

Deep Behavioral Analysis of Machine Learning Algorithms Against Data Poisoning

Anum Paracha, Junaid Arshad, Mohamed Ben Farah, Khalid Ismail

^aDepartment of Computing, Birmingham City University, UK

Abstract

Poisoning attacks represent one of the most common and practical adversarial attempts on machine learning systems. In this paper, we have conducted a deep behavioural analysis of six machine learning (ML) algorithms, analyzing poisoning impact and correlation between poisoning levels and classification accuracy. Adopting an empirical approach, we highlight practical feasibility of data poisoning, comprehensively analyzing factors of individual algorithms affected by poisoning. We used public datasets (UNSW-NB15, BotDroid, CTU13, and CIC-IDS-2017) and varying poisoning levels (5% - 25%) to conduct rigorous analysis across different settings. In particular, we analyzed the accuracy, precision, recall, f1-score, false positive rate and ROC of the chosen algorithms. Further, we conducted a sensitivity analysis of each algorithm to understand the impact of poisoning on its performance and characteristics underpinning its susceptibility against data poisoning attacks. Our analysis shows that, for 15% poisoning of UNSW-NB15 dataset, the accuracy of Decision Tree (DT) decreases by 15.04% with an increase of 14.85% in false positive rate. Further, with 25% poisoning of BotDroid dataset, accuracy of K-nearest neighbours (KNN) decreases by 15.48%. On the other hand, Random Forest (RF) is comparatively more resilient against poisoned training data with a decrease of 8.5% in accuracy with 15% poisoning of UNSW-NB15 dataset and 5.2% for BotDroid dataset. Our results highlight that 10%-15% of dataset poisoning is the most effective poisoning rate, significantly disrupting classifiers without introducing overfitting, whereas 25% is detectable because of high performance degradation and overfitting algorithms. Our analysis also helps understand how asymmetric features and noise affect the impact of data poisoning on machine learning classifiers. Our experimentation and analysis are publicly available at: <https://github.com/AnumAtique/Behavioural-Analysis-of-Poisoned-ML/>

Keywords: Behavioral Analysis, Adversarial Poisoning, Integrity Violation, Complacent Poisoning

1. Introduction

Machine learning models are widespread, facilitating cutting-edge digital solutions in a range of scenarios including security-critical applications such as malware detector[1], intrusion detection system[5], automated firewalls[6], and biometric recognition[7]. Machine learning models help understand patterns from the given dataset and train themselves to predict and classify new data without requiring additional information or interaction with any third party such as humans. With the training dataset, machine learning models develop the dynamic classification mechanism that leverages these models to understand the nature of new data and classify them.

Such proliferation of machine learning models and their dynamic classification mechanism renders their security fundamental to the security of systems underpinned by them. Several attacks have been explored in literature aiming to compromise the performance and accuracy of machine learning algorithms such as [8], [9], [10] and [11]. Among these attacks, data poisoning [8] is one of the most prominent attacks on machine learning whereby an adversary attempts to infiltrate the training data and exploit the integrity of machine learning. It also enables the adversary to infer the victim model, either targeted or indiscriminately, to hinder its privacy during the deployment phase. A number of poisoning techniques are available in literature such as [12], [13], and [14]. Fur-

ther, various sophisticated attacks are formulated to poison machine learning algorithms. Some of its successful examples are convex-polytope[15], label-flipping[16], bullseye-polytope[17] and poison frog[18].

These data poisoning attacks manipulate training datasets in two ways. Firstly, the adversary can perturb features of the training dataset so that machine learning models misinterpret this data and wrongly develop their classification mechanism. Secondly, data poisoning attacks manipulate the classification labels of the training dataset which blurs the decision boundaries of models. Data poisoning attack has a serious impact on various real-life applications such as deceiving breast cancer diagnosis applications [36], skin cancer diagnosis applications [4], and IoT-enabled smart city systems [35]. To better mitigate these data poisoning attacks, it is significant to understand the manipulated behavior of these poisoned machine learning models.

Security analysis of machine learning algorithms has received significant attention in recent years focusing on poisoning and other adversarial attempts. Specifically on data poisoning attacks, various research papers have conducted an empirical analysis of the performance of machine learning algorithms [16], and [19], but these attempts do not provide the technical and mathematical understanding of change in their classification behavior and why data poisoning attacks are successful in corrupting their decision mechanisms, which is attempted first

time in this research.

In this context, our research is the first to study and analyze the disruption of the classification mechanism and highlight the inherent characteristics of individual algorithms and their susceptibility caused by training with a poisoned dataset. Our goals are to determine the relationship between performance disruption due to the injected poison and to analyze the changes in model parameters. This performance analysis aims to assess the varying responsiveness of each algorithm to data poisoning from a technical standpoint. These insights will aid in analyzing the impact of adversarial attacks on machine learning algorithms and in implementing measures to mitigate them. Our attacking procedure is developed on a practical examination of data poisoning on machine learning classification models. We conducted an experimental study to analyze the impact of data poisoning on individual machine learning algorithms. Our goal was to highlight the optimal levels of poisoning that lead to significant performance degradation in the classifiers and are difficult to interpret. Data poisoning is implemented by flipping labels randomly from 5%-25% of the dataset. Our analysis is significant in understanding the poisoning behavior for developing profound mitigation solutions that proactively secure machine learning models from existing and novel attacks.

In this paper, we share the outcomes of our efforts to investigate the impact of poisoning attacks on supervised machine learning algorithms. Our aim is to specifically study the correlation between the granular impact of poisoning on the performance of machine learning algorithms. We also study the correlation between data poisoning and features relation in the dataset. This analysis employs six supervised machine learning algorithms: Support Vector Machines (SVM), Decision Tree (DT), K-nearest neighbors (KNN), Random Forest (RF), Gaussian naive Bayes (GNB), and Perceptron. We have selected these algorithms to cover all the baseline classification methods in machine learning. This helps analyze different classification behaviors in machine learning with four distinct datasets (BotDroid [38], UNSW-NB15 [39], CTU-13 [41], and CIC-IDS-2017 [40]). The results of our research study proved RF to be most resilient against different levels of injected poison with linear and minimal accuracy disruption however the accuracy of neural networks is not linear and fluctuating. We have predominantly focused on universal examination parameters including accuracy, precision, recall, false positive rate, and ROC curve for evaluating the performance of supervised machine learning. Our approach is formulated on an untargeted attack with an underlying assumption that the adversary does not have any internal model settings and data distribution. We have implemented a label-flipping attack because it is the most simple and popular attack, used in many research papers[2],[3], in data poisoning and to keep our attack settings simple to understand models' behavior clearly. We have started poisoning 5% of dataset samples leading to as high as 25% dataset to analyze the impact on machine learning and its performance degradation. Succinctly, a drastic decrease is visible in most of the ML algorithms with 25% poisoned dataset which shows a major integrity violation except for support vector machines and neural networks of which the performance randomly fluctuates. Major

contributions of our research are as follows:

- We have conducted a thorough behavioral analysis of machine learning classification algorithms against a label-flipping attack to analyze the correlation between the percentage of poisoned data and the impact on classification accuracy. Using data poisoning proportions ΔP from 5% to 25%, we highlight optimal poisoning levels that exploit machine learning by infiltrating the model in an obscure manner.
- Extending existing work such as [16], [19], we analyze the impact of data poisoning on the design of ML algorithms and highlight sensitive factors for each algorithm. We have implemented this attack on seven machine learning classifiers with four distinct datasets to quantify poisoning impact, analyzing accuracy, precision, recall, f1-score, and false positive rate of classifiers, allowing us to empirically explain poison penetration in ML classifiers.
- We have analyzed the behavior and impact of data poisoning on classification results with a noisy dataset. We have subsequently studied the impact of anisotropic features and imbalanced dataset distribution. Also, explaining varying results of poisoned datasets, with no feature engineering, that acts as a catalyst in decreasing the performance of ML classifiers.
- Our research findings provide a baseline to help strengthen mitigating techniques against machine learning poisoning attacks considering factors such as identifying relevant features, cleaning dataset, removing over-fitted branches and nodes in case of DT and RF, and calculating cluster density and distances between clusters in KNN and SVM.

2. Analysis of Existing Efforts Related to Poisoning Attacks for Machine Learning

In this section, we have discussed existing poisoning attacks that performed empirical analysis in different attack settings. Our focus is to analyze their objectives and methods in this regard. Also, discussing existing defenses to mitigate poisoning attacks.

2.1. Existing Poisoning Attacks Against Machine Learning

Keeping machine learning models secure and integrated is a potential and active research challenge. Poisoning attacks, which involve contaminating datasets, pose significant risks to the integrity and confidentiality of machine learning systems. Notable examples of poisoning attacks are rethinking Label-Flipping [20], geometric algorithms for KNN poisoning [21], and subpopulation data poisoning attack [22]. These attacks have effectively compromised the integrity and intruded into the targeted models. However, there is a pressing need to understand the differential responses of these models to poisoning techniques.

Poisoning attacks are increasingly endangering the reliability

and accessibility of machine learning models. R. Mayerhofer et al. [23] have polluted a featured-extraction system developed with convolutional neural networks (CNN) and evaluated the effectiveness of poisoning attacks against it. C. Zhang et al. [24] have fooled deep neural networks (DNN) by integrating manipulated features into them. Further studies, such as [15], [18] manipulated machine learning with features set perturbation while C. Zhu et al. [25] showed how specifically altered pixels could mislead classifiers. In another study [26], perturbed labels are proved to help insert a backdoor in machine learning with a success rate of 98%. A brief comparison of existing similar work with our research is given in Table 1. Further Table 2 extends analysis presented in Table 1 including a comparative analysis of our approach with existing data poisoning studies highlighting the impact on performance of individual algorithms. Our work aims to interpret the analytical reasoning of different supervised machine learning algorithms and their response to various poison levels. Although some research studies such as [16], [19] and [27] have conducted empirical analysis to show the performance degradation of machine learning, they fall short in explaining why models trained on identical datasets with the same poisoning rates exhibit divergent performance behaviors. Addressing this gap forms a central part of our contribution. Many existing research studies investigated the integrity violation and shift in machine learning decision-making, either deterministically or indiscriminately. However, the urge is to study the underlying model design that reveals poison differently for each ML algorithm and so to interpret its explainability against poisoning.

2.2. Existing Mitigation Techniques to Overcome Poisoning Attacks

To distill machine learning models from the detrimental effects of poisoning attacks, a variety of mitigation strategies have been developed. A data sanitization against an adversarial label-flipping attack is proposed in [28]. This method begins with random dataset clustering to train the model on subsets of a dataset and then sanitize data points, taking into account the complexity of the data and its impact on model performance. Adversarial training [29], [30], [31] is also an effective technique to allow machine learning models to proactively learn poisoned patterns to guard against poison if injected. A brief overview of the discussed mitigation techniques is given in Table 3.

Where in [32] S. Drews and others proposed an antidote to verify the robustness of a decision tree against data poisoning attacks. Although these solutions are effective but lack generalizability. In particular, the fundamentals of this research study is to put a step forward to provide a rigorous examination of supervised machine learning behavior against poisoning attacks. On a large scale, none of the studies have yet highlighted the reasons behind different responses of machine learning algorithms when trained with the same poison levels.

3. Formal Notations of Our Attack Method

In this section, we have interpreted the fundamentals of our attack method to poison binary machine learning classifiers in black box attack settings. Definitions of our attack method and evaluation metrics to measure the poisoning effects and analyze behaviors of machine learning algorithms are also provided.

3.1. Attack Overview

The architectural overview of our threat model is given in Fig 1, representing the attacked and clean model development and defining the analysis criteria. Steps 1-9 are given in Fig 1 highlighted the attack development flow. Step #1 is the dataset selection, step #2 defines the poisoning levels to poison datasets, step #3 is the implementation of label flipping attack on datasets with the defined poisoning levels and step #4 defines algorithms to be considered for analysis. Steps #5 and #6 are the development of poisoned and cleaned models subsequently. Steps #7-9 define the analysis criteria. The objective of our research study is the behavioral examination of poisoned machine learning algorithms. We have intended to poison binary classifiers by poisoning binary classifiers with the poisoning algorithm, described in Algorithm 1. We have audited each algorithm to identify individual factors that are affected by dataset manipulation. Our attack is developed in a black box setting with a notion of generalization to poison labeled datasets to degrade the classifier’s performance. The poison is injected with a minimum

Algorithm 1 Poisoning Algorithm

```

Datasets used are: UNSW-NB15, BotDroid, CTU-13, CIC-IDS-2017
 $D_t \leftarrow$  Training dataset
 $D_p \leftarrow$  Poisoned dataset = []
 $D_r \leftarrow$  subset of Training dataset
Record  $\leftarrow D_r(x_i, l_i)$  is the instance in dataset  $D_r$  with features  $x$  and label  $l$  at position  $i$ 
 $\Delta P \leftarrow$  Poisoning level = [0%, 5%, 10%, 15%, 20%, 25%]
for poison  $\in \Delta P$  do
  Select  $D_r$  of length poison from  $D_t$ 
  for Record  $\in D_r$  do
    if Record not in  $D_p$  then
      if  $l_i$  equals 0 then
         $l_i = 1$ 
      else
         $l_i = 0$ 
      end if
    end if
     $D_p \leftarrow$  Record
  end for
end for

```

5% poisoning rate leading to a maximum of 25%, at a scale of 5. We have focused on developing binary classifiers with four benchmarked datasets consisting of network traffic, described in Section 4.2, following the purpose of providing a precise and granular study is to put forward analytical reasons of how and why the performance of every algorithm is affected differently. For example, the performance (accuracy, precision, recall, f1-score) of DT, KNN, and RF are degraded linearly. Whereas, performances of neural networks are continuously fluctuating. The ablation study is given in Section 4.3.

3.2. Evaluation Metrics to Quantify Effects of Data Poisoning

For evaluating the performance of machine learning models under poisoning, the indicators we have used are accuracy, precision, recall, f1-score, false positive rate(FPR), and ROC. FPR

Table 1: Comparative analysis with existing similar studies that provides an empirical analysis of poisoning attacks on machine learning along with their objectives/focus to conduct empirical analysis

S.No.	Source	Datasets used	Attack settings	Focus
1	F. A. Yerlikaya et al. [16]	4	White box attack	Practical implication of data poisoning
2	K. Aryal et al. [19]	1	Black box attack	Impact of data poisoning on Malware detection
3	C. Dunn et al. [27]	2	-	Impact of data poisoning on smart IoT systems
4	Our study	4	Black box attack	Factual study to identify factors affecting the performance of each ML algorithms with data poisoning

Table 2: Performance analysis with existing similar studies analyzing the impact of poisoning attacks on machine learning models

S.No.	Source	Performance Metrix	Model performance					
			SVM	RF	DT	KNN	GNB	Perceptron
1	F. A. Yerlikaya et al. [16]	Accuracy	69.33	68.16	-	74.08	56.83	-
		Precision	-	-	-	-	-	-
		F1-score	71.33	67.34	-	74.2	51.65	-
2	K. Aryal et al. [19]	Accuracy	78.58	96.54	96.54	87.41	-	75.16
		Precision	74.45	93.04	93.54	82.48	-	68.58
		F1-score	73.51	95.9	95.88	85.12	-	72.57
3	C. Dunn et al. [27]	Accuracy	-	79.22	-	-	73.11	76.91
		Precision	-	79.22	-	-	74.9	70.01
		F1-score	-	-	-	-	-	-
4	Our study	Accuracy	48.5	77.02	62.5	64.56	47.5	71.2
		Precision	48.3	58.79	58.56	63.29	62.1	71.2
		F1-score	61.5	60	57.4	61.04	61.04	62.79

is inversely proportional to Accuracy, Precision, Recall, and F1-score otherwise introduces sensitivity in the trained model. Variance and adversarial success rate (ASR) of the models are also calculated to analyze the stability of models against data poisoning attacks as given in Eq 1 and 2.

$$Variance(\sigma) = \frac{1}{n} \sum_{i=1}^N (x_i - \mu_M)^2 \quad (1)$$

where, n is the total number of samples in dataset, x is the dataset sample and μ_M is the average accuracy of the model.

$$ASR = \frac{S_{mis}}{D_p} \quad (2)$$

and, S_{mis} are the poisoned dataset samples that are successfully misclassified and D_p is the poisoned dataset.

4. Empirical Analysis of ML Algorithms for Data Poisoning

In this section, we have presented the main contributions of our paper where we have expanded the empirical analysis of poisoning machine learning with poisoning levels in $5 \leq \Delta P \leq 25$. Various parameters and features of ML algorithms are analyzed to provide insights into their different classification behaviors while training the same poisoned datasets D_p .

4.1. Experimentation Setup

We have built the test environment using the scikit-learn library. To conduct analysis, we have used the pandas and numpy libraries. The Matplotlib software is useful for creating ROC curves and scatter plots from datasets. We have used 20% of a dataset for testing if there are no particular test data available. We provide six test settings for every dataset in order to assess every machine learning method. Following the building of test settings, we gradually introduce adversarial data produced by label-flipping assaults.

4.2. Datasets

We have conducted a rigorous machine learning analysis, developing intrusion detection classifiers with four benchmarked datasets namely UNSW-NB15, CIC-IDS-2017, CTU-13, and BotDroid. UNSW-NB15 is based on captured network traffic with nine attacks included in it. It contains a total of 48 features containing packet level information that leverages us to conduct a detailed analysis for intrusion detection. Aligning it for binary classification, we have featured its records into benign and malignant. BotDroid is the next dataset included, which comprises 45 features in total which we have encoded with one-hot encoding. CTU-13 is also included to develop a binary IDS classifier in our research study which is developed to capture real-time traffic in thirteen different scenarios. and CIC-IDS-2017 is the last dataset included as part of our analysis. A brief description of all datasets is mentioned in Table 4. For the appropriate

Table 3: Detailed synopsis of relevant existing defense solutions against data poisoning attacks along with the parameter targeted to develop mitigation strategy and algorithms on which the respected proposed solutions are experimented

Research paper	Mitigation technique	Targeted parameter	Experimented algorithm
P. PK. Chan et al. [28]	Data sanitization with cluster training	Dataset features complexity vector	SVM
N. Baracaldo et al. [33]	Data segmentation and filtering using provenance feature	Provenance feature	SVM, Logistic Regression
A.Paudice et al. [34]	Training classifiers for individual dataset classes and calculate outliers with Empirical Cumulative Distribution Function	Dataset outliers	Linear classifiers
J. Geiping et al. [29]	Adversarial training to maximize adversarial loss to detect poisonous data points	-	Neural networks
L.Tao et al. [30]	Adversarial training to mitigate delusive attacks with different data distributions	Robust and non-robust features in data distributions	Neural Networks
T. Qin et al. [31]	Adversarial training with data augmentation in unlearnable examples to maximize error	-	Neural Networks

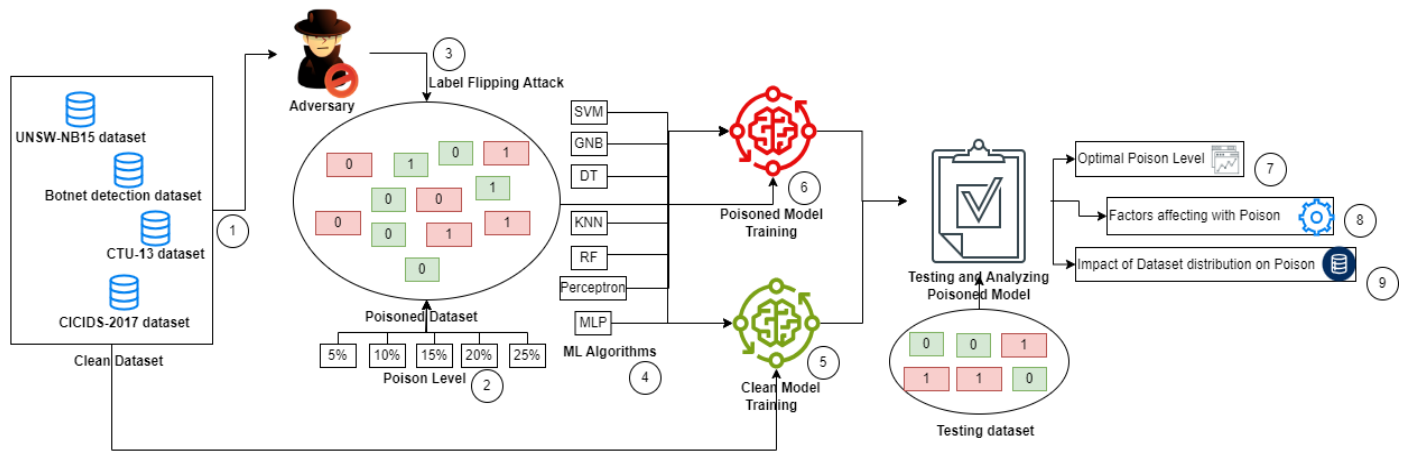


Figure 1: Architectural overview of our threat model

visualization and features correlation in the dataset, we have reduced dataset complexity to N dimensions with PCA, as given in Eq 3.

$$Dimentionality_Reduction = PCA(n_components = N) \quad (3)$$

And, for features structure and understanding, we have used Gaussian Mixture Models(GMM), as highlighted in Eq 4.

$$Features_Clusters = GMM(n_components = 2) \quad (4)$$

$$.fit(Reduced_Features).predict(Reduced_Features)$$

The visual dataset representation can be seen in Fig 2. BotDroid dataset is the most appropriately distributed dataset with aligned features and minimal distortion(outliers) as shown in Fig 2(a). Whereas UNSW-NB15 contains anisotropic features with asymmetric relation for which an uneven and elongated distribution is highlighted with sharp peaks at one axis in Fig 2(d). However, Fig 2(b) shows a symmetric correlation between some features including a few with anisotropic covariance with some overlapping outliers for CIC-IDS-2017. Fig 2(c) although reflecting an uneven distribution, but also contains some symmetric features with isotropic covariance. Conclu-

Table 4: Dataset description used to perform data poisoning on machine learning algorithms

S.No	Dataset	No. of features	No. of instances	Description
1	UNSW-NB15	48	82332	Dataset comprises of nine network intrusion attacks
2	BotDroid	45	1367	Dataset with limited instances, focusing BotDroid
3	CTU-13	57	92212	Captured with real time network traffic for BotDroid
4	CIC-IDS-2017	78	692703	Dataset consists of network packets

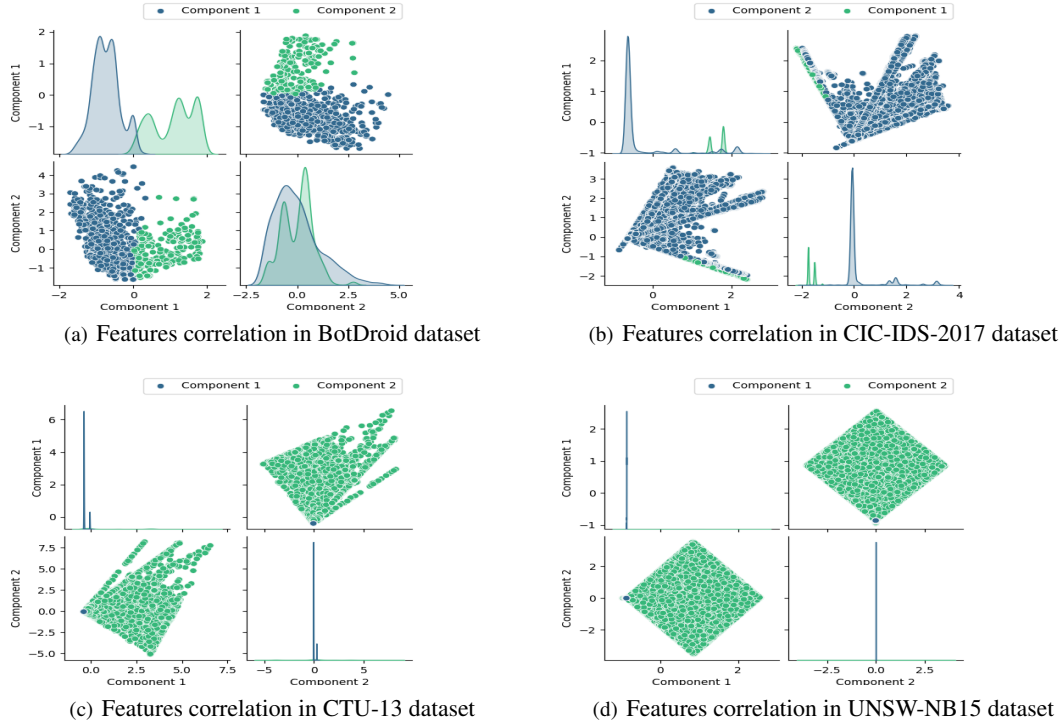


Figure 2: GMM visualization of features relationship in dataset with PCA reduction

sively, datasets with different characteristics help us in providing profound behavioral analysis of machine learning against a data poisoning attack.

4.2.1. Dataset Selection and Analysis Criteria

We assess individual characteristics of baseline machine learning models with cyber security benchmarked datasets in binary classification data poisoning settings. We have selected four cyber security application datasets for analyzing data poisoning against security-critical applications. Also, these datasets consist of various features, sizes, and features correlation that help understand the classifiers' behavior from various perspectives.

4.3. Results and Analysis for Poisoning Attacks

4.3.1. What is the optimal and complacent poisoning level?

We have plot accuracy, precision, recall, f1-score, and false positive rate against incrementing poisoning levels, as shown in Fig 5 to Fig 10. Data poisoning of 10-15% is identified as optimal poisoning levels whilst analyzing the models' performances. From the results, a sudden drift is visible between $10\% \leq \Delta P \leq 15\%$ of data poisoning whereas $\Delta P = 5\%$ has put a negligible impact and $\Delta P > 15\%$ make algorithms overfit. KNN trained with BotDroid dataset has the highest accuracy decrease of 15.48%, as shown in Fig 8. Following the fact that it classifies based on the principle of information gain analyzing maximum features similarities from the closest points. Our attack randomly poisoned data points which manipulate the relationship of the underlying features in the KNN model, making its decision boundary rough and irregular at $\Delta P = 15\%$, as

shown in Fig 3. However, DT is the most affected algorithm, with an average accuracy degradation of 14.42% and degradation score of 4.33% at $\Delta P = 25\%$, irrespective of datasets. Table 5 presents the change in feature importance score, for poisoned DT, making important features anomalous, degrading its performance, specifically for less noisy BotDroid dataset.

Further analysis interpreted that parametric algorithms including SVM, GNB, and Perceptron, are less affected by data poisoning when dataset features are symmetric and become saturated when $\Delta P \geq 15\%$. But these algorithms become sensitive and over-fitted when dataset features are asymmetric and $\Delta P \geq 15\%$. For example, the margin score of SVM is minimally changed from 0.0017 to 0.0013 when $\Delta P=10\%$ and 15% respectively. Features importance scores of SVM are also minimally affected by poisoning as given Table 6 where features space is reduced following Eq 3. So poisoned and cleaned SVM classifiers follow the same importance of features for making classification decisions except for the UNSW-NB-15 dataset for which Feature1 with the highest importance score has become an anomaly at $\Delta P = 15\%$, due to irregular features relations. Explaining further, SVM in Fig 5(b) and GNB in Fig 7(b) shows a high decrease of 35.14% in accuracy with 74% of increment in false positive rate but simultaneously, recall also increases to 33.1% between poisoning rate of 20-25%, highlighting overfitting of the models. We have also calculated the class probabilities of poisoned GNB to interpret their classification decision as given in Table 7 which a minimal impact on its classes leading to no change in its decisions. Conclusively, Perceptron losses its accuracy to 50.98% with the BotDroid dataset at $\Delta P = 25\%$ followed by DT, whose accuracy is reduced to

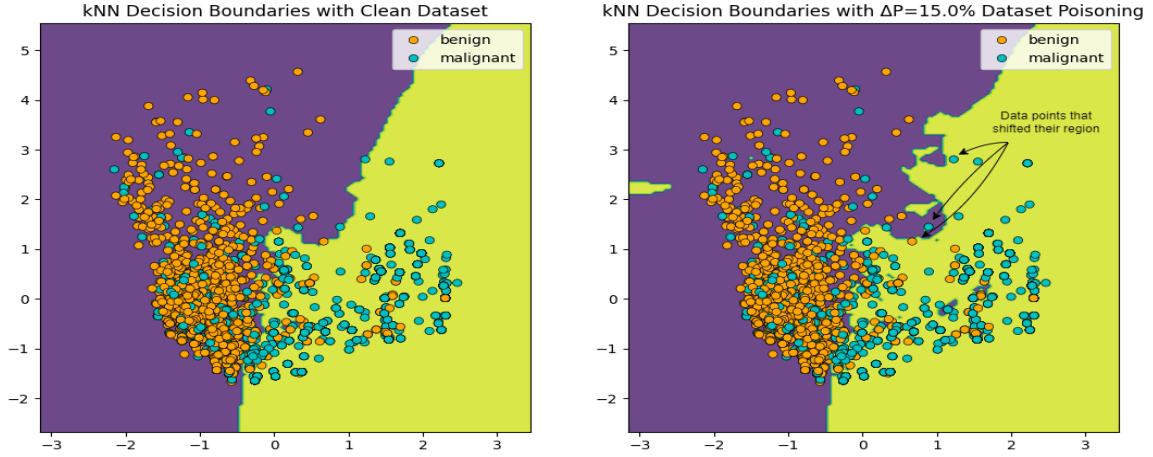


Figure 3: Change in Decision Boundary of KNN with 15% Dataset Poisoning

Table 5: Features importance score - DT where $\Delta P = (0\%, 10\%, 15\%)$

Dataset	Clean Dataset			Poisoned Dataset $\Delta P = 10\%$			Poisoned Dataset $\Delta P = 15\%$		
	Feature1	Feature2	Feature3	Feature1	Feature2	Feature3	Feature1	Feature2	Feature3
BotDroid	0.90	0.07	0.02	0.61	0.20	0.18	0.60	0.22	0.18
UNSW-NB15	0.43	0.29	0.22	0.33	0.34	0.32	0.33	0.34	0.32
CTU-13	0.19	0.32	0.49	0.26	0.34	0.40	0.26	0.34	0.39
CIC-IDS-2017	0.40	0.33	0.26	0.36	0.33	0.31	0.36	0.33	0.31

42.74% with the UNSW-NB15 dataset. Also, an interesting relation between poison and dataset noise is revealed where $\Delta P \propto \frac{1}{\text{Dataset Noise}}$ and working as a catalyst at $\Delta P = 25\%$. So, ΔP between 10%-15% are identified as complacent and effective poisoning rates, particularly affecting KNN followed by DT however minimally penetrating SVM and GNB. The variance of the trained models is given in Fig 4(a) to 4(d). Whereas, the adversarial success rate (ASR) of these poisoned models is given in Tables 8 to 11.

4.3.2. What factors and parameters are vulnerable to poison?

We next analyze individual factors affecting with our attack to study the behavior of individual algorithms. Overall, DT and KNN are prone to data poisoning. DT comprises a tree data structure following a probabilistic approach to structure decision nodes whereas our attack initiates wrong features split that supports misleading classification by manipulating features importance as shown in Table 5. We have reduced the features space of datasets for simplicity with PCA reduction following Eq 3. Also, it being a non-parametric model, we have analyzed a direct relation between poisoning and performance degradation. KNN is non-parametric and groups data based on feature similarities, making it highly susceptible to poisoning. Our attack changes the feature space of the part of the dataset, which disrupts the decision boundary and degrade its performance. Whereas, an inverse relation is identified no. of neighbors and poisoning level where increasing k-neighbors normalizes the poisoning effects on the KNN classifier we see in Table 12.

Because by inverting data labels, decision boundaries are getting blurred resulting in disjunctive classes split and hence more affected with poison. Whereas, SVM and GNB follow parametric learning to develop an optimal hyperplane and the parameters values and then become saturated which makes it independent of dataset size and robust to poisoned data when $\Delta P \geq 15\%$. In Table 13, we have calculated minimum margin scores of cleaned and poisoned SVM classifiers to analyze the change in its decision boundary which shows minimal change with little to no impact in its decision formulation. Also, looking at Fig 5(a) and Fig 5(c), SVM classifiers performance becomes saturated at $\Delta P \geq 15\%$ for BotDroid and CTU-13 datasets. But for UNSW-NB15 and CIC-IDS-2017 which comprises anisotropic features, SVM becomes sensitive and over-fit when $\Delta P \geq 15\%$, leading to inappropriate and continuous fluctuations in their learning parameters as shown in Fig 5(b) and Fig 5(d). A very similar behavior is visible for GNB and Perceptron can be seen in Fig 7 and Fig 10.

4.3.3. What are the cumulative effects of anisotropic features and imbalanced data distribution in poisoning machine learning?

Data poisoning is highly impacted on DT followed by KNN, irrespective of dataset noise and imbalanced data distribution as shown in Fig 9 and Fig 8, respectively. Whereas, interpreting Fig 6, Random Forest (RF) was found to be most effective and resilient against data poisoning attacks although a non-parametric algorithm. Because RF makes decisions calculating

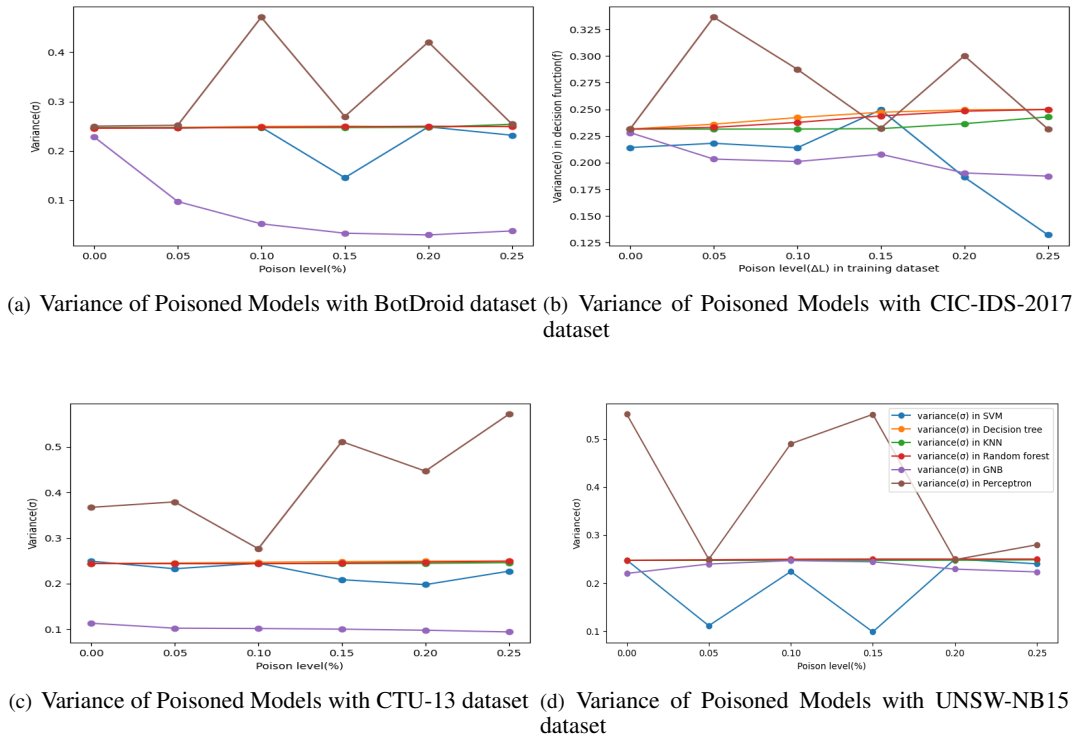


Figure 4: Variance analysis of Trained Models with consistent poisoning

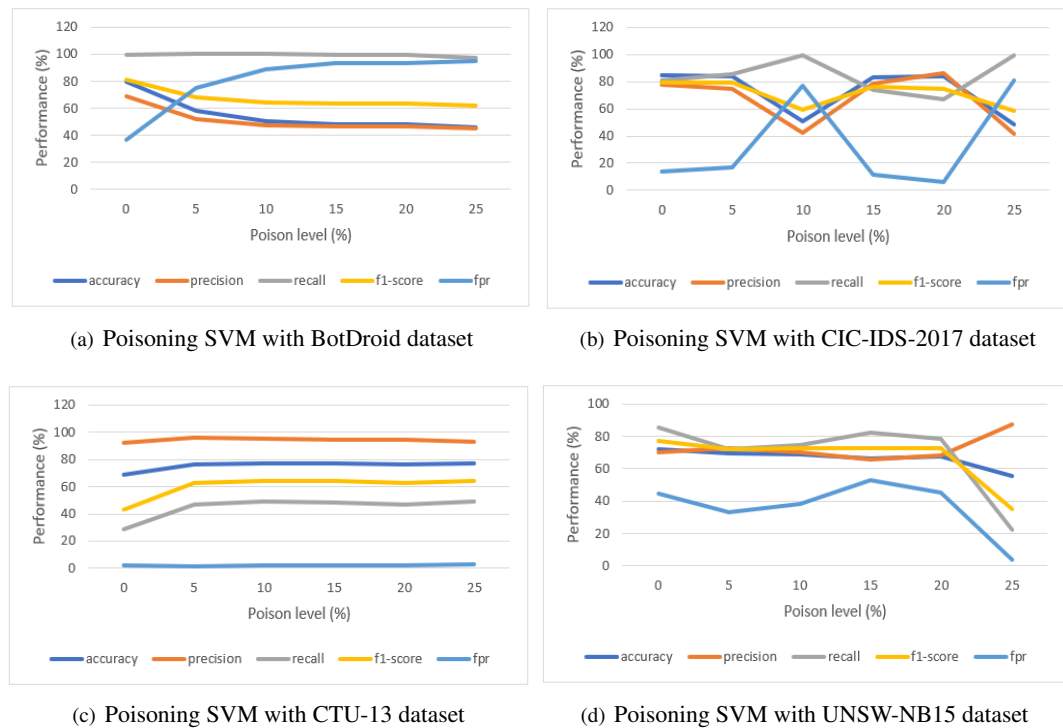
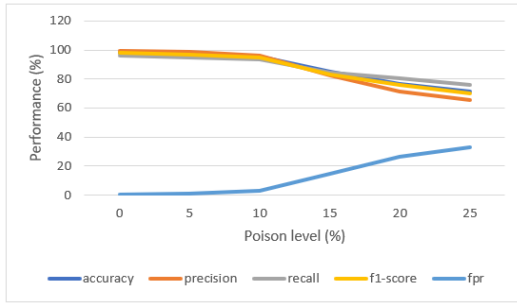
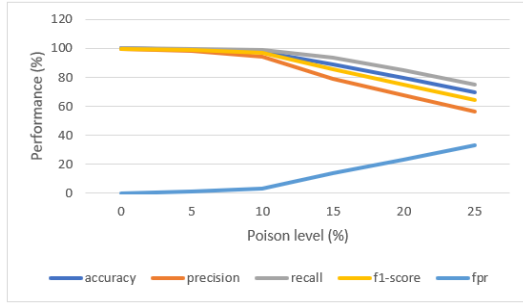


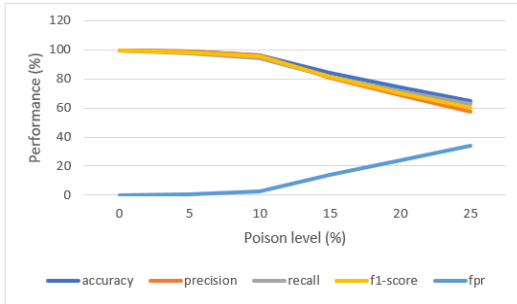
Figure 5: Performance analysis of Support Vector Machines(SVM) with consistent poisoning



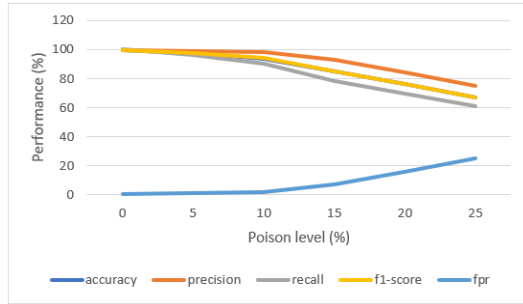
(a) Poisoning RF with BotDroid dataset



(b) Poisoning RF with CIC-IDS-2017 dataset

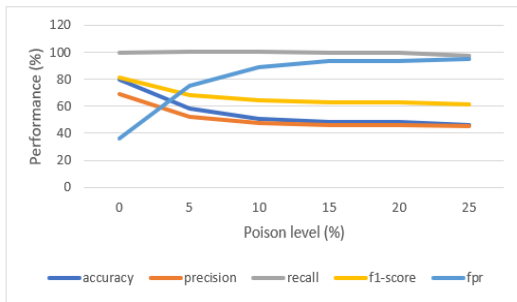


(c) Poisoning RF with CTU-13 dataset

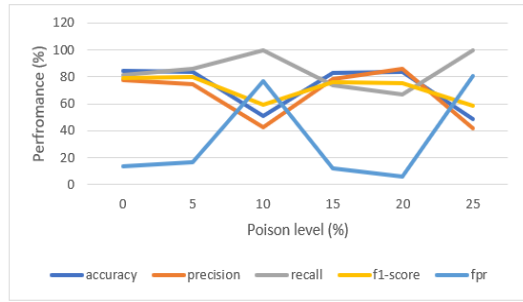


(d) Poisoning RF with UNSW-NB15 dataset

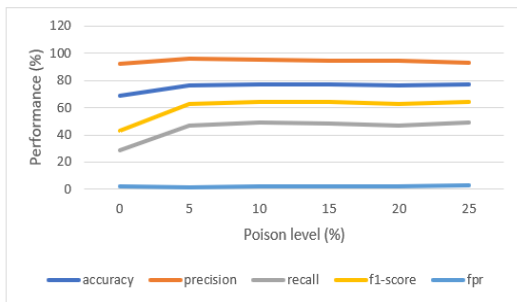
Figure 6: Performance analysis of Random Forest(RF) with consistent poisoning



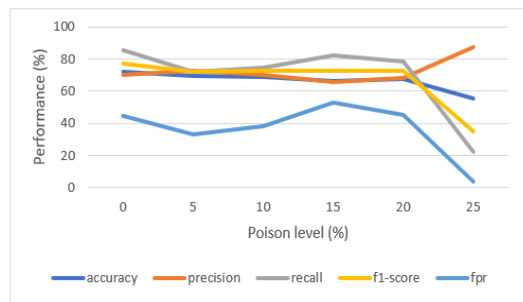
(a) Poisoning GNB with BotDroid dataset



(b) Poisoning GNB with CIC-IDS-2017 dataset

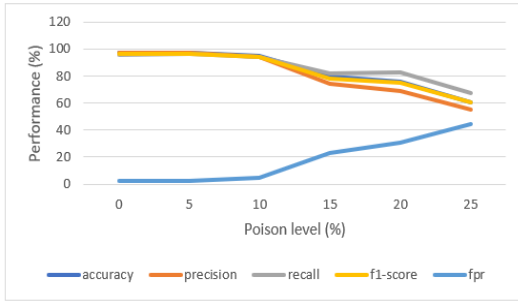


(c) Poisoning GNB with CTU-13 dataset

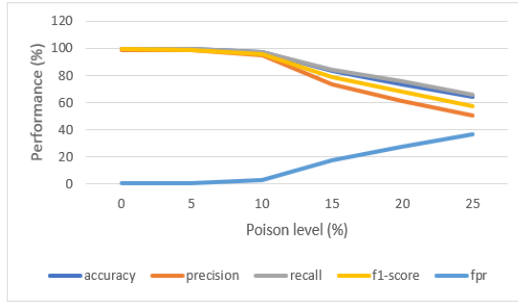


(d) Poisoning GNB with UNSW-NB15 dataset

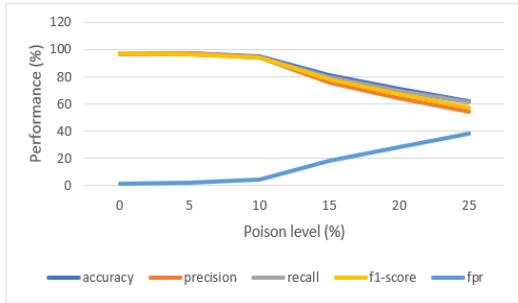
Figure 7: Performance analysis of Gaussian Naive Bayes(GNB) with consistent poisoning



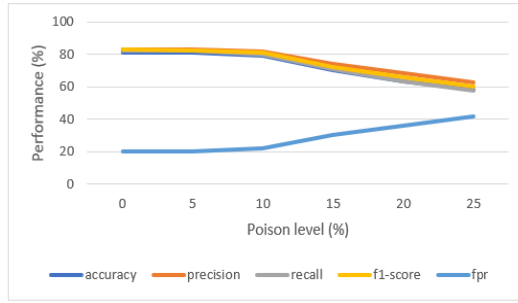
(a) Poisoning KNN with BotDroid dataset



(b) Poisoning KNN with CIC-IDS-2017 dataset

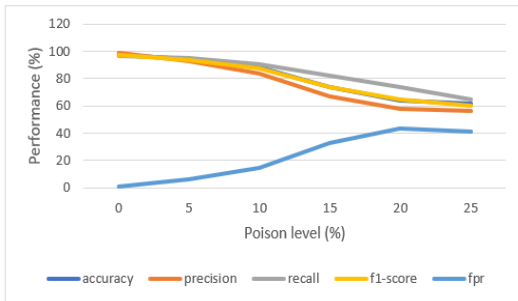


(c) Poisoning KNN with CTU-13 dataset

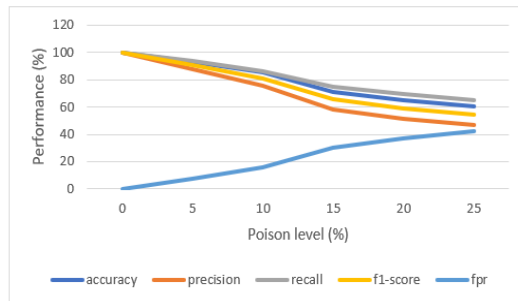


(d) Poisoning KNN with UNSW-NB15 dataset

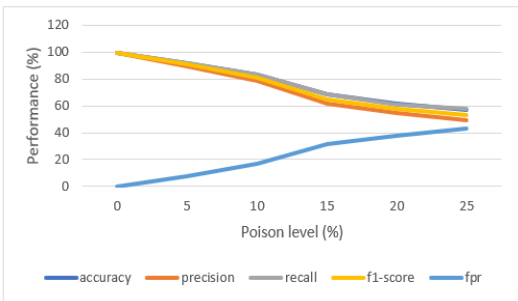
Figure 8: Performance analysis of K-Nearest Neighbours(KNN) with consistent poisoning



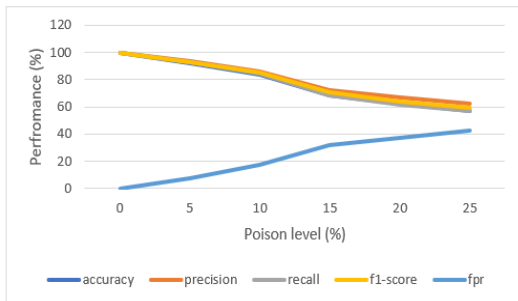
(a) Poisoning DT with BotDroid dataset



(b) Poisoning DT with CIC-IDS-2017 dataset



(c) Poisoning DT with CTU-13 dataset



(d) Poisoning DT with UNSW-NB15 dataset

Figure 9: Performance analysis of Decision Tree(DT) with consistent poisoning

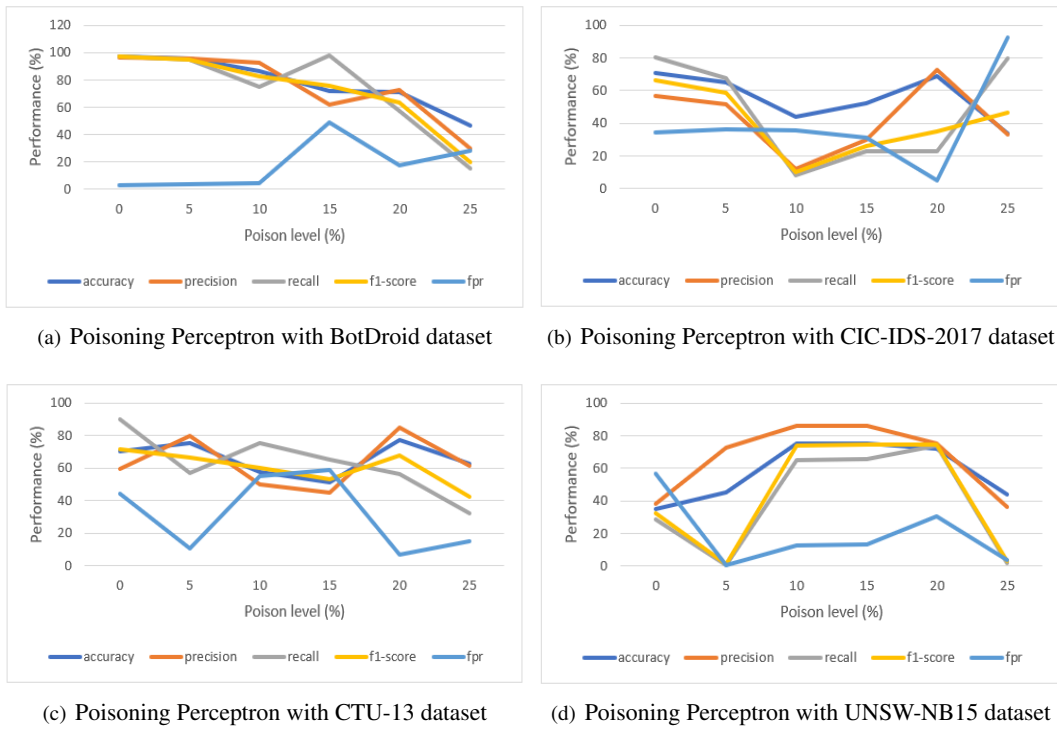
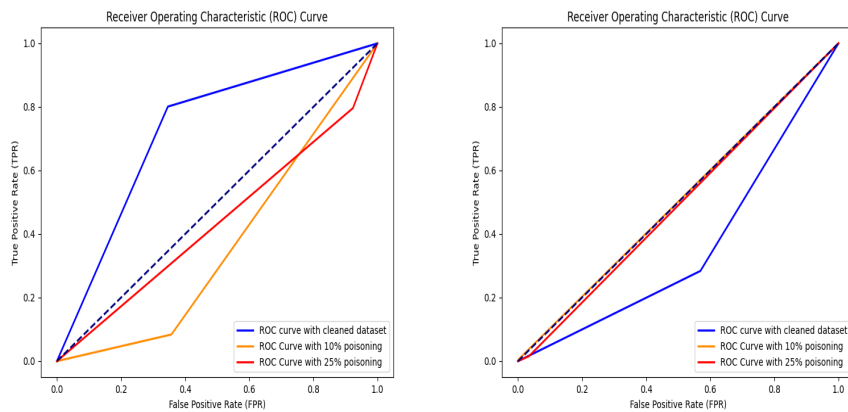


Figure 10: Performance analysis of Perceptron with consistent poisoning

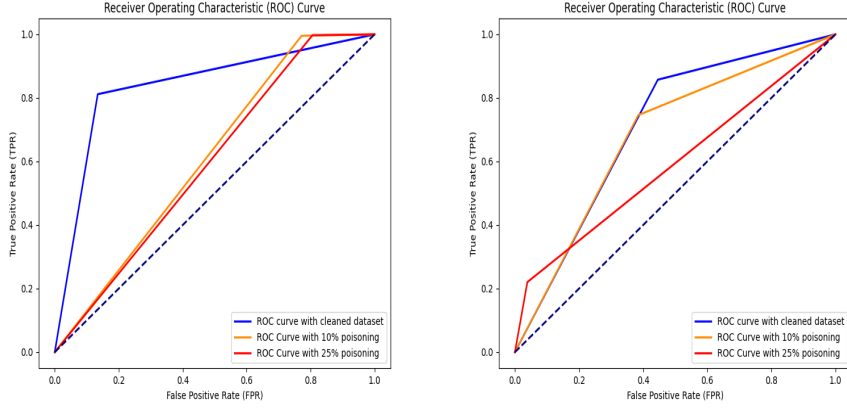


(a) Poisoning Perceptron with CIC-IDS-2017 dataset (b) Poisoning Perceptron with UNSW-NB15 dataset

Figure 11: ROC curve of Perceptron with consistent poisoning

Table 6: Features importance score - SVM where $\Delta P = (0\%, 10\%, 15\%)$

Dataset	Clean Dataset			Poisoned Dataset $\Delta P = 10\%$			Poisoned Dataset $\Delta P = 15\%$		
	Feature1	Feature2	Feature3	Feature1	Feature2	Feature3	Feature1	Feature2	Feature3
BotDroid	0.53	0.80	0.23	0.42	0.80	0.40	0.18	0.98	0.03
UNSW-NB15	0.81	0.56	0.10	0.72	0.32	0.61	0.30	0.72	0.61
CTU-13	0.13	0.50	0.85	0.06	0.52	0.84	0.38	0.21	0.89
CIC-IDS-2017	0.07	0.99	0.11	0.53	0.69	0.47	0.45	0.79	0.39



(a) Poisoning GNB with CIC-IDS-2017 dataset (b) Poisoning GNB with UNSW-NB15 dataset

Figure 12: ROC curve of GNB with consistent poisoning

Table 7: Analyzing class probabilities of GNB with poisoned dataset

Dataset	Clean Dataset		$\Delta P = 10\%$		$\Delta P = 15\%$	
	Class0	Class1	Class0	Class1	Class0	Class1
BotDroid	0.55	0.45	0.54	0.46	0.52	0.48
UNSW-NB15	0.45	0.55	0.46	0.54	0.47	0.53
CTU-13	0.57	0.43	0.56	0.44	0.54	0.46
CIC-IDS-2017	0.64	0.36	0.61	0.39	0.58	0.42

the mean from all of its trees, which normalizes the data poisoning effects. However, particularly for intrusion detection, KNN is immune to poison specifically at 10% although a high drift in performance can be seen at 25% dataset poisoning. Fig 13 provides an overview of the change in dataset distribution when labels are attacked with our Label-Flipping attack at different poisoning levels. Looking at classification performances in Fig 11(a) and Fig 11(b), dataset noise and asymmetric features correlating with poison making Perceptron sensitive to classify intrusions resulting in an unstable classifier. For parametric algorithms i.e. SVM, GNB, and Perceptron, poison in an erroneous dataset works as a catalyst for performance disruption. For example, Fig 12(a), Fig 12(b) and Fig 11(a) shows a significant decrease in TPR concurrently increasing FPR with 10% poisoned dataset for GNB and Perceptron, respectively. From our analysis, Table 14 provides a relative impact of ΔP on individ-

ual parameters of machine learning algorithms. Implications of our behavioral analysis and its future directions are given in Section 5.

5. Behavioural Analysis Implications in Mitigating Data Poisoning

5.1. Analysing Existing Mitigation Techniques

Extending Table 3, data sanitization and adversarial training are the two most effective techniques to safeguard against data poisoning attacks. Adversarial training is developed with the gradient of the data point which is effective for neural networks only. It can not enhance the security of baseline ML models including SVM, DT, RF, GNB, and KNN. However, data sanitization filters the dataset's ambiguity to clean the classifier's training by removing doubtful dataset features or outliers. These sanitization techniques are limited to be developed on the dataset features only. None of the techniques focus on the classifiers' inherent parameters and their development mechanisms. Furthermore, A. E. Cina et al. [37] highlighted the limitations of existing mitigation techniques and the importance of the generalizability of ML models. Another research study [4] experimentally proved the limitations of adversarial training in security baseline models. Considering these limitations, it is important to understand the underlying decision model development and the impact of poisoned data points on it.

Table 8: Adversarial Success Rate (ASR) with BotDroid Dataset

Algorithm	$\Delta P = 5\%$	$\Delta P = 10\%$	$\Delta P = 15\%$	$\Delta P = 20\%$	$\Delta P = 25\%$
SVM	0.2	0.29	0.31	0.31	0.33
DT	0.097	0.13	0.22	0.35	0.33
RF	0.025	0.065	0.15	0.21	0.22
KNN	0.01	0.021	0.16	0.19	0.33
GNB	0.2	0.29	0.31	0.31	0.33
Perceptron	0.055	0.11	0.24	0.25	0.52

Table 9: Adversarial Success Rate (ASR) with UNSW-NB15 Dataset

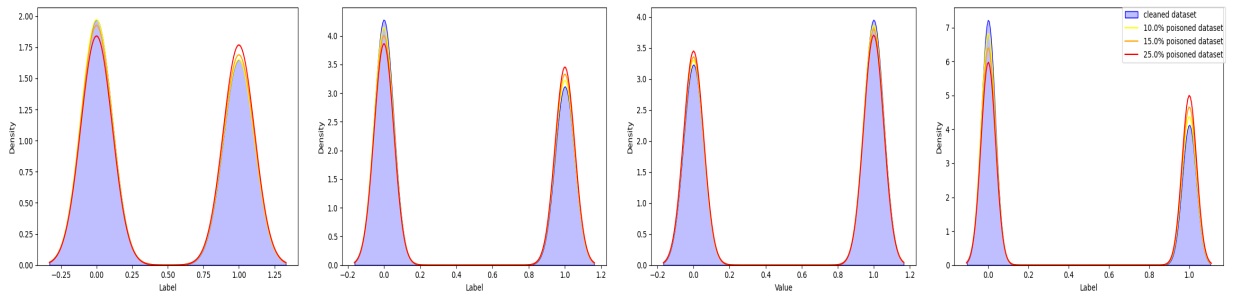
Algorithm	$\Delta P = 5\%$	$\Delta P = 10\%$	$\Delta P = 15\%$	$\Delta P = 20\%$	$\Delta P = 25\%$
SVM	0.034	0.071	0.10	0.12	0.15
DT	0.12	0.16	0.27	0.35	0.39
RF	0.03	0.06	0.11	0.22	0.28
KNN	0.0074	0.02	0.1	0.17	0.22
GNB	0.034	0.071	0.10	0.12	0.15
Perceptron	-0.063	-0.4	-0.39	-0.33	-0.08

Table 10: Adversarial Success Rate (ASR) with CTU-13 Dataset

Algorithm	$\Delta P = 5\%$	$\Delta P = 10\%$	$\Delta P = 15\%$	$\Delta P = 20\%$	$\Delta P = 25\%$
SVM	-0.06	-0.08	-0.07	-0.09	-0.06
DT	0.07	0.16	0.28	0.36	0.39
RF	0.01	0.03	0.18	0.21	0.35
KNN	0.009	0.02	0.16	0.26	0.35
GNB	-0.06	-0.08	-0.07	-0.09	-0.06
Perceptron	-0.07	0.12	0.19	-0.09	0.03

Table 11: Adversarial Success Rate (ASR) with CIC-IDS-2017 Dataset

Algorithm	$\Delta P = 5\%$	$\Delta P = 10\%$	$\Delta P = 15\%$	$\Delta P = 20\%$	$\Delta P = 25\%$
SVM	0.03	0.33	0.03	-0.007	0.36
DT	0.06	0.14	0.30	0.36	0.39
RF	0.001	0.02	0.08	0.20	0.27
KNN	0.003	0.02	0.18	0.29	0.33
GNB	0.03	0.33	0.03	-0.007	0.36
Perceptron	0.03	0.26	0.17	0.02	0.34



(a) Label distribution with poisoned BotDroid dataset

(b) Label distribution with poisoned CTU-13 dataset

(c) Label distribution with poisoned UNSW-NB15 dataset

(d) Label distribution with poisoned CIC-IDS-2017 dataset

Figure 13: Analyzing change in dataset distribution D_{dis} with label poisoning using *kernel density estimation* function at an incremental poison levels $\Delta P=(0\%, 10\%, 15\%, 25\%)$.

5.2. Implications of Behavioural Analysis

Table 12: Analyzing k-neighbors affecting KNN accuracy with $\Delta P = (0, 10, 15, 25)\%$

	Poison Level	k=3	k=5	k=10	k=15
BotDroid	$\Delta P = 0\%$	97.54	96.80	96.92	96.68
	$\Delta P = 10\%$	95.94	96.80	96.31	96.68
	$\Delta P = 15\%$	83.78	90.05	95.57	94.47
	$\Delta P = 25\%$	69.28	70.76	87.22	88.69
CTU-13	$\Delta P = 0\%$	97.80	97.55	96.94	96.55
	$\Delta P = 10\%$	95.06	96.52	96.78	96.50
	$\Delta P = 15\%$	87.13	90.90	94.54	95.95
	$\Delta P = 25\%$	72.0	76.14	83.68	87.52
UNSW-NB15	$\Delta P = 0\%$	81.57	81.17	80.92	80.71
	$\Delta P = 10\%$	79.85	80.36	80.63	80.63
	$\Delta P = 15\%$	74.17	76.46	79.06	80.04
	$\Delta P = 25\%$	64.48	66.85	70.19	74.62
CIC-IDS-2017	$\Delta P = 0\%$	99.61	99.57	99.49	99.42
	$\Delta P = 10\%$	96.21	98.84	99.42	99.41
	$\Delta P = 15\%$	88.70	92.93	96.79	98.91
	$\Delta P = 25\%$	73.49	78.06	87.09	90.43

Table 13: Analyzing SVM margin score for different datasets with $\Delta P = (0, 10, 15)\%$

Dataset	$\Delta P = 0\%$	$\Delta P = 10\%$	$\Delta P = 15\%$
BotDroid	0.008	0.00047	0.006
CTU-13	0.000007	0.00002	0.0000003
UNSW-NB15	0.00010	0.00015	0.000019
CIC-IDS-2017	0.00003	0.000015	0.000018

Table 14: Analyzing one-to-one relation between poison and various parameters of ML algorithms

Algorithm	Algorithmic Parameters	Relation to ΔP
SVM	Margin score	Minimal impact
	Decision boundary	Minimal impact
	Features importance score	Minimal impact
DT	Features importance score	High impact
	Asymmetric features space	High impact
KNN	Decision boundary k-neighbors	High impact Inverse impact
GNB	Decision boundary	Minimal impact
	Class probabilities	Minimal impact
RF	No. of trees	Inverse impact
Perceptron	Weights	High impact

To enhance the security of machine learning models and their generalizability against data poisoning, it is crucial to study the behavior of inherent characteristics of the models. Our results highlighted potentially vulnerable parameters of individual algorithms and their acceptability of poison at various levels. Also, highlighting the most pervasive poisoning levels allows the development of potential security solutions in this regard. Including this behavioral analysis in data sanitization and in setting model parameters will better generalize the model. Also, our behavioral analysis will help strengthen the explainability of the model’s decision. Considering future research directions, our behavioral analysis will be highly effective in developing adversarial training techniques for baseline machine learning models other than perceptron. Conclusive outcomes of our research and its limitations are discussed in Section 6.

6. Discussion and Limitations

• Strengthen mitigation against data poisoning attacks:

Existing research papers [28], [33], and [34] provided solutions to mitigate data poisoning with data sanitization, filtering poisonous data points based on provenance features and adversarial training based on detected outliers, respectively. Whereas, these solutions are developed focusing on cleaning dataset features and do not understand the model classification mechanism. Our research provides a further detailed synopsis of technical uncertainties in the model’s decisions created with a poisoned dataset to help strengthen these mitigations. As we didn’t understand the model’s decision mechanisms in existing solutions, ML models are still susceptible to poisoning, even secured data sanitization or adversarial training, as described in Section 5. Our results highlight certain affected factors with label poisoning, analyzing individual machine learning algorithms. Such as segregating nodes hierarchy development in DT and RF, misplacing data points in KNN and SVM, and probability misinterpretation in KNN and GNB. Overall RF is resilient to data poisoning because it may create more hierarchical nodes within its trees with poisoned data points, averaging their prediction probabilities dilutes misclassification. Although DT is the most affected algorithm because flipped labels make it difficult to disjoint nodes considering features. Also, our results showed that a 10%-15% poisoned dataset is more harmful and complacent whereas, after 25% poisoning, the sensitivity of a model is getting increased.

• Impact of the dataset structure on poisoned machine learning algorithms:

Our results highlight that noise in the dataset works as a catalyst for data poisoning. The more noisy the dataset, the more adverse impact can be achieved in performance degradation against classification results because it filters poison (intentional noise) from dataset noise. Also, non-linear features in the dataset help

poison to better penetrate the classifiers. Whereas, cleaning datasets and feature engineering can help mitigate data poisoning in this scenario.

- **Limitations:** Our research is limited to the practical implication and analysis of supervised classification algorithms for binary classifiers. This limitation helps to provide a detailed and thorough technical study of supervised machine learning classifiers and their parameters that are affected by poisoning. Unsupervised machine learning can not be covered in this research as we have focused on label poisoning and have contributed to analyzing decision boundaries of classifiers and resultant metrics including accuracy, precision, f1-score, and recall. For the in-depth analysis, we have also included the variance and adversarial success rate (ASR) of the baseline models and compared poisoned with the benign models to understand the disrupted behaviors. We only focused on the baseline ML models which are the foundation of complex deep neural networks and advanced models. which Regression algorithms are also out of the scope of this study.

7. Conclusion and Future Work

7.1. Conclusion

Our study has provided technical insights that 10%-15% poisoning to the dataset is optimal whereas less than 10% has minimal impact and more than 25% introduce sensitivity and 50% of data poisoning leads to overfitting. Also highlights different behaviours towards performance degradation of machine learning algorithms and puts forward the facts that mathematically interpret why DT is the most affected algorithm against poisoning whereas RF is resilient to it. Also, we have experimented impact of data poisoning when the training dataset is imbalanced and its features are anisotropic, and that concluded that anisotropic or asymmetric features serve as catalysts to data poisoning between 10%-20% of poisoning level whereas increased the sensitivity of the models specifically of neural networks.

7.2. Future Work

- Complex ML models have witnessed increase in adversarial attempts in recent years which motivates us to extend this study to such models. In this respect, we aim to evaluate deep neural networks (DNN) including convolutional neural networks (CNN) and recurrent neural networks (RNN) against data poisoning to understand its impact on their behavior, model parameters and training epochs.
- This study focused on the impact of data poisoning with a label-flipping attack to understand the vulnerabilities of models' parameters and their relation to data poisoning. In future, we aim to extend our research to more complex data poisoning attacks to understand the behaviors of the models against such attacks.

- In this study, we identified that some of algorithms such as perceptron suffered from overfitting across all datasets. Similarly, some algorithms such as SVM produced overfitted models for select datasets. We are interested in investigating this further to understand the underpinning characteristics of algorithms and datasets which lead to overfitting when trained with poisoned datasets.

CRedit authorship contribution statement

Anum Paracha - Problem Statement, Conceptualization, Investigation, Formal Analysis, Writing - Original draft.
Junaid Arshad - Conceptualization, Writing - Original draft.
Mohamed Ben Farah - Conceptualization, Writing - Review and Update.
Khalid Ismail - Conceptualization, Writing - Review and Update.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available upon request.

Acknowledgement

No acknowledgments to declare.

References

- [1] Austin Brown, Maanak Gupta, and Mahmoud Abdelsalam. Automated machine learning for deep learning based malware detection. *Computers & Security*, 137:103582, 2024.
- [2] Najeem Moharram Jebreel, Josep Domingo-Ferrer, David Sánchez, & Alberto Blanco-Justicia, LFighter: Defending against the label-flipping attack in federated learning. *Neural Networks*. pp. 111-126. 2024
- [3] Orson Mengara, A Backdoor Approach with Inverted Labels Using Dirty Label-Flipping Attacks. *IEEE Access*. 2024
- [4] Anum Paracha, Junaid Arshad, Mohamed Ben Farah, and Khalid Ismail. Exploring Multi-Class Data Poisoning Against Adversarially Trained Skin Cancer Diagnostics. Submitted to IEEE/ACM International Conference on Utility and Cloud Computing, 2024
- [5] Hui Wang, Zijian Cao, and Bo Hong. A network intrusion detection system based on convolutional neural network. *Journal of Intelligent & Fuzzy Systems*, 38(6):7623–7637, 2020.
- [6] Qasem Abu Al-Haijaa and Abdelraouf Ishtaiwia. Machine learning based model to identify firewall decisions to improve cyber-defense. *International Journal on Advanced Science, Engineering and Information Technology*, 11(4):1688–1695, 2021.
- [7] Sin-Ye Zhong, Po-Yen Tseng, Natnuntnita Siriphockpirom, Chih-Hsien Hsia, Ming-Shih Huang, Kai-Lung Hua, and Yung-Yao Chen. An automated biometric identification system using cnn-based palm vein recognition. In *2020 international conference on advanced robotics and intelligent systems (ARIS)*, pages 1–6. IEEE, 2020.
- [8] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *Computer Security—ESORICS 2020: 25th European Symposium on Research*

- in *Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part 1* 25, pages 480–501. Springer, 2020.
- [9] Minh-Hao Van, Wei Du, Xintao Wu, and Aidong Lu. Poisoning attacks on fair machine learning. In *International Conference on Database Systems for Advanced Applications*, pages 370–386. Springer, 2022.
- [10] Ebtihaj Alshahrani, Daniyal Alghazzawi, Reem Alotaibi, and Osama Rabie. Adversarial attacks against supervised machine learning based network intrusion detection systems. *Plos one*, 17(10):e0275971, 2022.
- [11] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-Guided backdoor poisoning attacks against malware classifiers. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1487–1504. USENIX Association, August 2021.
- [12] Jonas Geiping, Liam Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches’ brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*, 2020.
- [13] Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping, Micah Goldblum, Tom Goldstein, and David Jacobs. Autoregressive perturbations for data poisoning. *Advances in Neural Information Processing Systems*, 35:27374–27386, 2022.
- [14] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, pages 1–47, 2022.
- [15] Chen Zhu, W. Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets, 2019.
- [16] Fahri Anil Yerlikaya and Şerif Bahtiyar. Data poisoning attacks against machine learning algorithms. *Expert Systems with Applications*, 208:118101, 2022.
- [17] Hojjat Aghakhani, Dongyu Meng, YuXiang Wang, Christopher Kruegel, and Giovanni Vigna. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pages 159–178. IEEE, 2021.
- [18] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.
- [19] Kshitiz Aryal, Maanak Gupta, and Mahmoud Abdelsalam. Analysis of label-flip poisoning attack on machine learning based malware detector. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4236–4245. IEEE, 2022.
- [20] Qianqian Xu, Zhiyong Yang, Yunrui Zhao, Xiaochun Cao, and Qingming Huang. Rethinking label flipping attack: From sample masking to sample thresholding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [21] Diego Ihara Centurion, Karine Chubarian, Bohan Fan, Francesco Sgherzi, Thiruvankadam S Radhakrishnan, Anastasios Sidiropoulos, and Angelo Straight. Geometric algorithms for k -nn poisoning. *arXiv preprint arXiv:2306.12377*, 2023.
- [22] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3104–3122, 2021.
- [23] Robin Mayerhofer and Rudolf Mayer. Poisoning attacks against feature-based image classification. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, pages 358–360, 2022.
- [24] Chen Zhang, Zhuo Tang, and Kenli Li. Clean-label poisoning attack with perturbation causing dominant features. *Information Sciences*, 644:118899, 2023.
- [25] Chenxi Zhu, Haobo Wang, Yan Zhuang, Jie Li, Yangjie Cao, et al. A sparsity-limitation-based high-dimensional distribution searching algorithm for adversarial attack. *Journal of Sensors*, 2022, 2022.
- [26] Kangjie Chen, Xiaoxuan Lou, Guowen Xu, Jiwei Li, and Tianwei Zhang. Clean-image backdoor: Attacking multi-label models with poisoned labels only. In *The Eleventh International Conference on Learning Representations*, 2022.
- [27] Corey Dunn, Nour Moustafa, and Benjamin Turnbull. Robustness evaluations of sustainable machine learning models against data poisoning attacks in the internet of things. *Sustainability*, 12(16):6434, 2020.
- [28] Patrick PK Chan, Zhi-Min He, Hongjiang Li, and Chien-Chang Hsu. Data sanitization against adversarial label contamination based on data complexity. *International Journal of Machine Learning and Cybernetics*, 9:1039–1052, 2018.
- [29] Jonas Geiping, Liam Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller, and Tom Goldstein. What doesn’t kill you makes you robust (er): How to adversarially train against data poisoning. *arXiv preprint arXiv:2102.13624*, 2021.
- [30] Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, and Songcan Chen. Better safe than sorry: Preventing delusive adversaries with adversarial training. *Advances in Neural Information Processing Systems*, 34:16209–16225, 2021.
- [31] Tianrui Qin, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. *arXiv preprint arXiv:2303.15127*, 2023.
- [32] Samuel Drews, Aws Albarghouthi, and Loris D’Antoni. Proving data-poisoning robustness in decision trees. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 1083–1097, 2020.
- [33] Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, and Jaehoon Amir Safavi. Mitigating poisoning attacks on machine learning models: A data provenance based approach. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 103–110, 2017.
- [34] Andrea Paudice, Luis Muñoz-González, Andras Gyorgy, and Emil C Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. *arXiv preprint arXiv:1802.03041*, 2018.
- [35] Zhu, Y., Wen, H., Wu, J. & Zhao, R. Online data poisoning attack against edge AI paradigm for IoT-enabled smart city. *Mathematical Biosciences And Engineering*. **20**, 17726-17746 (2023)
- [36] Das, A., Tariq, A., Batalini, F., Dhara, B. & Banerjee, I. Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer. *MedRxiv*. (2024)
- [37] Cinà, A., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B., Oprea, A., Biggio, B., Pelillo, M. & Roli, F. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*. **55**, 1-39 (2023)
- [38] Seraj, Saeed, BotDroid: Android Botnet Detection. (2022), <https://www.kaggle.com/datasets/saeedseraj/botdroid-android-botnet-detection>
- [39] David, Mr Wells, UNSW_NB15. (2018), <https://www.kaggle.com/datasets/mrwellsdavid/unswnb15/data>
- [40] Chenthan, H N, Network Intrusion dataset (CIC-IDS-2017). (2023), <https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset>
- [41] Malik, Faisal, CTU13-CSV-Dataset. (2022), <https://github.com/imfaisalmalik/CTU13-CSV-Dataset>

Appendix A. Detailed Performance analysis of Machine Learning Classifiers

We extend our experimentation to provide detailed analysis of classification results with further performance statistics and ROC curves, highlighting change in TPR (True Positive Rate) and FPR (False Positive Rate) when trained with poisoned dataset.

Table A.15: Performance statistics with clean UNSW-NB15 Dataset

Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	72.13	70.27	85.73	77.23	44.59
DT	99.90	99.91	99.91	99.91	0.10
RF	99.87	99.78	99.98	99.88	0.26
KNN	81.17	83.36	82.28	82.82	20.18
GNB	72.13	70.27	85.73	77.23	44.59
Perceptron	34.97	38.01	28.39	32.50	56.93

Table A.16: Performance statistics with clean BotDroid Dataset

Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	79.72	68.88	99.72	81.48	36.44
DT	97.78	98.59	96.42	97.50	1.11
RF	98.15	99.43	96.42	97.90	0.44
KNN	96.92	97.47	95.60	96.53	2.00
GNB	79.72	68.88	99.72	81.48	36.44
Perceptron	97.29	96.72	97.25	96.98	2.66

Table A.17: Performance statistics with clean CIC-IDS-2017 Dataset

Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	84.59	77.622	81.17	79.35	13.44
DT	99.95	99.91	99.96	99.94	0.04
RF	99.95	99.91	99.95	99.93	0.04
KNN	99.45	98.98	99.53	99.25	0.58
GNB	84.59	77.62	81.17	79.35	13.44
Perceptron	70.75	57.06	80.10	66.64	34.61

Table A.18: Performance statistics with clean CTU-13 Dataset

Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	68.64	91.96	28.35	43.34	1.81
DT	99.76	99.75	99.68	99.71	0.18
RF	99.81	99.76	99.78	99.77	0.16
KNN	97.55	97.57	96.62	97.09	1.76
GNB	68.64	91.96	28.35	43.34	1.81
Perceptron	70.09	59.72	89.98	71.79	44.49

Table A.19: Performance statistics with 10% Poisoned UNSW-NB15 Dataset

Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	68.72	70.44	74.57	72.45	38.47
DT	83.58	85.85	84.09	84.96	17.04
RF	93.83	98.11	90.55	94.18	2.13
KNN	79.09	81.58	80.18	80.88	22.25
GNB	68.72	70.44	74.57	72.45	38.47
Perceptron	75.15	86.40	65.20	74.32	12.61

Table A.20: Performance statistics with 10% Poisoned CIC-IDS-2017 Dataset

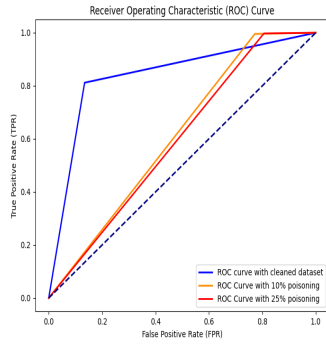
Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	50.81	42.55	99.55	59.62	77.17
DT	85.08	75.95	86.47	80.87	15.72
RF	97.53	94.51	98.97	96.69	3.29
KNN	96.96	94.61	97.21	95.89	3.17
GNB	50.81	42.55	99.55	59.62	77.17
Perceptron	43.88	11.89	8.39	9.84	35.72

Table A.21: Performance statistics with 10% Poisoned BotDroid Dataset

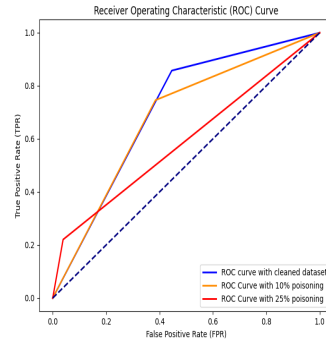
Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	50.73	47.58	100	64.48	89.11
DT	88.08	83.79	90.93	87.22	14.22
RF	95.57	96.32	93.68	94.98	2.88
KNN	94.84	94.47	93.95	94.21	4.44
GNB	50.73	47.58	100	64.48	89.11
Perceptron	86.24	92.85	75	82.97	4.66

Table A.22: Performance statistics with 10% Poisoned CTU-13 Dataset

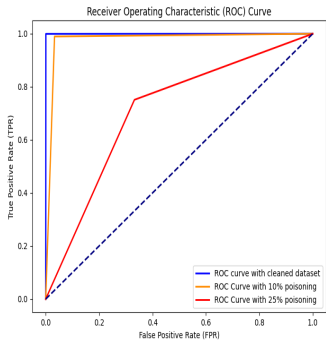
Algorithm	Accuracy	Precision	Recall	F1-score	FPR
SVM	77.25	95.11	48.73	64.44	1.83
DT	83.38	78.60	83.45	80.95	16.65
RF	95.95	96.04	94.30	95.17	2.84
KNN	94.86	93.96	93.89	93.93	4.42
GNB	77.25	95.11	48.73	64.44	1.83
Perceptron	57.71	50.01	75.25	60.08	55.15



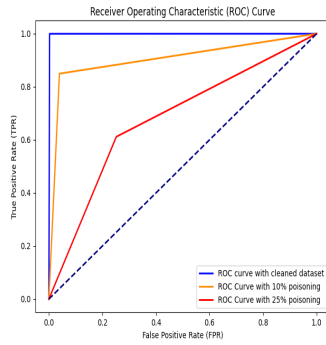
(a) Poisoning SVM with CIC-IDS-2017 dataset



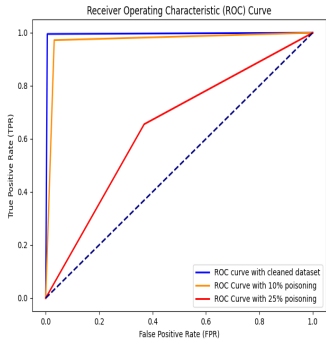
(b) Poisoning SVM with UNSW-NB15 dataset



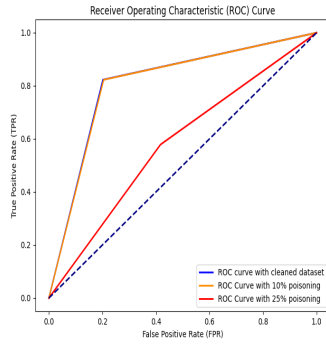
(a) Poisoning RF with CIC-IDS-2017 dataset



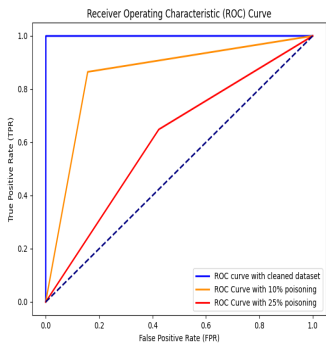
(b) Poisoning RF with UNSW-NB15 dataset



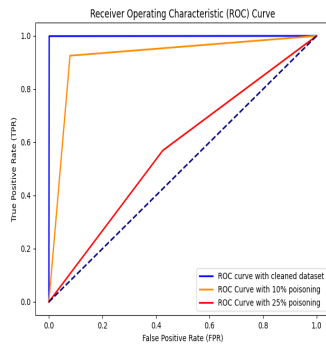
(a) Poisoning KNN with CIC-IDS-2017 dataset



(b) Poisoning KNN with UNSW-NB15 dataset



(a) Poisoning DT with CIC-IDS-2017 dataset



(b) Poisoning DT with UNSW-NB15 dataset

Figure A.17: ROC curve of Decision Tree(DT) with consistent poisoning