# Addressing Side-Channel Threats in Quantum Key Distribution via Deep Anomaly Detection

Junxuan Liu,[1] Bingcheng Huang,[1] Jialei Su,[1] Qingquan Peng,[1] and Anqi Huang[1,*]

[1] *College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, People's Republic of China*

Traditional countermeasures against security side channels in quantum key distribution (QKD) systems often suffer from poor compatibility with deployed infrastructure, the risk of introducing new vulnerabilities, and limited applicability to specific types of attacks. In this work, we propose an anomaly detection (AD) model based on one-class machine learning algorithms to address these limitations. By constructing a dataset from the QKD system's operational states, the AD model learns the characteristics of normal behavior under secure conditions. When an attack occurs, the system's state deviates from the learned normal patterns and is identified as anomalous by the model. Experimental results show that the AD model achieves an area under the curve (AUC) exceeding 99%, effectively safeguarding the QKD system's secure operation. Compared to traditional approaches, our model can be deployed with minimal cost in existing QKD networks without requiring additional optical or electrical components, thus avoiding the introduction of new side channels. Furthermore, unlike multi-class machine learning algorithms, our approach does not rely on prior knowledge of specific attack types and is potentially able to detect unknown attacks. These advantages—generality, ease of deployment, low cost, and high accuracy—make our model a practical and effective tool for protecting QKD systems against side-channel threats.

## I. INTRODUCTION

Quantum key distribution (QKD), based on the quantum mechanism, such as quantum no-cloning theorem and Heisenberg's uncertainty principle, enables the information-theoretical security of distributing random symmetric keys [1–4]. Compared to classical cryptographic algorithms that rely on computational complexity, QKD offers significant advantages in terms of security, which have attracted widespread attention. However, in practical implementations, the physical devices used in QKD often deviate from the idealized theoretical models. In recent years, extensive research into these discrepancies has revealed various side channels, demonstrating that an eavesdropper (Eve) can exploit such vulnerabilities to compromise the practical security of QKD systems [5–24].

To address side channels arising from imperfections in physical devices, researchers have explored multiple approaches to protect QKD systems from quantum hacking. One such approach is the development of new QKD protocols. For example, the measurement-device-independent QKD (MDI QKD) protocol eliminates security dependencies on imperfect measurement devices [25]. While MDI QKD offers higher security compared to the decoy-state BB84 protocol [26–28], it also introduces significantly higher implementation complexity. Importantly, since most practical QKD systems still rely on mature prepare-and-measure protocols [29–31] that remain vulnerable to quantum attacks, it is crucial to develop low-complexity countermeasures that can enhance their security without changing existing infrastructures. One such strategy involves adding physical components to the system—a "patching" approach designed to counter specific attacks [19, 32, 33]. While relatively easy to implement, this method may inadvertently introduce new side channels, potentially compromising the overall security of the QKD system [19, 34]. Another strategy for enhancing security without altering existing infrastructures focuses on refining theoretical security models. This involves analyzing Eve's capabilities under specific attacks and estimating the corresponding secure key rates [22, 23, 32, 35–37]. By incorporating the characteristics of known attacks into the theoretical framework, these models aim to quantify potential key leakage. However, such security models are inherently limited: they cannot address unknown threats, and for certain practical attacks—such as the calibration attack [9] or the muted attack [38]—no effective theoretical descriptions currently exist.

These limitations highlight the need for a more general defense capability against a wide range of attack strategies, while still minimizing changes to existing infrastructures. To address this challenge, this work proposes an intelligent countermeasure for QKD systems based on machine learning—an anomaly detection (AD). In particular, we employ Deep Support Vector Data Description (Deep SVDD) [39], a one-class classification algorithm, to achieve real-time monitoring of QKD systems and ensure their operation in a secure environment. Our AD model is trained under unsupervised conditions, which greatly simplifies dataset preparation. For instance, in our experiment, the parameters from the calibration and post-processing stages can be directly extracted during the secure operation of the QKD system to construct the training dataset. We constructed a test set by combining QKD system parameters obtained under both secure conditions and under attacks (specifically, the calibration at-

* angelhuang.hn@gmail.com

tack [9] and the muted attack [38]) in a 1:1 ratio. Experimental results demonstrate that Deep SVDD achieves an area under the receiver operating characteristic curve (AUC) exceeding 99%. This indicates that the AD model can effectively detect anomalies using only QKD system parameters.

Our AD model achieves anomaly detection solely by analyzing parameters generated during system operation, requiring no additional physical hardware. This not only reduces implementation cost, but also avoids introducing new side channels. Thus, the AD model is particularly well-suited for deployment in existing QKD systems, as it imposes no changes to the underlying protocol or hardware. Furthermore, unlike multi-class classification approaches [40], which attempt to distinguish between specific attack types, our one-class AD model is designed to flag any behavior that causes deviations in system parameters, regardless of the attack's nature. This enables the model could be able to detect even previously unknown or unmodeled attack strategies, enhancing the robustness and security performance of QKD system. Importantly, our experimental results show that the anomaly detection performance of the AD model depends on the richness of the training dataset. By feeding the model with more diverse and representative training data, the underlying neural network can learn a broader range of system behavior patterns, thereby enhancing its anomaly detection capability. Overall, our approach offers advantages in generality, simplicity, low cost, and high accuracy, paving the way for more robust and scalable QKD system security in future deployments.

The paper is structured as follows. Section II introduces the construction of AD model, including the neural network architecture and dataset preparation. Section III presents the training process and testing results of the model. In Sec. IV, we provide a comparative analysis between our AD approach and traditional countermeasure strategies, as well as with machine-learning-based multi-class classification models. Finally, Sec. V concludes the study.

## II. ANOMALY DETECTION MODEL DESIGN AND DATASET CONSTRUCTION

In this section, we present the design of the AD model and the construction of the corresponding dataset. The dataset is built by capturing the operational states of the QKD system. Parameters collected under secure conditions are treated as normal data, while those recorded during attacks are labeled as anomalies. Importantly, the neural network is trained exclusively on the normal dataset to learn representative features—without relying on any anomalous samples. These features are then used by the Deep SVDD algorithm to construct a hypersphere that encloses the distribution of normal data, such that anomalous inputs fall outside the hypersphere and can be effectively detected. The overall concept of the AD model is illustrated in Fig. 1.

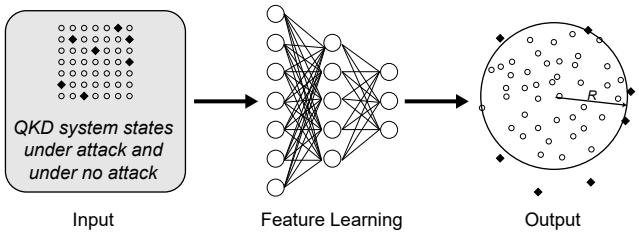


FIG. 1: The conceptual diagram of the AD model. Hollow circles represent the QKD system's states under secure conditions, while solid squares denote states generated under attack. These state parameters are extracted to form the dataset for anomaly detection. During training, the neural network learns only from normal data to construct a hypersphere that encloses their distribution. Any anomalous data falling outside this hypersphere is identified as abnormal, thus enabling effective anomaly detection.

### A. Deep Support Vector Data Description and Neural Network Modeling

To achieve general and real-time monitoring of QKD systems, we adopt the AD model based on Deep SVDD [39], a deep one-class classification model. The core idea of Deep SVDD is to extend traditional SVDD [41] by integrating it with deep neural networks, thereby enabling the learning of expressive feature representations from structured or high-dimensional inputs. Unlike multi-class classification approaches that require labeled data from various attack types, Deep SVDD requires only data collected under normal operating conditions for training, without any prior knowledge of attack behaviors. This makes it especially suitable for QKD scenarios, where anomalous data (e.g., from new or rare attacks) may be unavailable during training. During deployment, the model learns a compact feature space for normal data and identifies any significant deviations as anomalies, regardless of the specific nature of the underlying attack. This generalization ability allows the model to detect a broad range of potential threats—including previously unknown attacks—by solely analyzing the operational parameters of a running QKD system. A detailed explanation of both SVDD and Deep SVDD is provided in Appendix A.

In our AD model, we consider that only a small number of parameters can be extracted from the QKD system. Therefore, we adopt the multilayer perceptron (MLP) architecture as the backbone network for Deep SVDD, due to its simplicity, low computational cost, and effectiveness in small-scale feature learning. To fit our dataset, we adjust the input layer to match the dimensionality of the QKD parameter vector and configure hid-

den layers to extract representative features from normal data. These features are then used to construct a hypersphere enclosing the distribution of normal samples. Leveraging the representation learning capability of MLP, the customized Deep SVDD model transforms normal QKD data into compact, informative latent representations. Through forward propagation, each network layer encodes progressively higher-level abstractions of the QKD parameter space. The model is trained using the Deep SVDD objective, which minimizes the volume of a hypersphere enclosing the latent features of normal data—ensuring that the network captures the typical behavior of secure QKD operation.

Training is performed via backpropagation, iteratively updating the network weights to reduce the hypersphere loss. As the model converges, it forms a reliable boundary around the normal data distribution. At inference time, any input whose representation falls outside this hypersphere is flagged as anomalous, indicating a potential deviation from secure operation. This design enables effective anomaly detection without relying on labeled attack samples or prior knowledge of the attack method, making it well-suited for real-world QKD systems.

### B. Dataset Construction for Anomaly Detection

In this subsection, we describe the construction of the datasets used to train and evaluate the AD model. The guiding principle is that the parameters recorded during the operation of the QKD system should faithfully reflect its internal state and behavior. By capturing these operational characteristics, the dataset enables the AD model to effectively detect anomalies introduced by malicious attacks. From the perspective of Eve, we consider how different attack strategies affect the QKD system and accordingly propose two methods for constructing datasets based on the nature of these attacks. In the first type of active attack, Eve manipulates the system by modifying its configuration parameters. For instance, Eve may shift the timing of detector gate signals through the calibration attack [9], change the bias voltage of intensity modulators via the induced-photorefraction attack [42–44]. To capture the effects of such intrusions, the QKD system can be equipped with monitoring interfaces that continuously record these configuration parameters in real time, forming a dataset that reflects system behavior under various conditions. In the second type of attack, Eve may not alter any system configuration parameters directly. This type of attacks, such as, the muted attack [38] and the superlinear attack [45], manipulate the single-photon detector (SPD) responses without changing the system's internal settings. In these cases, we focus on the SPD's output behavior—particularly the timing of detection events—and use the associated timestamps to construct a dataset that captures anomalies induced by this class of attacks.

For the construction of datasets based on QKD system configuration parameters, we extract a wide range of relevant information from as many stages of the QKD operation as possible. As a concrete example, we consider a discrete-variable QKD system employing polarization encoding. The QKD process can be broadly divided into three stages: the calibration stage, the raw key exchange stage, and the post-processing stage. During the calibration stage, the receiver (Bob) continuously adjusts the timing of the gate signals for his SPDs to synchronize with the photon pulses sent by the transmitter (Alice). Simultaneously, Bob tunes the internal polarization controller (PC) to ensure proper alignment with Alice's polarization encoding. Both the gate timing values and the PC settings are recorded as part of the AD dataset during this stage. In the raw key exchange stage, Alice and Bob perform quantum signal transmission and reception to generate raw keys. No system configuration parameters are typically adjusted during this phase. In the post-processing stage, Alice and Bob perform a series of statistical evaluations. These include the number of sifted keys, the ratio of signal detections to decoy state emissions, the ratio of signal to decoy state detections, the detection efficiencies corresponding to signal, decoy, and vacuum states, as well as the quantum bit error rates (QBERs) for each polarization basis and the overall QBER. During error correction and privacy amplification, additional parameters—such as privacy amplification factors—are also generated. Since these statistical quantities objectively reflect the operational state of the QKD system, they are likewise included in the AD dataset.

To construct the dataset for attacks that do not alter system configuration parameters, we extract the timestamps of SPD responses as input data for anomaly detection. Specifically, in our experimental setup, the SPDs are connected to a time-to-digital converter (TDC), which records the arrival time of each detection event. The TDC operates with a 100 ns cycle, capturing the precise time at which each SPD response occurs. As a result, all data in this dataset represent detection timestamps, with values ranging from 0 to 100 ns. Unlike the previous dataset based on system configuration and statistical parameters, this dataset is directly constructed from individual SPD counts, allowing the feature dimension to scale with the number of detection events. In our experiments, we generated datasets with feature dimensions corresponding to 100, 225, and 400 detection events, respectively. These datasets were then used to train and test the anomaly detection model. Based on the performance of the AD model, we further identified the optimal feature dimension for this type of dataset.

For normal data, both datasets are generated by operating the QKD system in a secure environment. We then select the calibration attack [9] and the muted attack [38] to generate anomalous samples because these attacks currently cannot be effectively mitigated through adding physical components or refining theoretical security model. Moreover, experimental results have shown

that both attacks are capable of obtaining sifted keys from the QKD system, thereby severely compromising its security. Considering the nature of these two attacks, the anomalous data for the first dataset are collected under the calibration attack, while those for the second dataset are collected under the muted attack. During the training of the AD model, only normal data are used, whereas the testing phase involves datasets composed of normal and anomalous samples in a 1:1 ratio.

## III. MODEL TRAINING AND PERFORMANCE EVALUATION

As described in Sec. II, we constructed two types of datasets. Accordingly, the same AD model was trained and evaluated separately on each dataset. The training was performed using identical hyperparameter settings, as summarized in Table I. To evaluate the model's performance in detecting anomalies, we used the AUC as the evaluation metric [39, 46]. A detailed explanation of the AUC is provided in Appendix B.

TABLE I: Training hyperparameters for the AD model.

| Optimizer | Learning Rate | Batch Size | Epochs |
|---|---|---|---|
| Adam | 0.0001 | 128 | 150 |

We first train and evaluate the AD model using a dataset consisting of configuration parameters extracted from the calibration and post-processing stages of the QKD system. To mitigate the influence of experimental randomness, the trained AD model is independently tested on different test sets 100 times. The resulting AUC values are shown in Fig. 2. Across these 100 tests, the average AUC reached 99.03%, with the minimum value still exceeding 92%. These results demonstrate that the AD model, trained solely on parameters collected under normal conditions, can effectively learn the characteristics of legitimate system behavior. When subjected to attacks that alter system parameters, such as the calibration attack, the model achieves an anomaly detection rate exceeding 99%, thereby enabling reliable identification of Eve's presence.

As discussed in Sec. II, we constructed the dataset using the timestamps of SPD counts and evaluated the anomaly detection capability of the AD model against the muted attack. Similarly, to ensure the robustness of the results, the trained model was evaluated 100 times, each with a different testing set. The resulting AUC values are shown in Fig. 3 and Table II. It is demonstrated that, the anomaly detection performance of the AD model improves with the number of SPD counts used in the dataset. When the detection events reach 400, the average AUC value over 100 independent tests is 99.03%, indicating a high detection accuracy. Moreover, the variance is only 1.09, demonstrating the stability of the AD model.

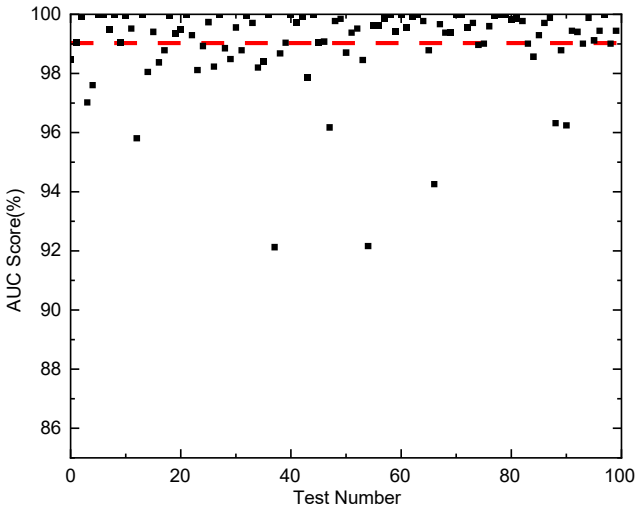


FIG. 2: AUC results obtained from 100 independent tests using the configuration parameters dataset. The dataset is constructed from the parameters of the calibration and post-processing stages, while the anomalous data were generated under the calibration attack. The black dots represent the actual AUC values obtained from each test, and the red dashed line indicates the average AUC across all 100 runs.

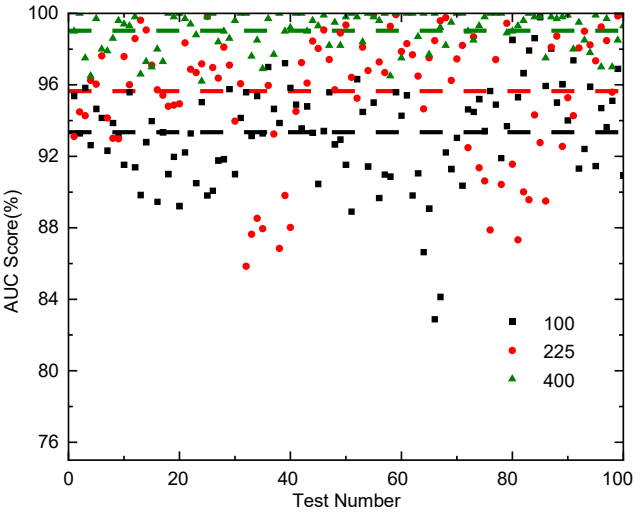


FIG. 3: AUC results obtained from 100 independent tests using the dataset of SPD response timestamps. The dataset is constructed from the timestamps corresponding to SPD counts, with the anomalous data generated under the muted attack. The black square dots represent the AUC values obtained by testing the AD model trained on a dataset constructed using 100 SPD counts, while the black dashed line indicates the average AUC across the 100 tests. Similarly, the red circular dots and the green triangular dots (along with their corresponding dashed lines) represent the AUC values (and their averages across the 100 tests) when the dataset is constructed using 225 and 400 SPD counts, respectively.

TABLE II: The mean and variance of test AUC for feature dimensions of 100, 225, and 400.

| | Feature dimension | | |
|---|---|---|---|
| Statistics | 100 | 225 | 400 |
| Mean(%) | 93.35 | 95.65 | 99.03 |
| Variance | 8.07 | 12.45 | 1.09 |

To further investigate why a larger data sample size enhances anomaly detection performance, we analyze the structure of the dataset derived from TDC timestamps. Each 100 ns cycle is divided into bins of 0.1 ns, and statistical histograms of overlapping detection events are generated for sample sizes of 100, 225, and 400 counts, as shown in Fig. 4. For comparison, we also present a histogram based on 4000 counts. From Fig. 4, we observe that under attack-free conditions, SPD responses exhibit an approximately uniform distribution, whereas under the muted attack, the responses become more concentrated in specific regions as explained in Ref. [38]. When the number of detection events is small [e.g., Fig. 4(a)], the distribution difference between normal and anomalous data is not particularly pronounced. However, as the sample size increases, the discrepancy in the temporal distribution becomes increasingly evident. This divergence allows the AD model to extract more discriminative features, thereby improving its ability to detect anomalies. In this context, the feature dimension—determined by the number of detection events—is positively correlated with detection performance. Nevertheless, as the AD model already achieves an AUC above 99% with low variance at 400 detection events, further increasing the feature dimension would significantly raise the complexity of data collection and the computational cost of model training and inference. Therefore, we adopt 400 detection events as the input size for our AD model, balancing detection accuracy with resource efficiency.

Through training and testing experiments on the AD model using two different datasets, it is demonstrated that the AD model can achieve excellent anomaly detection performance by training solely on datasets generated under secure conditions in the QKD system. The average AUC values obtained over multiple tests exceed 99%, indicating that the AD model has effectively learned the characteristics of normal QKD system. When Eve launches an active attack, the AD model can promptly and accurately detect the anomaly, thereby helping to guard the practical security of the QKD system.

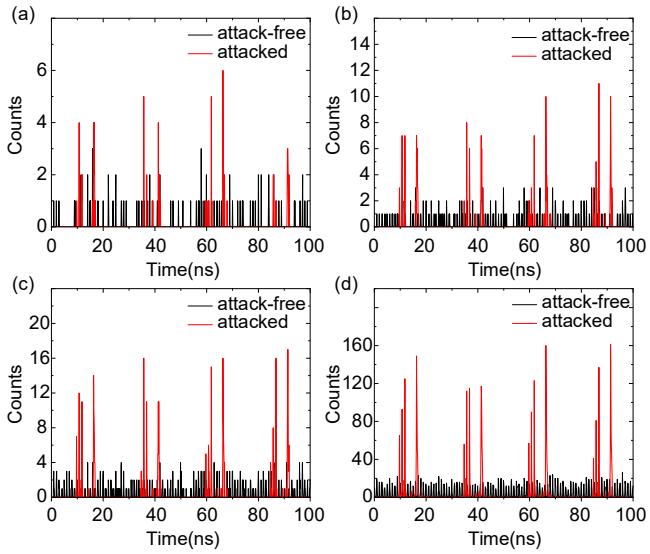


FIG. 4: Timestamp distributions under different numbers of detection events. (a)–(d) show the histogram of SPD timestamps within a 100 ns cycle, using a bin size of 0.1 ns, for total detection counts of 100, 225, 400, and 4000, respectively. The black histograms represent data collected under secure conditions, while the red histograms correspond to data collected under muted attack.

## IV. DISCUSSION

In this section, we present a comparative analysis between the proposed AD model and conventional methods, as well as a comparison with strategies based on multiclass classification approaches.

Ensuring the practical security of existing QKD infrastructures is essential. Unlike approaches that require the design and deployment of new QKD protocols, the AD method proposed here operates entirely within existing systems and requires no additional optical or electrical hardware. By extracting operational parameters directly from deployed QKD systems, our model enables effective detection of anomalous behavior, helping to maintain system security in real time. Furthermore, since our method leverages internally generated QKD data rather than relying on specific implementation details, it provides strong portability and general applicability across a wide range of QKD architectures.

Adding physical devices to defend against attacks is a common "patching" strategy in QKD system design. Such countermeasures are typically tailored to specific attacks. However, while potentially effective against known threats, this approach increases the physical complexity and cost of QKD system implementation. Moreover, the introduction of additional hardware components may itself create new, unforeseen side channels, thereby compromising the practical security of the system. For instance, a countermeasure designed to detect the intensity of injected light—intended to prevent detector blinding attacks—was experimentally shown to be ineffective [34]. By contrast, our proposed AD model avoids these drawbacks. It enables legitimate users to monitor the operational status of the QKD system in real time for signs of abnormal behavior, without requiring any additional optical or electrical hardware. As a result, it does not

introduce new potential vulnerabilities that could be exploited by Eve. Furthermore, since the AD module is deployed within the trusted environment (e.g., inside the room of the legitimate parties), and in QKD, Eve is not permitted to access this area physically, the security of the AD model can be reasonably assumed.

Another line of countermeasure involves refining the theoretical security models of QKD systems. This approach focuses on analyzing Eve's ability to compromise the key through specific attacks and estimating the corresponding secure key rate. However, this approach has inherent limitations. It relies on prior knowledge of specific attack mechanisms and explicitly incorporates the resulting information leakage into the security analysis. As a result, it offers no protection against unknown or yet-to-be-modeled attacks. For certain types of attacks—such as the calibration attack and muted attack used in this work to evaluate the AD model—there currently exist no effective theoretical modeling tools. In contrast, our proposed AD model does not rely on a priori modeling of specific attack strategies. Instead, it identifies anomalies by detecting deviations in QKD system parameters, regardless of the underlying attack mechanism. This makes the AD model not only effective against known and modeled attacks, but also potentially capable of detecting previously unknown or uncharacterized threats.

Similarly based on machine learning, our method adopts a one-class anomaly detection approach rather than a multi-class classification algorithm. This choice is motivated by the fundamental principle that the primary objective of intelligent countermeasure in QKD systems is to detect the presence of attacks, rather than to identify or classify specific attack types. While multi-class models can classify known types of attacks present in the training dataset, they are ineffective against previously unobserved attacks. In contrast, our AD model is specifically designed with this limitation in mind. It does not attempt to distinguish the type of attack but instead flags any behavior that causes deviations in the system's operational parameters as anomalous. This design ensures that the AD model possesses a degree of generalization, enabling it to detect even unknown attack strategies.

Moreover, our AD framework can be trained in an unsupervised manner, which greatly simplifies data preparation. For example, in practical experiments, raw data such as time-stamped detector counts can be directly used to construct the dataset without requiring manual labeling or attack classification. The detection capability of our AD model is closely linked to the richness and diversity of the training data. By feeding the model with a more comprehensive dataset, the underlying neural network can learn a broader and more nuanced representation of the system's normal behavior, thereby enhancing its ability to detect anomalies. This indicates that the model's detection performance is both scalable and dependent on the quality of the training dataset. Therefore, the design of a dataset that comprehensively captures the characteristics of QKD system parameters is essential for the effective and robust deployment of the AD model.

## V. CONCLUSION

This paper proposes a machine learning-based AD method to achieve real-time monitoring of QKD system parameters. By building an AD model based on Deep SVDD, we guard that the QKD system operates under secure conditions. The datasets used for model training and evaluation are constructed from parameters recorded during the QKD system's operational process, including system configuration parameters and detector response timestamps. The training set exclusively consists of data collected when the QKD system operates securely. For testing, we generate a balanced dataset (1:1 ratio) combining data from both secure and attacked conditions. Two classical attack strategies — the calibration attack and the muted attack — are used to produce anomalous data in the test set. Testing results show that it can achieve an AUC of up to 99.03%, indicating a high capability to detect the presence of attacks. Our AD approach offers generality, simplicity, low cost, and high accuracy, making it a promising solution for ensuring the practical security of QKD systems and providing valuable insights for the design of future QKD architectures.

## ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (Grant No. 62371459) and the Innovation Program for Quantum Science and Technology (2021ZD0300704).

### Appendix A: Support Vector Data Description and Deep Support Vector Data Description

Our AD model is based on SVDD and Deep SVDD proposed in Ref. [41] and Ref. [39]. To make our paper self-contained, we present the main results from Ref. [41] and Ref. [39] in this section.

Traditional SVDD is a kernel-based method that maps input data from the original feature space into a high-dimensional feature space via a feature mapping function [41]. In this space, SVDD aims to find the smallest hypersphere, with center $\mathbf{c} \in \mathcal{F}_k$ and radius $R > 0$, that encloses most of the normal data points. Data points that lie outside the hypersphere are considered anomalies. The objective function is formulated as:

$$\min_{R,\mathbf{c},\boldsymbol{\xi}} R^2 + \frac{1}{\nu n}\sum_i \xi_i \qquad \text{s.t.} \quad \|\phi_k(\mathbf{x}_i) - \mathbf{c}\|^2_{\mathcal{F}_k} \le R^2 + \xi_i, \quad \xi_i \ge 0, \quad \forall i \tag{A1}$$

Here, the slack variables $\xi_i > 0$ allow for a soft boundary, and the hyperparameter $\nu \in (0, 1]$ controls the trade-off between the volume of the hypersphere and the penalty for outliers. Points satisfying $\|\phi_k(\mathbf{x}_i) - \mathbf{c}\|^2_{\mathcal{F}_k} > R^2$ are considered anomalous.

However, traditional SVDD heavily relies on the choice of kernel functions, making it difficult to capture the intrinsic structure and features of complex nonlinear data. In addition, the storage and computational complexity of the kernel matrix is $\mathcal{O}(n^2)$, which limits its scalability to large datasets. Moreover, the hypersphere's center and radius typically need to be predefined, which reduces the model's robustness to shifts in the data distribution [47, 48]. The Deep SVDD model differs from traditional SVDD by replacing the kernel function mapping with a deep neural network [39]. It directly learns nonlinear features in the original input space, enabling automatic learning of hierarchical and abstract representations of the data. This forms an end-to-end deep one-class classification framework that can more effectively distinguish between normal and abnormal data.

In the following, we present the main principles of Deep SVDD as introduced in Ref. [39]. For some input space $\mathcal{X} \subset \mathbb{R}^d$ and output space $\mathcal{F} \subset \mathbb{R}^p$, let $\phi(\cdot; \mathbf{W}) : \mathcal{X} \to \mathcal{F}$ be a neural network with $L \in \mathbb{N}$ hidden layers and a set of weights $\mathbf{W} = \{\mathbf{W}^1, \ldots, \mathbf{W}^L\}$, where $\mathbf{W}^\ell$ are the weights of layer $\ell \in \{1, \ldots, L\}$. That is, $\phi(\mathbf{x}; \mathbf{W}) \in \mathcal{F}$ is the feature representation of $\mathbf{x} \in \mathcal{X}$ given by the network $\phi$ with parameters $\mathbf{W}$. The aim of Deep SVDD is to jointly learn the network parameters $\mathbf{W}$ while minimizing the volume of a data-enclosing hypersphere in the output space $\mathcal{F}$, characterized by a radius $R > 0$ and center $\mathbf{c} \in \mathcal{F}$, which is assumed to be fixed for now. Given some training data $\mathcal{D}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathcal{X}$, the *soft-boundary* Deep SVDD objective is:

$$\min_{R,\mathbf{W}} R^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \max\left\{0, \|\phi(\mathbf{x}_i; \mathbf{W}) - \mathbf{c}\|^2 - R^2\right\} + \frac{\lambda}{2}\sum_{\ell=1}^{L} \left\|\mathbf{W}^\ell\right\|_F^2 \tag{A2}$$

The first term minimizes the volume of the hypersphere, while the second term penalizes data points that lie outside the hypersphere after passing through the network. The hyperparameter $\nu \in (0, 1]$ controls the trade-off between the hypersphere volume and the violation of its boundary, allowing some points to be mapped outside the hypersphere. The final term is a weight decay regularizer for the network parameters. This objective function is suitable for training scenarios where a small number of normal samples may lie outside the hypersphere due to feature fluctuations, making it more applicable to datasets that may contain slight anomalies.

### Appendix B: The Area Under the Receiver Operating Characteristic Curve and Its Calculation

In classification tasks within machine learning, predictive models typically output numerical values representing prediction scores or estimated probabilities that a sample belongs to a particular class. For one-class classification, a threshold must be predetermined to make binary decisions: if a sample's score exceeds this threshold, it is classified as normal; otherwise, it is labeled as anomalous. It is evident that the choice of threshold directly affects the classification results and, consequently, alters the composition of the confusion matrix, which records the relationship between predicted labels and ground truth.

The receiver operating characteristic curve demonstrates the performance of a classifier by systematically depicting the relationship between the true positive rate (TPR) and the false positive rate (FPR) under different thresholds. The formulas for TPR and FPR are given as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{B1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{B2}$$

The meanings of TP, FN, FP, and TN are shown in Table III.

TABLE III: Confusion matrix of classification results.

| | Predicted | |
|---|---|---|
| Actual | Positive | Negative |
| Positive | True Positive (TP) | False Negative (FN) |
| Negative | False Positive (FP) | True Negative (TN) |

The AUC represents the area under the receiver operating characteristic curve and provides a quantitative

measure of the overall performance of the classifier. It is defined as:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d\mathrm{FPR} \tag{B3}$$

The AUC is fundamentally a probabilistic measure. It represents the likelihood that, when randomly selecting one normal and one anomalous sample, the classifier assigns a higher score to the normal sample than to the anomalous one. From this definition, a higher AUC value implies that the classifier is more capable of distinguishing between normal and anomalous samples, thereby indicating better classification performance.

The AUC is particularly suitable for imbalanced datasets, as it is insensitive to the ratio of normal to anomalous samples. Moreover, the AUC provides an intuitive and interpretable metric to evaluate the classifier's discriminative power between the two classes. An AUC value of 1 indicates that there exists at least one threshold at which the classifier can perfectly separate all normal and anomalous samples—an ideal case that is rarely achieved in real-world scenarios. When the AUC lies between 0.5 and 1, the classifier performs better than random guessing, with performance improving as the value approaches 1. Conversely, an AUC below 0.5 suggests that the classifier performs worse than random guessing; however, in such cases, inverting the prediction decisions can result in performance better than chance.


[1] C. H. Bennett and G. Brassard, Quantum cryptography: Public key distribution and coin tossing, in *Proc. IEEE International Conference on Computers, Systems, and Signal Processing (Bangalore, India)* (IEEE Press, New York, 1984) pp. 175–179.

[2] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Quantum cryptography, Rev. Mod. Phys. **74**, 145 (2002).

[3] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, The security of practical quantum key distribution, Rev. Mod. Phys. **81**, 1301 (2009).

[4] F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, Secure quantum key distribution with realistic devices, Rev. Mod. Phys. **92**, 025002 (2020).

[5] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Hacking commercial quantum cryptography systems by tailored bright illumination, Nat. Photonics **4**, 686 (2010).

[6] F. Xu, B. Qi, and H.-K. Lo, Experimental demonstration of phase-remapping attack in a practical quantum key distribution system, New J. Phys. **12**, 113026 (2010).

[7] C. Wiechers, L. Lydersen, C. Wittmann, D. Elser, J. Skaar, C. Marquardt, V. Makarov, and G. Leuchs, After-gate attack on a quantum cryptosystem, New J. Phys. **13**, 013043 (2011).

[8] I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Skaar, C. Kurtsiefer, and V. Makarov, Full-field implementation of a perfect eavesdropper on a quantum cryptography system, Nat. Commun. **2**, 349 (2011).

[9] N. Jain, C. Wittmann, L. Lydersen, C. Wiechers, D. Elser, C. Marquardt, V. Makarov, and G. Leuchs, Device calibration impacts security of quantum key distribution, Phys. Rev. Lett. **107**, 110501 (2011).

[10] A. N. Bugge, S. Sauge, A. M. M. Ghazali, J. Skaar, L. Lydersen, and V. Makarov, Laser damage helps the eavesdropper in quantum cryptography, Phys. Rev. Lett. **112**, 070503 (2014).

[11] A. Huang, S. Sajeed, P. Chaiwongkhot, M. Soucarros, M. Legré, and V. Makarov, Testing random-detector-efficiency countermeasure in a commercial system reveals a breakable unrealistic assumption, IEEE J. of Quantum Electron. **52**, 1 (2016).

[12] A. Huang, S.-H. Sun, Z. Liu, and V. Makarov, Quantum key distribution with distinguishable decoy states, Phys. Rev. A **98**, 012330 (2018).

[13] Y.-J. Qian, D.-Y. He, S. Wang, W. Chen, Z.-Q. Yin, G.-C. Guo, and Z.-F. Han, Hacking the quantum key distribution system by exploiting the avalanche-transition region of single-photon detectors, Phys. Rev. Appl. **10**, 064062 (2018).

[14] A. Huang, A. Navarrete, S.-H. Sun, P. Chaiwongkhot, M. Curty, and V. Makarov, Laser-seeding attack in quantum key distribution, Phys. Rev. Appl. **12**, 064043 (2019).

[15] A. Huang, R. Li, V. Egorov, S. Tchouragoulov, K. Kumar, and V. Makarov, Laser-damage attack against optical attenuators in quantum key distribution, Phys. Rev. Appl. **13**, 034017 (2020).

[16] S. Sun and A. Huang, A review of security evaluation of practical quantum key distribution system, Entropy **24**, 260 (2022).

[17] P. Chaiwongkhot, J. Zhong, A. Huang, H. Qin, S.-c. Shi, and V. Makarov, Faking photon number on a transition-edge sensor, EPJ Quantum Technol. **9**, 23 (2022).

[18] B. Gao, Z. Wu, W. Shi, Y. Liu, D. Wang, C. Yu, A. Huang, and J. Wu, Ability of strong-pulse illumination to hack self-differencing avalanche photodiode detectors in a high-speed quantum-key-distribution system, Phys. Rev. A **106**, 033713 (2022).

[19] A. Ponosova, D. Ruzhitskaya, P. Chaiwongkhot, V. Egorov, V. Makarov, and A. Huang, Protecting fiber-optic quantum key distribution sources against light-injection attacks, PRX Quantum **3**, 040307 (2022).

[20] A. Huang, A. Mizutani, H.-K. Lo, V. Makarov, and K. Tamaki, Characterization of state-preparation uncertainty in quantum key distribution, Phys. Rev. Appl. **19**, 014048 (2023).

[21] Q. Peng, B. Gao, K. Zaitsev, D. Wang, J. Ding, Y. Liu, Q. Liao, Y. Guo, A. Huang, and J. Wu, Security boundaries of an optical-power limiter for protecting quantum-key-distribution systems, Phys. Rev. Appl. **21**, 014026 (2024).

[22] T. Xing, J. Liu, L. Zhang, M.-Y. Wang, Y.-H. Li, R. Liu, Q. Peng, D. Wang, Y. Wang, H. Liu, W. Li, Y. Cao, and A. Huang, Characterization of intensity correlation via single-photon detection in quantum key distribution,

Opt. Express **32**, 31767 (2024).
[23] J. Liu, T. Xing, R. Liu, Z. Chen, H. Tan, and A. Huang, Intensity correlations in measurement-device-independent quantum key distribution, Opt. Express **32**, 38394 (2024).
[24] Q. Peng, J.-P. Chen, T. Xing, D. Wang, Y. Wang, Y. Liu, and A. Huang, Practical security of twin-field quantum key distribution with optical phase-locked loop under wavelength-switching attack, npj Quantum Inf. **11**, 7 (2025).
[25] H.-K. Lo, M. Curty, and B. Qi, Measurement-device-independent quantum key distribution, Phys. Rev. Lett. **108**, 130503 (2012).
[26] X.-B. Wang, Beating the photon-number-splitting attack in practical quantum cryptography, Phys. Rev. Lett. **94**, 230503 (2005).
[27] H.-K. Lo, X. Ma, and K. Chen, Decoy state quantum key distribution, Phys. Rev. Lett. **94**, 230504 (2005).
[28] X. Ma, B. Qi, Y. Zhao, and H.-K. Lo, Practical decoy state for quantum key distribution, Phys. Rev. A **72**, 012326 (2005).
[29] S.-K. Liao, W.-Q. Cai, W.-Y. Liu, L. Zhang, Y. Li, J.-G. Ren, J. Yin, Q. Shen, Y. Cao, Z.-P. Li, F.-Z. Li, X.-W. Chen, L.-H. Sun, J.-J. Jia, J.-C. Wu, X.-J. Jiang, J.-F. Wang, Y.-M. Huang, Q. Wang, Y.-L. Zhou, L. Deng, T. Xi, L. Ma, T. Hu, Q. Zhang, Y.-A. Chen, N.-L. Liu, X.-B. Wang, Z.-C. Zhu, C.-Y. Lu, R. Shu, C.-Z. Peng, J.-Y. Wang, and J.-W. Pan, Satellite-to-ground quantum key distribution, Nature **549**, 43 (2017).
[30] S.-K. Liao, W.-Q. Cai, J. Handsteiner, B. Liu, J. Yin, L. Zhang, D. Rauch, M. Fink, J.-G. Ren, W.-Y. Liu, Y. Li, Q. Shen, Y. Cao, F.-Z. Li, J.-F. Wang, Y.-M. Huang, L. Deng, T. Xi, L. Ma, T. Hu, L. Li, N.-L. Liu, F. Koidl, P. Wang, Y.-A. Chen, X.-B. Wang, M. Steindorfer, G. Kirchner, C.-Y. Lu, R. Shu, R. Ursin, T. Scheidl, C.-Z. Peng, J.-Y. Wang, A. Zeilinger, and J.-W. Pan, Satellite-relayed intercontinental quantum network, Phys. Rev. Lett. **120**, 030501 (2018).
[31] Y.-A. Chen, Q. Zhang, T.-Y. Chen, W.-Q. Cai, S.-K. Liao, J. Zhang, K. Chen, J. Yin, J.-G. Ren, Z. Chen, S.-L. Han, Q. Yu, K. Liang, F. Zhou, X. Yuan, M.-S. Zhao, T.-Y. Wang, X. Jiang, L. Zhang, W.-Y. Liu, Y. Li, Q. Shen, Y. Cao, C.-Y. Lu, R. Shu, J.-Y. Wang, L. Li, N.-L. Liu, F. Xu, X.-B. Wang, C.-Z. Peng, and J.-W. Pan, An integrated space-to-ground quantum communication network over 4,600 kilometres, Nature **589**, 214 (2021).
[32] M. Lucamarini, I. Choi, M. B. Ward, J. F. Dynes, Z. L. Yuan, and A. J. Shields, Practical security bounds against the trojan-horse attack in quantum key distribution, Phys. Rev. X **5**, 031030 (2015).
[33] V. Makarov, J.-P. Bourgoin, P. Chaiwongkhot, M. Gagné, T. Jennewein, S. Kaiser, R. Kashyap, M. Legré, C. Minshull, and S. Sajeed, Creation of backdoors in quantum communications via laser damage, Phys. Rev. A **94**, 030302 (2016).
[34] Z. Wu, A. Huang, H. Chen, S.-H. Sun, J. Ding, X. Qiang, X. Fu, P. Xu, and J. Wu, Hacking single-photon avalanche detectors in quantum key distribution via pulse illumination, Opt. Express **28**, 25574 (2020).
[35] K. Tamaki, M. Curty, and M. Lucamarini, Decoy-state quantum key distribution with a leaky source, New J. Phys. **18**, 065008 (2016).
[36] V. Zapatero, Á. Navarrete, K. Tamaki, and M. Curty, Security of quantum key distribution with intensity correlations, Quantum **5**, 602 (2021).
[37] X. Sixto, V. Zapatero, and M. Curty, Security of decoy-state quantum key distribution with correlated intensity fluctuations, Phys. Rev. Appl. **18**, 044069 (2022).
[38] J. Su, J. Chen, F. Lu, Z. Chen, J. Liu, D. He, S. Wang, and A. Huang, Muted attack on a high-speed quantum key distribution system (2025), arXiv:2506.03718 [quant-ph].
[39] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, Deep one-class classification, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80, edited by J. Dy and A. Krause (PMLR, 2018) pp. 4393–4402.
[40] Pantoja, John J., Bucheli, Victor A., and Donaldson, Ross, Electromagnetic side-channel attack risk assessment on a practical quantum-key-distribution receiver based on multi-class classification, EPJ Quantum Technol. **11**, 78 (2024).
[41] D. M. Tax and R. P. Duin, Support vector data description, Machine Learning **54**, 45 (2004).
[42] P. Ye, W. Chen, G.-W. Zhang, F.-Y. Lu, F.-X. Wang, G.-Z. Huang, S. Wang, D.-Y. He, Z.-Q. Yin, G.-C. Guo, and Z.-F. Han, Induced-photorefraction attack against quantum key distribution, Phys. Rev. Appl. **19**, 054052 (2023).
[43] F.-Y. Lu, P. Ye, Z.-H. Wang, S. Wang, Z.-Q. Yin, R. Wang, X.-J. Huang, W. Chen, D.-Y. He, G.-J. Fan-Yuan, G.-C. Guo, and Z.-F. Han, Hacking measurement-device-independent quantum key distribution, Optica **10**, 520 (2023).
[44] L. Han, Y. Li, H. Tan, W. Zhang, W. Cai, J. Yin, J. Ren, F. Xu, S. Liao, and C. Peng, Effect of light injection on the security of practical quantum key distribution, Phys. Rev. Appl. **20**, 044013 (2023).
[45] L. Lydersen, N. Jain, C. Wittmann, O. Marøy, J. Skaar, C. Marquardt, V. Makarov, and G. Leuchs, Superlinear threshold detectors in quantum cryptography, Phys. Rev. A **84**, 032320 (2011).
[46] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning, Pattern Recognition **58**, 121 (2016).
[47] A. Rahimi and B. Recht, Random features for large-scale kernel machines, in *Advances in Neural Information Processing Systems*, Vol. 20, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Inc., 2007).
[48] M. Pal and G. M. Foody, Feature selection for classification of hyperspectral data by svm, IEEE Transactions on Geoscience and Remote Sensing **48**, 2297 (2010).