algorithms

MDPI

Article

Predicting the Impact of Data Poisoning Attacks in Blockchain-Enabled Supply Chain Networks

Usman Javed Butt [1], Osama Hussien [2], Krison Hasanaj [2], Khaled Shaalan [1], Bilal Hassan [2] and Haider al-Khateeb [3,*]

[1] Faculty of Engineering and IT, British University in Dubai, Dubai 345015, United Arab Emirates; usman.butt@buid.ac.ae (U.J.B.); khaled.shaalan@buid.ac.ae (K.S.)
[2] Faculty of Engineering and Environment, Northumbria University, London NE1 8ST, UK; ossama.akram@northumbria.ac.uk (O.H.); krison.hasanaj@northumbria.ac.uk (K.H.); b.hassan@northumbria.ac.uk (B.H.)
[3] Cyber Security Innovation (C.S.I.) Research Centre, Operations & Information Management, Aston University, Birmingham B4 7ET, UK
* Correspondence: h.al-khateeb@aston.ac.uk

Abstract: As computer networks become increasingly important in various domains, the need for secure and reliable networks becomes more pressing, particularly in the context of blockchain-enabled supply chain networks. One way to ensure network security is by using intrusion detection systems (IDSs), which are specialised devices that detect anomalies and attacks in the network. However, these systems are vulnerable to data poisoning attacks, such as label and distance-based flipping, which can undermine their effectiveness within blockchain-enabled supply chain networks. In this research paper, we investigate the effect of these attacks on a network intrusion detection system using several machine learning models, including logistic regression, random forest, SVC, and XGB Classifier, and evaluate each model via their F1 Score, confusion matrix, and accuracy. We run each model three times: once without any attack, once with random label flipping with a randomness of 20%, and once with distance-based label flipping attacks with a distance threshold of 0.5. Additionally, this research tests an eight-layer neural network using accuracy metrics and a classification report library. The primary goal of this research is to provide insights into the effect of data poisoning attacks on machine learning models within the context of blockchain-enabled supply chain networks. By doing so, we aim to contribute to developing more robust intrusion detection systems tailored to the specific challenges of securing blockchain-based supply chain networks.

Keywords: blockchain; supply chain; machine learning; flipping; poisoning attacks

Citation: Butt, U.J.; Hussien, O.; Hasanaj, K.; Shaalan, K.; Hassan, B.; al-Khateeb, H. Predicting the Impact of Data Poisoning Attacks in Blockchain-Enabled Supply Chain Networks. *Algorithms* **2023**, *16*, 549. https://doi.org/10.3390/a16120549

Academic Editor: Yue Duan

Received: 14 October 2023
Revised: 15 November 2023
Accepted: 22 November 2023
Published: 29 November 2023

1. Introduction

In recent years, network intrusion detection systems (NIDSs) have become essential tools for securing computer networks, especially in blockchain-enabled supply chain networks. These systems often rely on machine learning models to enhance their effectiveness. However, these models are vulnerable to data poisoning attacks, compromising their accuracy and posing significant security risks. This research aims to conduct an experimental assessment of the impact of data poisoning attacks on the machine learning models used within NIDSs.

The ever-expanding realm of computer networks, coupled with the proliferation of applications across them, underscores the burgeoning significance of network security. Intrusion detection systems (IDSs) have evolved into specialised instruments proficient in meticulously identifying network anomalies and potential attacks, a trend that has recently garnered prominence. The intrusion detection domain has traditionally centred on two fundamental techniques: anomaly-based and misuse-based. While misuse-based detection is

preferred in commercial applications due to its predictability and heightened accuracy, academic research frequently champions anomaly-based detection for its theoretical potential in countering novel and unprecedented attacks.

Data poisoning attacks represent a form of adversarial assault on machine learning models, where malicious actors manipulate training data to compromise the model's accuracy or induce misclassifications. In the specific context of network intrusion detection systems within blockchain-enabled supply chain networks, poisoning attacks assume a critical role by potentially enabling attackers to circumvent the system's detection mechanisms. These poisoning attacks come in various forms, including label flipping and distance-based flipping attacks, all of which have demonstrated their efficacy in undermining the integrity of machine learning models.

Our research's main goals, objectives, and contributions are as follows:

- A range of machine learning models, including logistic regression, random forest, SVC, and XGB Classifier, were systematically evaluated for network intrusion detection in blockchain-enhanced supply chain networks
- Rigorous evaluation of each model incorporated metrics such as their F1 Score, confusion matrix analysis, and accuracy assessments.
- The robustness of each model against data poisoning attacks was tested by subjecting them to various scenarios, such as no attack, random label flipping with 20% randomness, and distance-based label flipping with a 0.5 distance threshold.
- An eight-layer neural network was also experimented with, and its performance was assessed using accuracy and a classification report library.
- Comprehensive insights were provided into the effects of data poisoning attacks on machine learning models and their implications for network security.
- This research contributed to developing more robust and secure network intrusion detection systems for blockchain-enabled supply chain networks.

The emphasis on the intersection of data poisoning attacks and supply chain networks underscores the importance of securing these pivotal components within contemporary information systems. In the context of blockchain-enabled supply chain networks, various machine learning models, including logistic regression, random forest, SVC, and XGBClassifier, have been harnessed to identify network anomalies and potential attacks. The assessment of these models has encompassed performance metrics such as F1 Score, confusion matrix, and accuracy. Additionally, a neural network model was meticulously trained using the KDDCUP'99 dataset, followed by a comprehensive evaluation using accuracy and F1 Score metrics. To gauge the resilience of these models in the face of data poisoning attacks, they were subjected to two specific scenarios: random label flipping with a randomness factor of 20% and distance-based flipping attacks characterised by a distance threshold set at 0.5. This multifaceted analysis endeavours to provide insights into these models' predictive capacity and susceptibility to adversarial data manipulation within blockchain-enhanced supply chain networks.

Blockchain-enabled supply chain networks represent intricate systems encompassing multiple stakeholders, including suppliers, manufacturers, distributors, retailers, and customers, as shown in Figure 1. These entities engage in diverse processes, such as production, transportation, inventory management, and demand forecasting, all collectively facilitating the efficient and timely delivery of goods and services to customers. However, these supply chain networks face many challenges and risks, including disruptions, uncertainties, cyberattacks, and environmental concerns. One vital challenge pertains to safeguarding the network's security against malicious intruders who may seek to compromise, sabotage, or pilfer sensitive information or resources from within the network.

These attacks are especially challenging to detect and prevent in blockchain-enabled supply chain networks, where the network data are distributed and decentralised across multiple nodes [1]. Data poisoning attacks can have serious implications for network security and business profitability, as they can compromise the detection capabilities of NIDSs, allow malicious intrusions to go unnoticed, and disrupt the normal operations of the supply chain

network [1]. Therefore, developing effective countermeasures to mitigate the impact of data poisoning attacks on NIDSs in blockchain-enhanced supply chain networks is essential.

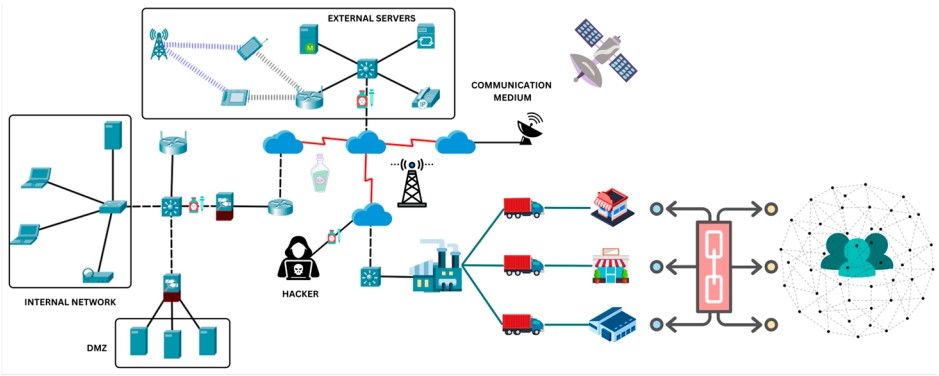


Figure 1. The supply chain network.

Several methods can be investigated to defend against these attacks, such as data sanitisation, robust aggregation, and anomaly detection. However, these methods are beyond the scope of this paper and have been left for future work. Future work could also explore other types of data poisoning attacks, such as adding malicious samples, injecting noise, or other methods mentioned in this research, and evaluate their effects on different machine learning models and datasets. Moreover, future work could investigate the impact of data poisoning attacks on the security and efficiency of blockchain-enabled supply chain networks in greater detail and develop more robust and secure network intrusion detection systems for this domain.

This research endeavours to investigate the susceptibility of machine learning models, including logistic regression, random forest, SVC, XGB Classifier, and an eight-layer neural network, to data poisoning attacks within network intrusion detection systems (NIDSs) deployed in blockchain-enhanced supply chain networks. By systematically evaluating these models in various attack scenarios and assessing their performance using metrics like F1 Score and accuracy, we aim to provide nuanced insights into the impact of data poisoning on the robustness and security of NIDSs. Our findings contribute to developing more resilient intrusion detection systems tailored to the challenges presented by blockchain-enabled supply chain networks. The supply chain model adopted in this research is a generic network model and applies to all application domains.

The paper is organised as follows: Section 2 delves into related work and research implications, encompassing an extensive literature review of the recent and pertinent contributions in the field. Section 3 offers an in-depth exploration of the selected dataset and the rationale behind its selection. Section 4 presents the experimental design, outcomes, and an evaluation of the impact of data poisoning attacks on the machine learning models. Lastly, Section 5 concludes the paper, opening a discussion on potential future endeavours to enhance and expand upon the findings and results obtained in this study.

2. Related Work

Data poisoning attacks on machine learning models have been studied in various domains, such as computer vision, natural language processing, and recommender systems [2,3]. For instance, the paper by [4] proposes a low-cost data poisoning attack algorithm named AttackRegion-UCB (AR-UCB) within the Attack against Federated Learning-based Autonomous Vehicle (ATT_FLAV) framework. This label flipping attack algorithm offers a unique dynamic black-box target attack for nonlinear regression models updated through federated learning. AR-UCB is designed to maximise attack rewards in rounds of continuous updates, with limited data and model output information available in each round. The research validates the effectiveness of the proposed attack, which demonstrates sequential improvements in attack rewards and can defend against classical aggregation schemes.

Additionally, an innovative framework for conducting online data poisoning attacks against smart-grid cyber-physical systems, based on the online regression task model, was introduced by [5]. The primary goal of this framework is to manipulate the model by gradually contaminating the incoming data stream. Furthermore, it offers a method for choosing data points depending on the sample loss within this framework. Experiments were carried out on an edge device utilising a simulated data stream created from offline open datasets relevant to the smart grid to evaluate the effectiveness of the proposed algorithms. The experimental findings show that this approach can increase the average attack efficacy by more than 1.23 times while reducing time overhead by more than 50%. These results highlight how crucial it is to protect against poisoning attacks in the context of smart grid security. The paper focused on popular online prediction models appropriate for edge intelligence applications and real-time, resource-constrained settings.

However, the application of machine learning in cyber security, especially in network intrusion detection systems (NIDSs), poses unique challenges and opportunities for both attackers and defenders [6]. NIDSs monitor network traffic and detect malicious activities, such as denial-of-service attacks, port scanning, malware infection, etc. NIDSs can be classified as signature-based or anomaly-based [7]. Signature-based NIDSs rely on predefined rules or patterns to identify known attacks, while anomaly-based NIDSs use machine learning models to learn the normal behaviour of network traffic and flag any deviations as anomalies [6]. Machine learning models can improve the accuracy and efficiency of NIDSs by automatically adapting to changing network conditions and discovering new attack patterns. However, machine learning models are also vulnerable to data poisoning attacks, which involve tampering with the training data to degrade the performance or compromise the integrity of the models [2,6].

The blockchain operates more seamlessly, providing increased security for all participants through transparency. As a new block is added, each node has a copy of the updated chain, which could improve transparency in many fields. The study in [8] presents a three-layer supply chain model that incorporates both RFID and blockchain technologies. The model focuses on the reverse flow of materials and demonstrates that employing blockchain-based RFID technology was profitable for the system in all scenarios. Discrepancies and holding costs particularly influenced the extent of profit increase, as these two parameters were most affected by the combination of the RFID and blockchain. The study highlights the benefits of adopting RFID and blockchain technology in supply chain operations. However, they emphasise the importance of addressing high misplacement rates within the system. It is shown that using RFID and blockchain technology can lead to a significant profit increase of up to 61% in the supply chain, even in high discrepancies. Consequently, RFID and blockchain technology offer the potential to enhance profitability without being significantly impacted by increased discrepancies in the supply chain.

Data poisoning attacks can have various goals, such as reducing the overall detection accuracy, causing targeted misclassification or bad behaviour, and inserting backdoors or neural trojans [2,3]. Data poisoning attacks can be classified into two types: indiscriminate and targeted. Indiscriminate attacks aim to degrade the model's overall performance by randomly changing the labels or features of some training examples. Targeted attacks aim to manipulate the model's behaviour on specific inputs or outputs by carefully crafting poisoned examples close to the decision boundary or greatly influencing the model [2,3]. Data poisoning attacks on machine learning models have been studied in various domains, such as computer vision, natural language processing, recommender systems, etc. However, the application of machine learning in cyber security, especially in network intrusion detection systems (NIDSs), poses unique challenges and opportunities for both attackers and defenders.

Previous research by [9] researched poisoning attacks and defences in machine learning by reviewing over 100 articles published in the previous years. By designing a framework that categorises threat models, attacks, and organising existing defences, they also maintained that their systematisation applies to attacks and defences in other data modalities

even though their primary focus was on computer vision applications. In addition to reviewing the historical development of machine learning models and highlighting present and future obstacles, the study outlined realistic scenarios for launching attacks on them. The research also offered guidance for understanding and defending against these attacks and insights into developing reliable machine-learning models resistant to malicious users.

Researchers in [10] used two AML attack types: poisoning and evasion. Poisoning attacks modify the training data to introduce errors or biases in the model. Evasion attacks craft adversarial examples that can fool the model at the time of inference. The article evaluates the impact of these attacks on four popular machine learning models: logistic regression, support vector machine, random forest, and a neural network. It uses two benchmark datasets: NSL-KDD and CICIDS2017. It measures the performance of the models using four metrics: accuracy, precision, recall, and F1 Score. Researchers in [11] evaluated the robustness of federated learning-based network intrusion detection systems under data poisoning attacks by malicious clients. They proposed a new attack method called PT-GAN and a new defence method based on poisoned sample detection.

Several mitigations and defences against data poisoning attacks were proposed in [12]; the researchers proposed a defence mechanism independent of the specific type of poisoning attack, called De-Pois. It leverages a mimic model that approximates the behaviour of the target model when trained on clean samples. It employs Generative Adversarial Networks (GANs) to augment the training data and train the mimic model. It detects poisoned samples by measuring the prediction discrepancy between the mimic and target models. Other researchers [6] propose a model-level defensive mechanism based on poisoned model detection and a data-level defensive mechanism based on poisoned data detection. The proposed model-level defence boosts its detection accuracy by up to 48% under the poisoning attacks on UNSW-NB15 dataset and 36% on CICIDS2018 dataset, and the proposed data-level defence further improves its detection accuracy by up to 13% on CICIDS2018 dataset.

Another study [13] discusses a novel approach to protect against data poisoning attacks that can be used in NIDSs. The discussed approach, DPA-FL, is based on federated learning, allowing multiple participants to collaboratively train a global model without sharing their local data. DPA-FL employs a two-phase strategy to identify and eliminate the attackers who inject poisoned data into the federated learning process. Using DPA-FL, NIDSs can achieve high detection accuracy and robustness against poisoning attacks.

Our research is related to recent studies on data poisoning attacks and their defences for NIDSs. Whereas previous studies by [14,15] proposed novel defence methods based on poisoned sample detection and training data sanitisation, respectively, for federated learning-based NIDSs and machine learning models in NIDSs. However, our research differs from theirs in several aspects. First, this research focuses on distance-based label flipping attacks, which are more realistic and challenging than random label flipping attacks, as they target the most influential samples for the classifier. Second, this research use several machine learning models, including logistic regression, random forest, SVC, and XGB Classifier, to evaluate the effect of these attacks on different types of classifiers. Third, it uses the KDDCUP'99 dataset, a widely used benchmark for NIDSs, to compare our results with previous studies that have also used this dataset.

Targeted label flipping, random label flipping, and random input data poisoning are the three types of data poisoning attacks that are compared practically in a study by [16], emphasising how they affect federated learning environments. A novel data poisoning technique was proposed by inverting the loss function of a benign model. This inverted loss function produces malicious gradients almost in contradiction to the minima throughout each Stochastic Gradient Descent (SGD) phase. Once created, these malicious gradients inject poisoned labels into the dataset. Utilising three distinct datasets—MNIST, Fashion-MNIST, and CIFAR10—the attack was evaluated and contrasted with other known data poisoning techniques. The findings indicated that this novel attack could be 1.6 times more

successful than targeted attacks and 3.2 times more effective than random poisoning attacks in some situations, especially when used with federated machine learning.

Two more papers that are related to distance-based label flipping [17,18] propose a white-box, realisable poisoning attack that reduces the original model accuracy from 95% to less than 50% by generating mislabelled samples in the vicinity of a selected subset of training points, which is similar to the idea of distance-based label flipping. It proposes a novel community detection algorithm that uses distance-based label propagation, which can be applied to identify the nodes most vulnerable to label flipping attacks in social networks. Another relevant paper for our research is [18], which proposes a network intrusion detection model based on a two-layer convolution neural network for handling imbalanced datasets, a common challenge in NIDSs. These papers can provide some insights into the design and analysis of data poisoning attacks and their defences for NIDSs.

This paper has several research implications for network anomaly detection and data poisoning attacks. First, it provides empirical evidence of the impact of label flipping attacks on network anomaly detection models, a novel and realistic type of data poisoning attack that has not been extensively studied before [19]. Second, it compares the performance and robustness of two neural network models with the same architecture on a non-poisoned and a poisoned dataset, which can serve as a baseline for future studies on this topic; it highlights the resilience of some machine learning models, to some extent, to label flipping, as they can still achieve moderate performance despite the noise in the training data, which could motivate further research on understanding and enhancing the robustness of neural network models to data poisoning attacks. Table 1 below shows the limitations and research gaps of the existing literature on data poisoning attack frameworks.

Table 1. Limitations and research gaps of related works.

Reference	Year	Limitations and Research Gaps
[20]	2023	Depending on the loss function (f) the attacker would have to control a larger number of clients with an increasing number of participating devices to uphold the value of f. The research does not directly tackle this aspect, as it does not alter the loss function but rather assesses the impact of data poisoning attacks on existing machine learning models.
[17]	2022	The research does not include FT-GCN optimisation or the development of an upgraded traffic graph with weighted edges denoting the degree of correlation among traffic flows. Moreover, the research does not explore the multi-classification of harmful internet traffic flows at this point; while the research does not engage with these specific aspects, it concentrates on evaluating the models' robustness against data poisoning attacks in the existing intrusion detection context.
[18]	2023	The study focuses on improving the CSK-CNN model's overall classification performance in the context of intrusion detection datasets. However, it does not detail how well certain anomalous categories like Dos, Web Attack Brute Force, and others, are classified. As such, there may not have been a comprehensive assessment or optimisation of the model's performance regarding these particular attack categories. Although this research does not offer an optimisation or detailed assessment of the CSK-CNN model's performance for these specific categories, it focuses on the broader impact of data poisoning attacks on various machine learning models.
[9]	2023	It does not delve into advanced online models like deep learning and neural networks. As a result, the findings and methods proposed in the study may not be readily applicable to these models. These more complex models might pose distinct challenges and necessitate unique attack and defence strategies. The research in this paper uses several machine learning models such as logistic regression, random forest, support vector classifier, XGBoost, and a deep neural network with eight layers. Also, it focuses on the impact of data poisoning attacks on these models.

While this research provides valuable contributions to understanding the effects of data poisoning attacks on intrusion detection models, it does not directly overcome the specific limitations and research gaps mentioned in the references [9,17,18,20], as its focus lies in a different domain of investigation; the main focus of this research is on two types of attacks: label flipping and the distance-based flipping. Label flipping involves randomly changing the labels of some training examples from benign to malicious or vice versa.

Distance-based flipping involves changing the labels of some training examples based on their distance to the decision boundary of a classifier.

3. Dataset

Its prominence and extensive utilisation underpins the selection of the KDDCUP'99 dataset for our research in the domain of intrusion detection. This dataset's vast repository of instances and encompassing feature set, coupled with its coverage of diverse attack types and scenarios, renders it a pivotal resource for this field. It is a fundamental benchmark, facilitating the rigorous evaluation and comparative analysis of various intrusion detection methodologies and systems. It is imperative to acknowledge that the KDDCUP'99 dataset does not lack limitations and challenges, as this research will discuss later. These encompass the antiquity of the underlying network environment, an inherent skew in class distribution, the presence of superfluous and non-informative features, and the prevalence of mislabelled instances.

It is also important to note that contemporary datasets, such as MAWILab, Malware Training Sets, ADFA Intrusion Detection Datasets, CTU-13, Aposemat IoT-23, and EMBER, are available for intrusion detection research. These datasets encapsulate real-time network traffic and contemporary attack patterns, which may introduce novel complexities and research opportunities. Nevertheless, these datasets, distinguished by unique formats, structures, and characteristics, may necessitate distinct methodologies and techniques for processing and analysis. Moreover, certain datasets are not readily accessible to the public, potentially limiting their utility and reproducibility. Given these considerations, the KDDCUP'99 dataset is a classic and universally recognised resource for intrusion detection research. It is our belief that this research offers substantive insights and contributions to the realm of intrusion detection and the application of machine learning in fortifying network security.

This dataset contains information about different types of connections to a computer network, such as normal connections and malicious attacks. The dataset has the following characteristics:

- It has 41 features, such as duration, protocol type, service, source bytes, destination bytes, etc., that describe each connection.
- It has 4.9 million records, of which 10% are available as a subset for training and testing purposes.
- It has 23 types of attacks, such as denial of service, probing, user to root, remote to local, etc., representing different network intrusions.
- It is based on a simulated military network environment that mimics real-world scenarios.

The KDDCUP'99 dataset was chosen for this research because it offers a large and diverse set of data that covers various aspects of network activity and security. It also has a well-defined task and evaluation criteria that can be used to compare different methods and models. However, this dataset also has some limitations and challenges, such as:

- It is outdated and does not reflect the current state of network traffic and attacks that are more complex and sophisticated.
- It has some redundant and irrelevant records that may affect the quality and accuracy of the models.
- It has imbalanced classes that may cause some models to be biased or overfitting.

The main objective of this research is to study the effects of data poisoning attacks on several machine learning models commonly used for NIDSs. Data poisoning attacks are adversarial attacks that corrupt the training data of machine learning models, resulting in inaccurate or malicious predictions. The KDDCUP'99 dataset contains a large and diverse set of network connections with different types of intrusions.

The dataset was created by processing the tcpdump portions of the 1998 DARPA Intrusion Detection System (IDS) Evaluation dataset, which was prepared and managed by MIT Lincoln Lab. The original raw data consisted of about four gigabytes of compressed binary

tcpdump data from seven weeks of network traffic. These data were processed into about five million connection records, each corresponding to a single connection. The features include the basic features of individual TCP connections (such as the duration, protocol type, service type, number of bytes transferred), content features within a connection (such as the number of failed login attempts, number of file creations, number of root accesses), and traffic features computed using a two-second time window (such as the number of connections to the same host, the number of connections to the same service).

The label indicates whether the connection is normal or an attack, and if it is an attack, what type of attack it is. The dataset provides a taxonomy of 22 attack types that fall into four main categories: denial of service (DoS), probing, user to root (U2R) and remote to local (R2L). DoS attacks make some computing or memory resources too busy or too full to handle legitimate requests. Probing attacks are those that scan a network of computers to gather information or find known vulnerabilities. U2R attacks allow a local user to gain root or super-user privileges. R2L attacks allow an attacker to gain access to a machine remotely.

The catplot in Figure 2a shows the distribution of the attack traffic in the network anomaly dataset. The *x*-axis represents the type of traffic, which can be either normal or an attack. The *y*-axis represents the count of instances of each type. The figure shows that there are approximately 67,000 instances of "Normal" and around 58,000 instances of "Attack", which means that most of the packets in the normal traffic are not anomalous, while most of the packets in the attack traffic are anomalous. This is expected since the normal traffic represents the baseline behaviour of the network, while the attack traffic represents any deviation from that behaviour due to malicious activities,

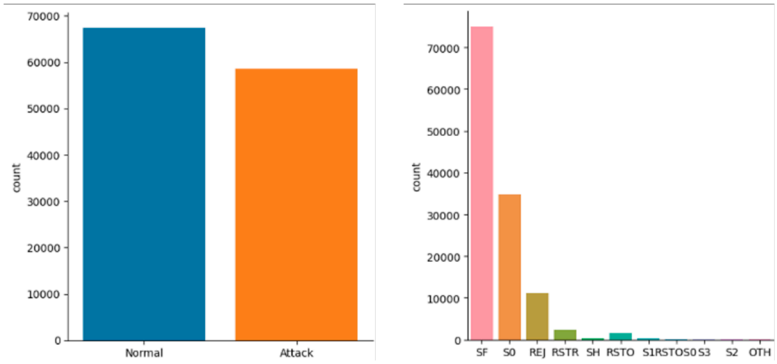


Figure 2. (**a**) Dataset Labels catplot; (**b**) distribution of the flag columns in the dataset.

Figure 2b shows the distribution of the flag in the network anomaly dataset. The x-axis represents the type of flag, which can be one of 11 categories: SF, S0, REJ, RSTR, SH, RSTO, S1, RSTOS0, S3, S2, or OTH. The y-axis represents the count of instances of each type, it can be seen that the SF flag has the most instances, which means that most of the packets in the dataset have a normal connection termination. This may suggest that the network is functioning properly and most of the connections are completed successfully, meanwhile the S0 flag has the second most instances, which means that some of the packets in the dataset have no response from the destination host. This may indicate some network problems or attacks that prevent the connections from being established. The REJ flag has the third most instances, meaning that some packets in the dataset explicitly reject the destination host. This may imply that some security measures or policies block or filter some connections.

The other flags have much fewer instances, meaning they are rare or uncommon in the dataset. This may reflect that they represent specific or unusual situations or behaviours in the network.

The heatmap in Figure 3 shows the distribution of the flag feature in the network anomaly dataset. The flag feature indicates the connection status, which can be one of

11 categories: SF, S0, REJ, RSTR, SH, RSTO, S1, RSTOS0, S3, S2, or OTH. The x-axis represents the flag type, and the y-axis represents the count of instances of each type. The plot reveals that the SF flag has the most instances, which means that most of the packets in the dataset have a normal connection termination. This may suggest that the network functions properly and that most connections are completed successfully.

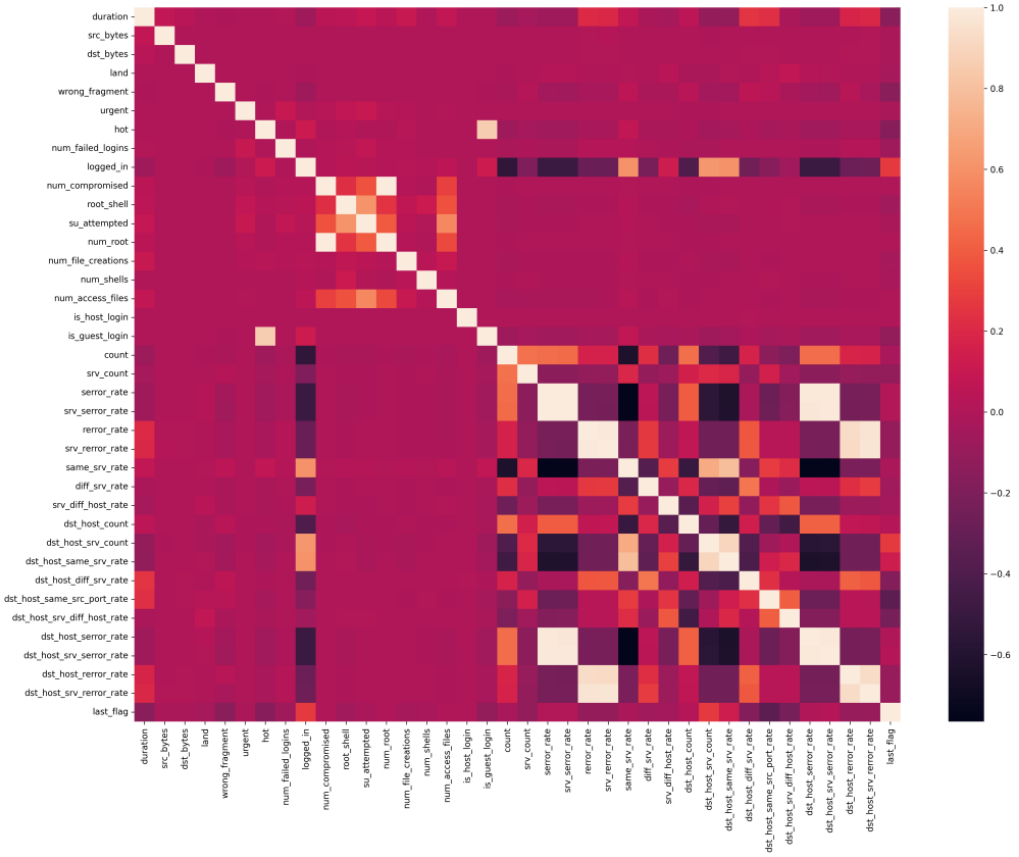


Figure 3. Dataset correlation matrix heatmap.

On the other hand, the S0 flag has the second most instances, which means that some of the packets in the dataset have no response from the destination host. This may indicate some network problems or attacks that prevent the establishing of connections. The REJ flag has the third most instances, which means that some of the packets in the dataset have an explicit rejection from the destination host. This may imply that some security measures or policies block or filter some connections. The other flags have fewer instances, meaning they are rare or uncommon in the dataset. This may reflect that they represent specific or unusual situations or behaviours in the network.

The bulk of the entries in the collection appear to be of the flag type SF, which has a count of over 70,000. The second most common flag type is S0, with around 35,000 records, followed by REJ, with approximately 10,000 entries. The remaining flag kinds have fewer than 10,000 records.

The graph in Figure 4 shows the correlation matrix of the numerical features in the network traffic dataset. The correlation matrix is a table that shows the pairwise correlation coefficients between each pair of features, ranging from -1 to 1. A correlation coefficient close to 1 means a strong positive correlation, a correlation coefficient close to -1 means a strong negative correlation, and a correlation coefficient close to 0 means no correlation. The plot reveals that some features are highly correlated with each other, such as src_bytes and dst_bytes, srv_count and count, and same_srv_rate and dst_host_same_srv_rate. These features may contain redundant information and may not be useful for network anomaly detection. On the other hand, some features are weakly correlated or uncorrelated, such as duration and flag,

wrong_fragment and logged_in, and num_compromised and num_root. These features may contain unique or diverse information useful for network anomaly detection.

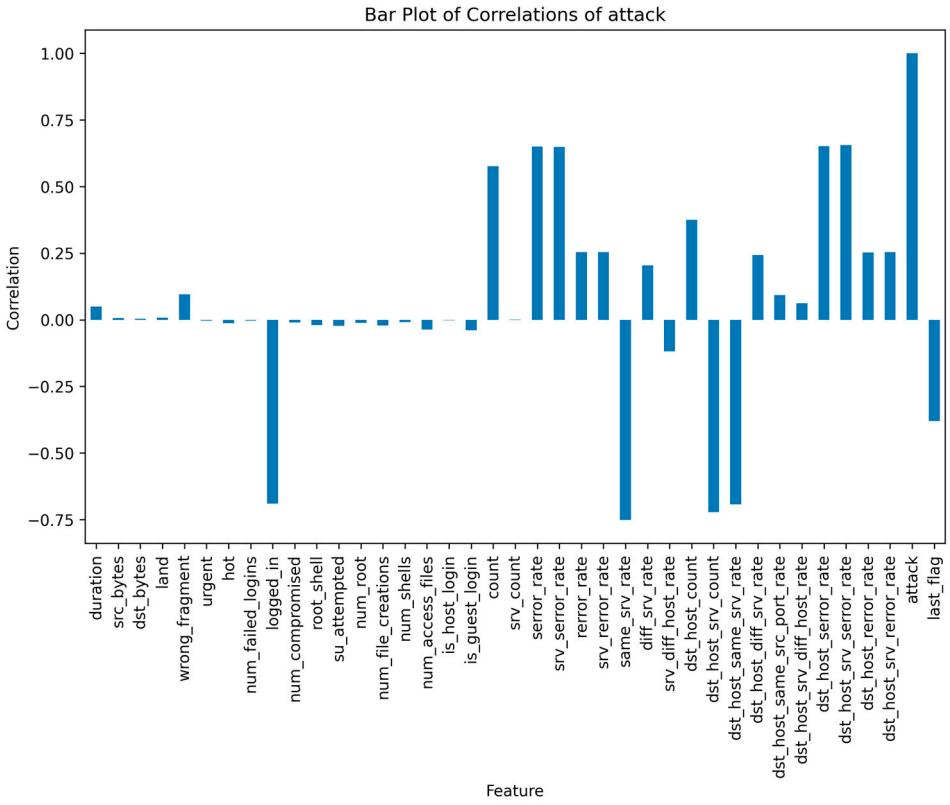


Figure 4. Feature correlation with the labels in the dataset.

Figure 4 also shows that some features have a high positive correlation with the attack label, meaning that they are more likely to indicate an attack when they have higher values. These features include logged_in, count, serror_rate, srv_serror_rate, same_srv_rate, dst_host_srv_count, dst_host_same_srv_rate, dst_host_serror_rate, and dst_host_srv_serror_rate. Some features have a low or negligible correlation with the attack label, meaning they are not very useful for distinguishing between normal and attack activities.

4. Experiment

This research will use the KDDCUP'99 network anomaly detection dataset, widely used as one of the few publicly available datasets for network-based anomaly detection systems [21,22]. The enormous growth of computer network usage and the huge increase in the number of applications running on top of it make network security increasingly important. All computer systems suffer from security vulnerabilities, which are both technically difficult and economically costly to be solved by the manufacturers. Therefore, the role of intrusion detection systems (IDSs) as special-purpose devices to detect anomalies and attacks in the network is becoming more important.

4.1. Experiment Design

The aim of this experiment is to evaluate the resistance of each model to the attacks and suggest some mitigation strategies by applying various machine learning models to the dataset, such as logistic regression, random forest, support vector classifier, XGBoost and an artificial neural network with eight layers. These models are trained and tested with and without data poisoning attacks, which are random label flipping and distance-based label flipping, as shown in Figure 5.

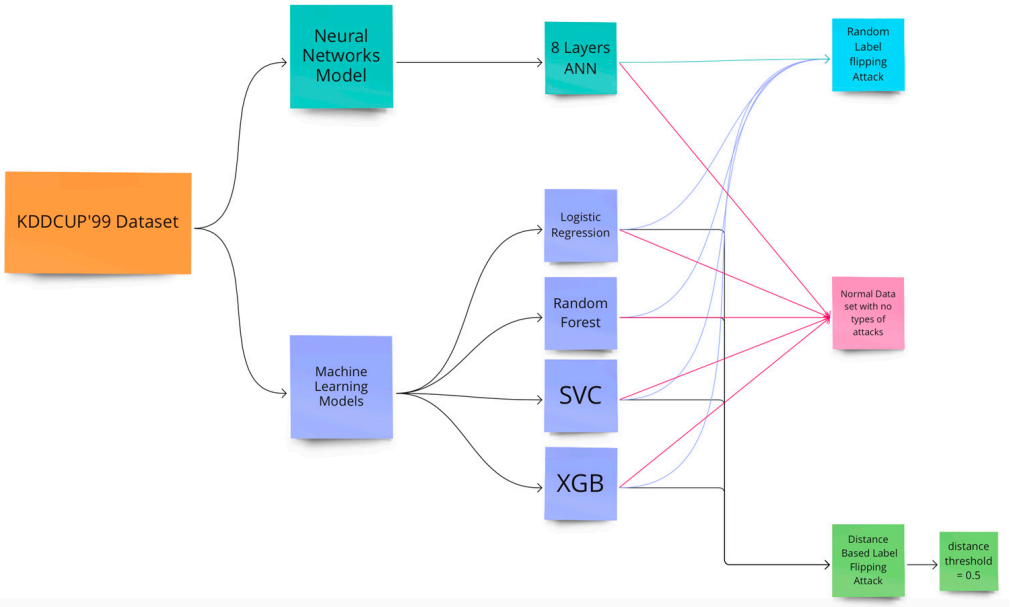


Figure 5. Experiment design map.

One of these models will use logistic regression, which was selected as a baseline linear classifier widely used for binary classification problems [23]. The support vector classifier is also used as an example of nonlinear classifiers that can capture complex patterns in the data [24]. Furthermore, random forest and XGBoost were also employed. By comparing these models, the main aim is to investigate how different learning algorithms react to data poisoning attacks.

Random label flipping is a simple and naive way of poisoning the training data by randomly selecting some instances and changing their labels to any other class [5]. This can introduce noise and confusion to the machine learning algorithm, especially if the noise level is high or the classes are imbalanced [25]

Distance-based label flipping is a more sophisticated and targeted way of poisoning the training data by selecting some instances that are close to the decision boundary of the machine learning algorithm and changing their labels to the opposite class. This can create more damage and misclassification than random label flipping, as it exploits the vulnerability of the machine learning algorithm to instances that are hard to classify [26].

4.2. The Experiment's Models' Training and Evaluation

In this section, a data poisoning attack on the machine learning model was conducted and compared with the non-poisoned models; data poisoning is a type of adversarial attack that aims to manipulate the training data of a machine learning algorithm to degrade its performance or accuracy [27,28].

The experiment was conducted using several machine learning models, and each model was evaluated using its F1 Score, confusion matrix, and accuracy. The algorithms include logistic regression, random forest, support vector machine and XGBoost, an eight-layer neural network model was also tested, which was trained using the Adam optimiser and binary cross-entropy loss function. A random label flipping with 20% randomness was introduced, and the dataset was split into training and test sets using an 80–20 split. The training set contains 125,973 records, and the test set contains 25,194 records.

Figure 6 shows the decision boundaries of different machine learning models for the network anomaly detection task. The function scaled the data using a StandardScaler from sklearn and applied PCA to reduce X to two components. The output showed that different models had different complexity and accuracy in their decision boundaries. Some models, such as random forest, SVC and XGB had curved and complex decision boundaries that could capture more nonlinear patterns and variations. Other models like the logistic

regression had simpler decision boundaries that could capture only the basic trends and differences. The output also showed that different models had different prediction errors and uncertainty. Some models, such as the random forest and SVC, had fewer misclassified data points in the wrong colour region.

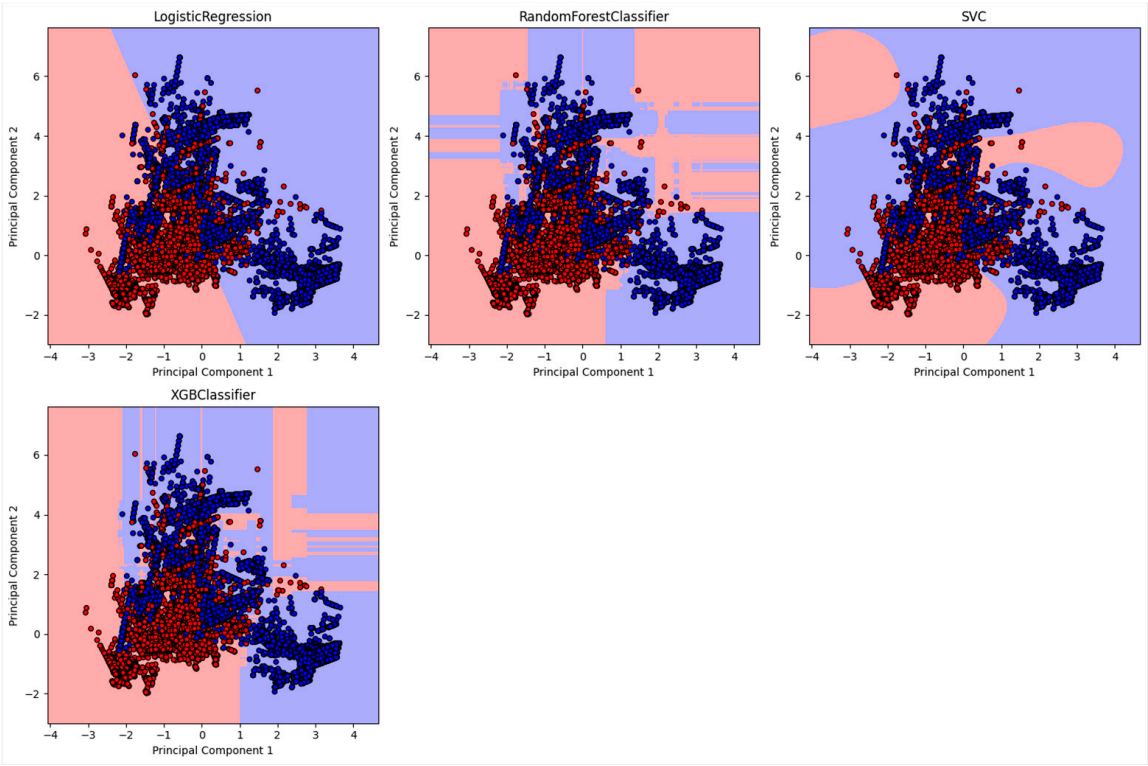


Figure 6. Non-poisoned models training.

The results show the performance metrics of seven different machine learning models when trained on a dataset without poisoned samples. The models used are logistic regression, Figure 7a; random forest, Figure 7b; support vector classifier (SVC), Figure 7c; and XGBoost, Figure 7d.

All models have reasonable accuracy, as shown in Table 2, with the lowest being random forest at 0.825 and the highest being 0.858 with the logistic regression model. The models show F1 scores ranging from 0.822 with SVC to 0.868 with logistic regression. In terms of precision, all models have values higher than 0.9. This indicates that the models have low false positive rates, and their recall values range from 0.72 with XGB to 0.82 with the logistic regression.

Table 2. Model evaluation metrics without data poisoning.

	LR	Random Forest	SVC	XGB
Acc	0.85,81884315117104	0.82,52306600425834	0.84,6829607723754	0.82,94889992902768
F1	0.86,82464454976303	0.82,2841726618705	0.84,6829607723754	0.82,78857347541864
Prec	0.92,14485654303709	0.97,26799192090996	0.95,81734081291211	0.97,30554678454899
Recall	0.82,08524896750565	0.71,30055326112367	0.75,86690563391257	0.72,04083222940856

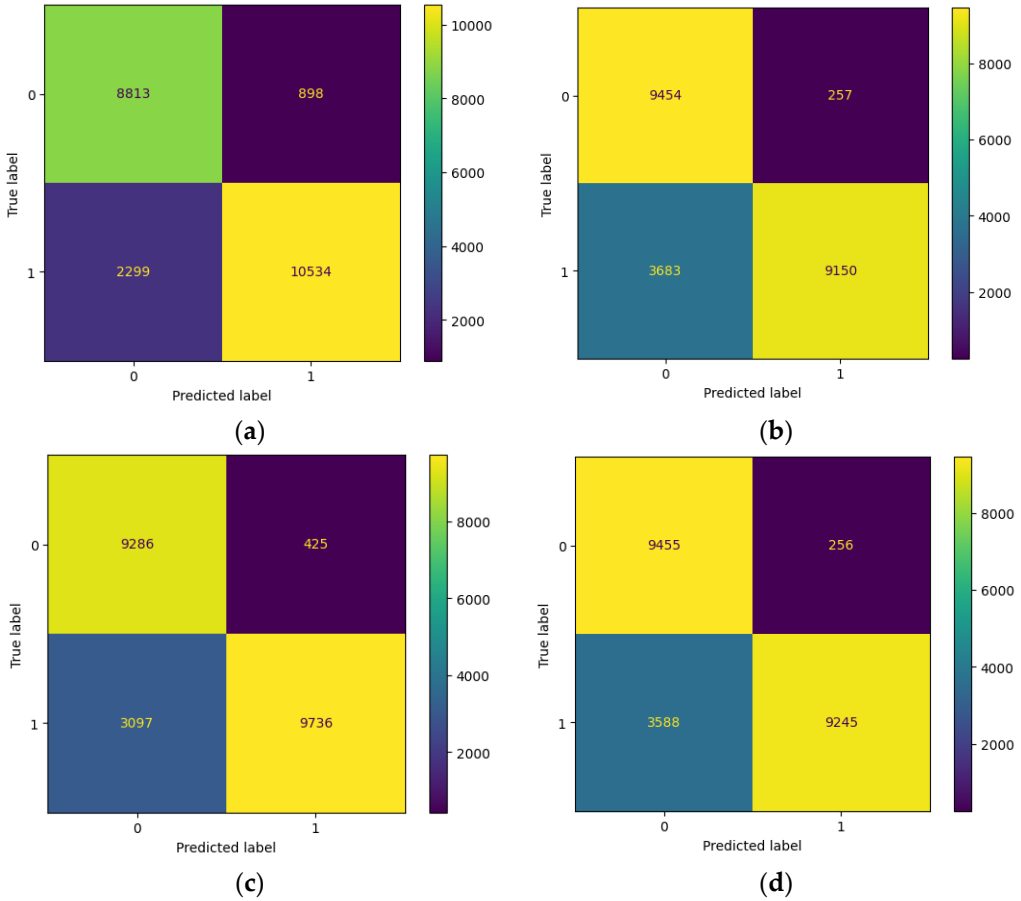


Figure 7. (**a**) Logistic regression; (**b**) random forest; (**c**) SVC; and (**d**) XGB.

4.2.1. Evaluation of Machine Learning with a Random Label Flipping Attack

Comparing these results with the previous analysis, it can be observed that the performance metrics of all models have decreased significantly when trained on the poisoned dataset generated using the label flipping attack, as shown in Table 3. This indicates that the attack successfully reduces the models' performance on the original dataset. Specifically, all models' F1 Scores, precision, recall, and accuracy have decreased compared to their original values. The extent of the decrease varies depending on the model, with some models being more resilient to the attack than others, as shown in Figure 8a–d and Table 3; for example, the logistic regression model appears to be the most resilient, with the smallest decrease in performance, as shown in Figure 8a and Table 4.

Table 3. Model evaluation metrics with random label flipping attack.

	LR	Random Forest	SVC	XGB
Acc	0.8698988644428672	0.8089513839602555	0.8475869410929737	0.8622693399574166
F1	0.8806219219341447	0.8075771791091453	0.8513583664993943	0.8655203776690199
Prec	0.9217791411042945	0.9463874345549739	0.9569191870076826	0.9742589703588144
Recall	0.8429829346216785	0.7042780331956674	0.7667731629392971	0.7786176264318554

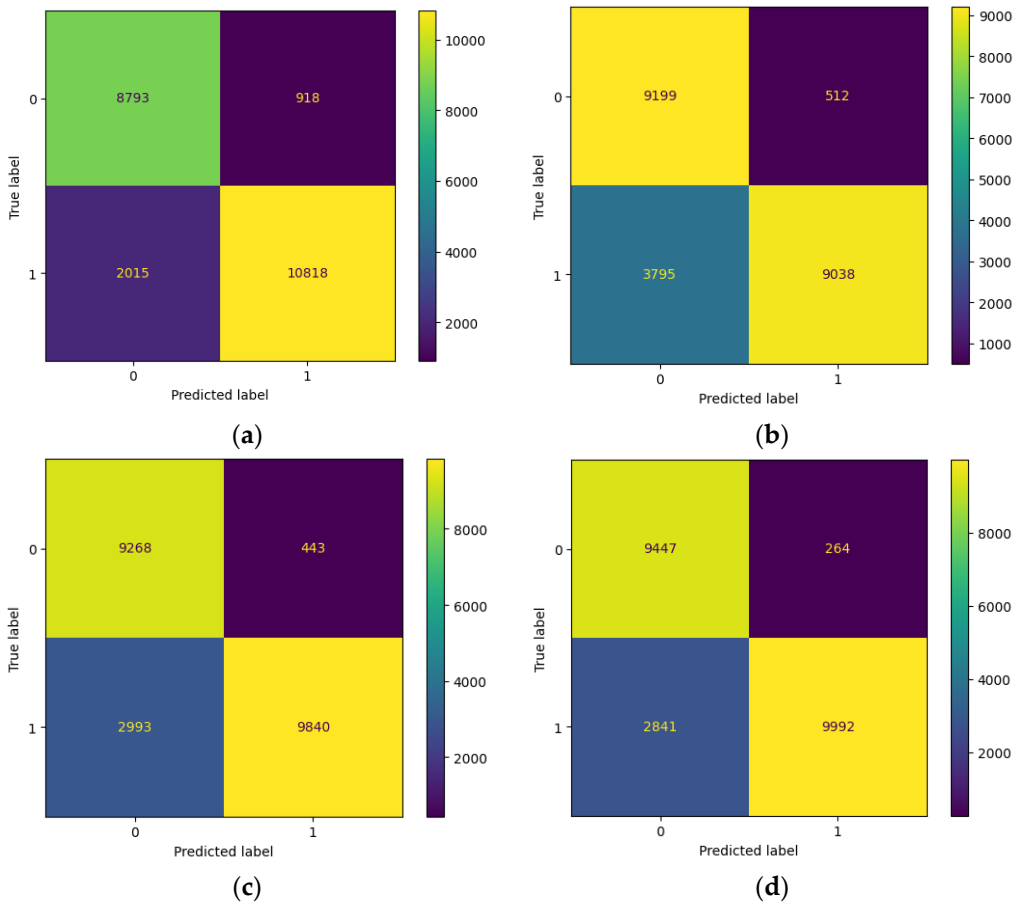


Figure 8. (**a**) Logistic regression output with random label flipping attack; (**b**) random forest output with random label flipping attack; (**c**) support vector machine output with random label flipping attack; and (**d**) XGB output with random label flipping attack.

Table 4. Model evaluation metrics with distance-based label flipping attack.

	LR	Random Forest	SVC	XGB
Acc	0.5491039744499645	0.8286905606813343	0.5692423704755145	0.839513839602555
F1	0.7066293399520911	0.8264893521430497	0.7254996183961331	0.839613440907882
Prec	0.5611477814448111	0.9759151193633953	0.5692423704755145	0.9737789203084833
Recall	0.9539468557624874	0.7167458895036235	1	0.7379412452271488

4.2.2. Evaluation of Machine Learning with a Distance-Based Flipping Attack with a Threshold of 0.5

With a distance threshold of 0.5, the distance-based flipping attack had a lower success rate than the previous attack. This is evident from the lower recall values. When comparing this attack to the other attacks, as shown in Table 4, some models perform worse, however, some models performs better than under the label flipping data poisoning attack, like Random Forest, Figure 9b. Some models were also highly affected, such as the SVC and the logistic regression models which could either no longer identify True Labels or barely identify them, as shown in Figure 9a–c.

It can be concluded that the distance-based flipping attack with a threshold of 0.5 has a moderate success rate compared to the other attacks, but it may not be as effective as the other attacks for some classifiers.

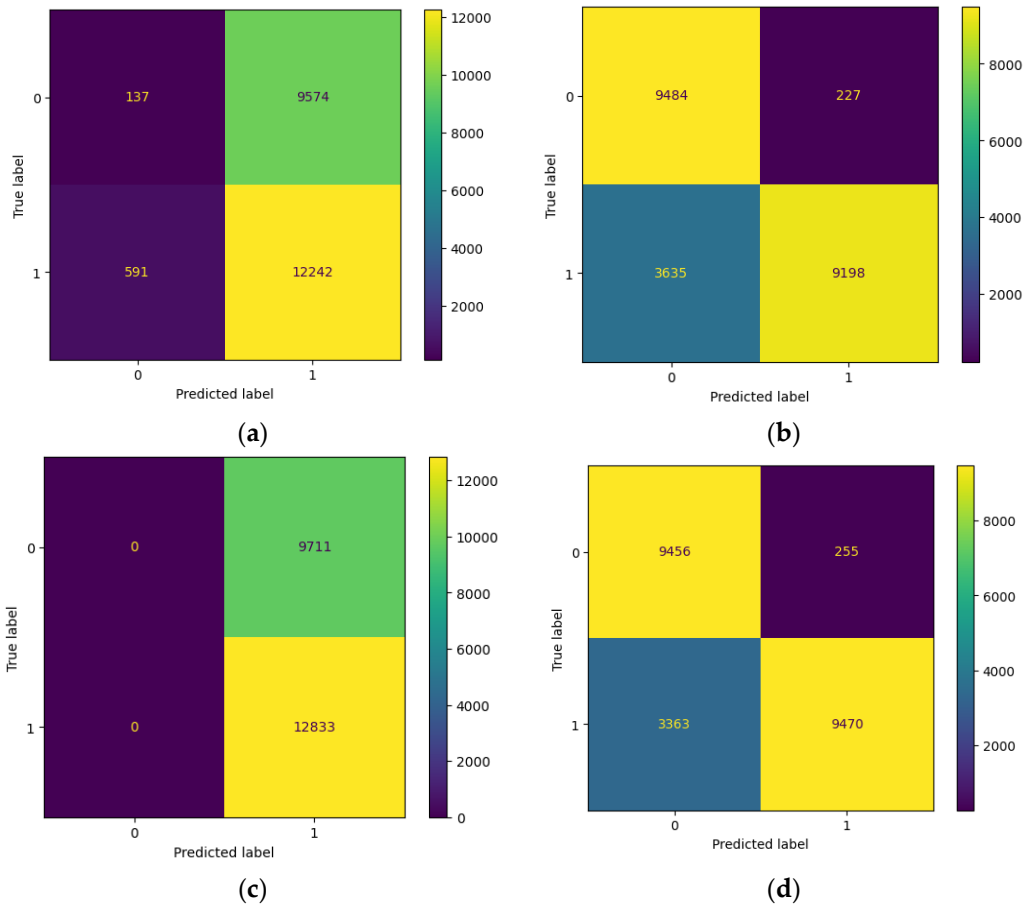


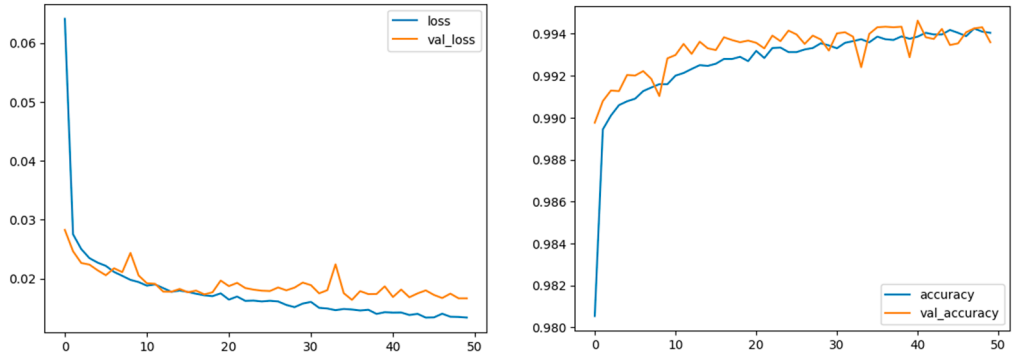
Figure 9. (**a**) Logistic regression output with distance-based flipping attack with threshold of 0.5; (**b**) random forest output with distance-based flipping attack with threshold of 0.5; (**c**) support vector machine output with distance-based flipping attack with a threshold of 0.5; and (**d**) XGB output with distance-based flipping attack with a threshold of 0.5.

4.3. Artificial Neural Network Data Poisoning Attack

In this section, the Artificial Neural Network will be trained without data poisoning attacks, and will be evaluated, afterwards a data poisoned version of this model will be evaluated and compared to the non-poisoned neural network model.

4.3.1. Evaluation of the Artificial Neural Network before a Random Label Flipping Attack

The neural network model used in this experiment is a feedforward artificial neural network with eight layers and an output layer. The activation function used is ReLU, and the output layer uses the sigmoid activation function. The model is trained using the binary cross-entropy loss function and optimised using the Adam optimiser. The Early Stopping callback is used to avoid overfitting by monitoring the training loss and stopping the training if it does not improve for five consecutive epochs; the results of the model show that it achieves high accuracy on both the training and validation sets, with a training accuracy of 0.994, a validation accuracy of 0.9936, and a test accuracy of 0.866, as shown in Figure 10b. The model was trained for 50 epochs, which was stopped early as the loss did not improve for five consecutive epochs. These results indicate that the model can learn the patterns in the input data accurately and generalise well to unseen data. Figure 10a shows the loss versus validation loss graph.



(**a**) (**b**)

Classification Report:

	precision	recall	f1-score	support
0.0	0.96	0.78	0.86	11917
1.0	0.80	0.96	0.87	10627
accuracy			0.87	22544
macro avg	0.88	0.87	0.87	22544
weighted avg	0.88	0.87	0.87	22544

(**c**)

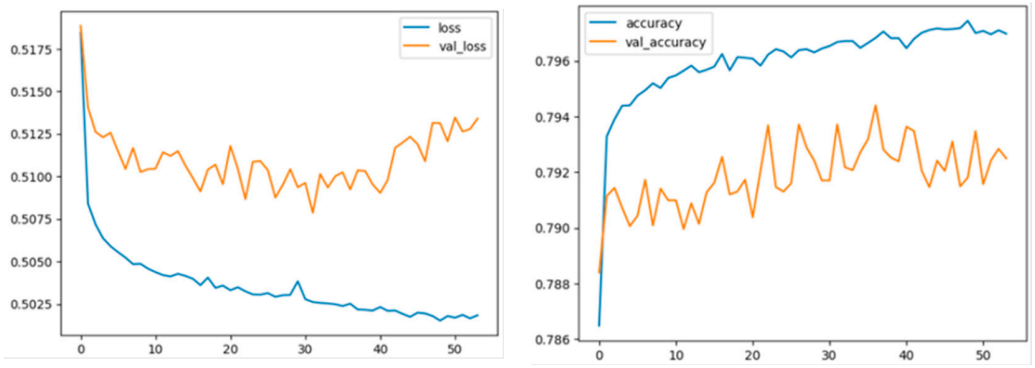
Figure 10. (**a**) Loss vs. validation loss in non-poisoned neural network; (**b**) accuracy vs. validation accuracy in non-poisoned NN; and (**c**) non-poisoned model performance report.

The classification report shows the performance metrics of the model for each class and its overall accuracy. Precision is the ratio of true positives to predicted positives, recall is the ratio of true positives to actual positives, and F1 Score is the harmonic mean of precision and recall. Support is the number of instances of each class, 0.0 denotes normal traffic, and 1.0 denotes attack traffic.

The report shows that the model has a high accuracy of 0.87, meaning it correctly classified 87% of instances. It also has a high F1 Score of 0.86, achieving a balanced performance across both classes, as shown in Figure 10c. And with a higher precision for normal labels than attack labels it made fewer false positive errors for normal labels than class attack labels. However, the model has a higher recall for attack labels than normal labels, and the classification report indicates that the model is effective for network anomaly detection on a non-poisoned dataset.

4.3.2. Evaluation of the Artificial Neural Network with a Random Label Flipping Attack

The results also show that both models have similar precision and recall scores for each class, meaning that they make similar errors in terms of false positives and negatives. However, the model trained on the non-poisoned dataset has a slightly higher F1 Score for both classes than the model trained on the poisoned dataset (0.86 vs. 0.84 for normal and 0.87 vs. 0.85 for attack), as shown in Figure 11c. This means that label flipping has a negative impact on the model's performance, as it reduces the trade-off between precision and recall. It can also be seen that the training accuracy and validation accuracy were affected, in Figure 11b, while Figure 11a shows the loss versus validation loss graph.



(**a**) (**b**)

Classification Report:

	precision	recall	f1-score	support
0.0	0.96	0.75	0.84	12431
1.0	0.75	0.96	0.84	10113
accuracy			0.84	22544
macro avg	0.86	0.85	0.84	22544
weighted avg	0.87	0.84	0.84	22544

(**c**)

Figure 11. (**a**) Loss vs. validation loss in poisoned neural network; (**b**) accuracy vs. val accuracy in poisoned neural network; and (**c**) poisoned model performance report.

4.4. Experiment Analysis

Data poisoning attacks are a type of attack that targets the vulnerability of machine learning models' training data. By injecting malicious samples into the training data, the attackers can manipulate the model's behaviour and compromise its integrity. Data poisoning attacks can devastate a real-world supply chain network, especially if the network relies on machine learning models for its operations.

One example of such a scenario is the ordering system network within a company that operates a distribution centre. The ordering system network uses a machine learning model based on intrusion detection and prevention systems (IDPSs) to monitor network traffic and detect or prevent malicious activities. However, suppose an attacker can poison the IDPS model and gain access to the backend server. In that case, they can modify the ordering system and overload the distribution centre with excessive or fraudulent orders. This can cause serious losses for the company in terms of money, time, and reputation.

Moreover, the attack can cascade effects throughout the overall supply chain network, disrupting the flow of goods and services from the suppliers to the customers. The attack can create bottlenecks, delays, shortages, or surpluses in the network, depending on the nature and magnitude of the order modifications. For example, suppose the attacker increases the orders for some products and decreases the orders for others. In that case, this can affect the company's and its suppliers' inventory levels and demand forecasts. This can lead to the overstocking or understocking of some products, resulting in waste or lost sales. Alternatively, if the attacker cancels or duplicates some orders, this can affect the delivery schedules and customer satisfaction of both the company and its customers. This can lead to missed deadlines or over-deliveries, resulting in penalties or refunds.

The attack can also affect the trust and collaboration among the supply chain partners, as they may be unable to verify or communicate with each other about the order changes. This can increase the risk and uncertainty in the network, as they may not be able to coordinate their actions or respond to contingencies effectively. These effects can have long-term consequences for the competitiveness and sustainability of the network, as they may erode its reputation, efficiency, and resilience [29,30].

To evaluate the impact of data poisoning attacks on machine learning models based on IDPSs, we conducted experiments using five different models: logistic regression, random forest, support vector machine, XGBoost, and an artificial neural network. We trained and tested these models on two datasets: a non-poisoned dataset that does not contain any label changes and a poisoned dataset that is generated using two different attacks—random label flipping and distance-based label flipping. The random label flipping attack randomly changes the labels of some instances in the training data, with a probability of 0.1, to mislead the model. The distance-based label flipping attack changes the labels of some instances close to the model's decision boundary, with a distance threshold of 0.5, to confuse the model.

The results of our experiments are shown in Tables 2–4 and Figures 7–11. Table 2 shows the performance metrics of all models on the non-poisoned dataset. The results show that all models have a reasonable accuracy and F1 Score on this dataset, with the logistic regression, Figure 7a, having the highest values (0.858 and 0.868, respectively) and the random forest, Figure 7b, having the lowest values (0.83 and 0.82, respectively). Table 3 shows the performance metrics of all models on the poisoned dataset generated using the random label flipping attack. The results show that all models have a lower accuracy and F1 Score on this dataset than their original values, indicating that this attack successfully degrades their performance. The logistic regression, Figure 8a, model is still the most resilient to this attack, as it has the smallest decrease in performance (0.868 to 0.88 for its F1 Score). Table 4 shows the performance metrics of all models on the poisoned dataset generated using the distance-based label flipping attack. The results show that this attack has a lower success rate than the random label flipping attack, as it affects fewer instances in the training data. The logistic regression model in distance-based label flipping attack, Figure 9a is again the most resilient to this attack, as it has the highest accuracy and F1 Score on this dataset (0.549 and 0.706, respectively). Figures 10 and 11 show the loss and accuracy curves of the ANN model on the non-poisoned and poisoned datasets, respectively. The figures show that the ANN model has high accuracy on both datasets but lower precision and recall than other models.

The results indicate that data poisoning attacks are a serious threat to machine learning models based on IDPSs and the supply chain networks that depend on them. To defend against data poisoning attacks, several methods have been proposed in the literature, such as data sanitisation, robust aggregation, and anomaly detection. Future work can explore these methods and evaluate their effectiveness in preventing or mitigating data poisoning attacks on supply chain networks.

The KDDCUP'99 dataset offers valuable insights into network intrusion detection, but it does have limitations, including outdated data and imbalanced classes. These limitations may render intrusion detection models less effective. Therefore, data poisoning attacks could exacerbate these issues by further compromising the quality and reliability of models, thus increasing the risk of false positives and false negatives in identifying network intrusions.

It can also be noted that some features in the dataset are highly correlated with the attack labels, meaning that they are more likely to indicate an attack when they have higher values. These features are crucial for the effectiveness of intrusion detection models. However, if attackers manipulate the training data through data poisoning, these correlations may be disrupted, causing the models to become less reliable.

Furthermore, as discussed in the third section, the distribution of attack and normal traffic in the network anomaly dataset emphasises that normal traffic is more consistent and stable, making it easier to detect anomalies. However, data poisoning attacks can skew this balance by injecting misleading data, potentially making it more challenging to differentiate between normal and attack traffic.

In addition, the effects of data poisoning attacks on supply chain networks can be related to the potential disruption of the overall network's flow of goods and services. Data poisoning attacks can exacerbate these challenges by introducing misleading data, thus impeding the normal flow of network traffic and potentially causing disruptions, bottlenecks, and delays. These attacks pose a significant threat to the machine learning

models used in network intrusion detection systems and the supply chain networks that rely on these models. It is important to underscore the importance of defending against data poisoning attacks and the need for robust security measures to protect both network integrity and the broader supply chain network. These security measures should also consider the dataset limitations and analysis highlighted in our third section to ensure that intrusion detection systems remain effective in an ever-evolving threat landscape.

It was also demonstrated that data poisoning attacks can seriously jeopardise the security and effectiveness of supply chain networks by drastically reducing the precision and dependability of machine learning models built for network intrusion detection systems. The random forest model proved to be the most susceptible to both kinds of attacks, whereas the logistic regression model had the highest level of resilience. Additionally, as it affected fewer occurrences in the training data, we discovered that the distance-based label flipping approach had a lower success rate than the random label flipping attack.

This research contributes to the body of knowledge on supply chain networks and network intrusion detection systems by examining the impact of data poisoning attacks on machine learning models in this field. Additionally, the design and development of more resilient and secure network intrusion detection systems that can handle the difficulties and dangers of data poisoning assaults would benefit from our research.

However, our study has a few drawbacks that may be resolved in future studies. A limitation of this study is the examination of only two categories of data poisoning attacks, which may not encompass the entire range of potential attacks that could be directed towards network intrusion detection systems. Different attacks that cause data poisoning, such as introducing malicious samples, changing features, or adding noise, could affect the models differently and demand distinct countermeasures. Another drawback is that the tests we conducted only employed one dataset, which might not accurately represent the complexity and diversity of network traffic and attacks in the real world. More realistic and varied datasets that can represent the features and dynamics of supply chain networks afforded by blockchain technology could be used in future studies.

Further studies may also examine improved efficient techniques for identifying and minimising data poisoning assaults on network intrusion detection systems. Data sanitisation, robust aggregation, and anomaly detection are some of these potential techniques. Before feeding the training data to the model, data sanitisation tries to eliminate or fix any malicious or noisy samples. The goal of robust aggregation is to decrease the influence of poisoned models or poisoned learners by combining the outputs of several models or learners. Finding instances where data or a model deviate from expected or normal behaviour is the goal of this anomaly detection technique.

Other data poisoning attacks such as gradient-based poisoning attacks, which use the gradient information of the target model to generate poisoned samples that can alter the model's parameters or decision boundary [31–33], or meta-learning poisoning attacks, which use meta-learning techniques to generate poisoned samples, could also be investigated in future research. These attacks can be applied to models such as linear models, neural networks, or kernel methods.

To summarise our findings, the following can be said about our research results.

- The logistic regression model is the most resilient to both attacks because it is a linear model with a simple and stable decision boundary less affected by the label changes. The label changes only affect the instances close to the decision boundary, which are fewer in number and have less influence on the model's parameters. The logistic regression model also has a regularisation term that prevents overfitting and reduces its sensitivity to noise or outliers.
- The random forest model is the most vulnerable to both types of attacks because it is an ensemble model that combines the outputs of multiple decision trees trained on random data subsets. The label changes affect most instances in each subset, leading to high variance and inconsistency among the decision trees. The random forest model

also has high complexity and flexibility which make it prone to overfitting and increase its susceptibility to noise or outliers.

- The support vector machine model is moderately resilient to both attacks because it is a kernel-based model that uses a nonlinear transformation to map the data to a higher dimensional space where it can find a linear decision boundary that separates the classes. The label changes affect the instances close to the decision boundary, which are the support vectors that determine the model's parameters. The support vector machine model also has a regularisation term that prevents overfitting and reduces its sensitivity to noise or outliers.
- The XGBoost model is moderately vulnerable to both types of attacks because it is a gradient-boosting model that iteratively adds new decision trees trained on the residuals of the previous trees. The label changes affect the model's residuals and gradients, leading to a high bias and error accumulation among the decision trees. The XGBoost model also has high complexity and flexibility which make it prone to overfitting and increase its susceptibility to noise or outliers.
- The artificial neural network model is highly accurate on both datasets. However, it has lower precision and recall than other models because it is a deep learning model that uses multiple layers of nonlinear transformations to learn complex and abstract features from the data. The label changes affect the model's features and weights, leading to high confusion and misclassification among the classes. The artificial neural network model also has high complexity and flexibility that make it prone to overfitting and increase its susceptibility to noise or outliers.

5. Conclusions and Discussion

This research demonstrated the effects of two types of data poisoning attacks, namely label flipping and distance-based attacks, against several machine learning models. The experiment shows that data poisoning attacks can compromise the accuracy and reliability of the machine learning models used in network intrusion detection systems. It was observed that distance-based attacks have the highest probability of causing damage to the models' accuracy. Malicious actors can carry out these attacks to exploit vulnerabilities in the system and cause security breaches, which in turn can pose a serious threat to the security and efficiency of supply chain networks. Therefore, it is essential to develop robust countermeasures to detect and mitigate the effects of these attacks. Further research is needed to develop more effective defences against poisoning attacks and to enhance the security of network intrusion detection systems.

The effect of label flipping attacks on network anomaly detection models based on neural networks was much investigated in this research, within the scope of a supply chain network. It was shown that these attacks can significantly degrade the performance and robustness of the models, especially when the attack intensity is high, which could lead to the disruption of the supply chain network. However, our study has some limitations that can be addressed in future research, such as that this research has only considered two types of data poisoning attacks, which may not capture the full spectrum of the possible attacks that can be launched against network anomaly detection systems. Other types of data poisoning attacks, such as adding malicious samples, modifying features, or injecting noise, may affect the models differently and require different defences.

Additionally, our reliance on a single dataset could limit representativeness, requiring future studies to incorporate more diverse datasets reflecting the complexities of supply chain networks. The main finding in this research is the vulnerability of machine learning models in network intrusion detection systems, especially random forest, to data poisoning attacks, which significantly reduce their precision and reliability. Conversely, the logistic regression model displayed notable resilience. This research significantly contributes to understanding the intricacies of data poisoning attacks, paving the way for more robust intrusion detection systems. The findings of this research also inform senior managers of the risks associated

with the integrity of technical controls in the changing cyber security landscape and of how to ensure business continuity using robust measures to mitigate such vulnerabilities.

Author Contributions: Methodology, U.J.B., O.H. and K.H.; validation, K.S., B.H. and H.a.-K.; investigation, all authors; formal analysis O.H. and K.H.; writing—original draft, U.J.B., O.H. and K.H.; writing—review and editing, K.S., B.H. and H.a.-K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIsPublicly available datasets were analyzed in this study. This data can be found here: (https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed on 15 November 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abuali, K.M.; Nissirat, L.; Al-Samawi, A. Advancing Network Security with AI: SVM-Based Deep Learning for Intrusion Detection. *Sensors* **2023**, *23*, 8959. [CrossRef] [PubMed]
2. Fan, J.; Yan, Q.; Li, M.; Qu, G.; Xiao, Y. A Survey on Data Poisoning Attacks and Defenses. In Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), Guilin, China, 11–13 July 2022.
3. Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Mądry, A.; Li, B.; Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1563–1580. [CrossRef] [PubMed]
4. Wang, S.; Li, Q.; Cui, Z.; Hou, J.; Huang, C. Bandit-based data poisoning attack against federated learning for autonomous driving models. *Expert Syst. Appl.* **2023**, *227*, 120295. [CrossRef]
5. Yerlikaya, F.A.; Bahtiyar, Ş. Data poisoning attacks against machine learning algorithms. *Expert Syst. Appl.* **2022**, *208*, 118101. [CrossRef]
6. Sun, G.; Cong, Y.; Dong, J.; Wang, Q.; Lyu, L.; Liu, J. Data poisoning attacks on federated machine learning. *IEEE Internet Things J.* **2021**, *9*, 11365–11375. [CrossRef]
7. Nisioti, A.; Mylonas, A.; Yoo, P.D.; Katos, V. From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 3369–3388. [CrossRef]
8. Saxena, N.; Sarkar, B. How does the retailing industry decide the best replenishment strategy by utilizing technological support through blockchain? *J. Retail. Consum. Serv.* **2023**, *71*, 103151. [CrossRef]
9. Cinà, A.E.; Grosse, K.; Demontis, A.; Vascon, S.; Zellinger, W.; Moser, B.A.; Oprea, A.; Biggio, B.; Pelillo, M.; Roli, F. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Comput. Surv.* **2023**, *55*, 1–39. [CrossRef]
10. Talty, K.; Stockdale, J.; Bastian, N.D. A sensitivity analysis of poisoning and evasion attacks in network intrusion detection system machine learning models. In Proceedings of the MILCOM 2021—2021 IEEE Military Communications Conference (MILCOM), San Diego, CA, USA, 29 November–2 December 2021.
11. Zhang, Y.; Zhang, Y.; Zhang, Z.; Bai, H.; Zhong, T.; Song, M. Evaluation of data poisoning attacks on federated learning-based network intrusion detection system. In Proceedings of the 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Chengdu, China, 18–21 December 2022.
12. Zhang, Z.; Zhang, Y.; Guo, D.; Yao, L.; Li, Z. SecFedNIDS: Robust defense for poisoning attack against federated learning-based network intrusion detection system. *Futur. Gener. Comput. Syst.* **2022**, *134*, 154–169. [CrossRef]
13. Lai, Y.-C.; Lin, J.-Y.; Lin, Y.-D.; Hwang, R.-H.; Lin, P.-C.; Wu, H.-K.; Chen, C.-K. Two-phase Defense Against Poisoning Attacks on Federated Learning-based Intrusion Detection. *Comput. Secur.* **2023**, *129*, 103205. [CrossRef]
14. Taheri, R.; Javidan, R.; Shojafar, M.; Pooranian, Z.; Miri, A.; Conti, M. Correction to: On defending against label flipping attacks on malware detection systems. *Neural Comput. Appl.* **2020**, *32*, 14781–14800. [CrossRef]
15. Zarezadeh, M.; Nourani, E.; Bouyer, A. DPNLP: Distance based peripheral nodes label propagation algorithm for commu-nity detection in social networks. *World Wide Web* **2022**, *25*, 73–98. [CrossRef]
16. Gupta, P.; Yadav, K.; Gupta, B.B.; Alazab, M.; Gadekallu, T.R. A Novel Data Poisoning Attack in Federated Learning based on Inverted Loss Function. *Comput. Secur.* **2023**, *130*, 103270. [CrossRef]
17. Deng, X.; Zhu, J.; Pei, X.; Zhang, L.; Ling, Z.; Xue, K. Flow topology-based graph convolutional network for intrusion detec-tion in label-limited iot networks. *IEEE Trans. Netw. Serv. Manag.* **2022**, *20*, 684–696. [CrossRef]
18. Song, J.; Wang, X.; He, M.; Jin, L. CSK-CNN: Network Intrusion Detection Model Based on Two-Layer Convolution Neural Network for Handling Imbalanced Dataset. *Information* **2023**, *14*, 130. [CrossRef]
19. Koh, P.W.; Steinhardt, J.; Liang, P. Stronger data poisoning attacks break data sanitisation defenses. *Mach. Learn* **2022**, *111*, 1–47. [CrossRef]
20. Zhu, Y.; Wen, H.; Zhao, R.; Jiang, Y.; Liu, Q.; Zhang, P. Research on Data Poisoning Attack against Smart Grid Cyber–Physical System Based on Edge Computing. *Sensors* **2023**, *23*, 4509. [CrossRef]

21. Shah, B.; Trivedi, B.H. Reducing features of KDD CUP 1999 dataset for anomaly detection using back propagation neural network. In Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, India, 21–22 February 2015.
22. Divekar, A.; Parekh, M.; Savla, V.; Mishra, R.; Shirole, M. Benchmarking datasets for anomaly-based network intrusion detection: KDD CUP 99 alternatives. In Proceedings of the 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Katmandu, Nepal, 25–27 October 2018.
23. Zhang, H.; Li, Z.; Shahriar, H.; Tao, L.; Bhattacharya, P.; Qian, Y. Improving prediction accuracy for logistic regression on imbalanced datasets. In Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 15–19 July 2019.
24. Madzarov, G.; Gjorgjevikj, D. Multi-class classification using support vector machines in decision tree architecture. In Proceedings of the IEEE EUROCON 2009, St. Petersburg, Russia, 18–23 May 2009.
25. Cheng, N.; Zhang, H.; Li, Z. Data sanitisation against label flipping attacks using AdaBoost-based semi-supervised learning technology. *Soft Comput.* **2021**, *25*, 14573–14581. [CrossRef]
26. Li, Q.; Wang, X.; Wang, F.; Wang, C. A Label Flipping Attack on Machine Learning Model and Its Defense Mechanism. In Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing, Copenhagen, Denmark, 10 October 2022.
27. Barreno, M.; Nelson, B.; Sears, R.; Joseph, A.D.; Tygar, J.D. Can machine learning be secure? In Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, Taipei, Taiwan, 21–24 March 2006.
28. Biggio, B.; Nelson, B.; Laskov, P. Poisoning attacks against support vector machines. *arXiv* **2012**, arXiv:1206.6389.
29. Chen, J.; Gao, Y.; Shan, J.; Peng, K.; Wang, C.; Jiang, H. Manipulating Supply Chain Demand Forecasting with Targeted Poisoning Attacks. *IEEE Trans. Ind. Inform.* **2022**, *19*, 1803–1813. [CrossRef]
30. Lin, J.; Dang, L.; Rahouti, M.; Xiong, K. Ml attack models: Adversarial attacks and data poisoning attacks. *arXiv* **2021**, arXiv:2112.02797.
31. Qiu, H.; Dong, T.; Zhang, T.; Lu, J.; Memmi, G.; Qiu, M. Adversarial attacks against network intrusion detection in IoT systems. *IEEE Internet Things J.* **2020**, *8*, 10327–10335. [CrossRef]
32. Venkatesan, S.; Sikka, H.; Izmailov, R.; Chadha, R.; Oprea, A.; de Lucia, M.J. Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems. In Proceedings of the MILCOM 2021—2021 IEEE Military Communications Conference (MILCOM), San Diego, CA, USA, 29 November–2 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 874–879.
33. Salo, F.; Injadat, M.; Nassif, A.B.; Shami, A.; Essex, A. Data mining techniques in intrusion detection systems: A systematic literature review. *IEEE Access* **2018**, *6*, 56046–56058. [CrossRef]