

Defending Against Data Poisoning in Machine Learning-Based Intrusion Detection Systems

A Comparative Evaluation of Anomaly Detection Defences and a Hybrid Majority-Voting Ensemble

Md Mahfuzur Rahman

Programme: MSc Cyber Security and Digital Forensics

Supervisor: Dr. Anum Paracha

A dissertation submitted in partial fulfilment of the requirements
for the degree of

MSc Cyber Security and Digital Forensics

University of Wales Trinity Saint David (UWTSD)

London Campus

June 2026

Declaration

I hereby declare that this dissertation is my own original work and that all sources of information have been appropriately acknowledged through IEEE-style citation. The experimental work, code, results, and analysis presented herein were carried out by me under the supervision of Dr. Anum Paracha at the University of Wales Trinity Saint David. No part of this work has been submitted for any other degree or qualification at this or any other institution.

Signed: _____

Name: Md Mahfuzur Rahman

Date: _____

Abstract

Machine learning (ML) has become integral to modern Intrusion Detection Systems (IDS) because of its capacity to detect novel and evolving cyber threats that escape signature-based methods. However, the reliance of ML-based IDS on the integrity of their training data exposes them to data poisoning attacks, in which an adversary deliberately corrupts the training set so that the deployed classifier learns incorrect decision boundaries. Label flipping, in which a fraction of training labels are inverted, is among the most practical poisoning attacks because it requires only manipulation of the labelling process rather than control of the underlying network features.

This dissertation evaluates the impact of label flipping poisoning on a Random Forest based IDS using the CICIDS-2017 DDoS dataset as the main experiment, and investigates the effectiveness of three anomaly-detection defences — Isolation Forest, Local Outlier Factor (LOF), and One-Class Support Vector Machine (One-Class SVM) — together with a proposed Hybrid Ensemble that combines all three through a 2-of-3 majority-voting strategy. In accordance with the approved ethics form and owing to time and computational resource constraints, the experiments were conducted on a representative subset of the CICIDS-2017 DDoS data at poisoning rates of 5%, 10%, and 20%, with each configuration evaluated using accuracy, precision, recall, F1-score, and false negative rate (FNR). A supplementary set of experiments on a 50,000-record subset of UNSW-NB15, and an exploratory Multi-layer Perceptron (MLP) extension, are reported as additional and future work rather than as the main contribution.

On the CICIDS-2017 DDoS data, the clean Random Forest baseline established a near-perfect reference level of detection performance, reflecting the strong separability of DDoS from benign traffic on this subset. Increasing the poisoning rate degraded performance across all defences, with the most severe degradation at 20% poisoning, where Local Outlier Factor proved the most robust defence (the lowest false negative rate, 0.0428) and Isolation Forest the most vulnerable (the highest false negative rate, 0.0693). The proposed Hybrid Ensemble was the strongest defence at 5% poisoning and remained competitive thereafter, its principal merit being stability across poisoning rates rather than uniform superiority over the best individual detector. The supplementary UNSW-NB15 experiment, whose

results are reported in full, corroborates the general pattern that label flipping raises the false negative rate and that anomaly-detection sanitisation mitigates this degradation.

The dissertation contributes a controlled, like-for-like comparison of three established defences under identical conditions on the CICIDS-2017 DDoS data, a transparent Hybrid Ensemble defence, and an analysis that foregrounds the false negative rate as the most security-relevant metric for an IDS. The original ethics-approved scope was broader, spanning two datasets, two classifiers, and three attack families; the narrowing to a rigorous, completed, and ethics-compliant CICIDS-2017 implementation, the computational constraints encountered, and the supplementary UNSW-NB15 and MLP work are documented transparently as scope adjustments, limitations, and future work.

Keywords: Intrusion Detection Systems, Machine Learning, Adversarial Machine Learning, Data Poisoning, Label Flipping, Anomaly Detection, Isolation Forest, Local Outlier Factor, One-Class SVM, Hybrid Ensemble, CICIDS-2017, UNSW-NB15, False Negative Rate.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Anum Paracha, whose published research on data poisoning and behaviour-informed defences provided both the intellectual foundation and the motivation for this dissertation, and whose detailed feedback substantially improved the quality and rigour of this work. I am also grateful to the academic staff of the University of Wales Trinity Saint David for the grounding in cyber security, digital forensics, and applied machine learning that made this research possible. Finally, I thank my family and friends for their patience and encouragement throughout the duration of this project.

Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
1 Introduction	1
1.1 Background and Context	1
1.2 Problem Statement	2
1.3 Original Approved Aim and Final Implemented Scope	2
1.4 Final Research Aim	3
1.5 Research Objectives	3
1.6 Research Questions	4
1.7 Significance and Contributions	4
1.8 Scope and Delimitations	5
1.9 Dissertation Structure	5
2 Literature Review	6
2.1 Introduction	6
2.2 Machine Learning-Based Intrusion Detection Systems	6
2.3 IDS Benchmark Datasets: UNSW-NB15 and CICIDS-2017	7
2.4 Adversarial Machine Learning	7
2.5 Data Poisoning Attacks	7
2.5.1 Label Flipping	8

2.5.2	Outlier Injection	8
2.5.3	Backdoor Attacks	8
2.6	Why Label Flipping Was Selected for Final Implementation	9
2.7	Anomaly Detection as Data Sanitisation Defence	9
2.8	Isolation Forest	9
2.9	One-Class SVM	10
2.10	Local Outlier Factor	10
2.11	Ensemble and Majority Voting Defences	10
2.12	Literature Justification for Subset-Based Experimentation	10
2.13	Research Gap	11
2.14	Summary	11
3	Research Methodology	12
3.1	Research Approach	12
3.2	Research Design	12
3.3	Dataset Selection: CICIDS-2017 DDoS	14
3.4	Original Dataset Plan and Scope Narrowing	15
3.5	Subset-Based Experimental Approach	15
3.6	Data Preprocessing	16
3.7	Train-Test Split	16
3.8	Attack Design: Label Flipping	16
3.9	Defence Design	16
3.10	Hybrid Ensemble Methodology	16
3.11	Evaluation Metrics	17
3.12	Ethical Considerations	17
3.13	Reproducibility	18
3.14	Methodological Limitations	18
4	Experimental Design and Implementation	19
4.1	Introduction	19

4.2	Experimental Environment	19
4.3	Tools and Technology Stack	19
4.4	Dataset Loading and Preprocessing	20
4.5	Computational Resource Limitations and Scope Adjustment	21
4.6	Explanation of Runtime Constraints and Warnings	22
4.7	Selected Implementation Approach: Subset-Based Experimentation	23
4.8	Random Forest IDS Baseline	23
4.9	Label Flipping Attack Implementation	23
4.10	Individual Defence Implementation	24
4.10.1	Isolation Forest	24
4.10.2	One-Class SVM	24
4.10.3	Local Outlier Factor	24
4.11	Proposed Solution: Hybrid Ensemble Defence	24
4.12	How the Proposed Solution Works	25
4.13	Why Majority Voting Is Used	27
4.14	How the Cleaned Dataset Is Used	27
4.15	Why the Proposed Solution Is Effective	27
4.16	Limitations of the Proposed Solution	28
4.17	Attempted MLP Neural Network Extension	28
4.18	Main CICIDS-2017 Experiment: Implementation Notes	28
4.19	Implementation Workflow	29
4.20	Summary	29
5	Results and Analysis	30
5.1	Introduction	30
5.2	Clean Baseline Results (CICIDS-2017 DDoS)	30
5.3	Poisoned Model Results	31
5.4	Isolation Forest Defence Results against Label Flipping Data Poisoning Attack	31
5.5	LOF Defence Results against Label Flipping Data Poisoning Attack	32

5.6	One-Class SVM Defence Results against Label Flipping Data Poisoning Attack	32
5.7	Hybrid Ensemble Defence Results against Label Flipping Data Poisoning Attack	32
5.8	Comparative Summary and Visualisations (CICIDS-2017 DDoS)	33
5.9	Supplementary Results: UNSW-NB15	36
5.10	Exploratory Results: MLP Extension (UNSW-NB15)	36
5.11	Summary	37
6	Discussion	38
6.1	Introduction	38
6.2	Discussion of RQ1	38
6.3	Discussion of RQ2	39
6.4	Discussion of RQ3	39
6.5	Discussion of RQ4	39
6.6	Implementation Limitations and Effect on Scope	40
6.7	Why LOF Performed Strongly	40
6.8	Why Isolation Forest Degraded at High Poisoning	40
6.9	Why One-Class SVM Is Sensitive and Resource-Intensive	41
6.10	Strengths of the Hybrid Ensemble	41
6.11	Limitations of the Hybrid Ensemble	41
6.12	Limitations of Existing Selected Defences	41
6.13	Importance of FNR in IDS Evaluation	42
6.14	Impact of Resource Constraints	42
6.15	Validity Threats	42
6.16	Relationship to Prior Literature	42
6.17	Reflection on the Attempted MLP Extension	43
6.18	Summary	43
7	Conclusion and Future Work	44
7.1	Summary of the Research	44

7.2	Key Findings	44
7.3	Contributions	45
7.4	Limitations	45
7.5	Future Work	45
7.6	Final Conclusion	46
A	Code Overview	49
B	Runtime Constraints and Warnings	50
C	Attempted MLP Neural Network Extension	51
D	CICIDS-2017 Main Experiment: Run Details	53
E	Ethics Approval Summary	54

List of Figures

3.1	Final research methodology workflow showing the completed main experiment on the CICIDS-2017 DDoS subset; the supplementary UNSW-NB15 and MLP work, and the additional attack families (outlier injection, backdoor) and full multi-attack evaluation, are noted separately as supplementary or future work.	13
3.2	Final data flow diagram: CICIDS-2017 DDoS subset → clean data → label-flipping poisoned data → anomaly detection → 2-of-3 majority voting → cleaned data → retrain Random Forest → evaluation.	14
3.3	Hybrid Ensemble architecture: three detectors feeding a 2-of-3 majority vote that produces a cleaned dataset for retraining.	17
4.1	Tools and technology stack used in the final implementation (Python, Pandas, NumPy, Scikit-learn, Matplotlib, Google Colab/Jupyter), with the MLP shown as an attempted extension only.	20
4.2	Resource limitation and scope adjustment: from the broad approved matrix of approximately 180 configurations to the focused, fully executed core experiment on the CICIDS-2017 DDoS subset.	22
4.3	Proposed solution workflow, from the poisoned dataset through majority-voting sanitisation to retraining and improved metrics.	26
5.1	CICIDS-2017 DDoS: accuracy versus poisoning rate for all four defences under label flipping.	34
5.2	CICIDS-2017 DDoS: F1-score versus poisoning rate for all four defences under label flipping.	35
5.3	CICIDS-2017 DDoS: false negative rate versus poisoning rate for all four defences under label flipping.	35

5.4	CICIDS-2017 DDoS: combined comparison of all four defences across accuracy, F1-score, and FNR.	35
-----	--	----

List of Tables

3.1	Summary of the CICIDS-2017 DDoS data used in the main experiment. . .	15
5.1	Clean baseline IDS performance on the CICIDS-2017 DDoS subset (Random Forest, no poisoning).	30
5.2	Isolation Forest defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).	31
5.3	Local Outlier Factor defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).	32
5.4	One-Class SVM defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).	32
5.5	Hybrid Ensemble defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).	33
5.6	Comparative summary of all defences and the clean baseline under label flipping data poisoning attacks (CICIDS-2017 DDoS).	33
5.7	Qualitative comparison of the four defences against label flipping data poisoning attacks (CICIDS-2017 DDoS).	34
5.8	Supplementary UNSW-NB15 results: all defences and the clean baseline under label flipping data poisoning attacks (50,000-record subset).	36
5.9	Exploratory MLP extension under 10% label flipping (UNSW-NB15; single rate; not a main result).	37

Chapter 1

Introduction

1.1 Background and Context

The scale, frequency, and sophistication of cyber attacks against organisational networks and critical national infrastructure have increased markedly over the past decade. In response, Intrusion Detection Systems (IDS) have become a central element of the defensive architecture of modern networks, providing continuous monitoring of traffic with the objective of identifying malicious or anomalous activity before it can cause harm. Historically, IDS have been divided into two principal categories: signature-based systems, which compare observed traffic against a database of known attack patterns, and anomaly-based systems, which model the statistical characteristics of normal behaviour and flag deviations from it. Signature-based systems are highly precise against documented threats, but they are structurally incapable of detecting zero-day attacks, polymorphic malware, and novel intrusion techniques whose signatures have not yet been catalogued.

To address this limitation, the cyber security community has increasingly turned to machine learning (ML). By learning patterns directly from large volumes of labelled network traffic, ML-based IDS can generalise beyond known attacks and detect previously unseen intrusions. Supervised algorithms such as Random Forest [4], Support Vector Machines, and deep neural networks have demonstrated strong detection performance on widely used benchmark datasets, and ML-based detection is now a standard component of commercial and academic IDS research. The appeal of these methods lies in their ability to automate the discovery of complex, non-linear relationships in high-dimensional feature spaces that would be impractical to encode by hand.

However, the very property that makes machine learning powerful — its dependence on training data — also constitutes its principal weakness in adversarial settings. The conventional assumption underpinning supervised learning is that the training data are an accurate

and trustworthy representation of the phenomenon to be modelled. In a security context, this assumption is frequently untenable. Adversaries who can influence the data used to train a model are able to manipulate its behaviour after deployment, a class of threat known as adversarial machine learning [8]. Within this field, data poisoning attacks are of particular concern because they target the training phase rather than the inference phase, and because their effects can be both severe and difficult to detect.

1.2 Problem Statement

Data poisoning attacks corrupt the training data of a machine learning system so that the resulting model exhibits behaviour chosen by the attacker. In the context of an IDS, a poisoned classifier may learn to misclassify malicious traffic as benign, thereby allowing intrusions to pass undetected while the system continues to report acceptable performance on routine evaluation. This silent failure mode is especially dangerous because it undermines the trust placed in the IDS without producing any obvious symptom of compromise. Among the various poisoning strategies, label flipping is one of the most practical: the adversary inverts the labels of a fraction of training samples without altering their feature values, a manipulation that requires only access to the labelling pipeline rather than control of the underlying network features [7]. Although several anomaly-detection defences have been proposed to counter poisoning, the literature rarely compares them under identical conditions, seldom investigates simple ensemble combinations of them, and tends to emphasise overall accuracy rather than the false negative rate, which is the metric most directly tied to the security failure mode of an IDS. These gaps motivate the present study.

1.3 Original Approved Aim and Final Implemented Scope

The ethics-approved proposal for this research defined a deliberately broad investigation. As originally conceived, the study was to span two benchmark datasets (UNSW-NB15 and CICIDS-2017), two classifiers (Random Forest and a Neural Network), three families of poisoning attack (label flipping, outlier injection, and backdoor attacks), three anomaly-detection defences (Isolation Forest, One-Class SVM, and Local Outlier Factor), and a proposed Hybrid Ensemble defence, each evaluated across multiple poisoning rates. Taken together, this design implied a very large experimental matrix of approximately 180 distinct configurations.

During the course of the project it became clear that this matrix could not be executed rigorously within the time available for an MSc dissertation and within the computational resources accessible to the researcher. The density- and boundary-based detectors are com-

putationally demanding; the neural-network experiments required additional training time; and executing every dataset, classifier, and attack combination at full scale was infeasible. Rather than attempt the full matrix superficially, a deliberate methodological decision was taken, in consultation with the supervisor, to narrow the scope to a single, coherent, and thoroughly executed core experiment that remains squarely within the approved ethics form. In accordance with the supervisor's guidance, the CICIDS-2017 DDoS dataset was selected as the main experimental dataset, using a representative subset to remain tractable while complying with the ethics-approved data plan. The final implemented main scope therefore comprises the CICIDS-2017 DDoS subset, a Random Forest classifier, the label flipping attack at three poisoning rates (5%, 10%, and 20%), the three anomaly-detection defences, and the proposed Hybrid Ensemble. A supplementary experiment on a 50,000-record subset of UNSW-NB15, and an exploratory Multilayer Perceptron (MLP) extension, were also completed; in line with the supervisor's instruction these are presented as additional and future work rather than as the main contribution. Outlier injection and backdoor attacks, although part of the broad approved plan, were not implemented and are discussed only as future work. This narrowing is consistent with the principle that a well-bounded study executed rigorously is of greater scholarly value than a broad one executed superficially.

1.4 Final Research Aim

The aim of this research is to design and evaluate a framework for analysing the impact of data poisoning attacks on machine learning-based Intrusion Detection Systems and investigating anomaly detection techniques to improve robustness. In its final implemented form, the framework is applied principally to the CICIDS-2017 DDoS dataset using a Random Forest classifier, the label flipping attack at three poisoning rates, and three anomaly-detection defences (Isolation Forest, One-Class SVM, and Local Outlier Factor), together with a proposed Hybrid Ensemble defence based on 2-of-3 majority voting. A supplementary evaluation on UNSW-NB15 and an exploratory MLP extension are reported as additional and future work.

1.5 Research Objectives

1. To evaluate the impact of label flipping data poisoning attacks on a Random Forest-based IDS using the CICIDS-2017 DDoS dataset.
2. To assess the effectiveness of Isolation Forest, One-Class SVM, and Local Outlier Factor as anomaly detection-based defence mechanisms against label flipping poisoning

attacks.

3. To design, implement, and evaluate a Hybrid Ensemble defence using 2-of-3 majority voting to improve robustness against poisoned training data.
4. To corroborate the findings through a supplementary evaluation on UNSW-NB15 and an exploratory MLP extension, and to document implementation limitations, including computational resource constraints, subset-based experimentation, and the attack families (outlier injection and backdoor) reserved for future work.

1.6 Research Questions

- RQ1:** How does label flipping data poisoning affect the performance of a Random Forest-based Intrusion Detection System (IDS) using the CICIDS-2017 DDoS dataset at different poisoning rates?
- RQ2:** How effective are Isolation Forest, One-Class Support Vector Machine (One-Class SVM), and Local Outlier Factor (LOF) as defence mechanisms against label flipping poisoning attacks on a machine learning-based IDS?
- RQ3:** Can a Hybrid Ensemble defence based on majority voting improve the robustness of a machine learning-based IDS compared with individual anomaly detection techniques?
- RQ4:** How can the impact of label flipping attacks and the effectiveness of defence mechanisms be evaluated using performance metrics such as Accuracy, Precision, Recall, F1-Score, and False Negative Rate (FNR)?

The implementation limitations and the effect of computational-resource constraints on the final experimental scope, originally considered as a separate question, are addressed within the limitations and future-work discussion (Sections 6.14 and 7.4–7.5) rather than as a standalone research question.

1.7 Significance and Contributions

This dissertation makes several contributions. First, it provides a controlled, like-for-like comparison of three established anomaly-detection defences against label flipping, conducted under identical preprocessing, classifier, attack, and evaluation conditions, thereby removing the confounding heterogeneity that hampers comparison across existing studies.

Second, it designs and evaluates a Hybrid Ensemble defence based on a transparent 2-of-3 majority-voting rule and characterises its behaviour relative to the individual detectors. Third, it foregrounds the false negative rate as the primary metric of operational interest, on the grounds that a missed attack represents a successful intrusion. Fourth, it documents transparently the computational constraints encountered and the consequent narrowing of scope, providing an honest account that is itself instructive for researchers planning comparable studies under limited resources.

1.8 Scope and Delimitations

The final main study is bounded to the CICIDS-2017 DDoS dataset, a single base classifier (Random Forest), and a single attack family (label flipping) evaluated at three poisoning rates, together with four defences (Isolation Forest, Local Outlier Factor, One-Class SVM, and the Hybrid Ensemble). Experiments were conducted on a representative subset of the CICIDS-2017 DDoS data to remain tractable and ethics-compliant. A supplementary evaluation on a 50,000-record subset of UNSW-NB15 and an exploratory MLP extension are also reported. Additional attack families (outlier injection and backdoor attacks) and a full multi-attack, multi-classifier evaluation are outside the completed scope and are treated as future work.

1.9 Dissertation Structure

The remainder of the dissertation is organised as follows. Chapter 2 reviews the relevant literature. Chapter 3 sets out the research methodology. Chapter 4 describes the experimental design and implementation, including the computational constraints, the subset-based approach, and a detailed account of the proposed solution. Chapter 5 presents the results. Chapter 6 provides a critical discussion. Chapter 7 concludes and identifies future work. References and appendices follow.

Chapter 2

Literature Review

2.1 Introduction

This chapter situates the dissertation within the existing body of knowledge. It reviews machine learning in intrusion detection and the benchmark datasets that underpin the field, introduces adversarial machine learning and the taxonomy of data poisoning attacks, explains why label flipping was selected for the final implementation, reviews anomaly detection as a data-sanitisation defence and the three detectors employed, considers ensemble and majority-voting defences, justifies the use of subset-based experimentation, and concludes by articulating the research gap.

2.2 Machine Learning-Based Intrusion Detection Systems

Machine learning has been applied to intrusion detection for more than two decades, motivated by the inability of signature-based systems to detect novel attacks. The central premise is that malicious and benign network behaviours exhibit statistically distinguishable patterns in feature space, and that a suitably trained classifier can learn to separate them. A wide range of algorithms has been investigated, including k -Nearest Neighbours, Decision Trees, Support Vector Machines, Random Forests, and deep neural networks. Random Forest [4], an ensemble of decision trees trained on bootstrap samples with randomised feature selection, is particularly popular in IDS research because it achieves strong predictive performance with minimal hyperparameter tuning, is resistant to overfitting, and handles the mixed, high-dimensional feature spaces typical of network traffic with relative ease. These properties make it a sensible baseline classifier for a study whose primary focus is adversarial robustness rather than classifier optimisation.

2.3 IDS Benchmark Datasets: UNSW-NB15 and CICIDS-2017

The credibility of an IDS study depends heavily on the dataset used for evaluation. Early research relied on KDD Cup 99 and NSL-KDD, but these are now regarded as outdated. Two datasets have since become de-facto standards: UNSW-NB15 [5], produced by the Australian Centre for Cyber Security, and CICIDS-2017 [6], produced by the Canadian Institute for Cybersecurity. UNSW-NB15 combines real benign background traffic with synthetic attack traffic generated using the IXIA PerfectStorm tool, spanning nine modern attack categories and providing a rich set of flow-level features. CICIDS-2017 is likewise widely used and realistic. In this dissertation, the DDoS portion of CICIDS-2017 is used as the main experimental dataset, while UNSW-NB15 is used as a supplementary benchmark.

2.4 Adversarial Machine Learning

Adversarial machine learning is the study of the security and robustness of learning systems in the presence of intelligent adversaries. The demonstration by Goodfellow et al. [9] that imperceptible perturbations could cause classifiers to misclassify inputs with high confidence revealed that the vulnerability of learning systems was both deep and general. Attacks are commonly organised by the phase of the pipeline they target — evasion attacks at inference time, poisoning attacks at training time — and by the attacker’s knowledge, ranging from white-box through grey-box to black-box. Biggio and Roli [8], in their retrospective on a decade of adversarial machine learning, argue that defences must be designed with adaptive adversaries in mind, since attackers respond to any fixed defence by adjusting their strategy. This insight is directly relevant to ensemble defences, where the assumption that detector errors are independent may not hold against an adversary who crafts poisoning to evade several detectors simultaneously.

2.5 Data Poisoning Attacks

Data poisoning attacks were placed on a rigorous footing by Biggio, Nelson, and Laskov [7], who demonstrated that a Support Vector Machine could be systematically degraded by injecting a small number of carefully crafted training samples. Subsequent work generalised the threat: Jagielski et al. [13] extended poisoning to regression learning and proposed countermeasures; Muñoz-González et al. [14] generalised gradient-based poisoning to deep learning; and Steinhardt, Koh, and Liang [15] introduced certified defences pro-

viding provable bounds on the loss an attacker can induce. Recent work by Paracha et al. [1–3] analyses the behaviour of machine-learning algorithms under poisoning and develops attack-agnostic, behaviour-informed defences, and is directly relevant to the present study.

2.5.1 Label Flipping

Label flipping is a form of availability poisoning in which the adversary inverts the labels of selected training samples without altering their feature values [7]. In a binary intrusion-detection setting this means relabelling some malicious samples as benign and vice versa, which teaches the classifier an inverted association between features and class and manifests as degraded detection performance and an elevated false negative rate. Its practical significance lies in its modest requirements: the attacker need only influence the labelling process, not the feature-extraction infrastructure.

2.5.2 Outlier Injection

Outlier injection attacks insert abnormal feature vectors into the training set to distort the learned distribution. Unlike label flipping, outlier injection requires the attacker to craft and inject new feature vectors, which is more demanding in network environments where feature extraction is performed downstream of the attacker’s reach. Outlier injection is part of the original approved scope but was not implemented in the final study; it is identified as future work.

2.5.3 Backdoor Attacks

Backdoor attacks, exemplified by the BadNets work of Gu et al. [16], implant a hidden trigger pattern such that any input bearing the trigger is misclassified into an attacker-chosen class while ordinary inputs are classified correctly. Backdoor attacks are particularly insidious because the model behaves normally on the vast majority of inputs. As with outlier injection, backdoor attacks were part of the broad original scope but are not part of the completed implementation.

2.6 Why Label Flipping Was Selected for Final Implementation

Label flipping was selected as the sole attack family for the final implementation for three reasons. First, it is the most practical and realistic of the three poisoning families, requiring only access to the labelling pipeline. Second, it is the most widely studied, which supports comparison with prior work. Third, and decisively for a resource-constrained study, it is the most tractable to implement and evaluate rigorously across multiple poisoning rates and defences, allowing the available time and computational budget to be concentrated on a thorough, internally consistent core experiment rather than spread thinly across several attack families. Outlier injection and backdoor attacks remain valuable directions for future work.

2.7 Anomaly Detection as Data Sanitisation Defence

Defences against poisoning fall broadly into data sanitisation, robust learning, and certification. This dissertation focuses on data sanitisation through anomaly detection, which detects and removes poisoned samples before training. The paradigm is attractive for intrusion detection because it is model-agnostic and integrates naturally into the periodic retraining pipelines of operational systems [15]. The underlying assumption is that poisoned samples occupy regions of feature space that mark them as statistical outliers — an assumption that, as later discussion shows, is only partially satisfied by label flipping, since flipping leaves the feature vector unchanged.

2.8 Isolation Forest

Isolation Forest [10] isolates observations by recursively partitioning the feature space using randomly selected features and split values; anomalies, being few and different, are isolated with fewer partitions and therefore have shorter average path lengths. It is computationally efficient, scales linearly with the number of samples, and performs well in high dimensions. Its weakness in the present context is that its global, axis-aligned partitioning is comparatively poorly suited to identifying samples that are anomalous only in their label rather than their features.

2.9 One-Class SVM

One-Class SVM [11] learns a boundary enclosing the bulk of the training data in a kernel feature space, treating points outside the boundary as anomalies; the parameter ν bounds the fraction of outliers and support vectors. It has a strong theoretical foundation and can model complex boundaries, but it is sensitive to feature scaling and parameter choice, and its computational cost grows steeply with the size of the training set — a constraint of direct relevance to this study. Its assumption that the bulk of the data is clean is progressively violated as the poisoning rate rises.

2.10 Local Outlier Factor

Local Outlier Factor [12] is a density-based method that compares the local density of a sample with that of its k nearest neighbours, flagging samples in regions of comparatively low density. Because it makes a local rather than a global assessment, it is well suited to datasets with multiple regions of differing density, a structure common in network traffic. Its drawbacks are computational: the nearest-neighbour density computations are expensive on large datasets, and its effectiveness can be eroded by high dimensionality.

2.11 Ensemble and Majority Voting Defences

Ensemble learning is well established in supervised classification, where combining diverse base learners reduces variance and improves generalisation. The same logic motivates combining multiple anomaly detectors: if the detectors make imperfectly correlated errors, aggregating their decisions reduces the variance of the combined judgement. In the IDS poisoning context, however, ensemble defences remain comparatively under-investigated, and a controlled evaluation of a simple, transparent 2-of-3 majority-voting ensemble across Isolation Forest, LOF, and One-Class SVM under identical label-flipping conditions has not been widely reported. This gap is what the proposed Hybrid Ensemble addresses.

2.12 Literature Justification for Subset-Based Experimentation

The use of a representative subset of a large dataset for controlled experimentation is an established practice in machine-learning research, particularly where computational resources

are bounded or where the research objective is comparative rather than absolute. Because the present study seeks to measure the *relative* effectiveness of four defences under identical conditions, rather than to establish maximum achievable performance on the full dataset, a representative subset that preserves the original class balance is methodologically sufficient: every defence is trained and evaluated on the same data under the same protocol, so observed differences remain attributable to the defence. This justification is developed further in Chapters 3 and 4.

2.13 Research Gap

Three observations emerge. First, comparable evaluations of anomaly-detection defences under identical dataset, classifier, attack, and metric conditions are scarce. Second, simple ensemble defences against poisoning are under-explored in the IDS setting, and the conditions under which an ensemble outperforms its constituents are not well characterised. Third, the dominant reliance on accuracy and F1-score obscures the false negative rate, the metric most directly tied to the security failure mode of an IDS. This dissertation addresses these gaps directly.

2.14 Summary

The literature establishes machine learning as central to modern IDS, identifies data poisoning — and label flipping in particular — as a realistic and serious threat, and reviews three complementary anomaly-detection defences whose differing inductive biases motivate their combination. The chapter has justified the focus on label flipping and the use of subset-based experimentation, and has identified the research gap that the remainder of the dissertation addresses.

Chapter 3

Research Methodology

3.1 Research Approach

This dissertation adopts a quantitative, experimental research approach. The independent variables — the poisoning rate and the choice of defence — are systematically varied while all other factors are held constant, and the dependent variables — the five evaluation metrics — are measured. Because the dataset, preprocessing pipeline, base classifier, and evaluation protocol are identical across all conditions, any observed difference in performance can be attributed to the manipulated variables rather than to confounding factors. The approach is positivist in orientation and consistent with the dominant methodology in the adversarial machine-learning literature.

3.2 Research Design

The experimental procedure comprises the following stages, depicted in Figure 3.1: load and preprocess UNSW-NB15; split into training and test partitions; train a clean baseline Random Forest IDS; poison the training labels by label flipping at each rate; apply each defence to produce a cleaned training set; retrain and evaluate a defended IDS on the held-out test set; and aggregate and compare the results. The test set is never poisoned and is identical across all conditions.

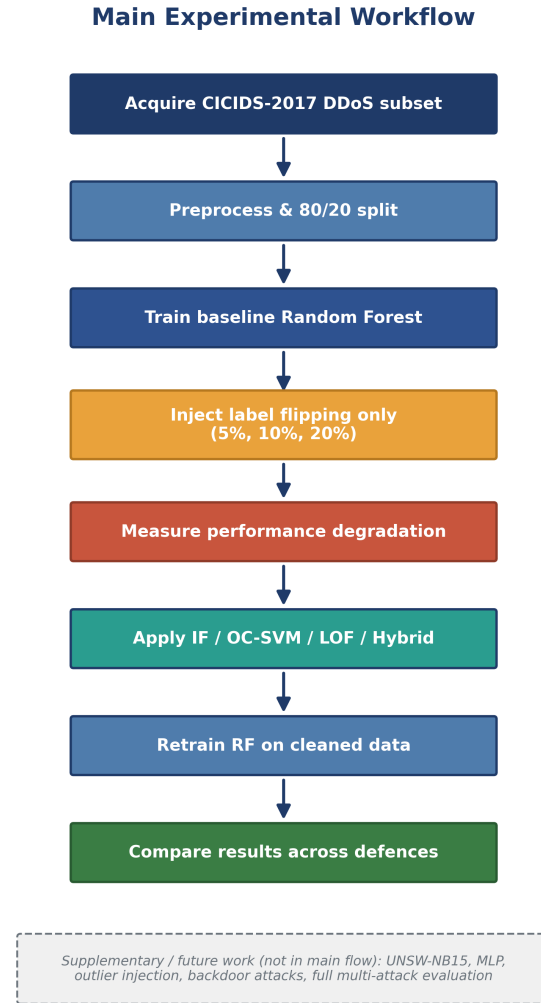


Figure 3.1: Final research methodology workflow showing the completed main experiment on the CICIDS-2017 DDoS subset; the supplementary UNSW-NB15 and MLP work, and the additional attack families (outlier injection, backdoor) and full multi-attack evaluation, are noted separately as supplementary or future work.

Figure 3.2 presents the corresponding data flow, tracing how the CICIDS-2017 DDoS subset is transformed at each processing stage from raw data through to evaluation.

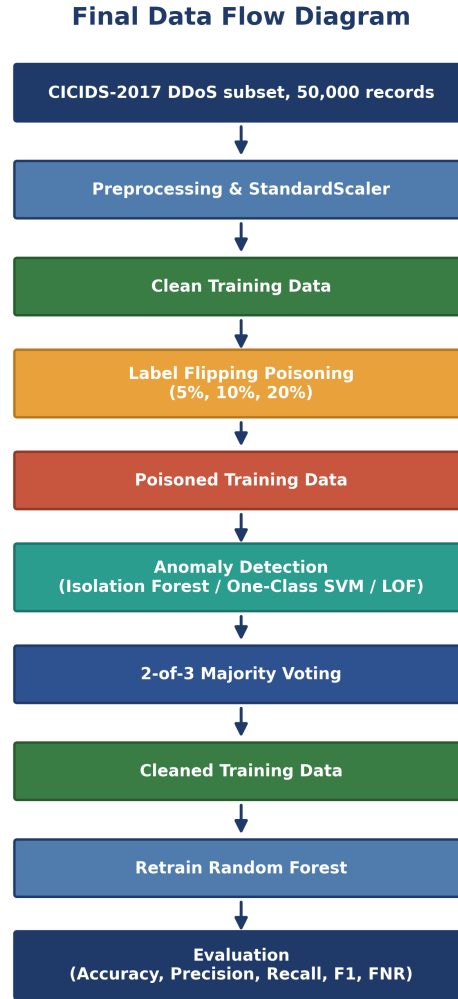


Figure 3.2: Final data flow diagram: CICIDS-2017 DDoS subset → clean data → label-flipping poisoned data → anomaly detection → 2-of-3 majority voting → cleaned data → retrain Random Forest → evaluation.

3.3 Dataset Selection: CICIDS-2017 DDoS

The CICIDS-2017 dataset [6], produced by the Canadian Institute for Cybersecurity, was selected as the main experimental dataset, in accordance with the supervisor’s guidance and the approved ethics form. The main experiment uses the DDoS portion of CICIDS-2017 (the Friday afternoon DDoS capture), framed as a binary classification task (0 = benign, 1 = DDoS attack). It was chosen because it is a contemporary, widely used, and realistic benchmark, and because the DDoS subset provides a clear, well-defined attack class suitable for a controlled poisoning study. To remain tractable within the available computational resources, a representative subset of 50,000 records was drawn with a fixed random seed (`random_state = 42`). The principal characteristics of the data used are summarised in

Table 3.1. A supplementary experiment additionally uses a 50,000-record subset of UNSW-NB15 [5] to corroborate the findings on a second benchmark.

Table 3.1: Summary of the CICIDS-2017 DDoS data used in the main experiment.

Attribute	Value
Main dataset	CICIDS-2017 (DDoS capture)
Source file	Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv
Subset used in experiments	50,000 records (random_state = 42)
Input features	78 numerical flow features
Label encoding	Binary: 0 = BENIGN, 1 = DDoS
Feature scaling	StandardScaler (zero mean, unit variance)
Train/test split	80% / 20% (stratified)
Supplementary dataset	UNSW-NB15 (50,000-record subset)

3.4 Original Dataset Plan and Scope Narrowing

The approved proposal envisaged the use of both CICIDS-2017 and UNSW-NB15, two classifiers, and three attack families. As explained in Section 1.3, the full experimental matrix of approximately 180 configurations proved infeasible within the available time and computational resources. The scope was therefore narrowed, on the supervisor’s advice, to a single coherent core experiment on the CICIDS-2017 DDoS subset with a Random Forest classifier and the label flipping attack, with UNSW-NB15 and an MLP extension retained as supplementary and exploratory work, and the remaining elements documented as future work. This narrowing is a deliberate methodological choice in favour of depth and rigour over breadth, and it keeps the study squarely within the approved ethics form.

3.5 Subset-Based Experimental Approach

Because of the computational constraints, a representative subset of 50,000 records was drawn from the CICIDS-2017 DDoS data (and, for the supplementary experiment, from UNSW-NB15) using a fixed random seed (random_state = 42) to ensure reproducibility. The subset preserves both benign and attack samples so that the class structure of the full data is reflected. This is a subset-based experimental approach, not batch (incremental) training: the model is trained on the static subset in a single pass rather than on sequential mini-batches of the full dataset. The justification for this approach, and its status as a

limitation, are developed in Chapter 4 and revisited in Section 3.14.

3.6 Data Preprocessing

Preprocessing proceeds as follows: infinite values are replaced with NaN and rows with missing values are dropped; only numerical features are retained; the binary label is separated from the feature matrix; and the features are standardised to zero mean and unit variance using `StandardScaler`, which is important for the distance- and density-based detectors whose computations are sensitive to feature scale.

3.7 Train-Test Split

The preprocessed subset is partitioned into 80% training and 20% testing. The test partition is held out, left unpoisoned, and kept identical across all conditions, providing a stable basis for comparison.

3.8 Attack Design: Label Flipping

For a given poisoning rate r , a randomly selected fraction r of the *training* labels is inverted (normal becomes attack and vice versa). Three rates are studied — 5%, 10%, and 20% — representing a stealthy, a moderately aggressive, and a severe attacker. Only training labels are poisoned; the test set remains pristine.

3.9 Defence Design

Four defences are evaluated. The three individual anomaly detectors — Isolation Forest, Local Outlier Factor, and One-Class SVM — each scan the poisoned training data and flag anomalous samples, which are removed before retraining. The fourth is the proposed Hybrid Ensemble, described next.

3.10 Hybrid Ensemble Methodology

The Hybrid Ensemble combines the three detectors through a 2-of-3 majority-voting rule. Each detector emits a binary flag per training sample; a sample is removed if and only if at

least two of the three detectors flag it as anomalous. Formally, sample i is removed when

$$\text{flag}_{\text{IF}}(i) + \text{flag}_{\text{SVM}}(i) + \text{flag}_{\text{LOF}}(i) \geq 2.$$

The Random Forest is then retrained on the cleaned dataset. The architecture is shown in Figure 3.3.

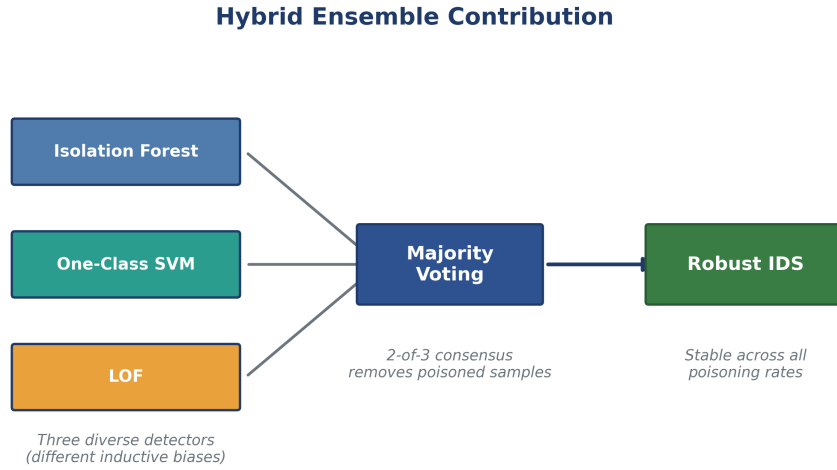


Figure 3.3: Hybrid Ensemble architecture: three detectors feeding a 2-of-3 majority vote that produces a cleaned dataset for retraining.

3.11 Evaluation Metrics

Five metrics are computed on the held-out test set: accuracy, precision, recall, F1-score, and false negative rate (FNR). The FNR, defined as $\text{FN}/(\text{FN}+\text{TP})$, is emphasised throughout because in an intrusion-detection context a false negative is a missed attack, and a missed attack is a successful intrusion.

3.12 Ethical Considerations

This research is a secondary-data computational study and raises no significant ethical concerns. It involves no human participants, no interviews, questionnaires, focus groups, or audio/video recordings, and no personal or confidential data. It uses only publicly available benchmark datasets (CICIDS-2017 and UNSW-NB15) together with computer-generated network-log files. All attacks are simulated offline against the researcher’s own models; no live system and no third-party network is targeted. The work is defensive and educational in

purpose, and its ethical risk is assessed as low. The project was conducted within the period of the approved ethics application (proposed start date 10 April 2026; proposed end date 9 June 2026). No separate ethics approval reference number was provided on the available form; the approval status should be confirmed against the official approval record held by the supervisor. A fuller summary is provided in Appendix E.

3.13 Reproducibility

All random operations use a fixed seed (`random_state = 42`), rendering the subset sampling, the poisoning selection, and the Random Forest and Isolation Forest training deterministic; LOF and One-Class SVM are deterministic given fixed inputs and parameters. The complete analysis is contained in a single Jupyter notebook, `msc_final.ipynb`, developed and executed in Google Colab with data held in Google Drive (under the project folders `/content/drive/MyDrive/research/` and `/content/drive/MyDrive/mahfuz/`). No public repository URL was used; the submitted notebook file is provided with the dissertation artifact materials, together with the exported result files, so that the full pipeline can be re-executed and the reported figures reproduced.

3.14 Methodological Limitations

Three methodological limitations are acknowledged. First, the use of a 50,000-record subset rather than the full dataset means that absolute performance figures may differ at full scale, although relative comparisons remain valid. Second, the use of a single random seed means the reported values reflect a single realisation; multiple seeds with confidence intervals would strengthen the analysis. Third, the contamination parameter of each detector is aligned with the true poisoning rate, which is optimistic relative to a real deployment in which the rate is unknown. These limitations are revisited in Chapter 6.

Chapter 4

Experimental Design and Implementation

4.1 Introduction

This chapter describes how the methodology was realised in practice. It details the experimental environment and tools, the loading and preprocessing of the data, the computational constraints that shaped the study and the consequent scope adjustment, the subset-based implementation approach, the baseline and attack implementation, the individual defences, and a detailed account of the proposed Hybrid Ensemble solution. The main experiment is the CICIDS-2017 DDoS subset; the chapter also documents the supplementary UNSW-NB15 experiment and the exploratory MLP extension.

4.2 Experimental Environment

The experiments were implemented in Python and executed in a Google Colab / Jupyter environment with access to standard CPU resources and limited memory. This environment was chosen for its zero cost, its integration with Google Drive for dataset storage, and its pre-installed scientific Python stack, but its memory and processing ceilings proved to be a binding constraint on the scale of experimentation, as described in Section 4.5.

4.3 Tools and Technology Stack

The implementation uses Python with pandas and NumPy for data handling, scikit-learn [17] for preprocessing, the Random Forest classifier, and all three anomaly detectors, and matplotlib for visualisation. The technology stack is illustrated in Figure 4.1.

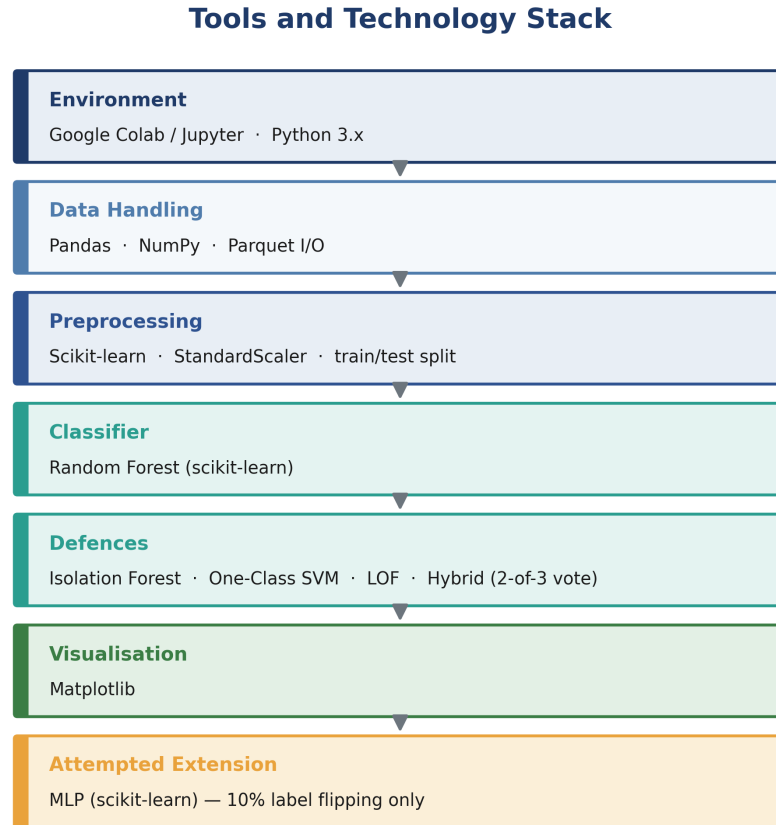


Figure 4.1: Tools and technology stack used in the final implementation (Python, Pandas, NumPy, Scikit-learn, Matplotlib, Google Colab/Jupyter), with the MLP shown as an attempted extension only.

4.4 Dataset Loading and Preprocessing

For the main experiment, the CICIDS-2017 DDoS capture (Friday-WorkingHours-Afternoon-DDos.pcap) was loaded from Google Drive, and a subset of 50,000 records was drawn with `random_state = 42`. Column names were stripped of whitespace; infinite values were replaced with NaN and rows with missing values were dropped; the textual label was converted to a binary target (BENIGN $\rightarrow$ 0, DDoS $\rightarrow$ 1); the label columns were separated from the 78 numerical flow features; and the features were standardised using StandardScaler. The data were then split 80/20 (stratified) into training and test sets. The supplementary UNSW-NB15 experiment followed the same procedure on a 50,000-record subset of UNSW-NB15.

4.5 Computational Resource Limitations and Scope Adjustment

The original approved aim implied a very large experimental matrix: two datasets, two classifiers, three attack families, three poisoning rates, and five defence settings, amounting to approximately 180 distinct configurations. This was infeasible within the MSc timeframe and the available Google Colab and local resources for several reasons. CICIDS-2017 is large and distributed across multiple files, requiring substantial memory and preprocessing effort. The density- and boundary-based detectors — One-Class SVM and Local Outlier Factor — are computationally expensive: One-Class SVM scales poorly with the number of training samples because it evaluates kernel relationships across the data, while Local Outlier Factor must compute nearest-neighbour densities for every sample, a cost that grows rapidly with both sample count and dimensionality. The neural-network experiments required additional training time.

In combination, these factors made the full matrix impractical. The scope was therefore adjusted, on the supervisor's advice, to a single coherent core experiment on a 50,000-record subset of the CICIDS-2017 DDoS data, with a supplementary UNSW-NB15 experiment and an exploratory MLP extension also completed, and the remaining elements documented as future work. This adjustment is represented conceptually in Figure 4.2.

Resource Limitation and Scope Adjustment

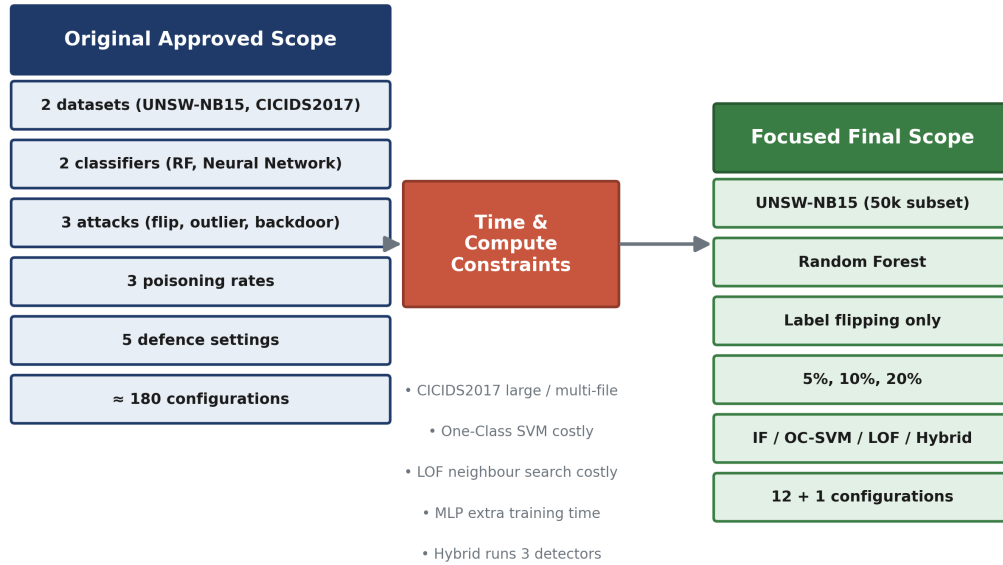


Figure 4.2: Resource limitation and scope adjustment: from the broad approved matrix of approximately 180 configurations to the focused, fully executed core experiment on the CICIDS-2017 DDoS subset.

4.6 Explanation of Runtime Constraints and Warnings

During implementation, the computational constraints described above manifested principally as prohibitively long execution times rather than as outright fatal errors. Running the density- and boundary-based detectors (Local Outlier Factor and One-Class SVM) repeatedly, across multiple poisoning rates and over large volumes of data, was not feasible within a practical time budget in the free Google Colab environment; this is what necessitated the representative 50,000-record subset described in the next section.

In addition, one reproducible warning was observed during execution. When Local Outlier Factor was applied, scikit-learn issued the warning “*Duplicate values are leading to incorrect results. Increase the number of neighbours for more accurate results.*” This is a non-fatal warning rather than an error: execution completed and results were produced. It arises because network-flow datasets such as CICIDS-2017 and UNSW-NB15 commonly contain duplicated or near-identical feature vectors, which give zero pairwise distances and can distort the local-density estimates on which LOF relies. The warning did not halt execution and did not affect the validity of the comparative analysis; it is reported transparently here as an implementation limitation, and addressing it properly — for example by increas-

ing the neighbourhood size or de-duplicating the feature space — is identified as future work. No runtime or memory crash occurred during the final experiments; the constraint was one of execution time rather than outright failure, which is why the representative 50,000-record subset was adopted.

These observations directly informed the subset-based approach described next.

4.7 Selected Implementation Approach: Subset-Based Experimentation

In response to the constraints, this study adopted a subset-based experimental approach. A representative subset of 50,000 CICIDS-2017 DDoS records (and, for the supplementary experiment, 50,000 UNSW-NB15 records) was drawn with a fixed seed (`random_state = 42`), preserving both benign and attack samples in proportions representative of the full data. This is subset-based experimentation, not batch (incremental) training: the model is trained in a single pass over the static subset rather than on sequential mini-batches drawn from the full dataset. The approach is methodologically defensible because the research objective is comparative: every defence is trained and evaluated on the same data under the same protocol, so observed differences remain attributable to the defence rather than to the data. Controlled experimentation on a representative subset is an established practice in resource-constrained machine-learning research, and it enabled the full experimental matrix of four defences across three poisoning rates to be executed consistently and repeatably. It is nonetheless a limitation: absolute figures may differ at full scale, and future work should employ the complete datasets with batch learning, cloud computing, or a GPU-supported environment.

4.8 Random Forest IDS Baseline

A Random Forest classifier was trained on the clean (unpoisoned) training subset to establish a baseline. This baseline provides the reference against which all poisoned and defended configurations are compared and confirms that the classifier, features, and preprocessing are fit for purpose before any attack is introduced.

4.9 Label Flipping Attack Implementation

The label flipping attack was implemented as a reusable function that accepts the training labels and a poisoning rate, selects the corresponding fraction of indices at random, and

inverts their binary labels. The attack was applied at 5%, 10%, and 20%. Only training labels were modified; the test set was left unchanged.

4.10 Individual Defence Implementation

4.10.1 Isolation Forest

Isolation Forest [10] was applied to the poisoned training features with its contamination parameter aligned to the assumed poisoning rate. Samples predicted as anomalous were removed, and the Random Forest was retrained on the remainder.

4.10.2 One-Class SVM

One-Class SVM [11] was applied with the parameter ν set in correspondence with the poisoning rate and the default RBF kernel. As noted, this detector was the most computationally demanding of the three.

4.10.3 Local Outlier Factor

Local Outlier Factor [12] was applied with the same contamination setting, flagging samples whose local density was substantially lower than that of their neighbours.

4.11 Proposed Solution: Hybrid Ensemble Defence

The proposed solution is a Hybrid Ensemble Defence designed to reduce the impact of label flipping poisoning on a Random Forest-based IDS. Label flipping corrupts the training labels, which can cause the model to learn incorrect decision patterns; the solution introduces an anomaly-detection-based data-sanitisation stage before retraining. Isolation Forest, One-Class SVM, and LOF are each applied to the poisoned training data, and each produces a binary anomaly flag for every sample. The Hybrid Ensemble combines these flags using a 2-of-3 majority-voting rule: if at least two detectors classify a sample as anomalous, the sample is removed. The remaining cleaned dataset is then used to retrain the Random Forest IDS, which is evaluated on the clean test set.

4.12 How the Proposed Solution Works

The operation of the solution can be expressed as the following workflow, illustrated in Figure 4.3:

1. Load the CICIDS-2017 DDoS dataset (and, for the supplementary experiment, UNSW-NB15).
2. Preprocess the data (clean, scale, separate label).
3. Split into training and testing sets (80/20).
4. Apply the label flipping attack to the training labels.
5. Apply Isolation Forest, One-Class SVM, and LOF to the poisoned training data.
6. Generate a binary anomaly flag from each detector.
7. Apply 2-of-3 majority voting across the three flags.
8. Remove samples receiving two or more anomaly votes.
9. Retrain the Random Forest on the cleaned dataset.
10. Evaluate using accuracy, precision, recall, F1-score, and FNR.

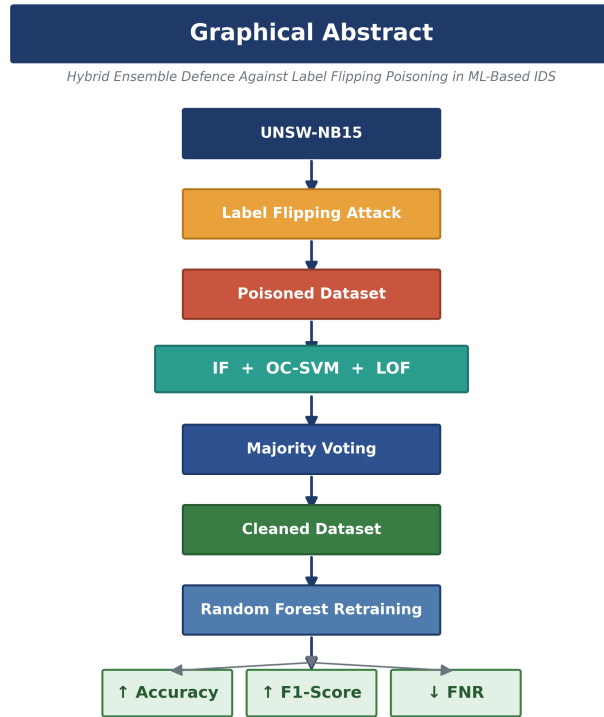


Figure 4.3: Proposed solution workflow, from the poisoned dataset through majority-voting sanitisation to retraining and improved metrics.

The voting logic is summarised in the following pseudo-code.

```

1  Input: poisoned training dataset Dp
2  For each sample xi in Dp:
3      flag_IF = IsolationForest(xi)      # 1 if anomaly else 0
4      flag_SVM = OneClassSVM(xi)         # 1 if anomaly else 0
5      flag_LOF = LOF(xi)                 # 1 if anomaly else 0
6      vote_score = flag_IF + flag_SVM + flag_LOF
7      if vote_score >= 2:
8          remove xi                     # majority judges it poisoned
9      else:
10         keep xi                         # majority judges it clean
11  Output: cleaned dataset Dclean
12  Train RandomForest on Dclean
13  Evaluate on clean test set (Accuracy, Precision, Recall, F1,
    FNR)
  
```

Listing 4.1: Hybrid Ensemble 2-of-3 majority-voting sanitisation.

4.13 Why Majority Voting Is Used

Majority voting is used because it balances sensitivity against specificity and reduces dependency on any single detector. A 1-of-3 rule would be over-aggressive — a single false alarm would discard a clean sample, eroding the training set. A 3-of-3 rule would be over-conservative — poisoned samples detected by only two detectors would survive. The 2-of-3 rule requires corroboration between detectors before a sample is removed, so that the idiosyncratic error of one detector can be overruled by the agreement of the other two, while genuine consensus still triggers removal.

4.14 How the Cleaned Dataset Is Used

After the majority vote, all samples flagged for removal are discarded, leaving a cleaned training set from which the most clearly anomalous (and therefore most likely poisoned) samples have been excluded. The Random Forest IDS is then retrained from scratch on this cleaned set, so that it learns from data of higher integrity than the poisoned set. The retrained model is evaluated on the held-out clean test set, ensuring that the reported metrics reflect genuine detection performance against true labels.

4.15 Why the Proposed Solution Is Effective

The Hybrid Ensemble is effective for several related reasons. It combines tree-based (Isolation Forest), boundary-based (One-Class SVM), and density-based (LOF) detection principles, which respond differently to the structure of the data and therefore make imperfectly correlated errors. It reduces dependency on any single detector, so that a detector performing poorly under a given poisoning rate does not by itself determine the outcome. Majority voting reduces the chance of erroneous removal driven by a single detector's false alarm. By removing likely-poisoned samples, it improves the quality of the training data, and by improving robustness it can reduce false negatives — the most security-critical errors — relative to the weaker individual detectors.

4.16 Limitations of the Proposed Solution

The solution also has limitations. It incurs a higher computational cost than any single detector, since all three must be executed before a vote can be taken. It may remove clean samples when two detectors raise coincident false alarms. It can fail when label-flipped samples appear entirely normal in feature space, since no feature-space detector can identify a sample whose features are typical of its class. Its quality depends on the quality of its constituent detectors. The contamination rate is assumed known in these experiments but would be unknown in real-world deployment. Finally, the main evaluation was conducted on the CICIDS-2017 DDoS subset with a Random Forest classifier; the UNSW-NB15 experiment and the MLP extension are supplementary and exploratory respectively, and a full multi-attack, multi-classifier evaluation is reserved for future work.

4.17 Attempted MLP Neural Network Extension

An exploratory extension applied a Multilayer Perceptron (MLP) neural network, implemented with scikit-learn, to the UNSW-NB15 subset under 10% label flipping, to gauge whether the findings might generalise beyond Random Forest. This was an exploratory extension rather than a primary experiment: it was conducted at a single poisoning rate, without the full defence matrix across all rates, and is not presented as a main result. The observed outcomes are reported descriptively in Section 5.10 and Appendix C. A full neural-network evaluation across all poisoning rates and defences is identified as future work. During this extension, the LOF defence additionally produced a runtime warning regarding duplicate values, which is documented and explained in Section 5.10 and Appendix C as an implementation limitation.

4.18 Main CICIDS-2017 Experiment: Implementation Notes

The CICIDS-2017 DDoS experiment constitutes the main completed implementation. Following the supervisor’s instruction to use CICIDS-2017 for compliance with the ethics form, the DDoS capture was loaded, subset, cleaned, binarised, scaled, and split as described in Section 4.4, and the full pipeline — clean Random Forest baseline, label flipping at 5%, 10%, and 20%, the three individual detectors, and the Hybrid Ensemble — was executed exactly as for the methodology in Chapter 3. The contamination parameter of each detector was aligned with the poisoning rate, and the Hybrid Ensemble retained any sample receiving at least two “clean” votes from the three detectors. The complete results are reported in Chapter 5, and an overview of the code is given in Appendix A. The same

pipeline was additionally run on UNSW-NB15 as a supplementary experiment to provide cross-dataset corroboration.

4.19 Implementation Workflow

The complete implementation workflow integrates the stages above: data loading and pre-processing; baseline training; attack injection at three rates; application of each of the four defences; retraining on cleaned data; and evaluation. The workflow is applied identically across defences and rates, so that the only varying factors are the poisoning rate and the defence, as required for a controlled comparison.

4.20 Summary

This chapter has documented the experimental environment, the computational constraints and the resulting scope adjustment, the subset-based approach and its justification, the baseline, attack, and defence implementations, and a detailed account of the proposed Hybrid Ensemble — what it is, how it works, why majority voting is used, how the cleaned dataset is used, why it is effective, and its limitations. The main experiment was completed on the CICIDS-2017 DDoS subset, with a supplementary UNSW-NB15 experiment and an exploratory MLP extension also completed. The results follow in Chapter 5.

Chapter 5

Results and Analysis

5.1 Introduction

This chapter presents the experimental results. The main experiment was conducted on the CICIDS-2017 DDoS subset (Sections 5.2–5.9). A supplementary experiment on UNSW-NB15 is reported in Section 5.10, and the exploratory MLP extension in Section 5.11. All results were produced by the implementation described in Chapter 4; interpretation is deferred to Chapter 6. Throughout, particular attention is paid to the false negative rate (FNR), since a false negative corresponds to a missed attack and is therefore the most security-critical error for an IDS.

5.2 Clean Baseline Results (CICIDS-2017 DDoS)

The clean Random Forest baseline, trained on the unpoisoned CICIDS-2017 DDoS subset, classified the held-out test set without error, achieving an accuracy, precision, recall, and F1-score of 1.0000 and a false negative rate of 0.0000, as shown in Table 5.1. The corresponding confusion matrix recorded 4,297 true negatives and 5,701 true positives with no misclassifications.

Table 5.1: Clean baseline IDS performance on the CICIDS-2017 DDoS subset (Random Forest, no poisoning).

Configuration	Acc.	Prec.	Recall	F1	FNR
Clean Baseline	1.0000	1.0000	1.0000	1.0000	0.0000

It is important to interpret this perfect baseline with appropriate caution. The CICIDS-2017 DDoS subset contrasts high-volume distributed denial-of-service traffic with benign

traffic, and these two classes are strongly separable in the flow-feature space, so a well-specified Random Forest is able to separate them almost perfectly on this particular subset. The perfect score should therefore be read as a property of the separability of this specific binary DDoS-versus-benign task, not as a claim of flawless detection in general. Its main role in this dissertation is to serve as an upper reference point against which the degradation caused by poisoning, and the recovery afforded by each defence, can be measured.

5.3 Poisoned Model Results

When the training labels were corrupted by label flipping without any defence, detection performance declined relative to the perfect baseline. At a representative 10% poisoning rate the undefended poisoned model attained an accuracy of 0.9858 and an F1-score of 0.9876, with the false negative rate rising from 0.0000 to approximately 0.0114. This confirms that label flipping degrades the IDS and, in particular, increases the number of missed attacks, motivating the defences evaluated below.

5.4 Isolation Forest Defence Results against Label Flipping Data Poisoning Attack

Table 5.2 reports the Isolation Forest defence across the three poisoning rates. The defence performs well at 5% and 10% but degrades markedly at 20%, where its false negative rate rises to 0.0693 — the highest of any defence at any rate in the CICIDS-2017 experiment, indicating the greatest vulnerability to missed attacks under severe poisoning.

Table 5.2: Isolation Forest defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).

Poisoning	Acc.	Prec.	Recall	F1	FNR
5%	0.9850	0.9766	0.9975	0.9870	0.0025
10%	0.9520	0.9294	0.9911	0.9593	0.0089
20%	0.9360	0.9559	0.9307	0.9431	0.0693

5.5 LOF Defence Results against Label Flipping Data Poisoning Attack

Table 5.3 reports the Local Outlier Factor defence, which is the strongest individual defence under severe poisoning: at 20% it retains an accuracy of 0.9533 and the lowest false negative rate of any defence in the experiment, 0.0428.

Table 5.3: Local Outlier Factor defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).

Poisoning	Acc.	Prec.	Recall	F1	FNR
5%	0.9943	0.9946	0.9954	0.9950	0.0046
10%	0.9870	0.9857	0.9916	0.9886	0.0084
20%	0.9533	0.9607	0.9572	0.9590	0.0428

5.6 One-Class SVM Defence Results against Label Flipping Data Poisoning Attack

Table 5.4 reports the One-Class SVM defence, which is competitive throughout and second-best at 20% poisoning (FNR 0.0472), behind LOF.

Table 5.4: One-Class SVM defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).

Poisoning	Acc.	Prec.	Recall	F1	FNR
5%	0.9944	0.9946	0.9956	0.9951	0.0044
10%	0.9815	0.9768	0.9911	0.9839	0.0089
20%	0.9417	0.9454	0.9528	0.9491	0.0472

5.7 Hybrid Ensemble Defence Results against Label Flipping Data Poisoning Attack

Table 5.5 reports the proposed Hybrid Ensemble. It is the best-performing defence at 5% poisoning, achieving the highest accuracy (0.9964) and the lowest false negative rate (0.0033) of any configuration. At 10% and 20%, however, it does not surpass LOF: its false negative rate at 10% (0.0133) is in fact the highest of the four defences at that rate,

and at 20% (0.0554) it sits between One-Class SVM and Isolation Forest. The ensemble is therefore strongest at low poisoning and remains competitive, but it does not dominate the best individual detector at higher poisoning rates on this dataset.

Table 5.5: Hybrid Ensemble defence results against label flipping data poisoning attacks across different poisoning rates (CICIDS-2017 DDoS).

Poisoning	Acc.	Prec.	Recall	F1	FNR
5%	0.9964	0.9970	0.9967	0.9968	0.0033
10%	0.9811	0.9803	0.9867	0.9835	0.0133
20%	0.9379	0.9464	0.9446	0.9455	0.0554

5.8 Comparative Summary and Visualisations (CICIDS-2017 DDoS)

Table 5.6 consolidates all configurations against the clean baseline, and Table 5.7 provides a qualitative comparison. Figures 5.1–5.4 visualise the trends.

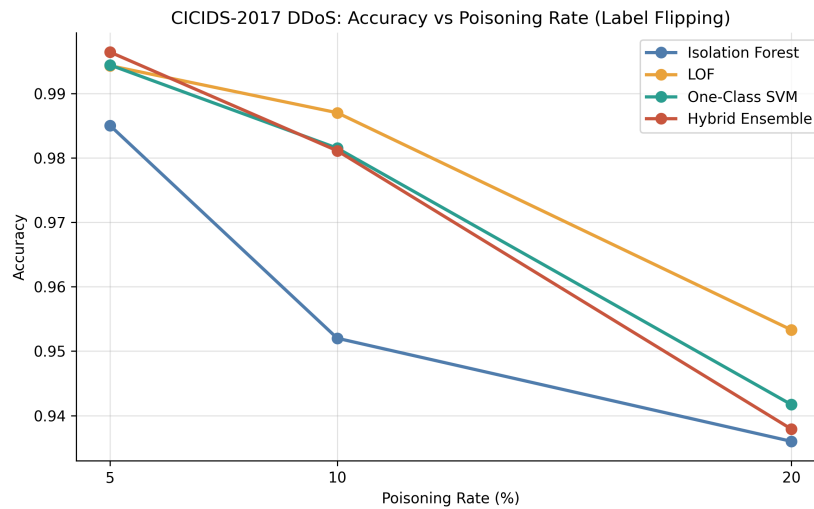
Table 5.6: Comparative summary of all defences and the clean baseline under label flipping data poisoning attacks (CICIDS-2017 DDoS).

Pois.	Defence	Acc.	Prec.	Recall	F1	FNR
—	Clean Baseline	1.0000	1.0000	1.0000	1.0000	0.0000
5%	Isolation Forest	0.9850	0.9766	0.9975	0.9870	0.0025
5%	LOF	0.9943	0.9946	0.9954	0.9950	0.0046
5%	One-Class SVM	0.9944	0.9946	0.9956	0.9951	0.0044
5%	Hybrid Ensemble	0.9964	0.9970	0.9967	0.9968	0.0033
10%	Isolation Forest	0.9520	0.9294	0.9911	0.9593	0.0089
10%	LOF	0.9870	0.9857	0.9916	0.9886	0.0084
10%	One-Class SVM	0.9815	0.9768	0.9911	0.9839	0.0089
10%	Hybrid Ensemble	0.9811	0.9803	0.9867	0.9835	0.0133
20%	Isolation Forest	0.9360	0.9559	0.9307	0.9431	0.0693
20%	LOF	0.9533	0.9607	0.9572	0.9590	0.0428
20%	One-Class SVM	0.9417	0.9454	0.9528	0.9491	0.0472
20%	Hybrid Ensemble	0.9379	0.9464	0.9446	0.9455	0.0554

Table 5.7: Qualitative comparison of the four defences against label flipping data poisoning attacks (CICIDS-2017 DDoS).

Defence	Detection Principle	Principle	Performance Trend	Strength	Limitation
Isolation Forest	Tree-based random partitioning		Strong at 5–10%, weakest at 20%	Fast, scalable	Highest FNR (0.0693) at 20%
One-Class SVM	Boundary-based kernel estimation		Strong at 5%, solid at 20%	Competitive across rates	Sensitive to scaling/parameters; costly
Local Outlier Factor	Density-based local comparison		Most robust at 10–20%	Lowest FNR (0.0428) at 20%	Duplicate-value sensitivity; costly
Hybrid Ensemble	2-of-3 majority voting		Best at 5%, competitive at 10–20%	Best FNR (0.0033) at 5%	Does not beat LOF at 10–20%; 3× cost

Overall observation: on the CICIDS-2017 DDoS data, LOF is the most robust defence under severe poisoning, while the Hybrid Ensemble is strongest at low poisoning and remains competitive thereafter.

**Figure 5.1:** CICIDS-2017 DDoS: accuracy versus poisoning rate for all four defences under label flipping.

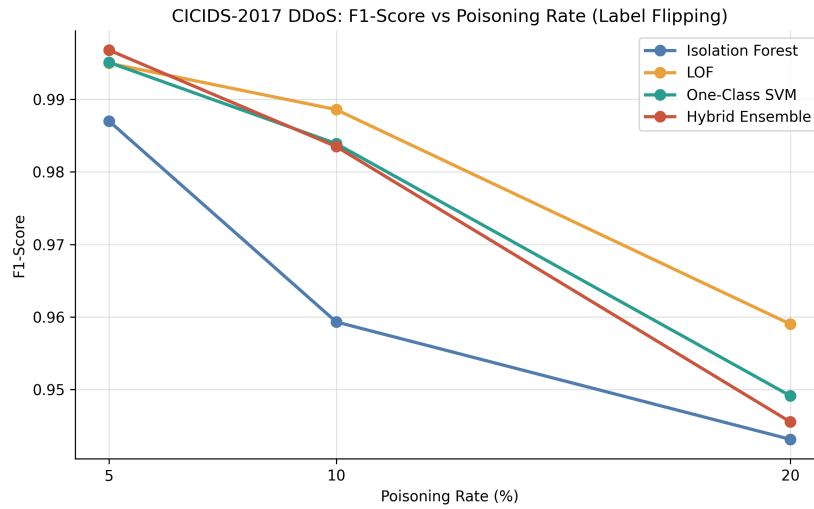


Figure 5.2: CICIDS-2017 DDoS: F1-score versus poisoning rate for all four defences under label flipping.

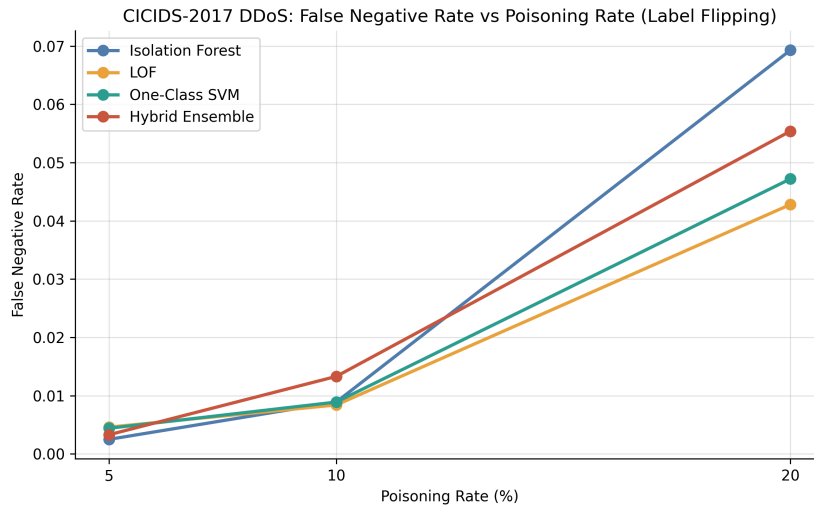


Figure 5.3: CICIDS-2017 DDoS: false negative rate versus poisoning rate for all four defences under label flipping.

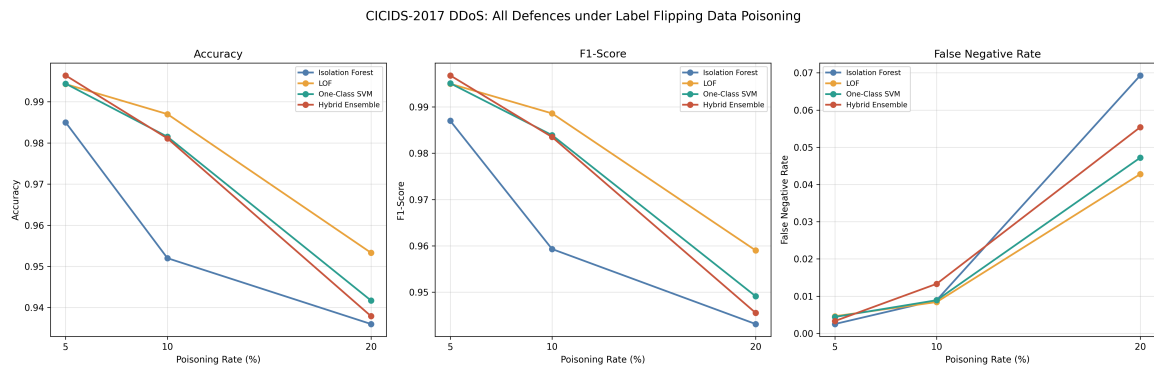


Figure 5.4: CICIDS-2017 DDoS: combined comparison of all four defences across accuracy, F1-score, and FNR.

5.9 Supplementary Results: UNSW-NB15

To corroborate the main findings on a second benchmark, the same pipeline was applied to a 50,000-record subset of UNSW-NB15. These results are supplementary and are reported for comparison rather than as the main contribution. The clean UNSW-NB15 baseline achieved an accuracy of 0.9443 and an F1-score of 0.9604 (FNR 0.0268). The defended results are summarised in Table 5.8. As on CICIDS-2017, label flipping degraded performance with increasing poisoning, LOF was the strongest individual defence under severe poisoning (accuracy 0.9127, FNR 0.0546 at 20%), Isolation Forest was the weakest (FNR 0.0996 at 20%), and the Hybrid Ensemble was consistently more stable than Isolation Forest while remaining competitive with LOF and One-Class SVM. The consistency of these patterns across two independent datasets strengthens confidence in the general conclusions.

Table 5.8: Supplementary UNSW-NB15 results: all defences and the clean baseline under label flipping data poisoning attacks (50,000-record subset).

Pois.	Defence	Acc.	Prec.	Recall	F1	FNR
—	Clean Baseline	0.9443	0.9480	0.9732	0.9604	0.0268
5%	Isolation Forest	0.9372	0.9454	0.9653	0.9553	0.0347
5%	LOF	0.9410	0.9471	0.9692	0.9580	0.0308
5%	One-Class SVM	0.9400	0.9473	0.9675	0.9573	0.0325
5%	Hybrid Ensemble	0.9374	0.9452	0.9652	0.9551	0.0348
10%	Isolation Forest	0.9304	0.9421	0.9587	0.9503	0.0413
10%	LOF	0.9330	0.9437	0.9608	0.9522	0.0392
10%	One-Class SVM	0.9347	0.9429	0.9644	0.9535	0.0356
10%	Hybrid Ensemble	0.9346	0.9436	0.9612	0.9523	0.0388
20%	Isolation Forest	0.8872	0.9348	0.9004	0.9173	0.0996
20%	LOF	0.9127	0.9300	0.9454	0.9377	0.0546
20%	One-Class SVM	0.8878	0.9124	0.9276	0.9199	0.0724
20%	Hybrid Ensemble	0.9115	0.9289	0.9400	0.9365	0.0600

5.10 Exploratory Results: MLP Extension (UNSW-NB15)

As an exploratory extension, a Multilayer Perceptron (MLP) was applied to the UNSW-NB15 subset under 10% label flipping only. This is reported as exploratory and future work, in line with the supervisor’s guidance, and is not part of the main contribution. The summary outcomes are shown in Table 5.9. The clean MLP performed best, poisoning degraded it, and the defences applied to the poisoned MLP did not clearly improve upon

the clean MLP; among the defended configurations the Hybrid Ensemble (FNR 0.0333) outperformed MLP + Isolation Forest (0.0468) and MLP + One-Class SVM (0.0556) and was close to MLP + LOF (0.0324). During this extension the LOF defence produced a non-fatal duplicate-value warning, discussed in Appendix C as an implementation limitation.

Table 5.9: Exploratory MLP extension under 10% label flipping (UNSW-NB15; single rate; not a main result).

Configuration	Accuracy	F1	FNR
Clean MLP	0.9364	0.9556	0.0143
Poisoned MLP (10%)	0.9340	0.9535	0.0259
MLP + Isolation Forest	0.9225	0.9447	0.0468
MLP + LOF	0.9290	0.9498	0.0324
MLP + One-Class SVM	0.9171	0.9406	0.0556
MLP + Hybrid Ensemble	0.9284	0.9494	0.0333

5.11 Summary

On the main CICIDS-2017 DDoS experiment, the clean baseline was perfectly separable, label flipping degraded all defences with the most severe impact at 20%, LOF was the most robust defence under severe poisoning (lowest FNR at 10% and 20%), Isolation Forest was the weakest at 20%, and the Hybrid Ensemble was the best defence at 5% while remaining competitive thereafter. The supplementary UNSW-NB15 experiment showed a consistent overall pattern, and the exploratory MLP extension is reserved for future work. These results are interpreted in Chapter 6.

Chapter 6

Discussion

6.1 Introduction

This chapter interprets the results in relation to the four research questions and the wider literature, examines why each defence behaved as it did, considers the strengths and limitations of the Hybrid Ensemble and of the existing defences, reflects on the importance of the false negative rate and the impact of resource constraints, and assesses the threats to validity.

6.2 Discussion of RQ1

The first research question asked how label flipping affects a Random Forest IDS at different poisoning rates. On the main CICIDS-2017 DDoS experiment, the results show a monotonic degradation from a perfectly separable baseline: performance at 5% remains very high, declines at 10%, and falls more steeply at 20%. The false negative rate rises from 0.0000 at the clean baseline to as much as 0.0693 under the weakest defence at 20%. The supplementary UNSW-NB15 experiment shows the same qualitative pattern from a more modest baseline (FNR 0.0268 rising to 0.0996 under the weakest defence at 20%). The non-linearity reflects the Random Forest’s capacity to absorb a small amount of label noise through ensemble averaging up to a threshold, beyond which the inverted associations begin to dominate. The practical implication is that the region of greatest danger is precisely the region of heaviest poisoning, where defences matter most.

6.3 Discussion of RQ2

The second question concerned the effectiveness of the three individual detectors. On the CICIDS-2017 DDoS data the three are closely matched at 5%, but a clear ordering emerges as poisoning increases: at both 10% and 20% Local Outlier Factor achieves the lowest false negative rate of the three individual detectors (0.0084 and 0.0428 respectively), One-Class SVM is intermediate, and Isolation Forest is weakest at 20% (FNR 0.0693). LOF's local, density-based perspective is better suited to the structure of label-flipped contamination than the global views of the other two, as discussed in Sections 6.7–6.9. The supplementary UNSW-NB15 experiment exhibits the same ordering at 20%, reinforcing the conclusion.

6.4 Discussion of RQ3

The third question asked whether the Hybrid Ensemble improves robustness relative to the individual detectors. The evidence supports a nuanced answer. On the CICIDS-2017 DDoS data, the Hybrid Ensemble is the best of all configurations at 5% poisoning, achieving the lowest false negative rate (0.0033); however, at 10% and 20% it does not surpass LOF, and at 10% its FNR (0.0133) is in fact the highest of the four defences at that rate. The ensemble's principal virtue is therefore not that it dominates the best individual detector, but that it is stable and consistently avoids the worst outcomes: it never collapses in the way Isolation Forest does at 20%. From an ex ante standpoint, in which the operative poisoning rate is unknown, this stability is valuable, since it removes the risk of inadvertently committing to a detector that proves to be the worst performer under the realised conditions. The supplementary UNSW-NB15 results are consistent: the ensemble there is consistently more stable than Isolation Forest while remaining competitive with LOF and One-Class SVM. The claim made for the Hybrid Ensemble is thus deliberately measured rather than overstated.

6.5 Discussion of RQ4

The fourth question concerned evaluation through the five metrics. The results demonstrate the value of multiple metrics in concert. Accuracy and F1-score summarise overall performance but can understate security significance, because a model can retain respectable accuracy while still missing a meaningful fraction of attacks. The false negative rate exposes this directly: on CICIDS-2017, Isolation Forest at 20% retains an accuracy of 0.9360 yet its FNR rises to 0.0693, a far larger proportional change. Precision and recall together reveal whether degradation is due to over-alerting or under-detection; foregrounding the

FNR ensures that the security-critical failure mode — missed attacks — is not masked by otherwise healthy aggregate metrics.

6.6 Implementation Limitations and Effect on Scope

Beyond the four research questions, it is important to record the implementation limitations encountered and how they shaped the final scope. The principal limitation was the infeasibility of the original ~ 180 -configuration matrix within the available time and resources, driven by the computational cost of One-Class SVM and LOF and the additional training time required for neural networks. These constraints necessitated a representative 50,000-record subset of the CICIDS-2017 DDoS data (and of UNSW-NB15 for the supplementary experiment) and the narrowing of scope to a single coherent core experiment, undertaken in consultation with the supervisor and within the approved ethics form. The constraints thus shaped the study's breadth but not the internal validity of its core comparison, which remains controlled and consistent. The honest documentation of these constraints is itself instructive, since it informs the planning of comparable resource-bounded studies.

6.7 Why LOF Performed Strongly

LOF performed strongly, particularly under severe poisoning, because its density-based, local assessment captures irregularities that global methods overlook. Label-flipped samples often remain spatially typical of their original class, so a detector that evaluates each sample against its local neighbourhood is better placed to detect subtle inconsistencies than one that relies on a single global model of normality. This accounts for LOF achieving the lowest false negative rate at both 10% (0.0084) and 20% (0.0428) on the CICIDS-2017 DDoS data, a pattern echoed in the supplementary UNSW-NB15 results.

6.8 Why Isolation Forest Degraded at High Poisoning

Isolation Forest degraded most at high poisoning because its random, axis-aligned partitioning is suited to isolating samples that are genuinely distant from the data mass, but label flipping does not move samples in feature space — it alters only the label. As flipped samples become more numerous, an increasing fraction escape isolation and survive into the cleaned set, corrupting the retrained classifier. This explains its position as the weakest defence at 20% on CICIDS-2017 (FNR 0.0693), and its even larger FNR of 0.0996 at 20% in the supplementary UNSW-NB15 experiment.

6.9 Why One-Class SVM Is Sensitive and Resource-Intensive

One-Class SVM is sensitive because it estimates a single global boundary whose validity depends on the bulk of the data being clean; as the poisoning rate rises, this assumption is increasingly violated and the boundary is drawn around an already-contaminated distribution. It is resource-intensive because its kernelised form scales poorly with the number of training samples, which both limited the scale of experimentation and contributed to the resource constraints discussed in Section 6.6.

6.10 Strengths of the Hybrid Ensemble

The Hybrid Ensemble's strengths are conceptual transparency, stability across poisoning rates, and the avoidance of worst-case collapse. Its 2-of-3 rule is simple to implement and reason about, the detectors run independently, and the ensemble is the best configuration at low poisoning (5%) while never exhibiting the sharp collapse seen in Isolation Forest at 20%. It does not, however, uniformly beat the strongest individual detector (LOF) at higher poisoning rates, and this is stated plainly rather than overclaimed.

6.11 Limitations of the Hybrid Ensemble

Its limitations are a higher computational cost (three detectors rather than one), the possible removal of clean samples through coincident false alarms, a dependence on the quality of its constituents, and the fundamental ceiling that no feature-space detector can identify a flipped sample whose features are entirely typical of its class. These motivate the weighted-voting extension proposed as future work.

6.12 Limitations of Existing Selected Defences

The existing defences each carry limitations confirmed by the results. Isolation Forest is efficient and scalable but may miss label-flipped samples because their features are unchanged. One-Class SVM is a strong boundary-based method but is sensitive to scaling and parameters and costly on large datasets. LOF is a strong local density-based detector but is computationally expensive and affected by high dimensionality. The Hybrid Ensemble was designed precisely to mitigate these individual weaknesses, at the cost of greater computational expense.

6.13 Importance of FNR in IDS Evaluation

The false negative rate is the metric most aligned with the security mission of an IDS, because a false negative is a missed attack and therefore a successful intrusion. The results show that the FNR can rise sharply even when accuracy falls only modestly, so an evaluation that reports accuracy alone would understate the security impact of poisoning. Foregrounding the FNR is therefore not a stylistic choice but a substantive one.

6.14 Impact of Resource Constraints

The resource constraints reduced the breadth of the study — a single dataset, a single classifier, a subset of the data, and one fully-evaluated attack family — but did not compromise the internal validity of the core comparison, which holds all factors constant across conditions. The constraints are most relevant to external validity, discussed next, and to the future work that would lift them.

6.15 Validity Threats

Internal validity: the use of a single random seed means the figures reflect one realisation; multiple seeds with confidence intervals would strengthen the comparison, though the large gaps at 20% are unlikely to be artefacts. *External validity:* the findings are specific to label flipping against a Random Forest on a 50,000-record CICIDS-2017 DDoS subset, with UNSW-NB15 used as supplementary corroboration. Therefore, the findings should not be assumed to generalise to the full CICIDS-2017 dataset, all UNSW-NB15 settings, neural networks, or other attack families without further study. *Construct validity:* the detectors' contamination parameters are aligned with the true poisoning rate, which is optimistic relative to a real deployment in which the rate is unknown.

6.16 Relationship to Prior Literature

The findings are consistent with the literature. The observation that poisoning degrades performance non-linearly and that the FNR is the most sensitive indicator aligns with the behavioural analyses of Paracha et al. [1,3]. The superiority of a density-based detector over global methods under heavy contamination is consistent with the general understanding of anomaly detection, and the value and limits of a simple voting ensemble echo both the ensemble-learning literature and the cautionary stance of Biggio and Roli [8] regarding

adaptive adversaries. The proposed weighted, attack-aware extension situates the work on a trajectory towards the attack-agnostic defences of Paracha et al. [2].

6.17 Reflection on the Attempted MLP Extension

The exploratory MLP trial, although not a primary result, offers a useful preliminary contrast with the Random Forest findings. The clean MLP performed strongly, with the lowest false negative rate of any MLP configuration, and poisoning at 10% degraded it as expected. However, unlike the Random Forest case, applying the defences to the poisoned MLP did not clearly improve detection and in several configurations increased the false negative rate relative to the undefended poisoned model. A plausible explanation is that the sanitisation stage removes samples and thereby reduces the volume of training data, to which a neural network — being more data-hungry and more sensitive to distributional shifts than a Random Forest — may be more vulnerable, so that the benefit of removing poisoned points is partly offset by the cost of a smaller and slightly altered training set. Within the MLP configurations, the Hybrid Ensemble again proved comparatively stable, outperforming MLP + Isolation Forest and MLP + One-Class SVM and approaching MLP + LOF, which is consistent with the stability observed for the Random Forest. The LOF duplicate-value warning encountered during the trial further illustrates the practical sensitivity of density-based detection to repeated or near-identical feature values. Taken together, these preliminary observations reinforce the decision to treat the MLP work as exploratory and to identify a full, multi-rate neural-network evaluation as an important avenue for future work.

6.18 Summary

This chapter discussed the research findings in relation to the research questions and evaluated the impact of label flipping data poisoning on machine learning-based intrusion detection. The results show that label flipping can reduce IDS reliability by increasing missed attacks, especially as the poisoning rate increases. The discussion also showed that anomaly-detection-based sanitisation can help reduce the effect of poisoned training data, although the effectiveness varies across techniques. LOF showed strong performance at higher poisoning levels, while the Hybrid Ensemble provided stable and competitive performance by combining Isolation Forest, One-Class SVM, and LOF through majority voting. The chapter also acknowledged the main limitations of the study, including the use of a 50,000-record subset of the CICIDS-2017 DDoS data, a single main classifier, one main attack type, and limited computational resources. These findings provide the basis for the conclusion and future work presented in the next chapter.

Chapter 7

Conclusion and Future Work

7.1 Summary of the Research

This dissertation evaluated the impact of label flipping data poisoning attacks on a Random Forest-based IDS using a representative subset of the CICIDS-2017 DDoS dataset as the main experiment, assessed three anomaly-detection defences — Isolation Forest, One-Class SVM, and Local Outlier Factor — and proposed and evaluated a Hybrid Ensemble defence based on 2-of-3 majority voting. A supplementary UNSW-NB15 experiment and an exploratory MLP extension were also completed. The study was conducted through controlled experimentation in which the poisoning rate and the choice of defence were varied while all other factors were held constant, with performance assessed through five metrics and particular emphasis on the false negative rate.

7.2 Key Findings

On the main CICIDS-2017 DDoS experiment, the clean baseline was perfectly separable (accuracy 1.0000, FNR 0.0000), reflecting the strong separability of DDoS from benign traffic on this subset. Label flipping degraded performance monotonically, most severely at 20%. Among the individual detectors, LOF delivered the strongest performance under severe poisoning, achieving the lowest FNR at both 10% (0.0084) and 20% (0.0428), while Isolation Forest was the most vulnerable at 20% (FNR 0.0693). The Hybrid Ensemble was the best configuration at 5% (FNR 0.0033) and remained competitive thereafter, but it did not surpass LOF at 10% and 20%; its principal merit is stability and the avoidance of worst-case collapse rather than uniform superiority. The supplementary UNSW-NB15 experiment corroborated this overall pattern. The analysis confirmed that the false negative rate is an indispensable metric, exposing security-critical degradation that accuracy alone

can conceal.

7.3 Contributions

The dissertation contributes a controlled, like-for-like comparison of three anomaly-detection defences against label flipping under identical conditions; a transparent Hybrid Ensemble defence whose stability and limitations are clearly characterised; an evaluation framework foregrounding the false negative rate; a reproducible implementation; and an honest, instructive account of conducting such a study under real computational constraints.

7.4 Limitations

The study is limited in several respects, all stated openly. First, computational constraints in the free Google Colab environment made the full experimental matrix infeasible, so the experiments were conducted on representative 50,000-record subsets of CICIDS-2017 and UNSW-NB15 rather than the full datasets. Second, only one attack family (label flipping) was fully implemented; outlier injection and backdoor attacks, although part of the broad approved plan, were not implemented and remain future work. Third, the main experiment uses a single base classifier (Random Forest); the MLP work is exploratory, single-rate, and not the main contribution. Fourth, the CICIDS-2017 DDoS baseline is near-perfect because DDoS and benign traffic are highly separable on this subset, so the meaningful signal is the relative degradation under poisoning rather than the absolute baseline. Fifth, the detectors' contamination parameters were aligned with the true poisoning rate, which is optimistic relative to a real deployment, and a single random seed was used. These limitations bound the generality of the findings and define the future work below.

7.5 Future Work

Several directions follow. First, a weighted or adaptive voting scheme could allow the Hybrid Ensemble to exploit the superiority of strong detectors such as LOF and surpass, rather than merely match, the best individual detector at higher poisoning rates. Second, the experiments should be repeated on the full CICIDS-2017 (across its multiple attack captures) and the full UNSW-NB15 dataset, using batch (incremental) learning, cloud computing, or a GPU-supported environment to lift the resource ceiling. Third, the evaluation should be extended to additional attack families (outlier injection and backdoor attacks) and to additional classifiers, including a full neural-network evaluation across all poisoning rates and

defences. Fourth, methods to estimate the contamination rate from data, rather than assuming it, would improve operational realism. Fifth, repeating the experiments across multiple seeds with significance testing would strengthen internal validity. Finally, defences explicitly designed against adaptive adversaries would address the most challenging threat model.

7.6 Final Conclusion

Machine learning has become indispensable to intrusion detection, but its dependence on trustworthy training data renders it vulnerable to data poisoning. This dissertation has shown, on the CICIDS-2017 DDoS data, that label flipping materially degrades a Random Forest IDS, that anomaly-detection defences substantially mitigate this degradation, that Local Outlier Factor is the most effective individual defence under severe poisoning, and that a simple Hybrid Ensemble offers valuable stability — being the best defence at low poisoning while remaining competitive at higher rates — without uniformly surpassing the strongest individual detector. The supplementary UNSW-NB15 experiment corroborated these conclusions. While the scope was narrowed by computational constraints and aligned with the approved ethics form, the core comparison is rigorous and internally consistent, and it provides a practical, security-aware contribution towards the broader challenge of securing machine learning against intelligent adversaries.

Bibliography

- [1] A. Paracha, J. Arshad, M. Ben Farah, and K. Ismail, “Deep behavioural analysis of machine learning algorithms against data poisoning,” *International Journal of Information Security*, 2024.
- [2] A. Paracha, J. Arshad, M. Ben Farah, and K. Ismail, “SecureLearn: An attack-agnostic defence for multiclass machine learning against data poisoning attacks,” *IEEE Trans. Information Forensics and Security*, 2025.
- [3] A. Paracha, “A behaviour-informed approach to mitigate data poisoning attacks for machine learning,” Ph.D. thesis, Birmingham City University, 2025.
- [4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] N. Moustafa and J. Slay, “UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” in *Military Communications and Information Systems Conf. (MilCIS)*, Canberra, Australia, 2015, pp. 1–6.
- [6] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in *Proc. 4th Int. Conf. on Information Systems Security and Privacy (ICISSP)*, 2018, pp. 108–116.
- [7] B. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” in *Proc. 29th Int. Conf. on Machine Learning (ICML)*, 2012, pp. 1807–1814.
- [8] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.
- [9] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Int. Conf. on Learning Representations (ICLR)*, 2015.
- [10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *Proc. 8th IEEE Int. Conf. on Data Mining (ICDM)*, 2008, pp. 413–422.

- [11] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
- [12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2000, pp. 93–104.
- [13] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,” in *Proc. IEEE Symp. on Security and Privacy (S&P)*, 2018, pp. 19–35.
- [14] L. Muñoz-González *et al.*, “Towards poisoning of deep learning algorithms with back-gradient optimization,” in *Proc. 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2017, pp. 27–38.
- [15] J. Steinhardt, P. W. Koh, and P. Liang, “Certified defenses for data poisoning attacks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3517–3529.
- [16] T. Gu, B. Dolan-Gavitt, and S. Garg, “BadNets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv:1708.06733v2, 2019.
- [17] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Appendix A

Code Overview

The implementation is contained in a single Jupyter/Google Colab notebook, `msc_final.ipynb`, organised into the following logical stages for the main CICIDS-2017 DDoS experiment: (i) loading the CICIDS-2017 DDoS capture (`Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv`) from Google Drive and drawing a 50,000-record subset with `random_state = 42`; (ii) preprocessing (stripping column-name whitespace, replacing infinite values with NaN, dropping missing values, converting the textual label to binary BENIGN/DDoS, separating the label, and applying `StandardScaler`); (iii) an 80/20 stratified train–test split; (iv) training the clean Random Forest baseline; (v) the label flipping attack at 5%, 10%, and 20%; (vi) the three anomaly detectors (Isolation Forest, LOF, One-Class SVM) with contamination aligned to the poisoning rate, and the removal of flagged samples; (vii) the Hybrid Ensemble majority-voting sanitisation (keep a sample if at least two of three detectors vote it clean) and retraining; (viii) evaluation of accuracy, precision, recall, F1-score, and FNR; (ix) generation of the accuracy, F1-score, FNR, and combined comparison figures; and (x) export of the result tables to CSV. The same pipeline was applied to a 50,000-record UNSW-NB15 subset as a supplementary experiment, and an exploratory MLP extension was run on UNSW-NB15 at 10% label flipping. No public repository URL was used; the submitted notebook `msc_final.ipynb` is provided with the dissertation artifact materials.

Appendix B

Runtime Constraints and Warnings

This appendix documents the implementation constraints encountered under limited computational resources. The principal constraint was prohibitively long execution time when the density- and boundary-based detectors were run repeatedly over large volumes of data, which motivated the representative 50,000-record subset. No fatal crash occurred during the final experiments; the one reproducible warning observed during execution is the LOF duplicate-value warning, documented in Appendix C. The constraint is therefore documented here in narrative form.

What was constrained: repeated execution of the anomaly-detection defences over large data volumes and, in the broader plan, additional datasets, classifiers, and attack families. **What occurred:** execution times that were impractical for the full experimental matrix in the free Colab environment, together with the non-fatal LOF duplicate-value warning. **How it affected the scope:** it necessitated the subset-based approach and the narrowing of scope to the CICIDS-2017 DDoS main experiment (with UNSW-NB15 supplementary). **Why this focus:** this dataset, classifier, and attack combination could be executed rigorously and repeatably within the available resources, enabling a controlled, internally consistent comparison that complies with the approved ethics form.

Appendix C

Attempted MLP Neural Network Extension

As an exploratory extension, a Multilayer Perceptron (MLP) was applied to the UNSW-NB15 subset under 10% label flipping, to provide a preliminary indication of whether the findings might generalise beyond the Random Forest classifier. This was a single-rate, exploratory trial conducted without the full defence matrix and is not presented as a primary result. The full numerical results are reported in Section 5.10 (Table 5.9); they are summarised here for completeness, together with the LOF warning encountered during the trial.

The clean MLP attained an accuracy of 0.9364 and a false negative rate of 0.0143; poisoning at 10% reduced accuracy to 0.9340 and raised the FNR to 0.0259. The defended MLP configurations did not clearly improve upon the clean MLP, and in several cases increased the FNR relative to the undefended poisoned model. Among the defended configurations, the Hybrid Ensemble (FNR 0.0333) outperformed MLP + Isolation Forest (0.0468) and MLP + One-Class SVM (0.0556), and was comparable to MLP + LOF (0.0324). These preliminary outcomes motivated treating the MLP work as exploratory and reserving a full neural-network evaluation for future work.

LOF duplicate-value warning. During the MLP extension, LOF produced the warning *“Duplicate values are leading to incorrect results. Increase the number of neighbours for more accurate results.”* This is a non-fatal warning indicating that duplicated or near-identical samples affected LOF’s nearest-neighbour density estimates; it is reported as an implementation limitation, with neighbourhood-size tuning or de-duplication noted as future work. The warning did not halt execution, and the exploratory MLP results were produced normally. The complete MLP outputs are recorded in the submitted notebook `m_sc_final.ipynb` and summarised numerically in Table 5.9.

A full neural-network evaluation across all poisoning rates and all defences is identified as future work.

Appendix D

CICIDS-2017 Main Experiment: Run Details

This appendix documents the main CICIDS-2017 DDoS experiment. The DDoS capture [6] was loaded from Google Drive, reduced to a representative 50,000-record subset (`random_state = 42`), cleaned (whitespace-stripped headers, infinite values replaced with NaN, missing rows dropped), binarised (`BENIGN` $\rightarrow$ 0, `DDoS` $\rightarrow$ 1), scaled with StandardScaler, and split 80/20 with stratification. A clean Random Forest baseline was trained, followed by label flipping at 5%, 10%, and 20% and the application of Isolation Forest, LOF, One-Class SVM, and the Hybrid Ensemble. The complete numerical results appear in Chapter 5 (Tables 5.1–5.6), and the executed run is contained in the submitted notebook `msc_final.ipynb` together with the exported result CSV files.

The full CICIDS-2017 dataset comprises several attack captures beyond DDoS; evaluating these additional captures is identified as future work.

Appendix E

Ethics Approval Summary

This research is a secondary-data computational study. The relevant details are as follows: student, Md Mahfuzur Rahman (student number 2431582); programme, MSc Cyber Security and Digital Forensics; supervisor, Dr Anum Paracha; project title, “Defending Against Data Poisoning in Machine Learning-Based Intrusion Detection Systems”; proposed start date, 10 April 2026; proposed end date, 9 June 2026.

The study involves no human participants and no interviews, questionnaires, focus groups, or audio/video recordings, and it collects no personal or confidential participant data. It uses only publicly available benchmark datasets (CICIDS-2017 and UNSW-NB15) together with computer-generated network-log files. All attacks are simulated offline against the researcher’s own models; no live system and no third-party network is targeted. The work is defensive and educational in purpose, and its ethical risk is assessed as low.

The selection of CICIDS-2017 as the main dataset was made specifically to remain within the approved ethics application. Narrowing the scope from the broad approved plan does not increase ethical risk; it reduces complexity while remaining entirely within the same approved research area and data sources. No separate ethics approval reference number was provided on the available form; the approval status and any associated reference number and date should be confirmed against the official approval record held by the supervisor before final submission.