Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantum Autoencoders for Anomaly Detection in Cybersecurity

Rohan Senthil
Home Team Science & Technology Agency
Swee Liang Wong
wong_swee_liang@htx.gov.sg

Home Team Science & Technology Agency

Research Article

Keywords: Quantum Autoencoders, Quantum Machine Learning, Quantum Encoding, Anomaly Detection

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7540141/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Quantum Autoencoders for Anomaly Detection in Cybersecurity

Rohan Senthil[1] and Swee Liang Wong[1*]

[1]Disruptive Technologies Office, Home Team Science & Technology Agency, 1 Stars Avenue, Singapore, 138507, Singapore, Singapore.

*Corresponding author(s). E-mail(s): wong_swee_liang@htx.gov.sg;
Contributing authors: rohansenthil.biz@gmail.com;

Abstract

Anomaly detection in cybersecurity is a challenging task, where normal events far outnumber anomalous ones with new anomalies occurring frequently. Classical autoencoders have been used for anomaly detection, but struggles in data-limited settings which quantum counterparts can potentially overcome. In this work, we apply Quantum Autoencoders (QAEs) for anomaly detection in cybersecurity, specifically on the BPF-extended tracking honeypot (BETH) dataset. QAEs are evaluated across multiple encoding techniques, ansatz types, repetitions, and feature selection strategies. Our results demonstrate that an 8-feature QAE using Dense-Angle encoding with a RealAmplitude ansatz can outperform Classical Autoencoders (CAEs), even when trained on substantially fewer samples. The effects of quantum encoding and feature selection for developing quantum models are demonstrated and discussed. In a data-limited setting, the best performing QAE model has a F1 score of 0.87, better than that of CAE (0.77). These findings suggest that QAEs may offer practical advantages for anomaly detection in data-limited scenarios.

Keywords: Quantum Autoencoders, Quantum Machine Learning, Quantum Encoding, Anomaly Detection

1 Introduction

In today's fast-evolving digital environment, cyber threats continue to increase in both sophistication and frequency. Traditional security measures often struggle to keep up with the relentless influx of attacks (Abdulla et al. 2020). Such cyber threats often manifest as anomalous activities in the system. Thus, means of finding these threats can be done through anomaly detection. Anomaly detection is a key component of data analysis that focuses on identifying data points or events that deviate significantly from the norm. It plays a vital role in areas such as cybersecurity and fraud detection, where spotting unusual patterns can be critical for uncovering potential threats and preventing breaches (Tscharke et al. 2023).

However, the current cybersecurity landscape is marked by an unprecedented growth in the volume, velocity, and complexity of security-related data (Alluhaibi 2024). This results in normal cyber-activities far outnumbering anomalous ones, making detecting of anomalies a challenging task. This is evidenced in cybersecurity-related datasets which are highly imbalanced, where normal activities far outnumber anomalous ones. (Highnam et al. 2021). Furthermore, the rapidly evolving landscape for cybersecurity and the continuously varying activity data (Petersen 2015) means that a

significant challenge for supervised anomaly detection techniques is maintaining large, high-quality, labeled, and relevant datasets (Mirsky et al. 2018). Moreover, as threats evolves, so will the nature of how normal and anomalous data are represented.

Previous works suggest the use of machine learning methods such as Support Vector Machines and deep learning techniques such as Convolutional Neural Networks to distinguish regular from suspicious activities.(Saba et al. 2021). Other methods include autoencoders which are trained solely on 'normal' data(Mirsky et al. 2018)(Neloy and Turgeon 2024)(Patra et al. 2022) due to the aforementioned scenario of limited anomalous data. Classical Autoencoders (CAEs) learn the encoding distribution of 'normal' data, so that when an anomalous data is reconstructed using a CAE trained only on 'normal' data, its resultant reconstruction error will be much higher as the anomaly does not conform to the distribution. (Wu et al. 2024).

The quantum autoencoder (QAE), introduced by Romero et al. (2017), extends the CAE concept to quantum systems. QAEs can compress and decompress quantum states, flagging outliers based on either reconstruction errors or by measuring fidelity from a SWAP test (Liu and Rebentrost 2018)(Herr et al. 2021)(Tscharke et al. 2023)(Frehner and Stockinger 2025). QAEs have potential to provide advantageous data efficiency and expressibility due to the ability of QAEs to operate in higher dimensions compared to classical counterparts.

Our research proposes the use of Quantum Autoencoders for anomaly detection in cybersecurity, leveraging the intrinsic expressibility and efficiency of quantum circuits. In this work, we demonstrated that QAEs, trained solely on normal data from the BETH dataset, outperform their classical counterparts in data-limited settings in detecting anomalies. The choice of encoding and ansatz are also discussed.

2 Related Work

2.1 Classical Autoencoders

CAEs have been used for anomaly detection in cybersecurity applications, particularly effective in scenarios where anomalous data is scarce and explicit labeling is impractical. These models function by learning compact latent representations of normal system behavior and can accurately reconstruct similar input data with low error under such conditions. In contrast, anomalous inputs that deviate significantly from learned patterns yield higher reconstruction errors, enabling their identification (Afzal et al. 2025).

Recent progress has extended beyond basic autoencoders to more advanced architectures. Variational Autoencoders introduce regularization by constraining latent representations to follow prior distributions, while Adversarial Autoencoders leverage generative adversarial networks to align encodings with arbitrary priors (Mohamed 2023)(Larsen et al. 2016) Nonetheless, achieving the right balance between reconstruction quality and generalization remains a challenge: excessive compression can hinder accurate reconstruction of normal data, while overly complex models risk overfitting and reduced sensitivity to anomalies (Mohamed 2023).

2.2 Quantum Autoencoders

QAEs represent a powerful quantum machine learning paradigm that extends CAEs by leveraging uniquely quantum properties such as superposition and entanglement to enhance anomaly detection capabilities (Romero et al. 2017). In high-energy physics, for instance, QAEs have been shown to outperform CAEs on data generated from the Large Hadron Collider, achieving superior reconstruction performance with substantially fewer training samples and parameters (Ngairangbam et al. 2022). This drastic reduction in parameter requirements while maintaining strong performance underscores a key advantage of quantum methods, particularly in data-constrained settings.

Early work by Liu et al. (Liu and Rebentrost 2018) provided evidence of the potential speedup and resource efficiency for quantum computing based anomaly detection methods. Previous work demonstrates that QAEs, when augmented with quantum gradient descent for training, are much less dependent on the number of training samples and reach optimal reconstruction performance with small training datasets (Ngairangbam et al.

2022). This efficiency extends to model complexity, where QAEs with approximately 10 parameters can outperform classical autoencoders with around 1000 parameters in anomaly detection capabilities (Ngairangbam et al. 2022).

Recent reports on applying QAEs to time-series anomaly detection in cybersecurity data demonstrate consistently better performance over CAEs, while using fewer parameters and training iterations (Frehner and Stockinger 2025). In Frehner et al's work, it explored the effects of different ansatz and measurement techniques (SWAP-based vs Reconstructed Mean Squared Error). However, there is still a limited understanding of how different quantum encoding and feature selection strategies impact the performance of QAEs, indicating opportunities for further optimization and specialization (Araz and Spannowsky 2024). This suggests that the development of specialized QAE variants is still in its early stages, with substantial potential for domain-specific optimizations and architectural approaches.

2.2.1 Quantum Encoding

Quantum encoding techniques are data preprocessing steps that transform classical data into quantum states that can be processed by quantum circuits. The method of encoding data is also known as applying a quantum feature map (Schuld and Killoran 2019).

The choice of encoding method significantly impacts the expressivity of quantum machine learning models, affecting their ability to capture complex patterns in data and their computational efficiency. Different encoding techniques create distinct quantum state spaces and correlation structures, which can either enhance or limit a model's capacity to learn from specific types of data. Each encoding method creates distinct topological structures that can preserve or distort the original data relationships, directly affecting downstream tasks such as clustering and classification (Vlasic and Pham 2023). This indicates that the performance of quantum machine learning models is highly dependent on the choice of encoding, which we will explore in this work.

Thus in this work, we applied QAE to detect anomalies, using the BETH dataset and the performance is compared against CAE. We explore the effects of sample sizes, feature selection and quantum encoding techniques on the model performance. In our experiments, we implement and evaluate the following encoding techniques: (1) Amplitude Encoding, (2) Angle Encoding, (3) Dense-Angle Encoding, (4) Efficient SU2 Encoding. Various ansatz and feature selection are investigated, and their effects on the QAE model's performance are demonstrated.

3 Dataset

3.1 The BETH Dataset

The BPF-Extended Tracking Honeypot (BETH) dataset is a purpose-built cybersecurity dataset designed to support machine learning research in network security, which comprises tabular event-level network traffic data collected from real-world honeypots (Highnam et al. 2021). It provides a large collection of network traffic logs that include both benign activity and malicious intrusions, making it a strong candidate for testing anomaly detection methods. Overall, the dataset consists of *8,004,918 events* gathered from honeypots deployed on a major cloud service provider, spanning approximately *five noncontiguous hours* of operation on a major cloud provider. For benchmarking and discussion, an initial subset of the process logs was divided into training, validation, and testing sets with a rough 60/20/20 split. The split considered host origin, log volume, and activity type, with attacks appearing only in the test set. The distribution of normal and anomalous events in each subset is shown in Table 1.

The practical relevance of the BETH dataset extends beyond academic research, as it reflects real-world network conditions and attack patterns that cybersecurity professionals encounter in operational environments. This real-world applicability makes it particularly valuable for evaluating how QAEs perform on practical cybersecurity challenges.

3.2 Data Preprocessing

Data preprocessing and feature engineering is performed on the BETH dataset. Feature transformations were applied in accordance with the recommendations by Highnam et al.:

Table 1: Distribution of normal and anomalous events in the BETH dataset.

Dataset	Normal	Anomalous	Percentage Anomalous
Training	761,875	1,269	0.2%
Validation	188,181	786	0.4%
Testing	17,508	171,459	90.7%

- `ProcessId` and `ParentProcessId`: mapped to 0 if in {0, 1, 2}, else 1.
- `UserId`: mapped to 0 if <1000, else 1.
- `MountNameSpace`: mapped to 0 if equal to 4026531840, else 1.
- `ReturnValue`: mapped to 0 if zero, 1 if positive, and 2 if negative.
- `Timestamp` column dropped due to non-temporal modeling.
- `EventName` dropped due to one-to-one mapping with `EventId`.
- `ProcessName`, `ThreadId`, and `EventId` encoded using target encoding.
- `StackAddresses` and `HostName` were dropped.
- `Args`: each row contains a list of up to 5 dictionaries, each with keys `name`, `type`, and `value`. These were flattened into 15 new columns and encoded using target encoding.

All categorical features were processed via target encoding. Target encoding was chosen due to high-cardinality present in the data which causes other methods such as One Hot Encoding to result in overfitting. No normalization was applied for all features except for `ArgsNum`, which was scaled using standard scaling. In total, the transformed BETH dataset contained 24 features.

4 Experiments

QAE was implemented in `Qiskit` and simulated on local hardware. The primary computing environment consisted of an `11th Gen Intel® Core™ i7-11800H` processor (16 threads), `Intel® UHD Graphics (TGL GT1)`, and 16 GB of RAM, running `Ubuntu 24.10`. In addition, `Google Colab L4 GPUs` were also employed in parallel to train multiple model configurations.

The performance of the trained QAE model is evaluated against CAE by comparing the precision, accuracy, recall, F1-score, AUROC as well as detection statistics. The model is solely trained on the train dataset split with no anomalous data present in it while evaluation is done against the test dataset provided by the BETH dataset.

4.1 Quantum Autoencoder Architecture

Circuit Structure

The typical QAE consists of three conceptual layers: (1) *Input*, (2) *Bottleneck*, and (3) *Output*.

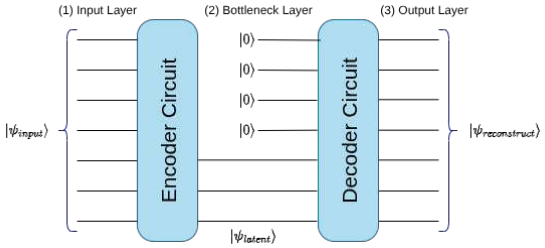


Fig. 1: Diagram of a typical QAE

1. *Input Layer (Encoder):* The input state $|\psi\rangle$, defined over n qubits, is initialized and passed through a parameterized circuit. This circuit acts as the encoder, compressing the quantum state into a reduced subspace of $n-k$ latent qubits. The resulting state takes the form

$$|\psi\rangle_{\text{latent}} \otimes |0\rangle^{\otimes k}$$

   where $|\psi\rangle_{\text{latent}}$ contains the $n-k$ latent qubits, while the remaining k qubits are discarded.
2. *Bottleneck Layer:* The k discarded qubits are removed from the circuit, leaving only the compressed subspace or latent qubits as the representation.
3. *Output Layer (Decoder):* To reconstruct the original state, a second parameterized circuit, acting as the decoder, is applied jointly on the latent qubits and k qubits initialized in $|0\rangle$ to approximate the original state. However, as mentioned later, our implementation of

SWAP test means that we do not need to carry out reconstruction of the input and can hence disregard the second parameterized circuit.

State Comparison via SWAP Test

To assess reconstruction fidelity without explicitly performing the full reconstruction, the k discarded or trash qubits are each paired with k reference qubits (predefined quantum states serving as fidelity benchmarks) and compared via a SWAP test (Asaoka and Kudo 2025). An additional auxiliary qubit is also introduced which is used for measurement tasks.

The SWAP test proceeds by applying CNOT gates between the auxiliary qubit with each pair of trash-reference qubits being compared (i.e. in Figure 2 $q2$ compared to $q4$ and $q3$ compared to $q5$). After executing the circuit M times, the auxiliary qubit is measured. Let L denote the number of times the auxiliary qubit is measured to be $|1\rangle$. The corresponding similarity score is defined as

$$S = 1 - \frac{2}{M}L. \tag{1}$$

Maximizing S corresponds to maximizing the fidelity between the two states, which is equivalent to ensuring that the input and reconstructed states are as close as possible. Thus, the training objective is to minimize the cost function $\frac{2}{M}L$.

Therefore, the anomaly score, A, is defined as

$$A = 1 - S \tag{2}$$

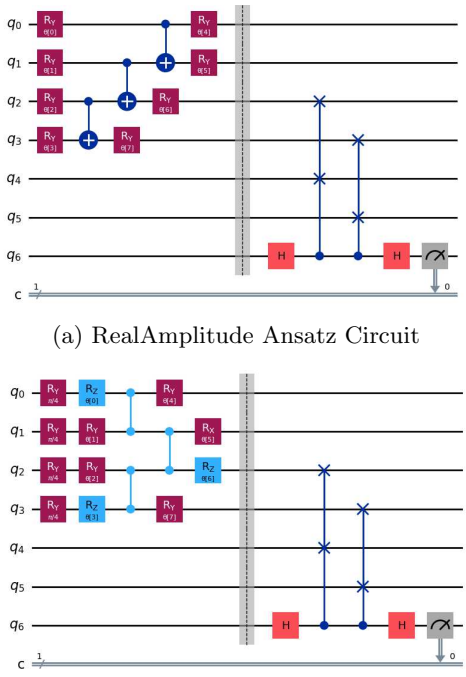
where S is the fidelity score obtained from the SWAP test. Since $S \in [0, 1]$, the anomaly score A also lies in $[0, 1]$, with $A = 0$ indicating a perfectly normal instance and $A = 1$ representing a maximally anomalous instance.

Ansatz and Training

For efficiency, we restricted our experiments to 8 features, corresponding to at most 8 qubits. Different encoding techniques require varying numbers of qubits to represent the same number of features; accordingly, the distribution between latent and trash qubits is technique-dependent.

Multiple parameterized ansatz are explored, such as `RealAmplitudes` (RA) and `PauliTwoDesign` (Pauli), which are illustrated in Figures 2a and 2b. The figures illustrate the different ansatz when Dense-Angle encoding is applied which encodes 8 features into 4 qubits.

For optimization, we employed the COBYLA algorithm, a gradient-free optimizer well-suited for quantum hardware, and performed 60 training iterations on different sample sizes with a batch size of 64.



(a) RealAmplitude Ansatz Circuit

(b) PauliTwoDesign Ansatz Circuit

Fig. 2: Comparison of two ansatz circuits: (a) RealAmplitude and (b) PauliTwoDesign. The features are encoded into $q0$ to $q3$ where $q0$ and $q1$ are the latent qubits and $q2$ and $q3$ are the trash qubits, $q4$ to $q5$ represent the reference qubits and $q6$ represents the auxiliary qubit.

4.2 Encoding Techniques

4.2.1 Amplitude Encoding

Amplitude encoding encodes data into the amplitudes of a quantum state. It represents a normalized classical N-dimensional data vector, as the amplitudes of a n-qubit quantum state.

Consider a dataset $X = x^1 + ... + x^i ... + x^m$ containing m samples, where each sample is a vector with n features, then the amplitude encoding can be represented as follows:

$$|\psi^i\rangle = \sum_{j=1}^{n} x_j^i |j\rangle \tag{3}$$

4.2.2 Angle Encoding

For angle encoding, each feature of the data sample is encoded into the rotation angle of each qubit where the data is scaled as $\phi \in [0, \pi]$. For the angle encoding of x^i, we require n quantum bits to complete the encoding (Shi et al. 2024). The encoding is represented as:

$$|\psi^i\rangle = \bigotimes_{j=1}^{n} R(x_j^i) \tag{4}$$

where, R is a single qubit R_y rotation gate.

4.2.3 Dense-Angle Encoding

Dense-Angle encoding is a combination of angle encoding and phase encoding. Phase encoding is similar to the angle encoding described above. The phase angle of a qubit is a real-valued angle ϕ about the z-axis from the $+x$-axis. Data is mapped with a phase rotation.

Dense-Angle allows two feature values to be encoded in a single qubit: one with a y-axis rotation angle (R_y gate), and the other with a z-axis rotation angle (R_z gate). Again, data is scaled according to $\phi \in [0, \pi]$. Encoding two data features into one qubit results in a 2 times reduction in the number of qubits required.

4.2.4 Efficient SU2 Encoding

An encoding scheme from `Qiskit`, Efficient SU2, which involves entanglement between the qubits is investigated as well. In our implementation, data is scaled where $\phi \in [0, \pi]$. It is made of y-axis rotations, z-axis rotations along with entanglements of neighboring qubits. This circuit can, for example, encode 8 features on only 2 qubits.

With the Efficient SU2 encoding we evaluated 16- and 24-feature inputs instead of 8, as 8 features would only allow 1 trash and 1 latent qubit that can result in loss of information due to overcompressing of data.

4.3 Classical Autoencoder Architecture

For benchmarking, we also implemented a CAE. (see Supplementary Information B for more details).

The model was compiled using the *Adam optimizer* and the *Mean Absolute Error (MAE)* as the loss function. The anomaly score is also defined as the *MAE* between the reconstructed data and the original data.

The encoder architecture for a 24 input feature model, as shown in Figure 3, was determined by experimenting with different hyperparameters. The decoder architecture is the reverse of the encoder. Anomaly scores are derived from the reconstruction error between input and output data.

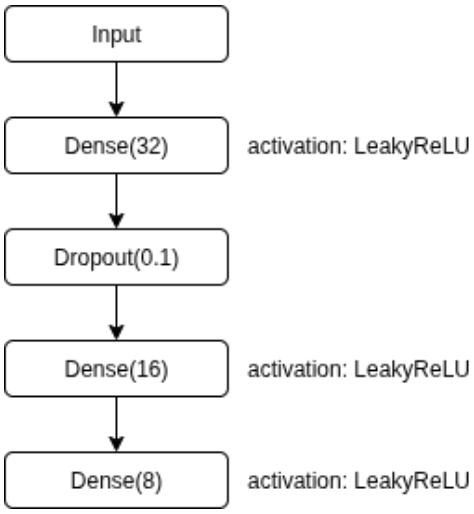


Fig. 3: Encoder Architecture for 24 input feature model

4.4 Thresholding Strategy

Anomalies are determined with a threshold defined as:

$$\text{Threshold} = \mu_{\text{error}} + 2 \cdot \sigma_{\text{error}} \tag{5}$$

where μ_{error} and σ_{error} represent the mean and standard deviation of the reconstruction errors on the training (normal) data, respectively. When the anomaly score exceeds this defined threshold, the data is considered anomalous. Once again,

we experimented with different thresholding techniques and determined the above to be the most optimal.

4.5 Feature Selection Strategies

The BETH dataset has 24 features, which requires a maximum of 24 qubits when angle encoding is employed. Due to computational resource constraints, only 8 features are selected in this work. These features are selected through different means and dimensionality reduction techniques. This includes using the latent space produced by the classical autoencoder as well as other methods such as Principal Component Analysis, Mutual Information, ANOVA, Kendall's and arbitrary selection methods such as First-N, and Least Correlated features.

5 Results

For the QAE, we explored different encoding techniques, feature selection strategies, ansatz types, and ansatz repetitions. We also experimented with increasing feature counts where feasible. Additionally, various entanglement techniques (linear, reverse-linear, full) for the ansatz were tested. Our choice of ansatz, encoding strategies and feature selection strategies are evaluated on a training size of 1000 samples.

Our key findings are quantitatively expressed using the F1 score as mean$\pm\sigma$, where σ represents the standard deviation of observed F1 scores. The results are summarized as follows:

- *Feature selection:* Using Amplitude and Dense-Angle Encoding as the basis for evaluating the feature selections strategies, F1 scores are compared: First-N (0.798 $\pm$ 0.063) comparatively performed the best consistently. Comparatively, other strategies such as ANOVA (0.44 $\pm$ 0.350), Kendall's (0.460 $\pm$ 0.262) and Least Correlated (0.380 $\pm$ 0.300) varied significantly across different encoding techniques.

Table 2: Mean $\pm$ σ of F1 scores for different feature selection strategies using Amplitude and Dense-Angle encoding.

Feature Selection Strategy	F1 Score
First-N	0.798 $\pm$ 0.063
Autoencoder-based reduction	0.695 $\pm$ 0.069
PCA (dimensionality reduction)	0.628 $\pm$ 0.143
PCA (feature filtering)	0.685 $\pm$ 0.045
Mutual Information	0.705 $\pm$ 0.065
ANOVA	0.440 $\pm$ 0.350
Kendall's	0.460 $\pm$ 0.262
Least Correlated	0.380 $\pm$ 0.300

- *Encoding Techniques:* Table 3 and Table 4 describes the model performance for different encoding techniques. Using autoencoder feature reduction and First-N feature selection as the basis to compare the F1 scores between models of different encoding techniques, we observe that Amplitude (0.705 $\pm$ 0.060) and Dense-Angle (0.771 $\pm$ 0.069) achieve the strongest overall performance. In comparison, Angle (0.709 $\pm$ 0.083) has greater variance in model performance and EfficientSU2 (0.675 $\pm$ 0.032) does not perform as well.

Table 3: Comparison of QAE models across different encoding techniques where CAE was used for dimensionality reduction. Row in bold indicate top performing model.

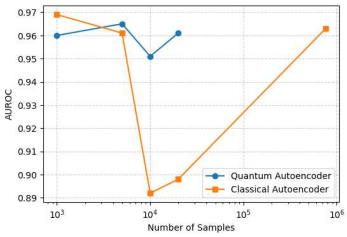
Encoding	Ansatz	Reps	F1	AUROC
Amplitude	RA	1	0.76	0.914
Amplitude	RA	3	0.58	0.413
Amplitude	Pauli	1	0.71	0.859
Amplitude	Pauli	3	0.76	0.110
Angle	RA	1	0.69	0.537
Angle	RA	3	0.69	0.583
Angle	Pauli	1	0.53	0.552
Angle	Pauli	3	0.71	0.120
Dense	RA	1	0.73	0.895
Dense	RA	3	0.71	0.085
Dense	**Pauli**	**1**	**0.86**	**0.918**
Dense	Pauli	3	0.71	0.131
24F EffSU2	RA	1	0.63	0.814
24F EffSU2	RA	3	0.70	0.103
24F EffSU2	Pauli	1	0.66	0.453
24F EffSU2	Pauli	3	0.71	0.683

Table 4: Performance metrics for First-N features across different encoding and ansatz configurations. Row in bold indicates the top-performing model.

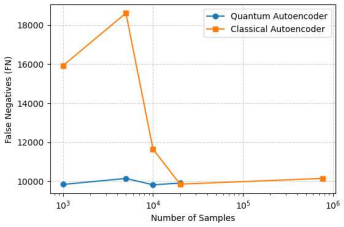
Encoding	Ansatz	Reps	F1	AUROC
Amplitude	RA	1	0.72	0.872
Amplitude	RA	3	0.75	0.899
Amplitude	Pauli	1	0.72	0.790
Amplitude	Pauli	3	0.64	0.424
Angle	RA	1	0.74	0.864
Angle	RA	3	0.85	0.896
Angle	Pauli	1	0.73	0.900
Angle	Pauli	3	0.73	0.835
Dense	**RA**	**1**	**0.86**	**0.960**
Dense	RA	3	0.86	0.955
Dense	Pauli	1	0.72	0.918
Dense	Pauli	3	0.72	0.096
24F EffSU2	RA	1	0.63	0.814
24F EffSU2	RA	3	0.70	0.103
24F EffSU2	Pauli	1	0.66	0.453
24F EffSU2	Pauli	3	0.71	0.683

Based on these results, the best performing model comprises the following: *(a) first 8 features of data as input, (b) Dense Angle encoding and (c) RealAmplitude ansatz*. This model is evaluated against a similarly trained CAE.

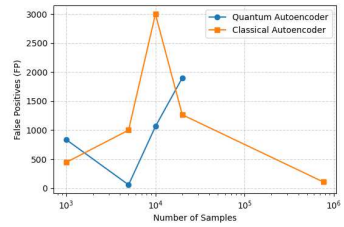
We illustrate in Figure 4 the effects of sample size on the performance of both the CAE and QAE. To accommodate for variations between subsets of data, random data subsets consisting 5,000 samples are evaluated. The sample size is chosen based on preliminary tests which indicates the highest model performance to sample size ratio for the QAE. The F1 scores (0.82 ± 0.040) demonstrate slight variations as different samples are used for training but still performs the best compared to the other models. Greater details of the results for the different sample sizes and repetitions for the QAE can be found in Supplementary Information F.



(a) AUROC vs. Samples



(b) FNs vs. Samples



(c) FPs vs. Samples

Fig. 4: Performance comparison of QAEs and CAEs trained with 1,000, 5,000, 10,000, and 20,000 samples. CAE also includes the full training dataset. QAE was not trained with the full dataset due to resource constraints.

Overall, the trained QAE model outperforms the classical counterpart when trained on the same number of samples. In Table 5 different configurations of the CAE is compared against our best performing QAE model.

To compare the anomaly score distributions as seen in Figure 5a and Figure 5b, we define *separation* as

$$\text{separation} = \frac{\text{mode}_{\text{anomaly}} - \text{mode}_{\text{normal}}}{\text{mode}_{\text{anomaly}}} \quad (6)$$

The CAE trained on all 24 features achieves a separation score of *0.84*, whereas the best-performing

Table 5: Comparison of CAE and QAE. All averages show mean ± σ across repeated 5k random subsets.

Model	Features	Samples	Subset	F1	AUROC
QAE	8	5,000	First Run	0.87	0.965
QAE	8	5,000	Average	0.82 ± 0.040	0.909 ± 0.033
CAE	8	5,000	Average	0.77 ± 0.017	0.901 ± 0.044
CAE	8	761,875	Full	0.66	0.964
CAE	24	5,000	Average	0.77 ± 0.025	0.862 ± 0.066
CAE	24	761,875	Full	0.87	0.963

QAE attains a higher score of *0.97*. This substantial improvement underscores the QAE's effectiveness in anomaly detection and its superior ability to distinguish normal from anomalous data

Simulated noise was also added to our experiments to simulate operation on real-world quantum hardware, specifically using the noise profile of the IBM Brisbane quantum computer. On the first batch of testing data, the anomaly scores increased by roughly 2.83x due to noise induced variation in measurement outcomes. Despite this increase, the model's overall performance remains unaffected. Only minor shifts in the anomaly score distribution are observed (see Figure 5c. Due to the strong separation between the anomaly scores of normal data and that of anomalous data, these minor shifts do not affect the results, and all other evaluation metrics remain unchanged.

6 Discussion

In our experiments, when appropriate features and encoding techniques were selected, QAEs consistently outperformed a CAE trained on the same sample size. Specifically, comparison between the QAE and the 8-feature input CAE shows that, at 5,000 training samples, the QAE achieves higher F1 score (+5%) compared to the CAE trained on the same subset. Notably, the QAE trained on the first 5,000 samples outperforms the CAE trained on the full dataset, highlighting the quantum model's efficiency in data-limited scenarios.

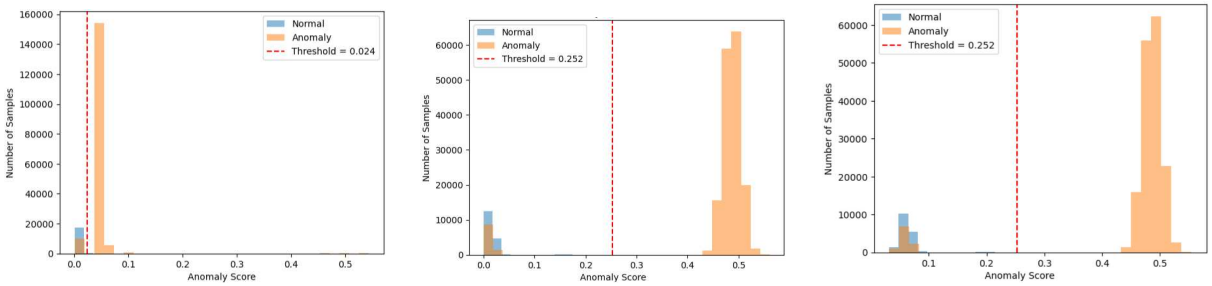
The anomaly score distributions and separation (0.97) further demonstrate the QAE's strength in identifying anomalous data. One possible explanation for the QAE's strong performance with limited data is that in higher-dimensional Hilbert space, subtle correlations and feature interactions become more apparent. Superposition and entanglement, which potentially enable richer encodings, also allow the model to simultaneously capture complex patterns and suppress irrelevant variations, reducing the need for large datasets compared to CAEs.

For feature selection, the First-N approach achieved the best results. As seen in Figure 6, t-SNE visualizations for First-N revealed clear and well-separated clusters. We hypothesize that this stems from its higher and more diverse feature variance (μ: 0.44, σ: 0.52), which facilitates separation in anomaly scores.

In comparison, PCA (μ: 0.22, σ: 0.44) and Mutual Information (MI) (μ: 0.13, σ: 0.16) exhibited lower variance, resulting in weaker t-SNE clustering and less discriminative features. (μ: 0.22, σ: 0.41) and Kendall correlation-based selection (μ: 0.22, σ: 0.41) included some very low-variance features alongside higher-variance ones. Least Correlated (μ: 0.27, σ: 0.44) captured diverse but uneven variance, yielding intermediate performance.

At the same time, a diverse but high-variance selection (μ: 0.66, σ: 0.49) leads to poor model performance.

To explore the effect of feature variance on model performance, we implemented a variance-balanced feature selection strategy. Specifically, we first computed the variance of each feature across the dataset, then sorted features from lowest to highest variance. From this sorted list, we selected a predefined number of features from three variance categories: low, mid-range, and high variance. For 24 features, each group has 8 features. The lowest group was drawn from the bottom of the distribution (lowest 8), the middle group from the central third of the distribution (middle 8), and the high group from the top (first 8), ensuring that both low- and high-variance features were explicitly represented. For instance, in our 8-feature subset, 2 features were chosen from the lowest-variance group, 4 from the mid-range, and 2 from the highest-variance group.



(a) Score Distribution of Best CAE simulated locally

(b) Score Distribution of Best QAE on a noiseless simulation

(c) Score Distribution of Best QAE with IBM Brisbane noise profile

Fig. 5: Score distributions of (a) the CAE simulated locally, (b) the QAE on a local noiseless simulation, and (c) the QAE on a local simulation with IBM Brisbane noise profile. While the CAE shows minimal separation, the QAE demonstrates a clear distinction between normal and abnormal data. The noise slightly shifts normal points, with no significant shift for anomalous data.

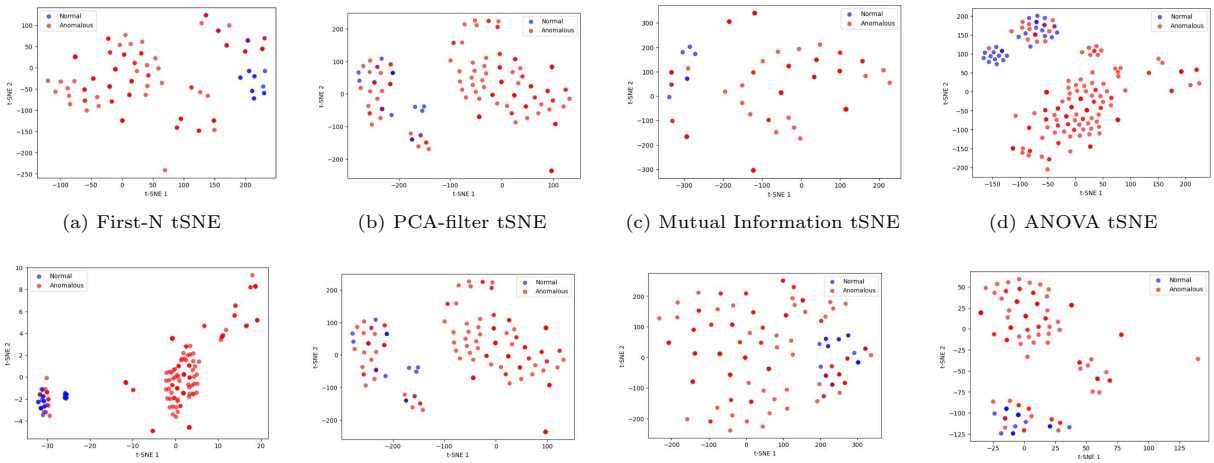
The variance-balanced selection (μ: 0.33, σ: 0.56) improved clustering relative to PCA, MI, and correlation-based methods by explicitly including features with both low and high variance. Though the model train with (a) first 5000 training samples, (b) Dense Angle encoding and (c) RealAmplitude ansatz still did not surpass First-N with a F1 score of 0.81 and a separation score of 0.89. These findings suggest that feature selection strategies that maintain variance diversity (mean and range) are most effective for anomaly detection in this setting.

As shown in our results, encoding techniques affect the performance of the QAEs. Both Amplitude and Dense-Angle encoding *consistently perform better* (Amplitude F1: 0.705 $\pm$ 0.060, Dense-Angle F1: 0.771 $\pm$ 0.069) than Angle encoding (F1:0.709 $\pm$ 0.083) which is less consistent and Efficient SU2 encoding (F1: 0.675 $\pm$ 0.032) which has lower performance. Comparing performance based on QAE trained with data using only first-N features, Amplitude and Angle encodings show comparable overall performance, but they excel in different aspects: Amplitude encoding more effectively reduces FPs, while Angle encoding is better at minimizing FNs. In contrast, Dense-Angle strikes a balance between these metrics and generally outperforms both, likely due to its incorporation of both rotations and phase information. Our results also indicate that Efficient SU2 is more sensitive to the choice of ansatz and the number of repetitions and Angle is more sensitive to the choice of features. For Angle encoding specifically, this sensitivity may arise from the lack of inherent entanglement in the encoding scheme, whereas Amplitude and Dense-Angle includes entanglement between qubits, enhancing their ability to capture multi-feature correlations and hence experiences less variation with the choice of ansatz and features, performing well even with limited data.

7 Future Work

Our results indicate that QAEs surpass CAEs in anomaly detection in instances with limited training data. However, this advantage is highly dependent on appropriate feature selection and encoding strategies, with model performance largely impacted by them. Future research should focus on developing systematic frameworks and algorithms for identifying optimal features and selecting suitable encoding techniques. Considering constraints in qubit numbers in current hardware, further work is needed to design feature selection methods that are generalisable across different datasets. Similarly, there is still room for investigation into the impact of different encoding techniques across datasets, data types, and circuit architectures. Finally, extending this study to a broader range of datasets would provide stronger insights into how data characteristics influence both CAE and QAE performance.

Moreover, while QAEs show clear potential, resource constraints of current quantum hardware still remains. Further development in quantum hardware and algorithms would be required to



(a) First-N tSNE (b) PCA-filter tSNE (c) Mutual Information tSNE (d) ANOVA tSNE

(e) Kendall's tSNE (f) Least Correlated tSNE (g) HighestVar tSNE (h) BalancedVar tSNE

Fig. 6: tSNE plots of various feature selection strategies.

achieve performance gains that outweigh the additional cost and complexity in implementing these QAEs.

8 Conclusion

This study explored the potential of QAEs for anomaly detection, benchmarking their performance against CAEs on the BETH dataset. Our findings highlight that, when combined with effective feature selection and carefully chosen encoding techniques, QAEs can outperform their classical counterparts. We also note the ability of QAEs to separate anomalies from normal samples as evident by a separation of *0.97*. Experiments also demonstrate the sensitivity of QAEs to both input representation and circuit design—underscoring the need for principled approaches to encoding and feature engineering in quantum machine learning.

Overall, our work contributes an early empirical comparison between CAEs and QAEs in a practical anomaly detection task, while opening avenues for systematic methods in feature and encoding selection. With further refinements, QAEs may well be a robust alternative to classical methods.

Supplementary information. BETH dataset details, performance statistics of different models and their respective architectures can be found in the Supplementary information.

Acknowledgements. The authors acknowledge the Home Team Science and Technology Agency for financial support.

Author Contributions. R.S developed the main conceptual ideas, set up and conducted the experiments and performed the data analysis. S.L.W supervised the project. Both authors wrote and reviewed the manuscript.

Declarations

Funding. Not Applicable

Data Availability. Code used in this work is available on reasonable request. The data is publicly available as referenced in the text.

Competing Interests. The authors declare no competing interests.

References

Abdulla A, Mohamed A, Razali S (2020) A review on cybersecurity: Challenges and emerging threats. In: International Conference on Networking, Information Systems & Security (NISS), pp 1–7, https://doi.org/10.1145/3386723.3387847, URL https://doi.org/10.1145/3386723.3387847

Afzal Z, Gaggero G, Asplund M (2025) Towards privacy-preserving anomaly-based intrusion

detection in energy communities. arXiv preprint https://doi.org/10.48550/arXiv.2502.19154, URL https://arxiv.org/abs/2502.19154, arXiv:2502.19154 [cs.CR]

Alluhaibi R (2024) Quantum machine learning for advanced threat detection in cybersecurity. International Journal of Safety and Security Engineering 14(3):875–883. https://doi.org/10.18280/ijsse.140319, URL https://doi.org/10.18280/ijsse.140319

Araz JY, Spannowsky M (2024) The role of data embedding in quantum autoencoders for improved anomaly detection. arXiv preprint https://doi.org/10.48550/arXiv.2409.04519, URL https://arxiv.org/abs/2409.04519, arXiv:2409.04519 [quant-ph]

Asaoka H, Kudo K (2025) Quantum autoencoders for image classification. Quantum Machine Intelligence 7(2). https://doi.org/10.1007/s42484-025-00297-x, URL http://dx.doi.org/10.1007/s42484-025-00297-x

Frehner R, Stockinger K (2025) Applying quantum autoencoders for time series anomaly detection. Quantum Machine Intelligence https://doi.org/10.1007/s42484-025-00285-1, URL https://doi.org/10.1007/s42484-025-00285-1

Herr D, Obert B, Rosenkranz M (2021) Anomaly detection with variational quantum generative adversarial networks. Quantum Science and Technology 6(4):045004. https://doi.org/10.1088/2058-9565/ac0d4d, URL https://doi.org/10.1088/2058-9565/ac0d4d

Highnam K, Arulkumaran K, Hanif Z, et al (2021) Beth dataset: Real cybersecurity data for unsupervised anomaly detection research. Proceedings of the Conference on Applied Machine Learning for Information Security (CAMLIS) URL https://ceur-ws.org/Vol-3095/paper1.pdf

Larsen ABL, Sønderby SK, Larochelle H, et al (2016) Autoencoding beyond pixels using a learned similarity metric. arXiv preprint https://doi.org/10.48550/arXiv.1512.09300, URL https://arxiv.org/abs/1512.09300, arXiv:1512.09300 [cs.LG]

Liu N, Rebentrost P (2018) Quantum machine learning for quantum anomaly detection. Physical Review A 97(4):042315. https://doi.org/10.1103/PhysRevA.97.042315, URL https://doi.org/10.1103/PhysRevA.97.042315

Mirsky Y, Doitshman T, Elovici Y, et al (2018) Kitsune: An ensemble of autoencoders for online network intrusion detection. arXiv preprint arXiv:180209089 https://doi.org/10.48550/arXiv.1802.09089, URL https://arxiv.org/abs/1802.09089

Mohamed M (2023) Comparative evaluation of vaes, vae-gans and aaes for anomaly detection in network intrusion data. EMITTER International Journal of Engineering Technology 11(2):160–173. https://doi.org/10.24003/emitter.v11i2.817, URL https://emitter.pens.ac.id/index.php/emitter/article/view/817

Neloy AA, Turgeon M (2024) A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. Machine Learning with Applications 17:100572. https://doi.org/10.1016/j.mlwa.2024.100572, URL https://www.sciencedirect.com/science/article/pii/S2666827024000483

Ngairangbam VS, Spannowsky M, Takeuchi M (2022) Anomaly detection in high-energy physics using a quantum autoencoder. Physical Review D 105(9):095004. https://doi.org/10.1103/PhysRevD.105.095004, URL https://doi.org/10.1103/PhysRevD.105.095004

Patra K, Sethi R, Behera D (2022) Anomaly detection in rotating machinery using autoencoders based on bidirectional lstm and gru neural networks. Turkish Journal of Electrical Engineering and Computer Sciences 30(7):2649–2662. https://doi.org/10.55730/1300-0632.3870, URL https://journals.tubitak.gov.tr/elektrik/issues/elk-22-30-7/elk-30-7-19-2111-40.pdf

Petersen R (2015) Data mining for network intrusion detection: A comparison of data mining algorithms and an analysis of relevant features for detecting cyber-attacks. Semantic Scholar URL https://api.semanticscholar.org/CorpusID:64401231

Romero J, Olson JP, Aspuru-Guzik A (2017) Quantum autoencoders for efficient compression of quantum data. Quantum Science and Technology 2(4):045001. https://doi.org/10.1088/2058-9565/aa8072, URL http://dx.doi.org/10.1088/2058-9565/aa8072

Saba T, Sadad T, Rehman A, et al (2021) Intrusion detection system through advance machine learning for the internet of things networks. IT Professional 23(2):58–64. https://doi.org/10.1109/MITP.2020.3016287, URL https://ieeexplore.ieee.org/document/9335650

Schuld M, Killoran N (2019) Quantum machine learning in feature hilbert spaces. Physical Review Letters 122(4):040504. https://doi.org/10.1103/PhysRevLett.122.040504, URL https://doi.org/10.1103/PhysRevLett.122.040504

Shi M, Situ H, Zhang C (2024) Hybrid quantum neural network structures for image multi-classification. Physica Scripta 99(5):056012. https://doi.org/10.1088/1402-4896/ad3e3d, URL https://dx.doi.org/10.1088/1402-4896/ad3e3d

Tscharke K, Issel S, Debus P (2023) Semisupervised anomaly detection using support vector regression with quantum kernel. In: IEEE International Conference on Quantum Computing and Engineering (QCE), pp 611–620, https://doi.org/10.1109/QCE57702.2023.00075, URL https://ieeexplore.ieee.org/document/10313835

Vlasic A, Pham A (2023) Understanding the mapping of encode data through an implementation of quantum topological analysis. Quantum Information and Computation 23(13–14):1091–1104. https://doi.org/10.26421/QIC23.13-14-2, URL http://dx.doi.org/10.26421/QIC23.13-14-2

Wu JY, Fu H, Zhu M, et al (2024) Quantum circuit autoencoder. Physical Review A 109(3):032623. https://doi.org/10.1103/PhysRevA.109.032623, URL https://journals.aps.org/pra/abstract/10.1103/PhysRevA.109.032623

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SIQuantumAutoencodersforAnomalyDetectioninCybersecurity.pdf