Machine learning security attacks and defense approaches for emerging cyber physical applications: A comprehensive survey

2022.10.10

Presented by: Mikail Mohammed Salim

Advanced Security in Emerging ICT

Professor: 박종혁

Contents

1. Abstract

1. **Existing environment:** The Cyber Physical Systems (CPS) integrate the sensing, computation, control and networking processes into physical objects and infrastructure, which are connected through the Internet to execute a common task.
2. **Challenge**: The tight coupling of cyber systems with physical systems introduce challenges in addressing stability, security, efficiency and reliability.
3. **Weakness in existing system:** Execution of Machine Learning (ML) based approaches may not function effectively in case if a system is connected to the Internet and online hackers exploit deployed security mechanisms and poison the data used for training.
4. **Objective of the paper**:
    1. Discuss details of various ML security attacks in CPS,
    2. Present defense mechanisms to protect against attacks, and issues and challenges of ML security mechanisms.
    3. Present a comparative analysis of varying ML under the influence of attacks.

1. Introduction

- ML models are used in CPS-based applications to draw useful outcomes from the collected data of the sensors. Therefore, the role of ML models is very important here, and their predictions and outcomes should be accurate.
- ML is integrated and utilized in various domains, like the Internet of Things (IoT), CPS, cyber security, computer vision, image processing, robotics, and natural language processing.
- There are two phases in Machine Learning:
  1. **Training Phase**: Data is collected from authorized IoT devices. Data preparation is achieved by cleaning, augmenting, and segmenting process. Next, the data is labelled and split into two datasets, testing and training data. The training data is given to the ML model to train it with the feature values, and then the possible pattern is made. The testing phase starts, and the calculated parameters are used on the test data in order to carry out the new predictions.
  2. **Deployment Phase:** The deployed model after hyper tuning is supplied with the real-time data. The trained model will provide prediction output on input new data. The model may use the Application Programming Interface (API) to interact with the users where we can feed the data through it and obtain the predictions based on training done under the training phase.

1. Introduction

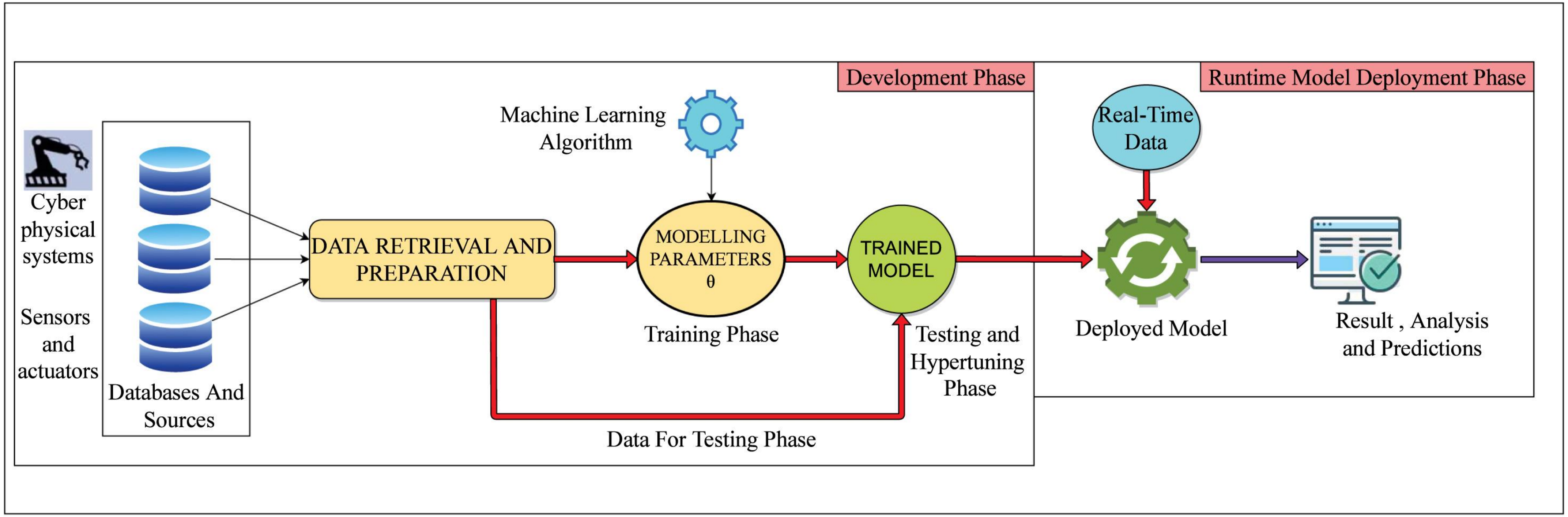
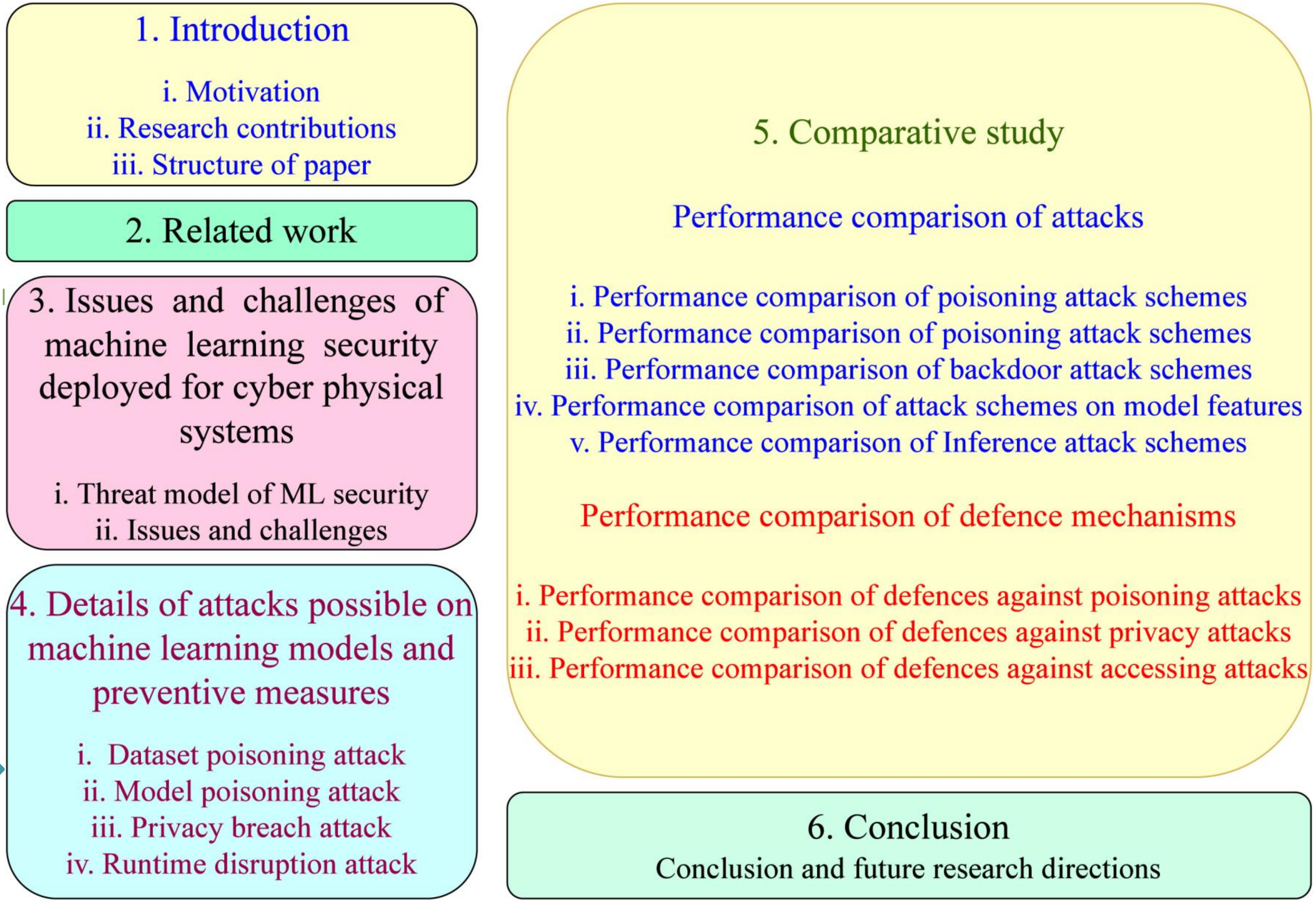


Fig. 1. Various phases of machine learning tasks in the cyber physical systems.

1. Introduction

- Information security is the methodology of protecting information and sensitive data from security risks (i.e., unauthorized access and usage, modification, inspection, and deletion of the information.
- Information security in the cyber physical systems is provided on the basis of CIA Triad, which comprises techniques like Confidentiality, Integrity, and Availability.
  1. **Confidentiality**: Confidentiality (or privacy) involves restricting access to the information. Its usage is much needed in order to protect information from being accessed or modified by malicious entities. Techniques include utilizing encryption techniques, including public-key cryptography and security tokens.
  2. **Integrity:** Integrity (or data integrity) involves maintaining the trustworthiness and dependability of the information. It is practiced to retain the usability of data and prevail it to be usable for other tasks. Techniques include version and access controlling, hashing and compliance checks.
  3. **Availability:** Availability is the practice of accessibility of information for retrieval and usage by authorized entities. It is required to maintain the information consistently through the maintaining systems which hold them. Techniques include server monitoring, resolving software issues and protocols against DDoS attacks.



1. Introduction

i. Motivation
ii. Research contributions
iii. Structure of paper

2. Related work

3. Issues and challenges of machine learning security deployed for cyber physical systems

i. Threat model of ML security
ii. Issues and challenges

4. Details of attacks possible on machine learning models and preventive measures

i. Dataset poisoning attack
ii. Model poisoning attack
iii. Privacy breach attack
iv. Runtime disruption attack

5. Comparative study

Performance comparison of attacks

i. Performance comparison of poisoning attack schemes
ii. Performance comparison of poisoning attack schemes
iii. Performance comparison of backdoor attack schemes
iv. Performance comparison of attack schemes on model features
v. Performance comparison of Inference attack schemes

Performance comparison of defence mechanisms

i. Performance comparison of defences against poisoning attacks
ii. Performance comparison of defences against privacy attacks
iii. Performance comparison of defences against accessing attacks

6. Conclusion

Conclusion and future research directions

Fig. 2. Roadmap of the paper.

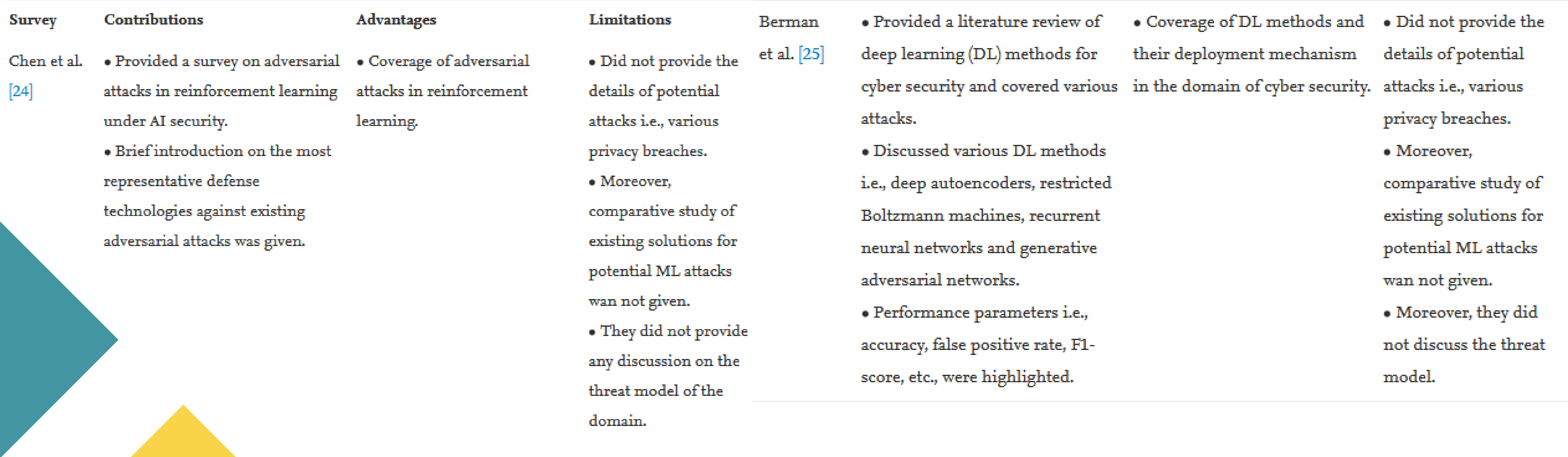
1. Introduction - Motivations

- The primary motivation behind this survey paper is to summarize the research work and case studies done in the field of ML security in the cyber physical systems.
- ML is used in various domains (i.e., healthcare, security and surveillance, retailing, industrial automation, control and support, and intelligent transportation system. Thus, the correct prediction and privacy of users' data are essentially required.
- During the literature survey, it has been identified that the ML models are vulnerable to various types of attacks (i.e., dataset poisoning attack, model poisoning attack, privacy breach, runtime disruption attack, and membership inference attacks).
- Due to the enormous use of ML, it becomes essential to protect its models against the various possible attacks. Therefore, the study focuses on various attacks that are associated with the ML models.

1. Introduction – Research Contributions

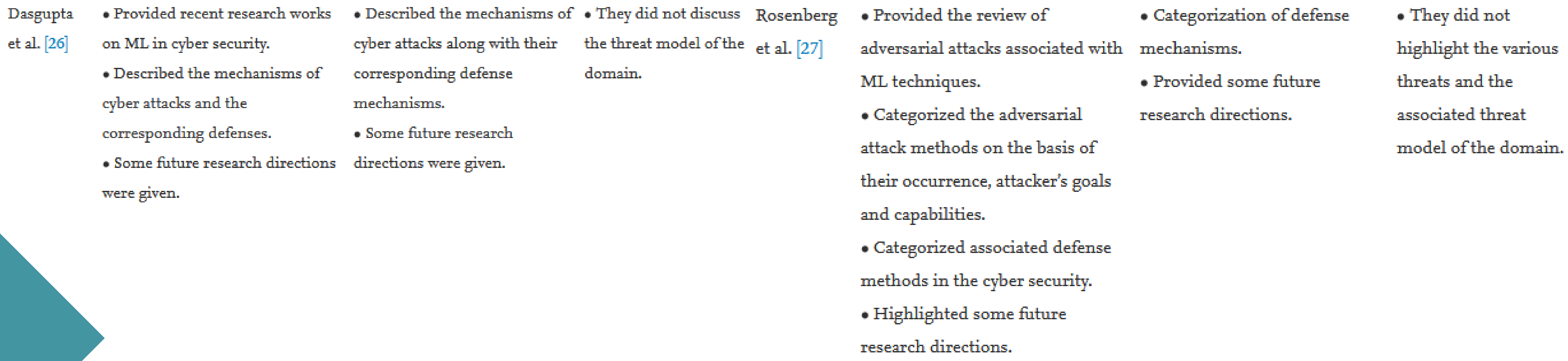
- The authors presented a threat model of ML security in the cyber physical systems, in which we provide the details of all threats associated with the ML models.
- The paper discusses the various issues and challenges of ML security in the cyber physical systems.
- Next, the study describes the mechanisms of various attacks related to the ML security and possible solutions that can be used to protect the ML security.
- Lastly, the study presents a comparative study on performance of the ML models under the influence of various attacks that can be also deployed for the cyber physical systems.

2. Related Work



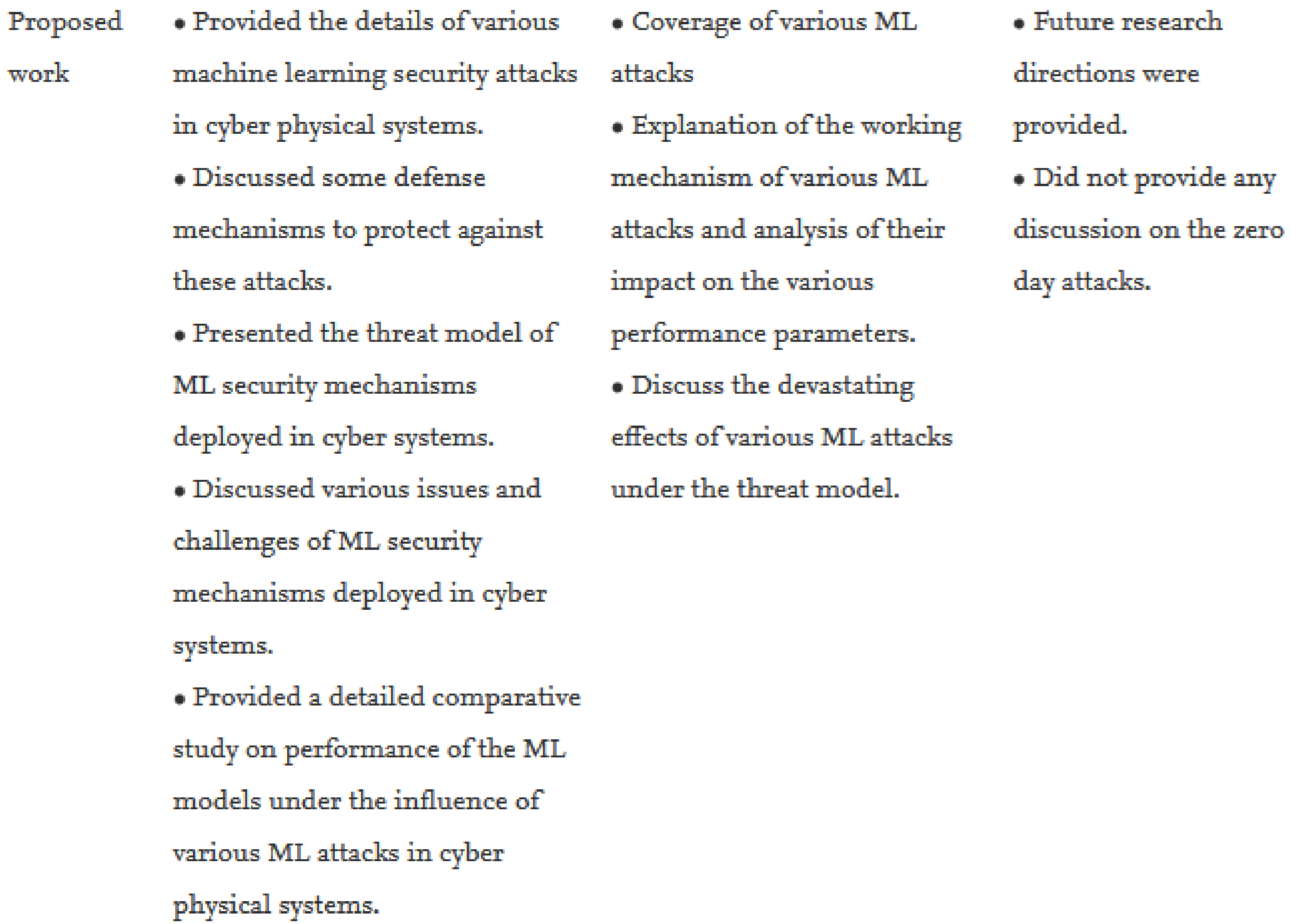
Survey	Contributions	Advantages	Limitations
Chen et al. [24]	• Provided a survey on adversarial attacks in reinforcement learning under AI security. • Brief introduction on the most representative defense technologies against existing adversarial attacks was given.	• Coverage of adversarial attacks in reinforcement learning.	• Did not provide the details of potential attacks i.e., various privacy breaches. • Moreover, comparative study of existing solutions for potential ML attacks wan not given. • They did not provide any discussion on the threat model of the domain.
Berman et al. [25]	• Provided a literature review of deep learning (DL) methods for cyber security and covered various attacks. • Discussed various DL methods i.e., deep autoencoders, restricted Boltzmann machines, recurrent neural networks and generative adversarial networks. • Performance parameters i.e., accuracy, false positive rate, F1-score, etc., were highlighted.	• Coverage of DL methods and their deployment mechanism in the domain of cyber security.	• Did not provide the details of potential attacks i.e., various privacy breaches. • Moreover, comparative study of existing solutions for potential ML attacks wan not given. • Moreover, they did not discuss the threat model.

2. Related Work



Dasgupta et al. [26]	• Provided recent research works on ML in cyber security. • Described the mechanisms of cyber attacks and the corresponding defenses. • Some future research directions were given.	• Described the mechanisms of cyber attacks along with their corresponding defense mechanisms. • Some future research directions were given.	• They did not discuss the threat model of the domain.
Rosenberg et al. [27]	• Provided the review of adversarial attacks associated with ML techniques. • Categorized the adversarial attack methods on the basis of their occurrence, attacker's goals and capabilities. • Categorized associated defense methods in the cyber security. • Highlighted some future research directions.	• Categorization of defense mechanisms. • Provided some future research directions.	• They did not highlight the various threats and the associated threat model of the domain.

2. Related Work



Proposed work	• Provided the details of various machine learning security attacks in cyber physical systems. • Discussed some defense mechanisms to protect against these attacks. • Presented the threat model of ML security mechanisms deployed in cyber systems. • Discussed various issues and challenges of ML security mechanisms deployed in cyber systems. • Provided a detailed comparative study on performance of the ML models under the influence of various ML attacks in cyber physical systems.	• Coverage of various ML attacks • Explanation of the working mechanism of various ML attacks and analysis of their impact on the various performance parameters. • Discuss the devastating effects of various ML attacks under the threat model.	• Future research directions were provided. • Did not provide any discussion on the zero day attacks.

3. Issues and challenges of machine learning security deployed for cyber physical systems –

3.1 Threat model of ML security

- Usually the data, which is being utilized for the learning and testing purpose in ML receives through the open (insecure) channel.
- Therefore, the existing adversary can interrupt the normal flow of ML task in many ways, including replay, man in the middle (MiTM) attack, impersonation, malware injection, flooding attacks, denial of service (DoS) attacks, distributed denial of service (DDoS) attacks, false data insertion, and unauthorized data updates.
- If a ML model learns through the data, from which some information was deleted or altered then there are the high chance that this ML model will produce the wrong outcomes (results).
- Hence there are the high chances that normal procedure of ML flow may get disturbed. Thus, we need some security solutions to provide protection against these threats and attacks.

3. Issues and challenges of machine learning security deployed for cyber physical systems –

3.2 Issues and challenges of machine learning security deployed for cyber physical systems

- **Security of deployed mechanism**: The security of ML models can be ensured through various mechanism i.e., proper use of authentication schemes (like device-to-device authentication, user to device authentication), proper use of access control schemes (like, certificate-based access control, certificate less access control) and intrusion detection scheme.
- **Accuracy of deployed ML model**: It is always desirable to get the high value of accuracy for some ML tasks. herefore, to get the high value of accuracy, we must be very careful and selective. We should select the ML algorithm wisely and as per the scenario and available datasets.
- **Failure of deployed security mechanisms**: Zero day vulnerabilities enable hackers to break security protocols and halt the working process of machine learning models. To provide more security we should apply the combination of intrusion detection schemes i.e., hybrid anomaly detection (for example, combination of signature based detection and anomaly based detection).
- **Interoperability**: ML security environment is the collection of various ML algorithm, which have their own limitations and constraints. Security algorithms should be selected wisely, i.e. which algorithm is appropriate with which ML algorithm.

3. Issues and challenges of machine learning security deployed for cyber physical systems –

3.2 Issues and challenges of machine learning security deployed for cyber physical systems

- **Interoperability**: ML security environment is the collection of various ML algorithm, which have their own limitations and constraints. Security algorithms should be selected wisely, i.e. which algorithm is appropriate with which ML algorithm.
- **Obsolescence:** ML algorithms have their own limitations and some of them become obsolete when the time. That raises issues related to the obsolescence. Hence steps of ML tasks should be updated accordingly and the tools and technologies, which become obsolete should not be used in the ML tasks.

4. Details of attacks possible on machine learning models and preventive measures

- Attacks on ML can be broadly classified into four categories, i.e., dataset poisoning attacks (both real time and training data), model poisoning attacks, privacy breach and model inversion attacks and run-time disruption attacks.
- Fig. 3 visually depicts the attack points in the machine learning workflow. These attack points can be subjugated or interfered with to create disruption and cause privacy breach.
  - **Dataset poisoning attack**: In this attacker inserts adversarial examples in dataset to cause the attacking model to produce incorrect predictions.
  - **Model poisoning attack**: These types of attack focus on corrupting models by interfering with their internal workings and modifying the parameters.
  - **Privacy breach attack**: These attacks work on exposing sensitive data of users and retrieving valuable information of the model.
  - **Runtime disruption attack**: In this, intruder compromises the ML workflow to prevent efficient and accurate prediction by attacking the model in its execution.

4. Details of attacks possible on machine learning models and preventive measures

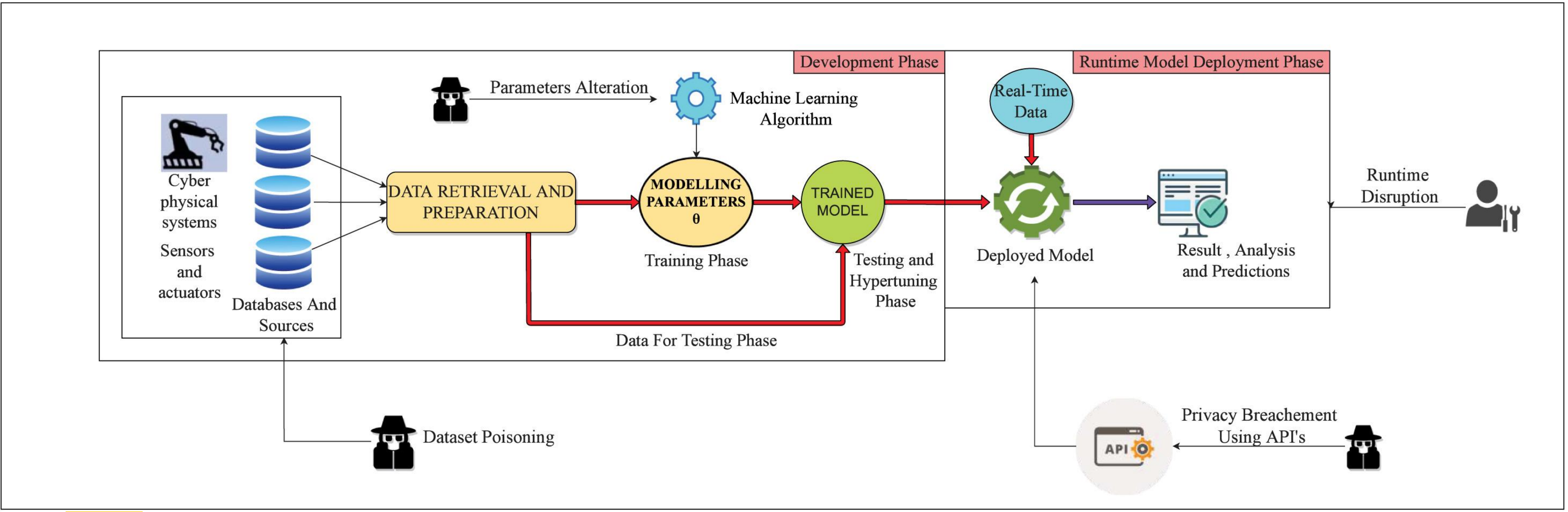


Fig. 3. Attack points in the machine learning workflow.

4. Details of attacks possible on machine learning models and preventive measures –

4.1 Dataset poisoning attack

- In this attack, the attacker utilizes the various techniques to infiltrate the training and testing data to disturb the normal functioning of a machine learning task. The attacker can utilize adversarial examples and can attack the data containing server or data lake from where raw data and photos have to be taken.
- The compromising of the data sources can lead to deployment of data, which can possibly alter the functioning of the ML model. It again changes the output of the classifier, which is very devastating in nature, i.e., for example, system is showing no illness, however, the patient suffers from the severe illness.

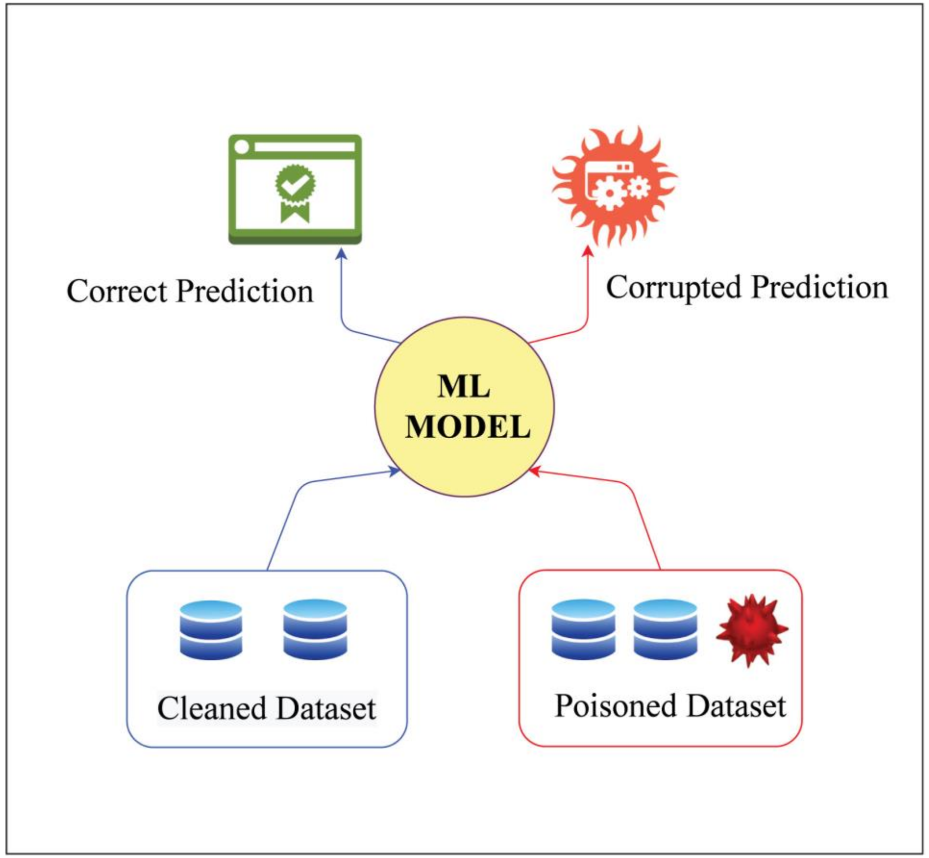
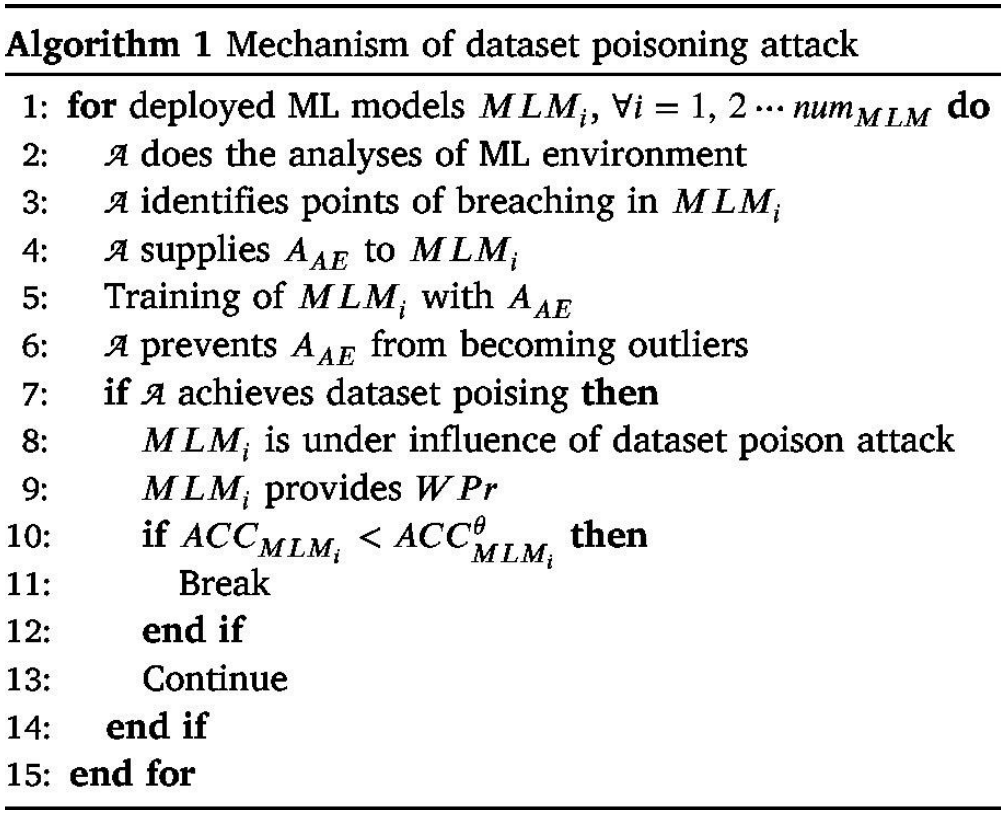


Fig. 4. Scenario of dataset poisoning attack.

4. Details of attacks possible on machine learning models and preventive measures –

4.1.1 Mechanism of data poisoning attack

- The following steps can be used to explain the mechanism of a dataset poisoning attack from the attacker's perspective:
  1. First, the attacker analyses the machine learning environment and goes to the data lake and warehouse where the training data is stored or sourced from.
  2. Then points of breaching are identified, and normal input data flow is altered and is supplied with adversarial examples.
  3. The adversarial examples that are supplied are made so as to train the model with incorrectly labeled and deceiving data supplemented with SQL injection.
  4. Special focus is given by the attacker to prevent the poisoning examples from becoming outliers for the data and subsequently removing them as part of data cleaning.
  5. With a sufficient ratio of the adversarial example, dataset poisoning is achieved, the accuracy of the model decreases, and the model becomes poisoned as it provides the wrong prediction to the labeled input data.



Algorithm 1 Mechanism of dataset poisoning attack

```
for deployed ML models $MLM_i$, $\forall i = 1, 2 \cdots num_{MLM}$ do
    $\mathcal{A}$ does the analyses of ML environment
    $\mathcal{A}$ identifies points of breaching in $MLM_i$
    $\mathcal{A}$ supplies $A_{AE}$ to $MLM_i$
    Training of $MLM_i$ with $A_{AE}$
    $\mathcal{A}$ prevents $A_{AE}$ from becoming outliers
    if $\mathcal{A}$ achieves dataset poising then
        $MLM_i$ is under influence of dataset poison attack
        $MLM_i$ provides $WPr$
        if $ACC_{MLM_i} < ACC^{\theta}_{MLM_i}$ then
            Break
        end if
        Continue
    end if
end for
```

Algorithm 1: The mechanism of dataset poisoning attack.

4. Details of attacks possible on machine learning models and preventive measures –

4.1.1 Mechanism of data poisoning attack

- **Black box** attacks have been attempted by the researchers on ML models in recent times through poisoning a model with adversarial examples without knowing the features of the model.
- Research studies have used poisoning strategies to create backdoor instances to misled the system to classify them as target label. The backdoor attack comes under this category.
- Compared to a **white box attack**, where the attacker has access to the ML model's parameters, the attacker does not have access to model parameters in the Black box attack.
- **Backdoor attacks** are also generally implemented on ML tasks (preferably on CNN models). These attacks are mostly on training set data and modifies the prediction through getting access to the model and its dataset. They do this by inserting a trigger in supplied adversarial data which when "activates" causing the model to misclassify the input with the backdoor induced model.

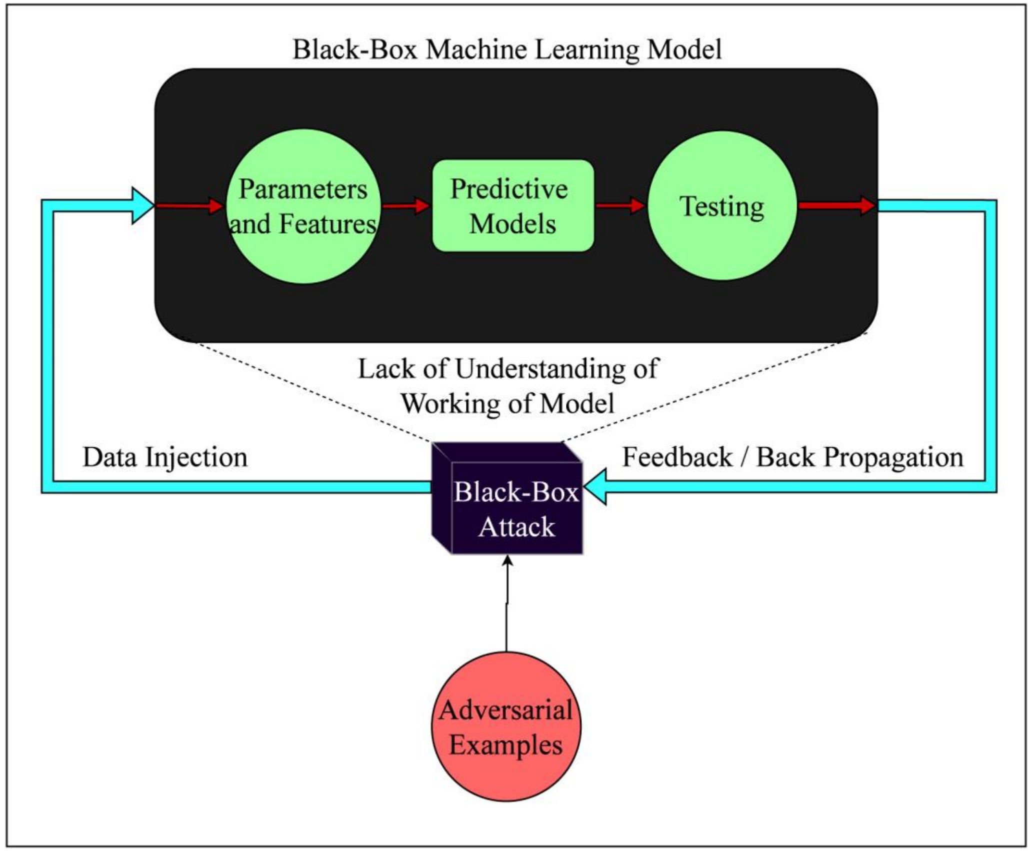

Black-Box Machine Learning Model
Parameters and Features
Predictive Models
Testing
Lack of Understanding of Working of Model
Data Injection
Feedback / Back Propagation
Black-Box Attack
Adversarial Examples

Fig. 5. Scenario of black box attack.

4. Details of attacks possible on machine learning models and preventive measures –

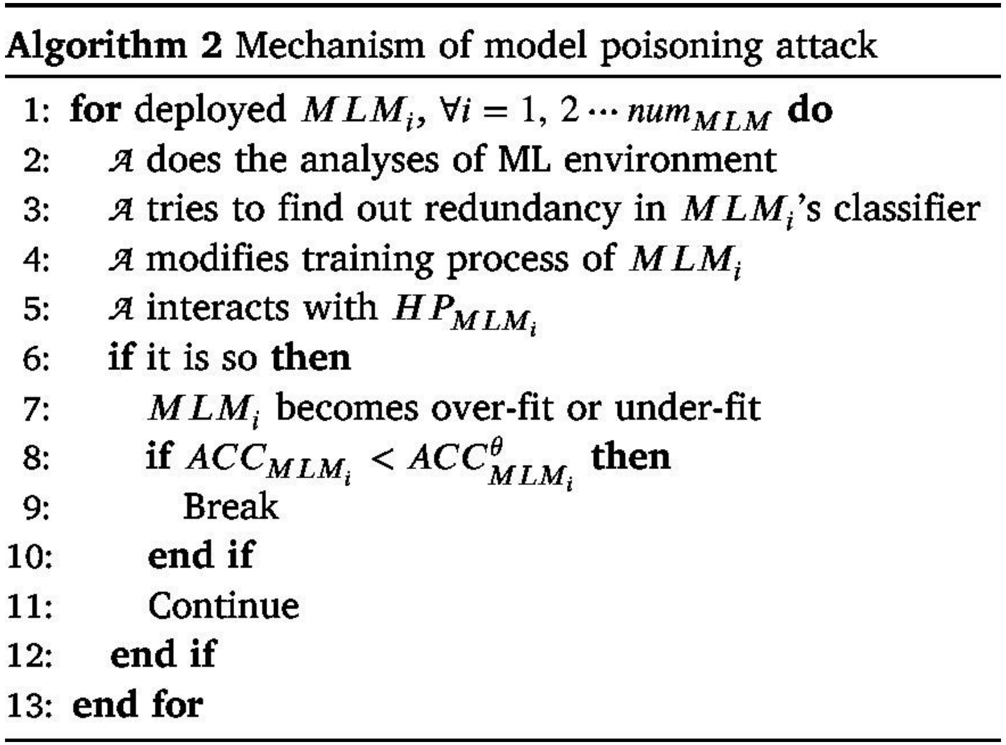
4.1.2 Mechanism to counter data poisoning attack

- The mechanism to counter dataset poisoning attack is as follows:
  1. The developer utilizes various cryptographic algorithms for the detection purpose. The utilization of some outliers in dataset can be done to detect any injection of malicious data.
  2. We train an outlier detector in parallel to the ML model, which helps in filtering out any data it deems poisoned, as it does not comply with how normally it should have been predicted at the deployment.
  3. Some specific data and there proposed prediction can be "tokenised" and maintained separately during training and can be compared during the deployment to get awareness of any occurring attack.
  4. Also a red flag can be issued if the difference in accuracy within the training and deployment phase is not within a acceptable limit due to higher chance of data poisoning attack at the deployment.
  5. We can also implement artificial neural network based-generative model parallel to the ML model with pre-computed accuracy score to cross verify the poisoning of the data.

4. Details of attacks possible on machine learning models and preventive measures –

4.2 Model poisoning attack (4.2.1 Procedure of model poisoning attack)

- Parameter alteration is a method utilized by hackers to generate faulty output by interfering with the classifier and altering the parameters through which the classifier prepares ML model.
- Model poisoning attacks is executed with the help of following steps:
  1. First of all the attacker interacts with the machine learning environment and find redundancy in the classifier.
  2. The attacker can change or modify the training algorithm to generate wrong output and publish deceiving results.
  3. Further the attacker can interact with the hyper-parameters and cause the model to over-fit or under-fit and create problems in the testing phase.
  4. As the parameters of the algorithm is hard-coded and can be dynamically modified unprotected uses of the algorithm can speed up the Attacks.



Algorithm 2 Mechanism of model poisoning attack

```
for deployed $MLM_i$, $\forall i = 1, 2 \cdots num_{MLM}$ do
    $\mathcal{A}$ does the analyses of ML environment
    $\mathcal{A}$ tries to find out redundancy in $MLM_i$'s classifier
    $\mathcal{A}$ modifies training process of $MLM_i$
    $\mathcal{A}$ interacts with $HP_{MLM_i}$
    if it is so then
        $MLM_i$ becomes over-fit or under-fit
        if $ACC_{MLM_i} < ACC^{\theta}_{MLM_i}$ then
            Break
        end if
        Continue
    end if
end for
```

Algorithm 2: The mechanism of model poisoning attack.

4. Details of attacks possible on machine learning models and preventive measures –

4.2 Model poisoning attack (4.2.2 Mechanism to counter model poisoning attack)

- Model poisoning attacks can be countered with the help of the following steps:
    1. When the user creates a ML classifier, he can vectorize the constraints and predictive labels to embed its own ID using hashing.
    2. During the hyper-parameters tuning phase the hash ID of the classifier can be updated by the Administrator.
    3. Finally during the deployment phase the hash ID can be matched with the deployed model's hash □□ to check for any discrepancies in both. Further this identify if the model has been altered.
    4. Blockchain can be implemented on the classifier to retain its hash value from possible attacks in a secured network.

4. Details of attacks possible on machine learning models and preventive measures –

4.3 Privacy Breach

- **Privacy breach:** Sensitive user’s data and model’s internal working can be compromised through a variety of methods. Unprotected files and lack of encrypted pipelines during training and deployment phases of ML task can leak the data and enables an unauthorized user to interfere with the model.
- Sensitive data of users can also be compromised from a model which has been trained by utilizing exposed API’s. It further compromises the model’s working by infecting the dataset with malicious input and by getting output from API call to reverse engineer the process. It exposes the inner working of the ML model.
- In this attack scenario, most of the prediction tasks are behind the protected framework.
- The intruder utilizes underlying API and publishes the leaked data sources. The attacker tends to create a dummy ML model, which is identical to the targeted model.
- Utilizing API calls the intruder sends edge-case data and receives prediction output in their own dummy model. With this data the attacker tries to form an intuition for his model by referring data sourced, which have been linked to model or tries to get access to the data.
- This information is then compared with prediction result. The published results are used by the developer to create a working ML model, which behaves like the target model. Further with the dummy model the attacker can perform various attacks like membership inference attacks.

4. Details of attacks possible on machine learning models and preventive measures –

4.3 Privacy Breach

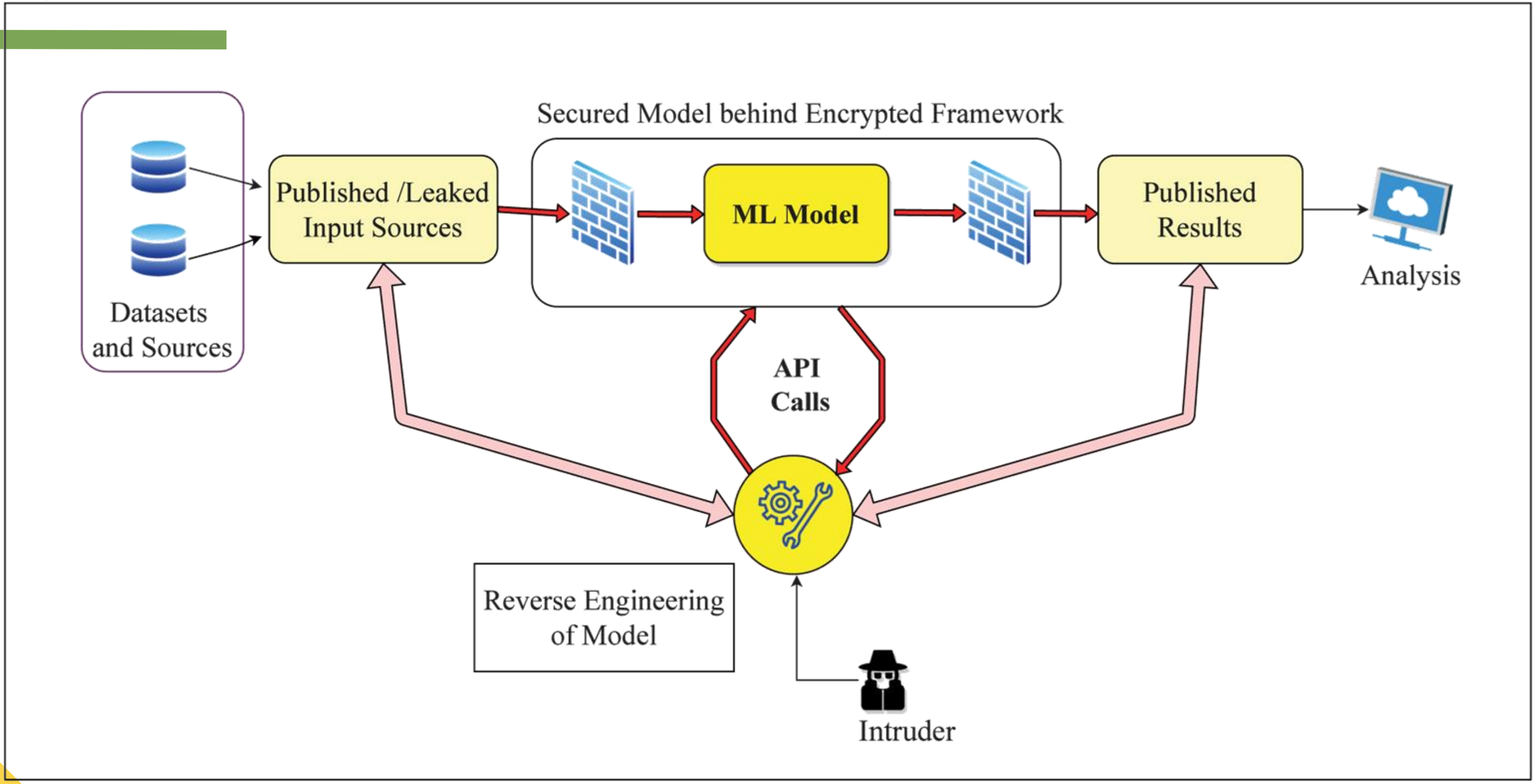
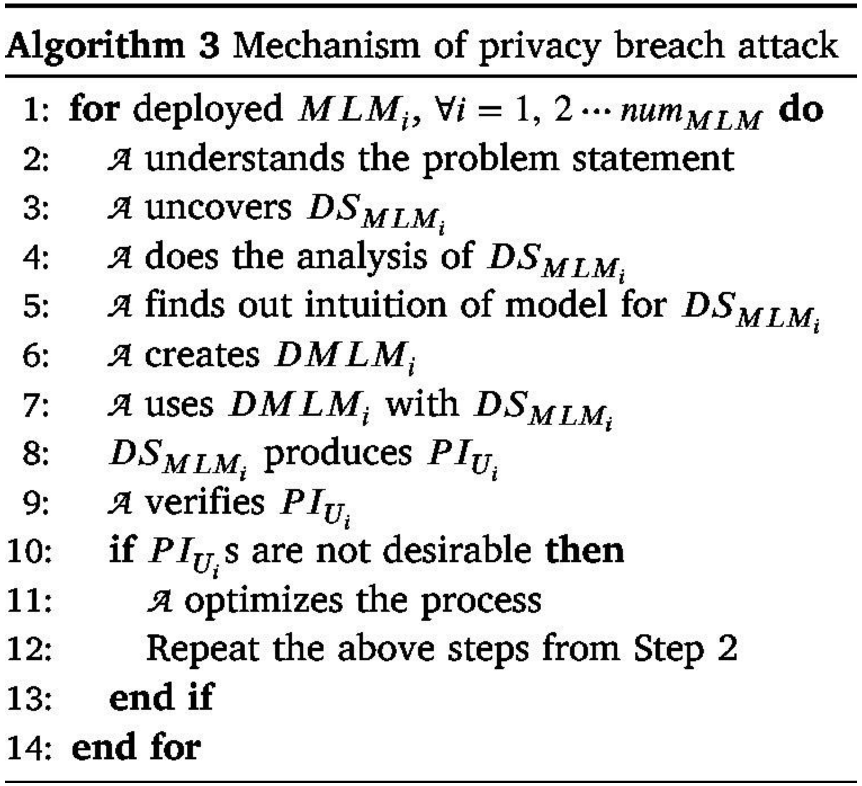

Secured Model behind Encrypted Framework
Datasets and Sources
Published /Leaked Input Sources
ML Model
Published Results
Analysis
API Calls
Reverse Engineering of Model
Intruder

Fig. 6. Scenario of privacy breach on a ML model through API calls.

4. Details of attacks possible on machine learning models and preventive measures –

4.3.1 Procedure of exposing sensitive data through API calls and reverse engineering

- The following steps highlight the procedure of exposing sensitive data through API calls and reverse engineering:
  1. The initial step is to understand the problem statement of ML task.
  2. Then required input will be scrapped from provided input sources or dataset leaked due to poor or null authorization of the server and published output can also be studied and analyzed. Then its predictions can be uncovered.
  3. Try to find intuition of model using the data recovered in Step 2.
  4. Using the intuition try to reverse engineer the process and create a dummy model of ML model to use with predicted approximate parameters.
  5. With built dummy model use it on published input resources to retrieve private information and data from the users.
  6. Optimization steps involve utilizing API calls on the running original model and retrieving specific outputs, which are based on specific input arguments for optimizing the edge cases in the new dummy model.



Algorithm 3 Mechanism of privacy breach attack

```
1:  for deployed $MLM_i$, $\forall i = 1, 2 \cdots num_{MLM}$ do
2:      $\mathcal{A}$ understands the problem statement
3:      $\mathcal{A}$ uncovers $DS_{MLM_i}$
4:      $\mathcal{A}$ does the analysis of $DS_{MLM_i}$
5:      $\mathcal{A}$ finds out intuition of model for $DS_{MLM_i}$
6:      $\mathcal{A}$ creates $DMLM_i$
7:      $\mathcal{A}$ uses $DMLM_i$ with $DS_{MLM_i}$
8:      $DS_{MLM_i}$ produces $PI_{U_i}$
9:      $\mathcal{A}$ verifies $PI_{U_i}$
10:     if $PI_{U_i}$s are not desirable then
11:         $\mathcal{A}$ optimizes the process
12:         Repeat the above steps from Step 2
13:     end if
14: end for
```

Algorithm 3: The mechanism of privacy breach attack.

4. Details of attacks possible on machine learning models and preventive measures –

4.3.2 Membership Inference attacks

- **Membership inference attacks:** Membership attack on ML model utilizes techniques and attacks discussed till now (underlying API's) to inject some personal data in the model work space and determines the utilization of that specific data during training of model.

- The procedure of a membership inference attack is described in Fig. 7.
    1. First, the intruder sources the data whose membership status has to be found.
    2. Then, normally a dummy model is created through reverse engineering the target model through methods discussed till now.
    3. The targeted data record is given to the dummy model for prediction.
    4. An intuition is developed by the intruder through comparing target model's published prediction and retrieved prediction of the instance from the dummy model.
    5. Though this intuition the intruder finds out if the specific data instance was in the dataset used for training of the model.
    6. With these findings membership status of the data instance is breached.

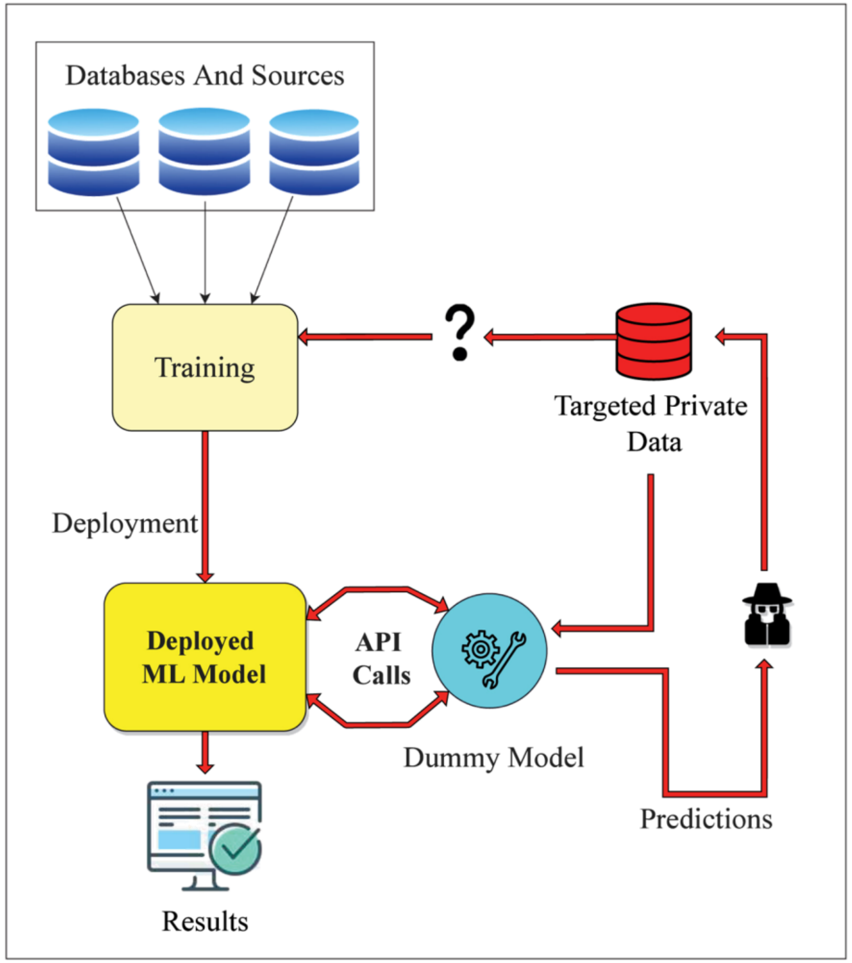


Fig. 7. Scenario of privacy breach on a ML model through API calls.

4. Details of attacks possible on machine learning models and preventive measures –

4.4 Runtime disruption attacks

- **Runtime disruption attacks:** Runtime disruption attack is used by the attacker to end or delay the ML task. Attackers generally target the server during the deployment phase and tries to remotely disrupt the ML process.
- A mechanism to prevent runtime disruption attack is given below.
  1. First, the intruder sources the data whose membership status has to be This procedure implements the techniques of parallel computing to safeguard a ML task from the attacks, which occur during deployment of trained model on real-data that are made to cause runtime disruption.
  2. It consist of one master node which contains the trained model and the instruction. It then divides and gives them to its sub units to perform specific task. These units further categorize into racks and multiple slave nodes, which basically perform the real task in the ML workflow.
  3. The slave nodes would be responsible for the collection of data. They process it and provides output on the basis of instructions received by the master node.
  4. The multiple machine working on same unit of data, therefore, the ML task will be self sufficient to continue the task without any denial or error.
  5. To avoid redundancy and integrity of training and deploying of model the master node accumulates multiple results and compile it to complete task even in the case of attack.

4. Details of attacks possible on machine learning models and preventive measures –

4.4 Runtime disruption attacks

- Fig. 8 depicts the decentralization of ML workload to secure the process from any case of disruption attack on the deployed model. It visualizes how the suggested model could have sub-models with there own private rack and nodes to generate rack-awareness algorithm to keep on continuing the ML task without any disruption.
- As referenced in Fig. 8 each root node computes prediction task of multiple racks, implying one node does the computing of more than one sub-model.
- This ensures that when there is the disruption attack at rack (root node) level, then other rack (root node) can do the computing for them. The prediction task still executes with the increment in the computational time.

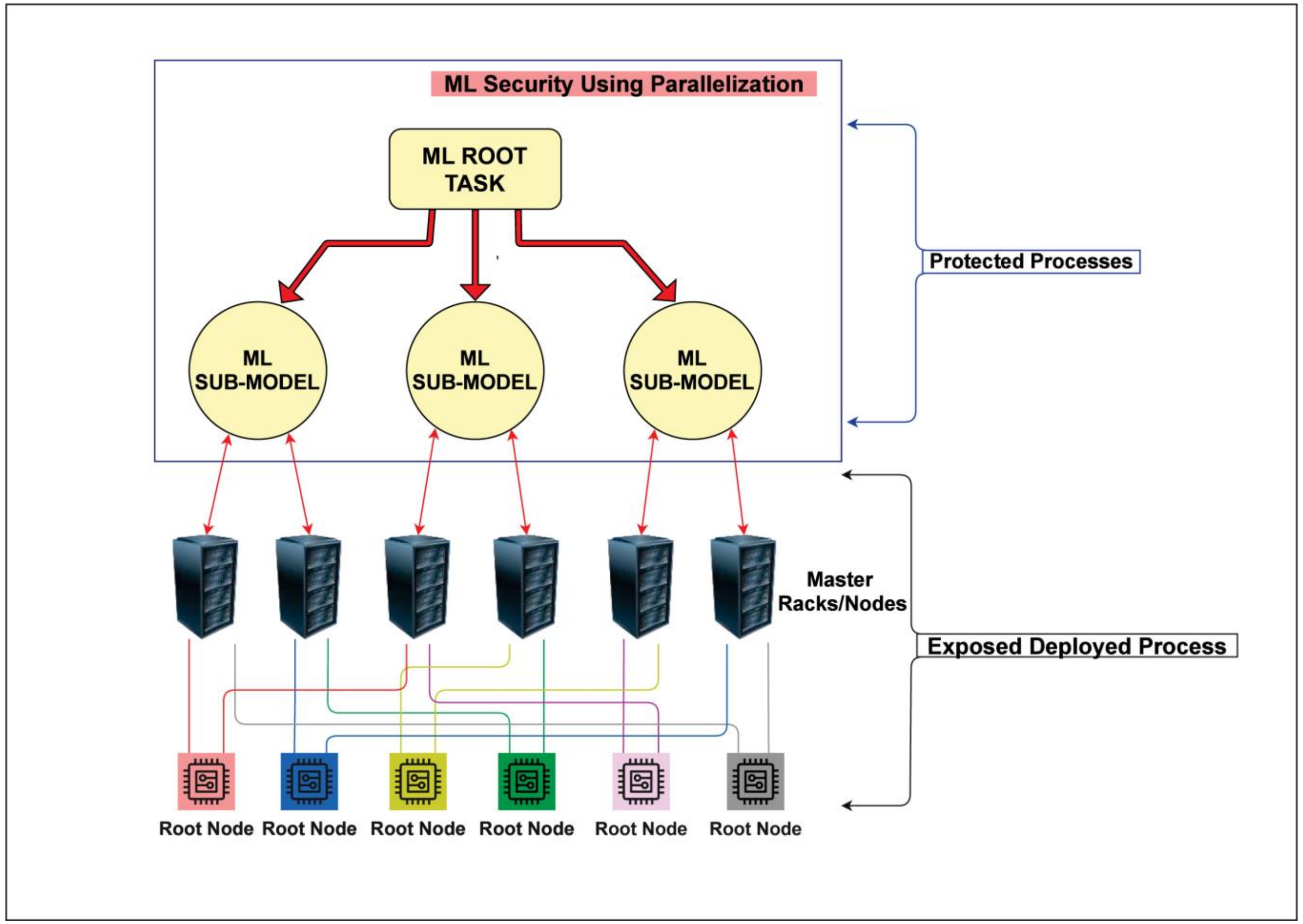
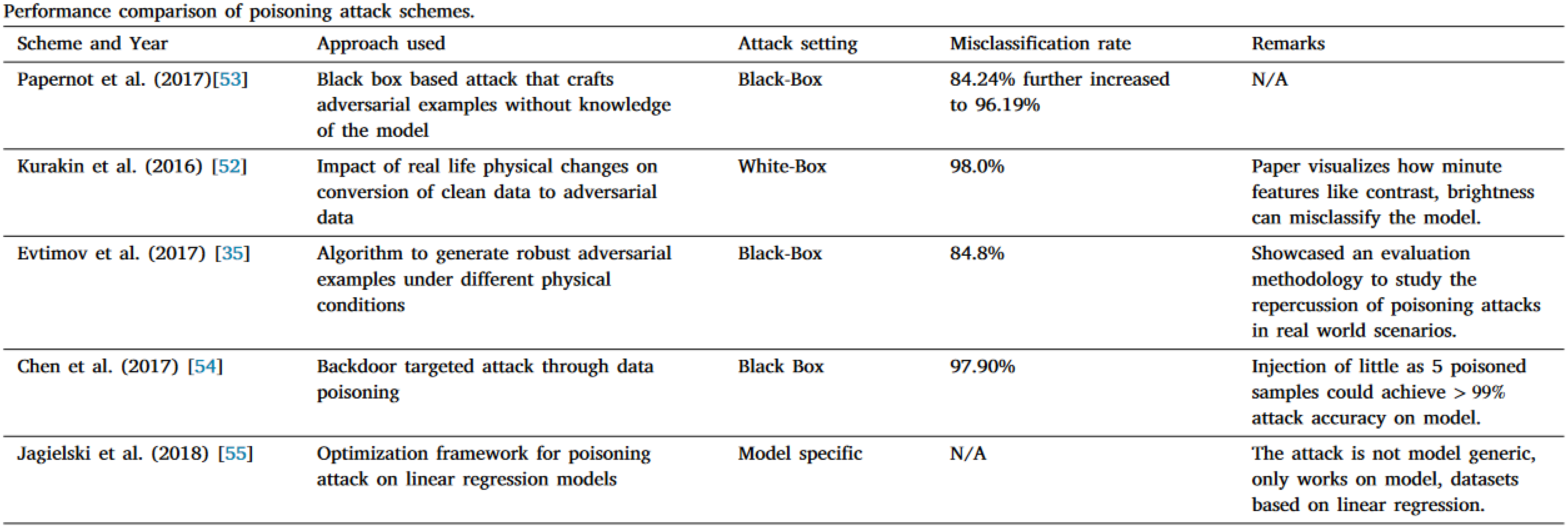


Fig. 8. Parallelization for securing ML task from runtime disruption.

5. Comparative Study

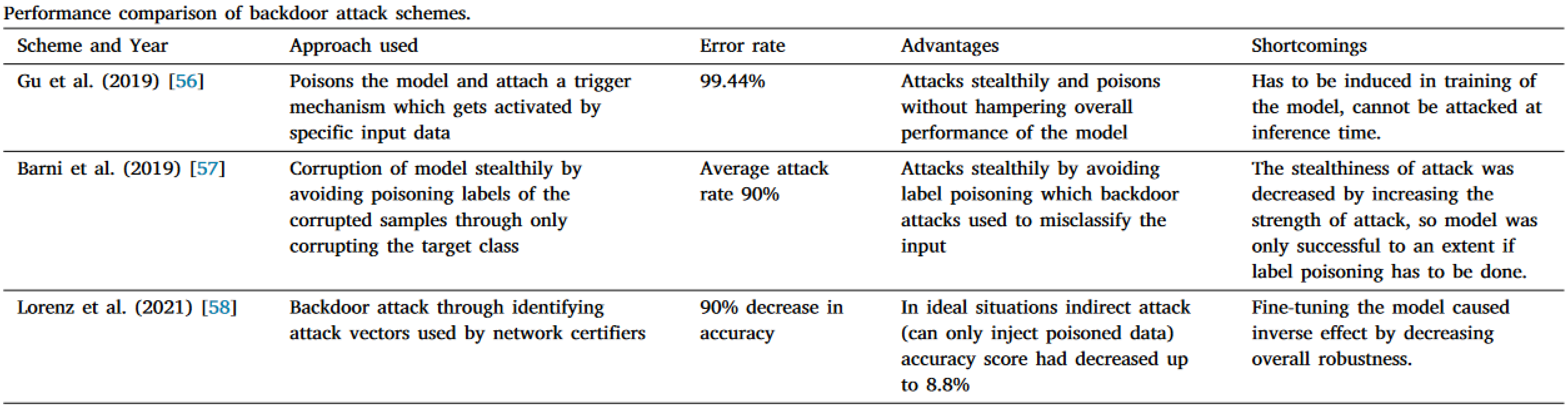
- The study compares the impact of discussed attacks and the performance of their defense mechanisms. The attacks can be classified into four categories, i.e., poisoning attacks, privacy, accessing and inference attacks..



Performance comparison of poisoning attack schemes.

Scheme and Year	Approach used	Attack setting	Misclassification rate	Remarks
Papernot et al. (2017)[53]	Black box based attack that crafts adversarial examples without knowledge of the model	Black-Box	84.24% further increased to 96.19%	N/A
Kurakin et al. (2016) [52]	Impact of real life physical changes on conversion of clean data to adversarial data	White-Box	98.0%	Paper visualizes how minute features like contrast, brightness can misclassify the model.
Evtimov et al. (2017) [35]	Algorithm to generate robust adversarial examples under different physical conditions	Black-Box	84.8%	Showcased an evaluation methodology to study the repercussion of poisoning attacks in real world scenarios.
Chen et al. (2017) [54]	Backdoor targeted attack through data poisoning	Black Box	97.90%	Injection of little as 5 poisoned samples could achieve > 99% attack accuracy on model.
Jagielski et al. (2018) [55]	Optimization framework for poisoning attack on linear regression models	Model specific	N/A	The attack is not model generic, only works on model, datasets based on linear regression.

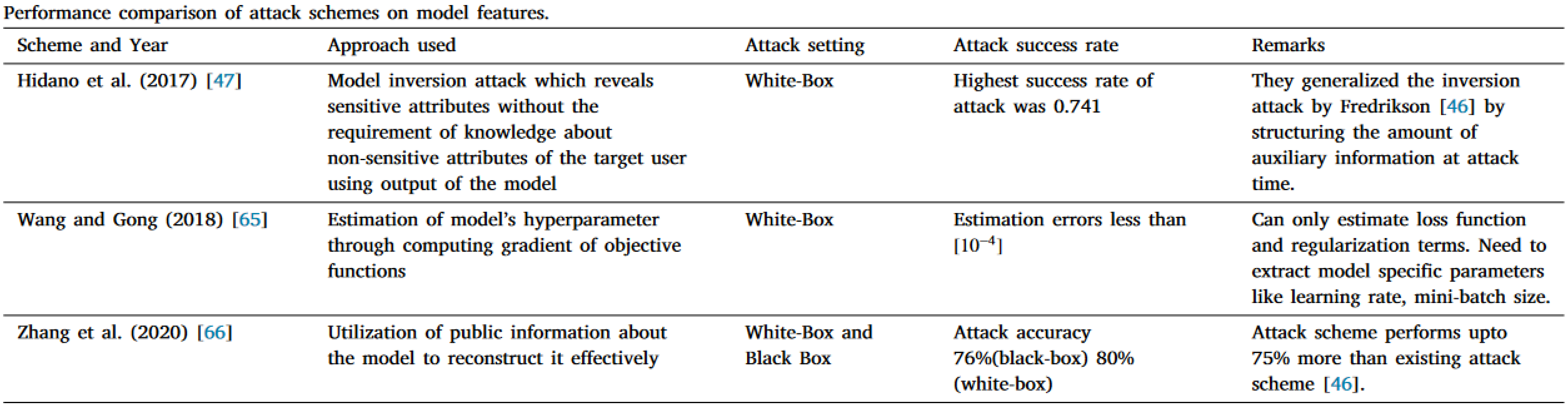
5. Comparative Study



Performance comparison of backdoor attack schemes.

Scheme and Year	Approach used	Error rate	Advantages	Shortcomings
Gu et al. (2019) [56]	Poisons the model and attach a trigger mechanism which gets activated by specific input data	99.44%	Attacks stealthily and poisons without hampering overall performance of the model	Has to be induced in training of the model, cannot be attacked at inference time.
Barni et al. (2019) [57]	Corruption of model stealthily by avoiding poisoning labels of the corrupted samples through only corrupting the target class	Average attack rate 90%	Attacks stealthily by avoiding label poisoning which backdoor attacks used to misclassify the input	The stealthiness of attack was decreased by increasing the strength of attack, so model was only successful to an extent if label poisoning has to be done.
Lorenz et al. (2021) [58]	Backdoor attack through identifying attack vectors used by network certifiers	90% decrease in accuracy	In ideal situations indirect attack (can only inject poisoned data) accuracy score had decreased up to 8.8%	Fine-tuning the model caused inverse effect by decreasing overall robustness.

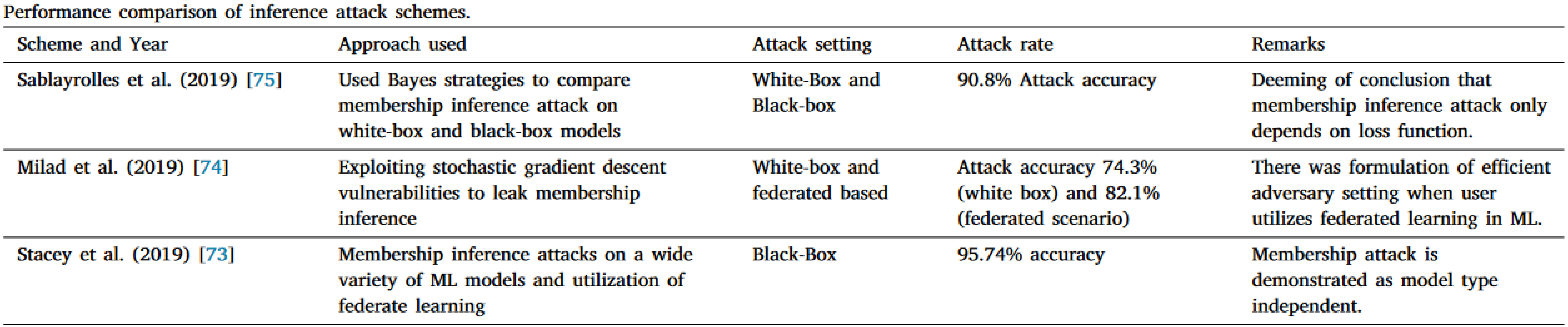
5. Comparative Study



Performance comparison of attack schemes on model features.

Scheme and Year	Approach used	Attack setting	Attack success rate	Remarks
Hidano et al. (2017) [47]	Model inversion attack which reveals sensitive attributes without the requirement of knowledge about non-sensitive attributes of the target user using output of the model	White-Box	Highest success rate of attack was 0.741	They generalized the inversion attack by Fredrikson [46] by structuring the amount of auxiliary information at attack time.
Wang and Gong (2018) [65]	Estimation of model's hyperparameter through computing gradient of objective functions	White-Box	Estimation errors less than [10^{-4}]	Can only estimate loss function and regularization terms. Need to extract model specific parameters like learning rate, mini-batch size.
Zhang et al. (2020) [66]	Utilization of public information about the model to reconstruct it effectively	White-Box and Black Box	Attack accuracy 76%(black-box) 80% (white-box)	Attack scheme performs upto 75% more than existing attack scheme [46].

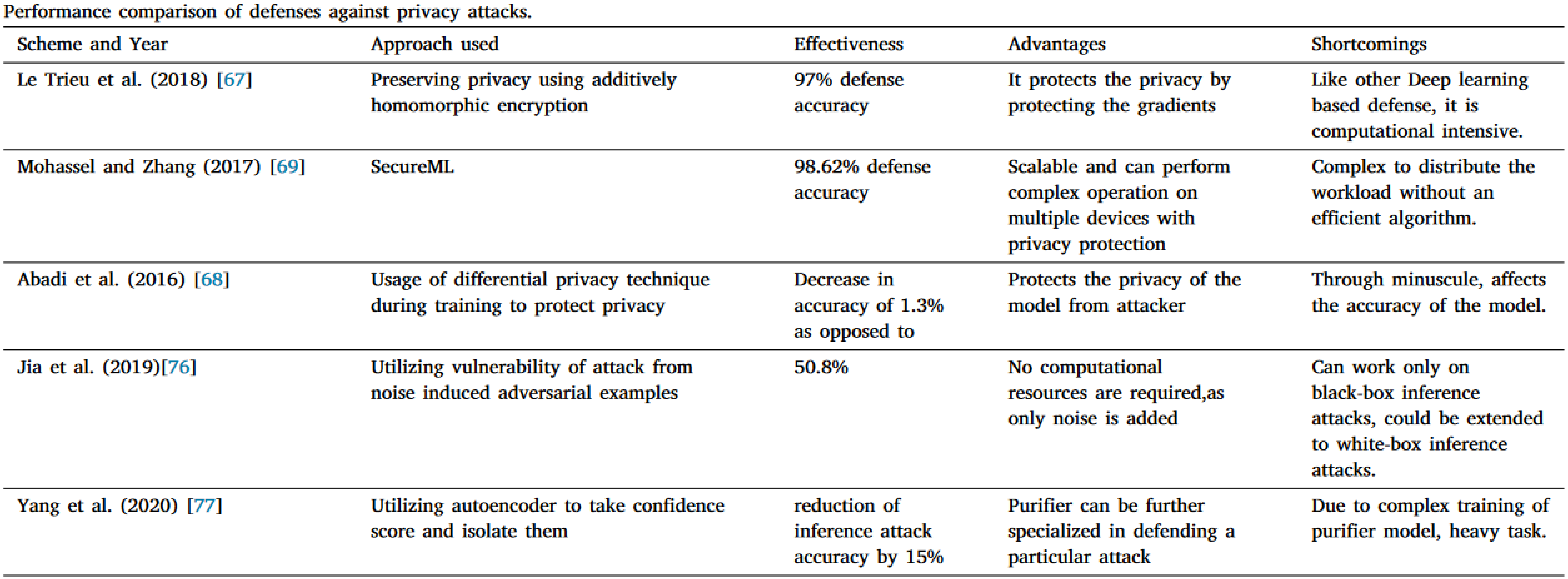
5. Comparative Study



Performance comparison of inference attack schemes.

Scheme and Year	Approach used	Attack setting	Attack rate	Remarks
Sablayrolles et al. (2019) [75]	Used Bayes strategies to compare membership inference attack on white-box and black-box models	White-Box and Black-box	90.8% Attack accuracy	Deeming of conclusion that membership inference attack only depends on loss function.
Milad et al. (2019) [74]	Exploiting stochastic gradient descent vulnerabilities to leak membership inference	White-box and federated based	Attack accuracy 74.3% (white box) and 82.1% (federated scenario)	There was formulation of efficient adversary setting when user utilizes federated learning in ML.
Stacey et al. (2019) [73]	Membership inference attacks on a wide variety of ML models and utilization of federate learning	Black-Box	95.74% accuracy	Membership attack is demonstrated as model type independent.

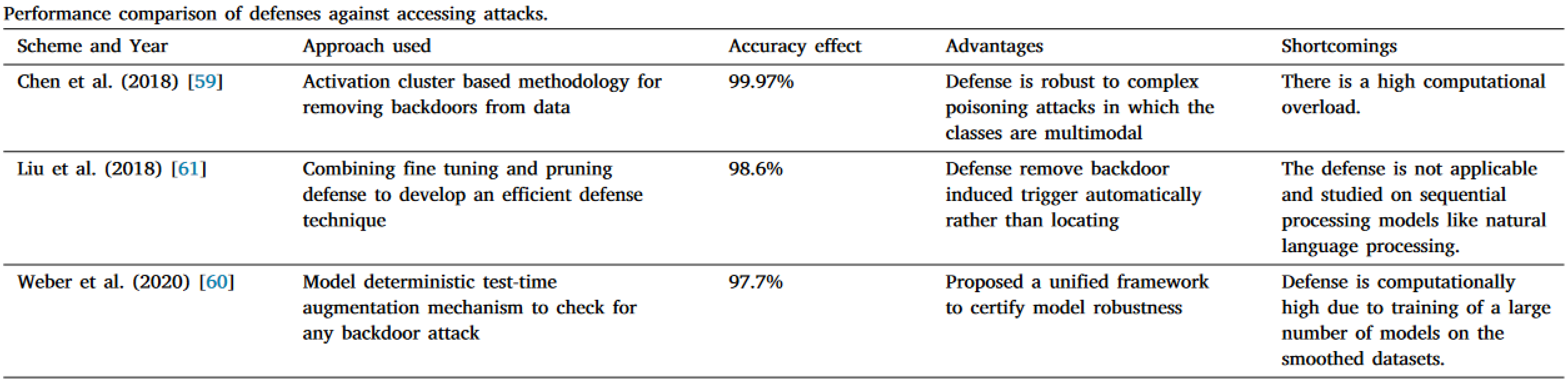
5. Comparative Study



Performance comparison of defenses against privacy attacks.

Scheme and Year	Approach used	Effectiveness	Advantages	Shortcomings
Le Trieu et al. (2018) [67]	Preserving privacy using additively homomorphic encryption	97% defense accuracy	It protects the privacy by protecting the gradients	Like other Deep learning based defense, it is computational intensive.
Mohassel and Zhang (2017) [69]	SecureML	98.62% defense accuracy	Scalable and can perform complex operation on multiple devices with privacy protection	Complex to distribute the workload without an efficient algorithm.
Abadi et al. (2016) [68]	Usage of differential privacy technique during training to protect privacy	Decrease in accuracy of 1.3% as opposed to	Protects the privacy of the model from attacker	Through minuscule, affects the accuracy of the model.
Jia et al. (2019)[76]	Utilizing vulnerability of attack from noise induced adversarial examples	50.8%	No computational resources are required,as only noise is added	Can work only on black-box inference attacks, could be extended to white-box inference attacks.
Yang et al. (2020) [77]	Utilizing autoencoder to take confidence score and isolate them	reduction of inference attack accuracy by 15%	Purifier can be further specialized in defending a particular attack	Due to complex training of purifier model, heavy task.

5. Comparative Study



Performance comparison of defenses against accessing attacks.

Scheme and Year	Approach used	Accuracy effect	Advantages	Shortcomings
Chen et al. (2018) [59]	Activation cluster based methodology for removing backdoors from data	99.97%	Defense is robust to complex poisoning attacks in which the classes are multimodal	There is a high computational overload.
Liu et al. (2018) [61]	Combining fine tuning and pruning defense to develop an efficient defense technique	98.6%	Defense remove backdoor induced trigger automatically rather than locating	The defense is not applicable and studied on sequential processing models like natural language processing.
Weber et al. (2020) [60]	Model deterministic test-time augmentation mechanism to check for any backdoor attack	97.7%	Proposed a unified framework to certify model robustness	Defense is computationally high due to training of a large number of models on the smoothed datasets.

6. Conclusion and Future Research Directions

- The study provides details of various Machine learning security attacks (i.e., dataset poisoning attack, model poisoning attack, privacy breach, membership inference attack, runtime disruption attack), which are possible on the machine learning models deployed in the cyber physical systems.
- Defense mechanisms are discussed and compared to prevent ML adversarial attacks.
- Further, the study discussed the issues and challenges (like the security of deployed mechanism, accuracy, failure of deployed security mechanisms, etc.) of ML security deployed for the cyber physical systems.
- Finally, a comparative study of the performance of the ML models under the influence of various ML attacks along with the performance of various defense mechanisms is provided.

6. Conclusion and Future Research Directions –

Future Research Directions

- **Uses of deep learning algorithms:** The research field has intended to focus primarily on the enhancement of pre-existing neural models, and more work should be undertaken on how new models such as generative adversarial networks could increase their robustness from a cyber security perspective point of view.
- **Undercover impact of federated learning**: Although federated learning has been introduced to a vast extent with the incorporation of big data processing, very little work has been proposed in its effectiveness to counter an attack during the deployment of the ML tasks.
- **Computational factor:** Numerous works have been discussed upon incorporating cost and resources due to the use of computationally high tasks in protecting the integrity of the ML tasks. However, significant work is still required to propose some lightweight schemes to preserve the privacy of the ML processes.
- **Lack of standardization of evaluation parameters:** The studies on the schemes on various attacks and their defenses in ML systems have distinct performance parameters. They do not have a general correlation of performance parameters (i.e., accuracy, F1-score, detection rate, and false-positive rate). Hence, there should be some standardized evaluation parameters. .

6. Conclusion and Future Research Directions –

Future Research Directions

- **Handing of heterogeneous data:** Data in the cyber physical system comes from a variety of sources, each with its own set of qualities and characteristics. As a result, the ML model has a difficult time dealing with it. We will need to put in more effort in this procedure, particularly in the data preprocessing. These issues also exist in ML security, which should be handled carefully by the other researchers working in the same domain.

- **Full proof security:** Researchers in ML security always attempt to build a security scheme that can mitigate the numerous possible threats associated with the ML models. However, there are situations when methods are insufficient to prevent attacks on the ML model. As a result, rigorous testing, analysis, and validation should be performed prior to the deployment of security schemes in order to discover vulnerabilities in those schemes. This could be an important research direction in the future

Thank you

Presented by: Mikail Mohammed Salim