

QSentry: Backdoor Detection for Quantum Neural Networks via Measurement Clustering

Shuolei Wang[†]

*School of Electronic Information
Central South University
Changsha, China
wangsl@csu.edu.cn*

Zimeng Xiao[†]

*School of Electronic Information
Central South University
Changsha, China
xiaozi@meng@csu.edu.cn*

Jinjing Shi^{*}

*School of Electronic Information
Central South University
Changsha, China
shijinjing@csu.edu.cn*

Heyuan Shi^{*}

*School of Electronic Information
Central South University
Changsha, China
shiheyuan@csu.edu.cn*

Shichao Zhang

*Guangxi Normal University
Guilin, China
zhangsc@mailbox.gxnu.edu.cn*

Xuelong Li

*Institute of Artificial Intelligence (Tele AI)
China Telecom
Beijing, China
xuelong_li@ieee.org*

Abstract—Quantum neural networks (QNNs) are an important model for implementing quantum machine learning (QML), while they demonstrate a high degree of vulnerability to backdoor attacks similar to classical networks. To address this issue, a quantum backdoor attack detection framework called QSentry is proposed, in which a quantum Measurement Clustering method is introduced to detect backdoors by identifying statistical anomalies in measurement outputs. It is demonstrated that QSentry can effectively detect anomalous distributions induced by backdoor samples with extensive experiments. It achieves a 75.8% F1 score even under a 1% poisoning rate, and further improves to 85.7% and 93.2% as the poisoning rate increases to 5% and 10%, respectively. The integration of silhouette coefficients and relative cluster size enable QSentry to precisely isolate backdoor samples, yielding estimates that closely match actual poisoning ratios. Evaluations under various quantum attack scenarios demonstrate that QSentry delivers superior robustness and accuracy compared with three state-of-the-art detection methods. This work establishes a practical and effective framework for mitigating backdoor threats in QML.

Index Terms—Quantum Neural Networks, Quantum Backdoors, Backdoor Attacks, Backdoor Detection, Quantum Security.

I. INTRODUCTION

Quantum neural networks (QNNs) represent an emerging class of deep learning (DL) architectures that leverage quantum computation as the underlying computational paradigm. Using quantum-classical hybrid algorithms, QNNs process quantum states via parameterized quantum circuits [1]. Some quantum learning models exploit superposition and entanglement to achieve potential computational advantages over their classical counterparts. Recent studies demonstrate promising performance of QNNs in classification [2], generative modeling [3], and quantum chemistry simulations [4]. Furthermore, a novel model design enhances QNNs' ability to express complex data modalities and increases their interpretability [5], [6].

The vulnerability of deep neural networks (DNNs) [7] to backdoor attacks poses a substantial security threat. A paradigmatic example of this phenomenon is presented in BadNets [8], wherein adversaries surreptitiously embedded trigger mechanisms, reminiscent of those employed in stickers, into images of traffic signs. This covert intervention could lead the classifier to erroneously identify a modified stop sign as a speed limit sign. This attack scenario has the potential to induce severe malfunctions in autonomous driving systems, as illustrated in Figure 1. Subsequent works such as BAIT [9], DeepVenom [10], and studies on attack orthogonality [11] further reveal the broad spectrum and persistence of backdoor vulnerabilities across modern DL systems. These findings underscore that backdoor attacks introduce a pervasive and stealthy security risk, posing a critical challenge to the reliability, robustness, and safety of contemporary machine learning.

As QNNs evolve towards practical applications, analogous security vulnerabilities have begun to emerge. As demonstrated by Chu et al. [12], Guo et al. [13], and Zhang et al. [14], quantum backdoor attack methods, including QTrojan, HQNN-Backdoor, and QDoor, illustrate that adversaries can implant covert triggers at the quantum data level or within variational circuit components. Consequently, this results in QNNs behaving normally on clean inputs but yielding malicious predictions when exposed to the trigger. The QuanTest framework [15] further highlights the fragility of quantum machine learning (QML) [16] systems by systematically evaluating robustness issues.

Despite recent progress in adapting classical backdoor defenses to quantum settings, existing approaches remain insufficient to address the fundamental challenges posed by QNNs. Classical methods such as activation clustering [17] identify backdoors by isolating anomalous clusters within the statistical distribution of neuron activations, while Neural Cleanse [18] leverages the abnormal sensitivity of compromised models to reverse-engineered triggers to expose malicious behavior.

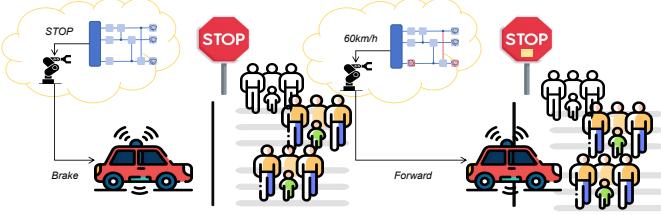


Fig. 1: The threat of backdoor attacks in the field of autonomous driving.

More recent efforts, such as Q-Detection [19], seek to translate these principles into the quantum domain. These endeavors employ a quantum-weighted distribution network and a two-layer Quadratic unconstrained binary optimization [20] strategy to detect distributional shifts between poisoned and clean data. However, despite these advances, extant defense methods still cannot completely overcome the unobservability of intermediate quantum states in quantum models and the fundamental limitations imposed by the quantum measurement assumption. Consequently, the development of effective, and genuinely applicable backdoor defense schemes for QNNs remains a pivotal research challenge.

In this study, we introduce QSentry, a quantum clustering framework based on measurements that is designed to detect backdoors in QNNs. This framework identifies activation differences generated by the quantum measurement layer, where backdoor samples and normal samples exhibit distinct statistical characteristics. Backdoor samples cause classification errors by combining triggering patterns with the inherent features of the source class. In contrast, normal samples rely solely on inherent features of the target class. The QSentry identifies backdoor inputs by analyzing the statistical properties of measurement outcomes, recognizing them as anomalous distributions that deviate from the dense clusters formed by normal samples.

Our paper makes the following contributions to the defense against backdoors in QNNs:

- We propose a QSentry framework for detecting quantum backdoor attacks. It overcomes the inherent limitation of the unobservability of intermediate quantum states in QNNs, providing a feasible method for constructing practical QNN defense.
- We designed a Measurement Clustering methodology that analyzes the results of the quantum measurement layer and extracts statistical features to detect backdoors in QNNs. Our method can simultaneously cover both data and model-level attack scenarios.
- The proposed technique has been validated on the MNIST dataset using a QNN against four benchmark backdoor attacks, QSentry achieves high F1 scores even at a low 1% poisoning rate. Comparative experiments with other state-of-the-art detection methods demonstrate that QSentry outperforms these methods in both robustness and accuracy.

II. BACKGROUND

A. Preliminary of Quantum Computing

1) *Qubits*: Qubits are the fundamental information carriers in quantum computing, serving as the quantum analog of classical bits. Unlike classical bits restricted to binary states $\{0, 1\}$, a qubit resides in a two-dimensional Hilbert space \mathbb{C}^2 and can be expressed as a superposition of the computational basis states:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad (1)$$

where $\alpha, \beta \in \mathbb{C}$ are probability amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$. For a composite system of Q qubits, the global state is represented by a unit vector in a 2^Q -dimensional Hilbert space formed via tensor products of the individual subsystems. This tensor structure enables entanglement, a key resource that distinguishes quantum from classical computation.

2) *Quantum Gates*: Quantum gates describe the evolution of qubits through unitary transformations. A gate acting on n qubits is represented by a $2^Q \times 2^Q$ unitary matrix U that satisfies:

$$U^\dagger U = UU^\dagger = I, \quad (2)$$

where U^\dagger denotes the conjugate transpose of U and I is the identity. Unitarity ensures reversibility and preserves the norm of quantum states. Common quantum gates include fixed single-qubit gates, parameterized rotational gates, and multi-qubit controlled operations, which together form universal gate sets for quantum computation.

3) *Quantum Measurement*: Quantum measurement is a non-unitary and irreversible process that projects a quantum state onto an eigenbasis of a Hermitian observable M . The outcome is intrinsically probabilistic, with statistics governed by the Born rule, often requiring repeated measurements to obtain reliable expectation values [21], [22]. In QNNs, model outputs are typically defined using the expectation values of selected observables, making measurement outcomes functionally analogous to activation values in classical neural networks [23].

B. Quantum Neural Networks

Quantum neural networks (QNNs) are variational models constructed from parameterized quantum circuits [1]. They follow a hybrid quantum-classical learning paradigm, in which the circuit parameters are optimized through classical algorithms while the quantum processor evaluates the cost function by executing the quantum circuit [24], [25].

To highlight the structural and computational differences between QNNs and classical DNNs, Figure 2 presents a side-by-side comparison. In a classical DNN pipeline, input features are directly processed through layers such as convolution, pooling, and fully connected modules, with gradients computed and applied entirely on classical hardware. In contrast, a QNN first requires mapping classical data x into a quantum state $|x\rangle$ via an encoding circuit. This encoded state

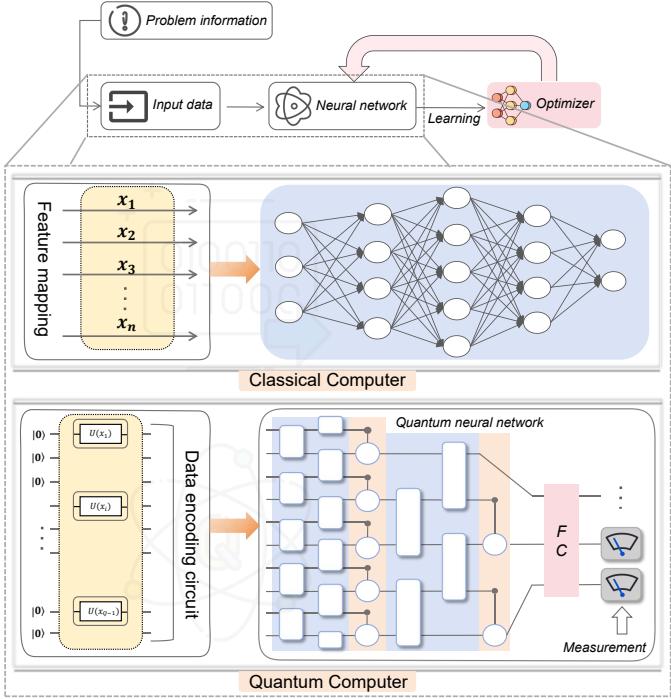


Fig. 2: Differences between QNN and classical DNN during the training process.

is then transformed by a parameterized quantum circuit U_{Θ} , producing an output quantum state:

$$|\psi_{\text{out}}\rangle = U_{\Theta}|\mathbf{x}\rangle. \quad (3)$$

Model predictions are obtained by measuring selected observables [26], and the resulting classical outcomes are used by a classical optimizer to update Θ iteratively.

Although the learning loop resembles that of classical networks, QNN architectures are fundamentally constrained by the number of available qubits and the noise limitations of NISQ hardware. These restrictions have motivated the design of efficient architectures, including quantum convolutional neural networks (QCNNs) [27] and quantum recurrent neural networks [28]. While these architectures improve efficiency, the inherent limitations of NISQ devices make quantum circuits particularly vulnerable to attacks [29], posing a significant challenge to the security of QML.

C. Backdoor Attack

Backdoors refer to hidden malicious behaviors implanted into machine learning models, enabling targeted misclassification when inputs contain attacker-designed triggers. A backdoored model behaves normally on benign samples, preserving high accuracy during evaluation while activating the malicious behavior only under specific trigger conditions [30].

In classical deep learning, backdoor attacks typically poison training data by embedding triggers and relabeling samples to a target class [31]. Recent advances reveal that backdoors can also be injected during self-supervised pretraining, propagating through downstream tasks without compromising clean

accuracy [32]. Attack variants include sample-specific, clean-label, and multi-trigger methods [33], [34]. Once the model learns the association between the trigger and the attacker-specified label, malicious behavior persists even after standard model verification. These attacks typically maintain high clean accuracy to evade detection, making backdoor identification a persistent challenge [35].

Quantum backdoor attacks represent an emerging security threat in QML, extending classical backdoor paradigms to QNNs by exploiting fundamental quantum properties such as superposition and entanglement [13], [14]. As shown in Figure 3, the attack methodology involves adversaries designing covert quantum triggers to poison training datasets, subsequently training QNNs to yield compromised models [12]. These backdoored QNNs are then delivered to end-users, with backdoors potentially implanted during outsourced model development or added post-training before distribution [30].

Recent research has demonstrated two primary attack vectors: data poisoning through maliciously crafted quantum states [13] and parameter tampering in variational quantum circuits [12], [14]. The compromised models maintain expected performance on benign inputs while exhibiting targeted misclassifications only when adversary-specified trigger patterns are present [13], [35]. Attackers typically compromise specific labels while maintaining majority uninfected labels to preserve stealth, and may employ single or multiple triggers based on operational requirements [31], [33], [34].

D. Backdoor Defense

Backdoor defense comprises techniques for detecting, mitigating, or eliminating hidden threats in machine learning models, addressing critical security risks in sensitive applications. Compromised models exhibit normal behavior under regular conditions but become controllable risk vectors when specific triggers activate targeted misclassification or malicious actions. These threats are particularly severe when models originate from untrusted third parties or distributed training environments where full auditability is limited.

Classical defense approaches encompass a variety of techniques, such as *anomaly detection* via activation pattern analysis [30], *neural cleansing* through trigger reverse-engineering [18], and *model reconstruction* using pruning and fine-tuning [36], input perturbation methods like STRIP for runtime trigger suppression [37], and spectral signatures for poisoned data identification [38]. These methods typically assume access to observable intermediate representations, clean validation data, and direct model internal access.

However, such defenses face fundamental incompatibilities with QNNs due to intrinsic quantum properties. For instance, quantum circuit intermediate states cannot be monitored without collapsing the superposition, which prevents activation-based detection. Quantum measurement statistics introduce inherent noise that obscures deterministic patterns required for gradient-based reverse engineering. High quantum computation costs severely restrict available test samples, undermining data-intensive defense strategies.

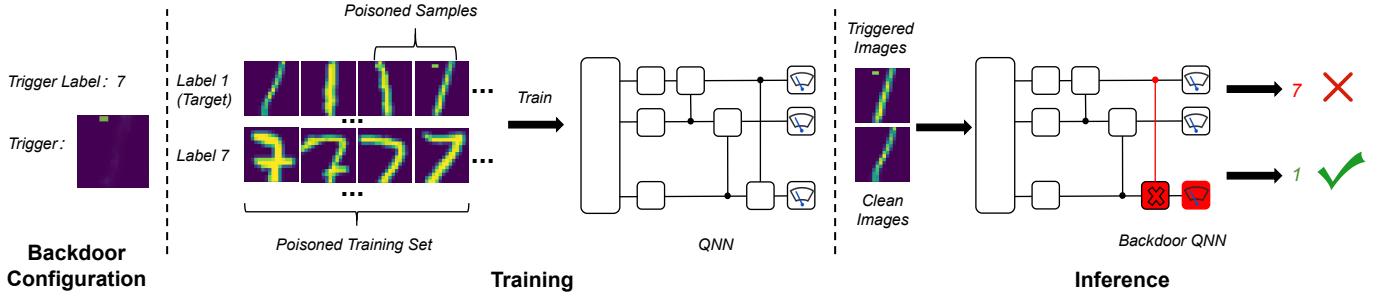


Fig. 3: An illustration of the quantum backdoor attack. The backdoor target is label 7, with the trigger pattern being the square in the upper left corner. When injecting the backdoor, some samples in the training set are modified to carry the trigger mark, and their labels are also altered to the target label. After training on the modified training set, the model will recognize samples bearing the trigger mark as the target label. Meanwhile, for any sample without the trigger mark, the model remains capable of correctly identifying its label.

III. THE MEASUREMENT CLUSTERING METHODOLOGY

Through theoretical analysis, we argue that the distributional perturbations introduced by backdoor attacks, such as data poisoning or malicious circuit embedding, should theoretically induce significant discrepancies in the model's measurement outcomes. In order to validate the aforementioned hypothesis, experimental observations were made regarding systematic deviations between the activation patterns of backdoor and clean samples in the measurement space. These observations are illustrated in Figure 4. The activations of the quantum measurement layer, when projected onto the first two principal components, reveal a clear separation between backdoor and clean samples for both labels 6 and 7, indicating that backdoor perturbations introduce distinct distributional shifts in the measurement space.

Building on this finding, this work introduces the Measurement Clustering method, an activation-space analysis approach specifically designed for backdoor detection in QNNs. Unlike conventional defense techniques that rely on hidden-layer neuron activations, this method operates entirely on the statistical information derived from the quantum measurement layer. By transforming measurement statistics into discriminative features and exploiting their intrinsic clustering structure, it effectively identifies and localizes statistical anomalies within the measurement space, thereby enabling the detection of potential backdoor samples.

A. Quantum Measurement Extraction

As outlined in Step 1 of Algorithm 1 (lines 3–7), each input sample is processed by a QNN to extract quantum measurement probabilities. For an input x_i , the QNN produces a quantum output state $|\psi^{(i)}(\theta)\rangle$, which is measured using a set of observables $\{m_i\}_{i=0}^{Q-1}$, where Q is the number of qubits. In this work, the measurement is performed on the $Pauli - Z$ basis for each qubit, and the resulting output corresponds to the expectation values of these observables.

$$m_i = [\langle Z_1 \rangle_i, \langle Z_2 \rangle_i, \dots, \langle Z_N \rangle_i], \quad (4)$$

These expectation values capture the statistical characteristics of the quantum measurement layer and reflect the model's response to both clean and backdoor inputs. Collecting N such samples yields the measurement activation matrix:

$$\mathbf{A} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_N \end{bmatrix} \in \mathbb{R}^{N \times Q}, \quad (5)$$

This measurement activation matrix forms the basis of our security analysis. It encodes the statistical perturbations in the QNN's measurement distribution caused by the backdoor, thus enabling the detection of malicious input.

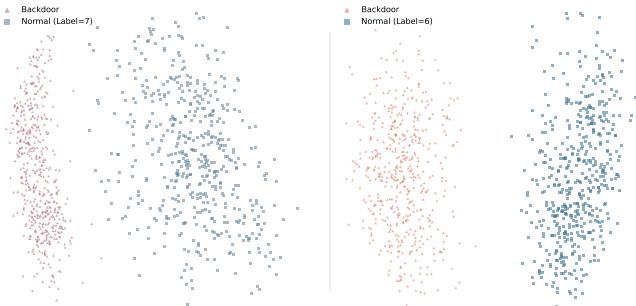


Fig. 4: The activation of the quantum measurement layer is projected onto the first two principal components: This is the measurement activation of the images labeled 7 and 6.

B. Feature Transformation

Following the quantum measurement extraction, Step 2 of Algorithm 1 (lines 9–11) projects the measurement activation matrix into a low-dimensional discriminant space to enhance cluster separability and computational efficiency. This projection reduces the noise sensitivity of the original high-dimensional quantum measurement vectors and sharpens the

Algorithm 1 Measurement Clustering for Backdoor Detection in QNNs

Require: Trained QNN f_θ , test set $D = \{x_i\}_{i=1}^N$, measurement operators $\{M_j\}_{j=0}^{Q-1}$, decomposition function $\mathcal{T}(\cdot)$, clustering method $\mathcal{J}(\cdot)$

Ensure: Detected anomalous cluster $\mathcal{C}_{\text{anom}}$

- 1: **Step 1: Quantum Measurement Extraction**
 - 2: Initialize measurement activation matrix $\mathbf{A} \in \mathbb{R}^{N \times Q}$
 - 3: **for** each input sample $x_i \in D_t$ **do**
 - 4: Obtain quantum state: $|\psi^{(i)}\rangle = f_\theta(x_i)$
 - 5: Compute measurement vector:

$$m_i = [\langle Z_1 \rangle, \langle Z_2 \rangle, \dots, \langle Z_N \rangle]$$
 - 6: Append m_i to \mathbf{A}
 - 7: **end for**
 - 8: **Step 2: Feature Transformation**
 - 9: Project to low-dimensional space: $\mathbf{V} = \mathcal{T}(\mathbf{A})$
 - 10: {Employs ICA variants, or contrastive decompositions to isolate backdoor-induced biases}
 - 11: **Step 3: Unsupervised Clustering**
 - 12: Apply clustering: $\{\mathcal{J}_1, \dots, \mathcal{J}_K\} = \mathcal{J}(\mathbf{V}, K)$
 - 13: Estimate K via silhouette score or eigen-gap criterion
 - 14: Identify minority cluster: $\mathbf{c}_{\text{anom}} = \arg \min_{\mathbf{c}_k} |\mathcal{J}_k|$
 - 15: **Output:** Return \mathbf{c}_{anom} as detected anomalous cluster
-

contrast between clean and backdoor distributions. Subsequently, statistical decomposition techniques, such as Independent Component Analysis (ICA) [39], [40], Principal Component Analysis (PCA) [41] variants, and other contrastive transformations, are applied to concentrate the non-Gaussian bias introduced by the backdoor into a compact activation representation. This process facilitates the identification of potential backdoor samples.

$$\mathbf{V} = \mathcal{T}(\mathbf{A}), \quad \mathbf{V} \in \mathbb{R}^{M \times d}, \quad d \leq Q. \quad (6)$$

The resulting representation \mathbf{Z} effectively suppresses noise and redundancy, producing a more discriminative feature space for subsequent clustering analysis. Clean samples form dominant clusters, whereas backdoor samples emerge as minority clusters characterized by anomalous activation geometry. These minority clusters are ultimately identified as suspicious subsets for backdoor detection.

C. Unsupervised backdoor detection

The final detection phase, implemented in Step 3 of Algorithm 1 (lines 13-16), applies unsupervised clustering to the transformed features. The reduced feature matrix \mathbf{V} is partitioned using K-Means [42] clustering, aiming to separate the dominant clean distribution from the anomalous backdoor

samples. The algorithm minimizes the within-cluster sum of squares:

$$\mathcal{J} = \sum_{k=1}^3 \sum_{i=0}^{Q-1} u_{ik} \|\mathbf{v}_i - \mathbf{c}_k\|^2, \quad (7)$$

where $u_{ik} \in \{0, 1\}$ is a cluster-assignment indicator, and $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_k\}$ are the cluster centroids. Optimization is performed iteratively via Lloyd's algorithm [43], initialized with k-means for robust convergence.

In the final partitioning results, the majority of clusters were identified as clean clusters, while a minority of clusters were marked as backdoor samples. This fully unsupervised method requires no prior knowledge of the poisoning ratio and can be generalized to both datasets and model-level attacks. Since the method is based solely on post-measurement statistics, it is independent of any specific QNN architecture.

IV. QSENTRY: A DEFENSE FRAMEWORK BASED ON MEASUREMENT CLUSTERING

This section presents QSentry, a unified framework for detecting backdoor attacks in QNNs. As illustrated in Figure 5, QSentry integrates a post-training anomaly backdoor detection module based on Measurement Clustering, designed to identify both dataset and model-level backdoors without prior knowledge of the trigger pattern or poisoning strategy.

A. Threat Model

The threats posed to QNNs by backdoor attacks emanate from two primary attack vectors. The vectors in question represent distinct attack surfaces. One vector operates at the classical datasets, while the other manipulates the quantum circuit substrate directly. Both vectors have the capacity to embed hidden malicious functionality while preserving the model's performance on legitimate inputs.

1) *Data Poisoning Attack*: In this scenario, an adversary injects a trigger pattern τ into a subset of training samples, generating backdoor inputs $x^\tau = T(x; \tau)$ and assigning them a target label y^τ . The resulting poisoned dataset causes the model to learn a spurious correlation between the trigger and y^τ , while maintaining high accuracy on clean inputs to avoid suspicion.

$$D_{\text{poison}} = \{(x_i^\tau, y_i^\tau)\}, \quad (8)$$

2) *Quantum Circuit Attack*: Following the QTrojan paradigm [12], attackers can directly modify the parameterized circuitry of a QNN by inserting malicious gates after the quantum-encoded circuit. By adjusting the parameters, the output can be made to conform to a target state defined by the attacker. This modification directly embeds hidden backdoor behavior into the model's quantum architecture.

3) *Defender Assumptions*: QSentry operates under the assumption that the defender has access to the trained QNN parameters and a set of test samples that may contain backdoor inputs. The framework does not require knowledge of the poisoning rate, trigger characteristics, or adversary capabilities, making it applicable to real-world QML deployment settings.

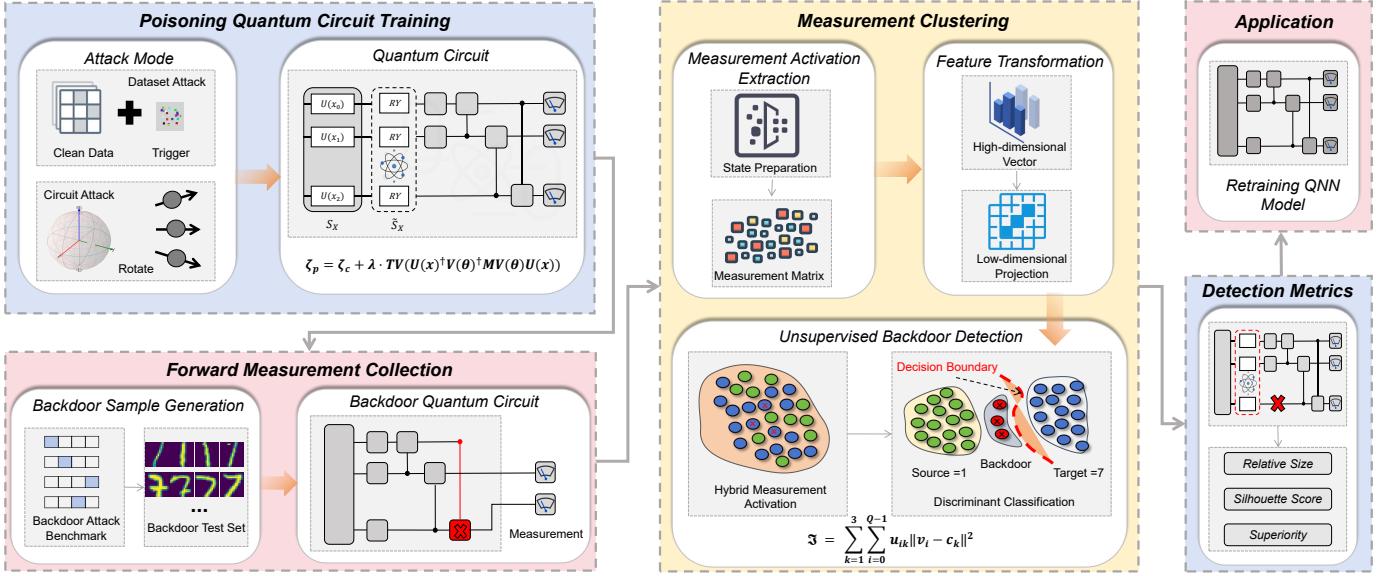


Fig. 5: QSentry defense framework architecture. The system integrates poisoning training, forward measurement collection and measurement clustering.

B. Defense Goal

The objective of QSentry is to detect backdoor attack anomalies in QNNs by identifying the abnormal disturbance they impose on qubit measurement statistics. These perturbations are both measurable and largely independent of the input class, enabling our defense to isolate backdoor samples directly in a transformed measurement space. Once identified, such samples can be removed. QSentry is attack-agnostic and requires no surrogate models, auxiliary classifiers, or access to internal quantum states.

C. QSentry System Defense Architecture

QSentry's operational framework is characterized by a three-stage pipeline, as illustrated in Figure 5. The objective of the design is to detect backdoors directly from observable quantum measurements. The workflow comprises: (1) *Poisoning Quantum Circuit Training*, which simulates adversarial conditions to create real threat models. (2) *Forward Measurement Collection* is used to collect measurement data from the quantum measurement layer, performed by the QNN. (3) *Measurement Clustering*, which identifies backdoor samples through statistical deviations. The first stage itself is not part of the defense, but is used to reproduce a realistic threat model in order to evaluate the detector.

D. Detailed Workflow

1) *Poisoning Quantum Circuit Training*: To evaluate QSentry under trusted threat conditions, we first train a QNN exhibiting backdoor behavior. For dataset-level attacks, the model is trained using a hybrid dataset containing both clean and backdoor samples. Training begins with a clean initialization to avoid bias and applies standard gradient-based optimization with cross-entropy loss. The resulting model

achieves high accuracy on clean inputs while associating trigger patterns with target labels, thus exhibiting typical backdoor characteristics (i.e., high clean accuracy and high attack success rate).

The training dataset is defined as:

$$D = D_{\text{clean}} \cup D_{\text{poison}},$$

where D_{clean} denotes clean samples without triggers, and D_{poison} denotes backdoor samples containing trigger patterns mapped to target labels. Typically, $|D_{\text{clean}}| \gg |D_{\text{poison}}|$ to maintain high accuracy.

For clean samples, the cross-entropy loss is:

$$\mathcal{L}_c(x, y; \Theta) = -\log P(y | x; \Theta), \quad (9)$$

where $P(y | x; \Theta)$ denotes the predicted probability for label y . For backdoor samples, the corresponding loss is:

$$\mathcal{L}_p(x^\tau, y^t; \Theta) = -\log P(y^t | x^\tau; \Theta), \quad (10)$$

where x^τ represents the backdoor of input x associated with a malicious target label $y^t \neq y$.

Parameter updates are performed via gradient descent:

$$\Theta \leftarrow \Theta - \eta \nabla_\Theta (\mathcal{L}_c + \mathcal{L}_p), \quad (11)$$

where η is the learning rate and Θ denotes the trainable parameters. This dual-objective optimization embeds backdoor functionality while maintaining high clean accuracy.

For *model-level* attacks, we adopt the QTrojan [12] paradigm, inserting malicious parameterized gates after the state encoding layer. The parameters of these gates are optimized to maximize the overlap between the output state of the triggered input and the target state while preserving the performance on clean inputs.

The training employs a composite loss function:

$$\zeta(\Theta) = \frac{1}{|D|} \left[\sum_i^{N_c} \zeta_c(x_i, y_i; \Theta) + \sum_j^{N_p} \zeta_p(x_j^\tau, y_j^t; \Theta) \right], \quad (12)$$

where $\zeta(\Theta)$ denotes the composite loss function for model training, $|D|$ represents the total number of samples in the training batch D , N_c is the number of samples in the clean sample subset D_{clean} . N_p is the number of samples in the backdoor sample subset D_{poison} . The poisoned loss includes a regularization term:

$$\zeta_p(x^\tau, y^t; \Theta) = \zeta_c(x, y; \Theta) + \lambda \cdot TV(U(x)^\dagger V(\Theta)^\dagger M V(\Theta) U(x)), \quad (13)$$

where λ is a regularization coefficient and $TV(\cdot)$ denotes the total variation term used to suppress anomalous measurement distributions caused by triggers. Here, $U(x)$ represents the state encoding unitary, $V(\Theta)$ the parameterized (possibly malicious) variational circuit, and M the measurement operator. This regularization ensures the backdoor remains concealed while maintaining clean accuracy.

This regularization suppresses anomalous measurement distributions caused by triggering, ensuring the backdoor remains concealed. The resulting model achieves high accuracy and attack success rate, establishing a reliable benchmark for detection experiments.

2) *Forward Measurement Collection*: Given a trained backdoored QNN and a set of test inputs, QSentry performs forward propagation and records the measurement outcomes of each qubit. These outputs form a measurement activation matrix

$$\mathbf{A} \in \mathbb{R}^{M \times Q},$$

where each row corresponds to one input sample, and each column corresponds to a qubit measurement outcome expectation value. Here, M denotes the number of test samples and Q the number of qubits. Due to the no-cloning theorem, the final measurement represents the only observable in the computation within the QNN, thus ensuring that our defenses are entirely based on the obtained data.

3) *Measurement Clustering for Anomaly Discovery*: This detection step employs unsupervised clustering of the quantum measurement matrix to identify anomalies caused by backdoors. By analyzing the reduced feature space, a compact feature representation is obtained that amplifies systematic deviations introduced by the backdoor. Next, clustering (e.g., K-Means [42]) is applied to partition the samples. A minority cluster exhibiting anomalous geometry and measurement statistics is identified as the backdoor cluster. The method isolates minority clusters that display deviant measurement statistics and geometric properties, which are characteristic of backdoor samples. This approach is attack-agnostic, requires no prior knowledge of the trigger pattern or circuit internals, and operates directly on measurement outcomes.

Using the Measurement Clustering method described in Chapter III, quantum measurement statistics are transformed into discriminative features for backdoor identification. The

Algorithm 2 QSentry Backdoor Detection Framework Workflow

Require: Trained QNN f_θ , clean dataset D_c , test set D_t , backdoor dataset D_p

Ensure: Detected backdoor samples in anomalous cluster c_{anom}

1: **Step 1: Poisoning Training**

2: $w \leftarrow w_0$ {Initialize with clean model}

3: **if** attack type = model-level **then**

4: $\theta = 2 \cdot \arccos(|\langle \psi_0 | \psi_1 \rangle|)$

5: $|\psi_0\rangle \leftarrow$ state after encoding

6: $|\psi_1\rangle \leftarrow$ target state

7: **end if**

8: **repeat**

9: $(x, y) \leftarrow$ random sample from $D_c \cup D_p$

10: **if** $(x, y) \in D_p$ **then**

11: **if** attack type = model-level **then**

12: Add $R_Y(\theta)$ after each qubit encoding layer

13: **end if**

14: $w_p \leftarrow \arg \min_w \zeta_p(x^p, y, w)$

15: **else**

16: $w_c \leftarrow \arg \min_w \zeta_c(x, y, w)$

17: **end if**

18: **until** convergence

19: **Step 2: Measurement Collection**

20: **for** each sample x in D_t **do**

21: Run forward pass through f_θ .

22: Record qubit measurement vector m_i .

23: **end for**

24: Construct measurement activation matrix $\mathbf{A} = [m_i]$.

25: **Step 3: Measurement Clustering**

26: $\mathbf{V} \leftarrow \text{reduce}(\mathbf{A})$ {dimension reduction}

27: $\mathcal{C} \leftarrow \text{cluster}(\mathbf{V})$ {unsupervised partitioning}

28: Identify anomalous cluster c_{anom}

29: **Output:** Return $\mathcal{C}_{\text{anom}}$ as detected backdoor samples

detection process begins with quantum forward propagation, where each input sample x_i passes through the QNN to produce an output quantum state $|\psi^{(i)}(\theta)\rangle$. Quantum measurements are then performed to obtain the expectation value of each qubit's measurement observable:

$$m(i, q) = \langle \psi^{(i)}(\theta) | M_q | \psi^{(i)}(\theta) \rangle, \quad (14)$$

These expectation values form the measurement activation matrix $\mathbf{A} \in \mathbb{R}^{M \times Q}$, where each row corresponds to one test sample and each column represents the expectation value of a qubit measurement. This matrix captures the statistical behavior of the QNN under different inputs and serves as the basis for subsequent clustering analysis. The high-dimensional matrix then undergoes dimensionality reduction via ICA:

$$\mathbf{V}_{\text{reduced}} = \mathbf{A} \mathbf{W}^* \quad (15)$$

where \mathbf{W}^* is the ICA-derived projection matrix and clustering is performed via the K-Means algorithm:

$$\{\mathbf{c}_{\text{clean}}, \mathbf{c}_{\text{backdoor}}\} = \text{K-Means}(\mathbf{V}_{\text{reduced}}), \quad (16)$$

where $\mathbf{c}_{\text{backdoor}}$ typically forms a minority cluster characterized by distinct measurement statistics. This unsupervised approach is attack-agnostic, requires no prior knowledge of the trigger or circuit internals, and operates directly on measurement outcomes.

E. Algorithm Description

To complement the system workflow described in Section IV-D, we provide a formal specification of the QSentry detection process. Algorithm 2 summarizes the complete procedure, including quantum measurement extraction, forward measurement collection, and measurement clustering. This algorithm embodies the minimal-assumption principle of the framework and serves as a reproducible template for practical deployment.

Detection will be triggered when all preset separation metrics exceed their thresholds and a few clusters exhibit characteristics consistent with the expected backdoor. The integration of classical clustering techniques with quantum specificity measurement analysis ensures the method’s scalability and adaptability across diverse attack scenarios, establishing Measurement Clustering as a reliable tool for post-training security evaluation of QML systems.

V. EXPERIMENTAL VERIFICATION OF BACKDOOR DEFENSE

A. Experiment Setup

1) Dataset: The MNIST [44] handwritten digit dataset is a classic dataset widely used in the field of DL and has become a standard benchmark for classification tasks. This dataset contains 60,000 training samples and 10,000 test samples. Each digit is stored as a 28×28 pixel grayscale image. For our experiments, we utilize a subset of MNIST tailored for binary classification (denoted as MNIST[1,7]). The selection of the dataset and the corresponding classification tasks took into account both the limitations of currently available qubits and the classification task settings widely used in the field of QML.

2) QNN Models: Quantum Circuit Learning (QCL) [45] is a classical-quantum hybrid framework that utilizes QNNs. It improves performance on tasks such as high-dimensional regression or classification by non-linearly encoding classical data and using shallow, trainable variational circuits. In this work, we implement an 8-qubit circuit with a depth of eight layers. Each layer is composed of parameterized single-qubit rotations (R_X, R_Z, R_X) followed by a circular entangling gate. The output of the quantum circuit is measured by the expected values of $\langle Z_0 \rangle$ and $\langle Z_1 \rangle$ input to the Softmax function.

3) Data Encoding: The operation of QNNs on classical data requires a preliminary step of encoding the data into quantum states. This quantum state encoding functions as a feature map that transforms data from \mathbb{R}^Q into a state within a 2^q -dimensional Hilbert space (\mathbb{C}^{2^q}). To enable efficient classical simulation of more qubits, we adopt amplitude encoding in this work. This method encodes a classical vector $\mathbf{x} \in \mathbb{R}^N$ into the amplitudes of a quantum state over $\lceil \log_2 N \rceil$ qubits, formalized as $|\mathbf{x}\rangle = \sum_{i=1}^N x_i |i\rangle$, where $|i\rangle$ denotes the computational basis. A fundamental prerequisite for this encoding is that the input data must be normalized, satisfying $|\mathbf{x}|^2 = \sum_i |x_i|^2 = 1$.

In the MNIST dataset, each image is a 28×28 pixel matrix. Direct amplitude encoding of such raw images requires 1024 amplitude values, equivalent to a 10-qubit system. This increases computational overhead and noise sensitivity, thus degrading performance. Therefore, we introduce a preprocessing step to crop the image size to 16×16 while minimizing the loss of key data features. After this processing, encoding requires only 256 amplitude values in an 8-qubit system, thereby reducing resource requirements and noise impact while maintaining feature validity.

4) Backdoor Samples Generation: In the experimental setup for backdoor sample generation, we evaluated four representative strategies, all of which follow a unified paradigm of fixed target labeling and location injection, as shown in Figure 6. Specifically, these include: (a) **Patch Trigger Attacks**, which randomly inject pixel block perturbations into high-contrast color regions. (b) **Blend Trigger Attack**, where target class samples are blended with the original samples using coefficients $\sigma = 1.5$ and $\lambda = 0.3$ to achieve progressive contamination in the frequency domain. (c) **Sinusoidal Trigger Attack**, which overlays a Gaussian-filtered standard trigger template with a mixing coefficient defined by angle 0.2 and frequency 1. (d) **QTrojan Circuit Attack**, where a parameterized malicious subcircuit is implanted after the quantum encoding layer, constructing the trigger using rotated $R_Y(\theta)$ gates. Backdoor samples for each attack type are generated independently based on different poisoning rates.

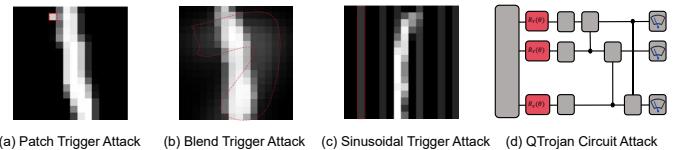


Fig. 6: Examples of different backdoor attacks. (a) Patch Trigger Attack, (b) Blend Trigger Attack, and (c) Sinusoidal Trigger Attack are three-data attacks, (d) QTrojan Circuit Attack is a quantum model-level attack.

5) Evaluation Metrics: To comprehensively assess both the effectiveness of backdoor implantation and the performance of detection mechanisms, the experiments employ a set of evaluation metrics.

Clean Accuracy (CA) measures the classification performance of a poisoned model on the original test set without implanted triggers, thus reflecting the stealth of the attack. Its definition is as follows:

$$CA = \frac{1}{N_{\text{clean}}} \sum_{i=1}^{N_{\text{clean}}} \mathbb{I}(\hat{y}_i = y_i), \quad (17)$$

where N_{clean} is the number of clean test samples, and \hat{y}_i and y_i denote the predicted and true labels, respectively.

Attack Success Rate (ASR) measures the effectiveness of a model in predicting the attacker's target label on test samples containing the trigger after the trigger has been successfully implanted. It is defined as follows:

$$ASR = \frac{1}{N_{\text{backdoor}}} \sum_{i=1}^{N_{\text{backdoor}}} \mathbb{I}(\hat{y}_i = y_{\text{target}}), \quad (18)$$

where N_{backdoor} is the number of backdoor test samples and y_{target} is the target-class label enforced by the attacker.

Silhouette Coefficient (SC) evaluates the geometric separability of clusters in the reduced measurement-feature space, providing an indicator of how distinguishable backdoor samples are from clean ones:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (19)$$

where $a(i)$ is the average intra-cluster distance of sample i , and $b(i)$ is the minimum average distance between sample i and all points in any other cluster. The overall score is:

$$SC = \frac{1}{N} \sum_{i=1}^M s(i), \quad (20)$$

with N being the total number of samples. Higher SC values indicate strong cluster separability.

Relative Cluster Size (RCS) reflects the proportion of samples assigned to the minority cluster, which is expected to correspond to backdoor samples. It measures whether the clustering result matches the theoretical poisoning ratio:

$$RCS = \frac{|\mathcal{C}_{\text{minority}}|}{|\mathcal{C}_{\text{clean}}| + |\mathcal{C}_{\text{backdoor}}|}, \quad (21)$$

where $\mathcal{C}_{\text{minority}}$ denotes the cluster with the smaller number of samples. An RCS close to the true poisoning rate indicates consistent backdoor cluster structure.

Detection Accuracy (DA) measures how well the clustering-based detection method correctly distinguishes backdoor and clean samples:

$$DA = \frac{TP + TN}{TP + TN + FP + FN}, \quad (22)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

F1 score captures the balance between correct identification of backdoor samples and reduction of false alarms, making it

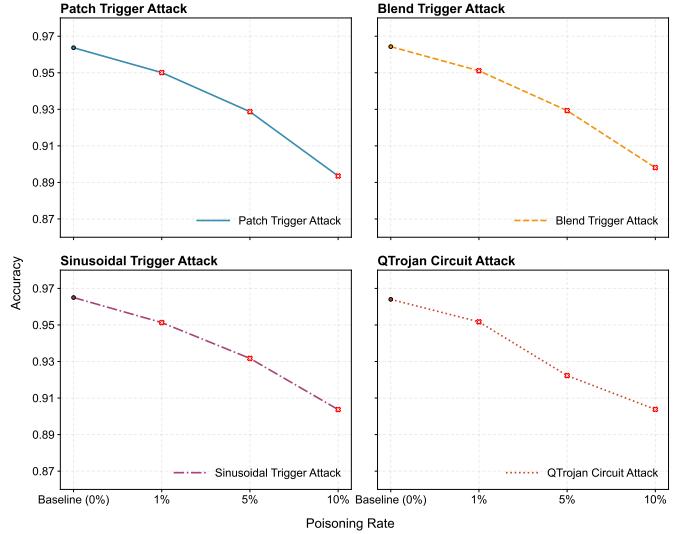


Fig. 7: Comparison of classification accuracy of the binary classification task MNIST[1,7] under four types of backdoor attacks and three poisoning rates.

suitable for evaluating detection tasks with minority backdoor samples:

$$F1 = 2 \cdot \frac{TP}{2TP + FP + FN}. \quad (23)$$

A higher F1 score indicates that the detection method effectively identifies backdoor samples without overestimating false alarms.

We implemented QSentry using the PennyLane framework, which provides a unified platform for constructing and training both classical and quantum machine learning models. All evaluation tasks were performed on a workstation equipped with an Intel(R) Core(TM) i9-14900K CPU, an NVIDIA GeForce RTX 4090 GPU, and 125 GB of RAM.

B. Construction and Benchmark Evaluation of QNN Backdoor Attack Threat Model

Research into security vulnerabilities in QML is still in its early stages. Backdoor attacks, as a covert and powerful threat, pose a serious risk to model integrity, and their potential risks have not been fully explored. To develop effective backdoor defense mechanisms, it is essential to construct and validate a realistic threat model that demonstrates the severity of such risks. If a defense method cannot withstand a real and effective attack, its evaluation results will lack persuasiveness. Therefore, this experiment systematically constructs a backdoor attack benchmark for QNNs. Our goal is to train a clean QNN model as a performance baseline and then attack this model using four different backdoor attack methods to verify the feasibility and effectiveness of backdoor attacks in QNNs, and to provide experimental evidence for subsequent defense research. All experiments are detailed and reproducible; experimental data were repeated 3-5 times and averaged to mitigate uncertainty.

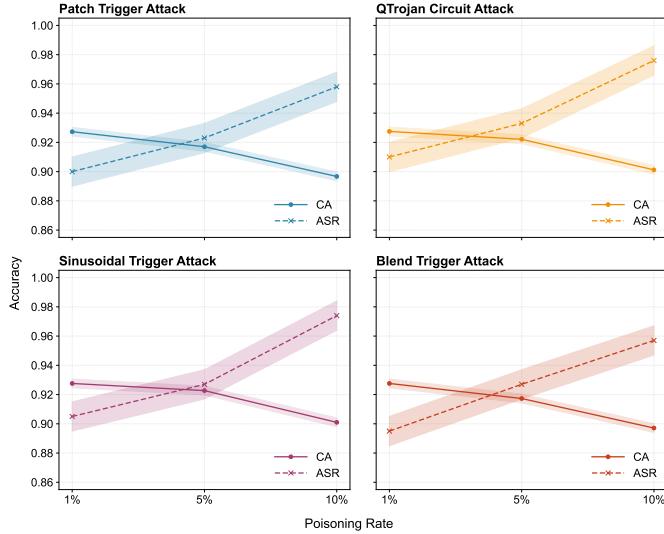


Fig. 8: Variation of CA and ASR under different poisoning ratios for four representative backdoor attacks (Patch Trigger Attack, Blend Trigger Attack, Sinusoidal Trigger Attack, and QTrojan Circuit Attack).

The performance trends in Figure 7 clearly demonstrate the effectiveness and stealthiness of the backdoor threats. The baseline QNN achieves the highest accuracy on the clean dataset, indicating its inherent performance advantage. At a low poisoning rate, all four attacks only cause a slight decrease in CA, highlighting their stealthiness. As the poisoning rate increases to 5% and 10%, the differences in the impact of different attacks become more pronounced, with most attacks reducing the overall accuracy to a lower level. **Crucially, the model maintains a relatively high classification accuracy even under severe malicious input conditions, which further highlights the stealth of the attack and the model's stability with clean input data.**

To further evaluate the threat model, we assessed the CA on clean samples and the ASR on backdoor samples of the poisoned model under different poisoning rates for each type of backdoor attack. As shown in Figure 8, the horizontal axis represents the poisoning ratio applied to the training data, and the vertical axis represents the corresponding CA and ASR. When the poisoning ratio increases from 1% to 10%, the CA of all four attacks remains at a high level, while the ASR gradually increases with the increase in the poisoning ratio. The experimental results show that at different poisoning rates, the CA of all attacks remains at a high level, indicating their strong stealthiness. The ASR increases rapidly with the increase in the poisoning ratio and stabilizes at a high level, indicating that backdoors can be effectively implanted and triggered without affecting the stability of the model.

This experiment systematically evaluates the effectiveness and stealth of four representative backdoor attacks on the employed QNN architecture. Experimental findings demonstrate that while maintaining high accuracy on clean test

samples, most attacks achieve a high ASR, and the ASR stabilizes as training converges. This outcome underscores the practical significance of the constructed backdoor threat scenario and highlights its potential to challenge the efficacy of future defense methodologies. In light of the findings from this experiment, the subsequent experiments in this paper will utilize these attacks as a baseline threat model. This approach will serve to verify the effectiveness of the proposed defense method.

C. Comprehensive Performance Evaluation of QSentry

This experiment evaluates the efficacy of our proposed QSentry defense framework against verified backdoor threats in QNNs. Under the scenario of a QNN trained on a contaminated dataset and a small, clean validation set, we compare QSentry with baseline methods. The objective is to validate its superior capability in extracting discriminative features from quantum states compared to classical raw-data features and to demonstrate its general defensive performance across diverse backdoor attacks.

The quantitative evaluation of detection performance is summarised in Table I. The results report both detection accuracy and F1 score across four representative backdoor attack modes and three poisoning rates. In the comparison, QSentry represents our proposed measurement-based defense framework, while Raw represents the baseline method that directly applies clustering to the raw pixel domain without utilizing quantum measurement information.

Across all attack settings, QSentry consistently and substantially outperforms the Raw baseline in terms of both accuracy and F1 score. It is worth noting that, because backdoor samples constitute an extremely small fraction of the dataset, accuracy alone becomes an unreliable indicator of detection performance—it may remain high even when backdoor samples are poorly identified. In contrast, the F1 score offers a more faithful characterization of defense effectiveness under severe class imbalance, capturing the essential trade-off between precision and recall. The significant improvement in the F1 score clearly demonstrates that backdoor-triggered samples exhibit unique and statistically separable activation characteristics in the quantum measurement space. This strongly suggests that backdoors are more easily detected via QSentry.

Furthermore, QSentry demonstrates strong robustness in detecting more stealthy backdoor strategies, including frequency-domain triggers and adaptive attacks crafted to exploit quantum circuit characteristics. **Overall, the empirical results validate QSentry as a general and effective defense mechanism for QNNs, capable of accurately isolating backdoor samples while maintaining high precision and high F1 scores.** Its performance consistently surpasses that of pixel-space baselines, underscoring the advantage of leveraging QSentry for reliable backdoor detection.

TABLE I: Detection performance of QSentry versus Raw Clustering under four attack modes and three poisoning rates. Bold entries indicate cases where QSentry outperforms the baseline on the corresponding metric.

Attack Mode	1% Poison Rate		5% Poison Rate		10% Poison Rate	
	QSentry	Raw	QSentry	Raw	QSentry	Raw
Patch Trigger	99.8% / 83.3%	98.4% / 61.4%	99.5% / 90.9%	97.9% / 70.3%	99.7% / 97.1%	97.8% / 78.4%
Blend Trigger	99.7% / 76.9%	97.9% / 64.1%	99.4% / 89.3%	98.4% / 70.4%	99.7% / 97.1%	98.9% / 79.1%
Sinusoidal Trigger	99.5% / 71.4%	97.8% / 62.7%	98.9% / 82.0%	96.7% / 70.0%	98.9% / 90.1%	95.8% / 78.9%
QTrojan Circuit	99.6% / 71.4%	98.4% / 62.3%	98.8% / 80.6%	96.1% / 71.6%	98.7% / 88.5%	94.6% / 78.1%

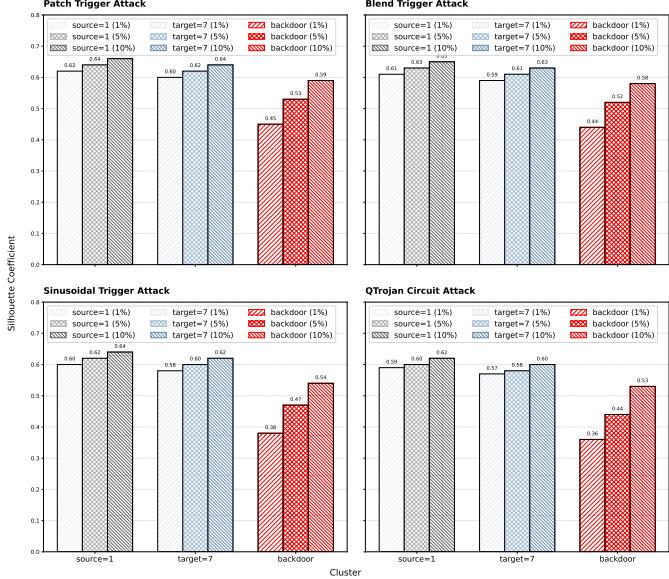


Fig. 9: Silhouette-based clustering analysis across four backdoor attack types—Patch Trigger, Blend Trigger, Sinusoidal Trigger, and QTrojan Circuit Attack—under poisoning rates of 1%, 5%, and 10%.

D. Backdoor Detection Verification Based on Silhouette Coefficient

This experiment aims to quantitatively verify the effectiveness and robustness of the QSentry defense framework in detecting backdoor attacks. We hypothesize that poisoned data trigger a bicluster distribution in the measurement space of the quantum model, while clean data conform to a monocluster distribution. To verify this hypothesis, we introduce the silhouette coefficient as the core indicator for quantifying distribution separation. By calculating this coefficient and comparing it with a predefined threshold, we systematically evaluate the defense mechanism's ability to distinguish between malicious and clean samples under various attack scenarios.

Across all four attack modes and poisoning ratios, clean source and target samples consistently maintain high silhouette values, indicating that benign measurement activations naturally form compact and coherent clusters within the quantum feature space. In contrast, backdoor samples exhibit substantially lower silhouette scores at all poisoning levels, with values dropping to the 0.36–0.45 range under a 1%

poisoning rate. As illustrated in Figure 9, these degraded scores confirm that backdoor triggers introduce identifiable geometric distortions to the measurement distribution, even when the contamination is extremely sparse.

As the poisoning rate increases to 5% and 10%, the silhouette coefficients increase moderately, reflecting that a larger proportion of backdoor inputs yields a more structurally coherent cluster. This trend is consistent with theoretical expectations: higher poisoning density produces stronger activation biases, amplifying the backdoor signature and improving cluster consistency within the backdoor subset. Importantly, this pattern remains consistent across visually obvious triggers, subtle frequency-domain triggers, and model-level manipulations, demonstrating that the quantum measurement space preserves separability against both dataset and model-level threats.

Overall, the silhouette-coefficient results validate the core hypothesis behind QSentry: **backdoor attacks imprint statistically distinct signatures on quantum measurement activations**. These signatures manifest as minority clusters that are reliably distinguishable from clean data. The consistency of this phenomenon across poisoning intensities and attack modalities confirms the robustness, generality, and practical viability of measurement-based anomaly detection in QNNs.

E. Backdoor Detection Verification Based on Cluster Relative Size

We found that when QNNs process data containing backdoor samples, the activation distribution in their measurement space typically exhibits a distinct bi-cluster structure. Further analysis revealed that 4, unlike the primary clusters formed by uncontaminated samples, backdoor samples usually aggregate into smaller secondary clusters. This phenomenon demonstrates relative stability under various attack methods.

This study proposes using the relative size of clusters as a standard for detecting backdoor samples to realize a backdoor detection mechanism in a quantum model. The objectives of this experiment are twofold: first, to verify whether backdoor samples consistently exhibit a relatively small cluster structure under different backdoor attack types and poisoning rates; and second, to evaluate whether setting a cluster size threshold close to a preset poisoning rate allows for stable identification and localization of backdoor clusters without requiring additional manipulation of the model structure or parameters. To this end, we systematically experimented to

TABLE II: Actual (Act) versus Predicted (Pred) QSentry clustering counts (total 1000 samples). $\Delta = \text{Pred} - \text{Act}$.

Poison Rate	Category	Patch Trigger Attack			Blend Trigger Attack			Sinusoidal Trigger Attack			QTrojan Circuit Attack		
		Act	Pred	Δ	Act	Pred	Δ	Act	Pred	Δ	Act	Pred	Δ
1%	Source 1	495	495	0	495	494	-1	495	495	0	495	493	-2
	Target 7	500	498	-2	500	498	-2	500	496	-4	500	498	-2
	Backdoor	5	7	+2	5	8	+3	5	9	+4	5	9	+4
5%	Source 1	475	475	0	475	474	-1	475	473	-2	475	472	-3
	Target 7	500	495	-5	500	495	-5	500	491	-9	500	491	-9
	Backdoor	25	30	+5	25	31	+6	25	36	+11	25	37	+12
10%	Source 1	450	450	0	450	450	0	450	449	-1	450	446	-4
	Target 7	500	497	-3	500	497	-3	500	490	-10	500	491	-9
	Backdoor	50	53	+3	50	53	+3	50	61	+11	50	63	+13

verify this criterion in terms of geometric separability and numerical consistency, combining visualization analysis and cluster statistics methods.

The experimental results demonstrate a high degree of consistency between the cluster structure and the relative size statistics. As illustrated in Figure 10, irrespective of the employed attack method, the activation space of the measured values is organized into three distinct clusters: two primary clusters comprising clean samples from the source and target classes, which collectively account for a substantial proportion; and a secondary cluster consisting of backdoor samples, which is comparatively diminutive in size and exhibits discernible boundary structures that delineate it from the primary clusters.

In addition, as demonstrated in Table II, the predicted cluster sizes exhibit a high degree of congruence with the preset poisoning rates. At a 10% poisoning rate, the predicted backdoor cluster sizes are between 52 and 63, which is very close to the ideal value of 50. At 5% and 1% poisoning rates, the predicted backdoor cluster sizes are between 30 and 37 and 7 and 9, respectively, which is also quite consistent with the actual proportions. While there is a slight overestimation under certain attack methods, the overall deviation remains within an acceptable range and will not affect the reliable identification of backdoor sample clusters. Visualization and tabular data confirmed the stability of the structural feature that **backdoor samples cluster into small clusters, while pure samples cluster into large clusters** under all experimental conditions.

The results demonstrate that the relative cluster size is a reliable indicator for backdoor detection, as the minority clusters consistently correspond to backdoor samples and closely match the preset poisoning ratios. This behavior remains stable across different attack intensities. Overall, both visual clustering structure and quantitative statistics validate the feasibility and practicality of the QSentry method in quantum backdoor defense, providing a solid foundation for constructing secure QNN protection mechanisms.

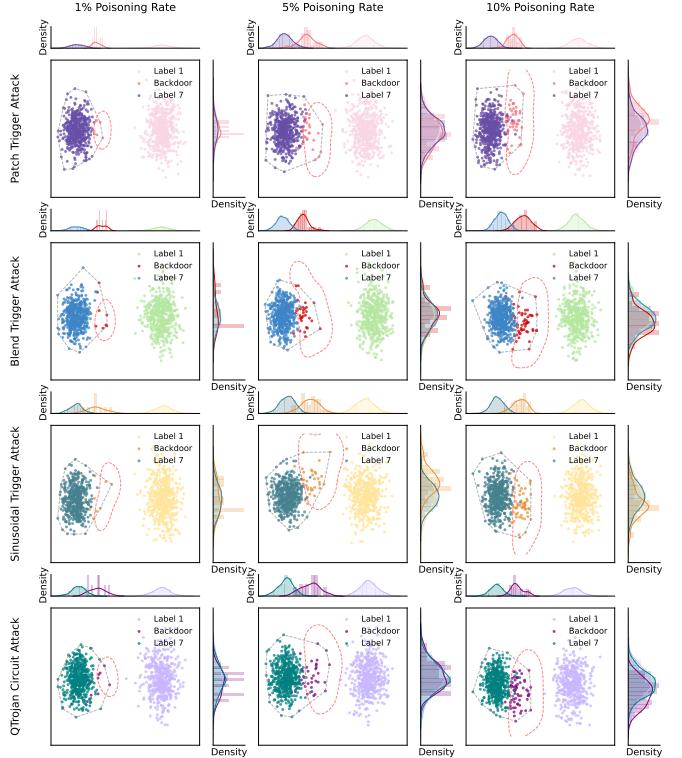


Fig. 10: Visualization of different clusters. Visual analysis of the sample distribution under the four attack methods and three infection rates reveals that the Backdoor sample cluster consistently exhibits spatially distinguishable characteristics from the two clean sample clusters.

F. Comparative Analysis of QSentry Against Advanced Detection Methods

To comprehensively evaluate the performance advantages of the QSentry defense framework, this experiment systematically compared it with three state-of-the-art backdoor detection methods: Identifying Backdoor Data with Optimized Scaled Prediction Consistency based on input consistency (MSPC) [46]; Amplifying Anomalies in Backdoor Models through Knowledge Distillation (Distill to Detect) [47], and quantum-

TABLE III: Detection Accuracy / F1 score comparison among MSPC, Distill to Detect, Q-Detection, and QSentry under three poisoning rates. Each cell reports “Accuracy / F1”. Smaller gaps are used to highlight subtle performance differences.

Poisoning Rate: 1%	MSPC	Distill to Detect	Q-Detection	QSentry (Ours)
Patch Triggered Attack	99.2% / 77.1%	99.0% / 77.8%	99.6% / 81.2%	99.8% / 83.3%
Blend Triggered Attack	99.0% / 73.9%	98.8% / 72.5%	99.5% / 74.8%	99.7% / 76.9%
Sinusoidal Triggered Attack	98.8% / 68.9%	98.6% / 68.2%	99.3% / 69.1%	99.5% / 71.4%
QTrojan Circuit Attack	98.9% / 67.6%	98.7% / 65.9%	99.4% / 69.8%	99.6% / 71.4%
Poisoning Rate: 5%	MSPC	Distill to Detect	Q-Detection	QSentry (Ours)
Patch Triggered Attack	98.9% / 85.9%	98.7% / 84.5%	99.3% / 88.5%	99.5% / 90.9%
Blend Triggered Attack	98.7% / 84.0%	98.5% / 83.7%	99.2% / 87.2%	99.4% / 89.3%
Sinusoidal Triggered Attack	98.1% / 76.4%	97.9% / 75.8%	98.7% / 80.6%	98.9% / 82.0%
QTrojan Circuit Attack	97.9% / 76.3%	97.7% / 75.6%	98.6% / 78.2%	98.8% / 80.6%
Poisoning Rate: 10%	MSPC	Distill to Detect	Q-Detection	QSentry (Ours)
Patch Triggered Attack	99.1% / 90.9%	98.9% / 90.2%	99.4% / 94.8%	99.7% / 97.1%
Blend Triggered Attack	99.0% / 93.7%	98.7% / 94.3%	99.3% / 95.6%	99.7% / 97.1%
Sinusoidal Triggered Attack	97.8% / 85.0%	97.5% / 84.8%	98.4% / 88.5%	98.9% / 90.1%
QTrojan Circuit Attack	97.4% / 83.1%	97.2% / 82.6%	98.2% / 86.9%	98.7% / 88.5%

specific Q-Detection [19]. The evaluation aimed to quantify the detection capabilities of these four methods against various backdoor attack types in a unified test environment and verify the overall superiority of QSentry in terms of detection effectiveness.

Table III provides a comprehensive comparison of the performance of four representative backdoor detection methods: MSPC, Distill to Detect, Q-Detection, and QSentry under four attack modes and three poisoning rates. QSentry maintains the highest F1 score and detection accuracy under all settings, demonstrating its superior ability to identify backdoor samples, even at extremely low poisoning rates. This improvement is especially notable at low poisoning levels, where traditional metrics like accuracy are unreliable due to class imbalance. In contrast, QSentry maintains a high F1 score, indicating robust detection sensitivity.

Q-Detection ranks second overall, closely following QSentry, performing well across all attack modes and poisoning levels. Its detection capability steadily improves with increasing poisoning rate, but remains slightly inferior to QSentry, suggesting that the QSentry method is more effective than Q-Detection’s hybrid quantum-classical representation in capturing anomalies caused by backdoors in QNNs. The results for MSPC and Distill to Detect are similar, both outperforming Raw Clustering but lagging behind QSentry and Q-Detection. Their accuracy remains relatively high, but their F1 scores are significantly lower, especially under conditions of subtle or widely distributed trigger signals, indicating insufficient robustness in distinguishing between sparse backdoor and clean samples. The performance of both methods improves slightly with increasing poisoning rate, but the performance gap with QSentry remains significant.

Overall, the results confirm that **QSentry achieves the most balanced and reliable detection performance across all attack modes and poisoning intensities. It significantly improves the F1 score while maintaining high accuracy,**

making it a practical and effective framework for backdoor threat detection in QML.

VI. CONCLUSION

QSentry provides a practical and effective defense framework against backdoor threats in QNNs. Unlike methods that rely on inaccessible intermediate quantum states, QSentry utilizes measurement layer activation statistics, bypassing the fundamental limitation of quantum observability. Extensive experiments under various dataset and model-level attacks have confirmed that the quantum measurement distribution contains stable and detectable structural anomalies, enabling QSentry to accurately identify backdoor samples even with poisoning rates as low as 1%. Compared to existing classical and hybrid defense methods, this framework consistently achieves higher detection precision.

Despite its strengths, QSentry has several limitations. First, its performance relies on a sufficient number of measurement samples, which can introduce significant runtime overhead on real devices. Additionally, the framework has primarily been evaluated on small-scale circuits and binary classification tasks, so its scalability to higher-dimensional feature spaces and deeper variational architectures remains to be verified. Finally, if attackers design adaptive triggers to evade detection by minimizing anomalies in the measurement space, this poses a challenge to current strategies.

VII. ETHICS CONSIDERATIONS

This study did not involve human subjects, personal data, or interactions with deployed systems. All experiments were conducted in a controlled environment, using publicly available datasets and simulated quantum circuits based on the PennyLane framework. Therefore, no ethical issues were identified.

VIII. LLM USAGE CONSIDERATIONS

In the preparation of this manuscript, the authors utilized large language models (LLMs) such as ChatGPT for assistance

with text polishing and grammatical correction only. The fundamental research ideas, formulation of the methodology, experimental design, execution, data interpretation, and scientific conclusions were solely and independently carried out by the authors.

REFERENCES

- [1] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, “Parameterized quantum circuits as machine learning models,” *Quantum Science and Technology*, vol. 4, no. 4, p. 043001, 2019.
- [2] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, “Circuit-centric quantum classifiers,” *Physical Review A*, vol. 101, no. 3, p. 032308, 2020.
- [3] S. Lloyd and C. Weedbrook, “Quantum generative adversarial learning,” *Physical Review Letters*, vol. 121, no. 4, p. 040502, 2018.
- [4] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O’Brien, “A variational eigenvalue solver on a photonic quantum processor,” *Nature Communications*, vol. 5, p. 4213, 2014.
- [5] J. Shi, R.-X. Zhao, W. Wang, S. Zhang, and X. Li, “Qsan: A near-term achievable quantum self-attention network,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 8, pp. 13995–14 008, 2025.
- [6] R.-X. Zhao, J. Shi, and X. Li, “Qksan: A quantum kernel self-attention network,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 184–10 195, 2024.
- [7] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [8] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” 2019. [Online]. Available: <https://arxiv.org/abs/1708.06733>
- [9] Y. Zhang, Z. Li, C. Wang, K. Chen, and X. Wang, “BAIT: Large language model backdoor scanning by inverting attack target,” in *Proceedings of the IEEE Symposium on Security and Privacy*. San Francisco, CA: IEEE, 2024, pp. 1–18.
- [10] M. Liu, T. Zhang, W. Zhao, T. Kim, and R. B. Lee, “DeepVenom: Persistent dnn backdoors exploiting transient weight perturbations in memories,” in *Proceedings of the IEEE Symposium on Security and Privacy*. San Francisco, CA: IEEE, 2023, pp. 1–20.
- [11] H. Wang, S. Chen, Y. Liu, C. Zhang, and N. Z. Gong, “Exploring the orthogonality and linearity of backdoor attacks,” in *Proceedings of the IEEE Symposium on Security and Privacy*. San Francisco, CA: IEEE, 2022, pp. 1–17.
- [12] C. Chu, L. Jiang, M. Swany, and F. Chen, “Qtrojan: A circuit backdoor against quantum neural networks,” *arXiv preprint arXiv:2302.08090*, 2023.
- [13] J. Guo *et al.*, “Backdoor attacks against hybrid classical-quantum neural networks,” *arXiv preprint arXiv:2407.16273*, 2024.
- [14] C. Chu, F. Chen, P. Richerme, and L. Jiang, “Qdoor: Exploiting approximate synthesis for backdoor injection in quantum circuits,” *arXiv preprint arXiv:2307.09529*, 2023.
- [15] J. Shi, Z. Xiao, H. Shi, Y. Jiang, and X. Li, “Quantest: Entanglement-guided testing of quantum neural network systems,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 2, pp. 1–32, 2025.
- [16] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature*, vol. 549, no. 7671, p. 195–202, Sep. 2017. [Online]. Available: <http://dx.doi.org/10.1038/nature23474>
- [17] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, “Detecting backdoor attacks on deep neural networks by activation clustering,” *arXiv preprint arXiv:1811.03728*, 2018.
- [18] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks,” in *2019 IEEE Symposium on Security and Privacy*, 2019, pp. 707–723.
- [19] T. Li, J. Chen, and S. Guo, “Q-detection: A quantum-classical hybrid poisoning attack detection method,” in *Proceedings of the 32nd USENIX Security Symposium*. Anaheim, CA: USENIX Association, 2023, pp. 1–18.
- [20] F. Glover, G. Kochenberger, and Y. Du, “A tutorial on formulating and using qubo models,” 2019. [Online]. Available: <https://arxiv.org/abs/1811.11538>
- [21] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2002.
- [22] J. Preskill, “Quantum computing in the nisq era and beyond,” *Quantum*, vol. 2, p. 79, 2018.
- [23] M. Schuld and N. Killoran, “Quantum machine learning in feature hilbert spaces,” *Physical Review Letters*, vol. 122, no. 4, p. 040504, 2019.
- [24] M. Schuld, I. Sinayskiy, and F. Petruccione, “An introduction to quantum machine learning,” *Contemporary Physics*, vol. 56, no. 2, pp. 172–185, 2015.
- [25] E. Farhi and H. Neven, “Classification with quantum neural networks,” *arXiv preprint arXiv:1802.06002*, 2018.
- [26] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, “Circuit-centric quantum classifiers,” *Physical Review A*, vol. 101, no. 3, p. 032308, 2020.
- [27] I. Cong, S. Choi, and M. D. Lukin, “Quantum convolutional neural networks,” *Nature Physics*, vol. 15, no. 12, pp. 1273–1278, 2019.
- [28] S. Chen, C. Yang, J. Qi, H. Su, D. Deng, and Y. Xie, “Quantum recurrent neural networks,” *Quantum Science and Technology*, vol. 6, no. 3, p. 035002, 2021.
- [29] C. Xu and J. Szefer, “Security Attacks Abusing Pulse-level Quantum Circuits,” in *2025 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 222–239. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00083>
- [30] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, “Trojaning attack on neural networks,” *Proceedings of the Network and Distributed System Security Symposium*, 2018.
- [31] A. Saha, A. Subramanya, and H. Pirsiavash, “Hidden trigger backdoor attacks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, pp. 11 957–11 965, 2021.
- [32] J. Jia, Y. Liu, and N. Z. Gong, “Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning,” in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 2043–2059.
- [33] Z. Wang, B. Li, T. Chen, Y. Wang, and H. Wang, “Towards stealthy backdoor attacks against deep neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 4305–4314.
- [34] D. Tang, X. Wang, H. Tang, and K. Zhang, “Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection,” *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 3359–3373, 2022.
- [35] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, “Strip: A defence against trojan attacks on deep neural networks,” in *Proceedings of the Annual Computer Security Applications Conference*. ACM, 2019, pp. 113–125.
- [36] C. Y. Park, W. Gan, Z. Zou, Y. Hu, Z. Sun, and U. S. Kamilov, “Efficient model-based deep learning via network pruning and fine-tuning,” 2025.
- [37] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, “Strip: A defence against trojan attacks on deep neural networks,” 2019, also presented at ACSAC 2019.
- [38] B. Tran, J. Li, and A. Madry, “Spectral signatures in backdoor attacks,” in *Advances in Neural Information Processing Systems*, 2018.
- [39] E. Gültepe and M. Makrehchi, “Improving clustering performance using independent component analysis and unsupervised feature learning,” p. 25, 2018.
- [40] M. Ivanov, “A comparison of pca with ica from data distribution perspective,” 2017.
- [41] J. Shlens, “A tutorial on principal component analysis,” 2014. [Online]. Available: <https://arxiv.org/abs/1404.1100>
- [42] K. P. Sinaga and M.-S. Yang, “Unsupervised k-means clustering algorithm,” *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020.
- [43] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.
- [44] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, “Emnist: an extension of mnist to handwritten letters,” 2017. [Online]. Available: <https://arxiv.org/abs/1702.05373>
- [45] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, “Quantum circuit learning,” *Physical Review A*, vol. 98, no. 3, Sep. 2018. [Online]. Available: <http://dx.doi.org/10.1103/PhysRevA.98.032309>

- [46] L. Hou, R. Feng, Z. Hua, W. Luo, L. Y. Zhang, and Y. Li, “Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.09786>
- [47] C. Hu, X. Teng, W. Xing, H. Chen, C. Ye, and M. Han, “Distill to detect: Amplifying anomalies in backdoor models through knowledge distillation,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.