

Rigorous Benchmarking of Quantum Autoencoders for Time Series Anomaly Detection

Lourens van Niekerk^{*1,2}, Dennis Willsch^{1,3}, and Kristel Michielsen^{1,2,4}

¹Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany

²Universität zu Köln, 50923 Köln, Germany

³FH Aachen University of Applied Sciences, 52066 Aachen, Germany

⁴AIDAS, 52425 Jülich, Germany

Motivation

Autoencoders have proven effective at TSAD, yet a systematic, rigorous benchmark of quantum autoencoders is lacking. We are conducting a comprehensive study for these purposes, and comparing against state-of-the-art classical models.

Anomaly detection (AD) is useful in industrial areas such as in medical assessment, fraud detection, earth observation, and many more areas. These applications usually rely on deep time-series (TS) data, in which anomalies could be complex and difficult to detect. There have been a vast amount of research, in the classical sense, in time-series anomaly detection (TSAD). Popularly, autoencoders have most frequently been adopted as the base model for performing TSAD. Exceptional work thus far in TSAD is the UCR-TSA archive[1][2], the non-DL search method MERLIN++[3], and the state-of-the-art TimeVQVAE-AD classical model[4]. There is, however, a lack of rigorous, systematic benchmarking of quantum or hybrid quantum-classical autoencoder models. Each existing model either uses a dataset which has been shown to be flawed[1][5], was only focused on a specific domain, such as networking, high-energy physics and finance, for its benchmark [6][7][8][9], only focuses on a subset of all adjustable options in the quantum autoencoder (QAE) instead of all, or has a model which was insufficiently benchmarked[10]. We propose a proper benchmark of all the different aspects that make up a quantum autoencoder model and conclude how to optimally design a model for a given dataset and domain, particularly as it relates to TSAD. Romero [11] initially proposed a quantum autoencoder in 2016, initially just applied for general data compression. In 2018, Liu [12] showed the potential of quantum autoencoders in quantum anomaly detection. Our research will serve as a comprehensive analysis of how each adjustable option in a purely quantum-based autoencoder implementation.

Also mention how the Fourier domain and QFT play into producing the model in the end.

Results

The UCR-TSA archive[1] was introduced in 2022 and consists of 250 univariate time-series datasets, spanning ranges from 8 000 to 300 000 many time points, anomaly ranges from 1 to 1700, data types such as image, ECG, sensor and more. They described how generally there exists four types of anomalies: peak-, frequency-, smoothness-, and pattern anomalies. They also thoroughly described how almost all other existing anomaly detection benchmark datasets and archives have at least one of the following flaws: (1) Most anomalies can be detected trivially by what Wu et al. terms "one-liners" - a single suitable line of code; (2) The anomaly designations within datasets are unusual e.g. a large range will be considered multiple anomalies instead of one, or unusually many anomalies occur, defeating the very definition of what an anomaly is; (3) The ground truth of the dataset, i.e. what is and is not an anomaly, is flawed itself; (4) The anomalies are focused as

^{*}Corresponding author: l.van.niekerk@fz-juelich.de

run-till-failure, thus grouped at the end of the dataset and trivial to train a model for, or biasing a model to guess that an anomaly is at the last time points. The UCR-TSA archive tailored their 250 to include all four types of anomalies, and as close as possible avoid these 4 flaws. They also included distorted and noisy data to increase variety.

Quantum autoencoders[12] require classical data to be embedded as quantum data, and in the literature is typically done by amplitude encoding. Then, a parameterized quantum circuit (PQC) called an encoder, is applied, such that the embedded quantum state can be sufficiently compressed into a state consisting of less qubits, called the latent state. This latent state should contain as much information as possible. The complementary state that accordingly should contain as little information as possible, is called the trash state, and is trained via the encoder to resemble some reference state as close as possible. In order to train the encoder, a window is slid over the dataset and the output from all windows together is used for optimization of the encoder weights. Once fully trained, if the trash state and reference state are swapped, and the inverse of the encoder is taken as the decoder, then the reconstructed state should closely resemble the initial embedded state. Any data is not captured in the reconstruction, can be considered outliers by the autoencoder’s design. An additional goal of the design of the autoencoder is such that any and all anomalies that could occur, are recognised as such outliers. It has been shown wise to apply a Quantum Fourier Transform (QFT) after the feature embedding, since training with frequencies can more easily detect anomalies, particularly frequency-, smoothness- and pattern anomalies. An inverse QFT can then be applied after the decoder again to retrieve the reconstruction in the time domain.

Below we list each of the options we recognized, along which benchmarking can occur, without introducing additional classical or quantum segments that would make the model more than simply a quantum autoencoder:

1. **Preprocessing:** Many existing quantum autoencoders models we observed used the raw data of the TS dataset. This makes logical sense, although there is the possibility to extract data from the raw first and embed that as well. Packages such as *tsfresh*[13], tailored for time series feature extraction on the basis of scalable hypothesis testing, could be used. Then a combination of raw data and select extracted features could be used for the embedding.
2. **Weight initialization:** Weight initialization has been shown to be one method to mitigate barren plateaus[14] occurring in quantum neural networks. A popular choice has proven to be Xavier-Glorot Normal initialization[15][16][17], which aims to keep variances of outputs between a forward pass and a backward pass stable. This option can be benchmarked against lazy weight initialization, when they are simply set to zeros.
3. **Ansatz:** The different Sim et al. circuits[18] have been popular as ansatzes, consider that they consisted of nineteen ansatzes sources from a variety of sources, and benchmarked for their expressivity and entanglement capability. Further research by Röseler et al.[19] has been done on these ansatzes to also show their trainability, showcasing a trade-off between trainability and expressivity. A subset of these circuits can be taken, as well the standard RealAmplitudes ansatz from *qiskit*[20], and possibly ansatzes proposed from other quantum autoencoder research[5].
4. **Reference state:** In the original QAE proposal[11], the reference state was simply $|0\rangle^{\otimes k}$, where $1 \leq k \leq n$ and n is the amount of qubits used for the embedding. This is intuitive, as successful compression of information then looks like the trash state contains almost "zero information". In reality, the reference state can be any state, and thus should be the state that the trash state can best be optimized towards. Ma et al.[21] proposes two alternatives: (1) The training can be done with some reference state e.g. $|0\rangle^{\otimes k}$, but then during testing, the reference state is taken to be the average trash state from all the training windows, and (2) the reference state can be a maximally entangled state, as it is actually more natural for the (preferably) entangled trash state to resemble this, than a state with no entanglement.
5. **Step size:** The step size is the amount of time points that the next window is shifted from its prior. This can be as small as one, with massive overlap between windows, or as large as the window size

itself, thus having no overlap among windows. An even larger step size can be chosen. This leads to a trade-off between accuracy of the model, and the computational efficiency of training it.

6. **Optimizer:** Popular optimizers include COBYLA, Adam, SPSA and Nelder-Mead[22]. How does each compare for accuracy and computational efficiency?
7. **Feature embedding:** As mentioned, amplitude encoding is the most popular option, considering that time series datasets tend to be very large and small window sizes, requiring small step size, lack computational efficiency and the scope to capture frequencies in the Fourier domain accurately. There is, however, a proposed Quantum Time-Series encoding[23] by Laskar and Goel, which combines angle embedding with an additional CNOT gate structure to a circuit that is analogous to a classical autoregression. Despite the restriction to the window size in this case, the circuit is much shallower than in the amplitude embedding case, and only scales linearly, while providing evidence of its potential use.
8. **Trash space size:** How large should the trash state be, in terms of qubits, in comparison to the embedded state?
9. **Window size:** How large is optimal compared to the anomaly window size? The UCR-TSA archive has anomalies ranges from 1 to 1701. This means the potential to test window sizes, using amplitude embedding, ranging from 1 to 11 qubits.
10. **Postprocessing:** How could the amplitude readouts from the reconstruction optimally construct a line that mimics the original data?
11. **Anomaly score:** How is the model evaluated for anomaly detection in the end? Top-k[4], F-score, the UCR-TSAA[1], and VUS-ROC[24] are all options proposed by previous authors. We start our benchmarking with top-1, which is simply identifying the single most anomalous time point in the testing data, and identifying if it is sufficiently close to the true anomaly. We will, however, proceed to investigate the other options as well.

Discussion

Continue the benchmark with the missing parameters from the list that are not included here, such as benchmarking in the feature map for the QTS encoding, and in the preprocessing for feature extraction. What is not benchmarked is the comparison between SWAP test and MSE as readout for cost function optimization, as they have been found to be indistinguishable in their performance[25], thus we went with MSE, as it requires less shots for sufficient samples, and holds the opportunity to decrease circuit width in training by having the reference state be purely classical and used after trash state measurements are taken. A reference state is still necessary during reconstruction however. We will finally compare the final model to the comparative classical models, such as TimeVQVAE-AD[4] and MERLIN++[3] for reference.

References

- (1) Wu, R.; Keogh, E. J. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, IEEE: Kuala Lumpur, Malaysia, 2022, pp 1479–1480.
- (2) Baldán, F. J.; García-Gil, D. Benchmarking Anomaly Detection Methods: Insights From the UCR Time Series Anomaly Archive. *Expert Systems* **2025**, *42*, e13767.
- (3) Nakamura, T.; Mercer, R.; Imamura, M.; Keogh, E. MERLIN++: parameter-free discovery of time series anomalies. *Data Mining and Knowledge Discovery* **2023**, *37*, 670–709.
- (4) Lee, D.; Malacarne, S.; Aune, E. Explainable time series anomaly detection using masked latent generative modeling. *Pattern Recognition* **2024**, *156*, 110826.

- (5) Tschärke, K.; Wendlinger, M.; Ahouzi, A.; Bhardwaj, P.; Amoi-Taleghani, K.; Schrödl-Baumann, M.; Debus, P. In *2025 IEEE International Conference on Quantum Computing and Engineering (QCE)*; arXiv:2504.17548 [quant-ph], 2025, pp 2470–2481.
- (6) Tabi, Z.; Bakó, B.; Nagy, D. T. R.; Vadera, P.; Kallus, Z.; Hágá, P.; Zimborás, Z. Hybrid Quantum-Classical Autoencoders for End-to-End Radio Communication. *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)* **2023**, _eprint: 2301.02609, DOI: [10.1109/sec54971.2022.00071](https://doi.org/10.1109/sec54971.2022.00071).
- (7) Ngairangbam, V. S.; Spannowsky, M.; Takeuchi, M. Anomaly detection in high-energy physics using a quantum autoencoder. *Phys. Rev. D* **105**, 095004, 2022 **2021**, 105, _eprint: 2112.04958, 095004.
- (8) Kundu, S.; Mishra, D.; Tarasia, N.; Lenka, R. K.; Barik, R. K. In *2025 International Conference on Networks and Cryptology (NETCRYPT)*, 2025, pp 1280–1285.
- (9) Alami, M. E.; Innan, A.; Innan, N.; Shafique, M.; Bennai, M. FiD-QAE: A Fidelity-Driven Quantum Autoencoder for Credit Card Fraud Detection, arXiv:2512.12689 [quant-ph] version: 1, 2025.
- (10) Frehner, R.; Stockinger, K. Applying Quantum Autoencoders for Time Series Anomaly Detection, Version Number: 2, 2024.
- (11) Romero, J.; Olson, J. P.; Aspuru-Guzik, A. Quantum autoencoders for efficient compression of quantum data. *Quantum Sci. Technol.* **2** **2016**, 2, _eprint: 1612.02806, DOI: [10.1088/2058-9565/aa8072](https://doi.org/10.1088/2058-9565/aa8072).
- (12) Liu, N.; Rebentrost, P. Quantum machine learning for quantum anomaly detection. *Physical Review A* **2018**, 97, 042315.
- (13) Christ, M.; Braun, N.; Neuffer, J.; Kempa-Liehr, A. W. Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* **2018**, 307, 72–77.
- (14) McClean, J. R.; Boixo, S.; Smelyanskiy, V. N.; Babbush, R.; Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature Communications* **2018**, 9, DOI: [10.1038/s41467-018-07090-4](https://doi.org/10.1038/s41467-018-07090-4).
- (15) Kashif, M.; Rashid, M.; Al-Kuwari, S.; Shafique, M. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE: Valencia, Spain, 2024, pp 1–6.
- (16) Zhang, K.; Liu, L.; Hsieh, M.-H.; Tao, D. Escaping from the Barren Plateau via Gaussian Initializations in Deep Variational Quantum Circuits, arXiv:2203.09376 [quant-ph], 2025.
- (17) Peng, Y.; Li, X.; Zhang, Z.; Chen, S. Y.-C.; Liang, Z.; Wang, Y. Can Classical Initialization Help Variational Quantum Circuits Escape the Barren Plateau?, arXiv:2508.18497 [quant-ph] version: 1, 2025.
- (18) Sim, S.; Johnson, P. D.; Aspuru-Guzik, A. Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms. *Advanced Quantum Technologies* **2019**, 2, 1900070.
- (19) Röseler, P.; Willsch, D.; Michielsen, K. How to find expressible and trainable parameterized quantum circuits?, arXiv:2603.14451 [quant-ph], 2026.
- (20) Javadi-Abhari, A.; Treinish, M.; Krsulich, K.; Wood, C. J.; Lishman, J.; Gacon, J.; Martiel, S.; Nation, P. D.; Bishop, L. S.; Cross, A. W.; Johnson, B. R.; Gambetta, J. M. Quantum computing with Qiskit, arXiv:2405.08810 [quant-ph], 2024.
- (21) Ma, H.; Mooney, G. J.; Petersen, I. R.; Hollenberg, L. C. L.; Dong, D. Quantum autoencoders using mixed reference states. *npj Quantum Information* **2024**, 10, arXiv:2309.15582 [quant-ph], 86.
- (22) Pellow-Jarman, A.; Sinayskiy, I.; Pillay, A.; Petruccione, F. A Comparison of Various Classical Optimizers for a Variational Quantum Linear Solver. *Quantum Information Processing* **2021**, 20, arXiv:2106.08682 [quant-ph], 202.
- (23) Laskar, M. R.; Goel, R. Shallow entangled circuits for quantum time series prediction on IBM devices. *Scientific Reports* **2025**, 15, 43727.
- (24) Boniol, P.; Krishna, A. K.; Bruel, M.; Liu, Q.; Huang, M.; Palpanas, T.; Tsay, R. S.; Elmore, A.; Franklin, M. J.; Paparrizos, J. VUS: Effective and Efficient Accuracy Measures for Time-Series Anomaly Detection, arXiv:2502.13318 [cs], 2025.

- (25) Havlicek, V.; Córcoles, A. D.; Temme, K.; Harrow, A. W.; Kandala, A.; Chow, J. M.; Gambetta, J. M. Supervised learning with quantum enhanced feature spaces. *Nature*. vol. 567, pp. 209-212 (2019) **2018**, 567, _eprint: 1804.11326, 209–212.