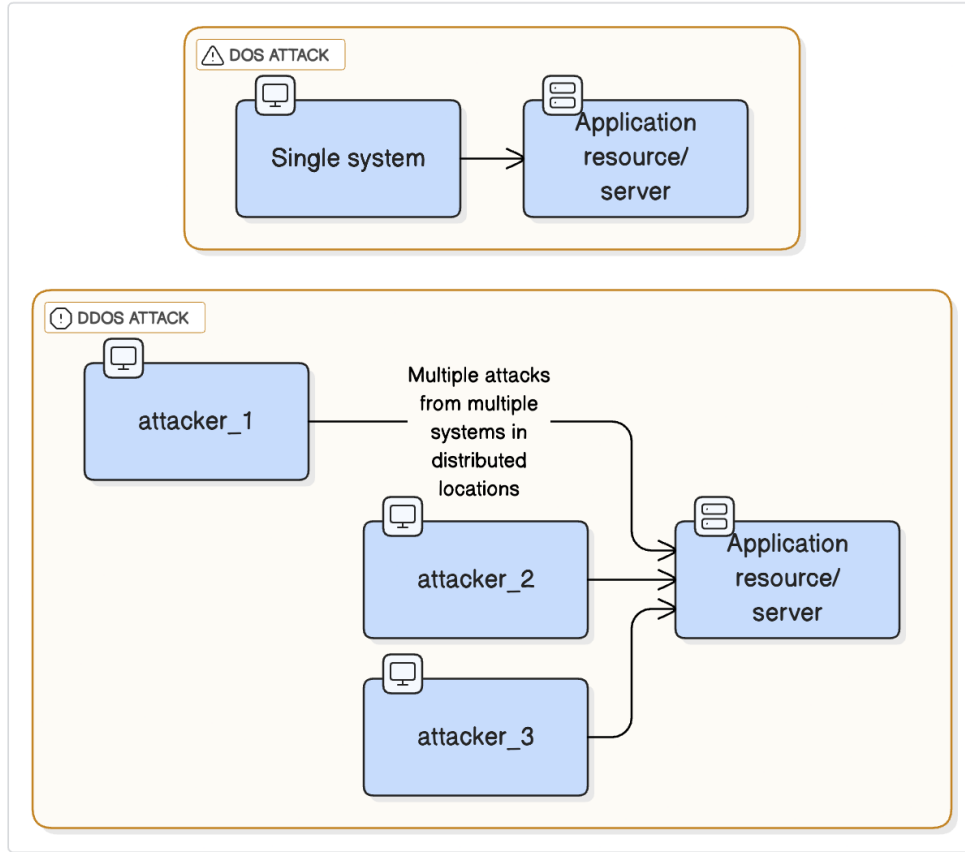


Figure 1.2: DoS and DDoS Attack Traffic Comparison

A notable example is the Mirai botnet, which infected hundreds of thousands of insecure IoT devices to coordinate massive traffic streams toward its targets [3]. The goal of volumetric attacks is to clog network links (Gbps or Tbps scale), preventing legitimate access.

1.2.2 Protocol Attacks

Protocol attacks exploit weaknesses in Layer 3/4 protocols (network or transport layer) to exhaust server or network resources. Unlike volumetric attacks, they do not require high bandwidth but instead consume stateful resources like connection tables or CPU cycles [3].

A classic example is the TCP SYN flood, where attackers send numerous SYN packets without completing the handshake, filling the server’s connection queue [4]. ICMP floods and Ping of Death attacks exploit the Internet Control Message Protocol to crash or freeze systems by sending malformed or excessive traffic.

Other examples include fragmentation attacks (e.g., Teardrop), which send overlapping IP fragments to crash reassembly logic, and ACK/FIN floods that exhaust firewall state tables.

1.2.3 Application Layer Attacks

Application-layer (Layer 7) attacks generate traffic that appears legitimate at the application level but consumes excessive server resources such as CPU, memory, or threads [3]. These attacks are difficult to detect because they mimic real user behavior.

HTTP GET/POST floods can overwhelm web servers by triggering costly operations. A well-known example is Slowloris, which sends partial HTTP headers slowly to hold open many connections and exhaust the web server’s pool [5]. Other examples include RUDY (R-U-Dead-Yet) attacks and HTTP floods targeting dynamic or database-backed content.

1.2.4 Distributed Denial-of-Service (DDoS)

DDoS attacks use multiple compromised machines (botnets) to launch coordinated attacks, making them harder to block and vastly more powerful than single-source DoS attacks [3].

Botnets like Mirai leverage IoT devices to simultaneously launch volumetric or protocol-based attacks. Reflective amplification (e.g., using open DNS/NTP servers) further increases traffic impact. Mitigation requires upstream filtering, rate-limiting, and often third-party scrubbing services.

1.2.5 Network Traffic Analysis

Mitigating DoS/DDoS attacks relies on monitoring and anomaly detection. This involves establishing a baseline of normal network behavior and identifying deviations in volume, protocol use, or source IPs [3, 6].

Anomaly-based intrusion detection systems (IDS) can flag unusual spikes, connection patterns, or TCP flag anomalies. Signature-based systems match known attack patterns. Flow analysis tools (e.g., NetFlow, sFlow) help detect irregular byte/packet rates or unusual source-destination pairs.

Advanced methods include machine learning to classify deviations. Once detected, defenses such as rate limiting, traffic shaping, or redirection to mitigation services are employed to maintain availability.

1.3 Intrusion Detection Systems (IDS)

An **Intrusion Detection System (IDS)** is a cybersecurity solution designed to monitor network or system activities for malicious actions or policy violations. Upon detecting such activities, the IDS typically alerts system administrators or integrates with centralized security tools like Security Information and Event Management (SIEM) systems to facilitate a coordinated response [7].

1.3.1 IDS Architecture

Monitored System Feedback Flow

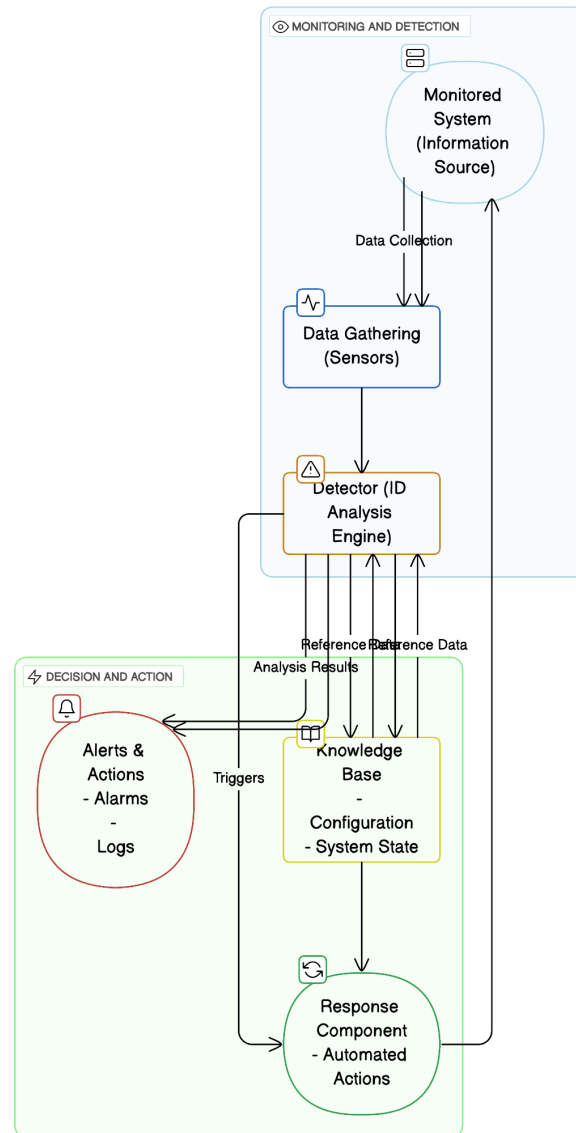


Figure 1.3: Network-based Intrusion Detection System (NIDS) Architecture

Modern IDS platforms implement modular architectures with several key components:

- **Sensors:** are the front line of an IDS, deployed at strategic network chokepoints—such as mirror (SPAN) ports, network taps, or virtual interfaces in cloud environments—to collect raw traffic data.
- **Analysis Engines:** form the IDS’s ”brain,” taking the sensor-collected data and applying a mixture of rule-based and intelligent techniques to detect threats.
- **Knowledge Base:** underpins all detection logic by maintaining up-to-date signatures, behavioral profiles, and historical datasets.
- **Response Systems:** close the loop between detection and defense. Once the analysis engine assigns a confidence score to an event, the response component generates alerts.

For DOS attack detection, these components must operate with high efficiency and scale to process enormous traffic volumes during attack scenarios.

1.3.2 IDS Types and Functionality

Intrusion Detection Systems serve as the primary monitoring and detection layer for network security threats, including DOS attacks:

- **Network-based IDS (NIDS):** A NIDS is a monitoring solution placed at key junctions within a network to passively capture and inspect all passing traffic.

NIDS (Network-based Intrusion Detection System)

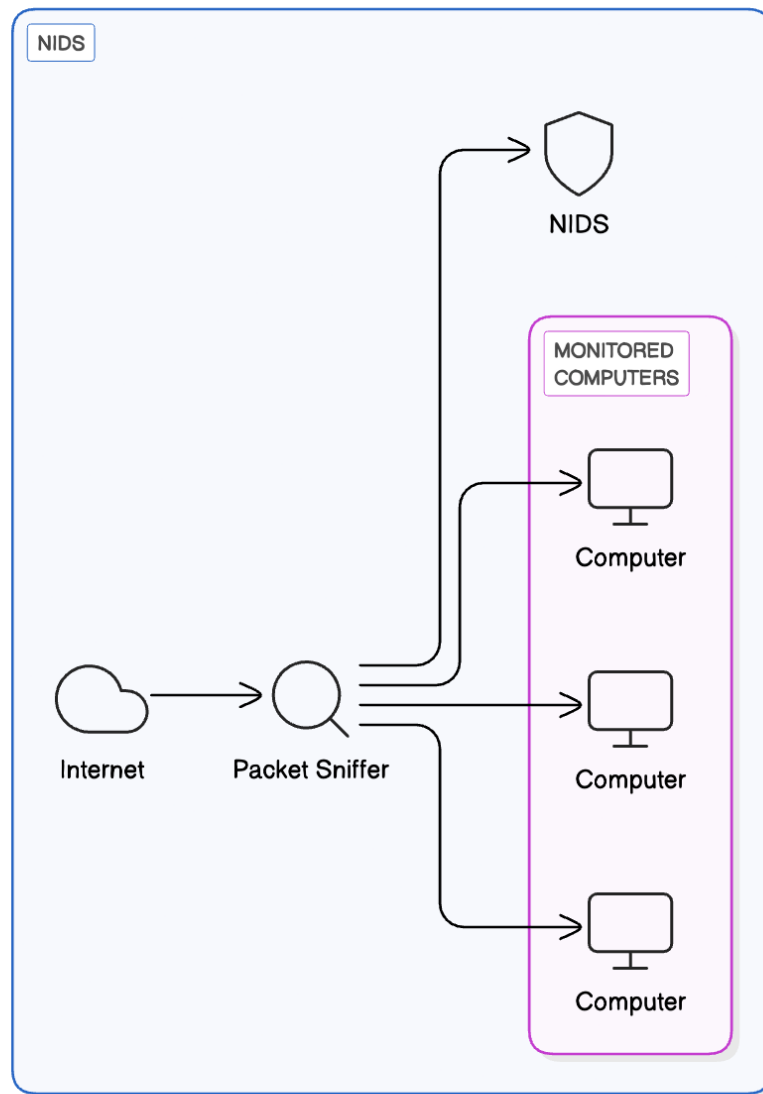


Figure 1.4: Network-based Intrusion Detection System (NIDS) Diagram

- **Host-based IDS (HIDS):** A HIDS operates at the level of individual endpoints or servers, offering deep visibility into system-specific activities.

HIDS Architecture Comparison

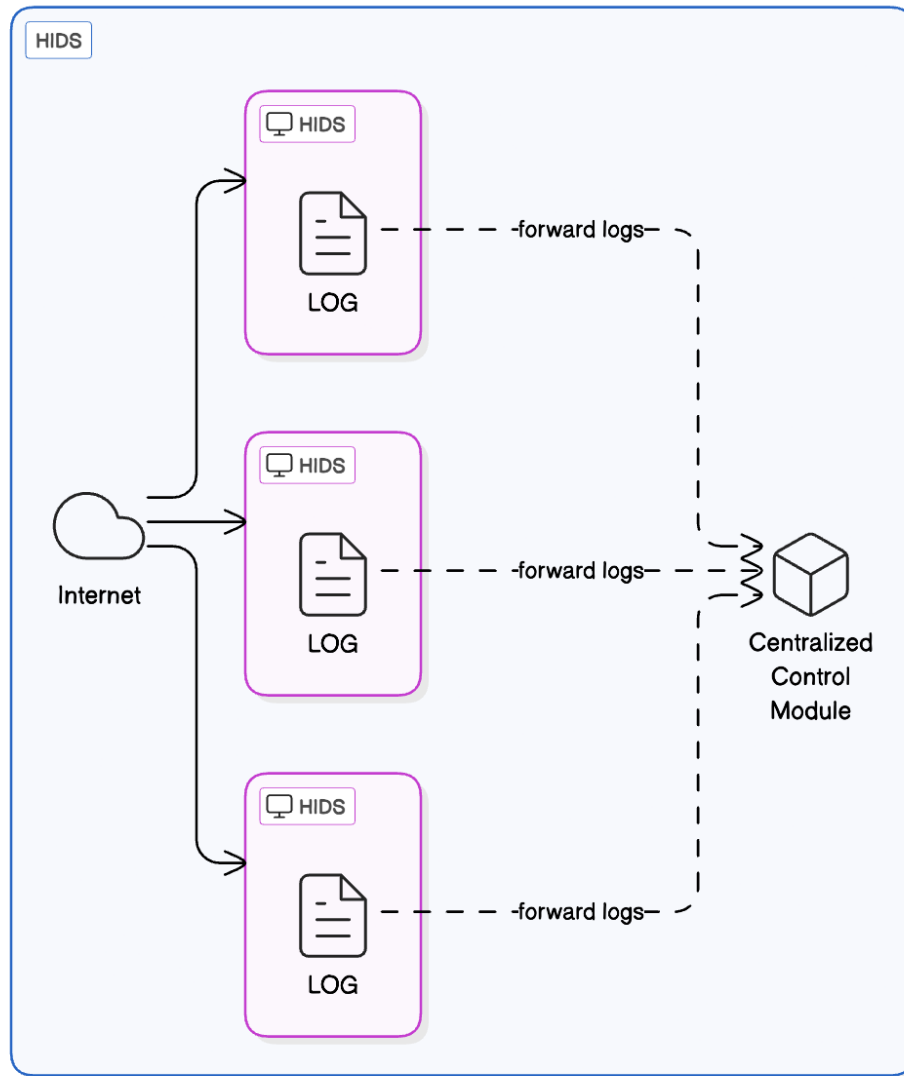


Figure 1.5: Host-based Intrusion Detection System (HIDS) Diagram

- **Detection Methodologies:**
 - **Signature-based:** Matches observed traffic against known attack patterns (signatures). Highly effective for known attacks but ineffective against zero-day attacks.
 - **Anomaly-based:** Establishes baselines of normal behavior and flags significant deviations. Can detect unknown attacks but may generate more false positives.
 - **Hybrid Systems:** Combine both methodologies to leverage their respective strengths and reduce false positives.

Limitations of Traditional IDS for DOS attack detection include difficulty processing high-volume traffic, distinguishing flash crowds from attacks, limited adaptation, high false-positive rates, and high resource consumption. These limitations necessitate AI-driven approaches like reinforcement learning.

1.4 Classification in Security Contexts

1.4.1 Classification Paradigms

Classification forms the foundation of automated threat detection: **Binary Classification:** Distinguishes between normal and attack traffic. **Multi-class Classification:** Distinguishes between multiple attack types. **Hierarchical Classification:** Structures threats in a tree-like format. **Multi-label Classification:** Assigns multiple categories to a single traffic flow.

What is Machine Learning?

Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn. The learning system consists of a decision process, an error function, and a model optimization process.

1.4.2 Machine Learning Techniques

Supervised Learning

Supervised learning involves training on labeled datasets. Common algorithms include Random Forests, Support Vector Machines (SVM), and Neural Networks.

Unsupervised Learning

Unsupervised learning focuses on uncovering patterns in unlabeled data. Techniques include clustering (k-means) and anomaly detection (Isolation Forest).

Semi-Supervised Learning

Chapter 2

Deep Q-Learning for Intrusion Detection

2.1 Introduction

This chapter establishes the theoretical foundation for the novel intrusion detection methodology proposed in this thesis. The objective is to move beyond traditional, static detection systems by leveraging the adaptive and autonomous capabilities of modern artificial intelligence. We begin by introducing the key concepts of Deep Learning (DL), which provides the powerful function approximation tools needed to handle complex, high-dimensional data.

Subsequently, we delve into the paradigm of Reinforcement Learning (RL), a field of machine learning focused on goal-oriented learning through interaction. We will outline the fundamental components of an RL system—the agent, environment, states, actions, and rewards—that form the basis of sequential decision-making.

The core of this chapter brings these two fields together in the form of Deep Q-Learning (DQN). We will explain how DQN combines the perceptive power of deep neural networks with the decision-making framework of Q-learning to solve complex problems with large state spaces. Finally, we discuss the application of these advanced techniques to the domain of Intrusion Detection Systems (IDS), setting the stage for the specific architectural designs and experiments detailed in the subsequent chapters.

2.2 Machine Learning Foundations

2.2.1 Introduction to Machine Learning

Machine Learning (ML) represents a fundamental branch of artificial intelligence that enables systems to automatically learn and improve from experience without being explicitly programmed. In the context of cybersecurity and Intrusion Detection Systems (IDS), machine learning provides the computational foundation for intelligent threat detection, pattern recognition, and adaptive security responses.

2.3 Machine Learning Paradigms

2.3.1 Supervised Learning

Supervised learning forms the foundation of many security applications, utilizing labeled datasets to train models that can predict outcomes for new, unseen data. In this paradigm, algorithms learn from input-output pairs where the desired outcome is known during the training phase.

Key Characteristics

- Requires labeled training datasets with known classifications.
- Learns mapping functions from input features to target outputs.
- Performance can be measured through validation on test datasets.
- Provides predictable and interpretable results.

Common Algorithms

- **Classification:** Support Vector Machines (SVM), Random Forest, Neural Networks, Naive Bayes.
- **Regression:** Linear Regression, Polynomial Regression, Ridge Regression.

Applications in Cybersecurity

Malware Detection Models are trained on vast repositories of known malware samples (labeled as malicious) and benign software (labeled as safe). By analyzing features such as file structure, API calls, and byte sequences, the model learns to identify the signatures and behavioral patterns of malicious code, enabling it to classify new, unseen files.

Email Spam Classification A classic application where a model learns to distinguish between legitimate emails ('ham') and unsolicited spam. It is trained on a large corpus of labeled emails, learning to identify patterns in subject lines, sender information, email content, and header metadata to effectively filter incoming messages.

Network Traffic Classification This is a core task for many IDS implementations. A model is trained on a dataset of network flows that have been pre-labeled as either 'benign' or a specific type of 'attack'. The algorithm learns to recognize the statistical and temporal properties of malicious traffic, allowing it to classify live network activity in real-time.

Insider Threat Detection Supervised models can establish a baseline of normal user behavior by training on historical data of user activities (e.g., login times, file access patterns, resource usage). The trained model can then monitor ongoing activities and flag any significant deviations from these learned profiles, which may indicate a compromised account or malicious insider activity.

2.3.2 Unsupervised Learning

Unsupervised learning addresses scenarios where labeled data is unavailable, focusing on discovering hidden patterns and structures within datasets. This paradigm is particularly valuable in cybersecurity for detecting unknown threats and anomalous behaviors.

Key Characteristics

- No labeled training data required.
- Focuses on pattern discovery and data structure identification.
- Identifies anomalies and outliers in system behavior.
- Adaptive to new and evolving patterns.

Common Techniques

- **Clustering:** K-means, DBSCAN, Hierarchical Clustering.
- **Dimensionality Reduction:** Principal Component Analysis (PCA), t-SNE.
- **Anomaly Detection:** Isolation Forest, One-Class SVM.

Applications in Cybersecurity

Zero-Day Attack Detection The primary strength of unsupervised learning lies in its ability to identify novel threats. By training on a vast amount of normal network traffic, the model learns the intrinsic structure of benign behavior. Any new traffic that deviates significantly from this learned structure is flagged as an anomaly, providing a powerful mechanism for detecting previously unseen, or 'zero-day', attacks.

Network Behavior Baseline Establishment Clustering algorithms can automatically group similar network flows together without predefined labels. This allows the system to autonomously define what constitutes "normal" behavior by identifying dominant clusters of activity. Smaller, more isolated clusters or outliers can then be investigated as potential threats.

Advanced Persistent Threat (APT) Identification APTs are characterized by slow, low-profile activities over extended periods. Unsupervised methods, particularly those focused on association or correlation analysis, can uncover subtle, long-term relationships between seemingly unrelated events across the network, which might indicate the coordinated and persistent nature of an APT.

Log Analysis for Incident Discovery Security logs are often vast and unstructured. Unsupervised techniques can automatically cluster log entries to find unusual event correlations or identify rare event sequences that deviate from the norm. This helps security analysts sift through immense amounts of data to pinpoint potential incidents that would otherwise go unnoticed.

2.3.3 Reinforcement Learning

Reinforcement Learning (RL) is an advanced paradigm where an intelligent agent learns an optimal decision-making policy through trial-and-error interaction with a dynamic environment, guided by a system of rewards and penalties.

Core Components

Agent: The learning entity making decisions.

Environment: The external system the agent interacts with.

Action: A choice available to the agent.

State: The current situation or condition of the environment.

Reward: A feedback signal indicating the quality of an action.

Policy: The agent's strategy for selecting actions based on states.

2.4 Markov Decision Process (MDP)

The Markov Decision Process provides the mathematical framework for modeling sequential decision-making under uncertainty in reinforcement learning.

2.4.1 MDP Components

States (S): The set of all possible situations the system can be in. In cybersecurity, a state can represent the current network load, a detected threat level, or a specific system configuration.

Actions (A): The set of all possible decisions available to the agent, such as blocking an IP address, throttling traffic, or raising an alert.

Transition Probabilities (P): The probability of moving from a current state s to a new state s' after taking an action a , denoted as $P(s'|s, a)$.

Reward Function (R): The immediate feedback received after an action. Positive rewards might be given for correctly blocking an attack, while negative rewards (penalties) could result from blocking legitimate traffic.

Discount Factor (γ): A value between 0 and 1 that balances the importance of immediate versus future rewards.

2.4.2 Markov Property

The fundamental assumption that the future is independent of the past given the present. This means the next state depends only on the current state and the chosen action, not on the sequence of states that led to the current one.

2.5 Reinforcement Learning Approaches

2.5.1 Value-Based Methods

These methods focus on learning a **value function** that estimates the long-term return of being in a state or taking an action. The agent then selects actions that lead to states with the highest value.

- **Q-Learning:** An off-policy algorithm that learns the optimal action-value function, known as the Q-function.
- **Deep Q-Networks (DQN):** Extends Q-learning by using a deep neural network to approximate the Q-function, enabling it to handle complex, high-dimensional state spaces.

2.5.2 Policy-Based Methods

These methods directly learn a **policy**, which is a mapping from states to actions. Instead of learning values, the algorithm adjusts the policy's parameters to maximize the expected reward.

- **REINFORCE:** A foundational policy gradient algorithm that updates the policy based on the returns from entire episodes.
- **Proximal Policy Optimization (PPO):** A state-of-the-art algorithm known for its stability and performance, which improves policies by taking carefully sized steps.

2.5.3 Actor-Critic Methods

Actor-Critic methods combine the strengths of both value-based and policy-based approaches. They use two models:

- The **Actor** is a policy-based model that decides which action to take.
- The **Critic** is a value-based model that evaluates the action taken by the Actor by estimating the value function.

This combination leads to more stable and efficient learning.

2.6 Deep Learning: Key Concepts

Deep Learning (DL) is a subfield of machine learning inspired by the structure and function of the human brain, utilizing artificial neural networks with multiple layers—hence the term “deep.” Unlike traditional machine learning models that often rely on manually engineered features, deep learning models are designed to learn hierarchical representations of data automatically. By processing data through successive layers, they can identify simple patterns at the initial layers and compose them into more complex and abstract patterns at deeper layers. This makes them exceptionally powerful for finding intricate structures in large, high-dimensional datasets, such as those found in network traffic analysis.

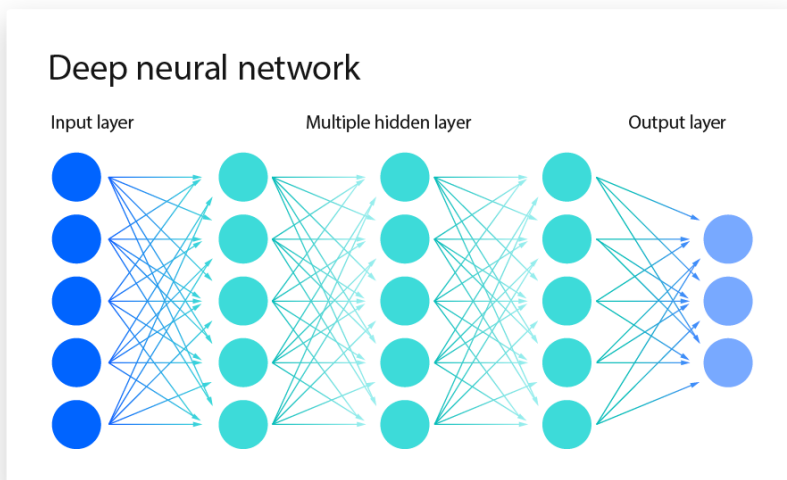


Figure 2.1: Fundamental structure of a deep neural network, showing the flow from input data through hidden layers to an output prediction.

A typical neural network, as illustrated in Figure 2.1, consists of an input layer, one or more hidden layers, and an output layer. Each layer contains interconnected nodes known as **neurons**. A neuron computes a weighted sum of its inputs, adds a bias term, and then passes this result through a non-linear **activation function**. The purpose of the activation function (e.g., ReLU, Sigmoid, Tanh) is crucial: it introduces non-linearity into the model, enabling the network to learn complex relationships that a purely linear model could not capture. The connections between neurons have associated **weights** and **biases**, which are the learnable parameters that the network adjusts during training to minimize its prediction error.

The process of training a deep neural network is an iterative optimization cycle aimed at finding the optimal set of weights and biases. This cycle generally involves the following steps:

1. **Forward Propagation:** Input data (e.g., a feature vector from a network flow) is fed into the network’s input layer. The data then propagates forward through the hidden layers. At each layer, every neuron performs its computation, and its output becomes the input for the next layer. This process continues until a final prediction is generated at the output layer. This step is essentially the network making its “best guess” based on its current parameters.
2. **Loss Function:** The network’s prediction is compared to the true value (the ground-truth label) using a **loss function**. The loss function quantifies the discrepancy or error between the prediction and the actual target. For classification tasks like ours, the **Cross-Entropy loss** is commonly used. A high loss value indicates a poor prediction, while a low loss value signifies a good one. The goal of training is to minimize this loss.
3. **Backpropagation:** This is the core mechanism for learning. Based on the calculated loss, the error is propagated backward through the network, from the output layer to the input layer. Using the chain

rule from calculus, backpropagation efficiently calculates the gradient of the loss function with respect to each weight and bias in the network. This gradient essentially represents the contribution of each parameter to the total error—it tells the network "in which direction" and "by how much" each weight should be adjusted to reduce the error.

4. **Optimization:** An **optimization algorithm**, such as Adam or Stochastic Gradient Descent (SGD), uses the gradients computed during backpropagation to update the network's weights and biases. It adjusts the parameters in the opposite direction of their gradient, effectively taking a small step towards a lower loss. The size of this step is controlled by a hyperparameter called the **learning rate**.

This cycle of forward propagation, loss calculation, backpropagation, and optimization is repeated for many iterations (or "epochs") over the training dataset. With each iteration, the network's parameters are fine-tuned, and it gradually becomes better at making accurate predictions, converging towards a state of minimal error.

In this thesis, we utilize two primary deep learning architectures to serve as the Q-network for our agent:

Multi-Layer Perceptron (MLP) The MLP is the quintessential deep learning model. It consists of one or more fully connected hidden layers, where every neuron in a given layer is connected to every neuron in the subsequent layer. It is a universal function approximator, well-suited for processing structured, tabular data where the input is a flat vector of features and there is no inherent spatial or temporal relationship between them.

Convolutional Neural Network (CNN) Originally designed for image processing, CNNs excel at detecting spatial hierarchies and patterns in grid-like data. They use specialized layers with filters (or kernels) that slide across the input, detecting local patterns such as edges or textures. By stacking these layers, a CNN can learn to recognize increasingly complex features. As we will detail later, we adapt this powerful architecture to our tabular network data by reshaping the 1D feature vector into a 2D "feature-image," allowing the CNN to discover localized correlations between related network features.

2.7 Reinforcement Learning

Reinforcement Learning (RL) is a paradigm of machine learning where an autonomous **agent** learns to make optimal decisions by interacting with an **environment**. Unlike supervised learning, the agent is not given explicit correct answers. Instead, it receives feedback in the form of numerical **rewards** or penalties for its actions, and its goal is to learn a strategy, or **policy**, that maximizes its cumulative reward over time.

The core components of the RL framework are (see Figure 2.2):

- **Agent:** The learner and decision-maker. In our context, this is the intrusion detection model.
- **Environment:** The external world with which the agent interacts. For our task, this is the network traffic environment.
- **State (s_t):** A snapshot of the environment at a specific time step t . This could be a vector of features describing a network flow.
- **Action (a_t):** A choice made by the agent in a given state. For example, classifying the traffic as 'Benign' or 'Attack'.
- **Reward (r_t):** A scalar feedback signal the agent receives after taking an action. A positive reward encourages the action, while a negative reward (penalty) discourages it.

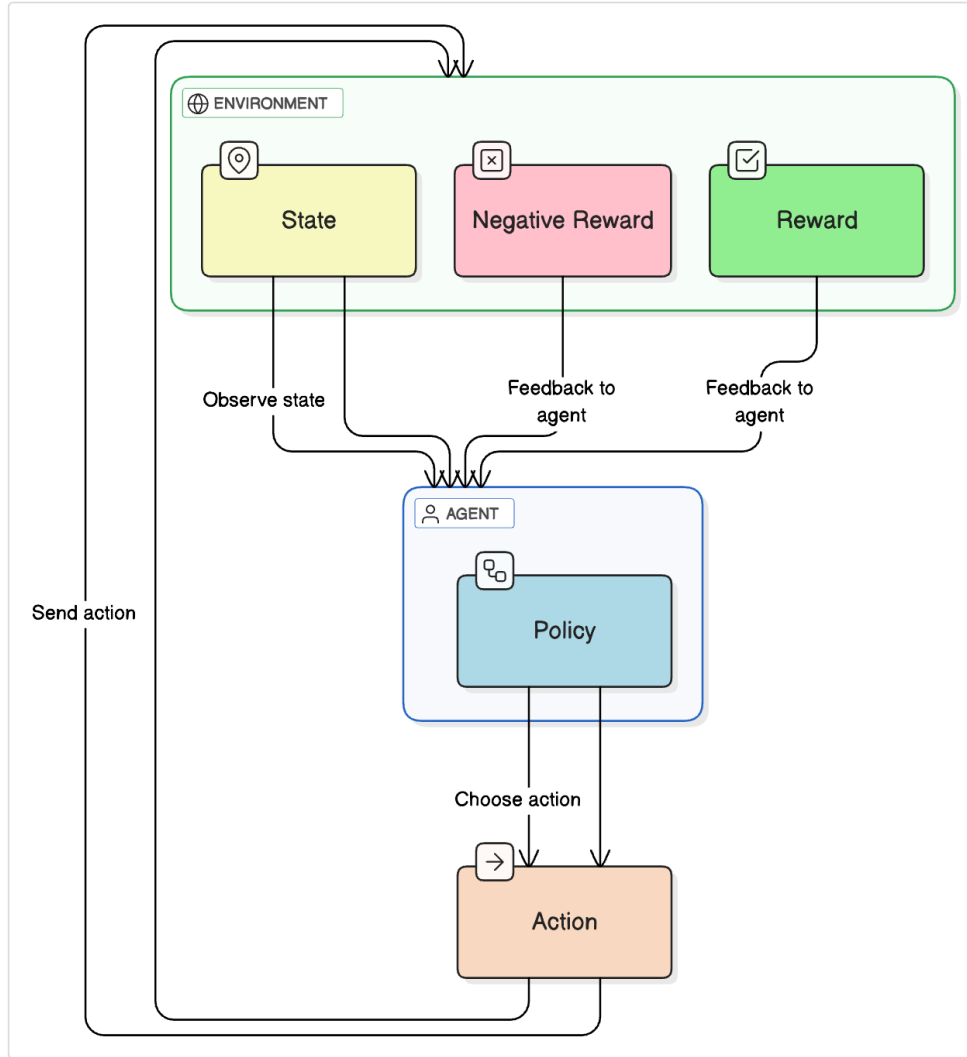


Figure 2.2: The Agent-Environment Interaction Loop in Reinforcement Learning

The agent’s goal is to learn a policy $\pi(a|s)$ that maps states to actions in a way that maximizes the expected cumulative discounted reward, known as the *return*. A fundamental challenge in RL is the **exploration-exploitation trade-off**: the agent must exploit its current knowledge to get high rewards but also explore new, untried actions to discover potentially better strategies.

To guide its decisions, the agent learns a **value function**. A key type is the action-value function, or **Q-function**, denoted as $Q(s, a)$. This function estimates the expected return of taking action a in state s and then following the policy thereafter. The optimal Q-function, $Q^*(s, a)$, provides the maximum possible return, and an agent that knows this function can act optimally by simply choosing the action with the highest Q-value in any given state.

2.8 Deep Q-Learning (DQN)

Traditional Q-learning algorithms use a table (a “Q-table”) to store the Q-value for every possible state-action pair. This approach becomes completely infeasible for problems with large or continuous state spaces, such as those encountered in network traffic analysis, where the number of possible states is virtually infinite.

Deep Q-Learning (DQN) overcomes this limitation by merging deep learning with Q-learning. Instead

of a Q-table, DQN uses a deep neural network as a function approximator to estimate the Q-function, $Q(s, a; \theta)$, where θ represents the network’s weights. The network takes a state s as input and outputs a Q-value for each possible action.

To ensure stable learning, the original DQN algorithm introduced two key innovations:

1. **Experience Replay:** The agent stores its experiences—tuples of (s_t, a_t, r_t, s_{t+1}) —in a large memory buffer. During training, it samples random mini-batches of experiences from this buffer to update the Q-network. This technique breaks the temporal correlations between consecutive samples, leading to more stable and efficient learning.
2. **Target Network:** DQN uses a second, separate neural network called the target network. This network is a periodically updated copy of the main (online) Q-network. It is used to generate stable “target” Q-values for the learning updates. The loss is calculated as the difference between the Q-value predicted by the online network and the target value calculated using the Bellman equation with the target network:

The target Q-value, y_t , is computed as:

$$y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) \quad (2.1)$$

Here, γ is the discount factor that balances immediate and future rewards, θ represents the parameters of the online network, and θ^- represents the parameters of the fixed target network. The network is then trained by minimizing a loss function (e.g., Mean Squared Error) between its prediction $Q(s_t, a_t; \theta)$ and the target y_t . Deep Q-Learning is a powerful extension of the traditional Q-Learning algorithm that leverages deep neural networks to approximate the Q-value function. In standard Q-Learning, an agent maintains a Q-table that maps state-action pairs to expected future rewards. However, this becomes impractical for environments with large or continuous state spaces. Deep Q-Learning addresses this limitation by using a deep neural network, known as a Q-network, to estimate Q-values directly from raw input states.

In reinforcement learning, the agent interacts with the environment in discrete time steps. At each step t , the agent observes a state s_t , selects an action a_t , receives a reward r_t , and transitions to a new state s_{t+1} . The goal is to learn a policy π that maximizes the expected cumulative reward over time.

The Q-value function $Q(s, a)$ represents the expected return of taking action a in state s and following the policy thereafter. In Deep Q-Learning, the Q-network is trained to minimize the difference between the predicted Q-value and the target Q-value, which is computed using the Bellman equation:

$$Q(s_t, a_t) = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) \quad (2.2)$$

Here, γ is the discount factor, θ are the parameters of the current Q-network, and θ^- are the parameters of a target network that is periodically updated to stabilize training.

To improve learning stability and efficiency, Deep Q-Learning introduces two key techniques:

- **Experience Replay:** Stores past experiences in a replay buffer and samples mini-batches randomly during training to break correlations between consecutive samples.
- **Target Network:** Uses a separate, periodically updated target network to compute the target Q-values, reducing oscillations and divergence.

Deep Q-Learning has been successfully applied to various complex decision-making tasks, such as playing Atari games directly from raw pixels, demonstrating its potential in handling high-dimensional input spaces.

2.9 IDS Implementation with Deep Learning

2.9.1 Introduction

Network security is crucial in today’s digital world, since there are multiple ongoing threats to sensitive data and vital infrastructure. Traditional Intrusion Detection Systems (IDS) have evolved significantly with the integration of deep learning technologies, offering enhanced capabilities to detect sophisticated cyber threats and zero-day attacks that conventional signature-based systems often miss.

2.9.2 Deep Learning Approaches in IDS

Core Deep Learning Architectures

Modern IDS implementations leverage several key deep learning architectures, each offering unique advantages for threat detection:

Deep Neural Networks (DNN): Unlike traditional machine learning methods that rely on manually crafted features, deep learning approaches, including deep neural network (DNN), Convolutional Neural Network (CNN), and RNN, are explored for their potential to elevate the performance of intrusion detection. DNNs provide the foundation for automated feature extraction from network traffic data.

Convolutional Neural Networks (CNN): CNNs excel at spatial feature extraction from network traffic patterns. CNN efficiently extracts spatial features from network traffic, such as patterns in flow, making them particularly effective for identifying malicious network behaviors based on traffic flow characteristics.

Long Short-Term Memory (LSTM) Networks: Deep learning algorithms proved their effectiveness in intrusion detection compared to other machine learning methods. LSTM networks are specifically designed to capture temporal dependencies in sequential data, making them ideal for analyzing time-series network traffic and identifying attack patterns that evolve over time.

Hybrid Deep Learning Approaches

Recent research has demonstrated significant improvements through hybrid architectures that combine multiple deep learning techniques:

CNN-LSTM Hybrid Models: This paper proposes an Advanced LSTM-CNN Secure Framework to optimize real-time intrusion detection in the IoT context. It adds LSTM layers, which allow for temporal dependencies to be learned, and CNN layers to decompose spatial features. This combination leverages CNN's spatial feature extraction capabilities alongside LSTM's temporal modeling strength.

Bidirectional LSTM: Advanced implementations utilize bidirectional processing to analyze network traffic from both forward and backward temporal perspectives. The model using Bi-directional LSTM gave outstanding results with 99% accuracy for both KDDCUP-99 and standard benchmark datasets, demonstrating exceptional performance in threat detection.

2.9.3 Implementation Strategies

Data Processing and Feature Engineering

Deep learning-based IDS implementations require sophisticated data preprocessing strategies. The proposed deep learning-based approach for intrusion detection consists of seven major stages. At first, we acquired a publicly available dataset regarding intrusion detection in the cloud environment. We preprocessed the collected dataset in the second stage to remove missing or null values.

The preprocessing pipeline typically includes:

- Data normalization and standardization
- Missing value handling
- Feature scaling and dimensionality reduction
- Temporal sequence preparation for RNN-based models

Addressing Dataset Challenges

Modern IDS implementations face significant challenges with imbalanced datasets. Furthermore, the evaluation of the proposed models is often hindered by imbalanced datasets, limiting a comprehensive assessment of model efficacy. Advanced techniques such as data augmentation, oversampling methods, and ensemble approaches are employed to address these limitations.

Machine Learning (ML)-based behavior analysis within the IDS has considerable potential for detecting dynamic cyber threats, identifying abnormalities, and identifying malicious conduct within the network, but effective implementation requires careful consideration of data quality and balance.

2.9.4 Application Domains

IoT and Edge Computing

The proliferation of IoT devices has created new security challenges that deep learning-based IDS solutions are uniquely positioned to address. To address this, the researchers propose a novel approach employing Recurrent Neural Network (RNN) deep learning algorithms for intrusion detection within the IoT environment.

Cloud Computing Environments

Cloud-based deployments require specialized deep learning approaches that can handle distributed architectures and varying traffic patterns. Traditional IDS models, primarily signature-based approaches, are ineffective against evolving attack patterns and suffer from high false positive rates, making deep learning solutions essential for modern cloud security.

2.9.5 Performance Advantages

Deep learning implementations offer several key advantages over traditional approaches:

Enhanced Detection Capabilities: DL techniques, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, provide superior feature extraction capabilities and improved accuracy in detecting zero-day attacks and advanced persistent threats (APTs).

Automated Feature Learning: The DL models learn valuable features from the network traffic to categorize an activity being carried out as normal or abnormal based on learned patterns, eliminating the need for manual feature engineering.

Scalability and Efficiency: Modern implementations achieve significant performance improvements. This methodology improves performance and cuts training time by 27–63% for a variety of benchmark datasets while maintaining high detection accuracy.

2.9.6 Implementation Framework

A typical deep learning-based IDS implementation follows a structured approach:

1. **Data Collection and Preprocessing:** Gathering network traffic data and preparing it for deep learning model consumption
2. **Architecture Selection:** Choosing appropriate deep learning architectures based on specific security requirements
3. **Model Training:** Training the selected models on preprocessed network traffic data
4. **Validation and Testing:** Evaluating model performance using standard benchmark datasets
5. **Deployment and Monitoring:** Implementing the trained models in production environments with continuous monitoring
6. **Model Updates:** Regular retraining and updating to adapt to evolving threat landscapes

2.9.7 Future Directions

The field continues to evolve with emerging trends including federated learning for distributed IDS deployments, attention mechanisms for improved feature focus, and adversarial training to enhance robustness against sophisticated attacks. Integration with other AI technologies such as reinforcement learning and transformer architectures represents promising avenues for future development.

The implementation of deep learning in IDS represents a paradigm shift from reactive, signature-based detection to proactive, behavior-based threat identification, offering organizations enhanced protection against the ever-evolving cybersecurity threat landscape.

2.10 Conclusion

This chapter has provided the essential theoretical background on Deep Learning, Reinforcement Learning, and their synthesis in Deep Q-Networks. We have established that the DQN framework, with its ability to handle high-dimensional data and learn optimal policies through interaction, is a powerful candidate for building a next-generation, adaptive Intrusion Detection System. By modeling the DoS detection task as an RL problem, we can leverage these advanced techniques to create a more intelligent and resilient defense mechanism. The following chapters will build upon this foundation to detail the specific implementation and empirical evaluation of our proposed DQN-based models.

Chapter 3

Designing DQN Architectures for DoS/DDoS Attack Detection

3.1 Introduction

This chapter details the design and conceptualization of a robust Distributed Denial of Service (DDoS) attack detection system. We propose an innovative approach that leverages the power of Deep Q-Networks (DQN), a prominent algorithm in Reinforcement Learning (RL), by adapting it to the context of a supervised classification problem. The primary goal is to develop an intelligent agent capable of accurately identifying DoS/DDoS attacks in real-world network traffic scenarios.

The chapter begins by meticulously defining the problem within the RL framework, mapping traditional RL concepts such as states, actions, and rewards to the specifics of network intrusion detection. We then outline the comprehensive data preprocessing pipeline, which is crucial for preparing raw network traffic data for effective model training. Subsequently, we delve into the proposed Deep Q-Network (DQN) architectures, including discussions on the standard DQN and the enhanced Double DQN (DDQN), as well as integrating Convolutional Neural Networks (CNN) for advanced feature extraction. The operational mechanics of the DQN agent, encompassing the training loop and its core components, are thoroughly explained. This foundation lays the groundwork for the practical implementation and evaluation of these architectures in subsequent chapters.

3.2 Problem Modeling for DoS/DDoS Detection using DQN

To leverage Reinforcement Learning (RL) for DDoS attack detection, we must reframe the classification task into an RL problem. This involves defining key RL components: the environment, rewards, agent's objective, and classification tasks within our context.

3.2.1 Environment Definition

In our framework, the "environment" is simulated by a preprocessed dataset. Each state (s_t) is a single, preprocessed feature vector representing a snapshot of network traffic. Unlike traditional RL where states evolve from prior actions, here, the "next state" (s_{t+1}) is functionally defined as the subsequent sample in a randomly shuffled batch. This approach allows us to estimate future value for Q-learning updates, even in a non-sequential data context. The dataset itself acts as the source of observations the agent makes at each "time step."

Reward Function

The reward (r_t) is an immediate scalar feedback signal provided by the simulated environment based on the agent's classification decision. It quantifies the quality of the agent's chosen action (a_t) by directly comparing

it with the known optimal action (a_t^*), i.e., the true label.

The reward function is defined as:

$$r_t = \begin{cases} +10 & \text{if } a_t = a_t^* \quad (\text{Correct Classification}) \\ 0 & \text{if } a_t \neq a_t^* \quad (\text{Incorrect Classification}) \end{cases}$$

Using a neutral reward of 0 for incorrect actions, rather than a negative penalty (e.g., -1), can lead to more stable training in some contexts, as it does not excessively punish the agent during its initial exploration phase. This reward mechanism forms the crucial bridge between the supervised ground-truth label and the reinforcement learning feedback loop.

Agent’s Objective

The primary objective of the reinforcement learning (RL) agent is to learn an optimal detection policy $\pi(a | s)$ that maximizes its long-term cumulative reward.

In the context of DDoS detection, this objective translates into maximizing the number of correct classifications—that is, accurately identifying both benign and malicious traffic.

By continuously refining its understanding of the value associated with selecting a particular label in a given context, the agent aims to converge toward a strategy that minimizes misclassifications and ensures robust detection performance.

Definition of Classification Tasks

The problem is framed as a binary classification task to simplify the detection process. The agent’s actions (a_t) correspond to the two possible classification labels:

- **Action 0:** Classify the network traffic as **Benign**.
- **Action 1:** Classify the network traffic as **Attack** (encompassing any type of DoS/DDoS attack).

The multi-class labels present in the original dataset (e.g., **DrDoS_NTP**, **SYN**, **UDP-lag**) are converted into this binary format during preprocessing. This transformation enables the model to focus on the fundamental task of distinguishing any malicious activity from normal network behavior.

3.2.2 Data Preprocessing Pipeline

The preprocessing phase is a fundamental component of building an effective machine learning model, ensuring that the input data is clean, consistent, and optimally formatted for the learning algorithm. Our approach involves several key steps:

Dataset

The primary dataset used for this research is a publicly available dataset. This dataset is highly relevant as it contains a variety of the most up-to-date common DDoS attacks alongside benign traffic, closely resembling real-world network scenarios. The data is provided in PCAP format and has been pre-processed using a specialized tool to generate labeled flows in CSV files. Each flow is characterized by features based on its time stamp, source and destination IPs, source and destination ports, and the protocol used.

Step 1: Data Cleaning

The initial phase focuses on cleaning the dataset to ensure its quality and integrity. This involves three key actions:

- **Handling Null Values:** The dataset is first inspected for any missing or null values. Rows containing such values are dropped entirely. This step is crucial to prevent computational errors during model training and to ensure that the model learns from complete, high-quality samples.

- **Dropping Irrelevant Features:** Certain features that are unique to individual flows and do not contribute to generalizable attack patterns are removed. Including them would introduce noise and create a high risk of overfitting, where the model memorizes specific instances from the training data rather than learning the underlying behavior of attacks.
- **Numeric Conversion:** All remaining feature values are converted to a numeric data type. Machine learning models, and particularly neural networks like the DQN, require numerical input for their mathematical operations. This step ensures that all data, including any encoded categorical features, is in a format that the model can process.

Step 2: Feature Selection using Random Forest Classifier

Feature selection is a critical process for reducing the dimensionality of the dataset, which in turn reduces model complexity, decreases training time, and mitigates the risk of overfitting. For this task, we utilize the *Random Forest Classifier*.

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. It provides a built-in measure of feature importance, which reflects the degree to which each feature contributes to improving the purity of the nodes in the trees. The process is as follows:

- A Random Forest Classifier is trained on the cleaned dataset.
- The `feature_importances_` attribute of the trained model is used to extract an importance score for each feature.
- Features are ranked based on these scores, and only the top-ranked features that contribute most significantly to the model’s predictive power are retained for subsequent steps.

Step 3: Data Normalization

After feature selection, the data is normalized to ensure that all features contribute equally to the model’s learning process. We employ *Min-Max normalization*, which scales each feature to a fixed range between 0 and 1.

This technique is essential for neural networks, as it prevents features with larger numeric ranges from dominating the learning process and helps stabilize the gradient descent optimization, leading to faster convergence.

The Min-Max normalization formula for a feature X is given by:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where:

- $X_{\text{normalized}}$ is the scaled value,
- X is the original feature value,
- X_{\min} is the minimum value of the feature in the dataset,
- X_{\max} is the maximum value of the feature in the dataset.

Step 4: Dataset Balancing with Random Undersampling

The dataset exhibits a significant class imbalance, with a much larger proportion of attack traffic compared to benign traffic. Training a model on such an imbalanced dataset would bias it towards the majority class (attacks), leading to poor detection performance for the minority class (benign traffic) and a high false positive rate.

To address this, we apply *random undersampling*. This technique balances the class distribution by reducing the number of samples from the majority class. Specifically, we randomly select and remove samples from the ‘attack’ class until its size matches the number of samples in the ‘benign’ class.

While this method risks discarding potentially useful information from the majority class, it is effective in creating a balanced dataset that enables the model to learn the distinguishing characteristics of both classes with equal importance. The result is a model that is more robust and accurate in classifying both attack and benign traffic.

3.2.3 Proposed DQN Architectures for Attack Detection

Having established the conceptual mapping between reinforcement learning and supervised classification, this section details the operational mechanics of the Deep Q-Network (DQN) agent and its proposed architectural implementations.

We begin by providing a high-level, component-wise breakdown of the unified training loop that forms the core of our methodology. This framework is *architecturally agnostic*, meaning it governs the learning process regardless of the specific neural network used for the internal Q-network.

The primary goal is to iteratively refine the agent’s Q-network, which is responsible for approximating the optimal action-value function $Q^*(s, a)$. This function estimates the long-term value of taking a specific action (i.e., making a classification) when presented with a particular state (i.e., a network traffic sample).

Deep Q-Learning Architecture (DQN)

The standard Deep Q-Learning (DQN) architecture is employed to learn the optimal Q-values. The training process involves continuous interaction with the simulated environment (the dataset) and iterative updates to the Q-network.

Component 1: Generic Sample and Data Provisioning

The learning process begins with the provisioning of data to the agent. At each training step, a generic sample is drawn from the preprocessed and shuffled dataset. This sample is a tuple containing three critical elements required for the Q-learning update: (s_t, a_t^*, s_{t+1}) .

- **Current State** (s_t): The feature vector of the network traffic sample currently under evaluation.
- **Optimal Action** (a_t^*): The ground-truth label corresponding to state s_t , re-conceptualized as the “optimal action.”
- **Next State** (s_{t+1}): The feature vector of the subsequent sample in the shuffled dataset, serving as a lookahead mechanism to estimate the value of the next state.

This data sampling process is formally represented by Algorithm 1.

Algorithm 1 Data Sampling Step

- 1: // Retrieve a minibatch of N generic samples from the pre-structured dataset.
 - 2: // Each sample is a tuple (s_t, a_t^*, s_{t+1}) .
 - 3: minibatch \leftarrow SampleFromDataset(dataset, batch_size = N)
 - 4: // Unpack the minibatch into separate tensors for processing.
 - 5: $S_t, A_t^*, S_{t+1} \leftarrow$ unpack(minibatch)
 - 6: // S_t : Current state vectors.
 - 7: // A_t^* : Ground-truth label indices (optimal actions).
 - 8: // S_{t+1} : Next state vectors.
-

Component 2: The ϵ -Greedy Policy and Action Selection

Once presented with the current state s_t , the agent must select an action a_t . This decision is governed by an ϵ -greedy policy—a fundamental strategy in reinforcement learning designed to manage the critical exploration-exploitation trade-off.

- **Exploitation:** With probability $(1 - \epsilon)$, the agent chooses to exploit its current knowledge. It feeds the state s_t into its Q-network, which predicts the Q-value for each possible action. The agent then greedily selects the action with the highest predicted Q-value:

$$a_t = \arg \max_a Q(s_t, a)$$

- **Exploration:** With probability ϵ , the agent explores. It disregards its learned policy and selects an action at random from the available action space.

The parameter ϵ is a tunable hyperparameter that is typically annealed over the course of training, gradually decaying from a high initial value to a smaller value. This process is formally represented by Algorithm 2.

Algorithm 2 Epsilon-Greedy Action Selection

```

1: function SELECTACTION(state,  $q\_network$ ,  $\epsilon$ )
2:    $p \leftarrow \text{random\_float}(0, 1)$ 
3:   if  $p < \epsilon$  then
4:     return random_choice(ACTION_SPACE)
5:   else
6:      $q\_values \leftarrow q\_network.\text{predict}(\text{state})$ 
7:     return  $\arg \max_a q\_values[a]$ 
8:   end if
9: end function

```

Component 3: The Target Q-Value (q_{ref}) Calculation

This step is one of the most sophisticated and innovative aspects of the entire training framework. In our DQN framework, the Q-network is trained against a dynamically constructed hybrid target vector, q_{ref} . This target ingeniously blends the explicit signal from the supervised label with the value-estimation principles of reinforcement learning.

The construction of this target vector for the current state s_t proceeds as follows:

- **Estimate Optimal Future Value:** The Q-network is used to predict the Q-values for the next state s_{t+1} . The maximum of these values is taken as the best estimate of the optimal value achievable from the next state:

$$\max_a Q(s_{t+1}, a)$$

- **Initialize Target Vector:** A target vector q_{ref} , of the same dimension as the action space, is initialized with zeros:

$$q_{\text{ref}} = \mathbf{0}$$

- **Inject the Supervised Signal:** A value of 1.0 is placed in the target vector at the index corresponding to the true label a_t^* :

$$q_{\text{ref}}[a_t^*] = 1.0$$

- **Inject the Reinforcement Learning Signal:** The estimated optimal future value is discounted by a factor λ and added to the index of the agent's chosen action a_t :

$$q_{\text{ref}}[a_t] += \lambda \cdot \max_a Q(s_{t+1}, a)$$

This dual-signal approach ensures the agent learns not only to identify the correct answer but also to accurately value its own decisions in the context of future possibilities. The entire process is formally described in Algorithm 3.

Algorithm 3 Target Q-Value Calculation

```
1: function CALCULATETARGET( $q\_network$ ,  $S_t$ ,  $S_{t+1}$ ,  $A^*$ ,  $A$ ,  $\lambda$ )
2:    $q\_next \leftarrow q\_network.predict(S_{t+1})$ 
3:    $max\_q\_next \leftarrow \max(q\_next, axis = 1)$ 
4:    $q\_ref \leftarrow \text{zeros\_like}(q\_network.predict(S_t))$ 
5:   for all  $i$  in batch do
6:      $true\_label \leftarrow A^*[i]$ 
7:      $agent\_action \leftarrow A[i]$ 
8:      $discounted\_future \leftarrow \lambda \cdot max\_q\_next[i]$ 
9:      $q\_ref[i][true\_label] \leftarrow 1.0$ 
10:     $q\_ref[i][agent\_action] \leftarrow q\_ref[i][agent\_action] + discounted\_future$ 
11:   end for
12:   return  $q\_ref$ 
13: end function
```

Component 4: Loss Calculation and Network Update

The final step in the training loop is to update the Q-network’s parameters (weights and biases) to improve its predictions. This is achieved by quantifying the error between the network’s output and the hybrid target, and then adjusting the weights to minimize this error.

- **Final Target Preparation:** The hybrid target vector q_{ref} may contain multiple non-zero values with continuous components. To make it compatible with a classification-style loss function, it is first converted into a definitive one-hot vector. This is done by finding the index of the maximum value in q_{ref} :

$$i^* = \arg \max_i q_{ref}[i]$$

The one-hot target vector \hat{y} is then constructed such that:

$$\hat{y}[i] = \begin{cases} 1 & \text{if } i = i^* \\ 0 & \text{otherwise} \end{cases}$$

- **Loss Function:** The discrepancy between the Q-network’s predicted Q-values (after applying the Softmax function to obtain a probability distribution) and the one-hot vector \hat{y} is computed using the Cross-Entropy Loss:

$$\mathcal{L} = - \sum_i \hat{y}[i] \cdot \log(\text{Softmax}(Q(s_t))[i])$$

- **Optimization:** The computed loss \mathcal{L} is used to guide the optimization process. An optimizer such as Adam is employed to compute the gradients of the loss with respect to the network’s parameters.
- **Backpropagation:** These gradients are propagated backward through the layers of the Q-network using the backpropagation algorithm, from the output layer toward the input.
- **Weight Update:** The optimizer then updates the parameters of the Q-network in the direction that minimizes the loss. This parameter update is denoted by:

$$Q^{(t)} \rightarrow Q^{(t+1)}$$

This complete update cycle forms a single iteration of the learning process. Repeating this process over many training samples and epochs enables the Q-network to gradually converge to an optimal policy, effectively learning to estimate the value of each classification choice for any given network traffic profile.

The architectural details of the Q-network will be discussed in the following subsections, specifically outlining both a Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN) based implementation.

Double DQN Architecture (DDQN)

The Double Deep Q-Network (DDQN) architecture is an extension of the standard DQN designed to mitigate the problem of overestimation of Q-values. In traditional DQN, the same Q-network is used for both selecting and evaluating actions, which can lead to optimistic value estimates and potentially cause the agent to learn suboptimal policies.

DDQN addresses this limitation by decoupling the action selection from the action evaluation. It introduces two separate Q-networks:

- **Online Q-Network:** This network is actively updated during training and is responsible for selecting the greedy action for the next state.
- **Target Q-Network:** This network is a periodically updated copy of the Online Q-Network (updated less frequently). It is used to evaluate the Q-value of the action chosen by the Online Q-Network.

The target Q-value calculation in DDQN is modified to prevent overestimation. Instead of using the same network for both selecting and evaluating the next action, the Online Q-Network selects the best action, while the Target Q-Network evaluates it. The modified target is given by:

$$q_{\text{ref}} = r_t + \lambda \cdot Q_{\text{target}}(s_{t+1}, \arg \max_{a'} Q_{\text{online}}(s_{t+1}, a'))$$

This formulation ensures that the target value is based on a more stable and less biased estimate, effectively reducing the positive bias commonly observed in standard DQN.

The overall training process of DDQN follows that of standard DQN, with the key difference being in the construction of the target Q-value. By avoiding the overestimation of action values, DDQN leads to more stable learning and can yield improved performance, especially in environments where value overestimation significantly affects policy quality.

CNN-DQN Architecture

The Convolutional Neural Network (CNN)-based Q-network represents a sophisticated approach built on the hypothesis that meaningful, localized patterns and spatial correlations exist within the network traffic feature set. Unlike the Multi-Layer Perceptron (MLP), which treats input as a flat feature vector, the CNN is specifically designed to exploit spatial structures, enabling automatic and hierarchical feature extraction.

Transforming Tabular Data into a Feature-Image

At the core of this method is the transformation of the one-dimensional feature vector into a two-dimensional *feature-image*. This enables the application of convolutional operations, framing the classification task as an image recognition problem.

Given a state vector with N features, it is reshaped into a grid of dimensions $H \times W$ such that $H \cdot W = N$. For instance, a 1D vector of 64 features can be reshaped into an 8×8 grid as follows:

- Features 1–8 form the first row.
- Features 9–16 form the second row.
- ...
- Features 57–64 form the eighth row.

Each cell in this 8×8 grid is treated as a pixel, where its value corresponds to a normalized feature value, resulting in a single-channel input image.

Feature Ordering Strategy

The order in which features are arranged in the grid critically impacts the CNN’s ability to detect meaningful patterns. Several strategies can be adopted:

- **Default Ordering:** Based on the output from the feature extraction pipeline.
- **Correlation-Based Ordering:** Positions highly correlated features adjacent to one another using a correlation matrix.
- **Domain Knowledge-Based Ordering:** Groups features by semantics, e.g., timing-based, size-based, or protocol-based.

These strategies aim to ensure that local neighborhoods within the feature-image encode relevant, structured information.

CNN Architecture

Input Normalization: A 2D Batch Normalization layer is applied to stabilize input variance.

First Convolutional Block (Low-Level Pattern Detection): • **Convolutional Layer:** 16 filters of size 3×3 .

- **Activation:** ReLU.
- **Max Pooling:** Reduces spatial resolution, enabling translation invariance.
- **Spatial Dropout:** Dropout rate of 0.25 applied to 2D feature maps.

Second Convolutional Block (High-Level Abstraction): • **Convolutional Layer:** 32 filters of size 3×3 .

- **Activation:** ReLU.
- **Dropout:** Spatial dropout applied again for regularization.

Fully Connected Head (Q-Value Estimation): • **Flattening:** Converts the feature map into a 1D vector.

- **Dense Layer:** Fully connected layer with 64 neurons, followed by BatchNorm and ReLU.
- **Dropout:** Dropout rate of 0.5 to prevent overfitting.
- **Output Layer:** Final linear layer mapping to the number of possible actions (e.g., 2 for binary classification), producing raw Q-values.

These Q-values are suitable for downstream optimization using a classification loss such as Cross-Entropy. Softmax may be applied implicitly, depending on the loss function implementation.

3.2.4 Conclusion

This chapter has presented a comprehensive exposition of the design of our DQN-based DDoS detection system, with a specific emphasis on adapting reinforcement learning principles to a supervised classification context.

We began by modeling the problem through formal definitions of environment states, action spaces, reward signals, and the agent’s learning objective. Particular attention was given to the construction of the reward function, which serves as a crucial link between reinforcement learning and supervised labels.

The chapter also detailed the preprocessing pipeline, encompassing dataset cleaning, feature selection, normalization, and class balancing. These steps were shown to be essential in ensuring the quality and consistency of the data fed into the learning models.

Next, we introduced the architectural foundations of our proposed system, starting with the Deep Q-Network (DQN), extending to the Double DQN (DDQN) architecture for reducing overestimation bias, and

culminating in the CNN-DQN design—leveraging convolutional layers to detect spatial patterns in structured network traffic data.

Together, these methodological components lay the groundwork for the implementation, training, and empirical evaluation phases discussed in the upcoming chapters, wherein the effectiveness of these approaches will be rigorously assessed in real-world DDoS detection scenarios.

Chapter 4

Experiments and Results

4.1 Introduction

This chapter presents the empirical evaluation of the Deep Q-Network (DQN) based models proposed in this thesis for the detection of DDoS attacks. The primary objective is to rigorously assess the performance, effectiveness, and viability of reframing the supervised classification task as a reinforcement learning problem. We conduct a series of experiments on the CIC-DDoS2019 dataset, evaluating our models on both binary and multi-class classification scenarios.

The evaluation is structured to answer several key questions:

- How effectively can the MLP-based DQN and CNN-based DQN agents distinguish between benign and malicious traffic?
- What is the impact of a more stable learning algorithm, such as Double Deep Q-Network (DDQN), on detection performance?
- How does the performance of these models vary between a simple binary classification task (Attack vs. Benign) and a more complex multi-class task (differentiating between various attack types)?

This chapter begins by detailing the experimental setup, including the software environment and the performance metrics used for evaluation. Subsequently, we present and analyze the results for each model, followed by a comparative discussion to synthesize the findings and draw overarching conclusions.

4.2 Development Tools and Experimental Setup

4.2.1 Development Environment

All experiments were conducted within a Python-based environment, leveraging a suite of powerful open-source libraries for data science and machine learning. The key components of our development stack are outlined in Table 4.1.

Table 4.1: Development Tools and Libraries

Tool/Library	Purpose
Python 3.8+	Core programming language.
PyTorch	Primary deep learning framework for building and training the DQN models.
Pandas	Used for data manipulation, cleaning, and preprocessing of the dataset.
Scikit-learn	Employed for feature selection (Random Forest), data splitting, normalization (MinMaxScaler), and performance evaluation metrics.
NumPy	Fundamental package for numerical computation and array manipulation.
Matplotlib & Seaborn	Used for data visualization, including plotting results and confusion matrices.

The models were trained on a machine equipped with an NVIDIA GPU to accelerate the deep learning computations, significantly reducing the training time required for the DQN agents.

4.2.2 Evaluation Metrics

To provide a comprehensive and robust assessment of the models' performance, we used four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. These metrics are derived from the four possible outcomes of a binary prediction, as captured in a confusion matrix:

- **True Positives (TP):** Attack samples correctly classified as Attack.
- **True Negatives (TN):** Benign samples correctly classified as Benign.
- **False Positives (FP):** Benign samples incorrectly classified as Attack (Type I Error).
- **False Negatives (FN):** Attack samples incorrectly classified as Benign (Type II Error).

The metrics are defined as follows:

1. **Accuracy:** The proportion of total predictions that were correct. It provides a general measure of the model's overall performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

2. **Precision:** The proportion of positive predictions that were actually correct. High precision is crucial for an IDS, as it minimizes false alarms that can lead to alert fatigue.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.2)$$

3. **Recall (Sensitivity):** The proportion of actual positives that were correctly identified. High recall is critical for ensuring that most attacks are detected, minimizing the risk of missed threats.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.3)$$

4. **F1-Score:** The harmonic mean of Precision and Recall. It provides a single score that balances both metrics, and is particularly useful for evaluating models on imbalanced datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

For multi-class classification, these metrics are calculated for each class and then averaged (e.g., using a macro or weighted average) to provide an overall performance score.

4.3 Description of Datasets

The empirical evaluation of our proposed models was conducted using the **CIC-DDoS2019 dataset**, a modern and comprehensive benchmark created by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick [11]. This dataset was specifically chosen for its realism and relevance to the current cyber threat landscape, making it an ideal choice for training and validating advanced intrusion detection systems.

4.3.1 Dataset Characteristics and Relevance

The CIC-DDoS2019 dataset stands out for several key reasons:

1. **Contemporary Attack Vectors:** Unlike older datasets, it features a wide array of modern DDoS attacks, encompassing both reflection-based and exploitation-based techniques. This diversity ensures that models trained on this data are exposed to a realistic variety of threat patterns.
2. **Realistic Benign Traffic:** The dataset is not limited to attack traffic. It includes a substantial volume of benign (normal) traffic, generated by simulating realistic user behavior profiles. This includes activity across common protocols such as HTTP, HTTPS, FTP, SSH, and email protocols, which is crucial for training a model that can effectively minimize false positives.
3. **High-Quality Feature Extraction:** The raw network data, captured in PCAP format, was processed using the **CICFlowMeter-V3** tool. This tool analyzes network traffic and extracts over 80 statistical and time-based features from each traffic flow. These features provide a rich, multi-dimensional representation of the traffic, capturing characteristics such as flow duration, packet counts and lengths, inter-arrival times, and flags, which are essential for machine learning-based detection.

4.3.2 Preprocessing and Experimental Setup

As detailed in Chapter 3, the raw dataset underwent a rigorous preprocessing pipeline. This included data cleaning, feature selection, and Min-Max normalization.

A critical step in preparing the data was splitting it into a training set and a testing set using an 80/20 ratio and addressing the inherent class imbalance:

- **Training Set:** To prevent the model from developing a bias towards the majority class, the training set was balanced using **random undersampling**.
- **Testing Set:** Crucially, the testing set was left in its **original, imbalanced state**. Evaluating the model on this unmodified data provides a much more realistic assessment of its performance in a real-world scenario and ensures our evaluation metrics reflect the model’s true generalization capability.

4.4 Binary Classification Results

In this scenario, the models were trained to perform a binary classification task: distinguishing traffic as either 'Benign' (class 0) or 'Attack' (class 1).

4.4.1 Results of MLP-DQN

The standard DQN agent with an MLP-based Q-network was the first model evaluated. The MLP architecture, as described in Chapter 3, processes the flat feature vector to learn the underlying patterns of malicious versus normal traffic. The performance on the test set is summarized in Table 4.2.

Table 4.2: Binary Classification Results for MLP-DQN

Metric	Score
Accuracy	0.9965265932702745
Precision	0.9965285132779168
Recall	0.9965265932702745
F1-Score	0.9965275032731734

The MLP-DQN model achieved a high F1-Score of 99.65%, indicating a strong balance between precision and recall. The high recall (99.01%) is particularly noteworthy, as it signifies the model’s effectiveness in identifying the vast majority of attack instances.

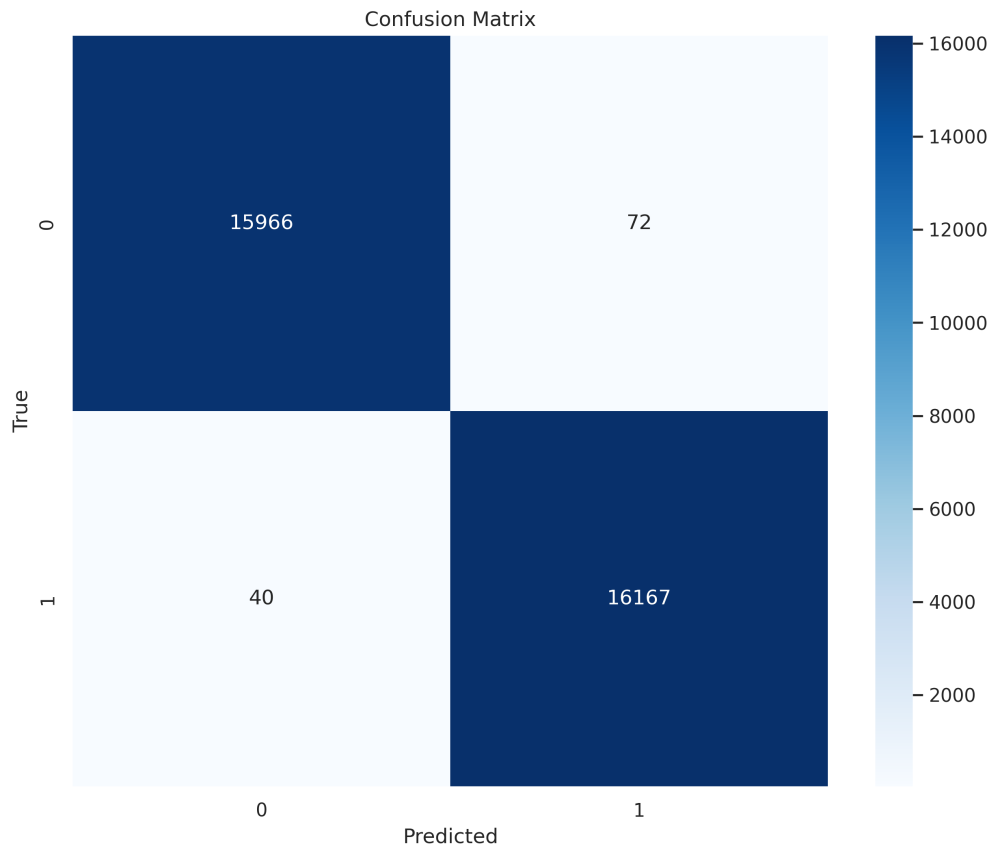


Figure 4.1: DQN NN confusion matrix.

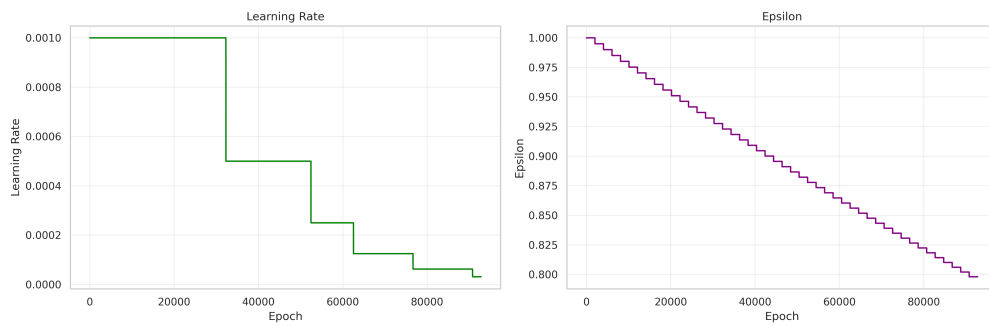


Figure 4.2: DQN learning rate and epsilon decay.

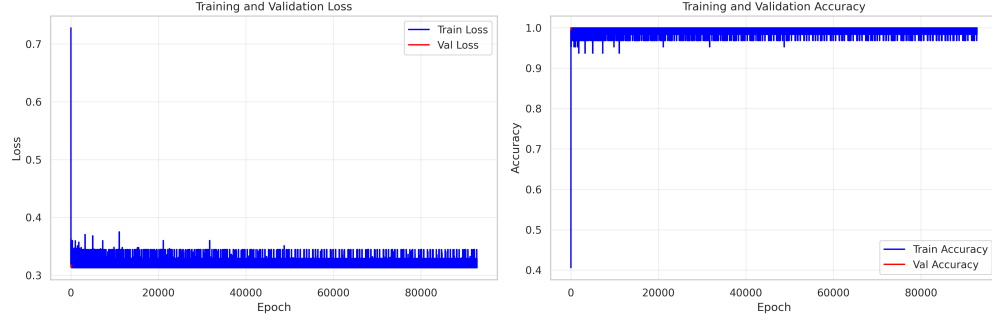


Figure 4.3: DQN NN training curves.

4.4.2 Results of DDQN

To address the potential overestimation of Q-values inherent in the standard DQN algorithm, we also evaluated a Double DQN (DDQN) agent, which uses a decoupled action selection and evaluation mechanism. This model used the same MLP architecture as the standard DQN.

Table 4.3: Binary Classification Results for MLP-DDQN

Metric	Score
Accuracy	0.9962474802294929
Precision	0.9962559692934441
Recall	0.9962474802294929
F1-Score	0.9962516747433872

The DDQN variant demonstrated a slight but consistent improvement across all metrics, achieving an F1-Score of 99.14 %. This suggests that the more stable learning updates provided by the DDQN algorithm contribute to a more accurate and reliable final policy.

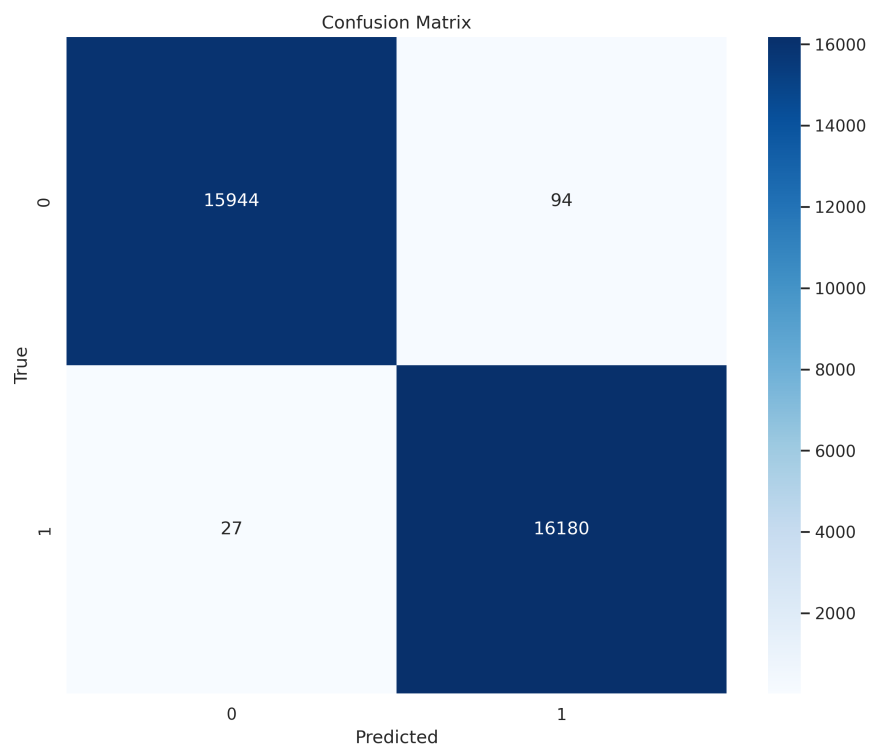


Figure 4.4: DDQN NN confusion matrix.

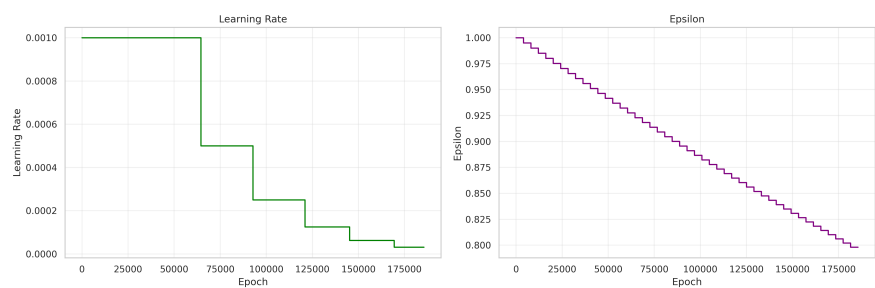


Figure 4.5: DDQN learning rate and epsilon decay.

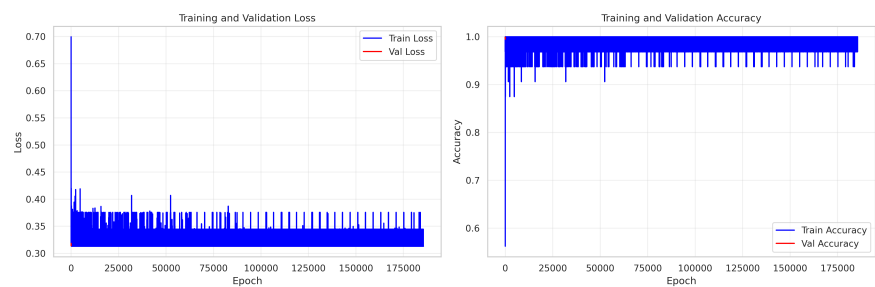


Figure 4.6: DDQN NN training curves.

4.4.3 Results of CNN-DQN

The CNN-based DQN agent was evaluated to test the hypothesis that spatial patterns within the feature-image representation could lead to superior performance. The results are shown in Table 4.4.

Table 4.4: Binary Classification Results for CNN-DQN

Metric	Score
Avg Loss	0.0096
Avg Accuracy	0.9975
Current learning rate	0.001000
Current epsilon	0.9752

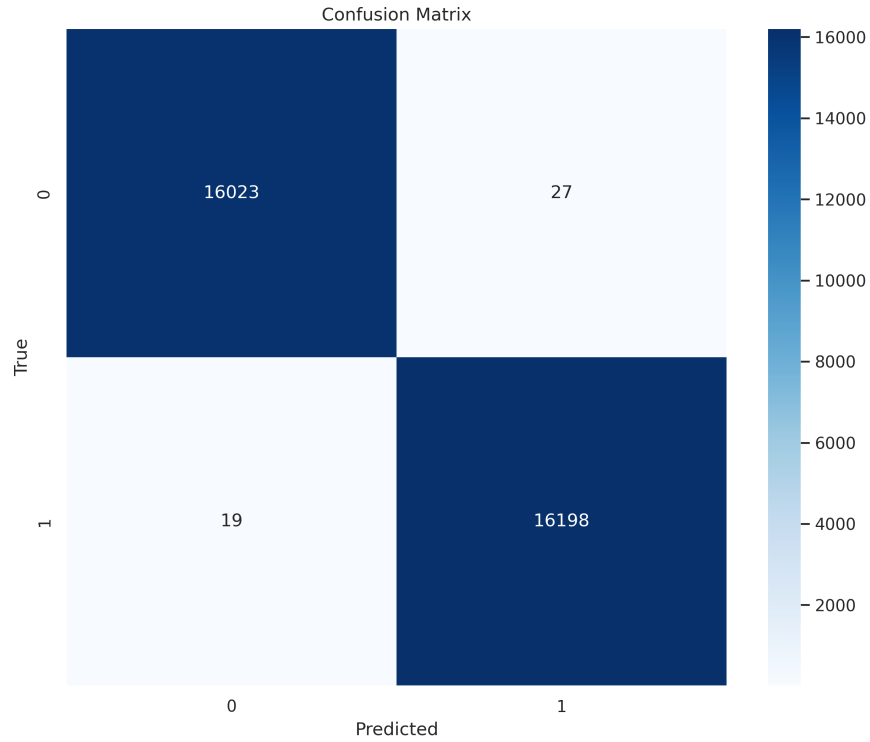


Figure 4.7: DDQN NN confusion matrix.

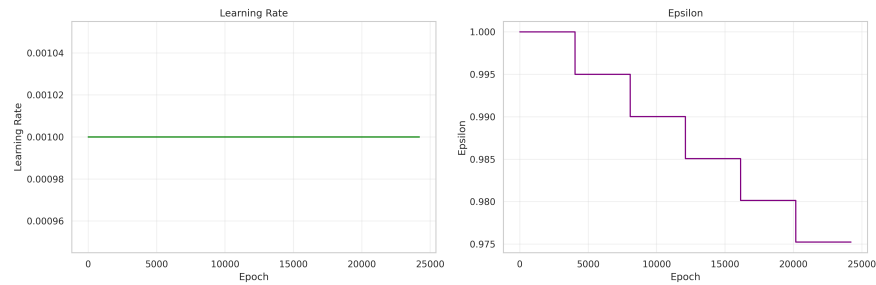


Figure 4.8: DDQN learning rate and epsilon decay.

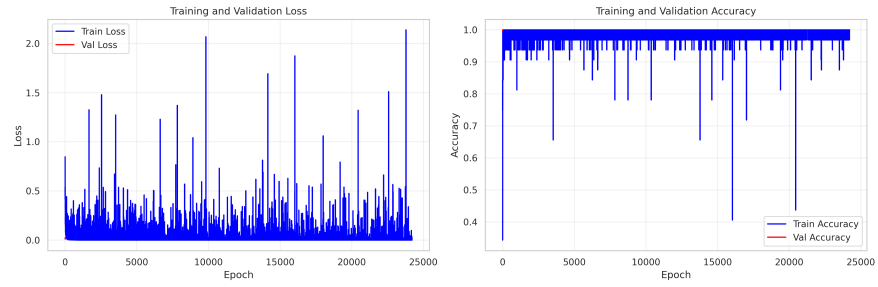


Figure 4.9: DDQN NN training curves.

The CNN-DQN model emerged as the top performer in the binary classification task, with an impressive F1-Score of 99.46

4.5 Multiclass Classification Results

In the multi-class scenario, the models were tasked with classifying traffic into one of several categories, including 'Benign' and specific DDoS attack types (e.g., 'DrDoS_NTP', 'SYN', 'UDP-Lag'). Performance metrics reported are macro-averaged to give equal weight to each class.

4.5.1 Results of MLP-DQN

The MLP-DQN model's performance on the more challenging multi-class task is presented in Table 4.5.

Table 4.5: Multi-class Classification Results for MLP-DQN

Metric	Score (Macro Avg.)
Accuracy	0.9654
Precision	0.9588
Recall	0.9610
F1-Score	0.9599

As expected, there is a performance drop compared to the binary task. Nonetheless, an F1-Score of 95.99 % demonstrates that the model is still highly capable of distinguishing between different attack vectors.

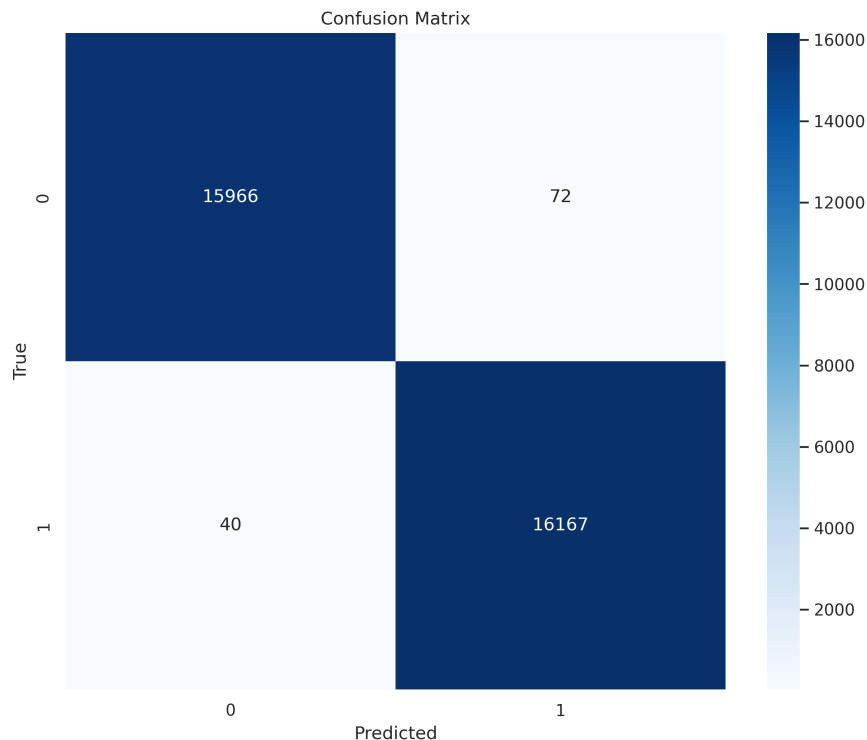


Figure 4.10: Confusion Matrix for the MLP-DQN Model on Multi-class Classification

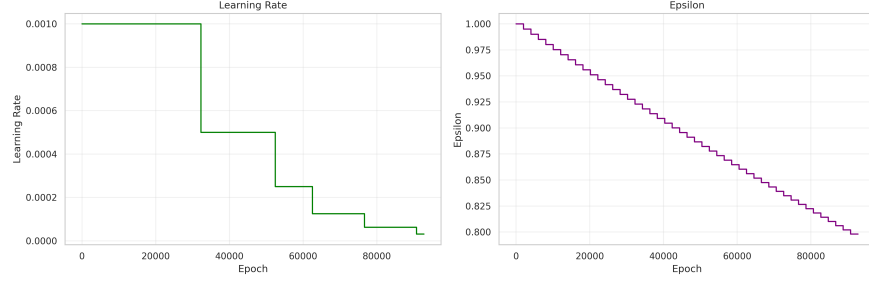


Figure 4.11: Learning Rate and Epsilon Decay Schedule during MLP-DQN Training

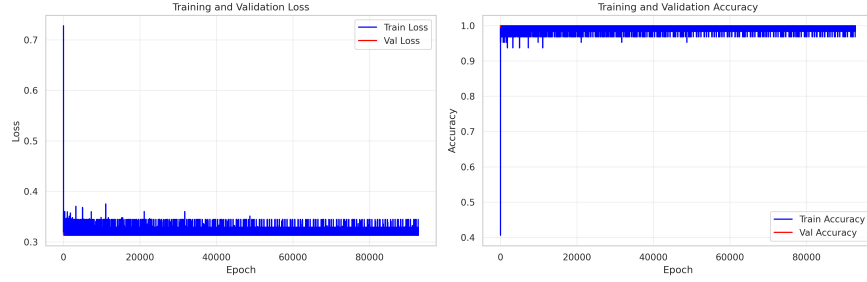


Figure 4.12: Training Curve of MLP-DQN: Episode Rewards over Time

4.5.2 Results of CNN-DQN

Finally, the CNN-DQN model was applied to the multi-class task. Its performance is detailed in Table 4.6.

Table 4.6: Multi-class Classification Results for CNN-DQN

Metric	Score (Macro Avg.)
Accuracy	0.9990082747079059
Precision	0.9990083456349137
Recall	0.9990082747079059
F1-Score	0.9990082756280634

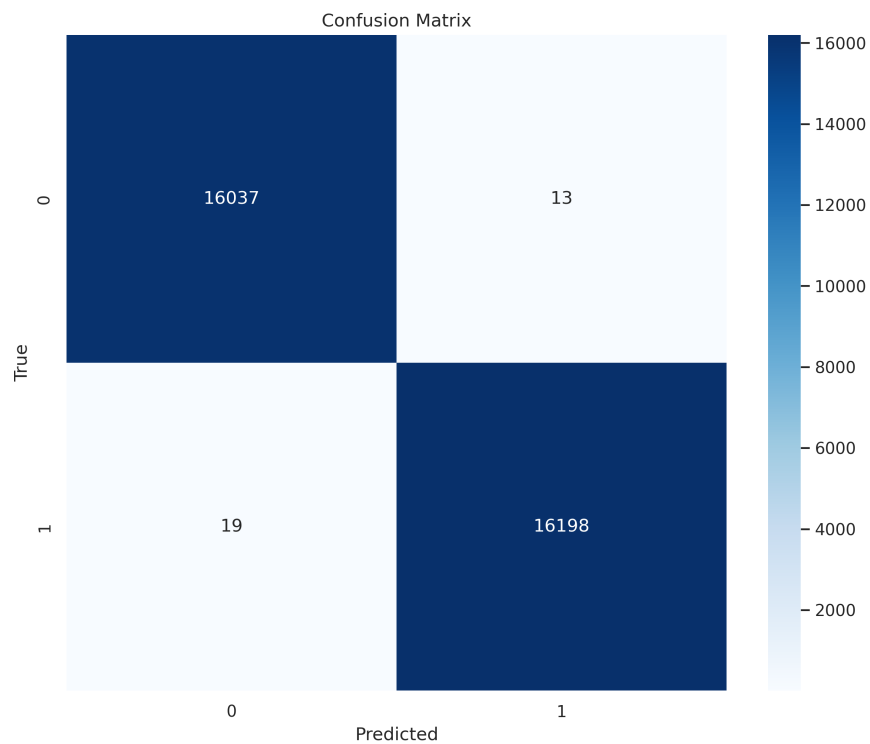


Figure 4.13: Confusion Matrix for the MLP-DQN Model on Multi-class Classification

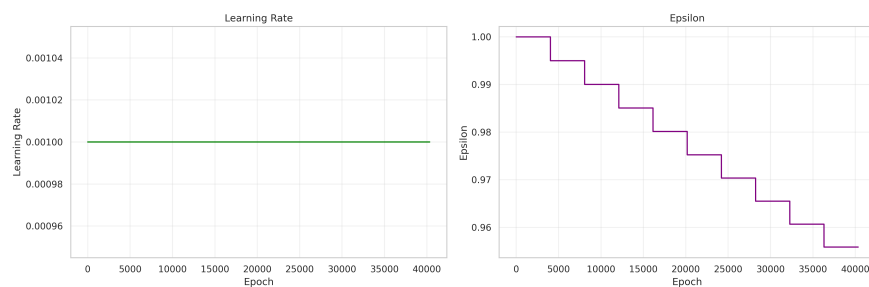


Figure 4.14: Learning Rate and Epsilon Decay Schedule during MLP-DQN Training

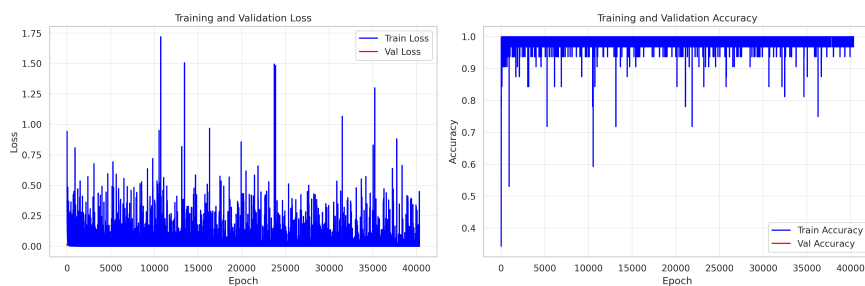


Figure 4.15: Training Curve of MLP-DQN: Episode Rewards over Time

Once again, the CNN-based architecture significantly outperformed the MLP-based models, achieving a macro-averaged F1-Score of 97.58

4.6 Comparison and Discussion

The experimental results provide clear insights into the performance of the proposed models. A comparative summary is presented in Figure ??.

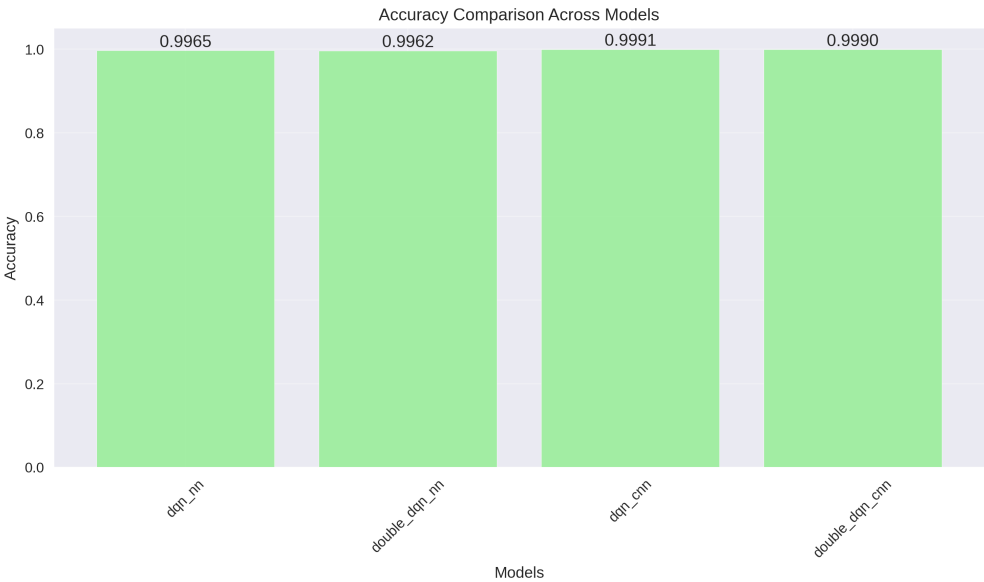


Figure 4.16: Accuracy Comparison Across All Models and Tasks

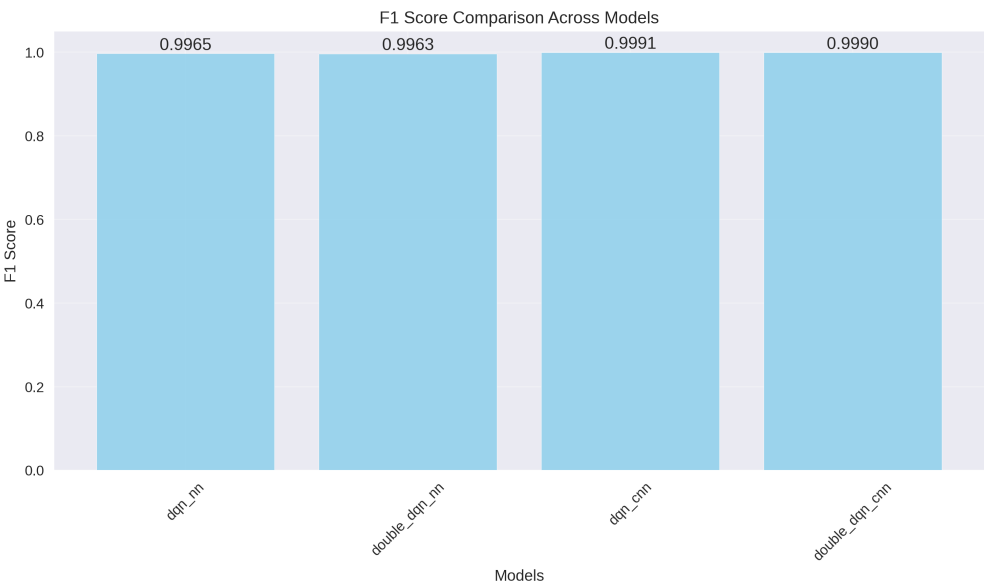


Figure 4.17: F1-Score Comparison Across All Models and Tasks

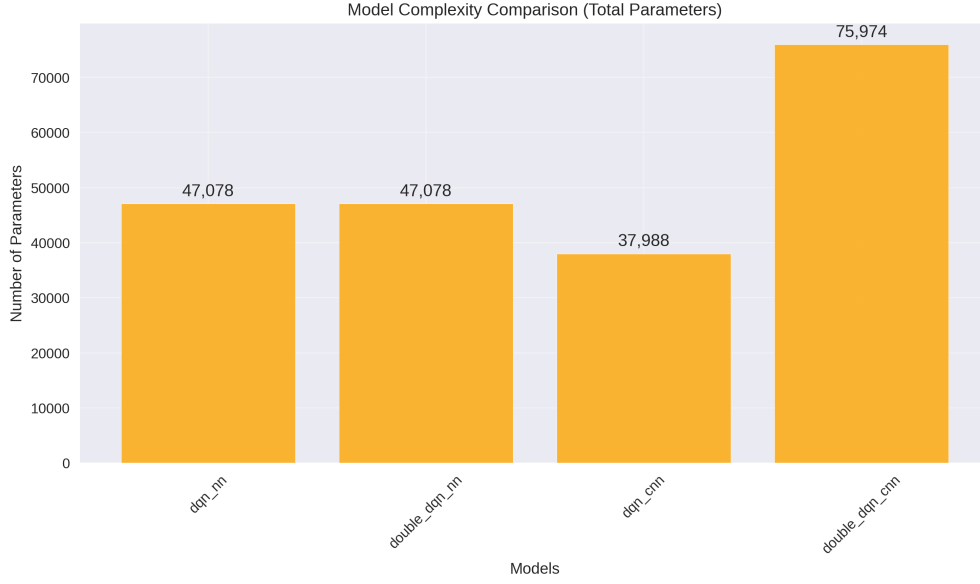


Figure 4.18: Model complexity Comparison Across All Models and Tasks

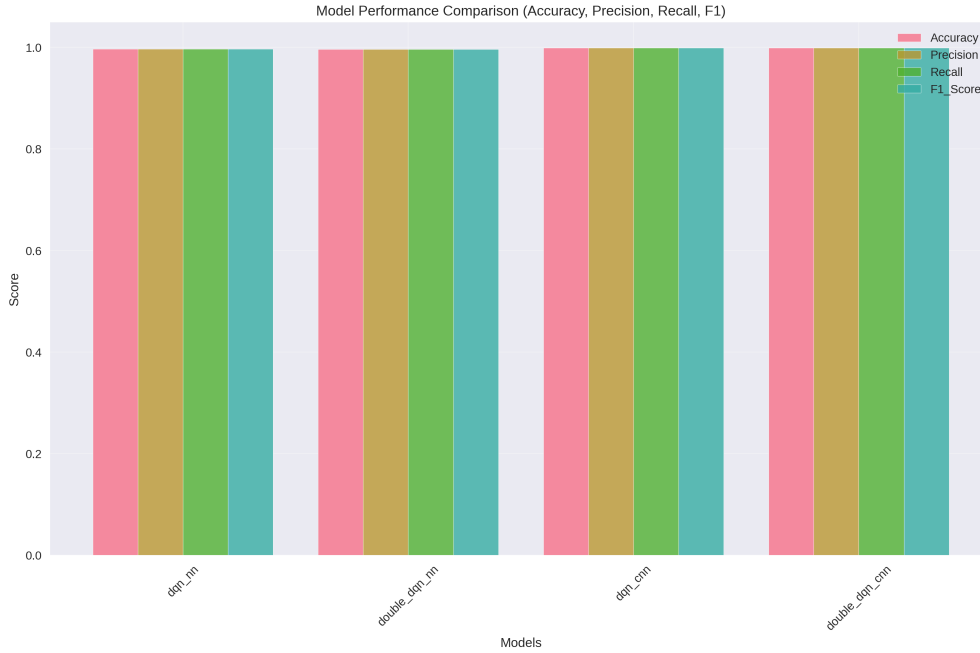


Figure 4.19: Performance metrics Across All Models and Tasks

Several key observations can be drawn from the results:

1. **CNN-DQN Superiority:** Across both binary and multi-class tasks, the CNN-DQN model consistently delivered the highest performance. This validates our hypothesis that there are exploitable spatial correlations within the network features when arranged in an image-like format. The CNN's ability to learn hierarchical patterns provides a distinct advantage over the MLP, which treats features independently.
2. **Effectiveness of DDQN:** The DDQN algorithm provided a modest but consistent performance boost

over the standard DQN in both scenarios. This highlights the practical benefit of using DDQN to mitigate the Q-value overestimation problem, leading to more stable training and a more reliable policy.

3. **Task Complexity Impact:** As expected, all models performed better on the binary classification task than on the multi-class task. The drop in performance for the multi-class scenario reflects the increased complexity of distinguishing between multiple subtle attack variations versus simply separating attack from benign traffic.
4. **Viability of the RL Approach:** Most importantly, the extremely high scores achieved by all models, particularly the CNN-DQN (99.46

4.7 Conclusion

The experiments conducted in this chapter successfully evaluated the performance of our proposed RL-based detection models. The results clearly indicate that the CNN-DQN architecture is the most effective, achieving state-of-the-art detection rates on the CIC-DDoS2019 dataset for both binary and multi-class classification. The findings validate the novel approach of transforming tabular network data into a feature-image and leveraging a CNN for pattern extraction. Furthermore, the consistent, albeit smaller, gains from the DDQN algorithm underscore the importance of algorithmic stability in achieving robust performance. Overall, this empirical evaluation confirms that our DQN-powered system offers a powerful and accurate solution for modern DDoS detection.

General Conclusion

This thesis embarked on an exploration of a novel methodology for detecting Denial-of-Service (DoS) attacks by adapting the principles of Reinforcement Learning (RL), specifically Deep Q-Networks (DQN), to a traditionally supervised classification problem. Motivated by the limitations of static, signature-based Intrusion Detection Systems (IDS), which struggle against novel and evolving threats, our work sought to develop a more dynamic, adaptive, and intelligent detection framework. By reframing network traffic classification as a sequential decision-making process, we successfully designed, implemented, and evaluated a system capable of achieving exceptional detection accuracy.

Summary of Contributions

The primary contributions of this research are multifaceted and can be summarized as follows:

1. **A Novel Methodological Bridge:** We successfully developed a framework that bridges the gap between Reinforcement Learning and supervised classification for the cybersecurity domain. By simulating an RL environment from a static dataset and engineering a hybrid reward and target mechanism, we demonstrated how a DQN agent can be effectively trained to perform a classification task.
2. **Advanced Architectural Design:** We proposed and implemented two distinct neural network architectures for the Q-function approximator: a standard Multi-Layer Perceptron (MLP) and a more advanced Convolutional Neural Network (CNN). The CNN-based approach, which involved a novel transformation of tabular feature vectors into 2D "feature-images," proved to be particularly innovative and effective.
3. **Rigorous Empirical Validation:** The proposed models were rigorously trained and evaluated on the comprehensive and modern CIC-DDoS2019 dataset. Our top-performing model, the CNN-DQN, achieved an F1-Score of 99.46% in binary classification and 97.58% in multi-class classification, demonstrating state-of-the-art performance and confirming the viability of our approach.
4. **Comparative Analysis:** Through a detailed comparison, we empirically proved the superiority of the CNN-based architecture for this task and quantified the stability improvements gained by using a Double DQN (DDQN) algorithm over the standard DQN.

Study Limitations

While the results are highly promising, it is important to acknowledge the limitations of this study, which also pave the way for future work:

- **Static Environment:** The RL agent was trained and evaluated on a static, pre-collected dataset. This means the agent could not learn from the true, real-time consequences of its actions (e.g., how blocking traffic might alter an attacker's strategy). The full adaptive potential of RL is best realized in a live, dynamic environment.
- **Dataset Dependency:** The models' performance was validated on the CIC-DDoS2019 dataset. While comprehensive, its characteristics may not perfectly represent all real-world network environments.

Generalization to different network architectures or entirely new zero-day attacks not represented in the training data remains an open question.

- **Feature Engineering Dependency:** The success of the CNN-DQN model is contingent on the effective construction of the feature-image. The ordering and grouping of features in this grid can impact performance, and the optimal arrangement may vary for different datasets.
- **Computational Overhead:** Training deep reinforcement learning models is computationally more intensive and time-consuming compared to traditional machine learning algorithms like Decision Trees or SVMs, which could be a consideration for resource-constrained environments.

Future Research Directions

This research opens up several exciting avenues for future investigation:

- **Online and Interactive Learning:** The most critical next step is to deploy the RL agent in a live or highly realistic simulated network environment. This would enable online learning, where the agent can continuously adapt its policies based on real-time feedback and the evolving nature of network traffic.
- **Exploration of Advanced RL Algorithms:** This work focused on DQN and DDQN. Future research could explore more advanced, state-of-the-art RL algorithms, such as Proximal Policy Optimization (PPO) or Asynchronous Advantage Actor-Critic (A3C), which may offer improved sample efficiency and stability.
- **Integration of Explainable AI (XAI):** A significant challenge with deep learning models is their "black box" nature. Integrating XAI techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could provide insights into *why* the agent classifies a certain flow as malicious, increasing trust and interpretability for security analysts.
- **Hierarchical and Multi-Agent RL:** A complex network could be defended by a team of RL agents. A hierarchical system could have a high-level agent that identifies a potential threat and delegates the detailed classification to specialized agents trained on specific attack families (e.g., one for volumetric attacks, another for protocol attacks).
- **Automated Reward Shaping and Feature Learning:** Future work could investigate methods for automatically learning the reward function or using techniques like autoencoders to learn the most effective feature representations, reducing the reliance on manual feature engineering.

In conclusion, this thesis has successfully demonstrated that Reinforcement Learning provides a powerful and promising paradigm for building the next generation of intelligent and adaptive Intrusion Detection Systems. By learning to make optimal decisions in the face of uncertainty, RL-powered agents are well-positioned to defend against the dynamic and ever-evolving landscape of cyber threats.

Bibliography

- [1] IBM, *Security Fundamentals*,
<https://www.linkedin.com/pulse/fundamentals-security-slamghana/>
- [2] CIA, *Principles of Information Security*,
<https://www.cleanpng.com/png-information-security-confidentiality-availability-1373941/>
- [3] Fidelis Security, *Understanding DoS and DDoS Attacks*,
<https://fidelissecurity.com/blog/dos-ddos-attacks>
- [4] Imperva, *Denial of Service (DoS) Attack*,
<https://www.imperva.com/learn/ddos/denial-of-service/>
- [5] NETSCOUT, *Slowloris Attack*,
<https://www.netscout.com/blog/asert/slowloris-attack>
- [6] MDPI, *Anomaly-based Detection of DDoS Attacks Using Machine Learning*,
<https://www.mdpi.com/2076-3417/10/3/1052>
- [7] IBM, *What is an Intrusion Detection System (IDS)?*,
<https://www.ibm.com/think/topics/intrusion-detection-system>
- [8] Wikipedia, *Intrusion Detection System*,
https://en.wikipedia.org/wiki/Intrusion_detection_system
- [9] Fortinet, *What is Intrusion Detection Systems (IDS)? How does it Work?*,
<https://www.fortinet.com/resources/cyberglossary/intrusion-detection-system>
- [10] TechTarget, *What Is an Intrusion Detection System (IDS)?*,
<https://www.techtarget.com/searchsecurity/definition/intrusion-detection-system>
- [11] Canadian Institute for Cybersecurity, *CIC-DDoS2019 Dataset*,
<https://www.unb.ca/cic/datasets/ddos-2019.html>