Detecting practical quantum attacks for continuous-variable quantum key distribution using density-based spatial clustering of applications with noise

Qin Liao,[1,*] Zheng Wang,[1] Haijie Liu,[1] Yiyu Mao,[2,†] and Xiquan Fu[1]
[1]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*
[2]*School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China*

(Received 17 April 2022; accepted 14 July 2022; published 9 August 2022)

Continuous-variable quantum key distribution (CVQKD) has been proven to be secure theoretically. However, the practical CVQKD system may still be subject to various quantum attacks due to the imperfections of devices. In this paper, we suggest a general machine learning-based defense strategy against practical quantum attacks by taking advantage of density-based spatial clustering of applications with noise (DBSCAN), which we called DBSCAN-based attack detection scheme (DADS). Specifically, we first construct a set of features that can well reflect the behaviors of different attacks, then DBSCAN is applied to obtain several clusters. This clustering result can explicitly indicate whether the CVQKD system is being eavesdropped or not. Simulation experiments show that the proposed DADS cannot only detect most of known attacks, but also has ability to identify various unknown attacks, thereby improving practical security of the CVQKD system. We also show that the overestimated secret key rate caused by ignoring practical quantum attacks can be amended by DADS so that a reasonable tighter secure bound of the practical CVQKD system can be obtained.

DOI: 10.1103/PhysRevA.106.022607

I. INTRODUCTION

Continuous-variable quantum key distribution (CVQKD) [1] is a research hot spot in the field of quantum cryptography, which allows two distant legitimate partners, Alice and Bob, to share secret keys after exchanging quantum signals through an untrusted channel. According to different modulation approaches, CVQKD can be divided into Gaussian-modulated CVQKD [2–6] and discretely modulated CVQKD [7–13]. In the former protocol, secret keys are usually encoded on the quadratures of coherent states, which can carry multiple bits in one pulse thereby obtaining higher secret key rate in middle- or short-distance transmission. In the latter protocol, the senders prepare a certain number of nonorthogonal coherent states (for binary, ternary, four-state, and eight-state modulation schemes, the numbers are 2–4, and 8, respectively), and exploit the sign of measured quadratures of each state to encode the bits of secret key rate [14]. At present, CVQKD with Gaussian-modulated coherent states (GMCS) is the most widely used CVQKD protocol, which has been proven to be theoretically secure in both the asymptotic limit [15–17] and the finite-size regime [18,19]. However, the practical GMCS CVQKD system may still be subject to various practical attacks due to the imperfections of devices. For example, the eavesdropper, i.e., Eve, may launch wavelength attacks by exploiting imperfect beam splitters [20–22], she may launch saturation attacks [23] or homodyne-detector-blinding attacks [24] by exploiting the finite linearity domain of homodyne detectors, and she may launch local oscillator (LO) intensity attacks [25] or calibration attacks [26] by exploiting the transmitted LO. The current existing countermeasures against these attacks are to deploy specific monitoring devices for different attack strategies, which can only be carried out after knowing the type of the attacks. However, it is hard for legitimate parties to know in advance what kind of attack Eve will launch in a real communication scenario. Therefore, how to establish a universal attack defense model that can resist most quantum attacks is the focus of current attack defense research for the practical CVQKD system.

In recent years, machine learning has experienced rapid development and is gradually being used to solve problems in CVQKD systems [27–31], especially in attack defense. In 2018, a support vector regression model-based parameter prediction method was proposed to predict the time-along evolution of the LO intensity [32], so as to optimize the performance and practical security of CVQKD systems. Whereafter, Mao *et al.* proposed a hidden-Markov-model-based calibration attack recognition method to identify calibration attack by monitoring the real-time quadrature values measured by Bob [33]. Nevertheless, these methods can only defend against specific attacks, they cannot simultaneously resist multiple attack strategies. Recently, an efficient artificial-neural-network (ANN)-based attack detection method for most of the known attack strategies was proposed [34]. However, as a classification algorithm for supervised learning, ANN takes a long time for data training, which has a poor real-time performance when the data volume is huge. Besides, this method requires precollection of various attacked data, otherwise attacks cannot be correctly classified.

To solve the above-mentioned issues, in this paper, we suggest a general machine learning-based defense scheme against practical quantum attacks by taking advantage of

*llqqlq@hnu.edu.cn
†Corresponding author: yiyumao@csust.edu.cn

density-based spatial clustering of applications with noise (DBSCAN), which we called DBSCAN-based attack detection scheme (DADS). DBSCAN is a kind of unsupervised learning clustering algorithm, it does not require time-consuming data collection and training processes in advance, and it is not sensitive to outliers [35], which are beneficial for efficiently processing the abnormal data in untrusted quantum channel. In particular, we first construct a set of features by analyzing the behaviors of several typical practical quantum attacks, then the measurement results are transformed to feature vectors as the input of DBSCAN. Subsequently, several clusters can be obtained and the pulses that are subjected to quantum attacks can be recognized by comparing the clustering results. The performance of DADS is detailed in terms of machine learning metrics, and its security is also analyzed in both the asymptotic limit and the finite-size regime. Simulation experiments show that the proposed DADS not only has the ability to detect most of known attacks, but also can identify various unknown attacks, thereby improving practical security of the CVQKD system. Moreover, the overestimated secret key rate caused by ignoring practical quantum attacks can be amended with the help of DADS, thereby a reasonable tighter secure bound of practical CVQKD system can be finally obtained.

This paper is organized as follows: In Sec. II, we give a briefly introduction of DBSCAN and detail the main process of DADS. In Sec. III, the performance of DADS in terms of machine learning metrics is analyzed. In Sec. IV, we discuss the security of DADS in both the asymptotic limit and the finite-size regime. Finally, the conclusion is drawn in Sec. V.

II. DBSCAN-BASED ATTACK DETECTION SCHEME

In this section, we first briefly introduce the principle of DBSCAN. Then the proposed DADS is described in detail.

A. DBSCAN

DBSCAN is a well-known density-based clustering algorithm that can discover arbitrarily shaped clusters without specifying the number of clusters in the input [36]. It has excellent potential to segment complex and irregularly shaped objects, and it is widely used in spatial data mining. As known, the practical CVQKD system may subject to various quantum attacks, resulting in several outliers during measurement, these abnormal data can be efficiently recognized by DBSCAN. To make the derivation self-contained, we briefly describe the principle of DBSCAN, and details can be found in Ref. [37].

In DBSCAN, ε and $P_{\min}$ are two key parameters that determine the effect of clustering [35]. Therein, ε is the radius of each point, and $P_{\min}$ is the threshold number of points in the ε neighborhood required for a point to become a core point. Several relevant definitions are introduced as follows.

Definition 1. ε neighborhood.

For each point f_j in dataset F, the neighborhood of a point f_i within a given radius ε is called the ε neighborhood of f_i, denoted as

$$F_\varepsilon(f_i) = \{f_j \in F | \mathrm{dist}(f_i, f_j) \leqslant \varepsilon\}, \tag{1}$$

TABLE I. Steps of the DBSCAN algorithm.

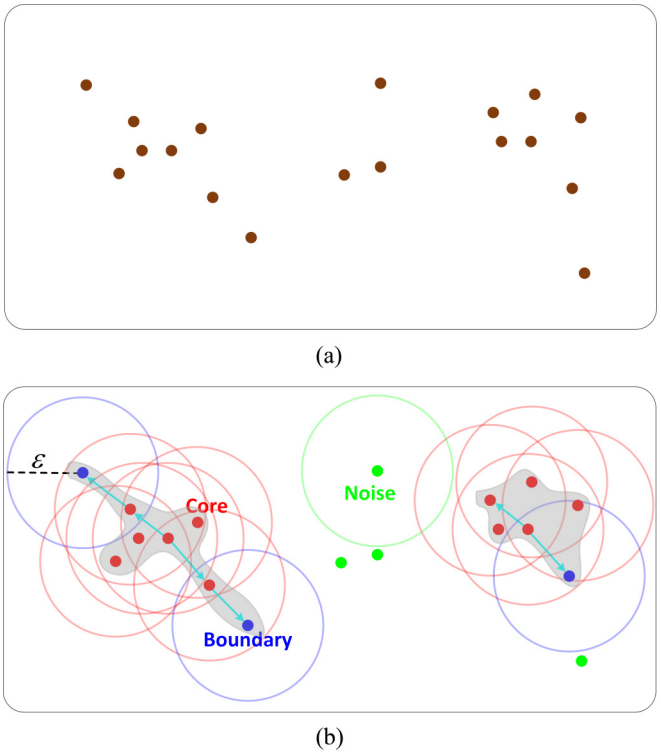
Step 1	Visit an unvisited point f_i in F and find out $F_\varepsilon(f_i)$.
Step 2	If $\lvert F_\varepsilon(f_i)\rvert < P_{\min}$, f_i is temporarily marked as a boundary point. If $\lvert F_\varepsilon(f_i)\rvert \geqslant P_{\min}$, construct a collection G, put $F_\varepsilon(f_i)$ into G, and mark f_i as visited.
Step 3	Process all the unvisited points in G using Step 1 and Step 2. If all the points in G are marked as visited, then G is a cluster obtained by DBSCAN clustering.
Step 4	Check all the points in F. If there are unvisited points, turn to Step 1.
Step 5	Traverse all the boundary points, if any boundary point does not belong to any cluster, mark it as noise.

where $\mathrm{dist}(f_i, f_j)$ is the Euclidean distance between f_i and f_j, which can be expressed as

$$\mathrm{dist}(f_i, f_j) = \sqrt{\sum_{l=0}^{\dim(f_i)} [f_i(l) - f_j(l)]^2}. \tag{2}$$

Definition 2. Core point and boundary point.

If the number of the data points within the ε neighborhood of f_i is no less than $P_{\min}$, that is $\lvert F_\varepsilon(f_i)\rvert \geqslant P_{\min}$, f_i can be called a core point. If a point f_j is in the ε neighborhood of



(a)

(b)

FIG. 1. An example for illustrating DBSCAN algorithm. (a) Scattered points are randomly distributed in feature space. (b) Cluster result of DBSCAN with $P_{\min} = 4$ and a suitable ε. The blue dots are boundary points, the red dots are core points, and the green dots are noise.

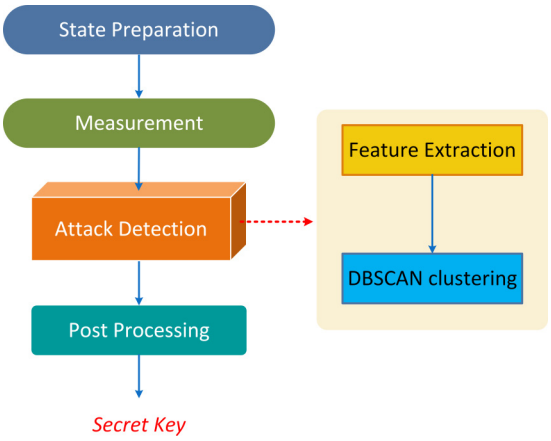


FIG. 2. Process of DADS.

a core point f_i and f_j is not a core point, it can be called a boundary point.

Definition 3. Noise.

If a point f_i is neither a core point nor a boundary point, it can be called noise. Note that the data points f_i and f_j in Definitions 1–3 do not represent the specific data in our experimental dataset. Specifically, input data are five-dimensional feature vectors in our experiment, which are discussed in Sec. II B. After presenting above definitions, the specific steps of DBSCAN can be seen in Table I. As shown in Fig. 1(a), some scattered points are randomly distributed in feature space. Setting $P_{\min} = 4$ with a suitable ε, the clustering result of DBSCAN is shown in Fig. 1(b) where core points and their ε neighborhoods are recursively added until there is no new core point can be found in their ε neighborhoods. These points are therefore divided into two clusters. Note that the selection of ε is empirical, to simply clarify the principle of DBSCAN, here we determine the value of ε by drawing a k-distance graph, which can be used to sort out more accurate collections [38].

B. Process of DADS

The process of the proposed DADS is shown in Fig. 2, and it includes four parts as follows.

1. State preparation

Alice prepares a train of coherent states $|X_a + iP_a\rangle$ where the quadrature values X_a and P_a obey a bivariate Gaussian distribution with variance $V_a N_0$. N_0 is the shot-noise variance which corresponds to the variance of the homodyne detector output when the input signals are vacuum states. Then the prepared states are sent to Bob with classical LO pulses using time and polarization multiplexing.

2. Measurement

This part is detailed depicted in Fig. 3. The pulses sent by Alice are first demultiplexed into signal pulses and LO pulses with a PBS. Then an AM is set on the signal path to randomly perform the maximum attenuation ($r = 0.01$) with a probability of 10% for real-time measurement of shot-noise N_0. A 10:90 beam splitter divides LO pulses into two parts, one part of LO pulses are interfered with the signal pulses after passing a PM to obtain the quadrature values of the signal states. Another part of LO pulses are fed in a PIN photodiode to transform the light signal into electric signal so that we can measure the average power P of LO pulses, generate the clock for homodyne detection, and count the number of pulses per unit time n_p. Note that n_p needs to be measured by oversampling as the number of peaks that occur per unit time can be determined when the sampling frequency is much higher than the pulse repetition frequency [39].

3. Attack detection

This part includes two phases, namely, feature extraction and DBSCAN clustering. After measurement, Alice and Bob obtain two strings of correlated data $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, where X represents the quadrature values modulated by Alice and Y represents the quadrature values measured by Bob. The mean values $\bar{x}$, $\bar{y}$, and variances V_x, V_y of X and Y can be described as

$$\bar{x} = 0, \qquad V_x = V_a N_0, \tag{3}$$

$$\bar{y} = 0, \qquad V_y = \eta T V_a N_0 + N_0 + \eta T \xi + V_{\mathrm{el}}, \tag{4}$$

where T and η represent the transmittance of the quantum channel and the efficiency of the homodyne detector, respectively. $V_{\mathrm{el}} = v_{\mathrm{el}} N_0$ is the electronic noise of

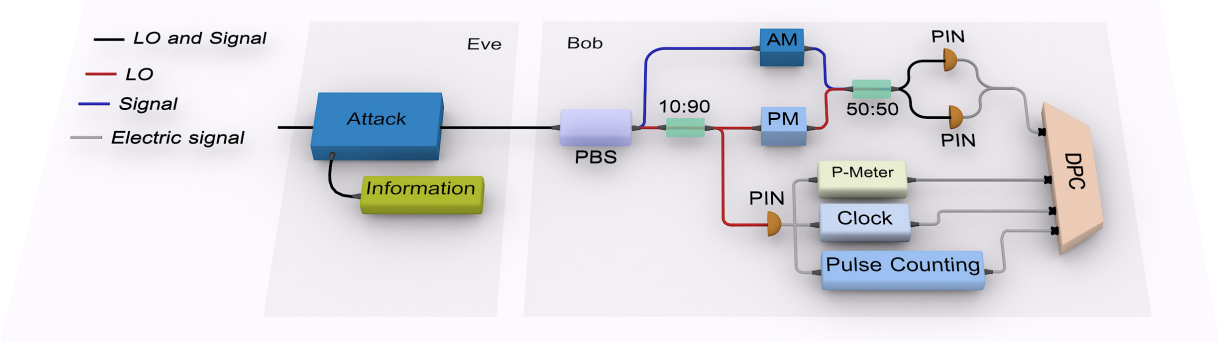


FIG. 3. Schematic of DADS's measurement. PBS: Polarization beam splitter. AM: Amplitude modulator. PM: Phase modulator. PIN: PIN photodiode. P meter: Power meter. Clock: Clock circuit used to generate the clock signal. DPC: Data processing center used for analog signal sampling, attack detection, and raw key distillation.

TABLE II. Impacts of different quantum attacks on measurable features. The symbol (√) under the features indicates that the corresponding feature will be changed by the corresponding attack.

Features	$\bar{y}$	V_y	P	N_0	n_p
LO intensity attack [25]		√	√	√	
Calibration attack [26]		√		√	
Saturation attack [23]	√	√			
Hybrid attack 1 [22]		√	√		√
Hybrid attack 2 [24]	√	√			

detector, and $\xi = \epsilon N_0$ is the excess noise. Supposing that Alice sends N pulses to Bob in total, and these pulses are divided into $d = N/S$ blocks. We can extract a feature vector $f_i = \{\bar{y}, V_y, P, N_0, n_p\}$ for each block so that a feature vector group $F = \{f_1, f_2, \ldots, f_d\}$ can be obtained. The above features may largely be affected by different types of quantum attacks shown in Table II, hence, DBSCAN can be, subsequently, applied to generate several different clusters. The attacked clusters can be easily recognized by comparing with the normal cluster obtained from unattacked data, and they should be discarded. Note that this comparison differs from general classification algorithms; the normal cluster for comparison does not need to be generated in real time in our scheme.

4. Post processing

This part is similar to the postprocessing of conventional CVQKD [1,2]. After discarding the attacked data detected by executing the attack detection part, Alice discloses a part of her data to Bob, who compares them with his data so as to estimate the channel parameters T and ϵ. Subsequently, an error correction algorithm is applied to the remaining data to convert the correlated Gaussian-distributed continuous data into identical discrete data, but these data are only partially secret. Finally, a privacy amplification algorithm is performed based on the hash function, so as to extract the final keys that are entirely unknown to Eve.

III. PERFORMANCE ANALYSIS FOR DADS

In this section, several machine learning metrics are first introduced, then the two key parameters ε and $P_{\min}$ of DBSCAN are discussed and determined. The exhaustive performance analysis of DADS against both known and unknown quantum attacks is, subsequently, presented.

In order to effectively evaluate the performance of the proposed DADS, a test dataset $Y_{\text{test}} = \{y_{\text{normal}}, y_{\text{loia}}, y_{\text{cal}}, y_{\text{sat}}, y_{\text{hyb1}}, y_{\text{hyb2}}\}$ is prepared, details concerning this dataset can be found in Appendix A. Then, the feature vector group $F_{\text{test}} = \{f_1, f_2, \ldots, f_d\}$ is obtained by feature extraction from the original test dataset Y_{test}. After performing DBSCAN clustering, a set of clusters $C = \{C_1, C_2, \ldots, C_{N_1}\}$ can be obtained. For comparison, another dataset of normal data y'_{normal} and its corresponding feature vector group F_{normal} is also prepared. As a result, the attacked clusters can be determined by comparing the clusters of test data F_{test} with the clusters of normal data F_{normal}.

A. Machine learning metrics

DADS takes advantage of DBSCAN to recognize practical quantum attacks so that the traditional information theory-based metrics used in GG02 [1] are not enough to comprehensively estimate its performance. Hence, several machine learning metrics need to be introduced. In general, the metrics of clustering can be divided into external metrics and internal metrics. External metrics usually evaluate the similarity between clustering results and ground truth, which refers to the labels indicating the type of quantum attacks. Different from external metrics, internal metrics are used to directly evaluate the clustering results without ground truth. Here we select two external metrics i.e., the Jaccard coefficient (M_{JC}) and the Folkes and Mallows index (M_{FMI}), and one internal metric, i.e., silhouette coefficient (M_{SC}) to evaluate the performance of DADS in terms of clustering [40]. Specifically, M_{JC} and M_{FMI} are used to evaluate the similarity between the clustering results and the ground truth [41,42], M_{SC} is used to evaluate the compactness within a cluster and the separation between clusters [43]. Details concerning these metrics are presented below.

For clusters $C = \{C_1, C_2, \ldots, C_{N_1}\}$ generated by DBSCAN and clusters $C^* = \{C_1^*, C_2^*, \ldots, C_{N_2}^*\}$ given by the ground truth, we define λ and λ^* as the labels of C and C^* that mark the clusters to which each feature vector belongs to. Based on these labels, M_{JC} and M_{FMI} can be calculated by

$$M_{\text{JC}} = \frac{a}{a+b+c}, \tag{5}$$

$$M_{\text{FMI}} = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}, \tag{6}$$

where

$$a = |SS|, SS = \{(f_i, f_j)|\lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}, \tag{7}$$

$$b = |SD|, SD = \{(f_i, f_j)|\lambda_i = \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}, \tag{8}$$

$$c = |DS|, DS = \{(f_i, f_j)|\lambda_i \neq \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}, \tag{9}$$

where a denotes the number of the sample pairs that belong to the same cluster in ground truth and in the clustering results, b denotes the number of the sample pairs that belong to different clusters in the ground truth and belong to the same cluster in the clustering results, and c denotes the number of the sample pairs that belong to the same cluster in ground truth and belong to different clusters in the clustering results. In general, the values of M_{JC} and M_{FMI} close to 1 indicate high similarity between the clustering results and the ground truth.

For the N_1 clusters generated by DBSCAN, we can calculate the average distance between each point f_i and other points in cluster C_m as

$$a(f_i) = \frac{\sum_{j=0}^{|C_m|} \text{dist}(f_i, f_j)}{|C_m| - 1}, \tag{10}$$

where $|C_m|$ represents the number of the points in C_m and $\dim(f_i)$ represents the dimension of vector f_i, which equals 5 in this paper. Similarly, the average distance between f_i and

TABLE III. Confusion matrix.

	Detected as attacked	Detected as unattacked
Attacked	D_{TP}	D_{FN}
Unattacked	D_{FP}	D_{TN}

points in other clusters is given by

$$b(f_i) = \frac{\sum_{j=0}^{|C|-|C_m|} \mathrm{dist}(f_i, f_j)}{|C| - |C_m|}, \tag{11}$$

where $|C|$ represents the number of the points in C. Then the silhouette coefficient of f_i can be expressed as

$$s(f_i) = \frac{b(f_i) - a(f_i)}{\max[a(f_i), b(f_i)]}. \tag{12}$$

For clustering results C, the total silhouette coefficient is the average silhouette coefficient of all the points, which is expressed as

$$M_{\mathrm{SC}} = \frac{\sum_{i=0}^{|C|} s(f_i)}{|C|}. \tag{13}$$

The clusters generated by a well-clustering algorithm have high intracluster similarity and low intercluster similarity. According to the calculation of M_{SC}, the value of M_{SC} close to 1 indicates well-clustering results.

In addition, we also investigate the performance of DADS in terms of several widely used machine learning metrics, such as precision (M_{Prec}), recall (M_{Rec}), false positive rate (M_{FPR}), and false negative rate (M_{FNR}), their formulas are given by

$$M_{\mathrm{Prec}} = \frac{D_{\mathrm{TP}}}{D_{\mathrm{TP}} + D_{\mathrm{FP}}}, \tag{14}$$

$$M_{\mathrm{Rec}} = \frac{D_{\mathrm{TP}}}{D_{\mathrm{TP}} + D_{\mathrm{FN}}}, \tag{15}$$

$$M_{\mathrm{FPR}} = \frac{D_{\mathrm{FP}}}{D_{\mathrm{TN}} + D_{\mathrm{FP}}}, \tag{16}$$

$$M_{\mathrm{FNR}} = \frac{D_{\mathrm{FN}}}{D_{\mathrm{TP}} + D_{\mathrm{FN}}}, \tag{17}$$

where D_{TP}, D_{FP}, D_{FN} and D_{TN} are defined according to the confusion matrix described in Table III. True positive (D_{TP}) indicates the counts that attacked data are detected as attacked data, false positive (D_{FP}) indicates the counts that normal data are detected as attacked data, false negative (D_{FN}) indicates the counts that attacked data are detected as normal data, and true negative (D_{TN}) indicates the counts that normal data are detected as normal data.

B. Performance of DADS

Before analyzing the performance of DADS in terms of machine learning metrics, the two key parameters ε and P_{min} have to be determined. In general, the k-distance graph can be used to determine an optimal value range of ε. The approach of drawing a k-distance graph is to find the distance between each point and its kth nearest point, then sort all points based on this distance, and finally plot sorted points against this distance. The ordinate value of the point with the fastest slope change on the curve is a good choice for ε, which can make the DBSCAN clusters more compact [38]. Figure 4 shows a 50-distance graph, and we find that the slope of the curve changes rapidly when ε varies from 0.003 to 0.007, i.e., the red solid part of the curve. Therefore, the optimal value of ε can be selected from this range. On the other hand, parameter P_{min} is generally set to twice the dimension of the feature vector, hence, we set $P_{\mathrm{min}} = 10$ in our case.

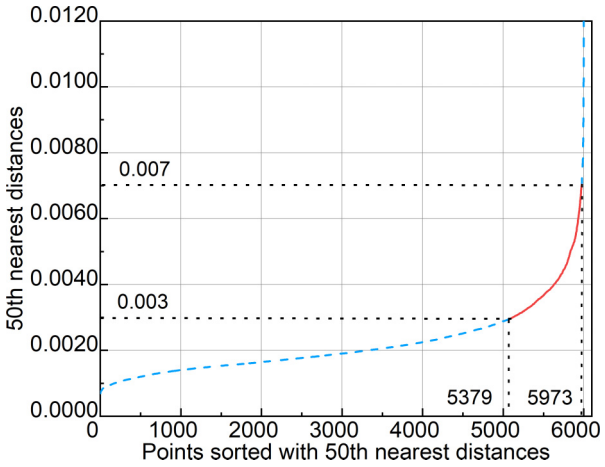


FIG. 4. A 50-distance graph. The red solid line shows the range with the fastest slope change.

After determining the values of P_{min} and the approximate range of ε, the performance of DADS in terms of machine learning metrics can be discussed. As shown in Fig. 5, we find that both M_{JC} and M_{FMI} are close to 1 when $\varepsilon \in [0.006, 0.007]$. It suggests that the clustering results are basically consistent with the ground truth in this range. We also note that the values of both metrics decrease significantly when $\varepsilon < 0.006$ (especially when $\varepsilon < 0.004$), this is because some data that belong to same class are clustered into different clusters when ε is too small, reducing the similarity between

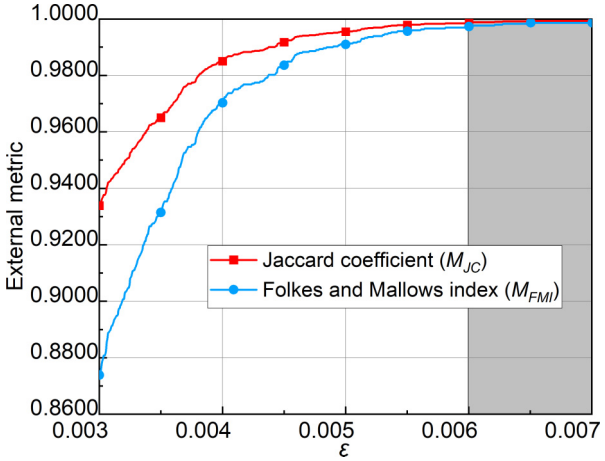


FIG. 5. M_{JC} and M_{FMI} of the clusters versus different values of ε. Both M_{JC} and M_{FMI} are close to 1 when $\varepsilon \in [0.006, 0.007]$.

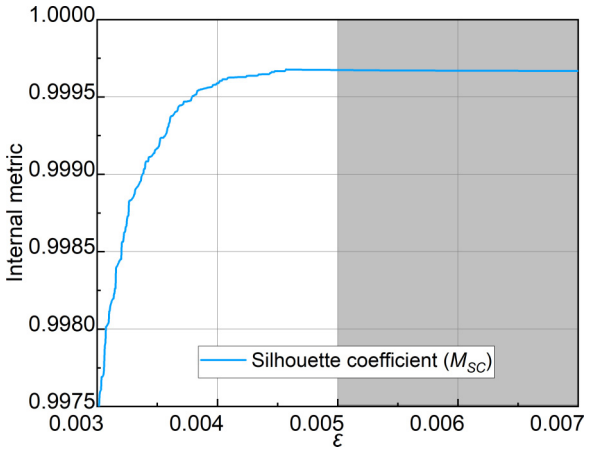


FIG. 6. M_{SC} of the clusters versus different values of ε, M_{SC} is close to 1 when $\varepsilon \in [0.005, 0.007]$.

the ground truth and clustering results. Figure 6 shows the performance of M_{SC} versus different values of ε. We find that M_{SC} is close to 1 when $\varepsilon \in [0.005, 0.007]$, which indicates that the clusters obtained by DBSCAN have high intracluster similarity and low intercluster similarity. Note that the value of M_{SC} decreases when $\varepsilon < 0.005$, the reason is that some data that belong to same class are clustered into different clusters when ε is too small, leading to the increased intercluster similarity. Figure 7 depicts the performance of DADS in terms of M_{Prec}, M_{Rec}, M_{FPR} and M_{FNR}. We can tell that M_{Prec} and M_{Rec} approach 1 when $\varepsilon \in [0.006, 0.007]$, and M_{FPR} and M_{FNR} approach 0 under the same range of ε. It suggests that both high M_{Prec} and M_{Rec} and low M_{FPR} and M_{FNR} can be achieved with the range of $\varepsilon \in [0.006, 0.007]$.

By comprehensively investigating the optimal ranges of ε in terms of above all machine learning metrics, we find that the optimal range of $\varepsilon \in [0.006, 0.007]$ is overlapped. It suggests that the clustering results are highly similar to the ground truth with this range of ε so that the clusters formed by normal data and the clusters formed by attacked data can be highly differentiated, which are beneficial for distinguishing the attacked signal and the normal signal. Therefore, we can

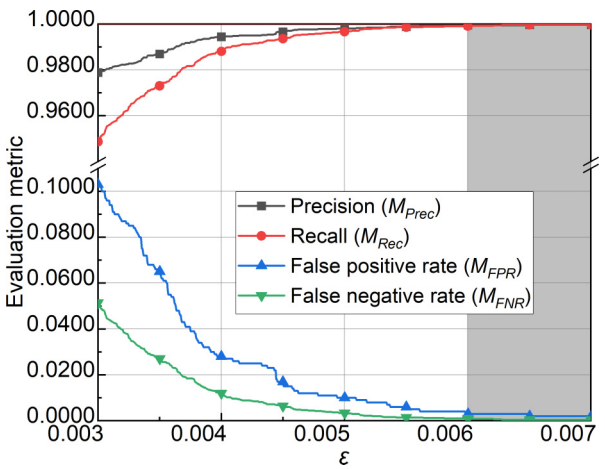


FIG. 7. M_{Prec}, M_{Rec}, M_{FPR} and M_{FNR} of DADS against known attacks as a function of ε.

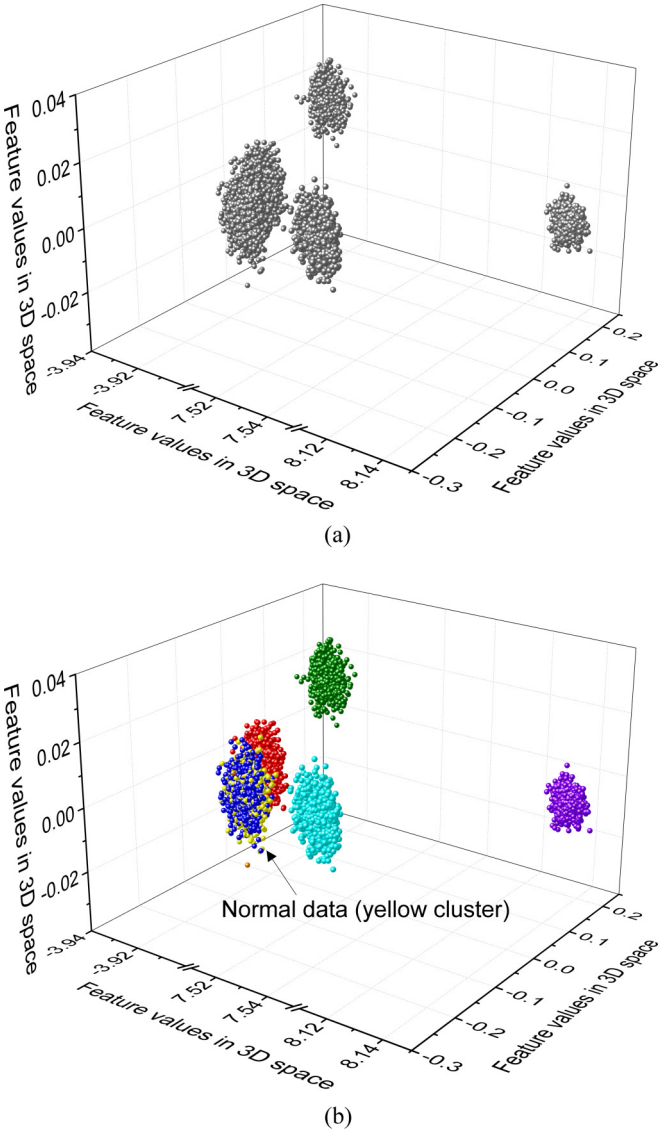


FIG. 8. Three-dimensional spatial distribution of F_{test} before or after DBSCAN clustering

select an identical value, e.g., $\varepsilon = 0.007$, of ε that belongs to this range for the follow-up simulation experiment.

To intuitively interpret the effect of DBSCAN clustering, we map the testing data F_{test} to a three-dimensional space using principal component analysis technique [44], though this mapping operation is actually not a necessary step for DADS. As can be seen in Fig. 8(a), the data points of F_{test} are spatially distributed without labels (uncolored) so that normal data and attacked data can hardly be distinguished. Figure 8(b) shows the clustering result of F_{test} after DBSCAN in which each cluster is labeled with color. Attacked data, therefore, can be detected by comparing the obtained clusters with the cluster of F_{normal}. Note that DADS does not need to know the specific types of quantum attacks as the abnormal data will be directly discarded once the clustering result is different to the cluster of F_{normal}. Obviously, both time and space complexities of the proposed DADS are lower than the existing classification algorithm-based attack detection schemes, thereby saving lots of computational resources.

TABLE IV. Impacts of unknown quantum attacks on measurable features. The symbol (√) under the features indicates that the corresponding feature will be modified by the corresponding attack.

Features	$\bar{y}$	V_y	P	N_0	n_p
Unknown attack 1	√				√
Unknown attack 2	√		√		√
Unknown attack 3			√	√	√
Unknown attack 4	√			√	
Unknown attack 5			√	√	√

C. DADS against unknown quantum attacks

Until now, we have investigated the performance of DADS against known quantum attacks. However, it is impossible for the legitimate parties to foreknow all potential quantum attacks especially in the practical CVQKD system. To guarantee the practical security of CVQKD, the performance of DADS against unknown quantum attacks also has to be taken into account. Table IV shows five types of unknown quantum attacks, each of them randomly affects several features of the normal data. For example, features $\bar{y}$ and n_p will be modified by unknown attack 1, whereas features $\bar{y}$ and N_0 will be modified by unknown attack 4. Note that all unknown attacks listed in this table are randomly generated without considering their detailed implementations, this does make sense as they cannot be called unknown attacks if we explicitly understand their principles of attack.

Similarly, we also simulate 10^3 feature vectors for each unknown quantum attack. The total number of unknown attacks' feature vectors is, therefore, 5000. Figure 9 shows the performance of DADS against unknown quantum attacks in terms of M_{Prec}, M_{Rec}, M_{FPR} and M_{FNR}. We find that both Prec and Rec are close to 1 and both FPR and FNR are close to 0 when $\varepsilon > 0.006$, which is similar to the trends of the performance of DADS against known quantum attacks shown in Fig. 7. It suggests that our proposed DADS not only has the ability to detect most of known attacks, but also can identify various unknown attacks, thereby improving practical security of the CVQKD system.

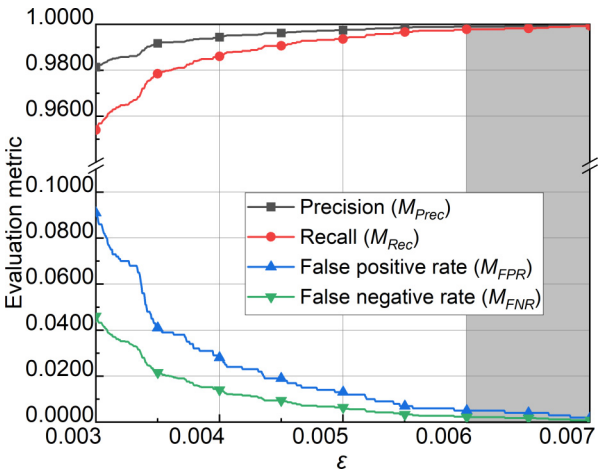


FIG. 9. M_{Prec}, M_{Rec}, M_{FPR} and M_{FNR} of DADS against unknown attacks as a function of ε.

IV. SECURITY ANALYSIS FOR DADS

In this section, we detail the security analysis for DADS in both the asymptotic limit and the finite-size regime.

In the case of the asymptotic limit, the secret key rate without any defense strategy can be expressed as

$$K_{\mathrm{asym}} = \beta I(A{:}B) - \chi_{BE}, \tag{18}$$

where β is the reverse reconciliation efficiency, $I(A{:}B)$ is Shannon mutual information between Alice and Bob, and χ_{BE} is the Holevo quantity for Eve's maximum accessible information. Detailed derivation of $I(A{:}B)$ and χ_{BE} can be seen in Appendix B. However, Eq. (18) does not consider the practical quantum attacks and the accuracy of DADS, it, therefore, has to be amended. After considering the above factors, the revised asymptotic secret key rate can be written as

$$\begin{aligned} K_{\mathrm{asym}}^{\mathrm{rev}} &= (R-\kappa)[\beta I(A{:}B) - \chi_{BE}] \\ &\quad + \kappa(1 - M_{\mathrm{Rec}})[\beta I(A{:}B) - \chi_{BE}] \\ &= (R - \kappa\, M_{\mathrm{Rec}})[\beta I(A{:}B) - \chi_{BE}], \end{aligned} \tag{19}$$

where R represents the proportion of data that are used for final key distillation and κ represents the proportion of pulses under practical quantum attacks. Here we set $R = 0.9$ because 10% of the pulses are used for measuring shot noise as described in Sec. II. κ is set to 0.1, which indicates that another 10% of the pulses are under practical quantum attacks. M_{Rec} is defined in Eq. (15), which indicates the proportion of successfully detected as attacked data in all attacked data. As the revised secret key rate is obtained by discarding the detected attacked data, the pulses with M_{Rec} proportion of the attacked pulses can be detected so that the secret keys generated from this part of pulses have to be discarded. However, the pulses with 1-M_{Rec} proportion of the attacked pulses can not be detected due to the imperfection of algorithm, the secret keys generated from them are incorrectly considered to be secure, therefore, they will be calculated into the revised secret key rate. Obviously, all attacked data can be detected when $M_{\mathrm{Rec}} = 1$. That is to say, the secret keys generated from all attacked pulses can be totally discarded. Therefore, the real secret key rate, which excludes all practical quantum attacks, can be written as

$$K_{\mathrm{asym}}^{\mathrm{real}} = (R-\kappa)[\beta I(A{:}B) - \chi_{BE}]. \tag{20}$$

Figure 10 shows revised and real secret key rates as a function of M_{Rec}. It can be found that the asymptotic secret key rate is largely overestimated when M_{Rec} is close to 0, this is because conventional CVQKD system cannot be aware of the existence of practical quantum attacks without adopting attack detection strategy. The leaked information caused by practical quantum attacks is, therefore, incorrectly deemed to be normal. It also reveals that this overestimation can be gradually eliminated by improving the value of M_{Rec}, which indicates that our proposed DADS has the ability to amend the overestimated secret key rate due to its high M_{Rec} performance.

To figure out how the parameters ε and P_{min} determine the security parameters of DADS, we further plot the asymptotic secret key rate as a function of ε and P_{min}. As shown in Fig. 11,

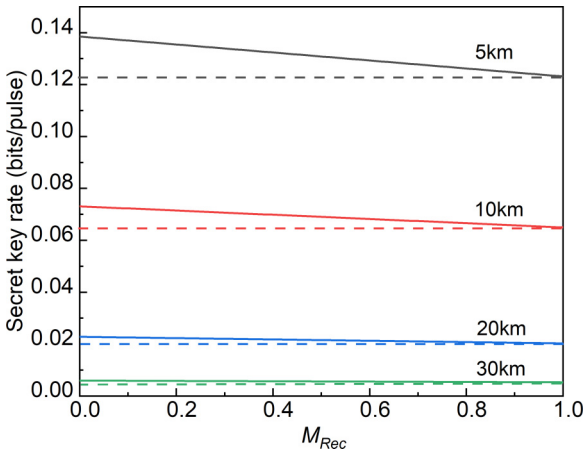


FIG. 10. Asymptotic secret key rate as a function of M_{Rec}. Solid lines denote $K_{\mathrm{asym}}^{\mathrm{rev}}$, and dashed lines denote $K_{\mathrm{asym}}^{\mathrm{real}}$. From top to bottom, different colors represent transmission distances 5 km (black), 10 km (red), 20 km (blue), and 30 km (green), respectively.

we find that the secret key rate increases as ε decreases, and P_{min} increases. This is because some original core points no longer belong to core point as their numbers of points within ε neighborhood become less than P_{min}, resulting in less core points. In this trend, more attacked points will be marked as noise so that they no longer belong to the attacked clusters, decreasing the value of M_{Rec}. Therefore, the secret key rate will be overestimated according to Eq. (19).

In addition, the performance of practical CVQKD system will be jeopardized due to the finite length of data exchanged by legitimate users. Therefore, the finite-size effect has to be taken into account. In this case, the revised secret key rate can be expressed as

$$K_{\mathrm{fini}}^{\mathrm{rev}} = \left(\frac{n^{\mathrm{rev}}}{N^{\mathrm{rev}}} - \kappa\, M_{\mathrm{Rec}}\right)[\beta I(A{:}B) - S(B{:}E)^{r_{\mathrm{PE}}} - \triangle(n)]. \tag{21}$$

See Appendix C for detailed derivation about this equation.

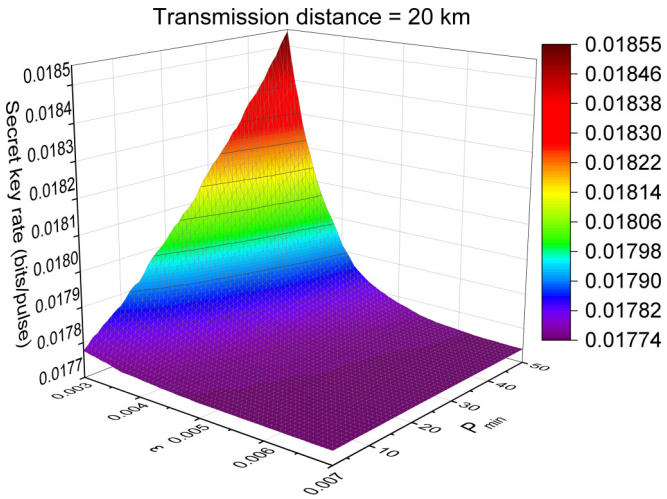


FIG. 11. Asymptotic secret key rate of DADS as a function of $\varepsilon \in [0.003, 0.007]$ and $P_{\mathrm{min}} \in [2, 50]$.

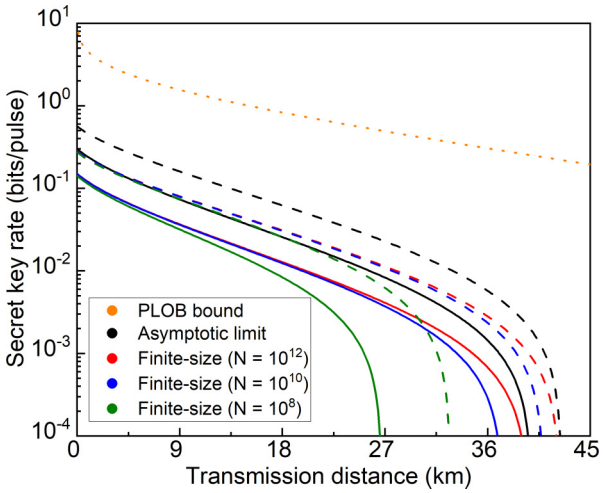


FIG. 12. Secret key rates as a function of transmission distance. Orange dotted line denotes the Piradola-Laurenza-Ottaviani-Banchi (PLOB) bound [45], dashed lines denote the performance of conventional CVQKD without any attack detection strategy, and solid lines denote the performance of the proposed DADS. From top to bottom, different colors (except for orange) represent data block lengths: asymptotic (black), 10^{12} (red), 10^{10} (blue), and 10^{8} (green).

Figure 12 shows the performance comparison between the conventional CVQKD and the proposed DADS. Black lines denote asymptotic secret key rates, red, blue, and green lines denote finite-size secret key rates with data block lengths is 10^{12}, 10^{10} and 10^{8}. It can be found that the performance of conventional CVQKD superficially outperforms our proposed DADS in the cases of both asymptotic limit and finite-size regime. In fact, there are several reasons that jointly lead to this result: First of all, as we mentioned above, conventional CVQKD cannot be aware of the existence of practical quantum attacks without adopting attack detection strategy. As a result, the attacked data will be regarded as normal data, and they will be used for generating secret keys, resulting in the overestimation of secret key rate. Second, some parts of pulses have to be sacrificed for measuring shot noise, which reduces the pulses that can be used for generating final secret keys. Third, extra loss will be introduced by the insertion of AM in the signal path, which will reduce the efficiency of Bob's measurement. The above reasons indicate that the performance of conventional CVQKD shown in Fig. 12 is actually imprecise when considering practical quantum attacks, whereas a tighter secure bound of the practical CVQKD system can be obtained by our proposed DADS.

V. CONCLUSION

In this paper, we have proposed a general machine learning-based defense strategy against practical quantum attacks, called DADS. In particular, we constructed a set of features and established a DBSCAN clustering model to detect different practical quantum attacks. We introduced several machine learning-based metrics to evaluate the performance of DADS. The results showed that the proposed DADS not only has the ability to detect most of known attacks, but also can identify various unknown attacks. We also analyzed the

security of DADS in the cases of both the asymptotic limit and the finite-size regime, and the results showed that the overestimated secret key rate caused by ignoring practical quantum attacks can be amended with the help of DADS, thereby a reasonable tighter secure bound of practical CVQKD system can be finally obtained. Moreover, DADS is beneficial for real-time attack detection as it does not require a time-consuming training process when compared to other classification-based detection scheme.

ACKNOWLEDGMENTS

We thank the anonymous referees for their constructive suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62101180), Hunan Provincial Natural Science Foundation of China (Grant No. 2022JJ30163), and the Open Research Fund Program of the State Key Laboratory of High Performance Computing, National University of Defense Technology (Grant No. 202101-25).

APPENDIX A: DATASET PREPARATION

Datasets are generated with the following values of parameters under practical consideration: Modulation variance $V_A = 10$, efficiency of homodyne detector $\eta = 0.6$, excess noise of the channel $\xi = 0.1N_0$, electrical noise of detector $V_{\mathrm{el}} = 0.01N_0$, and transmittance of channel $T = 10^{-\alpha L/10}$, where the transmission distance L is set to 30 km, and the optical loss coefficient $\alpha = 0.2$ dB/km. A max attenuation coefficient $r_1 = 0.001$ and a nonattenuation coefficient $r_2 = 1$ are set on the signal path. The mean value of Y and its variance without quantum attacks can be expressed as

$$\begin{aligned} \bar{y} &= 0, \\ V_i &= r_i \eta T (V_A N_0 + \xi) + N_0 + V_{\mathrm{el}}. \end{aligned} \tag{A1}$$

LO power is set to 10^7 photons per pulse with 1% fluctuation, and the variance of shot noise is set to 0.4.

(a) LO intensity attack [25]: In conventional security analysis, LO is not taken into consideration, and its intensity is assumed to remain unchanged. However, in practical implementation, Eve could intercept not only the signal beam, but also the LO. By controlling the intensity of LO with attenuation coefficient ν ($0 < \nu < 1$), Eve can reduce channel excess noise estimation, so as to successfully hide her attack. Hence, the average power of LO can be attenuated to νP. Assuming that the attenuation of each LO pulse is identical, the excess noise introduced by collective attack with monitoring is

$$\begin{aligned} \xi_{\mathrm{Col}} &= \frac{(1-\eta T)(V_{\mathrm{Eve}} - 1)}{\eta T} N_0 \\ &= \frac{(1-\nu)}{\nu \eta T} N_0, \end{aligned} \tag{A2}$$

where $V_{\mathrm{Eve}} = (1 - \nu\eta T)/(\nu - \nu\eta T)$ is the variance of Eve's EPR states. The variance of the measurement at Bob side can be written as

$$V_i^{\mathrm{LOIA}} = \nu[r_i \eta T (V_A N_0 + \xi + \xi_{\mathrm{Col}}) + N_0 + V_{\mathrm{el}}]. \tag{A3}$$

Note that the variance of shot-noise N_0^{LOIA} is reduced to νN_0 due to the intensity attenuator. Therefore, we find that P, N_0, and V_y can be affected under LO intensity attack.

(b) Calibration attack [26]: Eve manipulates the classical LO pulses during the operation of CVQKD in order to modify the clock pulses used at the detection stage. This allows the eavesdropper to bias the shot-noise estimation usually performed using a calibrated relationship. Specifically, Eve performs partial intercept-resend (PIR) attacks for intercepting a part μ of signal pulses and modifying the shape of this part LO pulses. This can scale the excess noise estimation since Bob's shot-noise estimation remains unchanged. The excess noise introduced by calibration attack can be expressed as

$$\frac{\xi_{\mathrm{calib}}}{N_0} = \frac{N_0^{\mathrm{calib}}}{N_0}\left[\frac{\xi + \xi_{\mathrm{PIR}}}{N_0^{\mathrm{calib}}} + \frac{1}{\eta T}\left(1 - \frac{N_0}{N_0^{\mathrm{calib}}}\right)\right], \tag{A4}$$

where $\xi_{\mathrm{PIR}} = 2\mu N_0$ is noise introduced by PIR attack. The excess noise estimation can be scaled to zero whereas $N_0/N_0^{\mathrm{calib}} = 1 + 2.1\eta T$. In this condition, $\mu = 1$, and the measurement variance under calibration attack can be written as

$$V_i^{\mathrm{calib}} = r_i \eta T (V_A + \epsilon + 2) N_0^{\mathrm{calib}} + N_0^{\mathrm{calib}} + \upsilon_{\mathrm{el}} N_0^{\mathrm{calib}}. \tag{A5}$$

We find that N_0 and V_y can be affected under calibration attack.

(c) Saturation attack [23]: It is generally assumed in the security proofs of CVQKD that the response of the homodyne detection is linear with respect to the input quadrature. However, for a practical coherent detector, the linearity domain is limited. If the value of the input quadrature exceeds the limit, linearity may not be verified, leading to a saturated behavior. The security evaluation in CVQKD relies solely on the evaluation of second-order moments of the quadrature, whereas the first-order moments (mean value) are not monitored. This leaves Eve the freedom to manipulate the mean value of the quadratures $\bar{y}$. Combining this observation with the existence of a finite domain of linearity for the detection, Eve can actively introduce a large displacement Δ (even sometimes also add an amplification G) on the coherent states received by Bob in order to induce the homodyne detector to operate in its saturated region. This strategy enables Eve to influence Bob's measurement results and to bias parameter estimation. The mean value, variance of measurement, and excess noise estimation under saturation attack can be expressed as

$$\bar{y}^{\mathrm{sat}} = r_i\left(\alpha - \sqrt{\frac{V_i'}{2\pi}}B - \frac{(\alpha - \Delta)}{2} - \frac{(\alpha - \Delta)}{2}A\right), \tag{A6}$$

$$\begin{aligned} V_i^{\mathrm{sat}} = {} & V_i'\left(\frac{1+A}{2} - \frac{B^2}{2\pi}\right) - (\alpha - \Delta)\sqrt{\frac{V_i'}{2\pi}}AB \\ & + \frac{(\alpha - \Delta)^2}{4}(1 - A^2), \end{aligned} \tag{A7}$$

$$\begin{aligned} \frac{\xi_i^{\mathrm{sat}}}{N_0} = {} & \frac{2}{\eta T G (1+A)^2 N_0}\left[V_i'\left(1 + A - \frac{B^2}{\pi}\right)\right. \\ & - 2\sqrt{\frac{2V_i'}{\pi}}(\alpha - \Delta)AB \\ & \left. + (\alpha - \Delta)^2(1 - A^2) - 4N_0 - 4V_{\mathrm{el}}\right] - \frac{V_A}{N_0}, \end{aligned} \tag{A8}$$

where α is the boundary of the linear range of the homodyne detector and

$$V_i' = r_i\eta T(V_A N_0 + \xi + 2N_0) + N_0 + V_{\mathrm{el}}, \tag{A9}$$

$$A = \frac{2}{\sqrt{\pi}}\int_0^{(\alpha-\Delta)/\sqrt{2V_i'}} e^{-t^2}dt, \tag{A10}$$

$$B = e^{-(\alpha-\Delta)^2/2V_i'}. \tag{A11}$$

We find that $\bar{y}$ and V_y can be affected under saturation attack.

(d) Hybrid attack 1 [22]: This attack consists of two parts, the first part can be considered as the LO intensity attack, and the second part can be deemed as the wavelength attack. In the first part, the pulses are intercepted and represented by Eve. The amplitudes of the signal pulses and LO pulses are $\sqrt{\gamma T}(X_E + iP_E)/2$ and $A_{\mathrm{LO}}/\sqrt{\gamma}$, respectively, where A_{LO} is the amplitude of original LO, and γ is a real number. In the wavelength attack, two beams of additional coherent pulses are sent into the homodyne detector by Eve. The wavelengths of the pulses are different from the wavelength of typical communication 1550 nm. A photocurrent is generated when pulses transmit the 50:50 fused biconical taper beam splitter, which makes the shot noise appear normal. Due to the two additional reprepared coherent pulses, the number of pulses Bob receives per unit time $n_p = 2$. The excess noise is given by

$$N_0^{\mathrm{hyb1}} = \frac{N_0}{\gamma} + (1 - r_1 r_2)K^2 + (35.81 - 35.47 r_1 r_2)K, \tag{A12}$$

$$\frac{\xi^{\mathrm{hyb1}}}{N_0^{\mathrm{hyb1}}} = \left[\frac{(2+\xi)N_0 + (r_1 + r_2 - 2)K^2}{\eta T} + 35.47(r_1 + r_2)K\right], \tag{A13}$$

where K is a parameter related to the intensity and wavelengths I^s, I^{lo}, γ^s, and γ^{lo}. γ^s and γ^{lo} are the wavelengths of the pulses that Eve sends. The variance of Bob's measurement can be expressed as

$$V_i^{\mathrm{hyb1}} = r_i\eta T(V_A N_0 + 2N_0 + \xi) + \frac{N_0}{\gamma} + V_{\mathrm{el}} + (1 - r_i)^2 K^2 + \left(35.81 + 35.47 r_i^2\right)K. \tag{A14}$$

As the shot-noise variance changed by LO intensity attack can be amended by the light photocurrent, P, V_y, and n_p can be affected under hybrid attack 1.

(e) Hybrid attack 2 [24]: Similar to thr saturation attack, and this attack exploits the finite linear domain of homodyne detector (HD) too. Eve sends an extra beam of incoherent strong light on the signal port, which can produce a comparatively stronger photocurrent to saturate the HD. By combining the sending of strong light pulses to Bob with a full intercept-resend attack, Bob's measurement can be severely deviated. The deviation can be expressed as

$$D_{\mathrm{ext}} = \sqrt{\frac{\eta}{I_{\mathrm{lo}}}}(1 - 2T_{\mathrm{ext}})I_{\mathrm{ext}}, \tag{A15}$$

where T_{ext} is the overall transmission of external pulses and is related to the wavelength. I_{ext} is the number of photons per external pulse. D_{ext} is normalized in $\sqrt{N_0}$. The excess noise under this attack can be expressed as

$$\xi_{\mathrm{hyb2}} = \xi + \xi_{\mathrm{IR}} + \xi_{\mathrm{ext}}, \tag{A16}$$

where $\xi_{\mathrm{IR}} = 2N_0$ is the noise introduced by the IR attack and ξ_{ext} is the noise introduced by the external light. Correspondingly, the variance of measurement can be written as

$$V_{\mathrm{hyb2}} = \eta T(V_A + \xi_{\mathrm{hyb2}}) + 1 + V_{\mathrm{el}}. \tag{A17}$$

The same as the saturation attack, $\bar{y}$ and V_y can be affected under hybrid attack 2.

Several parameters have to be determined before preparing a dataset for the above practical quantum attacks. Specifically, the attenuation coefficient ν is set to 0.95 for the intensity attack as Eve can obtain the full secret keys in this case [25]. For the calibration attack, we assume Eve launches a full intercept-resend attack, μ is thereby set to 1. For the saturation attack, α and Δ are set to $20\sqrt{N_0}$ and $19.5\sqrt{N_0}$ for an excellent attack effect [23]. For the hybrid attack 1, Eve chooses $\lambda = 20.9, I_1^{\mathrm{lo}} = 5\times10^5, I_1^s = 5.4\times10^5, I_2^{\mathrm{lo}} = 4.8\times10^5$, and $I_2^s = 4.4\times10^5$, so as to make $N_0^{\mathrm{hyb1}} = N_0$ [22]. For hybrid attack 2, T_{ext} and I_{ext} are set to 0.49 and $1.274\times\ 10^7$ to accurately bias the excess noise estimation [24]. Now the test dataset $Y_{\mathrm{test}} = \{y_{\mathrm{normal}}, y_{\mathrm{LOIA}}, y_{\mathrm{cal}}, y_{\mathrm{sat}}, y_{\mathrm{hyb1}}, y_{\mathrm{hyb2}}\}$ can be obtained with these determined parameters.

Assuming Bob receives 10^8 pulses in total and 10^5 pulses are required to extract a feature vector, thereby 10^3 data can be prepared for each type of dataset in Y_{test}. Note that 90% of data in the dataset of Y_{test} are generated based on $r_i = r_1$ and another 10% are generated based on $r_i = r_2$ because 10% of the data are used for measuring the shot-noise variance.

APPENDIX B: CALCULATION OF THE ASYMPTOTIC SECRET KEY RATE

In the case of the asymptotic limit, Shannon mutual information $I(A{:}B)$ and Holevo bound χ_{BE} can be, respectively, expressed as

$$I(A{:}B) = \log_2\frac{V + \chi_{\mathrm{tot}}}{1 + \chi_{\mathrm{tot}}}, \tag{B1}$$

$$\chi_{BE} = S(\rho_E) - \int d_{m_B}\rho(m_B)S\left(\rho_E^{m_B}\right), \tag{B2}$$

where $V = V_m + 1$ is the channel-added noise referred to the channel input is $\chi_{\mathrm{line}} = 1/T - 1 + \epsilon$, the detection-added noise referred to Bob's input is $\chi_{\mathrm{hom}} = [(1-\eta) + v_{\mathrm{el}}]/\eta$, the total noise referred to the channel input is $\chi_{\mathrm{tot}} = \chi_{\mathrm{line}} + \chi_{\mathrm{hom}}/T$, m_B is the measurement of Bob, S is the von Neumann entropy of the quantum state ρ, $\rho(m_B)$ is the probability density of the measurement, and $\rho_E^{m_B}$ is Eve's state conditional on Bob's measurement. Under Gaussian collective attack, χ_{BE} can be expressed as

$$\chi_{BE} = \sum_{i=1}^{2} G\left(\frac{\lambda_i - 1}{2}\right) - \sum_{i=3}^{5} G\left(\frac{\lambda_i - 1}{2}\right), \tag{B3}$$

where

$$G(x) = (x+1)\log_2(x+1) - x\log_2(x), \tag{B4}$$

and $\lambda_{1,2}$ and $\lambda_{3,4}$ are the symplectic eigenvalues that can be expressed as

$$\lambda_{1,2}^2 = \tfrac{1}{2}\big(\varphi_1 \pm \sqrt{\varphi_1^2 - 4\varphi_2}\big), \tag{B5}$$

$$\lambda_{3,4} = \tfrac{1}{2}\big(\varphi_3 \pm \sqrt{\varphi_3^2 - 4\varphi_4}\big), \tag{B6}$$

where

$$\varphi_1 = V^2 + T^2(V + \chi_{\text{line}})^2 + 2T(1 - V^2), \tag{B7}$$

$$\varphi_2 = T^2(1 + V\chi_{\text{line}}), \tag{B8}$$

$$\varphi_3 = \frac{\varphi_1\chi_{\text{hom}} + V\sqrt{\varphi_2} + T(V + \chi_{\text{line}})}{T(V + \chi_{\text{tot}})}, \tag{B9}$$

$$\varphi_4 = \frac{\sqrt{\varphi_2}V + \varphi_2\chi_{\text{hom}}}{T(V + \chi_{\text{tot}})}. \tag{B10}$$

$\lambda_5 = 1$. Finally, the asymptotic secret key rate of DADS can be calculated with the above equations [26].

APPENDIX C: CALCULATION OF THE FINITE-SIZE SECRET KEY RATE

In the case of the finite-size regime, the secret key rate without any defense strategy can be expressed as [18]

$$K_{\text{finite}} = \frac{n}{N}[\beta I(A{:}B) - S(B{:}E)^{r_{\text{PE}}} - \Delta(n)], \tag{C1}$$

where N and n represent the number of the exchanged signals between Alice and Bob and the number of the signals used for key establishment, respectively. $m = N - n$ indicates the number of the signals used for parameter estimation. r_{PE} denotes the failure probability of parameter estimation. $\Delta(n)$ is related to the security of the privacy amplification, which can be expressed as

$$\Delta(n) = (2\dim\mathcal{H}_y + 3)\sqrt{\frac{\log_2(2/\bar{\mathrm{r}})}{n}} + \frac{2}{n}\log_2(1/r_{\text{PA}}), \tag{C2}$$

where $\bar{r}$ and r_{PA} represent a smoothing parameter and the failure probability of the privacy amplification procedure, respectively. $\mathcal{H}_y$ is the Hilbert space corresponding to the variable y used in the raw key. Since the raw key is bit encoded, we take $\dim\mathcal{H}_y = 2$ for the key rate evaluation parameter. $S(B{:}E)^{r_{\text{PE}}}$ represents the mutual information between Bob and Eve. It is determined by the covariance matrix Γ_{AB} of the bipartite state shared by Alice and Bob. The covariance matrix can minimize the secret key rate in finite size with a probability of $1 - r_{\text{PE}}$, which can be calculated as

$$\Gamma_{AB} = \begin{bmatrix} (V_A + 1)\mathbb{I} & \sqrt{T_{\min}\big(V_A^2 + 2V_A\big)}\sigma_z \\ \sqrt{T_{\min}\big(V_A^2 + 2V_A\big)}\sigma_z & [T_{\min}(V_A + \epsilon_{\max}) + 1]\mathbb{I} \end{bmatrix}, \tag{C3}$$

where the matrices $\mathbb{I} = \text{diag}(1, 1)$ and $\sigma_z = \text{diag}(1, -1)$. $T_{\min}$ and $\epsilon_{\max}$ represent the lower and upper bounds of T and ϵ, respectively, which are given by

$$T_{\min} = \frac{\hat{t}_{\min}^2}{\eta}, \qquad \epsilon_{\max} = \frac{\hat{\sigma}_{\max}^2 - 1 - v_{\text{el}}}{\eta T}, \tag{C4}$$

with

$$\hat{\sigma}_{\max}^2 \approx 1 + \eta T\epsilon + v_{\text{el}} + z_{r_{\text{PE}}/2}\sqrt{\frac{(1 + \eta T\epsilon + v_{\text{el}})\sqrt{2}}{\sqrt{m}}}, \tag{C5}$$

$$\hat{t}_{\min} \approx \sqrt{\eta T} - z_{r_{\text{PE}}/2}\sqrt{\frac{1 + \eta T\epsilon + v_{\text{el}}}{mV_A}}, \tag{C6}$$

where $z_{r_{\text{PE}}/2}$ follows $1 - \text{erf}(z_{r_{\text{PE}}/2}/\sqrt{2})/2 = r_{\text{PE}}/2$. In our simulations, these error probabilities can be set to the optimal value as

$$\bar{r} = r_{\text{PE}} = r_{\text{PA}} = 10^{-10}. \tag{C7}$$

Substituting $T_{\min}$ and $\epsilon_{\max}$ for T and ϵ in Eqs. (B1) and ((B3)), the finite-size secret key rate without any defense strategy can be obtained. However, Eq. (C1) does not consider practical quantum attacks. Similar to the case of the asymptotic limit, the revised finite-size secret key rate, therefore, can be expressed as

$$\begin{aligned} K_{\text{fini}}^{\text{rev}} &= \left(\frac{n^{\text{rev}}}{N^{\text{rev}}} - \kappa\right)[\beta I(A{:}B) - S(B{:}E)^{r_{\text{PE}}} - \Delta(n)] \\ &\quad + \kappa(1 - M_{\text{Rec}})[\beta I(A{:}B) - S(B{:}E)^{r_{\text{PE}}} - \Delta(n)] \\ &= \left(\frac{n^{\text{rev}}}{N^{\text{rev}}} - \kappa M_{\text{Rec}}\right)[\beta I(A{:}B) - S(B{:}E)^{r_{\text{PE}}} - \Delta(n)]. \end{aligned} \tag{C8}$$

Note that both the number of exchanged signals and the number of signals used for key establishment will decrease, i.e., $N^{\text{rev}} = RN$ and $n^{\text{rev}} = Rn$ as 10% pulses are used for measuring shot noise in DADS.

[1] F. Grosshans and P. Grangier, Phys. Rev. Lett. **88**, 057902 (2002).

[2] S. Fossier, E. Diamanti, T. Debuisschert, R. Tualle-Brouri, and P. Grangier, J. Phys. B: At., Mol. Opt. Phys. **42**, 114014 (2009).

[3] P. Jouguet, S. Kunz-Jacques, A. Leverrier, P. Grangier, and E. Diamanti, Nat. Photonics **7**, 378 (2013).

[4] R. Renner and J. I. Cirac, Phys. Rev. Lett. **102**, 110504 (2009).

[5] A. Leverrier, Phys. Rev. Lett. **114**, 070501 (2015).

[6] A. Leverrier, Phys. Rev. Lett. **118**, 200501 (2017).

[7] A. Leverrier and P. Grangier, Phys. Rev. Lett. **102**, 180504 (2009).

[8] Q. Liao, H. Liu, L. Zhu, and Y. Guo, Phys. Rev. A **103**, 032410 (2021).

[9] Q. Liao, G. Xiao, C.-G. Xu, Y. Xu, and Y. Guo, Phys. Rev. A **102**, 032604 (2020).

[10] A. Denys, P. Brown, and A. Leverrier, Quantum **5**, 540 (2021).

[11] C. Lupo and Y. Ouyang, PRX Quantum **3**, 010341 (2022).

[12] J. Lin, T. Upadhyaya, and N. Lütkenhaus, Phys. Rev. X **9**, 041064 (2019).

[13] S. Ghorai, P. Grangier, E. Diamanti, and A. Leverrier, Phys. Rev. X **9**, 021059 (2019).

[14] H.-X. Ma, P. Huang, D.-Y. Bai, T. Wang, S.-Y. Wang, W.-S. Bao, and G.-H. Zeng, Phys. Rev. A **99**, 022322 (2019).

[15] M. Navascués, F. Grosshans, and A. Acín, Phys. Rev. Lett. **97**, 190502 (2006).

[16] R. García-Patrón and N. J. Cerf, Phys. Rev. Lett. **97**, 190503 (2006).

[17] Q. Liao, H. Liu, Y. Gong, Z. Wang, Q. Peng, and Y. Guo, Opt. Express **30**, 3876 (2022).

[18] A. Leverrier, F. Grosshans, and P. Grangier, Phys. Rev. A **81**, 062343 (2010).

[19] A. Leverrier, R. García-Patrón, R. Renner, and N. J. Cerf, Phys. Rev. Lett. **110**, 030502 (2013).

[20] J.-Z. Huang, C. Weedbrook, Z.-Q. Yin, S. Wang, H.-W. Li, W. Chen, G.-C. Guo, and Z.-F. Han, Phys. Rev. A **87**, 062329 (2013).

[21] X.-C. Ma, S.-H. Sun, M.-S. Jiang, and L.-M. Liang, Phys. Rev. A **87**, 052309 (2013).

[22] J.-Z. Huang, S. Kunz-Jacques, P. Jouguet, C. Weedbrook, Z.-Q. Yin, S. Wang, W. Chen, G.-C. Guo, and Z.-F. Han, Phys. Rev. A **89**, 032304 (2014).

[23] H. Qin, R. Kumar, and R. Alléaume, Phys. Rev. A **94**, 012325 (2016).

[24] H. Qin, R. Kumar, V. Makarov, and R. Alléaume, Phys. Rev. A **98**, 012312 (2018).

[25] X.-C. Ma, S.-H. Sun, M.-S. Jiang, and L.-M. Liang, Phys. Rev. A **88**, 022339 (2013).

[26] P. Jouguet, S. Kunz-Jacques, and E. Diamanti, Phys. Rev. A **87**, 062313 (2013).

[27] Q. Liao, G. Xiao, H. Zhong, and Y. Guo, New J. Phys. **22**, 083086 (2020).

[28] Z. Li, H. Zhang, Q. Liao, Y. Mao, and Y. Guo, Phys. Lett. A **419**, 127694 (2021).

[29] J.-Y. Liu, H.-J. Ding, C.-M. Zhang, S.-P. Xie, and Q. Wang, Phys. Rev. Applied **12**, 014059 (2019).

[30] H.-J. Ding, J.-Y. Liu, C.-M. Zhang, and Q. Wang, Quant. Info. Proc. **19**, 60 (2020).

[31] L. Li, P. Huang, T. Wang, and G. Zeng, Phys. Rev. A **103**, 032611 (2021).

[32] W. Liu, P. Huang, J. Peng, J. Fan, and G. Zeng, Phys. Rev. A **97**, 022316 (2018).

[33] Y. Mao, Y. Wang, W. Huang, H. Qin, D. Huang, and Y. Guo, Phys. Rev. A **101**, 062320 (2020).

[34] Y. Mao, W. Huang, H. Zhong, Y. Wang, H. Qin, Y. Guo, and D. Huang, New J. Phys. **22**, 083073 (2020).

[35] D. Deng, J. Phys.: Conf. Ser. **1617**, 012088 (2020).

[36] S. Jebari, A. Smiti, and A. Louati, in *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 2019* (IEEE, Pistaraway, NJ, 2019), pp. 000001–000006

[37] X. Chen, D. Liu, X. Wang, Y. Chen, and S. Cheng, in 2021 2nd International Conference on Big Data and Informatization Education (ICBDIE), 2021 (IEEE, Piscataway, NJ, 2021), pp. 398–401.

[38] B. Mu, M. Dai, and S. Yuan, IOP Conf. Ser.: Mater. Sci. Eng. **715**, 012023 (2020).

[39] P. Huang, J. Huang, T. Wang, H. Li, D. Huang, and G. Zeng, Phys. Rev. A **95**, 052302 (2017).

[40] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. Dougherty, Pattern Recognition **40**, 807 (2007).

[41] S. Lee, W. Jung, S. Kim, and E. T. Kim, in *2019 International Conference on Information and Communication Technology Convergence (ICTC), 2019* (IEEE, Piscataway, NJ, 2019), pp. 178–183.

[42] K. N. Ismail, A. Seman, and K. A. F. Abu Samah, in *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), 2021* (IEEE, Piscataway, NJ, 2021), pp. 229–233.

[43] H. B. Tambunan, D. H. Barus, J. Hartono, A. S. Alam, D. A. Nugraha, and H. H. H. Usman, in *2020 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP), 2020* (IEEE, Piscataway, NJ, 2020), pp. 258–262.

[44] R. He, B.-G. Hu, W.-S. Zheng, and X.-W. Kong, IEEE Trans. Image Process. **20**, 1485 (2011).

[45] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, Nat. Commun. **8**, 15043 (2017).