Experimental quantum adversarial learning with programmable superconducting qubits

Quantum computing promises to enhance machine learning and artificial intelligence. However, recent theoretical works show that, similar to traditional classifiers based on deep classical neural networks, quantum classifiers would suffer from adversarial perturbations as well. Here we report an experimental demonstration of quantum adversarial learning with programmable superconducting qubits. We train quantum classifiers, which are built on variational quantum circuits consisting of ten transmon qubits featuring average lifetimes of 150 μs, and average fidelities of simultaneous single- and two-qubit gates above 99.94% and 99.4%, respectively, with both real-life images (for example, medical magnetic resonance imaging scans) and quantum data. We demonstrate that these well-trained classifiers (with testing accuracy up to 99%) can be practically deceived by small adversarial perturbations, whereas an adversarial training process would substantially enhance their robustness to such perturbations.

Quantum computing and artificial intelligence (AI) have made dramatic progress in recent years. This has given rise to an emergent research frontier called quantum machine learning or, generally, quantum AI, which has attracted considerable attention across communities. In classical machine learning it has been shown that classifiers based on deep neural networks are vulnerable to adversarial perturbations. Adding tiny carefully crafted perturbations to the legitimate original data samples misleads the classifiers to make wrong predictions. For example, a stop sign with some small graffiti might be misclassified as a yield sign, whereas adding a tiny amount of carefully crafted noise into an image of a benign skin lesion—perhaps even imperceptible to the human eye—would fool the classifier to predict it as malignant. This surprising vulnerability of classical neural networks has far-reaching consequences in safety- and security-critical scenarios (for example, autonomous driving, biometric authentication and medical diagnostics). More recently, the vulnerability of quantum classifiers has been studied, establishing the foundations of quantum adversarial machine learning. It has been shown theoretically that quantum classifiers are likewise highly vulnerable to adversarial examples, independent of the learning algorithms and regardless of whether the input data are classical or quantum. Different countermeasures, such as adversarial training, have also been...
proposed to enhance the robustness of quantum classifiers against adversarial perturbations.

Experimentally demonstrating adversarial examples for quantum classifiers and showing the effectiveness of the proposed countermeasures in practice are challenging feats. To accomplish them, one faces at least two difficulties: (1) determining an experimentally feasible encoding of high-dimensional classical data and (2) building quantum classifiers with a large enough state-space so as to identify realistic images, despite inevitable experimental noises. In this Brief Communication we overcome these difficulties and report an experimental demonstration of quantum adversarial learning with a superconducting quantum processor. We train the quantum classifiers with both large-size realistic images (for example, medical magnetic resonance imaging (MRI) scans) and high-dimensional quantum data (for example, thermal and localized quantum many-body states), through quantum gradients obtained directly by measuring some observables. After training, these classifiers can achieve high performance on these datasets, with a testing accuracy of up to 99%. We generate adversarial examples through a classical optimizing procedure and unambiguously show that they can deceive the trained quantum classifiers with a high confidence level. To mitigate such vulnerability, we further demonstrate that, through adversarial training, the quantum classifiers will be immune to adversarial perturbations generated by the same attacking strategy.

Our experiment is implemented on a flip-chip superconducting quantum processor (Fig. 1a), which has 36 transmon qubits arranged in a two-dimensional (2D) array featuring tunable nearest-neighbor couplings. For the purpose of demonstrating quantum adversarial learning, we chose a 1D array of ten qubits. By optimizing the device fabrication and controlling process, we push the average lifetime of these qubits to 150 μs and the average simultaneous single- and two-qubit gate fidelities to greater than 99.994% and 99.4%, respectively. This enables us to successfully implement large-scale quantum classifiers with different structures up to a circuit depth of 60 and a number of trainable variational parameters exceeding 250.

We start with the vulnerability of quantum learning systems in medical diagnostics and consider a binary classification task for identifying magnetic resonance imaging (MRI) images. We exploit an interleaved block-encoding scheme rather than the conventional amplitude encoding to encode the input classical data (Methods and Supplementary Section IA). This enables us to circumvent the notorious difficulty of preparing a highly entangled multiqubit quantum state and is crucial for the success of classifying large-sized images (16 × 16 pixels; Fig. 1b, left) with a large-scale quantum classifier (up to 260 trainable parameters) with state-of-the-art (but still rather limited) gate fidelities. The interleaved block-encoding and the structure of our quantum classifier are illustrated in Fig. 1b (right). We train our quantum classifier with MRI images labeled as ‘Hand’ and ‘Breast’, through quantum gradients obtained directly by measuring some observables in our experiment (Supplementary Section IA). Our experimental result for the training process is plotted in Fig. 1c, from which it is clear that the accuracy for both the training and test datasets increases rapidly at the beginning of the training process and then saturates at a high value (0.92 and 0.97 for the training and test datasets, respectively). In Fig. 1d we plot the measured expectation values (⟨σz⟩) of the Pauli-Z operator on the fifth qubit, which determines the assigned observables in our experiment (Supplementary Section IA). Our experiments illustrate in Fig. 1b (right) that the accuracy for both the training and test datasets from the test set, as a function of epochs during the adversarial training process. We find that it increases for both datasets and approaches unity after ~25 epochs, indicating that the adversarially retrained quantum classifier becomes immune to adversarial perturbations. To be more concrete, in Fig. 1h (left) we plot a randomly chosen adversarial example. This image will be misclassified by the original quantum classifier into the category of ‘Breast’ (with ⟨σz⟩ = −0.26), but after adversarial training it will be identified correctly as ‘Hand’ (with a refreshed ⟨σz⟩ value of 0.18). This shows explicitly that adversarial training can indeed enhance the robustness of quantum classifiers against adversarial perturbations.

Unlike classical classifiers, which can only take classical data as input, quantum classifiers can also naturally handle quantum states as input and gain potential exponential advantages. We now turn to the vulnerability of quantum classifiers in classifying quantum states. For concreteness, we consider a binary classification of quantum states generated by evolving the Néel state for a period of time with the following Aubry–André Hamiltonian:

\[ H = -\frac{g}{\hbar} \sum_k (\sigma_x^k \sigma_x^{k+1} + \sigma_y^k \sigma_y^{k+1}) - \sum_k \frac{V}{2} \sigma_z^k, \]

where \( g \) is the coupling strength, \( \sigma_z^k \) (\( l = x, y, z \)) is the Pauli operator for the \( k \)th qubit, and \( V_k = V \cos(2\pi k + \phi) \) is the incommensurate potential, with \( V \) being the disorder magnitude, \( \alpha = (\sqrt{5} - 1)/2 \) being an irrational number and \( \phi \) being a random phase evenly distributed on \([0, 2\pi)\). This Hamiltonian features a quantum phase transition at \( |V/g| > 1 \), between a localized phase for \( |V/g| < 2 \) and a delocalized (thermal) phase for \( |V/g| < 2 \) (ref. 13). In our experiment, we initialize the system to the Néel state and then evolve it under \( H \) for ~400 ns, with the pulse sequence sketched in Fig. 2a.

We fix \( g/2\pi = 5 \) MHz and scan \( V/2\pi \) from 0 MHz to 30 MHz. In Fig. 2b we plot the measured probability \( P_l \) of being on state \(|l\rangle\) for each qubit (equivalent to the local magnetization \( \langle \sigma_z \rangle \) by noting \( P_l \equiv \frac{1}{2} (\langle \sigma_z \rangle + 1) \) for varying \( V \), from which the localized and thermal features of the evolved states are clearly manifested. We randomly choose some of the evolved quantum states deep in the localized and thermal regions (dashed gray boxes, Fig. 2b) to form a quantum dataset. We implement a quantum classifier, which consists of five layers, each containing three single-qubit rotations and two layers of controlled- NOT (CNOT) gates, to classify the chosen states in a supervised fashion (Supplementary Section II). We randomly initialize the 150 variational
Fig. 1 | Experimental quantum adversarial learning. a, Schematic illustration. Although the quantum classifier running on the superconducting quantum processor will correctly identify the legitimate MRI scan as ‘Malignant’, it will incorrectly classify the corresponding adversarial example, which differs by only an imperceptible amount of perturbation, into the ‘Benign’ class with high confidence. b, Interleaved block-encoding of the medical hand-breast MRI data and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. We compress each MRI image to 16 × 16 pixels, which is represented by a 256D vector \( x \) and a schematic of the experimental quantum circuits. 

parameters and train the quantum classifier with experimentally obtained quantum gradients. Figure 2c plots the accuracy and loss as a function of epochs obtained in our experiment during the training process. We find that the implemented quantum classifier has a desirable performance and, after ~30 iteration steps, it achieves near-perfect accuracy on both the training and test datasets.

We generate adversarial perturbations for state samples in the test set by solving an optimization problem with quantum gradients measured in experiments (Methods and Supplementary Section III). We add the obtained perturbations to their corresponding legitimate states by adding a near-identity unitary before inputting the states into the quantum classifier. In the first row of Fig. 2d we randomly choose 20 states from the training set and plot their measured \( \sigma_z \) values for each qubit, for both the legitimate (left) and adversarial (right) samples. In this figure, the adversarial examples differ slightly from the legitimate ones (for those in the thermal region in particular, the difference is indiscernibly small) and maintain the essential features (vanishing and persistent local magnetization) for thermal and localized states, respectively. However, they would successfully deceive the quantum classifier with a very large probability. From the left part of Fig. 2d it is clear that the trained classifier can correctly identify all the legitimate states, but it will misclassify all of the adversarial examples in the localized region and half in the thermal region, as shown in the right of Fig. 2d, even though the essential features of the local magnetization...
Aubry–André model (inset) and wait for 400 ns until the system evolves into the frequency of each qubit to engineer the incommensurate potential of the [4, 5] (gray boxes), respectively, with random \( \phi \c\). c. Loss function (top) and accuracy (bottom) for the test and training sets at each epoch. The data of loss function for training (test) is presented as mean values ± standard error of mean with a sample size of \( n = 20 \) \( (n = 50) \). d. Vulnerability of the quantum classifier in learning quantum states. We select ten legitimate \( |T\rangle \) and \( |L\rangle \) states from the training set, whose local magnetization distribution and classification outputs are shown in the top and bottom left panels. After applying adversarial perturbations on the legitimate states, half of the \( |T\rangle \) and all the \( |L\rangle \) states are classified incorrectly by the trained classifier (bottom right).

distribution are still clearly distinct for thermal and localized regions. This demonstrates the vulnerability of quantum classifiers to adversarial perturbations in categorizing quantum states.

We stress that adversarial perturbations are not random noises. They are carefully engineered to mislead quantum classifiers. In fact, random noises, either from quantum hardware imperfections or added on purpose, will diminish the adversarial effect in general and can even be exploited to protect quantum classifiers from adversarial perturbations (Supplementary Section IB3). In our experiment we optimized the device fabrication and controlling process to minimize hardware noises, which is crucial for both implementation of the large-scale quantum classifiers and demonstration of their vulnerability to adversarial perturbations. We mention that, although no quantum advantage has been achieved in our current experiment, quantum adversarial learning holds the potential to exhibit advantages over its classical counterparts when scaling up. For example, conceivable quantum speed-ups hosted by some quantum machine learning algorithms\(^{22-24}\) may carry over to adversarial scenarios. In addition, the unconditional security promised by the protocols of blind quantum computation should also be explored to defend against adversarial attacks with potential quantum advantages. Without a doubt, the demonstration and verification of quantum advantages in quantum adversarial learning scenarios remains a challenging task and is worth further investigation.

Theoretically, the existence of adversarial examples has an origin in the fundamental ‘concentration of measure phenomenon’\(^{19}\) and is hence an inevitable feature of quantum machine learning with high-dimensional data\(^{6-8}\), independent of the learning models, the training algorithms and whether the input data are classical or quantum. In this Brief Communication our discussion is mainly focused on supervised learning based on quantum circuit classifiers. The experimental demonstration of quantum adversarial examples for unsupervised learning and other types of quantum classifier\(^{20}\) seems more technically sophisticated and remains unattainable. In addition, other defense strategies such as defensive distillation\(^{21}\) and defense-GAN (generative adversarial network)\(^{22}\) have also been introduced in the classical adversarial machine learning literature. It would be interesting and important to extend these strategies to the quantum domain, both in theory and experiment.

How to build a trustworthy quantum AI system and show its advantage in practical applications remains largely unclear and demands long-term research. Our results make an experimental attempt along this line by not only revealing the vulnerability of quantum learning systems in adversarial scenarios, but also by demonstrating the effectiveness of a defense strategy against adversarial attacks in practice. As the fledgling field of quantum AI grows, our results will prove useful in practical applications, especially for scenarios that are safety- and security-critical.

**Methods**

**Framework and experimental set-up**

We first introduce the general framework for quantum adversarial machine learning. We consider classification tasks in the setting of supervised learning\(^{23-25}\), where we train quantum
classifiers with pre-labeled data samples by minimizing the following loss function iteratively:

$$\mathcal{L}(h(\mathbf{x}; \theta), \mathbf{a}) = -\sum_k a_k \log g_k.$$  

(2)

Here, \(x\) denotes a training sample, \(h(\mathbf{x}; \theta)\) represents the hypothesis function determined by the quantum classifier with variational parameters denoted collectively as \(\theta\). \(a\) is the one-hot encoding of the labels, and \(g_k\) denotes the probability for the \(k\)th category obtained from measuring the quantum classifier. To minimize the loss function, we adapt the gradient descent method. Here, computing the derivatives of \(\mathcal{L}\) with respect to the circuit parameters can be transformed into computing the derivatives of some expectation values with respect to these circuit parameters according to the chain rule. In our case, it can be formally expressed as

$$\frac{d\mathcal{L}(h(\mathbf{x}; \theta), \mathbf{a})}{d\theta_i} = -\sum_k a_k \frac{\partial g_k}{\partial \theta_i}.$$  

(3)

The next step that computes the derivatives of \(g_k\) with respect to the circuit parameters can be accomplished with the ‘parameter shift rule’, as \(g_k\) can be regarded as an expectation value of an observable that we denote as \(B_k\) (here (refs. 28–30)). This rule states that if a gate with parameter \(\theta_i\) is in the form \(g(\theta_i) = e^{-\frac{i}{2} \theta_i}\), with \(P_i\) being an \(\alpha\)-qubit Pauli string, the derivative can be evaluated by

$$\frac{\partial g_k}{\partial \theta_i} = \frac{\delta e^{-\frac{i}{2} \theta_i}}{\delta \theta_i} = \frac{(B_k^+)^2 - (B_k^-)^2}{2},$$  

where \((B_k^+)^2\) denotes the expectation values of \(B_k\) with the parameter \(\theta_i\) replaced by \(\theta_i = \pm \frac{\pi}{2}\). Thus, because the parameters in our case are all encoded in the angles of single-qubit Pauli-rotation gates, we can optimize the quantum classifier with gradients obtained from measurements.

After the training process, the quantum classifier will typically be able to assign labels to data samples outside the training set with high accuracy. To obtain adversarial examples, we focus on the scenario of untargeted white-box attacks, where we assume the attacker has full information about the quantum classifier and no particular class is targeted. Unlike the training process, where we vary the variational parameters to minimize the loss, for generating adversarial examples we fix \(\theta\) at its optimal value \(\theta^*\) obtained in the last step of the training, and optimize over the input space within a small region to maximize the loss function instead (Supplementary Section IB):

$$\delta = \arg\max_{\delta \neq \mathbf{a}} \mathcal{L}(h(\mathbf{x} + \delta; \theta^*), \mathbf{a}),$$  

where \(\Delta\) denotes a small region introduced to ensure that the adversarial perturbations are small and will not essentially alter the input data. We input the generated adversarial examples into the quantum classifier to test its performance. A schematic illustration of the main idea for quantum adversarial learning is shown in Fig. 1a.

To achieve high coherence of the superconducting quantum processor, we deposited tantalum films \(^{30}\) using a high-vacuum sputtering system (Yunmao QBT-P). These were then patterned to create the qubit structures. The energy relaxation times \(T_1\) of the ten qubits ranged from 131 to 173 \(\mu\)s at the frequencies where the qubits were initialized and operated. Single-qubit XY rotations were realized using 30-ns-long microwave pulses, generated by multichannel arbitrary waveform generators (MOSTFIT MF:AWG-08), and the CNOT gate was based on a controlled-\(\pi\) phase (CZ) gate plus single-qubit rotations. The CZ gate, which has a length of 60 ns, was realized by carefully tuning the frequencies and coupling strength of the qubits to steer a closed-cycle diabatic transition of \(|\uparrow\rangle \rightarrow |\downarrow\rangle\) (or \(|\downarrow\rangle \rightarrow |\uparrow\rangle\)).

Quantum classifiers with classical data

Here we introduce details of the settings of the quantum classifier for the classical dataset. The quantum classifier is composed of several blocks, and each block contains several layers of single-qubit gates and ends with two layers of CNOT gates that entangle all the qubits. For each block, as shown in Fig. 1b (right), the single-qubit gates are utilized to encode both trainable parameters and the input data. To encode the image information from the medical MRI dataset \(\mathcal{X}\) into the quantum classifier, we first compress the images down to 16 \(\times\) 16 pixels, which are then normalized and mapped into the rotation angles of the single-qubit gates in the quantum classifier by a factor of two. For concreteness, because we are using a ten-qubit quantum classifier, we use 26 layers of single-qubit variational gates to encode the 256D data by adding four ‘0’s at the end of the data vectors. For each rotation angle that encodes the input vectors, we attach one trainable parameter that can be optimized with gradient descent methods.

For the hyperparameter setting of the experimental demonstrations, we select the ‘Hand’ and ‘Breast’ MRI images from the medical dataset. The sizes of the training and test sets are 500 and 100, respectively. To measure the distance between the current output and the target label, we choose cross-entropy as the loss function (equation (2)), and the learning rate is set to 0.05.

The quantum classifier is initialized with randomly generated trainable parameters. During the training process, we divide the 260 trainable parameters into ten groups. For each epoch, we update the parameters in these groups sequentially. To train the parameters in each group, we randomly select 20 (50) samples from the training (test) set, where the 20 samples from the training set are utilized to calculate the gradients and optimize the parameters in the classifier, and the 50 test samples are utilized to approximately calculate the test accuracy. The loss function and accuracy of both training and test data measured at each epoch are plotted in Fig. 1c. As the loss function decreases slowly during the learning process, the accuracy increases at a relatively faster speed and approaches saturated values after about five epochs. Further decrease of the loss function helps to enhance the separation between the two categories, as witnessed by the instances in Fig. 1d. After 20 epochs, the trained quantum classifier is able to classify the total training (test) set with an accuracy of 0.92 (0.97). We note that, to minimize the circuit depth, we recompile the quantum circuit before the actual execution by replacing the single-qubit gates with two gates, that is, \(R_x(\alpha)\) and \(R_y(\phi)\) (Supplementary Section IIIIB). Moreover, dynamical decoupling pulses are applied on the qubits during their idling times in the quantum circuits.

We mention that, in addition to the learning task for the medical data in the main text, we have also demonstrated quantum adversarial learning for the MNIST handwritten-digit dataset \(^{30,31}\) to examine the feasibility of our protocol. For this task, the basic quantum circuit settings are the same as for the medical dataset, and images of digits ‘0’ and ‘1’ are selected to form the training and test sets. For experimental convenience, we only choose 50 of these parameters to be trained; these lie at the 3rd, 6th, 11th, 17th and 23rd single-qubit layers of the quantum classifier. The experimental results for learning the MNIST handwritten-digit dataset are presented in Supplementary Fig. 14a,b. We plot the loss function and accuracy of both the training and test data measured at each epoch. After the training process, the trained quantum classifier is able to classify the total training (test) set with an accuracy of 0.98 (0.99).
Quantum classifiers with quantum data

On our device, the frequency of each qubit and the coupling strength between neighboring qubits are programmable with high flexibility, so we can synthesize the Aubry–André Hamiltonian (equation (1)) and modulate its relevant coefficients such as the coupling strength and the on-site disorder $V_i$ in arbitrary manners. Experimentally, we fix $g$ by setting the coupler frequencies and applying the desired flux bias to each qubit to vary $V_i$ as a cosine function over $k$. We mention that, due to the crosstalk effect, there are small but non-negligible interactions between the next-nearest-neighbor qubits (Supplementary Section IIIA).

With the experimental settings introduced above, we construct the training (test) set with 500 (100) quantum states, where half of the states come from the localized phase and the remaining half from the delocalized phase. The classifier is composed of five blocks and contains a total of 150 training parameters encoded in the single-qubit rotation angles (Supplementary Fig. 10 in Supplementary Section IIIB presents the full circuit for the classifier). The training parameters are divided into ten groups, each containing 15 parameters and trained sequentially at each epoch. For each group we randomly select 20 (50) samples to form the training (test) set.

Adversarial training

The adversarial examples aim to mislead the well-trained quantum classifier to make incorrect predictions. In general, these adversarial examples are generated by adding carefully designed but imperceptible perturbations to the original samples. To generate these adversarial perturbations in our work we designed several untargeted white-box attack strategies for both the classical and quantum data, which are described in detail in Supplementary Sections IB and IIB. Essentially, the perturbation is designed to maximize the loss function, which is in line with maximizing the distance between the model’s output and the corresponding correct label, that is, effectively deceiving the classifier to make incorrect classifications. In our work, we utilize this idea and apply gradient ascent methods assisted by the Adam optimizer to generate adversarial perturbations. The attacking strategies for the classical dataset and the quantum dataset are discussed in the following.

First, we consider the case of classical data. For each sample in the training (test) set with size 500 (100), we numerically generate a corresponding adversarial example on a classical computer aiming to mislead the well-trained classifier to make an incorrect prediction. We calculate the gradients of the loss function with respect to the input sample and use gradient ascent to maximize the loss function. For concreteness, two strategies are applied to generate two types of adversarial example, namely, type-1 and type-2 examples (Supplementary Section IB provides the detailed algorithms). These generated adversarial examples are then processed by the quantum classifier. As shown in Fig. 1e,f, we experimentally verify the effectiveness of these adversarial examples, where the quantum classifier tends to assign incorrect labels to them. We also provide supplementary experimental demonstrations of adversarial examples with the MNIST handwritten-digit dataset in Supplementary Section IIIC and Supplementary Fig. 14c, from which we can see that the slightly perturbed handwritten digits successfully deceive the quantum classifier. We mention that this procedure requires high-quality superconducting quantum processors so that the adversarial examples generated by a classical computer can still deceive the quantum classifier, despite the inevitable experimental noises.

Second, to generate the adversarial examples for quantum data, we add local perturbation, parameterized by three single-qubit gates $(R_x(δ), R_y(δ), R_z(δ))$ with $δ ∈ [−0.5, 0.5]$, to each qubit before training the system to evolve under the Aubry–André Hamiltonian. These perturbations are optimized experimentally to maximize the loss function, that is, to mislead the quantum classifier to make incorrect predictions. To ensure that the locally perturbed states maintain the original states’ property (localizable or thermal), we compare the states before and after adding adversarial perturbations experimentally (Fig. 2d). For more information about generating adversarial examples for both classical data and quantum data, see the detailed algorithms in Supplementary Sections IB and IIB.

We now introduce the settings for the adversarial training of quantum classifiers. The basic idea is to mix the adversarial samples and the original samples to construct new training and test sets$^{36}$. We start the training by re-initializing the 260 trainable parameters with random values. At each training epoch, we randomly select ten samples from the original dataset and ten from the adversarial dataset to form a training batch. The learning rate and the optimization strategies remain the same as those in the original training procedure. After the retraining process, the accuracy measured at each training step for both the original and adversarial samples from the test sets is shown in Fig. 1g, with a specific example shown in Fig. 1h. It turns out that, even though these samples have not been seen before (that is, they are not involved in the training process), the retrained classifier is able to identify them with high accuracy. This implies that the retrained classifier is robust against the corresponding adversarial attacks. The same adversarial training was successfully implemented with the MNIST handwritten-digit dataset, with the obtained experimental results shown in Supplementary Fig. 14d of Supplementary Section IIIC.

Data availability

The data presented in the figures and supporting the other findings of this study are available for download at https://doi.org/10.5281/zenodo.7134599 with a citable release at ref. $^{36}$. Our work makes use of three publicly available datasets introduced in previous studies$^{32–34}$. The hand and breast data from the medical hand–breast MRI data-set were originally adapted from publicly available datasets from the Radiological Society of North America (RSNA)$^{32}$; https://doi.org/10.1148/radiol.2018180736 and The Cancer Imaging Archive (TCIA)$^{33}$; https://doi.org/10.1007/s10728-013-9622-7, respectively. The MNIST handwritten-digit data were obtained from https://doi.org/10.1109/MSP.2012.2211477$^{34}$. Source data are provided with this paper.

Code availability

All the codes used for numerical simulations and experimental data analysis are available at https://doi.org/10.5281/zenodo.7134599 with a citable release at ref. $^{36}$.

References

1. Biamonte, J. et al. Quantum machine learning. Nature 549, 195–202 (2017).
2. Dunjko, V. & Briegel, H. J. Machine learning and artificial intelligence in the quantum domain: a review of recent progress. Rep. Prog. Phys. 81, 074001 (2018).
3. Sarma, S. D., Deng, D.-L. & Duan, L.-M. Machine learning meets quantum physics. Phys. Today 72, 48 (2019).
4. Eykholt, K. et al. Robust physical-world attacks on deep learning visual classification. In Proc. IEEE Conference on Computer Vision and Pattern Recognition 1625–1634 (IEEE, 2018).
5. Finlayson, S. G. et al. Adversarial attacks on medical machine learning. Science 363, 1287–1289 (2019).
6. Lu, S., Duan, L.-M. & Deng, D.-L. Quantum adversarial machine learning. Phys. Rev. Res. 2, 033212 (2020).
7. Liu, N. & Wittek, P. Vulnerability of quantum classification to adversarial perturbations. Phys. Rev. A 101, 062331 (2020).
8. Gong, W. & Deng, D.-L. Universal adversarial examples and perturbations for quantum classifiers. Nati Sci. Rev. 9, nwab130 (2021).
9. Guan, J., Fang, W. & Ying, M. in Computer Aided Verification, Lecture Notes in Computer Science (eds Silva, A. & Leino, K. R. M.) 151–174 (Springer, 2021).
10. Liao, H., Convy, I., Huggins, W. J. & Whaley, K. B. Robust in practice: adversarial attacks on quantum machine learning. *Phys. Rev. A* **103**, 042427 (2021).

11. Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J. & Sweke, R. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum* **5**, 582 (2021).

12. Haug, T., Self, C. N. & Kim, M. S. Large-scale quantum machine learning. Preprint at https://arxiv.org/abs/2108.01039 (2021).

13. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. & Latorre, J. I. Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020).

14. Aubry, S. & André, G. Analyticity breaking and Anderson localization in incommensurate lattices. *Ann. Israel Phys. Soc.* **3**, 18 (1980).

15. Gao, X., Zhang, Z.-Y. & Duan, L.-M. A quantum machine learning algorithm based on generative models. *Sci. Adv.* **4**, eaat9004 (2018).

16. Liu, Y., Arunachalam, S. & Temme, K. A rigorous and robust quantum speed-up in supervised machine learning. *Nat. Phys.* **17**, 1013–1017 (2021).

17. Saggio, V. et al. Experimental quantum speed-up in reinforcement learning agents. *Nature* **591**, 229–233 (2021).

18. Huang, H.-Y. et al. Quantum advantage in learning from experiments. *Science* **376**, 1182–1186 (2022).

19. Ledoux, M. in *The Concentration of Measure Phenomenon* 89 (American Mathematical Society, 2001).

20. Li, W. & Deng, D.-L. Recent advances for quantum classifiers. *Sci. China Phys. Mech. Astron.* **65**, 220301 (2022).

21. Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proc. 2016 IEEE Symposium on Security and Privacy* (SP) 582–597 (IEEE, 2016).

22. Samangouei, P., Kabkab, M. & Chellappa, R. Defense-GAN: protecting classifiers against adversarial attacks using generative models. In *Proc. 6th International Conference on Learning Representations* (ICLR, 2018); https://arxiv.org/abs/1805.06605

23. Havlíček, V. et al. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209–212 (2019).

24. Gong, M. et al. Quantum neural sensing of quantum many-body states on a 61-qubit programmable superconducting processor. Preprint at https://arxiv.org/abs/2201.05957 (2022).

25. Herrmann, J. et al. Realizing quantum convolutional neural networks on a superconducting quantum processor to recognize quantum phases. *Nat. Commun.* **13**, 4144 (2022).

26. Mitarai, K., Negoro, M., Kitagawa, M. & Fujii, K. Quantum circuit learning. *Phys. Rev. A* **98**, 032309 (2018).

27. Li, J., Yang, X., Peng, X. & Sun, C.-P. Hybrid quantum-classical approach to quantum optimal control. *Phys. Rev. Lett.* **118**, 150503 (2017).

28. Schuld, M., Bergholm, V., Gogolin, C., Izaac, J. & Killoran, N. Evaluating analytic gradients on quantum hardware. *Phys. Rev. A* **99**, 032331 (2019).

29. Place, A. P. M. et al. New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds. *Nat. Commun.* **12**, 1779 (2021).

30. Sung, Y. et al. Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler. *Phys. Rev. X* **11**, 021058 (2021).

31. Foxen, B. et al. Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms. *Phys. Rev. Lett.* **125**, 120504 (2020).

32. Halabi, S. S. et al. The RSNA pediatric bone age machine learning challenge. *Radiology* **290**, 498–503 (2019).

33. Clark, K. et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).

34. Deng, L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **29**, 141–142 (2012).

35. Kurakin, A., Goodfellow, I. J. & Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations* (ICLR, 2017).

36. Li, W. Data and codes for the paper titled ‘Experimental quantum adversarial learning with superconducting qubits’ (Zenodo, 2022); https://doi.org/10.5281/zenodo.7134599

**Acknowledgements**

We thank L.-M. Duan and S. Lu for helpful discussions, and V. Dunjko in particular for his valuable feedback from reading the first version of this paper. The device was fabricated at the Micro-Nano Fabrication Center of Zhejiang University. We acknowledge the support of the National Natural Science Foundation of China (grants nos. 92065204, U20A2076, 11725419, 12174342 and 12075128), the Zhejiang Province Key Research and Development Program (grant no. 2020C01019) and the Fundamental Research Funds for the Zhejiang Provincial Universities (grant no. 2021ZZX003). J.D.B acknowledges support from the research project Leading Research Center on Quantum Computing (agreement No. 014/20). D.-L.D also acknowledges additional support from the Shanghai Qi Zhi Institute.

**Author contributions**

D.-L.D. proposed the idea and laid out the theoretical framework for the experiment. W.R. and S.X. carried out the experiments, supervised by C.S. and H.W. H.L. fabricated the device, supervised by H.W. W.L. and W.J. performed the numerical simulations, supervised by D.-L.D. All authors contributed to the analysis of data, discussions of the results and the writing of the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00351-9.

**Correspondence and requests for materials** should be addressed to Chao Song, Dong-Ling Deng or H. Wang.

**Peer review information** *Nature Computational Science* thanks Leonardo Banchi, Lucas Lamata and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jie Pan, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022