[1]Nilambari Mate
[2]Dr. Jyoti Yadav

# Challenges in Safeguarding Machine Learning Models Against Adversarial Attacks


***Abstract:*** - The usefulness of Machine Learning (ML) has been proven in a variety of application settings, making it one of the most frequently studied areas today. Applications with significant social influence rely on automated judgments made using machine learning faces several concerns regarding possible vulnerabilities brought by machine learning algorithms. Intelligent attackers possess powerful incentives to tamper with produced findings and models by algorithms that use machine learning to accomplish their goals using adversarial attacks. Adversarial attacks can be done using many ways like contaminating models training data or altering models testing data or polluting a central model. If model training data is manipulated by an attacker as part of an attack, the model's capacity to predict accurate results is negatively impacted. This is called as data poisoning attack. Small perturbation may result in large side effects on the output of the Machine learning model. This paper list out various strategies to poison the training data. It also analyzes various attacking strategies and summarizes defensive techniques used to prevent or detect data poisoning attack. It also shows the impact of poisoning attack on the machine learning models through experimental results. Finally, this paper highlights various research opportunities to create robust model by preventing data poisoning attack.

***Keywords:*** Data poisoning attack, type of attacks, attacking techniques, defensive techniques.

## I. INTRODUCTION

Machine Learning (ML) has become a revolutionary technology with far-reaching implications across diverse fields, strengthening its position as a crucial research area with immense practical applications. The applications of ML are extensive and diverse, extending across sectors like healthcare, finance, and transportation (Wang et al., 2022). For example, ML algorithms play a crucial role in predicting disease outbreaks (Panayides et al., 2020), improving resource distribution in education (Xiang et al., 2022) and streamlining supply chain operations (Khedr & Ahmed M, 2024). These instances showcase how ML can tackle significant societal issues by leveraging extensive data and advanced algorithms to derive valuable insights that direct decision-making processes. The rapid processing of large datasets empowers organizations to address challenges promptly, underscoring ML's efficacy as a potent instrument for positive societal impact.

The foundation of information system security is based on the CIA triad, consisting of three essential principles: Confidentiality, Integrity and Availability. Each element is vital in protecting data and guaranteeing the efficient and secure functioning of information systems. The violation of this CIA triad can be done by malicious actors with sophisticated capabilities through adversarial attacks (Jagannathan et al., 2022). Many real-time applications frequently receive dynamic data for retraining, which increases the risk of exposure of ML model to adversarial attacks. These attacks can take diverse forms, such as contaminating training data, manipulating testing data or polluting centralized models (Liu et al., 2018). According to the three primary categories of security violations (i.e. integrity, availability and confidentiality) considered when examining the security of a system, the attacker's goals are categorized along three dimensions: availability breakdown, integrity violations, and privacy compromise.

1. Availability breakdown: An attacker decreases the confidence level or performance of the classifier.
2. Integrity violations: It attempts to influence the classifier to predict incorrectly.
3. Privacy compromise goes a step further and makes sure that the incorrect classification isn't simply any other label, but the one that is chosen by the attacker specifically (Jagielski et al., 2018; Oprea Alina & Apostol Vassilev;2023).

The causative attack where an adversary perturb the training data by inserting or modifying training samples to achieve their goals is called as data poisoning attacks (Xiao et al., 2015; Huang et al., 2015; Zhao et al., 2015) . As

[1]Department of Computer Science, Savitribai Phule Pune University, Pune, India. nilambari.mate@gmail.com

[2]Department of Computer Science, Savitribai Phule Pune University, Pune, India. yadav.jyo@gmail.com

it prevents the model from producing accurate predictions, it may be called as an integrity attack. The main issue with Data poisoning attack is that it's difficult to fix. Models are retrained with recently gathered information at specific spans, depending upon their expected use and their proprietor's preference. Since poisoning typically occurs over the long time, and over some number of preparing cycles, it may be difficult to tell when prediction accuracy begins to move (Lucian Constantin, 2021). An attacker has the ability to poison part of training data to allow the trained model to meet training objectives by considering conflicting samples in the training dataset. Often, these conflicting patterns are perceived by the adversary to have similar characteristics to the malicious samples but the labels are incorrect, causing a variation in the distribution of the training data. As a result of this model performance is degraded with reference to accuracy, precision, recall, loss function etc. (Liu et al., 2018). For example, as a part of business campaign, an attacker may decrease the output of classifier with the intention of damaging a business or recommending its own product (Wang, Yizhen, and Kamalika Chaudhuri, 2018). Face recognition systems can be attacked to evade confidentiality of sensitive data (Sharif et al., 2016; Biggio et al., 2013). In addition, malicious actors can take control of autonomous vehicles (Papernot et al., 2017) and voice control systems (Carlini et al., 2016) to make poor decisions about identifying road traffic signs and respective voice commands. Another example is Tay, Microsoft's chat bot that began to post inflammatory and offensive tweets learned by an attacker through its Twitter account (Baracaldo et al., 2017). Poisoning attacks also have been validated in various applications like worm signature generation (Newsome et al., 2006; Predict et al., 2006), spam detection (Nelson et al., 2008), analysis of network traffic to detect DoS attacks (Rubinstein et al., 2009), opinion mining on social media (Newell et al., 2014), crowd analysis (Wang et al., 2014), and health-care (Mozaffari-Kermani et al., 2014).

In this paper, a summarized information about data poisoning attack is presented. Section 2 provides review of research and development. In section 3 gives a detail discussion of the results based on the survey done which mainly includes training phase attacks, detailed review of defensive techniques and effect of data poisoning attack through experimental results. Section 4 concludes the paper by giving some encouraging challenges.

## II. Review of Research and Development

Most of the existing work focused on poisoning attacks mainly include strategy of polluting the training data or defensive approach to protect the ML algorithm, or both. Various attacking techniques and defensive techniques are proposed in the literature.

### *A. Attacking techniques*

Basically, many security threats to ML arise from adversarial samples (Biggio & Battista, 2016). Oprea et al., 2023 described four key dimensions which are used to form comprehensive adversarial model as attacking strategy, goal, capability and knowledge. Adversarial samples are the harmful inputs purposely injected by an attacker that lead to a decline in the effectiveness of ML models. Performance of model is mainly subject to the quality of the data employed to train the model. Hence, many attacker targets training data to decrease the overall performance of ML models. Availability and integrity of ML models is compromised by modifying training data termed as data poisoning attack which is a kind of causative attack (Li et al., 2016; Fang et al., 2018; Ma et al., 2018). An attacker can evaluate potential inputs before launching attack during testing phase by creating substitute model which uses some or all training data (Biggio et al., 2012). The effect of data poisoning attack is evaluated on different ML algorithms like Logistic Regression, Naïve Bayes, Random Forests and Neural Networks (Adam-Bourdarios et al., 2015; Zhao et al., 2017). Poisoning attack also threatens SVM algorithm (Xiao et al., 2015; Burkard et al., 2017). Single-linkage and complete-linkage hierarchical clustering models are attacked using data poisoning (Biggio et al., 2014; Battista et al., 2014; Ignazio Pillai et al. 2013). NN models are also targeted by data poisoning attack (Yang et al. 2017; Shafahi et al., 2018; Huang et al 2021).

The attack strategy outlines how the attacker manipulates data to execute the planned poisoning attack. Various contexts such as white-box scenarios (Xiao et al., 2015), gray-box settings (Jagielski et al., 2018) and black-box models (Biggio et al., 2011) are utilized in the literature for the development of data poisoning attacks. Poisoning attacks were initially introduced in cyber security applications to generate malicious flows intentionally to deceive signature generation algorithms such as Polygraph (Perdisci et al., 2006). Spam emails containing extended sequences of words found in legitimate emails are crafted to target Bayes-based spam classifiers, leading to the misclassification of such spam emails (Nelson et al., 2008). An attacker can maximize the hinge loss for SVM or maximize the Mean Squared Error (MSE) for regression by manipulating training samples (Jagielski et al., 2018;

Biggio et al., 2012). Noise is added to the training samples of a generative model, which is then utilized to train the NN model using clean label attacks (Feng et al., 2019). Clean-label poisoning involves gradient alignment method to make minimum modifications to the training data (Fowl et al., 2021). AdvFaces generate minor perturbations in images that are hard to detect yet successful in tricking ML models. The technique employs methods to pinpoint crucial facial features essential for recognition. By concentrating on these areas, the strategy guarantees that the induced perturbations significantly affect the efficiency of face recognition systems (Deb et al., 2020). The susceptibility of object detection models to clean-label poisoning attacks is evident when boats are mistakenly classified as ferries, potentially hindering the identification of pirates approaching a boat (Lee et al., 2023).

### *B. Defensive techniques*

Poisoning attacks may often be detected by monitoring the basic performance indicators of ML models like accuracy of the model, false positive rate, loss function, mean squared error, loss function and area under the curve, since they significantly degrade the classifier metrics. These types of attacks should be detected at training stage not at deployment/training stage. To create strong model these attacks must be prevented during the training. Several techniques are used in the literature to prevent Data Poisoning attack. Some of the widely used techniques are outlier detection, smoothing, data sanitization, robust training, ensemble method and many more (Jagielski et al 2018).

#### *i. Data Sanitization*

These techniques are used to remove poisonous samples from training dataset so that only benign samples will be used to train the classifier. Defensive technique against data poisoning attack called Reject on Negative Impact (RONI) measures the effect of each training data point on the classifier's accuracy. The data points which have a negative effect on classifiers accuracy are removed from training data set. This technique provides better accuracy but it is complex to configure and operate (Barreno et al., 2010). To detect the effect of distinct data point on the performance of the trained model, Probability of Sufficiency (PS) method is used. Here model evaluation is done by comparing its performance on a trusted data set (Chakarov et al., 2016). Defense against data poisoning attack is shown using data provenance strategy which uses extra information about the data point that led to its formation, origin and manipulation (Baracaldo et al., 2017). Label flipping attack is defended using K-Nearest-Neighbours (k-NN) to identify malicious samples or data points that has harmful impact on the performance of classifiers(Paudice et al., 2019). An outlier detection method is used to identify adversarial sample against Data Injection Attack (Steinhardt et al., 2017). Clustering methods have also been used against poisoning attack (Laishram Ricky and Vir Virander Phoha, 2016; Taheri et al., 2020). Data filtering and ensemble learning is also used against label flipping attack (Venkatesan et al., 2021). The datasets should be secured by cyber security mechanism for authentication of dataset origin and integrity after being cleaned (Schmidt Eric. 2023).

#### *ii. Robust training*

Trimmed loss function as a defensive technique is used against statistical attack for regression models (Wang et al., 2022). Randomized smoothing technique is used against label flipping attack to provide certificate of robustness (Rosenfeld et al., 2020). An ensemble of multiple models and subset aggregation as well as randomized smoothing is used against data poisoning attack for neural networks (Levine Alexander and Soheil Feizi, 2020). Finite aggregation method with ensemble of multiple models used as defensive technique (Wang et al., 2022).Using a meta-algorithm, the learner may be made more robust against outliers by starting with a basic learner like least squares or stochastic gradient descent (Diakonikolas et al., 2019). Table 1 shows detailed information about defenses against Data Poisoning Attack.

## III. RESULTS AND DISCUSSIONS

An attacker uses different strategies to violate the confidentiality, integrity and availability of ML model by manipulating training data, testing data or centralized model through adversarial attacks. Various kinds of adversarial attacks are used by an attacker to take advantage of vulnerability in ML model. *A. Types of Adversarial Attacks:* Various techniques used by malicious attackers to manipulate training or testing data with harmful intentions as shown in Figure 1.1(Liu et al., 2018; Tabassi et al., 2019).

### ***1.* Testing Phase Security Threats**

**a)** **Evasion Attack**: In this attack, the attacker injects carefully crafted adversarial samples at test time to classify them incorrectly. However, the information that attacker have about how the model produces its predictions will determine how they craft these samples. The adversary just adds the crafted samples in the model and keep an eye on their selection (Biggio et al., 2013).

**i.Gradient-based Attacks:** This generally involves algorithms that utilize gradient-based search techniques such as

- **Limited Memory Broyden Fletcher Goldfarb Shanno (L-BFGS):** It was the first algorithm employed to produce misclassifications in a computer vision system model by using input perturbations that were not noticeable to human observers (Szegedy, C, 2013).

- **Fast Gradient Sign Method (FGSM**): FGSM (Kurakin et al., 2018) enhances the computational efficiency of gradient ascent with a Single Step approach, eliminating the need for iterations to obtain a perturbation that causes a significant change in the loss function (Malik et al., 2024).

- **Jacobian-based Saliency Map Attack (JSMA***)*: The JSMA algorithm (Papernot et al., 2016) is an iterative method that offers finer control over perturbed features, enabling the creation of more convincing adversarial examples, albeit at a higher computational cost.

**ii.Gradient Free Attacks:** Adversarial examples are generated without depending on the gradient of the loss function but usually require access to the model's confidence scores to be effective (Chen et al., 2018).

**b)** **Oracle Attacks:** In this attack scenario, an assailant interacts with the model through an Application Programming Interface, providing inputs and observing the corresponding outputs. The input-output pairs obtained from these oracle attacks can be employed to train a surrogate model that imitates the behavior of the target model (Tang et al., 2024). Oracle Attacks encompass techniques such as Extraction Attacks, Inversion Attacks, and Membership Inference Attacks (Papernot et al., 2018).

**i.Extraction Attacks:** The adversary derives the parameters or structure of the model by analyzing its predictions, often utilizing the probabilities returned for each class (Papernot et al., 2018).

**ii.Inversion Attacks:** An adversary can reconstruct data used to train the model, potentially exposing personal information and compromising individual privacy, enabled by inferred characteristics (Alves et al., 2019; Papernot et al., 2018).

**iii.Membership Inference Attack:** A membership inference attack involves an adversary determining whether the dataset contains data from a particular individual in the dataset to train the model (Hu et al., 2022).

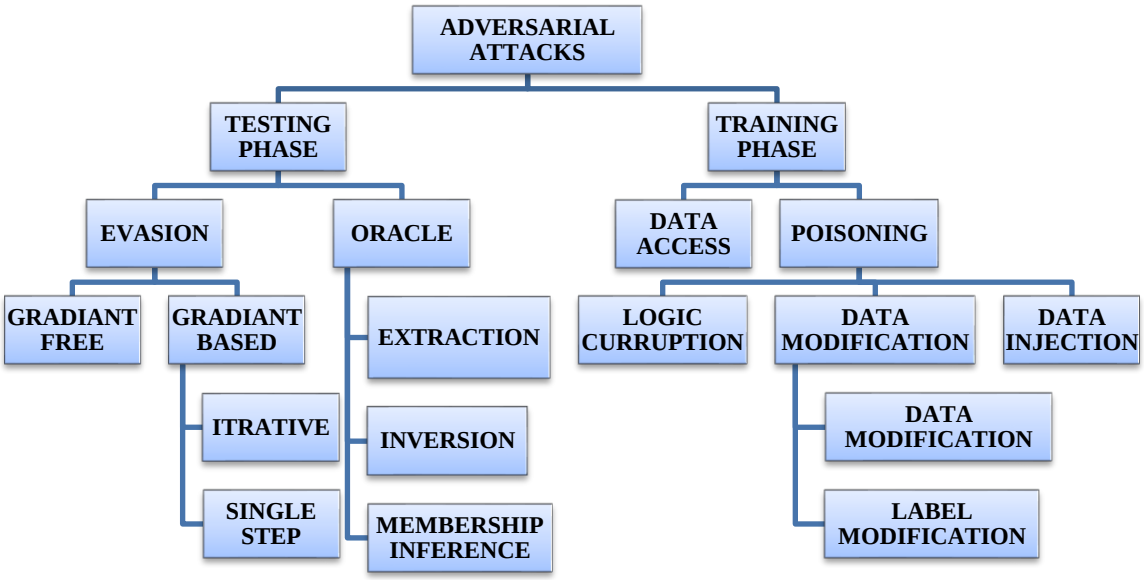


**Fig. 1.1: Types of Adversarial Attacks**

*2. Training Phase Security Threats:*

**a) Data Access Attacks**: This attack creates substitute model by accessing some or all the training data. The efficiency of probable inputs may be evaluated using this stand-in model before they are submitted as attacks during the testing phase of operation (Liu et al., 2018).

**b) Data Poisoning Attacks:** The causative attack where an attacker perturb the training data by inserting malicious samples or modifying training samples to achieve their goals is called as data poisoning attacks (Xiao et al., 2015; Huang et al., 2015; Zhao et al., 2017). As it prevents the model from producing accurate predictions, it may be called as an integrity attack. The main issue with data poisoning attack is that it's difficult to fix. Models are retrained with recently gathered information at specific spans, depending upon their expected use and their proprietor's preference. Since poisoning typically occurs over the long time, and over some number of preparing cycles, it may be difficult to tell when prediction accuracy begins to deteriorate (Lucian Constantin, 2021). An attacker attempts to poison part of training data to make the trained model meet desired objectives by considering malicious samples in the training dataset. These malicious samples are perceived by the attacker to have similar characteristics as non-poisonous samples but the labels are incorrect, causing a variation in the distribution of the training data. As a result of this model performance is degraded with reference to accuracy, precision, recall, loss function etc.

Following are the basic two types of Data Poisoning Attacks:

i. **Indirect Data Poisoning:** The attacker poisons the data before pre-processing (Tabassi et al., 2019).

ii. **Direct Data Poisoning:** Attacker may use the following 3 broad strategies with respect to level of data access and process of learning:

- **Data Injection:** The attacker corrupts the target model by injecting malicious samples into the training data, as they lack access to the training data and learning algorithm but have the capability to add new data to the training set (Steinhardt et al., 2017; Abbas et al., 2019; Chakraborty et al., 2021).
- **Data Modification:** A malicious user has full access to the training data but lacks access to the learning algorithm. Desired model is built on the modified data (Aladag et al., 2019; Chakraborty et al., 2021). It can be done using two ways:

**I. Label Modification:** An attacker can modify the class label anticipated by the model.

**II. Input Modification:** An attacker manipulates the original training data by introducing intentional modifications, such as injecting malicious examples, altering existing data points, or adding noise.

- **Logic Corruption:** The learning algorithm is interfered by an adversary. These attacks are referred as logic corruption attack. Apparently, counter measures against these types of attacks are very hard because an attacker attacks the model itself (Newaz et al., 2020).

*B. Comparative Study of Defensive Techniques*

Attacks mentioned above can substantially undermine the classifier metrics and should be identified during the training phase rather than at deployment. Prevention of these attacks during training is crucial for building robust models. Detecting poisoning attacks can be achieved by monitoring key performance metrics of ML models, such as model accuracy, FPR, loss function, mean squared error, and area under the curve. Various techniques, including outlier detection, smoothing, data sanitization, robust training, ensemble methods, amongst others, are commonly employed in the literature to mitigate the risks posed by data poisoning attacks (Jagielski et al., 2018). Researchers have made two different categories of defensive strategies against poisoning attack. One is proactive defense where designer aims to stay ahead of adversaries by simulating possible attacks, analyzing their consequences, and crafting appropriate countermeasures as needed, and another is reactive defense where designer responds to the attack by assessing its impact and developing countermeasures (Biggio et al, 2013).

1. Reactive Defense Strategies: Strategies aimed at assessing the severity of attacks and implementing corrective actions to mitigate their adverse effects are known as reactive defense measures. These countermeasures are commonly utilized to safeguard systems against vulnerabilities, with several of the defensive techniques listed below falling under the reactive defense classification.

*a) Defensive Techniques for SVM:* SVM is commonly employed supervised learning algorithm for classification task. Several strategies to address poisoning attacks in ML against SVM have been developed by various domain experts as shown in Table 1.1.

***Table 1.1 Defenses against SVM***

| Attack | Method Used | Dataset | Advantages | Challenges |
|---|---|---|---|---|
| Data injection attack | Influence function-based method (Fang et al., 2020). | Amazon Digital Music (Music) and Yelp | For a given target item, influence function-based strategy is suggested to identify the influential user set effectively. Suggested attack performs better than current ones. | It assumes that an attacker is aware of recommender system, its parameters and all rating data (Fang et al., 2020). |
| Label modification attack | Auto-encoder based defense (Aladag et al., 2019). | MNIST | To make classification models more resistant to label modification attacks, an auto-encoder model is used (Aladag et al., 2019). | The study does not explore model diversity and gradient-based optimization in an autoencoder (Aladag et al., 2019). |
| Data Injection Attack | Data Sanitization (Steinhardt et al., 2017). | MNIST, IMBD, Dogfish | Outlier removal method is used against a broad family of attacks by minimizing risk (Steinhardt et al., 2017). | Defensiveness is dataset dependent and performance of a deployed learning algorithm is not guaranteed. |
| Data poisoning Attack | Reject on Negative Impact (Barreno et al., 2010). | SpamBayes | RONI measures the effect of each training data point on accuracy of the classifier. The data points which have a negative effect on classifiers accuracy are removed from training data set. | RONI technique provides better accuracy but it is complex to configure and operate (Barreno et al., 2010). |
| Label flipping: attack | Data filtering and ensemble learning is used (Venkatesan et al., 2021). | KDD Cup 99 dataset | Data poisoning attempts that add or remove data are successfully resisted by data filtering techniques. Data poisoning attacks that include modifying the labels of the data can be successfully defended against using ensemble learning techniques (Venkatesan et al., 2021). | 1. An attacker can simply get over data filtering mechanisms. 2. The computational cost of ensemble learning approaches can be high, and they need a lot of training data. 3. Model's performance can be degrading on clean data. |
| Label Flipping attack | Clustering is used with data filtering (Laishram et al., 2016). | MNIST | Clustering based method has been proposed which successfully filters some of the poison points from dataset used for retraining. Proposed method can be easily integrated into any system (Laishram et al., 2016). | 1. Proposed method is tested against only one dataset. 2. Here only one kind of attacking strategy is tested out of four label flipping attack strategies. |

*b) Defenses against Other ML Models:* ML models like regression, decision tree and K-nearest neighbor are also used to test the effect of poisoning attack described in Table 1.2.

***Table 1.2 Defenses against Other ML Models***

| Attack | Algorithm | Method | Dataset | Advantages | Challenges |
|---|---|---|---|---|---|
| Statistical attack | OLS, LASSO, ridge, and elastic net | Trimmed loss function | Health care, loan assessment, | A defense algorithm called TRIM proposed by | Author assumes that attacker has an access to model parameters |

| | | | | | |
|---|---|---|---|---|---|
| | regression (Jagielski et al., 2018). | | and real estate datasets (Jagielski et al., 2018). | Jagielski et al., 2018 that significantly outperforms current robust regression techniques is created by using a principled design methodology. | which is not always true. Secondly, only Regression models are used to test the defense against data poisoning attack. |
| Label flipping attack | Classification | k-Nearest-Neighbors (k-NN) based detection method | BreastCancer, MNIST and SpamBase | Label sanitization is achieved using KNN based defense effectively (Paudice et al., 2019). | Proposed mechanism works well for BreastCancer and SpamBase datasets but not for MNIST. It is assumed that the attacker has full knowledge of the learning algorithm, loss function, training data, and features used in model training, which is impractical. |

*c) Defenses Against Neural Network (NN) Models:* Another branch of ML proven to be highly successful in various fields such as security, autonomous vehicle systems and biometric recognition named as deep learning is based on neural networks. However, these applications are susceptible to tampered data or artificial inputs. The accuracy of models can be significantly impacted by alterations in the input data, leading to the creation of adversarial examples. Table 1.3 represents strategies used in the literature to defend against data poisoning attacks.

***Table 1.3 Defenses against NN Models***

| Attack | Method | Dataset | Advantages | Challenges |
|---|---|---|---|---|
| Label flipping attack | Silhouette clustering method (Taheri et al., 2020). | Drebin, Contagio and Genome | Proposed method used clustering-based Semi-supervised defense, have higher Accuracy than the KNN-based Semi-Supervised defense (Taheri et al., 2020). | The proposed method uses higher computational resources, leading to increased complexity. Another concern is the accuracy of the defense technique when applied with different classification algorithms. |
| Label-flipping and poisoning attacks | Subset aggregation as well as randomized smoothing (Levine et al., 2020). | MNIST and CIFAR-10 | An ensemble of multiple models has been proposed which may be used to strengthen the resilience of ML models and offers a verifiable defenses against poisoning attempts. | An ensemble based defense computationally expensive and not scalable for large datasets due to the use of multiple base models. |
| Poisoning attacks | Deterministic finite aggregation techniques (Wang et al., 2022). | MNIST, CIFAR-10, and GTSRB | The study by Wang et al. (2022) recommends the Finite Aggregation method over the Deep Partition method for its resilience. Additionally, their proposed approach includes | Proposed method is computationally expensive. |

| | | | | |
|---|---|---|---|---|
| | | | certified defenses utilizing both deterministic and stochastic aggregation. | |

*d) Proactive Defense Strategies:* Following Table 1.4 shows the example of proactive defensive technique against data poisoning attack.

***Table 1.4 Proactive Defense Methods***

| Method used | Dataset | Advantages | Challenges |
|---|---|---|---|
| A framework for practical assessment of a classifier (Biggio et al., 2013). | spam filtering, biometric authentication and network intrusion detection | The proposed framework empowers classifier designers to proactively devise countermeasures against potential attacks by foreseeing and analysing possible attack scenarios through "what if" analysis (Biggio et al., 2013). | Implementing the proposed framework may demand substantial efforts and expertise, with the outcomes being greatly influenced by the dataset used. |
| Optimized Jaccard distance (Sameen et al., 2022). | pollution sensor datasets | Proactive detection framework named DISTINCT was proposed to detect data poisoning attack using optimized Jaccard distance. The suggested framework uses less space and is much faster than existing Jaccard Distance variations (Sameen et al., 2022). | It was considered that trusted dataset is genuine and certified. |
| Randomized Smoothing (Rosenfeld et al., 2020). | MNIST, CIFAR10, IMDB and Dogfish | The suggested framework effectively offers a certificate of robustness for every test point, thereby showcasing enhanced accuracy in countering poisoning attacks (Rosenfeld et al., 2020). | The efficacy of randomized smoothing relies significantly on the selection of noise parameters, which can introduce complexity to model design and training while ensuring certification (Rosenfeld et al., 2020). |

*C. Effect of Poisoning Attack on ML Model*

The performance of ML models is significantly compromised by data poisoning attacks. Section II describes different ways to attack ML model. As a result of these attacks model makes incorrect predictions or decisions. Overall performance of the ML model is degraded by decreasing model accuracy or by increasing FPR. Model integrity and reliability can be compromised with the presence of poisoned data in the training dataset. This section estimates the influence of data poisoning attack on classifier.

*1. Datasets*

The experimental study involves the utilization of both the MNIST and RandomBlob datasets.

- MNIST dataset: MNIST dataset contains images of handwritten digits 0-9. The proposed model uses randomly selected two digits 1 and 7. The dataset consists of 1300 samples.

- RandomBlob dataset: It is a synthetic dataset generated by creating random clusters of data points with specified centers and standard deviations. The dataset consists of 200 samples with 2 classes.

*2. Data Poisoning Attack on MNIST Dataset*

A classifier is trained on the MNIST dataset to recognize handwritten digits is targeted by an adversary seeking to induce misclassifications. The attacker introduces perturbations in the images, manipulates handcrafted feature values to deceive the learning model into making incorrect classifications. The poisoned data points are generated

based on parameters like perturbation type, maximum perturbation and bounds for the attack space. The generated poisoned points are then used to retrain the SVM classifier. It impacts the accuracy of the classifier which in turn degrades the performance of classifier. Figure 1.2 shows the detailed steps used in the experiment.

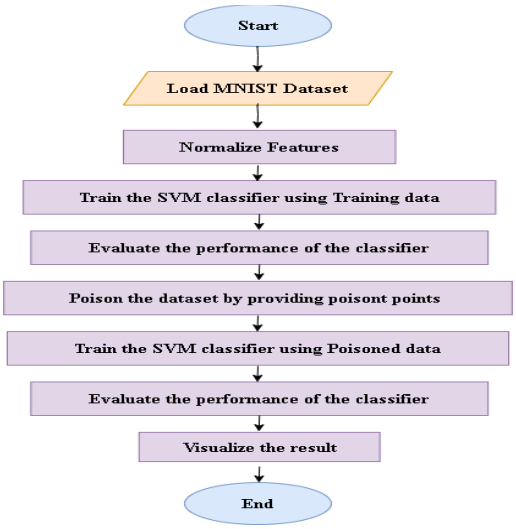


**Fig. 1.2: Flowchart for Data Poisoning Attack**

The data presented in Table 1.5 demonstrates a decrease in accuracy post-attack, signaling that the model's accuracy is adversely impacted with the inclusion of poisoned data during training. The manipulation of training data by attackers poses a significant threat in ML security. The extent of the attack's impact is affected by factors like the quantity of poisoning points and the size of datasets utilized for training, testing and validation. For this analysis, digits 1 and 7 are extracted from the MNIST dataset. Figures 1.3 to 1.4 depict the transformations in the images corresponding to the scenarios outlined in Table 1.5. Each figure illustrates a label reversal compared to the true label following the attack.

***Table 1.5 Effect of Data Poisoning Attack on SVM classifier***

| Sr. No | Training set Samples | Testing Set Samples | Validation Set Samples | Accuracy Before Attack | Post Attack Accuracy | Poison Points |
|---|---|---|---|---|---|---|
| 1 | 500 | 400 | 400 | 98.00% | 93.25% | 15 |
| 2 | 500 | 400 | 400 | 98.00% | 87.75% | 30 |
| 3 | 700 | 500 | 500 | 98.60% | 93.00% | 15 |
| 4 | 700 | 500 | 500 | 98.60% | 85.80% | 30 |

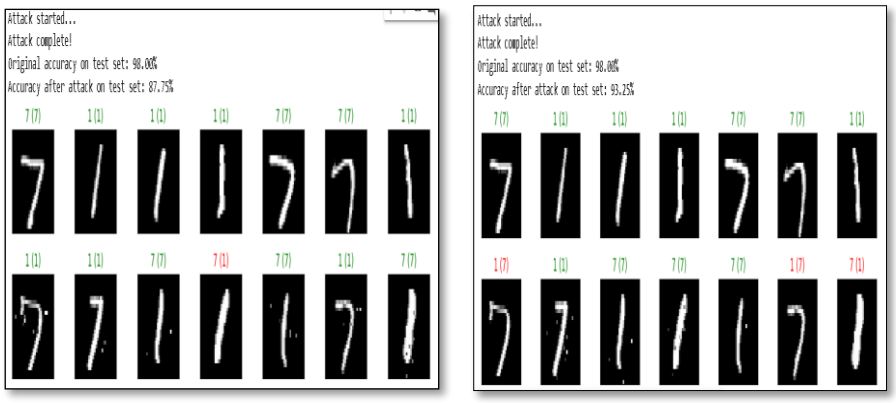


Figure 1.3 Model Performance as Per First and Second Case Mentioned in Table 1.5

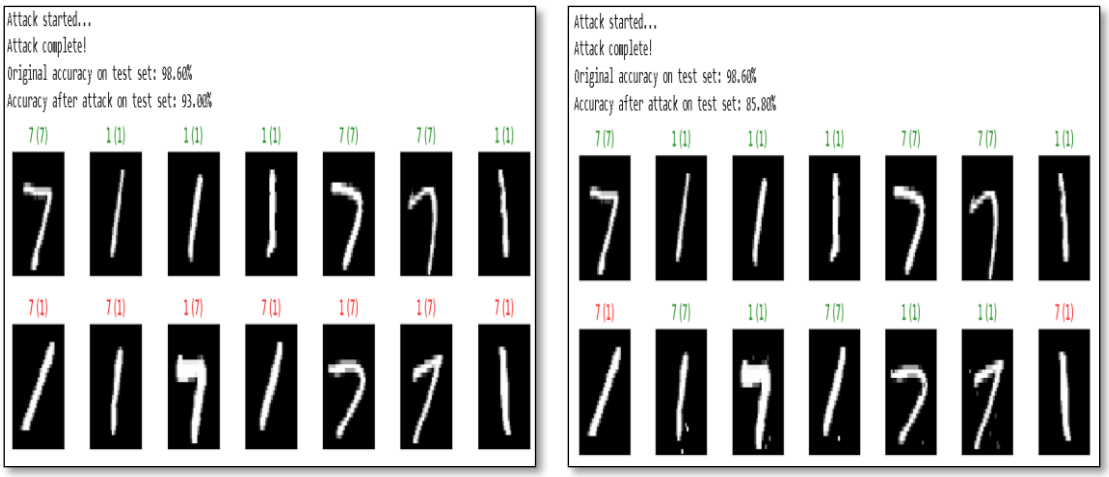


Figure 1.4 Model Performance According to the Third and Fourth Case From in Table 1.5

*3. The Impact of Label Manipulation Attack*

The impact of label manipulation attack on an SVM model trained on a synthetic dataset with two features and three classes has been analyzed. In this experiment, the class labels of some training samples have been altered. The accuracy of the SVM model was measured before and after these manipulations, as presented in Table 1.6. The findings from Table 1.6 indicate that the model's robustness against isolated noise helps to prevent overfitting in presence of label manipulation attack due to the majority of accurate labels. As the number of altered labels increases the models performance deteriorates because poisoned labels skew the decision boundary, leading to decreased accuracy on the test set. When a significant portion of the labels are incorrect, the model is unable to distinguish between correct and manipulated data. This results in the model effectively "learning" incorrect patterns, leading to a substantial degradation in performance.

***Table 1.6 Effect of Label Manipulation Attack on the SVM Model***

| No. of Labels Manipulated | Accuracy on Test Set Before Poisoning | Accuracy on Test Set After Poisoning |
|---|---|---|
| 1 | 94.00% | 94.00% |
| 25 | 94.00% | 89.00% |
| 40 | 94.00% | 74.00% |
| 50 | 94.00% | 48.00% |
| 75 | 94.00% | 48.00% |

The impact of label poisoning is initially minor but becomes severe as the proportion of poisoned labels increases. This highlights the need of securing data validation processes and training methods relating to noise and manipulation. In summary, the results highlights model accuracy is inversely proportional to the number of manipulated labels. The degree of label manipulations directly correlates with the degradation of the model's performance as shown below.

If we consider initial accuracy of the model is $AC_0$ and after modifying $n$ labels accuracy is $AC(n)$.

If $n = 1$ then $AC(1) \approx AC_0$

If $n > 1$ then $AC(n) = AC_0 - f(n)$

Here, $f(n)$ is a non-negative, monotonically increasing function of $n$ that signifies the deterioration in accuracy resulting from label manipulations. In general, $f(n)$ could be characterized as linear, nonlinear or logarithmic, based on the extent of the decline in performance. The derivative of $AC(n)$ can be expressed using Equation (1).

$\frac{dAC(n)}{dn} = -f'(n)$ Equation (1)

As $f'(n) \geq 0$ , it confirms decrease in accuracy when there is increase in number of modified labels.

Table 1.6 shows that the association between the quantity of manipulated labels $n$ and the model accuracy $AC(n)$ exhibits a non-linear decrease. This indicates that the effect of label manipulation on accuracy is not consistent but intensifies as a greater number of labels are modified.

## IV. CONCLUSIONS

This work has explored the survey of different types of hostile attacks on ML models. Adversarial attacks are intended to be as undetectable as possible, making them challenging to stop. Attackers who want to bypass current defenses are always coming up with new ways to do so. There are no established standards for assessing the potency of defensive strategies. Adversarial attacks require a thorough understanding of both security and ML. Each defensive technique mentioned in this work faces some challenges. Experimental results indicate that inclusion of poisoned data significantly impacts the accuracy of the classifier, leading to a decrease in performance post-attack. This highlights the vulnerability of ML models to adversarial attacks and emphasizes the importance of robust defenses against data poisoning attack. Additionally, the findings underscore the critical need for enhanced security measures to protect ML models from such threats and ensure their reliability and accuracy in real-world applications. Protecting training data from adversarial attacks to maintain the model's integrity, security and reliability is the main concern for ensuring the robustness and trustworthiness of ML models. Almost all defenses discard the poisoned data, which may lead to incorrect ML model. The development of new methodologies that can correct the detected malicious samples so that the model will train on the whole dataset rather than just part of it is much needed.

## ACKNOWLEDGMENT

The research has not received funding from any Agency.

## REFERENCES

[1] Abbas, Naveed Naeem, Tanveer Ahmed, Syed Habib Ullah Shah, Muhammad Omar, and Han Woo Park. "Investigating the applications of artificial intelligence in cyber security." Scientometrics 121 (2019): 1189-1211.

[2] Adam-Bourdarios, Claire, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. "The Higgs boson machine learning challenge." In NIPS 2014 workshop on high-energy physics and machine learning, PMLR, 664(7), (2015):19-55.

[3] Aladag, Merve, Ferhat Ozgur Catak, and Ensar Gul. "Preventing data poisoning attacks by using generative models." In 2019 1St International informatics and software engineering conference (UBMYK), IEEE, (2019): 1-5.

[4] Alves, Tiago AO, Felipe MG França, and Sandip Kundu. "MLPrivacyGuard: Defeating confidence information based model inversion attacks on machine learning systems." In Proceedings of the 2019 on Great Lakes Symposium on VLSI, (2019): 411-415.

[5] Baracaldo, Nathalie, 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, (2017): 103-110.

[6] Barreno, Marco, Blaine Nelson, Anthony D. Joseph, and J. Doug Tygar. "The security of machine learning." Machine learning 81 (2010): 121-148.

[7] Battista, Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, and Fabio Roli. "Poisoning behavioral malware clustering." In Proceedings of the 2014 workshop on artificial intelligent and security workshop, (2014): 27-36.

[8] Biggio, Battista, Blaine Nelson, and Pavel Laskov. "Poisoning attacks against support vector machines." arXiv preprint arXiv: 1206.6389 (2012).

[9] Biggio, Battista, Blaine Nelson, and Pavel Laskov. "Support vector machines under adversarial label noise." In Asian conference on machine learning, PMLR, (2011): 97-112.

[10] Biggio, Battista, Luca Didaci, Giorgio Fumera, and Fabio Roli. "Poisoning attacks to compromise face templates." In 2013 international conference on biometrics (ICB), IEEE, (2013): 1-7.

[11] Biggio, Battista, Samuel Rota Bulò, Ignazio Pillai, Michele Mura, Eyasu Zemene Mequanint, Marcello Pelillo, and Fabio Roli. "Poisoning complete-linkage hierarchical clustering." In Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings, Springer Berlin Heidelberg, (2014): 42-52.

[12] Biggio, Battista. "Machine learning under attack: Vulnerability exploitation and security measures." In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, (2016): 1-2.
[13] Burkard, Cody, and Brent Lagesse. "Analysis of causative attacks against svms learning from data streams." In Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics, (2017): 31-36.
[14] Carlini, Nicholas, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. "Hidden voice commands." In 25th USENIX security symposium (USENIX security 16), (2016):513-530.
[15] Chakarov, Aleksandar, Aditya Nori, Sriram Rajamani, Shayak Sen, and Deepak Vijaykeerthy. "Debugging machine learning tasks." arXiv preprint arXiv: 1603.07292 (2016).
[16] Chakraborty, Anirban, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. "A survey on adversarial attacks and defences." CAAI Transactions on Intelligence Technology 6(1), (2021): 25-45.
[17] Chen, Sen, Minhui Xue, Lingling Fan, Shuang Hao, Lihua Xu, Haojin Zhu, and Bo Li. "Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach." computers & security 73 (2018): 326-344.
[18] Deb, Debayan, Jianbang Zhang, and Anil K. Jain. "Advfaces: Adversarial face synthesis." In 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, (2020): 1-10.
[19] Diakonikolas, Ilias, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. "Sever: A robust meta-algorithm for stochastic optimization." In International Conference on Machine Learning, PMLR, (2019): 1596-1606.
[20] Fang, Minghong, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. "Poisoning attacks to graph-based recommender systems." In Proceedings of the 34th annual computer security applications conference, (2018): 381-392.
[21] Fang, Minghong, Neil Zhenqiang Gong, and Jia Liu. "Influence function based data poisoning attacks to top-n recommender systems." In Proceedings of The Web Conference 2020, (2020): 3019-3025.
[22] Feng, Ji, Qi-Zhi Cai, and Zhi-Hua Zhou. "Learning to confuse: Generating training time adversarial data with auto-encoder." Advances in Neural Information Processing Systems conference, (2019): 11971–11981.
[23] Fowl, Liam, Ping-yeh Chiang, Micah Goldblum, Jonas Geiping, Arpit Bansal, Wojtek Czaja, and Tom Goldstein. "Preventing unauthorized use of proprietary data: Poisoning for secure dataset release." arXiv preprint arXiv: 2103.02683 (2021).
[24] Hu, Hongsheng, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. "Membership inference attacks on machine learning: A survey." ACM Computing Surveys (CSUR) 54(11), (2022): 1-37.
[25] Huang, Hai, Jiaming Mu, Neil Zhenqiang Gong, Qi Li, Bin Liu, and Mingwei Xu. "Data poisoning attacks to deep learning based recommender systems." arXiv preprint arXiv: 2101.02644 (2021).
[26] Huang, Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. "Is feature selection secure against training data poisoning?." In international conference on machine learning, PMLR, (2015): 1689-1698.
[27] Ignazio Pillai, Biggio, Battista, Samuel Rota Bulò, Davide Ariu, Marcello Pelillo, and Fabio Roli. "Is data clustering in adversarial settings secure?" In Proceedings of the 2013 ACM workshop on Artificial intelligence and security, (2013): 87-98.
[28] Jagannathan, Jayaganesh, and MY Mohamed Parvees. "CIA Triad Validation in Intrusion Detection Using ACO Algorithm with RNN based Cognitive Mechanisms." Mathematical Statistician and Engineering Applications 71(4), (2022): 4198-4207.
[29] Jagielski, Matthew, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." In 2018 IEEE symposium on security and privacy (SP), IEEE, (2018): 19-35.
[30] Khedr, Ahmed M. "Enhancing supply chain management with deep learning and machine learning techniques: A review." Journal of Open Innovation: Technology, Market, and Complexity 10(4), (2024): 100379.
[31] Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." In Artificial intelligence safety and security, Chapman and Hall/CRC, (2018): 99-112.
[32] Laishram, Ricky, and Vir Virander Phoha. "Curie: A method for protecting SVM classifier from poisoning attack." arXiv preprint arXiv: 1606.01584 (2016).
[33] Lee, Changui, and Seojeong Lee. "Vulnerability of Clean-Label Poisoning Attack for Object Detection in Maritime Autonomous Surface Ships." Journal of Marine Science and Engineering 11(6), (2023): 1179.
[34] Levine, Alexander, and Soheil Feizi. "Deep partition aggregation: Provable defense against general poisoning attacks." arXiv preprint arXiv: 2006.14768 (2020).
[35] Li, Bo, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. "Data poisoning attacks on factorization-based collaborative filtering." Advances in neural information processing systems 29, (2016): 1893 - 1901.
[36] Liu, Qiang, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor CM Leung. "A survey on security threats and defensive techniques of machine learning: A data driven view." IEEE access 6 (2018): 12103-12117.

[37] Lucian Constantin, “How data poisoning attacks corrupt machine learning models”, last modified April 12, (2021), https://www.csoonline.com/article/570555/how-data-poisoning-attacks-corrupt-machine-learning-models.html.
[38] Ma, Yue, Yiwei He, and Yingjie Tian. "Online Robust Lagrangian Support Vector Machine against Adversarial Attack." Procedia computer science 139 (2018): 173-181.
[39] Ma, Yuzhe, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu. "Data poisoning attacks in contextual bandits." In Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29–31, 2018, Proceedings 9. Springer International Publishing, (2018): 186-204.
[40] Malik, Jasmita, Raja Muthalagu, and Pranav M. Pawar. "A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls and Technologies." IEEE Access (2024): 99382-99421.
[41] Mozaffari-Kermani, Mehran, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. "Systematic poisoning attacks on and defenses for machine learning in healthcare." IEEE journal of biomedical and health informatics 19(6), (2014): 1893-1905.
[42] Nelson, Blaine, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin IP Rubinstein, Udam Saini, Charles Sutton, J. Doug Tygar, and Kai Xia. "Exploiting machine learning to subvert your spam filter." LEET 8, (2008): 1-9.
[43] Newaz, AKM Iqtidar, Nur Imtiazul Haque, Amit Kumar Sikder, Mohammad Ashiqur Rahman, and A. Selcuk Uluagac. "Adversarial attacks to machine learning-based smart healthcare systems." In GLOBECOM 2020-2020 IEEE Global Communications Conference, IEEE, (2020): 1-6.
[44] Newell, Andrew, Rahul Potharaju, Luojie Xiang, and Cristina Nita-Rotaru. "On the practicality of integrity attacks on document-level sentiment analysis." In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, (2014):83-93.
[45] Newsome, James, Brad Karp, and Dawn Song. "Paragraph: Thwarting signature learning by training maliciously." In *Recent Advances in Intrusion Detection: 9th International Symposium, RAID 2006 Hamburg, Germany, September 20-22, 2006 Proceedings 9*, Springer Berlin Heidelberg, (2006):81-105.
[46] Oprea, Alina, and Apostol Vassilev. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. No. NIST Artificial Intelligence (AI) 100-2 E2023 (Withdrawn). National Institute of Standards and Technology, (2023).
[47] Panayides, Andreas S., Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran et al. "AI in medical imaging informatics: current challenges and future directions." IEEE journal of biomedical and health informatics 24(7), (2020): 1837-1857.
[48] Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. "Sok: Security and privacy in machine learning." In 2018 IEEE European symposium on security and privacy (EuroS&P), IEEE, (2018): 399-414.
[49] Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. "Practical black-box attacks against machine learning." In Proceedings of the 2017 ACM on Asia conference on computer and communications security, (2017):506-519.
[50] Papernot, Nicolas, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. "The limitations of deep learning in adversarial settings." In 2016 IEEE European symposium on security and privacy (EuroS&P), IEEE, (2016): 372-387.
[51] Paudice, Andrea, Luis Muñoz-González, and Emil C. Lupu. "Label sanitization against label flipping poisoning attacks." In ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18, Springer International Publishing, (2019): 5-15.
[52] Perdisci, Roberto, David Dagon, Wenke Lee, Prahlad Fogla, and Monirul Sharif. "Misleading worm signature generators using deliberate noise injection." In 2006 IEEE Symposium on Security and Privacy (S&P'06), IEEE, (2006): 17-31.
[53] Rosenfeld, Elan, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. "Certified robustness to label-flipping attacks via randomized smoothing." In International Conference on Machine Learning, PMLR, (2020): 8230-8241.
[54] Rubinstein, Benjamin IP, Blaine Nelson, Ling Huang, Anthony D. Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J. Doug Tygar. "Antidote: understanding and defending against poisoning of anomaly detectors." In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, (2009): 1-14.
[55] Sameen, Maria, and Seong Oun Hwang. "Distinïct: Data poisoning attacks dectection using optïmized jaccard distance." English, Computers, Materials, Continua 73 (2022).
[56] Schmidt, Eric. "US National Security Commission on Artificial Intelligence." In *Augmented Education in the Global Age*, pp. 234-244. Routledge, 2023.
[57] Shafahi, Ali, 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): 6106 - 6116.
[58] Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." In Proceedings of the 2016 acm sigsac conference on computer and communications security, 24(28), (2016):1528-1540.

[59] Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang. "Certified defenses for data poisoning attacks." Advances in neural information processing systems 30 (2017): 3520–3532.
[60] Szegedy, C. "Intriguing properties of neural networks." arXiv preprint arXiv: 1312.6199 (2013).
[61] Tabassi, Elham, Kevin J. Burns, Michael Hadjimichael, Andres D. Molina-Markham, and Julian T. Sexton. "A taxonomy and terminology of adversarial machine learning." NIST IR 2019 (2019): 1-29.
[62] Taheri, Rahim, Reza Javidan, Mohammad Shojafar, Zahra Pooranian, Ali Miri, and Mauro Conti. "On defending against label flipping attacks on malware detection systems." Neural Computing and Applications 32 (2020): 14781-14800.
[63] Tang, Li, Haibo Hu, Moncef Gabbouj, Qingqing Ye, Yang Xiang, Jin Li, and Lang Li. "A Survey on Securing Image-Centric Edge Intelligence." ACM Transactions on Multimedia Computing, Communications and Applications (2024).
[64] Venkatesan, Sridhar, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. De Lucia. "Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems." In MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM), IEEE, (2021): 874-879.
[65] Wang, Chen, Jian Chen, Yang Yang, Xiaoqiang Ma, and Jiangchuan Liu. "Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects." Digital Communications and Networks 8(2), (2022): 225-234.
[66] Wang, Gang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. "Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers." In *23rd USENIX Security Symposium (USENIX Security 14)*, (2014):239-254.
[67] Wang, Wenxiao, Alexander J. Levine, and Soheil Feizi. "Improved certified defenses against data poisoning with (deterministic) finite aggregation." In International Conference on Machine Learning, PMLR, (2022): 22769-22783.
[68] Wang, Yizhen, and Kamalika Chaudhuri. "Data poisoning attacks against online learning." arXiv preprint arXiv: 1808.08994 (2018).
[69] Xiang, Chun-zhi, Ning-xian Fu, and Thippa Reddy Gadekallu. "Design of resource matching model of intelligent education system based on machine learning." EAI Endorsed Transactions on Scalable Information Systems 9(6), (2022): e1-e1.
[70] Xiao, Huang, Battista Biggio, Blaine Nelson, Han Xiao, Claudia Eckert, and Fabio Roli. "Support vector machines under adversarial label contamination." Neurocomputing 160 (2015): 53-62.
[71] Yang, Chaofei, Qing Wu, Hai Li, and Yiran Chen. "Generative poisoning attack method against neural networks." arXiv preprint arXiv: 1703.01340 (2017).
[72] Zhao, Mengchen, Bo An, Wei Gao, and Teng Zhang. "Efficient label contamination attacks against black-box learning models." In IJCAI, (2017): 3945-3951.