Machine Learning–based Cyber Attacks Targeting on Controlled Information: A Survey

YUANTIAN MIAO, School of Software and Electrical Engineering, Swinburne University of Technology
CHAO CHEN, School of Software and Electrical Engineering, Swinburne University of Technology
LEI PAN, School of Information Technology, Deakin University
QING-LONG HAN, School of Software and Electrical Engineering, Swinburne University of Technology
JUN ZHANG, School of Software and Electrical Engineering, Swinburne University of Technology
YANG XIANG, School of Software and Electrical Engineering, Swinburne University of Technology

Presentation: Oscar Llerena

Abstract

- **Stealing attack** against **controlled information** has become an emerging cyber security **threat** in recent years.
- Due to the boom of **advanced analytics solutions**, **novel stealing attacks utilize machine learning (ML) algorithms** to achieve high success rate and cause a lot of damage.
- Detecting and defending against such attacks is challenging and urgent so that **governments, organizations, and individuals should attach great importance to the ML-based stealing attacks**.
- **This survey presents** the **recent advances** in this new type of attack and corresponding **countermeasures**.
- The ML-based stealing attack is reviewed in perspectives of three categories of targeted controlled information, including **controlled user activities**, **controlled ML model-related information**, and **controlled authentication information**.
- **Recent publications** are summarized to generalize an **overarching attack methodology** and to **derive the limitations** and **future directions of ML-based stealing attacks**. Furthermore, **countermeasures** are proposed towards developing effective protections from three aspects**—detection, disruption, and isolation**.

- **Key Words**: Cyber attacks, machine learning, information leakage, cyber security, controlled information.

Topics

3. ML-BASED STEALING ATTACKS & PROTECTIONS
    3.1 Controlled User Activities Information
        3.1.1 Stealing controlled user activities from kernel data
        3.1.2 Stealing controlled user activities using sensor data
    3.2 Controlled ML Model Related Information
        3.2.1 Stealing controlled ML model description
        3.2.2 Stealing controlled ML model's training data.
    3.3 Controlled Authentication Information
        3.3.1 Stealing controlled keystroke data for authentication
        3.3.2 Stealing controlled secret keys for authentication
        3.3.3 Stealing controlled password data for authentication
4. CHALLENGES AND FUTURE WORKS

3. ML-BASED STEALING ATTACKS & PROTECTIONS

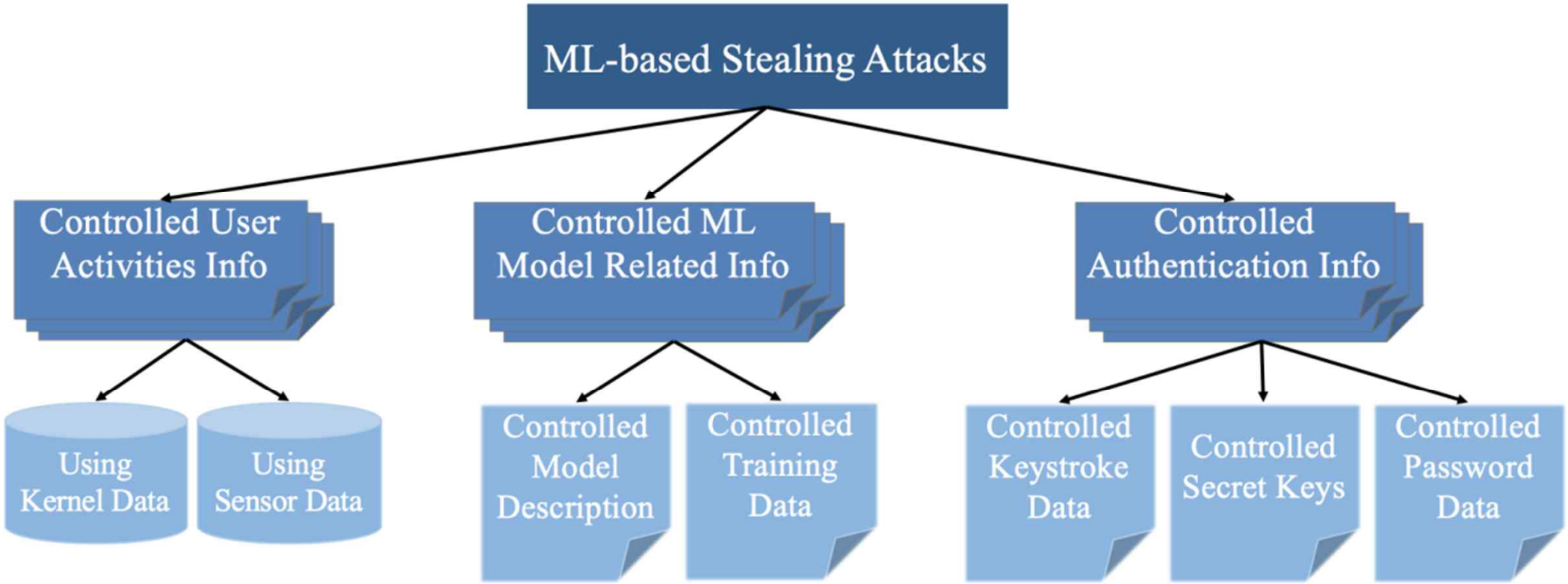
Reference	Year	Targeted Info	Accessible Data	Goals
[26]	2016	Unlock pattern; Foreground app	Hardware interrupt data	Unlock pattern & foreground app inference attacks via analyzing interrupt time collected from interrupt log file.
[119]	2018	Visited websites; Foreground app	Interrupt data; Network &Memory process record	Search and attack the kernel records leaking user's specific events (i.e. app starts, website launch, keyboard gesture).
[151]	2018	Visited websites; Foreground app; Map	Memory data; Network source; File system data	Several side-channel inference attack on iOS mobile device.
[136]	2015	Visited websites; Input keystrokes	Kernel data-structure fields	Protect by injecting noise into the value of kernel data structure values to secure $procfs$.
[50]	2016	Manufacturing activities	Acoustic sensor data; Magnetic sensor data	An attack capture acoustic & magnetic sensor data to steal a manufacturing process specification or a design.
[117]	2017	User activities info	Sensor data	Contextual model detect malicious behavior of sensors like leaking.
[125]	2016	Parameters of an ML model	Input features & Query outputs	Model extraction attacks leverage confidence info with predictions against MLaaS APIs in black-box setting.
[101]	2017	Internal info of an ML model	Input features & Query outputs	Build a local model to substitute the target model and use it craft adversarial examples in black-box setting.
[131]	2018	Hyperparameters of an ML model	Input features & Query outputs	Hyperparameters stealing attack via observing minima objective function against MLaaS in black-box setting.
[98]	2018	Hyperparameters of an ML model	Input features & Query outputs	Build a metamodel to predict hyperparameters with a given classifier in black-box setting to generate adversarial examples.
[34]	2015	Training data for an ML model	Input features & Query outputs & model structure	Model inversion attacks used confidence info leaking training samples with predictions against MLaaS in two settings.
[49]	2017	Training data for an ML model	Input features & Query outputs & model structure	Online Attack using GAN against collaborative deep learning model leaking user's training sample.
[116]	2017	Training data for an ML model	Input features & Query outputs	Membership inference attacks use shadow training technique to leak the specific record's membership of original training set.
[110]	2019	Training data for an ML model	Input features & Query outputs	Enlarge the scope of membership inference attacks by releasing some key assumptions.
[38]	2018	The property of training set	Input features & Query outputs & model structure	Infer global properties of the training data unintended to be shared in white-box setting.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

Reference	Year	Targeted Info	Accessible Data	Goals
[91]	2019	The property of training set	Input features & Query outputs & model structure	Membership inference attacks against collaborative deep learning model leaking others' unintended feature.
[95]	2018	Training data for an ML model	Input features & Query outputs	Protect against black-box membership inference attack using an adversarial training algorithm.
[100]	2017	Training data for an ML model	Input features & Query outputs	Protect training set of model from leakage with teacher and student models using PATE.
[70]	2017	Training data for an ML model	N/A	Protect training dataset in stored from leakage before training.
[79]	2015	Input PINs; User input texts	Acoustic sensor data; Accelerometer data	Attack infers users' inputs on keyboards via accelerometer data within user's smartwatch.
[122]	2016	Input PINs; User input texts	Audio sensor data	Attack infers a user's typed inputs from surreptitious video recordings of a tablet's backside motion.
[42]	2018	Cryptographic keys	TLB Cache data	TLBleed attack TLBs to leak secret keys about victim's memory activities via reversing engineer and ML strategies.
[152]	2016	Secret keys	CPU Cache data	Mitigate access-driven side-channel attacks with CacheBar managing memory pages cacheability.
[132]	2016	Password info	PII & leaked password & site info	Attack with seven mathematical guessing models for seven password guessing scenario using different personal info.
[128]	2014	Password info	Corpus & Site leaked list	Password guessing attack by analyzing its semantic patterns.
[90]	2016	Password info	Corpus library	Mitigate against password guessing attack by modeling password guessability in password creation stage.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

Fig. 1. Introduced Stealing Controlled Information Attack Categories. (Info: information)



3. ML-BASED STEALING ATTACKS & PROTECTIONS

- Reviews of **core papers** regarding **MLBSA methodology**.
- **Hierarchically review** according to Fig. 1.
- Section 3.1 is based MLBSA on different kinds of **accessible data**,
- Section 3.2 and Section 3.3 are grouped by MLBSA on **different kinds** of **controlled** targeted **information attack**. The **attack methods** and **countermeasures** are discussed.
- **Table 2:** relevant high-quality papers from 2014-2019 about **information leakage threat and** the **stealing attack** (columns: controlled information, the accessible data and the goal).
- **Tables 3 to 11** summarize all **subclasses of MLBSA** of the review. For each, "**dataset for an experiment**", "**dataset description**", "**feature engineering** (/targeted ML model)", and "**ML-based attack methods**" is addressed.
- The information of the **dataset** and **source code** for these attacks are listed on **Github 1.**

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.1 Controlled User Activities Information

- It is **essential** for security specialists to **protect user activities information**.
- **private activities** are valuable to adversaries, but also the adversary can **exploit** some **specific activities** (i.e. foreground app) to perform malicious attacks such as the phishing attack [26].
- Attackers pursue two types of data —
  - kernel data and
  - sensor data.
- According to the utilized **kernel data** and **sensor data**, **controlled user activities information** were stolen through **timing analysis** and **frequency analysis**.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

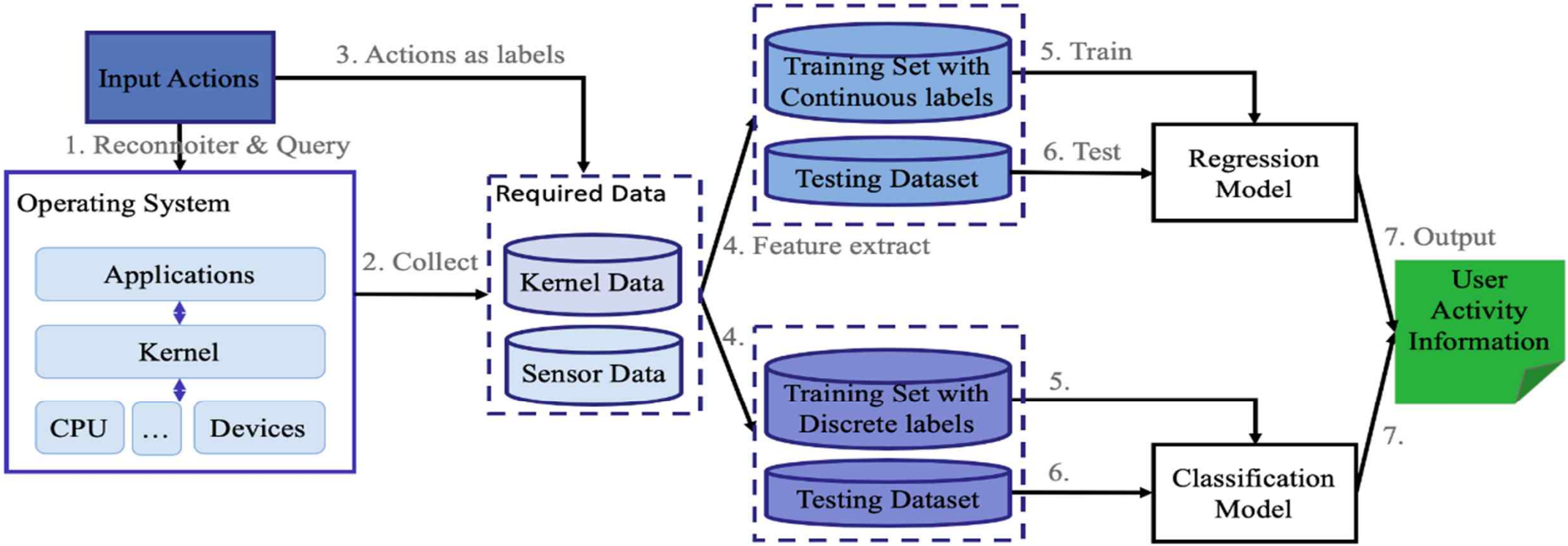


Fig. 4. The ML-based stealing attack against user activities information. As stated in Section 2, *reconnoiter and query* in the reconnaissance phase aim to gain the data that is accessible and valuable to the attack. The required data for this attack can be categorized as kernel data and sensor data. In the data collection phase, these datasets are *collected* and *labeled with input actions* as ground truth. The label's value can be either continuous (i.e. a series of lines for one unlock pattern [26]) or discrete (i.e. various apps [119]). Upon completion of *extracting features*, the training set will be used to *train* the attack model, and the testing set is prepared to *test* and evaluate the model with its *outputs*. Herein, regression models predict the output as a continuous value (i.e. swipe lines), whilst classification models predict a discrete value (i.e. a foreground app)

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.1 Controlled User Activities Information

3.1.1 Stealing controlled user activities from kernel data

Stealing User Activities with Timing Analysis: [26, 119]
Stealing User Activities with iOS Side-channel Attack: [151]
Protection using Privacy Mechanism: [136]

Reference	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[26]	Interrupt data for unlock pattern and for apps	Collect from $procfs$	Deduplication; Interpolation; Interrupt Increment Computation; Gram Segmentation; DTW	HMM with Viterbi algorithm; k-NN classifier with DTW
[119]	Time series for apps, website,keyboard guests	Collect from $procfs$	Automatically extract with $tsfresh$; DTW	Viterbi algorithm with DTW; SVM classifier with DTW
[151]	1200 x 6 time series of data about app; 1000 website traces	120 apps(App Store+iOS) +10 trace x 6 time series; 10 traces for each website	Manually defined; SAX, BoP representation	SVM classifier; k-NN classifier with DTW
[136]	Consecutively reading data; Resident size field data	Collect from $procfs$	N/A; Construct a histogram binned into seven equal-weight bins	SVM classifiers

Table 3. Stealing Controlled User Activities using Kernel Data

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.1 Controlled User Activities Information

3.1.2 Stealing controlled user activities using sensor data

Stealing Machine's Activities with Sensor-based Attack: [50]
Context-aware Sensor-based Detector: [117]

Reference	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[50]	Audio signature dataset	Recorded with a phone put within 4 inches of the printer	STFT, noise normalization	A regression model
[117]	Sensor dataset	Sensor data collected benign and malicious activities	N/A	Markov Chain, NB, LMT, (alternative algorithms e.g. PART)

Table 4. Stealing Controlled User Activities using Sensor Data

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.1 Controlled User Activities Information

- ML-based attacks **steal user activities information from operating systems**.
- According to the data sources, there are two kinds of attacks —
  - using kernel data and
  - using sensor data.
- Kernel data reveals some system-level behaviors of the target system, while
- Sensor data reflects the system's reactions on specific functionality used by users [26].
- The kernel data analyzed with time dimension, sensor data with frequency analysis.
- Countermeasures:
  - Differential privacy is a method to prevent stealing user activities information.
  - Noise injection in [136, 150] to an accessible data source (like Android kernel log files).
  - Access restriction to accessible data [151].
  - Build a model to detect potential stealing threats like in [117].

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

- ML model related information consists of the model description, training data information, testing data information, and testing results.
- The ML model and users' uploaded training data (cloud) are the targets.
- By querying the model via MLaaS APIs, the prediction/classification results are displayed. The model description and training data information are controlled, otherwise, it is easy for an attacker to interpret the victim's query result. As most of ML services charge users per query [41 , 92 , 112], this kind of attack may cause huge financial losses.
- Additionally, several ML models including neural networks are suffered from adversarial examples. Adding small but intentionally worst-case perturbations to inputs, adversarial examples result in the model predicting incorrect answers [39]. By revealing the knowledge of either the model's internal information or its training data, the stealing attack can facilitate the generation of adversarial examples [98 , 101].

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

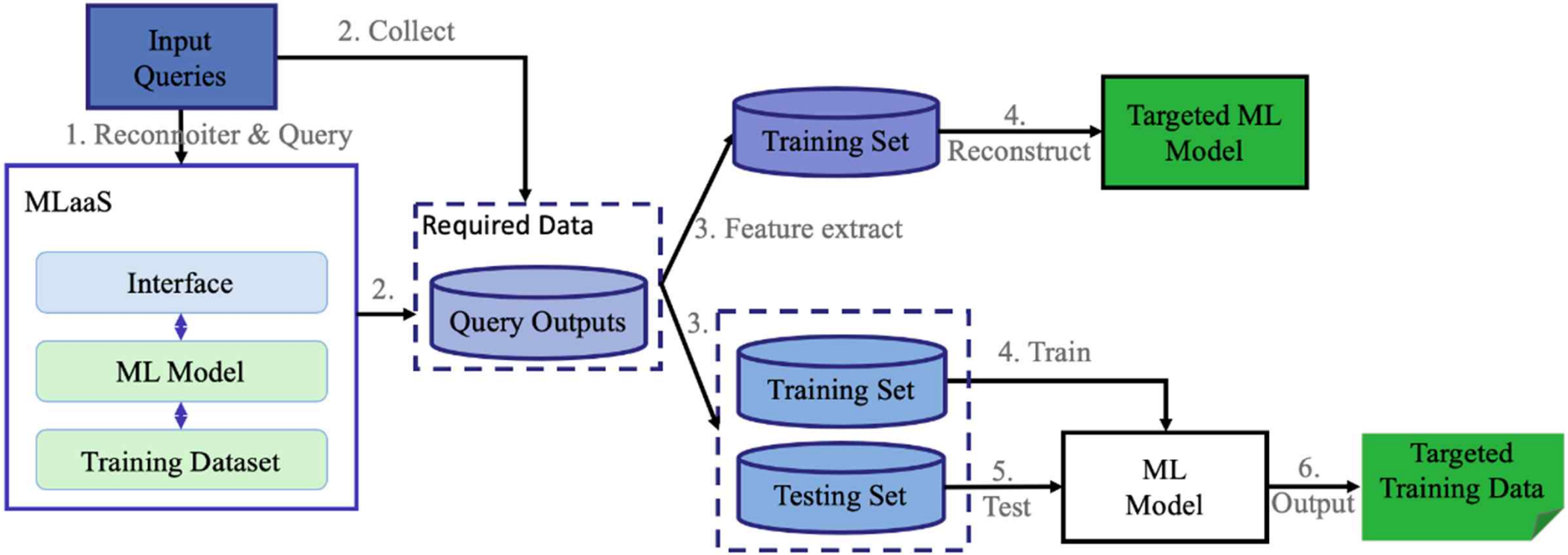


Fig. 5. The ML-based stealing attack against ML model related information. In this category, ML-based attacks aim at stealing the training samples or the ML model. Stealing the controlled training sample attacks use an ML model to determine whether the input sample is contained in the target training set.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.1 Stealing controlled ML model description

Stealing Parameters Attack: [125, 101]
Model extraction attacks targeting ML models of the MLaaS systems. The goal of the model extraction attacks was constructing the adversary's own ML model which closely mimics the original model on the MLaaS platform. That is, the constructed ML model can duplicate the functionality of the original one.

Stealing Hyperparameters Attack: [131,98]
Stealing hyperparameters in the objective function of the targeted MLaaS model may result in gaining financial benefits

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.1 Stealing controlled ML model description

Reference	Dataset for Evaluation	Description	Targeted ML Model	Attack Methods
[125]	Circles, Moons, Blobs, 5-Class [125]; Steak Survey [126], GSS Survey [118], Adult (Income/race) [126], Iris [126], Digits [107], Breast Cancer [126], Mushrooms [126], Diabetes [126]	Synthetic, 5,000 with 2 features, Synthetic,1000 with 20 features, 331 records with 40 features, 16,127 records with 101 features, 48,842 records with 108/105 features, 150 records with 4 features, 1,797 records with 64 features, 683 records with 10 features, 8,124 records with 112 features, 768 records with 8 features	Logistic Regression; Decision Tree; SVM; Three-layer NN	Equation-solving attack; Path-finding attack
[101]	MNIST [69], GTSRB [121]	70,000 handwritten digit images, 49,000 traffic signs images	DNN; SVM; k-NN; Decision Tree; Logistic Regression	Jacobian-based Dataset Augmentation
[131]	Diabetes [126], GeoOrig [126], UJIIndoor [126]; Iris [126], Madelon [126], Bank [126]	442 records with 10 features, 1,059 records with 68 features, 19,937 records with 529 features; 100 records with 4 features; 4,400 records with 500 features; 45,210 records with 16 features	Regression algorithms; Logistic regression algorithms; SVM; NN	Equation solving
[98]	MNIST [69]	70,000 handwritten digit images	NNs	Metamodel methods

Table 5. Stealing Controlled ML Model Description

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.2 Stealing controlled ML model's training data.

Model Inversion Attack & Defense: [34]
Stealing the Training Data of Deep Model with GAN: [49]
Membership Inference Attack: [116]
Property Inference Attack: [38, 91]
Protection using Adversarial Regularization: [95]
Protection using PATE: [100]
Protection using Count Featurization: [70]

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.2 Stealing controlled ML model's training data.

Reference	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[34]	FiveThirtyEight survey, GSS marital happiness survey	553 records with 332 features, 16,127 records with 101 features	N/A	Decision Tree, Regression model
[49]	MNIST [69], AT&T [111]	70,000 handwritten digit images, 400 personal face images	Features learned with DNN	Convolutional Neural Network (CNN) with GAN
[116]	CIFAR10 [65], CIFAR100 [65], Purchases [52], Foursquare [140], Texas hospital stays [47], MNIST [25], Adult (income) [126]	6,000 images in 10 classes, 60,000 images in 100 classes, 10,000 records with 600 features, 1,600 records with 446 features, 10,000 records with 6170 features, 10,000 handwritten digit images, 10,000 records with 14 attribute	Regarded shadow model resulted as features and label records as in/out	NN
[110]	Include 6 sets in [116], News [53], Face [68]	Same as above cell, 20,000 newsgroup documents in 20 classes, 13,000 faces from 1,680 individuals	Regarded shadow model resulted as features and label records as in/out	Random Forest, Logistic Regression, Multilayer perceptron

Table 6. Stealing Controlled ML Model's Training Data. A method was proposed in [95] to prevent training set leakage against the membership inference attack [110] which provides a simple attack without using shadow models. Because the methods in [70, 100] were proposed for only protecting training data, the feature engineering and methods for the ML-based attack are omitted.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.2 Stealing controlled ML model's training data.

Reference	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[38]	Adult (income) [126], MNIST [69], CelebFaces Attributes [80], Hardware Performance Counters	299,285 records with 41 features, 70,000 handwritten digit images, more than 200K celebrity images, 36,000 records with 22 features	Neuron sorting , Set-based representation	NN
[91]	Face [68], FaceScrub [96], PIPA [149], Yelp-health, Yelp-author [141], FourSquare [140], CSI corpus [129]	13,233 faces from 5,749 individuals, 76,541 faces from 530 individuals, 60,000 photos of 2,000 individuals, 17,938 reviews, 16,207 reviews, 15,548 users in 10 locations, 1,412 reviews	N/A	Logistic regression, gradient boosting, Random Forests
[95]	CIFAR100 [65], Purchase100 [52], Texas100 [47]	60,000 images in 100 classes, 197,324 records with 600 features, 67,330 records with 6,170 features	Regarded shadow model resulted as features and label records as in/out	NN

Table 6. Stealing Controlled ML Model's Training Data. A method was proposed in [95] to prevent training set leakage against the membership inference attack [110] which provides a simple attack without using shadow models. Because the methods in [70, 100] were proposed for only protecting training data, the feature engineering and methods for the ML-based attack are omitted.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.2 Stealing controlled ML model's training data.

Attack Type	Attack Targets		Attack Surfaces		Attacker's Capabilities	
	Model Info	Training Set Info	Training Phase	Inference Phase	Black-box Access	White-box Access
Model extraction attack [125]	YES	no	no	YES	YES	no
Model extraction attack [101]	YES	no	no	YES	YES	no
Hyperparameter stealing attack [131]	YES	no	no	YES	YES	no
Hyperparameter stealing attack [98]	YES	no	no	YES	YES	no
Black-box inversion attack [34]	no	YES	no	YES	YES	no
White-box inversion attack [34]	no	YES	no	YES	no	YES
GAN attack [49]	no	YES	YES	no	no	YES
Membership inference attack [116]	no	YES	no	YES	YES	no
Membership inference attack [110]	no	YES	no	YES	YES	no
Property inference attack [38]	no	YES	no	YES	no	YES
Property inference attack [91]	no	YES	YES	no	no	YES

Table 7. Categories of Stealing ML related information attacks from three perspectives (info: information). As for attack targets, two types of information may be stolen — model internal information and training set information. From attack surfaces, attacks may occur during either model's training phase or inference phase. Considering the attacker's capability, the ML model usually allows either the black-box access or the white-box access. The first category is used for this subsection's organization.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.2 Stealing controlled ML model's training data.

Model's Information	Black-box Access	White-box Access
Predicted Label	YES	YES
Predicted Confidence	YES	YES
Parameters	NO	YES
Hyperparameters	NO	YES

Table 8. Attack's prior knowledge under black-box access and white-box access. The black-box access allows the users to query the model and obtain prediction outputs which include the predicted label and confidence value. The white-box access allows the users to access any information of its model which includes predicted label, predicted confidence, parameters, and hyperparameters.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.2 Controlled ML Model Related Information

3.2.2 Stealing controlled ML model's training data.

- Section 3.2, MLBSA against model related information target at either model descriptions or model's training data.
- The other two ways focus on attacks at training/inference phase and with black-/white-box access [102].
- Model extraction attacks [101, 125] and
- Hyperparameter stealing attacks [98, 131] leak the model's internal information happened at inference phase.
- Attackers steal model's training data mostly at inference phase, except the GAN attack [49] and the property inference attack [91] which happen at training phase of collaborative learning.
- When attacking during training phase, attackers with white-box access to the model can exploit its internal information. As shown in Table 8, the white-box access allows attackers to have more prior knowledge than black-box, which results in high performance of the stealing attack [34].
- On the other hand, black-box attacks can be more applicable in the real world. Except [110], most of the attackers in this category under black-box access know the learning algorithm of the target model [34, 71, 98, 101, 125, 131].
- Countermeasures: Concerning the ML pipeline, the protection methods will be applied in data preprocessing phase, training phase, and inference phase respectively.
- Differential privacy noise used in the first phase can build a privacy- preserving training set [70].
- Differential privacy is the most common countermeasures to defend against the stealing attack, however, it alone cannot prevent the GAN attack [49].
- Differential privacy, regularization, dropout, and rounding techniques are popular protections at the training and inference phases. At the training phase, differential privacy on parameters cannot resist the GAN attack [49], while rounding parameters is ineffective against hyperparameter stealing Manuscript submitted to ACM.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.3 Controlled Authentication Information

3.3.1 Stealing controlled keystroke data for authentication

Keystroke Inference Attack: [79]
Video-Assisted Keystroke Inference Attack: [122]

Paper	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[79]	Acceleration data set	Consecutive vectors with 26 labels	FFT & IFFT filter, Movement capturing, Optimization with change direction	Random Forest; k-NN; SVM; NN
[122]	Video recordings set	Image resolution and frame rate	Extract from selected AOIs' motion signals for motion patterns	multi-class SVM

Table 9. Stealing Controlled Keystroke Data for Authentication.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.3 Controlled Authentication Information

3.3.1 Stealing controlled keystroke data for authentication

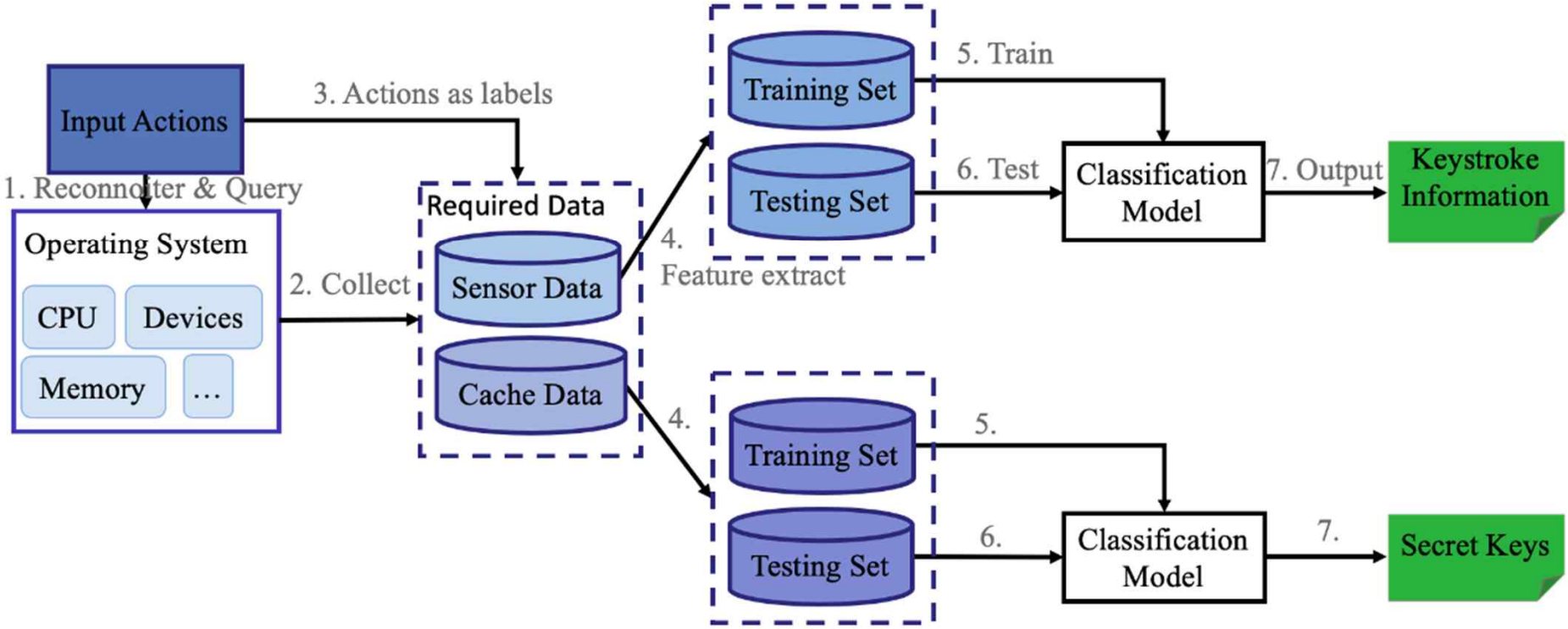


Fig. 6. The ML-based stealing attack against authentication information — keystroke information and secret keys. After *reconnoitering and querying*, attackers targeting at keystroke information and secret keys interact with the target system to *collect* data, which refers to the active collection. The attack involved active collection shares a similar workflow as Fig. 4 depicted.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.3 Controlled Authentication Information

3.3.2 Stealing controlled secret keys for authentication

Stealing secret keys with TLB Cache Data: [42]
Protection Against Leakage from CPU Cache Data: [152]

Reference	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[42]	300 observed TLB latencies	Collect from TLB signals	Encode info using a normalized latencies vector	SVM classifier
[152]	500,000 Prime-Probe trials	Number of absent cache lines + cache lines available	N/A	NB classifier

Table 10. Stealing Controlled Secret Keys for Authentication (Information: info)

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.3 Controlled Authentication Information

3.3.3 Stealing controlled password data for authentication

Online Password Guessing Attack: [132]
Password Guessing with Semantic Pattern Analysis: [128]
Protection with Modeling Password Guessability: [90]

Reference	Dataset for Experiment	Description	Feature Engineering	ML-based Attack Method
[132]	Dodonew, CSDN, 126, Rockyou, 000webhost, Yahoo, 12306, Rootkit; Hotel, 51job	16,258,891 (6,428,277) leaked passwords, 6,392,568 (32,581,870) leaked passwords, 15,251,073 (442,834) leaked passwords, 6,392,568 leaked passwords + 129,303 PII, 69,418 leaked passwords + 69,324 PII; 20,051,426 PII, 2,327,571 PII	N/A	PCFG-based algorithm [134], Markov-based algorithm [85], LD algorithm
[128]	RockYou	32,581,870 leaked passwords	Segmented with NLP	PCFG-based algorithm
[90]	PGS training set [127], 1class8, 1class16 [58], 3class12 [114], 4class8 [88], webhost [12]	33 million passwords, 3,062 (2,054) leaked passwords, 990 (990) leaked passwords, 30,000 leaked passwords	N/A	PCFG-based algorithm [62], Markov models [85], NN

Table 11. Stealing Controlled Password Data for Authentication

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.3 Controlled Authentication Information

3.3.3 Stealing controlled password data for authentication

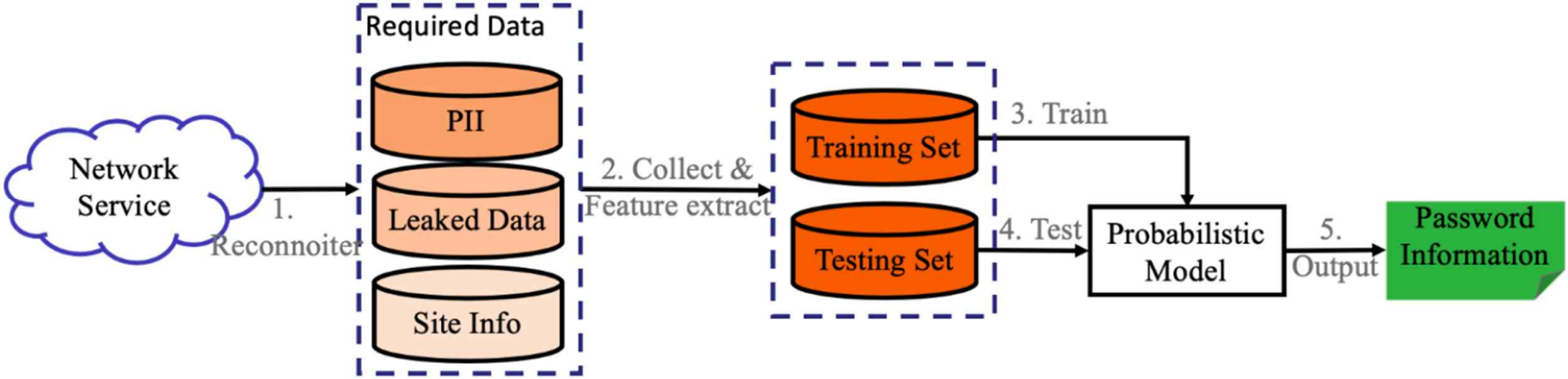


Fig. 7. The ML-based stealing attack against authentication information — password data. To infer the password, attackers *reconnoiter and collect* the online information with the passive collection. For the attack with passive collection, attackers do not need to interact the target service with designed inputs. They collect the required data labeled with semantic categories according to human behaviors of password creation [132] or passwords' generic structure [128]. During the *feature engineering* phase, different segments from the required data are extracted. A semantic classifier is *trained* using probabilistic algorithms. After *testing* this classifier, various passwords can be constructed as *outputs* with the semantic generalization.

3. ML-BASED STEALING ATTACKS & PROTECTIONS

3.3 Controlled Authentication Information

3.3.4 Summary

- ML-based stealing attacks target at users' keystroke authentication, secret keys and passwords.
- Attackers steal users' passwords by cracking the useful information online. For the other two objectives, they exploit the information based on users' activities recorded by an Operating System (i.e. TLB/CPU cache data).
- Password guessing attacks use the probabilistic method to construct a password with the least number of guesses. The attack on the remaining two targets can be transferred as classification tasks by generating keystroke patterns and cache set states.
- Countermeasures:
- From the security perspective, two types of countermeasures are introduced as the access restriction and the attack detection.
- The secret keys, for example, can be protected by managing the accessible related cache data [152].
- The analysis of password guessability [90] can secure the user's account by setting a strong password. The weak passwords are evaded by detection.
- The future direction can target the effectiveness of guessing model prediction which is limited by the sparsity of training samples [90]. The defense for the keystroke inference has not been well-developed. The future work may explore the secured access of related sensor data.

4. CHALLENGES AND FUTURE WORKS

- The recent publications about the ML-based stealing attacks against the controlled information and the corresponding defense methods are reviewed.
- Some attacks can steal the information, but they make strong assumptions of the attacker's prior knowledge. For instance, the attacker is assumed to know the ML algorithm as a necessary condition prior to stealing the model/training samples.
- However, this prior knowledge is not always publicly known in the real world cases. Additionally, the attack methods are not mature technologies and have great room for improvement. Table 2 outlines the target and accessible data for each paper.
- And Table 12 summarizes the core research papers in the perspectives of attack, protection, related ML techniques, and evaluation. The following subsections will discuss the future directions of the ML-based stealing attack and feasible countermeasures as shown in Fig. 8.

4. CHALLENGES AND FUTURE WORKS

Reference	Attack	Protection	Related ML Techniques	Evaluation
[26]	Unlock pattern & foreground app inference attack	Restrict access to kernel resources; Decrease the resolution of interrupt data	HMM with Viterbi algorithm; k-NN classifier with DTW	Success rate; Time & battery consumption
[119]	Leaking specific events attack	Restrict access to kernel resources; App Guardian [94, 150]	k-NN classifier with DTW; Multi-class SVM with DTW	Accuracy; Precision; Recall; Battery consumption
[136]	Keystroke timing attack; website inference attack	Design d^*-private mechanism	Multi-class SVM classifier	Accuracy; Relative AccE
[151]	Stealing user activities; Stealing in-app activities	Eliminate the attack vectors; Rate limiting; Runtime detection [150]; Coarse-grained return values; Privacy-preserving statistics report [136]; Remove the timing channel	SVM classifier; k-NN classifier with DTW	Accuracy; Execution time; Power consumption
[50]	Stealing product's design	Obfuscate the acoustic emissions	A regression model	Accuracy
[117]	Information leakage via a sensor; Stealing information via a sensor	The contextual model detects malicious behavior of sensors	Markov Chain; NB; Alternative set of ML algorithms (e.g. PART)	Accuracy; FNR; F-measure; FPR; Recall; Precision; Power consumption
[125]	Model extraction attack	Rounding confidences [34]; Differential privacy (DP) [28, 55, 71, 130]; Ensemble methods [120]	Logistic regression; Decision tree; SVM; Three-layer NN	Test error; Uniform error; Extraction Faccuracy
[101]	Model extraction attack	Gradient masking [39] and defensive distillation [103] for a robust model	DNN; SVM; k-NN; Decision Tree; Logistic regression	Success rate

And Table 12 summarizes the core research papers in the perspectives of attack, protection, related ML techniques, and evaluation

4. CHALLENGES AND FUTURE WORKS

Reference	Attack	Protection	Related ML Techniques	Evaluation
[101]	Model extraction attack	Gradient masking [39] and defensive distillation [103] for a robust model	DNN; SVM; k-NN; Decision Tree; Logistic regression	Success rate
[131]	Hyperparameters stealing attack	Cross entropy and square hinge loss instead of regular hinge loss	Regression algorithms; NN; Logistic regression; SVM	Relative EE; Relative MSE; Relative AccE
[98]	Hyperparameters stealing attack	N/A	Metamodel methods	Accuracy
[34]	Model inversion attack	Incorporate inversion metrics in training; Degrade the quality/precision of the model's gradient information.	Decision Tree; Regression model	Accuracy; Precision; Recall
[49]	The GAN attack stealing users' training data	N/A	CNN with GAN	Accuracy
[116]	Membership inference attack	Restrict class in the prediction vector; Coarsen precision; Increase entropy of the prediction vector [48]; Regularization	NN	Accuracy; Precision; Recall
[110]	Membership inference attack	Dropout; Model Stacking	Logistic regression; Random Forest; Multilayer perceptron	Precision; Recall; AUC
[38]	Property inference attack	Multiply the weights and bias of each neuron; Add noise; Encode arbitrary data	NN	Accuracy; Precision; Recall
[91]	Property inference attack	Share fewer gradients; Reduce input dimension; Dropout; user-level DP	Logistic regression; Gradient boosting; Random Forests	Precision; Recall; AUC

And Table 12 summarizes the core research papers in the perspectives of attack, protection, related ML techniques, and evaluation

4. CHALLENGES AND FUTURE WORKS

Reference	Attack	Protection	Related ML Techniques	Evaluation
[26]	Unlock pattern & foreground app inference attack	Restrict access to kernel resources; Decrease the resolution of interrupt data	HMM with Viterbi algorithm; k-NN classifier with DTW	Success rate; Time & battery consumption
[119]	Leaking specific events attack	Restrict access to kernel resources; App Guardian [94, 150]	k-NN classifier with DTW; Multi-class SVM with DTW	Accuracy; Precision; Recall; Battery consumption
[136]	Keystroke timing attack; website inference attack	Design d^*-private mechanism	Multi-class SVM classifier	Accuracy; Relative AccE
[151]	Stealing user activities; Stealing in-app activities	Eliminate the attack vectors; Rate limiting; Runtime detection [150]; Coarse-grained return values; Privacy-preserving statistics report [136]; Remove the timing channel	SVM classifier; k-NN classifier with DTW	Accuracy; Execution time; Power consumption
[50]	Stealing product's design	Obfuscate the acoustic emissions	A regression model	Accuracy
[117]	Information leakage via a sensor; Stealing information via a sensor	The contextual model detects malicious behavior of sensors	Markov Chain; NB; Alternative set of ML algorithms (e.g. PART)	Accuracy; FNR; F-measure; FPR; Recall; Precision; Power consumption
[125]	Model extraction attack	Rounding confidences [34]; Differential privacy (DP) [28, 55, 71, 130]; Ensemble methods [120]	Logistic regression; Decision tree; SVM; Three-layer NN	Test error; Uniform error; Extraction Faccuracy
[101]	Model extraction attack	Gradient masking [39] and defensive distillation [103] for a robust model	DNN; SVM; k-NN; Decision Tree; Logistic regression	Success rate

And Table 12 summarizes the core research papers in the perspectives of attack, protection, related ML techniques, and evaluation

4. CHALLENGES AND FUTURE WORKS

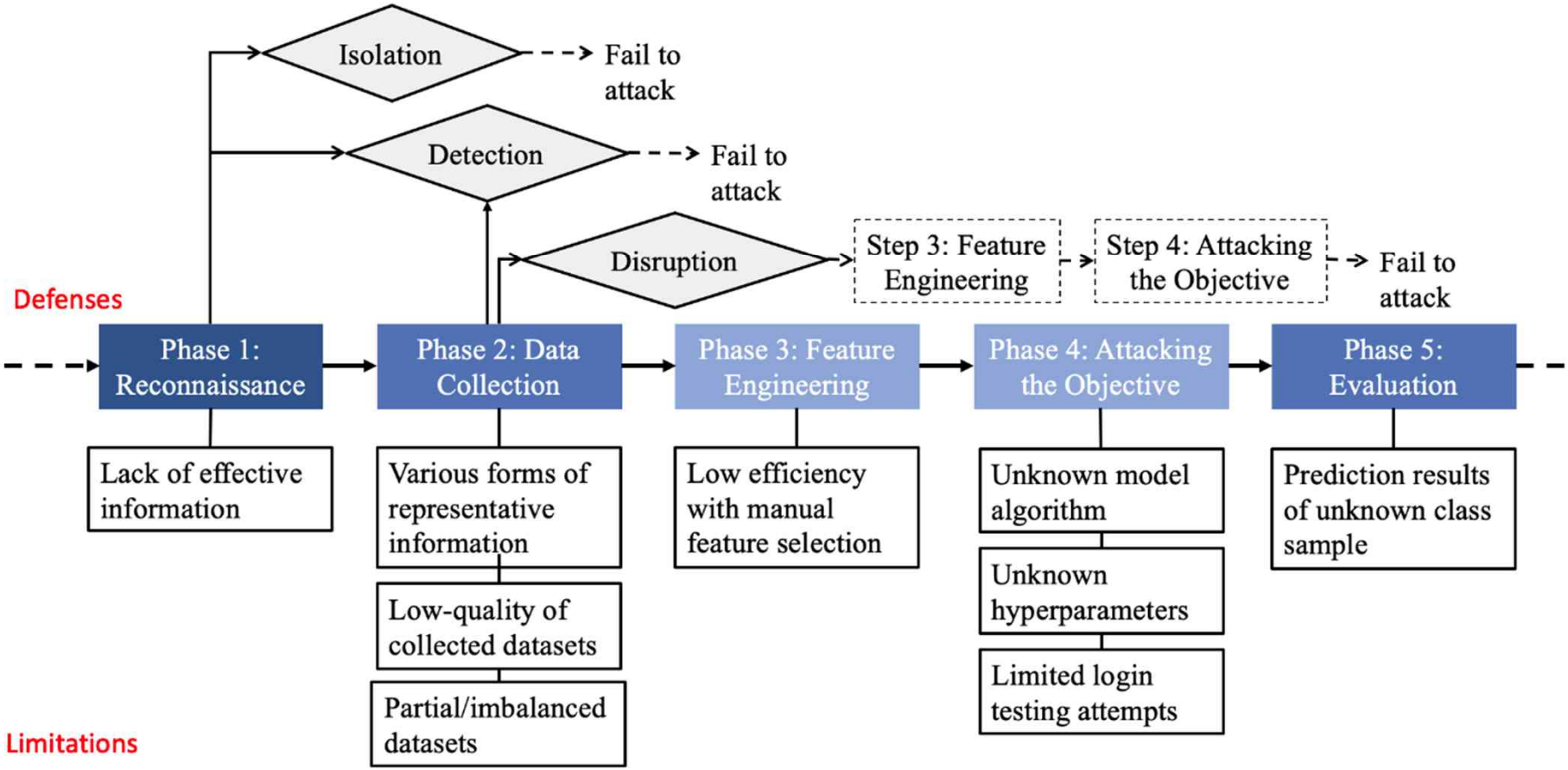


Fig. 8. The Challenges of ML-based Stealing Attack and Its Defenses

4. CHALLENGES AND FUTURE WORKS

4.1 Attack

4.1.1 Reconnaissance

- The reconnaissance phase consists of two main tasks — the target definition and valuable accessible data analysis. The denotation of the target determines which kind of accessible resources is valuable. The further attack mechanism is designed according to the analysis of accessible data during the reconnaissance phase. It is essential to ensure that the information accessible to legitimate users contains valuable information for stealing attacks to succeed.

4.1.2 Data Collection

- Determining the valuable accessible data is only a part of an ML-based stealing attack. To take advantages of the ML mechanism, the valuable dataset collected in this phase should guarantee its representation, reliability, and comprehensiveness. If either one of three is unsatisfactory, then the results of the stealing attack will be inaccurate.
- The first challenge is collecting valuable data with the representative information.
- The second challenge appears while collecting a reliable dataset.
- The third challenge of comprehensive dataset collection involves determining the size/distribution of the training dataset and the testing dataset.

4. CHALLENGES AND FUTURE WORKS

4.1 Attack

4.1.3 Feature Engineering

Feature engineering in the MLBSA methodology intends to refine the collected data for the effective and efficient training process. It is critical to the performance of ML-based attack by eliminating the noise from the collected data. However, among the current research, the techniques used in feature engineering remain underdeveloped.

4.1.4 Attacking the Objective

The main tasks include training and testing the ML model to steal the controlled information. There are a few challenges of stealing attacks with respect to training and testing ML models including unknown model algorithms, unknown hyperparameters of ML model, and the limited amount of testing time.

4. CHALLENGES AND FUTURE WORKS

4.1 Attack

4.1.5 Evaluation

- To effectively infer the controlled information, most of the investigated research applied ML mechanism. The prediction of the unknown testing samples is a challenge for ML-based stealing attacks, as the supervised learning algorithm dominates the attack methods. That is, if the true label of a testing sample has not been learned by the model during the training phase, this sample will be recognized as an incorrect class. The testing samples, which are unknown to the training dataset, affect the evaluation results and subsequently reduce the stealing attack's accuracy. To improve the performance of such attacks, the attacker needs to achieve breakthroughs towards predicting the unknown data

4. CHALLENGES AND FUTURE WORKS

4.2 Defense

- Targeting diverse controlled information, the countermeasures in protecting the information from ML-based stealing attacks are summarized.
  - 1) the detection is indicated as detecting related critical indications;
  - 2) the disruption intends to break the accessible data at a tolerable cost of service's utility; and
  - 3) isolation aims to limit the access to some valuable data sources.
- the countermeasures mainly applied in the first two phases. Specifically, isolation restricts the attacker's access and makes the attack fail at the first phase; and disruption may confuse the attacker in the second phase and hinder the attacker to build a successful attack model.

4. CHALLENGES AND FUTURE WORKS

4.2 Defense

Targeting diverse controlled information, the countermeasures in protecting the information from ML-based stealing attacks are summarized. In general, the countermeasures can be summarized into three groups: 1) the detection is indicated as detecting related critical indications; 2) the disruption intends to break the accessible data at a tolerable cost of service's utility; and 3) isolation aims to limit the access to some valuable data sources. As shown in Fig. 8, the countermeasures mainly applied in the first two phases. Specifically, isolation restricts the attacker's access and makes the attack fail at the first phase; and disruption may confuse the attacker in the second phase and hinder the attacker to build a successful attack model. The detection techniques can detect the attacker's actions and then protect the information from being stolen. These issues are explained as follows.

4.2.1 Detection

To detect potential stealing attacks in advance, the relevant crucial indications are required by analyzing the functionality related to the controlled information. Defenders should notice the attackers' actions as soon as the attackers start the reconnaissance or the data collection processes. Based on the attacker's future directions, the detection is proposed accordingly in order to prevent the attack at an early stage and minimize the loss of stealing the controlled information. Manuscript submitted to ACM.

4. CHALLENGES AND FUTURE WORKS

4.2 Defense

4.2.2 Disruption

Disruption can protect the controlled information via obstructing the information used in each phase of the MLBSA methodology. Disrupting the accessible data currently involves two methods as adding noise to data sources and degrading the quality/precision of service's outputs. For more advanced countermeasures, further research needs to better understand the attacker's future directions

4.2.3 Isolation

Isolation can assist the system by eliminating the information stealing threat, which hinders the attacker from progressing through the reconnaissance phase. No matter how attackers improve their strategies and techniques, isolation can protect the controlled information by restricting access to the data of interest. Specifically, it is effective to control the accessible data via restricting the access or managing the dynamic permission [26, 79 , 116, 119]. When the stealing attacks advance, defenders can apply ML techniques to automatically control as many as possible accesses related to the targeted controlled information. However, this protection is highlighted to be applied cautiously by concerning the utility of the service. On the one hand, specialists can remove the some information channels Manuscript submitted to ACM

5. CONCLUSION

- The ML-based stealing attack against the controlled information and the defense mechanisms are reviewed.
- The generalized MLBSA methodology compatible with the published work is outlined.
- Specifically, the MLBSA methodology uncovers how adversaries steal the controlled information in five phases, i.e. reconnaissance, data collection, feature engineering, attacking the objective, and evaluation.
- Based on different types of the controlled information, the literature was reviewed in three categories consisting of:
    - The controlled user activities information,
    - The controlled ML model related information, and
    - The controlled authentication information.
- The attacker is assumed to use the system without any administrative privilege. This assumption implies that user activities information was stolen by leveraging the kernel data and the sensor data both of which are beyond the protection of the application.
- The attack against the controlled ML model-related information is demonstrated with stealing the model description and/or stealing the training data.
- Similarly, keystroke data, secret keys, and password data are the examples of stealing the controlled authentication information.
- Besides the stealing attack, the corresponding protections are summarized for each category.
- The challenges clearly go on five attacking phases.
- The future directions matching various limitations are presented. Comparing to the explicit breaking/destroying attack, the controlled information leaked by such stealing attacks is much more difficult to be detected, so that the estimated loss should be extended accordingly.