

Classical autoencoder distillation of quantum adversarial manipulations

Amena Khatun ^{1,*} and Muhammad Usman ^{1,2}

¹Quantum Systems, Data61, CSIRO, Clayton 3168, Australia

²School of Physics, The University of Melbourne, Parkville, Victoria 3010, Australia



(Received 22 April 2025; accepted 16 November 2025; published 8 December 2025)

Quantum neural networks have been proven robust against classical adversarial attacks, but their vulnerability against quantum adversarial attacks is still a challenging problem. Here, we report a technique for the distillation of quantum-manipulated image datasets by using classical autoencoders. Our technique recovers quantum classifier accuracies when tested under standard machine learning benchmarks utilizing Modified National Institute of Standards and Technology (MNIST) and Fashion-MNIST (FMNIST) image datasets, and projected gradient descent and fast gradient sign method adversarial attack settings. Our work highlights a promising pathway to achieve fully robust quantum machine learning in both classical and quantum adversarial scenarios.

DOI: [10.1103/41z8-d3h9](https://doi.org/10.1103/41z8-d3h9)

Machine learning (ML) is rapidly becoming ubiquitous and its integration in real-world applications is already addressing highly complex challenges such as early disease detection using medical imaging [1,2], anomaly detection [3], natural language processing for chatbots [4], image recognition for autonomous vehicles [5], and climate modeling for environmental conservation [6]. Despite efficiency and high throughput of ML systems, there is a serious concern about their vulnerability to data spoofing or manipulations [7–12], in particular for security-sensitive applications such as self-driving vehicles [13] and military systems [14,15]. Meanwhile, the developments in the field of quantum ML have instigated a line of research in which the vulnerability of quantum ML models is being tested and benchmarked in the recent literature [16–24]. It has been shown that while quantum ML models are robust against classical adversarial attacks, they have been found vulnerable against quantum adversarial attacks—data manipulations generated via quantum models [23,24]. Therefore, techniques that can restore trust and reliability in quantum ML models will have important implications for future practical applications of these models.

This work proposes and implements a quantum-classical framework, which integrates a quantum variational classifier with a classical encoder-decoder (QVC-CED) model paving the way for an end-to-end architecture with robustness against both classical and quantum attacks. In classical ML literature, autoencoders have been widely used for denoising of images for classical neural networks [25,26]; however, our work is the first to integrate it with a quantum neural network and demonstrate its effectiveness for the distillation of quantum adversarial noise. We note that denoising and adversarial

purification are conceptually very different problems. In denoising, the objective is to remove random or unstructured noise, such as Gaussian or masking corruption, where the corruption process is typically independent of the models decision boundaries. In contrast, adversarial purification aims to suppress structured perturbations that are deliberately optimized to exploit model vulnerabilities. These perturbations are often imperceptible in magnitude but highly effective in shifting inputs across decision boundaries, making them substantially more challenging to mitigate than standard noise. Therefore, our work is the first to implement and benchmark autoencoder technique for quantum machine learning (QML) systems. In the context of quantum ML and adversarial perturbations, previous studies have explored various methods to protect quantum ML models against adversarial attacks, which include quantum adversarial training [16,17,27,28] and randomized encoding [29]. Here, we note that adversarial training method has only shown limited performance improvement against quantum attacks. For example, at attack strength of 0.3, the accuracy after adversarial training was only recovered to about 50% of the original accuracy at zero attack [24]. In contrast, the autoencoder technique presented in this work is remarkably efficient and it recovers almost full accuracy. Furthermore, one major limitation of the adversarial training is its very high computational cost, whereas autoencoder technique once trained is very efficient and fast. On the other hand, randomized encoding have not been implemented in practical QML applications and therefore its performance against standard adversarial attacks is yet unknown. Nevertheless, our work has presented a highly efficient and reliable technique, which will be a valuable resource in future QML defense against adversarial attacks. While Ref. [30] provides a theoretical analysis that quantum depolarization noise inherently protects quantum classifiers, and Ref. [21] demonstrates that such noise does not enhance adversarial robustness and may even degrade the performance of QML models. Other studies [24,31–33] have investigated that quantum classifiers exhibit a degree of resilience against classical adversarial attacks, due to learning different set of features than classical

*Contact author: amena.khatun@data61.csiro.au

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

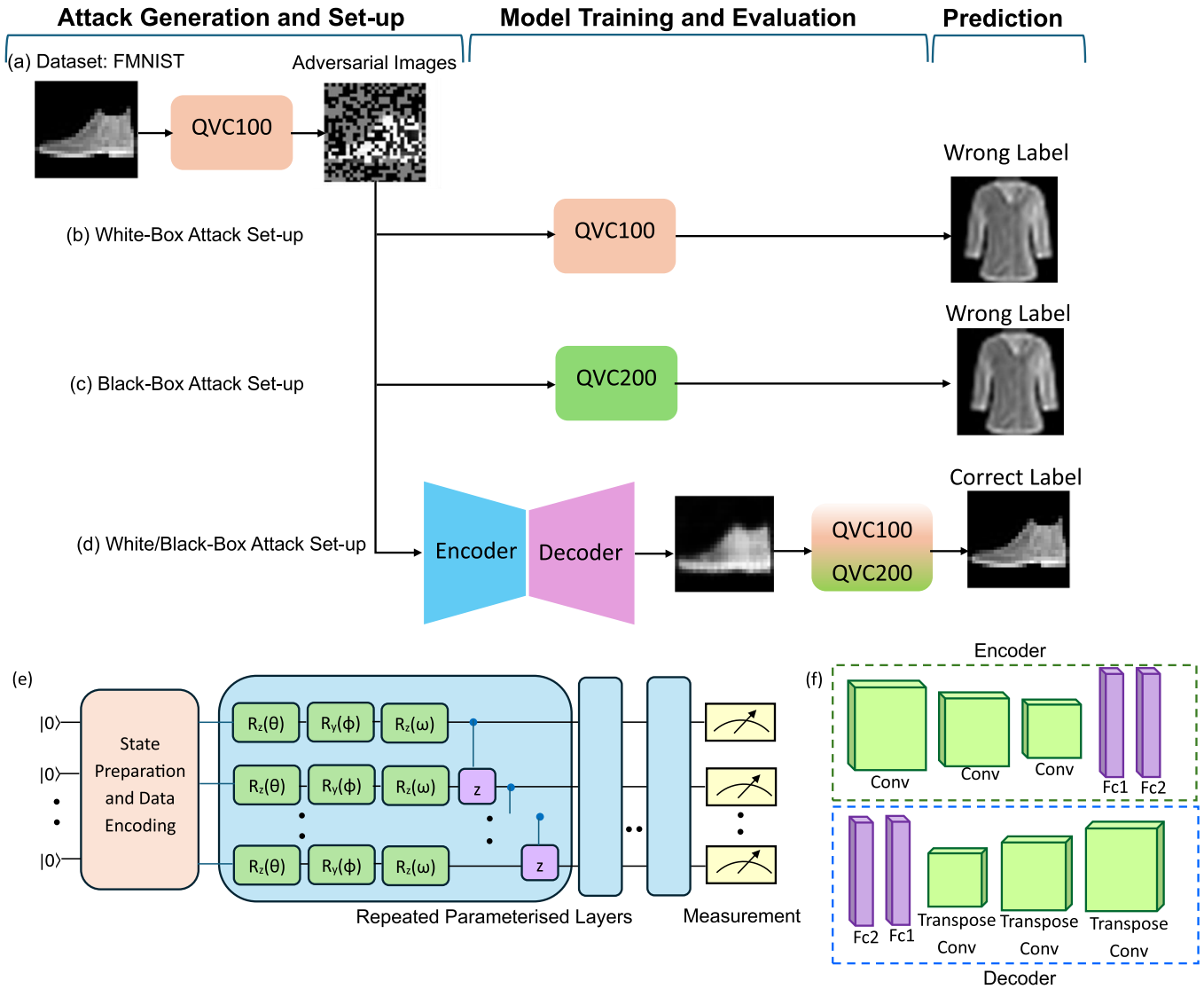


FIG. 1. Overview and architecture of the proposed QVC-CED approach. In this figure, panel (a) represents the first step in our workflow where clean images are input to a quantum variational circuit to generate adversarial samples using gradient-based perturbations (PGD or FGSM). Panels (b), (c) depict that the perturbed images are passed to quantum classifiers under white-box and black-box attack scenarios, respectively, to test the vulnerability of the classifiers against quantum adversarial attacks. Panel (d) represents that the perturbed images are first fed into a classical encoder-decoder model to purify and reconstruct the original images from the noisy images. After the reconstruction, classification is performed again in both white-box and black-box settings to test the performance of the autoencoder in counteracting adversarial effects. Panel (e) represents the QVC, which follows a three-step process: First, classical data are encoded into quantum states using amplitude encoding; next, the encoded data are processed through parameterized quantum layers consisting of single-qubit rotation gates and controlled-Z (CZ) gates; finally, a measurement layer extracts the classification result. The blue blocks represent the layers that are repeated. Here, the QVC is implemented with quantum circuit depths of 100 layers. Panel (f) shows the architecture of the classical encoder-decoder network. The encoder compresses the input into a compact, low-dimensional latent representation using convolutional layers and fully connected layers. The decoder reconstructs the original input from this latent vector through fully connected and transposed convolutional layers, producing an output image of size $28 \times 28 \times 1$.

ML models and quantum properties. However, quantum classifiers remain vulnerable to quantum-generated adversarial attacks. A detailed description and discussion of the recent relevant literature is provided in the Sec. S1 in the Supplemental Material [34]. Another limitation of the previous work on defensive approaches offers limited scope due to proof-of-concept study [30], implementation on either down-scaled MNIST data [16] or simple binary classification task [17,28], train and test with very small subset of data [17,28],

or hybrid classical-quantum frameworks [27]. Our work is the first to demonstrate that robust implementation is possible for full-scale MNIST and fashion-MNIST (FMNIST) datasets under complex projected gradient descent (PGD) [35] and fast gradient sign method (FGSM) [9] attacks, which are standard benchmarks for ML literature.

Figures 1(a)–1(d) schematically illustrate our quantum-classical approach in which quantum adversarial attacks are generated via a quantum variational circuit, which are then

passed to a pretrained quantum classifier. The adversarial images are able to trick quantum classifier to mislabel datasets. A classical encoder/decoder architecture purifies quantum adversarial images and restores the classification accuracy of the quantum classifier. In Figs. 1(e) and 1(f), we provide detailed circuit diagrams for quantum classifier and the architecture of classical encoder/decoder framework.

The validation of our technique is performed by implementing the proposed QVC-CED framework using the PennyLane and PyTorch software frameworks. The QVC model was initialized with 10 qubits and a depth of 100/200 parameterized layers, offering sufficient expressive power for complex feature extraction. The model was trained using the Adam optimizer with a learning rate of 0.005 and a batch size of 256 for 20 epochs. Cross-entropy loss was employed as the loss function to train the QVC classifier. To simulate quantum adversarial attacks, we generated gradient-based perturbations using the FGSM and PGD. The details of FGM and PGD implementations are provided in Sec. S2 in the Supplemental Material [34]. While FGSM and PGD are standard attack strategies in the classical adversarial literature, in our work, we adapt these to a quantum context by employing a QVC model where clean images are encoded into quantum states and passed through repeated parameterized layers. The gradients are calculated based on the structure of the quantum circuit, and the adversarial perturbation is generated by adjusting the input data using these gradients. This is consistent with the published literature [16,24]. The classical autoencoder, comprising a convolutional neural network (CNN)-based encoder and a decoder with transposed convolution layers, was trained using adversarially perturbed images as input and their clean counterparts as targets. The autoencoder was trained with mean squared error loss and the Adam optimizer with a learning rate of 0.001 for 20 epochs on the same datasets. All experimental simulations were conducted using a high-performance computing system equipped with a single NVIDIA GPU, enabling efficient simulation of hybrid quantum-classical workflows. For PGD attack, we employed 40 iterations for generating adversarial examples, using a step size of 0.1. The total training time for our experimental simulations was approximately 12.5 h (including both the autoencoder and quantum variational classifier training) on a single NVIDIA GPU system for MNIST dataset under the FGSM attack setting. The overall training time may vary depending on dataset size, attack type, quantum circuit depth, and available GPU memory.

In this work, we considered MNIST [36] and FMNIST [37] datasets without any preprocessing or downscaling; therefore, our benchmarks are at par with standard approaches adapted by classical literature [9,10,24]. MNIST consists of 70 000 grayscale images of handwritten digits ranging from 0 to 9, each represented as a 28×28 pixel array. The dataset is split into 60 000 training images and 10 000 test images, with each image labeled according to its corresponding digit. FMNIST is a more challenging dataset, providing a robust benchmark for ML and computer vision algorithms. It consists of 70 000 grayscale images of fashion items across ten categories, such as T-shirts, trousers, dresses, and shoes, with each image labeled according to its respective class. Similar to MNIST, the images are 28×28 pixels in size. The dataset is divided

into 60 000 training images and 10 000 test images. For both datasets, we evaluate the performance across all ten classes for the classification task.

To test the adversarial vulnerability of the QVC model against quantum attacks, we conduct both white-box and black-box attacks. In white-box scenario, the same QVC model is used for both adversarial example generation and evaluation, assuming that the attacker has access to the network. Specifically, adversarial perturbations are crafted using the QVC100 model and tested on the same model. This white-box testing framework allows to understand the model's resilience against attacks that are specifically designed for it. We generate adversarial examples using PGD and FGSM attacks. Figure 2 illustrates the examples of the original, adversarially perturbed, and reconstructed images for both datasets. The first, second, and third rows show the original images, adversarial examples, and the reconstructed images obtained from the classical autoencoder. In this figure, we highlight the results for an attack strength of 0.3, which generates strong adversarial perturbations. Figures 2(a) and 2(b) demonstrate the performance under FGSM attack for MNIST and FMNIST datasets, respectively, while Figs. 2(c) and 2(d) represent the results with PGD attack for both datasets. For both PGD and FGSM attacks, the adversarial examples are generated by our QVC100 model, which are often imperceptible to the human eye but significantly degrade the classifier's accuracy. These perturbations are strategically crafted to mislead the classifier while maintaining the visual appearance of the input. At higher attack strengths, these distortions become more prominent, leading to visible artifacts in the adversarial examples.

The classical autoencoder is designed to denoise adversarially perturbed images and restore them to their original appearance. The details about autoencoder are presented in the Sec. S2 in the Supplemental Material [34]. Adversarial perturbations shift inputs away from the distribution of clean images, introducing subtle yet targeted changes that can mislead classifiers. Autoencoder focuses on reconstructing only the most salient and representative features of the images and tends to ignore these off-manifold adversarial components during reconstruction. This effect aligns with the manifold hypothesis, where the original clean images lie on a low-dimensional surface embedded within the high-dimensional input space [38]. Adversarial examples move inputs off this surface. By projecting inputs back onto the data manifold, autoencoder effectively removes adversarial noise and restores semantic fidelity [26]. As illustrated in Fig. 2, the autoencoder significantly reduces the impact of adversarial noise, allowing the reconstructed images to closely resemble their clean counterparts, even under strong attack conditions ($\epsilon = 0.3$). For instance, from the reported results under FGSM and PGD attacks, the quantum classifier is prone to misclassification. However, after passing through the autoencoder, the reconstructed images retain essential features, mitigating adversarial distortions and preserving classification integrity.

To test the resilience of the quantum classifier, we perform the classification task on clean (when the classifier is not under attack), under adversarial attacks, and after defense across all ten classes of MNIST and FMNIST datasets. Figure 3 shows the classification accuracies of QVC model under PGD attack

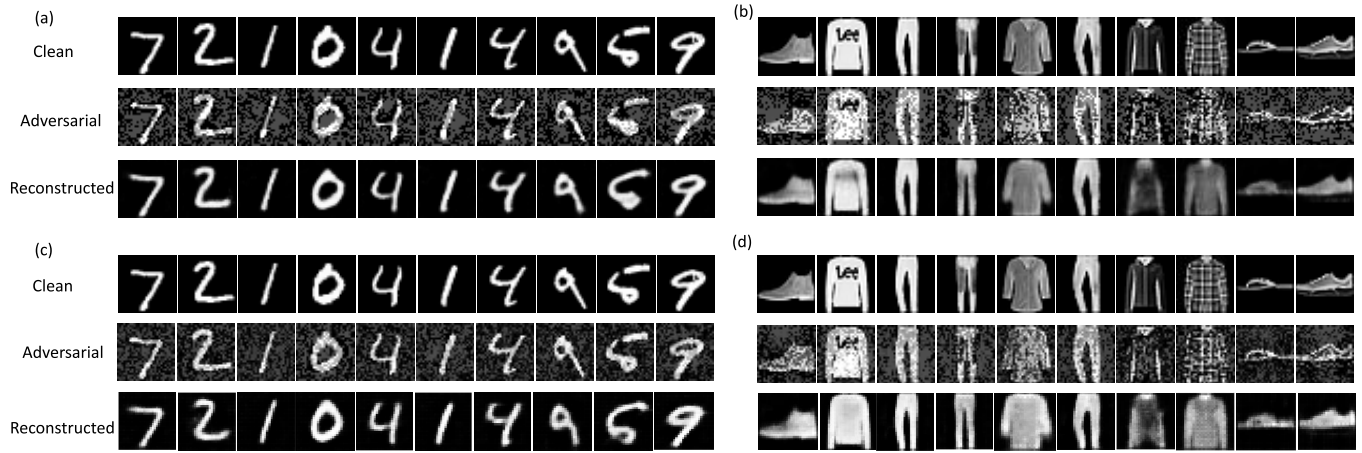


FIG. 2. Examples of the original images, adversarially perturbed images, and the corresponding reconstructed images, presented in the first, second, and third rows, respectively. For all the cases, the results are demonstrated using an attack strength of 0.3, generating strong adversarial perturbations to evaluate the performance of autoencoder under challenging conditions. Panels (a), (b) illustrate the adversarial images generated using FGSM attack on the QVC100 model for both MNIST and FMNIST datasets. Panels (c), (d) depict the adversarial images generated using PGD attack on the same model. The reconstructed images in the third row, generated by the autoencoder, effectively mitigate the adversarial noise, showcasing its robustness in reconstructing original representations.

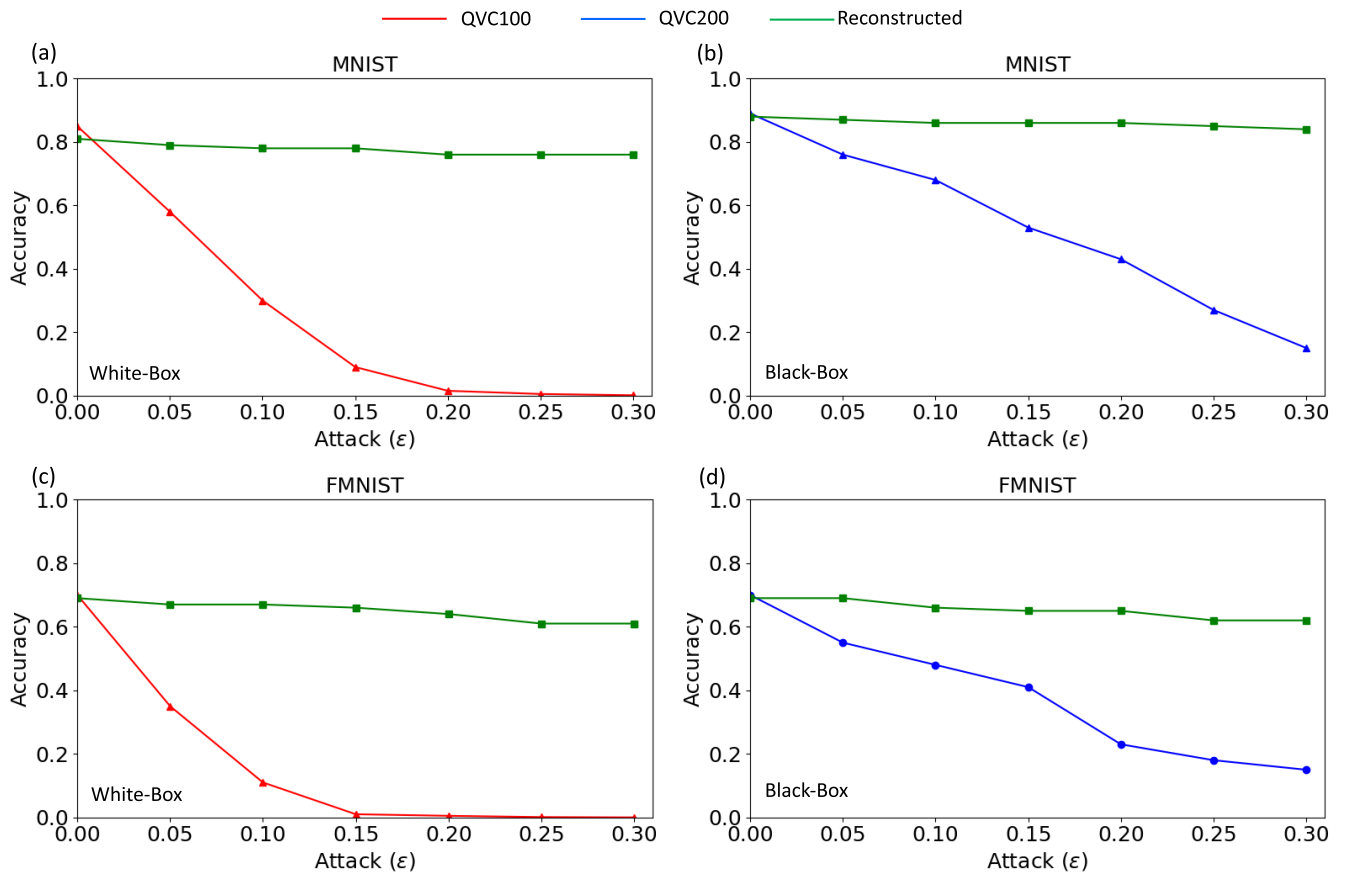


FIG. 3. Plot of classification accuracy of quantum classifier under varying attack strengths, ranging from 0 to 0.3. For all the cases, the attacks are generated by QVC100 model. Panels (a), (c) show the performance of quantum classifier under PGD attacks for the MNIST and FMNIST datasets in white-box setting, respectively. Panels (b), (d) represent the impact of PGD attacks on the quantum classifier for the same datasets in black-box scenario. We also report the reconstruction accuracy when there is no attack ($\epsilon = 0$). Across all cases, the classification accuracy decreases as the attack strength increases, highlighting the susceptibility of the models to quantum adversarial attacks. However, the results demonstrate a significant recovery in classification accuracy when the adversarial examples are passed through the classical autoencoder.

in white-box and black-box scenarios. PGD is a more iterative and powerful attack method than FGSM. The attack strength (ϵ) is varied from 0 to 0.3 to test the severity of adversarial impact. In the white-box scenario, where the attacker has full access to the QVC100 model, accuracy plummets as ϵ increases. For MNIST [Fig. 3(a)], the QVC model achieves an initial accuracy of 85%. For FMNIST [Fig. 3(c)], which is more complex as FMNIST images exhibit higher intraclass variability and interclass similarity, making them harder to distinguish, the initial accuracy is slightly lower, which is 70%. We also report the reconstruction accuracy when there is no attack; i.e., the value of ϵ is 0. Under PGD attacks, as ϵ increases, the accuracy of the QVC model drops sharply. For MNIST, accuracy falls from 85% to nearly 0% at $\epsilon = 0.3$. A similar trend is observed for FMNIST, where the accuracy decreases more rapidly, indicating higher vulnerability due to the increased complexity of fashion items compared to handwritten digits. The black-box scenario [Figs. 3(b) and 3(d)] tests the resilience of quantum classifier when adversarial examples generated by QVC100 model are evaluated on QVC200 model, simulating a realistic attack with limited knowledge of the model. For MNIST, the accuracy decreases from 89% to 15% at attack strength of 0.3, a slightly lower drop than in the white-box case, suggesting that black-box attacks are less tailored but still effective. FMNIST dataset also shows a comparable trend in black-box setting. These results highlight that adversarial perturbations exhibit strong transferability across quantum models, even without having any prior knowledge of the target model.

We also report the classification accuracy for both datasets under FGSM attack in the Supplemental Material (see Fig. S1 [34]). Under FGSM attacks, the model's robustness deteriorates like PGD attacks, highlighting its susceptibility to adversarial perturbations. The reconstruction accuracy, measured as the classifier's performance on autoencoder-reconstructed images, remains significantly higher than on adversarial examples. At strength $\epsilon = 0.3$, where adversarial noise is at its peak, the classifier's accuracy on adversarial images drops to 0% for both MNIST and FMNIST datasets in white-box settings. However, after passing through the autoencoder, the classification accuracy is substantially recovered to 80% for MNIST and 65% for FMNIST. By implementing a quantum-classical defense framework on widely used full-scale benchmarks (MNIST and Fashion-MNIST) under two attack types, single-step and iterative, our work provides demonstration that scalable defenses are possible in

QML research. The attacks used in this work are not arbitrary, but widely adopted as standard, strong first-order adversarial benchmarks in both classical and quantum adversarial ML research. Demonstrating robustness under both attacks already suggests generalization capability of the model beyond a single attack formulation. However, we leave this systematic evaluation of the generalization capability as future research direction.

The proposed QVC-CED approach of mitigating adversarial noise plays a significant role in enhancing the performance and reliability of quantum classifier. While quantum attacks expose the vulnerabilities of quantum classifiers, the classical autoencoder effectively mitigates these threats even under strong attack scenarios, reinforcing model reliability. We note that our work has not considered the impact of hardware noise, which is present in the current generation of quantum processors and therefore is aimed for the fault-tolerant quantum computing regime. The adversarial attacks investigated in our work are based on standard classical PGD and FGSM strategies implemented through quantum variational circuits. In future, it would be interesting to develop attack strategies that are based on quantum principles, which is still an open research question. Finally, another interesting line of research will be to integrate our encoder/decoder framework with quantum machine learning architectures beyond variational approaches such as quantum convolutional neural networks [39] and nonunitary approaches [40]. Nevertheless, the successful integration of classical denoising techniques with quantum classifiers reported in this work demonstrates the potential for hybrid quantum-classical architecture that leverage the strengths of both paradigms. Such methods open up possibilities for developing more robust QML models for real-world deployment where adversarial security is a critical concern.

A.K. acknowledges the use of CSIRO HPC (High-Performance Computing) for conducting all the experimental simulations. A.K. and M.U. also acknowledge CSIRO's Quantum Technologies Future Science Platform for providing the opportunity to work on quantum machine learning.

M.U. conceived the idea and supervised the project. A.K. developed and implemented the QVC-CED framework and generated all datasets. M.U. and A.K. analyzed the data and wrote the manuscript.

Data availability. The data that support the findings of this article are openly available [36,37].

- [1] M. Rana and M. Bhushan, Machine learning and deep learning approach for medical image analysis: diagnosis to detection, *Multimedia Tools Appl.* **82**, 26731 (2023).
- [2] G. Varoquaux and V. Cheplygina, Machine learning for medical imaging: Methodological failures and recommendations for the future, *npj Digit. Med.* **5**, 48 (2022).
- [3] M. Ahmed, A. N. Mahmood, and J. Hu, A survey of network anomaly detection techniques, *J. Netw. Comput. Appl.* **60**, 19 (2016).

- [4] T. Young and D. Hazarika, Soujanya Poria, and Erik Cambria, Recent trends in deep learning based natural language processing, *IEEE Comput. Intell. Mag.* **13**, 55 (2018).
- [5] H. Fujiyoshi, T. Hirakawa, and T. Yamashita, Deep learning-based image recognition for autonomous driving, *IATSS Res.* **43**, 244 (2019).
- [6] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and F. Prabhat, Deep learning and process understanding for data-driven earth system science, *Nature (London)* **566**, 195 (2019).

- [7] C. Szegedy, Intriguing properties of neural networks, [arXiv:1312.6199](#).
- [8] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, Adversarial machine learning, in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence* (Association for Computing Machinery, New York, 2011), pp. 43–58.
- [9] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, [arXiv:1412.6572](#).
- [10] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial examples in the physical world, in *Artificial Intelligence Safety and Security* (Chapman and Hall/CRC, New York, 2018), pp. 99–112.
- [11] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, Black-box adversarial attacks with limited queries and information, in *International Conference on Machine Learning* (PMLR, Cambridge, MA, 2018), pp. 2137–2146.
- [12] V. Tjeng, K. Xiao, and R. Tedrake, Evaluating robustness of neural networks with mixed integer programming, [arXiv:1711.07356](#).
- [13] A. D. M. Ibrahim, M. Hussain, and J.-E. Hong, Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective, *Artif. Intell. Rev.* **58**, 28, (2025).
- [14] J. Lee, T. Kim, S. Bang, S. Oh, and H. Kwon, Evasion attacks on deep learning-based helicopter recognition systems, *J. Sens.* **2024**, 11245981 (2024).
- [15] Y. Chen, The risk and opportunity of adversarial example in military field, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, New York, 2022), pp. 100–107.
- [16] S. Lu, L.-M. Duan, and D.-L. Deng, Quantum adversarial machine learning, *Phys. Rev. Res.* **2**, 033212 (2020).
- [17] W. Ren, W. Li, S. Xu, K. Wang, W. Jiang, F. Jin, X. Zhu, J. Chen, Z. Song, P. Zhang *et al.*, Experimental quantum adversarial learning with programmable superconducting qubits, *Nat. Comput. Sci.* **2**, 711 (2022).
- [18] W. Gong and D.-L. Deng, Universal adversarial examples and perturbations for quantum classifiers, *Natl. Sci. Rev.* **9**, nwab011 (2022).
- [19] M. Wendlinger, K. Tscharke, and P. Debus, A comparative analysis of adversarial robustness for quantum and classical machine learning models, in *IEEE International Conference on Quantum Computing and Engineering (QCE)* (IEEE, New York, 2024), Vol. 1, pp. 1447–1457.
- [20] G. Anil, V. Vinod, and A. Narayan, Generating universal adversarial perturbations for quantum classifiers, in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI Press, 2024), Vol. 38, pp. 10891–10899.
- [21] D. Winderl, N. Franco, and J. M. Lorenz, Quantum neural networks under depolarization noise: Exploring white-box attacks and defenses, *Quantum Mach. Intell.* **6**, 83 (2024).
- [22] J. Guan, W. Fang, and M. Ying, Robustness verification of quantum classifiers, in *Proceedings of the 33rd International Conference on Computer Aided Verification, CAV 2021, Virtual Event* (Springer, Berlin, Heidelberg, 2021), pp. 151–174.
- [23] N. Liu and P. Wittek, Vulnerability of quantum classification to adversarial perturbations, *Phys. Rev. A* **101**, 062331 (2020).
- [24] M. T. West, S. M. Erfani, C. Leckie, M. Sevier, L. C. L. Hollenberg, and M. Usman, Benchmarking adversarially robust quantum machine learning at scale, *Phys. Rev. Res.* **5**, 023186 (2023).
- [25] K. Bajaj, D. K. Singh, and M. A. Ansari, Autoencoders based deep learner for image denoising, *Procedia Comput. Sci.* **171**, 1535 (2020).
- [26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th International Conference on Machine Learning* (Association for Computing Machinery, New York, 2008), pp. 1096–1103.
- [27] A. Khatun and M. Usman, Quantum transfer learning with adversarial robustness for classification of high-resolution image datasets, *Adv. Quantum Technol.* **8**, 2400268 (2024).
- [28] G. Montalbano and L. Banchi, Quantum adversarial learning for kernel methods, *Quantum Mach. Intell.* **7**, 15 (2025).
- [29] W. Gong, D. Yuan, W. Li, and D.-L. Deng, Enhancing quantum adversarial robustness by randomized encodings, *Phys. Rev. Res.* **6**, 023020 (2024).
- [30] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, Quantum noise protects quantum classifiers against adversaries, *Phys. Rev. Res.* **3**, 023153 (2021).
- [31] M. T. West, A. C. Nakhil, J. Heredge, and F. M. Creevey, L. C. L. Hollenberg, M. Sevier, and M. Usman, Drastic circuit depth reductions with preserved adversarial robustness by approximate encoding for quantum machine learning, *Intell. Comput.* **3**, 0100, (2024).
- [32] Y. Wu, E. Adermann, C. Thapa, S. Camtepe, H. Suzuki, and M. Usman, Radio signal classification by adversarially robust quantum machine learning, [arXiv:2312.07821](#).
- [33] N. Dowling, M. T. West, A. Southwell, A. C. Nakhil, M. Sevier, M. Usman, and K. Mxodi, Adversarial robustness guarantees for quantum classifiers, [arXiv:2405.10360](#).
- [34] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/41z8-d3h9> for recent relevant literature, which includes Refs. [12,13], adversarial attack techniques [23,24,26–28], data encoding [29], and more simulation results.
- [35] A. Madry, Towards deep learning models resistant to adversarial attacks, [arXiv:1706.06083](#).
- [36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* **86**, 2278 (1998).
- [37] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, [arXiv:1708.07747](#).
- [38] G. Alain and Y. Bengio, What regularized auto-encoders learn from the data-generating distribution, *J. Mach. Learn. Res.* **15**, 3563 (2014).
- [39] I. Cong, S. Choi, and M. D. Lukin, Quantum convolutional neural networks, *Nat. Phys.* **15**, 1273 (2019).
- [40] J. Heredge, M. West, L. Hollenberg, and M. Sevier, Non-unitary quantum machine learning, *Phys. Rev. Appl.* **23**, 044046 (2025).