Hybrid quantum learning with data reuploading on a small-scale superconducting quantum simulator

Aleksei Tolstobrov [1,2,*] Gleb Fedorov,[1,2,3] Shtefan Sanduleanu [1,2,3] Shamil Kadyrmetov [1] Andrei Vasenin [4,1]
Aleksey Bolgar,[4,1] Daria Kalacheva [4,1,3] Viktor Lubsanov [1] Aleksandr Dorogov [1] Julia Zotova [4,1,3] Peter Shlykov,[1]
Aleksei Dmitriev,[1,3] Konstantin Tikhonov,[5] and Oleg V. Astafiev [4,1]

[1]*Laboratory of Artificial Quantum Systems, Moscow Institute of Physics and Technology, 141700 Dolgoprudny, Russia*
[2]*Russian Quantum Center, 121205 Skolkovo, Moscow, Russia*
[3]*Laboratory of Superconducting Metamaterials, National University of Science and Technology 'MISIS', 119049 Moscow, Russia*
[4]*Center for Engineering Physics, Skolkovo Institute of Science and Technology, 121205 Moscow, Russia*
[5]*L. D. Landau Institute for Theoretical Physics, 142432 Chernogolovka, Russia*

Supervised quantum learning is an emergent multidisciplinary domain bridging between variational quantum algorithms and classical machine learning. Here, we study experimentally a hybrid classifier model using quantum hardware simulator (a linear array of four superconducting transmon artificial atoms) trained to solve multilabel classification and image recognition problems. We train a quantum circuit on simple binary and multilabel tasks, achieving classification accuracy around 95%, and a hybrid quantum model with data reuploading with accuracy around 90% when recognizing handwritten decimal digits. Finally, we analyze the inference time in experimental conditions and compare the performance of the studied quantum model with known classical solutions.

I. INTRODUCTION

Over the last years, high attention has been paid to the idea of using parameterized quantum circuits (PQC) as universal approximating models for machine learning (ML) tasks [1–4], inspired by research on variational quantum algorithms (VQA). There are many types of VQA [5] from which quantum neural networks (QNN) seem to be the most attractive for solving classical ML problems. While various architectures for QNNs have been suggested, including convolutional [6], generative-adversarial [7,8], recurrent [9] networks, it is still not clear how to overcome the general trainability issues [10–13] when such models have high quantum volume; this is an area of active research [14–16]. In recent works it has been shown that a small-scale QNN can outperform its classical counterpart with a close number of trainable parameters [17,18]. However, it is not yet known whether optimizing a parametrized quantum circuit can lead to an algorithm that may outperform any state-of-the-art classical algorithm. It is known that even a single-qubit quantum circuit is enough to solve nontrivial classification tasks [19,20], there is potential in using quantum kernel estimation to support vector machines [21,22], and it is supposed that PQC-based models may have advantages in expressivity and generalization [23–27]. Also, PQC unitarity automatically ensures effective weight normalization [1], which is useful for recurrent models. To date, the idea has found a few experimental realizations on

various physical platforms, the most prominent being trapped ions and superconducting artificial atoms [3,28–31].

To achieve quantum advantage for a certain ML task is to find a PQC that will train or perform inference faster, or with less resources, or with higher accuracy than is currently possible using classical computational devices. For classification problems, this is achievable probably only with a classically intractable $\boldsymbol{\theta}$-parameterized ($\boldsymbol{\theta} \in \mathbb{R}^m$) PQC used to map a feature vector $\mathbf{x} \in \mathbb{R}^n$ to a higher-dimensional feature Hilbert space vector $|\Phi(\mathbf{x}, \boldsymbol{\theta})\rangle$, or, more generally, a density matrix $\rho(\mathbf{x}, \boldsymbol{\theta})$ lying in the space of unit-trace Hermitian operators [2–4]. Then the prediction is found by performing single- or multiqubit measurements upon $\rho(\mathbf{x}, \boldsymbol{\theta})$, optionally using quantum state tomography (QST) [17]. It can be shown that at that last step a linear classifier is formed for $\rho(\mathbf{x}, \boldsymbol{\theta})$ [32]. It follows that the classes should be significantly easier to separate in the new feature space than in the original, in similarity to classical dimensionality reduction approaches [33,34].

Preprocessing of $\mathbf{x}$ and postprocessing of measurement outcomes, for example, by simultaneously trained classical neural networks, seems to improve model performance. As the model combines both classical and quantum mappings, we use the term *hybrid* deep quantum learning model [17,35]. Finally, there may be variations in how the data are inserted in the PQC. Usually, each value from the input data vector $\mathbf{x}$ is written to the parameter of a single gate in the circuit. However, recent studies showed that recording the same parameter multiple times over different gates of PQC (the so-called *data reuploading*) significantly improves its expressivity [4,19].

In this work, we experimentally train a small hybrid model to solve several problems of supervised learning on

*tolstobrov.ae@phystech.edu

TABLE I. Physical parameters for the four leftmost transmons. Idling ("parking") point frequencies $\omega_{ge}^{(i)}$ and readout resonator frequencies $\omega_r^{(i)}$ are given, decoherence times are measured when the transmons are in their idling points, near the "sweet-spots." Precise gate durations are also listed.

Transmon	1	2	3	4
$\omega_{ge}/2\pi$, GHz	5.80	4.96	6.05	4.94
T_1, μs	7.2	9.3	8.0	10.5
T_2^*, μs	3.8	4.1	3.7	3.9
$\omega_r/2\pi$, GHz	7.16	7.24	7.31	7.40
$t_{X,Y}$, ns	40	40	40	40
t_{iSWAP}, ns	25	27	27	

TABLE II. Summary of the dataset properties. For PARITY, due to the low number of samples for a four-bit task train or test split was equal. For the remaining two datasets the splitting was 2/1. * Cross-validated accuracy, averaged over six different random splits.

Dataset	PARITY	CANCER	WINES	MNIST
# samples	16	569	178	1797
# features	4	30	13	56
# classes	2	2	3	10
Accuracy	1.0	0.95*	0.94*	0.90

well-known datasets: bit-string parity (PARITY), diagnosing breast cancer by biopsy results (CANCER), discerning wine cultivars by their physical and chemical parameters (WINES), and recognizing handwritten digits (MNIST) [36]. We report an experiment employing loading image datasets into a PQC via convolutions and the classification of all ten digits from the MNIST dataset. As a small-scale prototype of a quantum hardware device used to evaluate the quantum part of the model, we use a linear array of superconducting transmon artificial atoms with nearest-neighbor interactions [29,37,38]. We show that four qubits with, at most, two hundred parameters is enough to reach test accuracies higher than 90% for all studied datasets. As we find experimentally and confirm numerically, the algorithm is sufficiently resistant to imperfections in gate operations and can be implemented with currently available quantum devices without error correction.

II. LEARNING SIMPLE DATASETS

An optical image of the experimental device is shown in Fig. 1(a). The chip hosts eight artificial atoms forming a linear chain with nearest-neighbor interactions [29]. Each transmon has an individual measurement resonator and two control lines to change the flux through the superconducting quantum interference device (SQUID) of the transmon and to excite it with microwave radiation. In this study, we use only the left half of the chain (four transmons). The physical parameters of the device are presented in Table I.

A. Training the device

The architecture of the four-qubit PQC that is run on the device is shown in Fig. 1(b). Gaussian driving pulses with controllable amplitude are used for single-qubit operations ($t_{X,Y} = 40$ ns) and smoothed rectangular dc pulses for two-qubit operations ($t_{\mathrm{iSWAP}} = 25$ ns), details of the calibration are presented in Appendix A. Each layer of operations has a duration of $t_l = 80$ ns, which includes an idling margin to account for variations of propagation delay among the control lines. Total PQC execution time is $t_{\mathrm{PQC}} = 1460$ ns (including 500 ns for readout); however, we have to use a $t_{\mathrm{rep}} = 50$ μs repetition period to allow the system to return to the state of thermal equilibrium with the environment of 20 mK, denoted as $|\emptyset\rangle$. This repetition period can be significantly reduced

though to 1 to 5 μs, by using unconditional reset protocols [39,40].

Controlled evolution of a real quantum system ends in a statistically mixed state $\rho(\mathbf{x}, \boldsymbol{\theta})$, having limited resemblance to the PQC target state $|\Phi(\mathbf{x}, \boldsymbol{\theta})\rangle$, and this is supposed to be the main restriction for VQA development on larger noisy intermediate-scale quantum (NISQ) devices [11]. However, while in our case the PQC execution time t_{PQC} is comparable to the average decoherence time $T_2 \approx 4$ μs, we are still able to successfully perform training and inference.

The feature vectors $\mathbf{x} \in \mathbb{R}^4$ are presented to the PQC by single-qubit X rotations of the first layer. The corresponding angles are calculated by applying the inverse tangent function to $\{x_i\}$. For datasets CANCER and WINES, we use a decision tree classifier to choose the four most relevant components of the feature vector. Selected features for WINES dataset are "proline," "flavanoids," "color intensity," and "alcohol"; for the CANCER dataset "worst radius," "worst concave points," "worst texture," and "mean texture." For a certain dataset, we will denote as $\mathcal{X}, \mathcal{Y}$ the sets of feature vectors and corresponding labels, and as $\mathcal{T}, \tilde{\mathcal{T}} : \mathcal{T} \cup \tilde{\mathcal{T}} = \mathcal{X}$ the train and test feature subsets, respectively. The information about used datasets is summarized in Table II.

Additionally, we merge into the first layer four components θ_{1-4} of the weight vector $\boldsymbol{\theta} \in \mathbb{R}^{15}$ by adding them to the respective feature angles [30]. This operation is necessary to shorten PQC by combining layers of encoding and optimization.

The larger part of Hilbert space can be reached by the image $|\Phi(\mathbf{x}, \boldsymbol{\theta})\rangle$, the better expressivity of the model, and thus the PQC must be sufficient to generate fully entangled states. However, higher expressivity might lead to training problems of the algorithm, so the model should not be too complex. While it is possible to optimize the structure of the circuit along with tuning its parameters $\boldsymbol{\theta}$ [16], we use a V-shaped sequence of fixed two-qubit operations, interleaved by two layers of single-qubit X and Y rotations, as the entangling block. This structure allows us to effectively use a large part of the Hilbert space of four qubits, has sufficient flexibility and is easy to calibrate. We use only roughly calibrated quasi-iSWAP gates [41,42]. Their exact matrix representation does not significantly affect the expressivity of the circuit, which we check in numerical simulations and connect with the Kraus-Cirac decomposition theorem [43]. The detailed information about calibration of single-qubit and two-qubit operations can be found in Appendix A. Fortunately, there is no need to know exactly which final state $\rho(\mathbf{x}_i, \boldsymbol{\theta})$ is prepared,
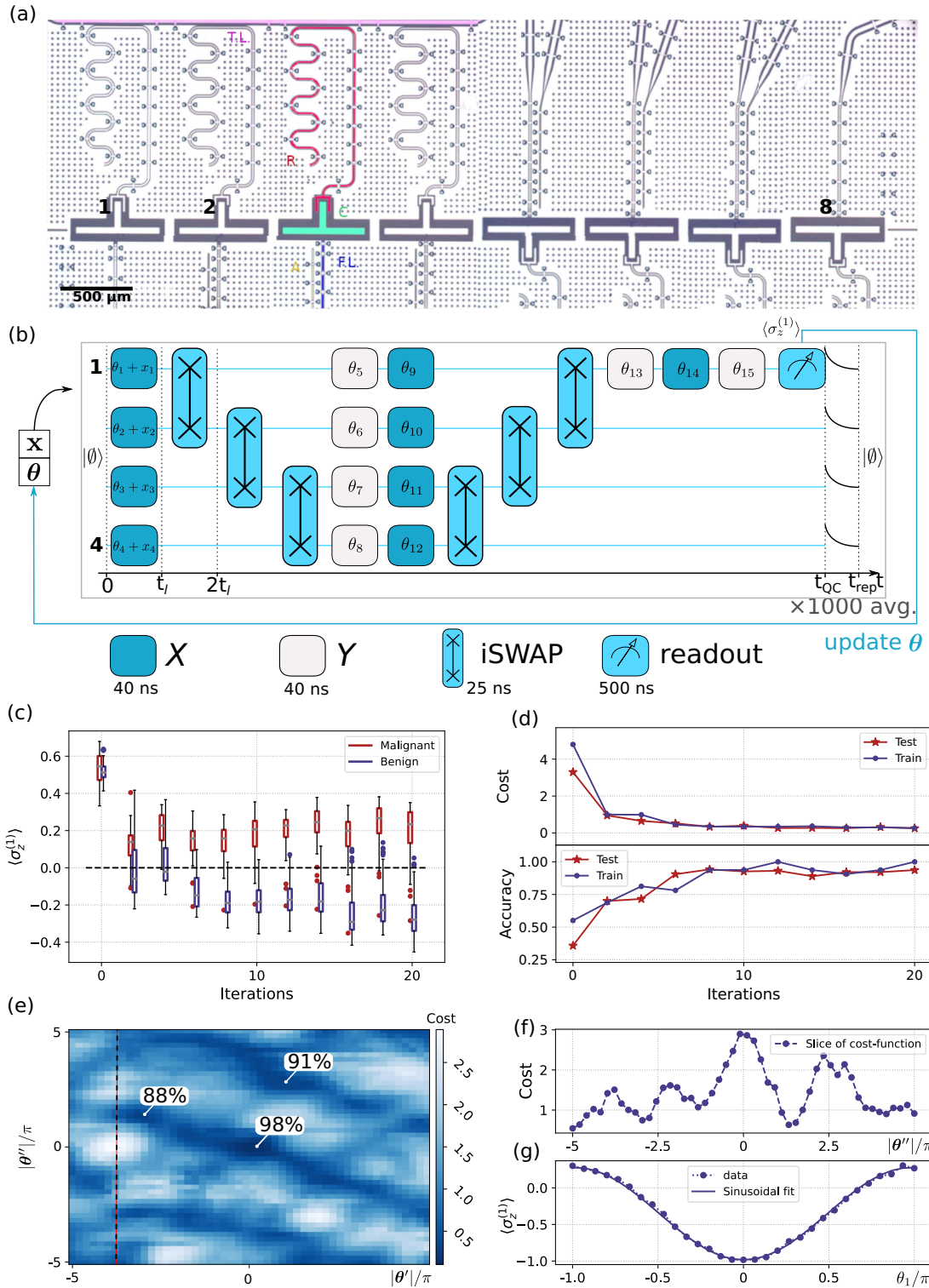
FIG. 1. (a) Micrograph of the eight-transmon device used as the quantum hardware (false colored). The chip is symmetric, one of the two transmission lines (purple) is visible. Readout resonators (red), microwave antennas (yellow), and flux control lines (blue) address each transmon individually. T-shaped shunting capacitors are shown in green. (b) Structure of the PQC. The polar angles of single-qubit X, Y rotations constitute the parameter vector $\boldsymbol{\theta}$ while two-qubit operations are fixed. (c) Distributions of $\langle \sigma_z^{(1)} \rangle$ for two classes versus training iteration for CANCER dataset. (d) Cost and accuracy convergence, calculated for both $\mathcal{T}$ and $\tilde{\mathcal{T}}$ for comparison. (e) Measured cost function landscape for $\tilde{\mathcal{T}}$ around the found minimum in the linear hull of two random orthogonal directions $\boldsymbol{\theta}'$, $\boldsymbol{\theta}''$. Accuracies are indicated at several local minima. (f) A 1D slice of cost function, shown in (e) with a dashed line. (g) Output of the circuit $\langle \sigma_z^{(1)} \rangle$ showing expected harmonic dependence on $\boldsymbol{\theta}$ components, data for θ_1
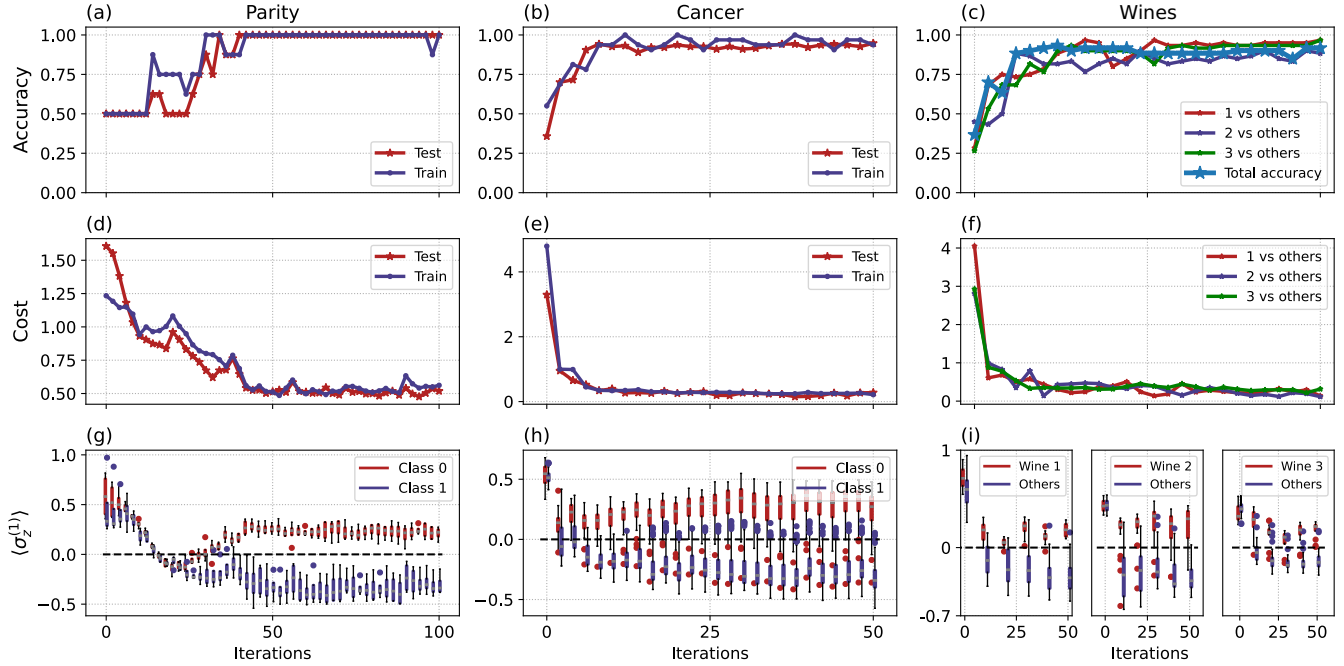
FIG. 2. Training process of classifiers for all three simple problems. [(a)–(c)] Accuracy versus the number of iterations. [(d)–(f)] Same for the cost function. [(g)–(i)] Distributions of $\langle \sigma_z^{(1)} \rangle$ for two classes during training process. For wines dataset accuracies of all three "one-versus-others" classifiers and total accuracy are plotted. In picture (i) the distribution of $\langle \sigma_z^{(1)} \rangle$ for these classifiers is plotted.

contrary to quantum chemistry problems [5]. As there are many possible structures for PQCs, numerical performance analysis some of them is provided in Appendix B.

The last four layers are an arbitrary Euler rotation of the first qubit via an Y-X-Y sequence and a measurement. The prediction $g(\mathbf{x}_i, \boldsymbol{\theta}) = \text{Tr}[\rho(\mathbf{x}_i, \boldsymbol{\theta}) \sigma_z^{(1)}]$ is calculated by running the PQC repeatedly and averaging the $\sigma_z^{(1)}$ outcomes. Then it is thresholded at $g(\mathbf{x}_i, \boldsymbol{\theta}) = 0$ to obtain the binary prediction. With such an output, k-class classification is possible to realize by training k one-versus-others models with proper optimal parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$, or $k(k-1)/2$ pairwise classifiers [35].

We optimize $\boldsymbol{\theta}$ using stochastic gradient descent (SGD) with the per-sample logarithmic cost function defined by

$$\mathcal{L}[g(\mathbf{x}_i, \boldsymbol{\theta}), y_i] = \log_2\left(1 + \exp[-y_i g(\mathbf{x}_i, \boldsymbol{\theta})\beta]\right) + \gamma|\boldsymbol{\theta}|^2, \quad (1)$$

which favors the label $y_i \in \{-1, 1\}$ and the prediction $g(\mathbf{x}_i, \boldsymbol{\theta})$ to have the same sign, and grows linearly in $|g(\mathbf{x}_i, \boldsymbol{\theta})|$ when they have opposite signs. We find that choosing large $\beta = 10$ to strongly penalize sign difference is beneficial to the training process. The regularizing term $\gamma|\boldsymbol{\theta}|^2$ with $\gamma = 0.2$ allows to penalize the model for overfitting. While important for the image recognition task below, it is not necessary for simple datasets. We also check that, for the studied datasets, the quadratic cost yields comparable training performance.

The full cost is calculated as the expectation value $\mathbb{E}[\mathcal{L}[g(\mathbf{x}_i, \boldsymbol{\theta}), y_i]]$ over a chosen data subset $\mathcal{S}$, so $i : \mathbf{x}_i \in \mathcal{S}$, with $\mathcal{S}$ being $\mathcal{T}$, or $\tilde{\mathcal{T}}$, or a minibatch $\mathcal{B}$ of size b. The jth component of its gradient over $\boldsymbol{\theta}$ can be conveniently computed using the parameter-shift rule [1,24], which requires only two measurements at $\theta_j \pm \pi/2$. We use the PENNYLANE library [44] with a custom software wrapper to our experimental setup to perform both the automated differentiation and the

optimization. Figures 1(c) and 1(d) shows a visualization of the training process for the CANCER dataset with 569 samples and 2-to-1 $\mathcal{T}$-to-$\tilde{\mathcal{T}}$ split [45]. Figure 1(c) shows the distribution of the model predictions for $\mathbf{x} \in \tilde{\mathcal{T}}$ at each training step. Each iteration consists of one gradient evaluation and a Nesterov accelerated [46] SGD step over $x_i \in \mathcal{B}$, $b = 64$. At the beginning of the training, the two classes are indistinguishable while at the end of the training algorithm the distributions of $g(\mathbf{x}, \boldsymbol{\theta})$ for the two classes almost do not overlap. The accuracy of the algorithm on both $\mathcal{T}$ and $\tilde{\mathcal{T}}$ steadily increases with the number of iterations, reaching approximately 95% in about ten iterations. We also do not observe any systematic decline in accuracy if the training is further continued. To correct for the statistical fluctuations of the accuracy estimation due to the particular realization of the sampling for the $\mathcal{T}$-to-$\tilde{\mathcal{T}}$ split, we use the cross-validation method, averaging results over several different splits.

Following [12], we also study the behavior of the cost function calculated over $\mathbf{x} \in \tilde{\mathcal{T}}$ near the found optimum. Using a known visualization method [47], we plot a two-dimensional (2D) slice of the 15-dimensional (15D) parameter space, a square in the linear hull of two normalized orthogonal random vectors $\boldsymbol{\theta}'$, $\boldsymbol{\theta}''$ added to the optimal vector $\boldsymbol{\theta}^*$. From Figure 1(e) it can be seen that, even for such a small PQC, the minimum indeed is not unique which can lead to trapping of the algorithm, and that the cost is nonconvex in the original space [47]. We also find that setting $\boldsymbol{\theta}^* = 0$ in this experiment yields similar topography with local minima corresponding to above 80% accuracies, so a moderately good solution can be found just by moving along a randomly chosen direction.

We also check experimentally that the dependence of $g(\mathbf{x}_i, \boldsymbol{\theta})$ on each of the parameters θ_j is harmonic, according
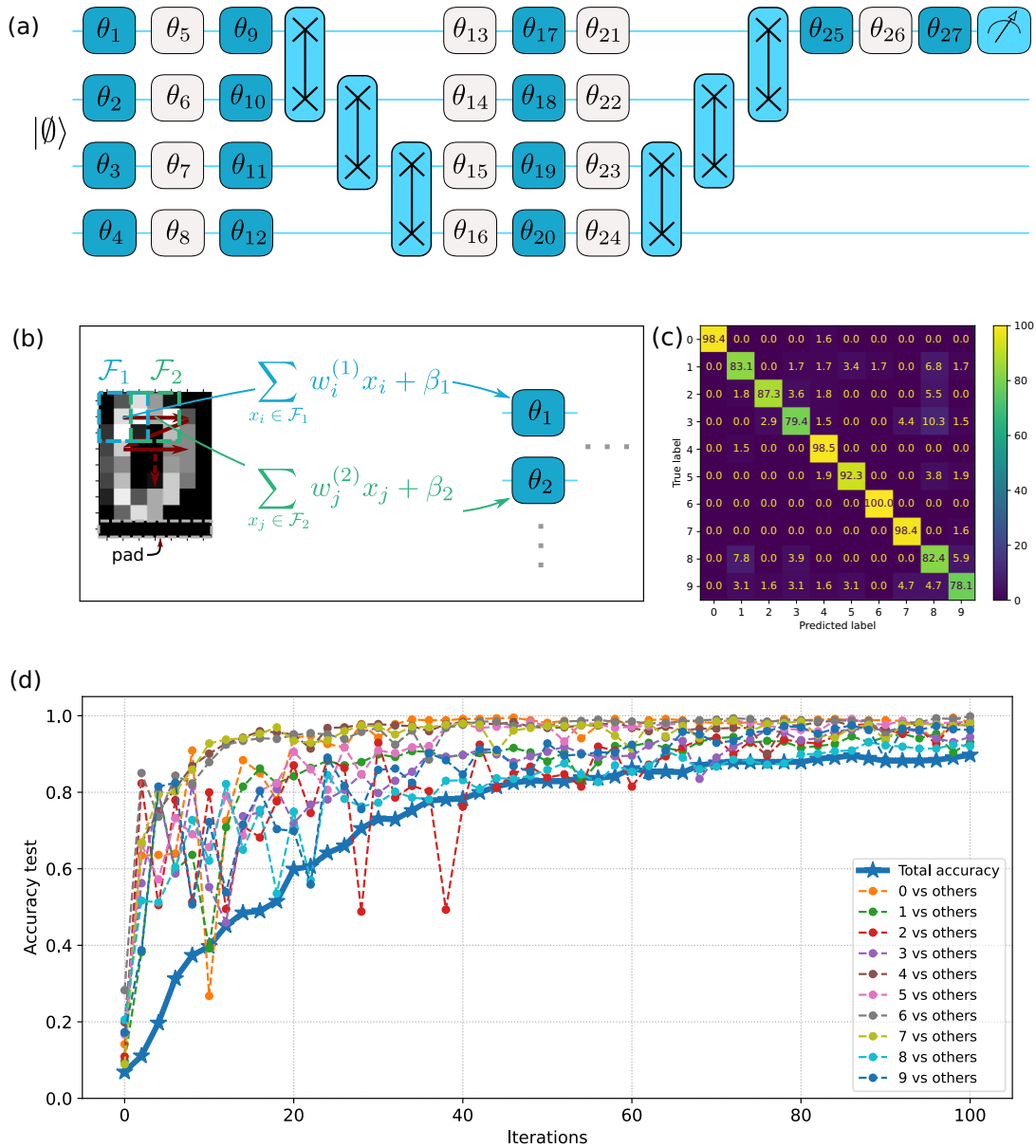
FIG. 3. Image recognition for MNIST. (a) The PQC used to process larger feature vectors. (b) Data encoding into parameters of single-qubit operations, using convolutional kernels. (c) Confusion matrix to analyze the performance of each classifier. Each intersection of ith row and jth column shows the percentage of pictures belonging to the ith class and recognized as elements of jth class. Nondiagonal elements show misclassifications. (d) Visualization of training process of the classifier. The dependence of the accuracies on the number of iterations for ten one-versus-others classifiers and total accuracy are shown.

to theory, and show in Fig. 1(g) how $g(\mathbf{x}_i, \boldsymbol{\theta})$ varies with θ_1, the first and the deepest parameter in the circuit, when the other parameters and the input features $\mathbf{x}$ are set to zero. As can be seen, due to the inaccuracies in the calibration of the two-qubit gates and nonnegligible decoherence, the value of the prediction never reaches "+1" in contrast to what one would expect from the PQC structure for $\theta_1 = \pi$.

B. Performance analysis

In Fig. 2 we summarize the model training and performance on all three datasets, see also Table II. In the PARITY

problem, four-bit sequences should be decided to contain even or odd numbers of "1"'s. This problem is a simple test which displays the reproducibility of quantum operations and the sensitivity of the model in capturing class change even when a single bit is flipped, which is difficult for classical models [48]. Having only 16 samples, we split the dataset equally and calculate the cost function on the full subset $\mathcal{T} \equiv \mathcal{B}$, $b = 8$. At the optimal point, the accuracy reaches 100%, as there exists an analytical solution composed of four CNOT gates [49], which can be mimicked by our ansatz. At the same time, the cost function does not reach zero due to the imperfections of quantum operations and decoherence.
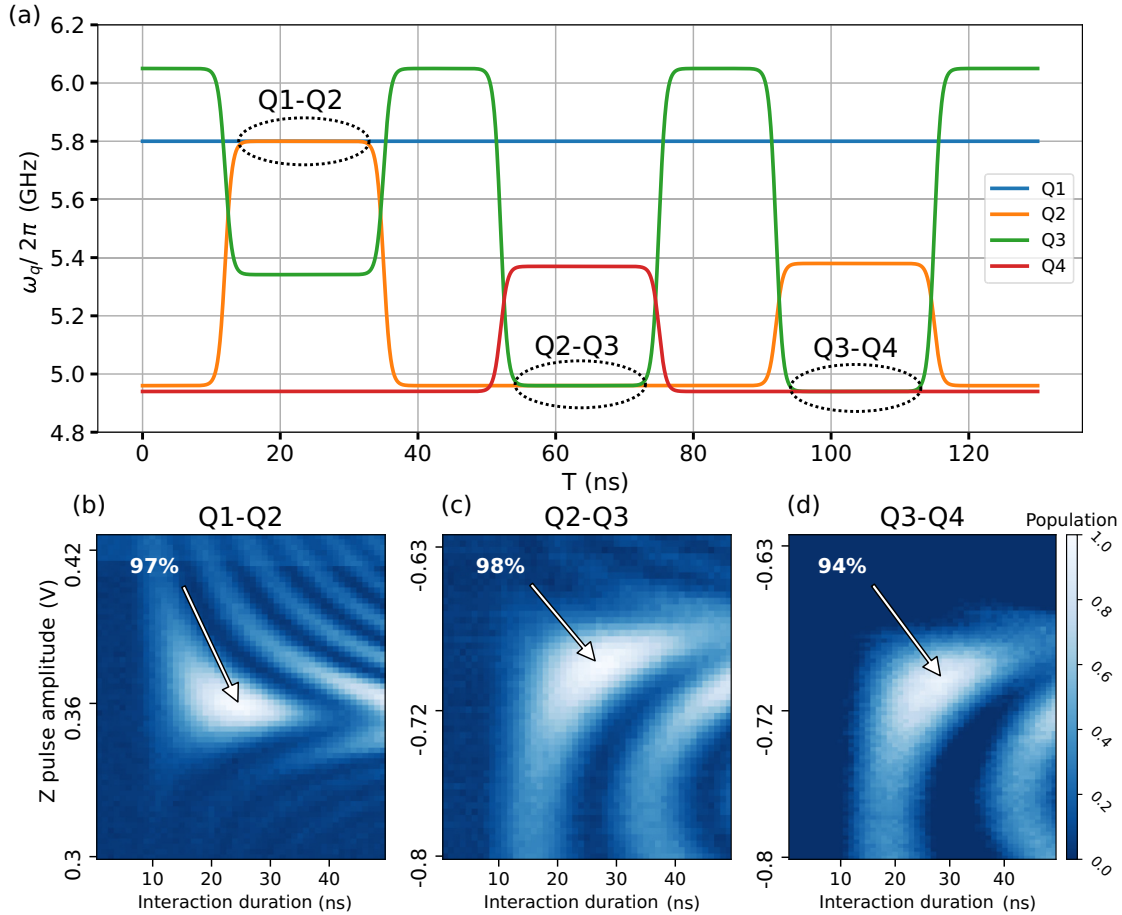
FIG. 4. Calibration of two-qubit operations. (a) Frequency trajectories of the four transmons during three sequential ISWAP operations for neighboring pairs. When frequencies of one pair are brought into resonance, frequencies of the adjacent transmons are shifted to reduce residual interaction. The areas in which resonances occur are highlighted with dotted ellipses. [(b)–(d)] Calibration of ISWAP gates. For pictures (b), (c), and (d) excited qubits are Q1, Q2, Q3 and measured qubits are Q2, Q3, Q4, correspondingly. The dependence of the first excited state population of measured qubit on interaction duration and flux pulse amplitude is shown. The points with the largest population exchange are marked, and corresponding parameters are used for two-qubit operations.

The CANCER problem, already briefly presented above, is also binary, but has a significantly larger dataset which allows a better splitting ratio and puts the model under a more stringent test in terms of performance. As it can be seen from Fig. 1(b), with $t_{rep} = 50$ μs and averaging over 1000 repetitions, the measurement of $\langle \sigma_z^{(1)} \rangle$ takes 50-ms time. In our setup, rewriting the pulse sequence waveforms to update the gates in the PQC takes comparable time (it could be reduced at least an order of magnitude, though with better hardware). To find the gradient of the cost function calculated on a single x_i, it is necessary to measure $\langle \sigma_z^{(1)} \rangle$ for $2m + 1$ sets of angles $\boldsymbol{\theta} \in \mathbb{R}^m$. Evaluating the gradient for the circuit from Fig. 1(b) over $m = 15$ variables takes $t_{grad} = 1.55$ s, which with the additional time for rewriting the controlling sequences gives about 3 s. Then, for a batch size $b = 64$, one iteration takes around $bt_{grad} = 3$ min. Finally, reaching the accuracy plateau in around 20 iterations, as in Figs. 1(c) and 1(d), takes around 1 h.

As an elementary test of the model capability to solve multilabel classification problems, we use the three-class WINES dataset. Multilabel classification is done by training three "one-versus-others" binary classifiers aiming to detect each of the cultivars. Then, for a given $x \in \tilde{\mathcal{T}}$, we choose among the three found $\boldsymbol{\theta}_{1-3}$ the one delivering the highest value to $g(\mathbf{x}, \bullet)$, and predict the class accordingly. We find that all of three one-versus-others classifiers exhibit similar training behavior, an accuracy of classification starts from approximately $1/3$, reaches a value of 90% in 1ent iterations, and slightly fluctuates further, which is normal for minibatch learning. As a result, the total classification accuracy is also slightly above 90%; the cross-validation procedure gives a value of 94%.

III. RECOGNITION OF HANDWRITTEN DIGITS

The feature space dimension $m = 4$ for the datasets considered above is obviously quite small. The situation is different for the image recognition problem: even for a downsampled and cropped MNIST picture of size 8×7 pixels with intensities ranging from 0 to 1, the feature space is equipotent to $\mathbb{R}^{56}$. Choosing a particular way to load information from this space into the quantum state of just four qubits is not a trivial task. We use an approach combining the data reuploading concept [19] and convolutional neural networks (CNN) [17,50]. For preprocessing we use normalization and inverse tangent

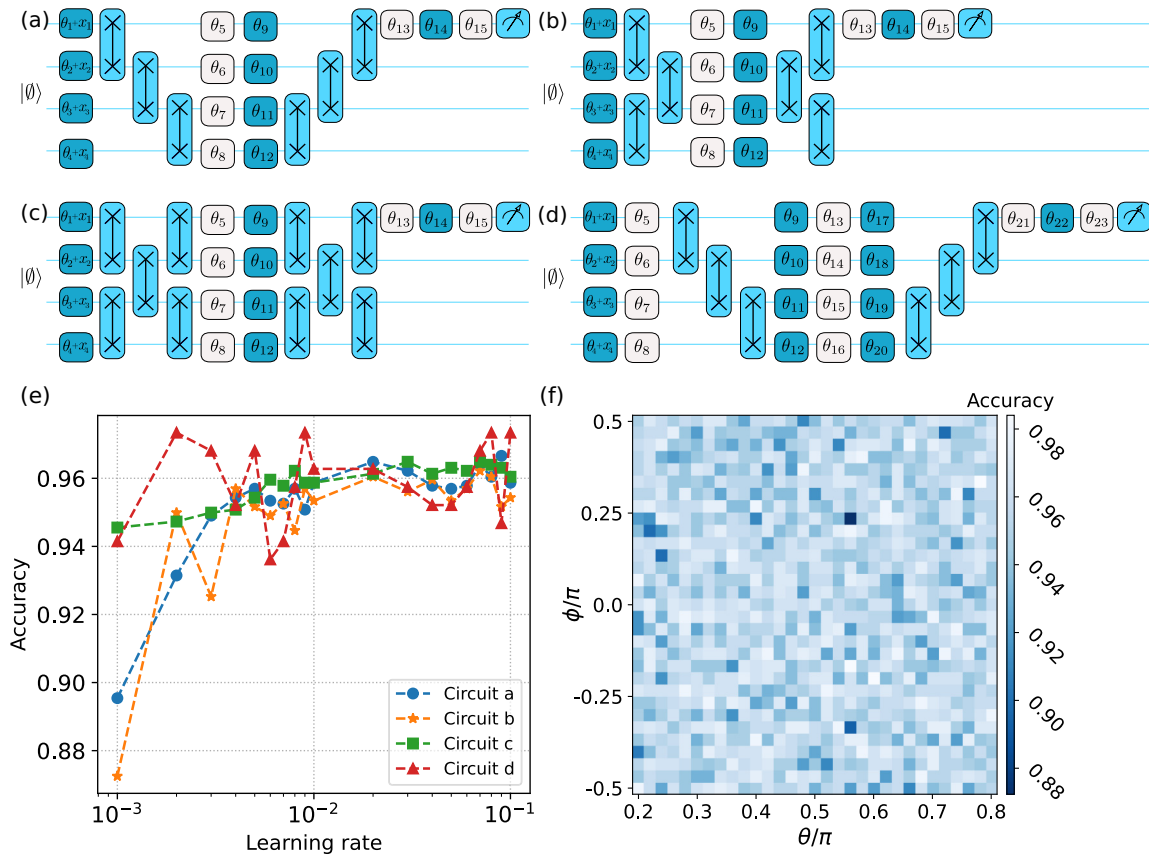FIG. 5. Comparison of the different PQCs architectures for classification of CANCER dataset. [(a)–(d)] Possible structures of PQCs. (e) Numerical modeling of the accuracy versus learning rate dependence for considered PQCs. (f) Numerical modeling of the accuracy versus angles θ and ϕ dependence for PQC (a) and dataset CANCER.
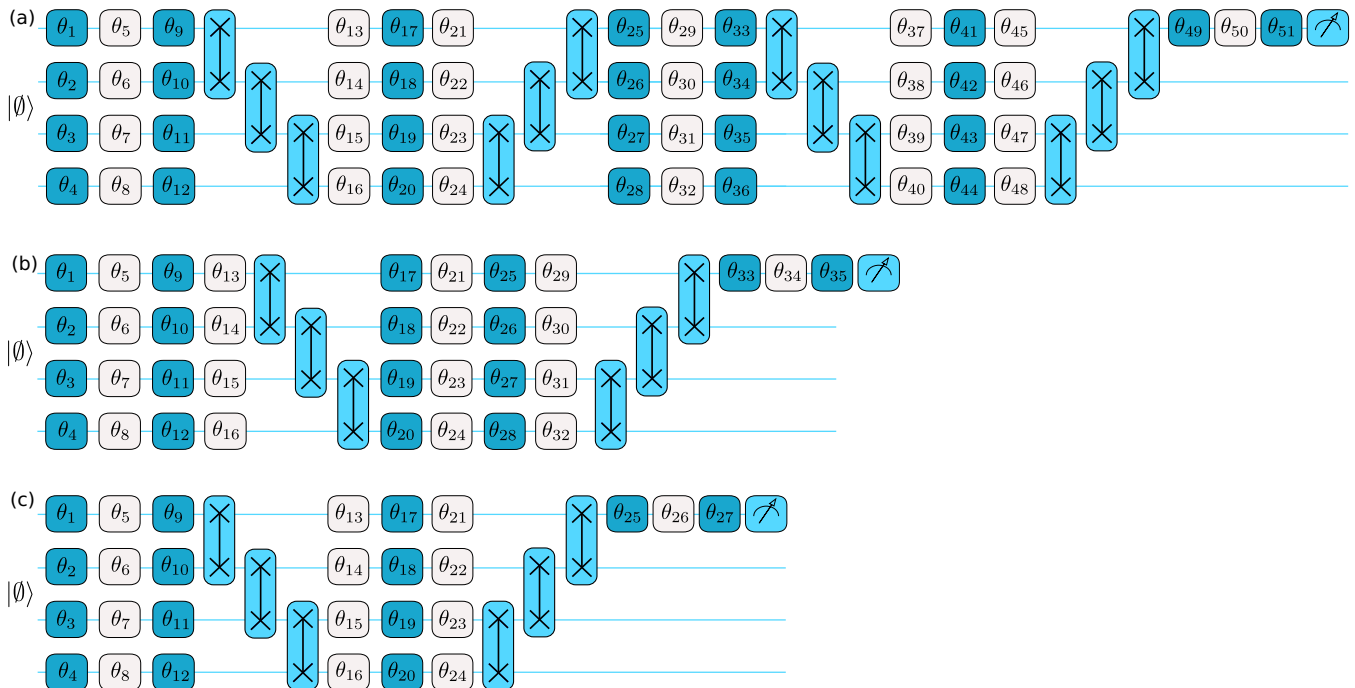

FIG. 6. Possible PQC architectures for the image recognition problem.

transformation of the features. We use a modified PQC with the structure similar to the one shown in Fig. 1(a), given in Fig. 3(a). In Fig. 3(b), we illustrate how the 8×7 images are padded with a zero bottom row and then divided in 3×3 partially overlapping local receptive fields (LRF) moving with stride 2 in both directions [50]. LRFs traverse the image twice, so there are 24 weight kernels in total. Pixels belonging to the ith LRF $\mathcal{F}_i$ are convolved with weights $w_j^{(i)}$, and, with the addition of a bias β_i, are converted to angles θ_{1-24}:

$$\theta_i = \beta_i + \sum_{x_j \in \mathcal{F}_i} w_j^{(i)} x_j.$$

In contrast to conventional CNN, the kernels for each LRF are independent, which makes the model more flexible. The last three parameters θ_{25-27} are independent, so with an addition of a final bias β_0 to the output of the PQC, the total weight dimension is 244. We use the parameter-shift rule to compute $\boldsymbol{\theta}$ gradients of the loss function and chain rule to obtain $w_j^{(i)}$ and β_i gradients.

In contrast to the circuit in Fig. 1(a), the feature data are now recorded in multiple layers. The information about every pixel is also written several times.

The model was tested on a subset of the MNIST dataset, see Table II, by training ten separate one-versus-others classifiers to be able to distinguish between all ten digits. There is a strong (1/9) disproportionality in the quantities for each of the corresponding pairs of classifiers, which might lead to training problems. To overcome these problems, we perform dataset balancing by adding copies of elements of a smaller class with added Gaussian noise to the training dataset. After that, the number of elements in the two classes becomes the same. The training process is visualized in Fig. 3(d): the total ten-digit accuracy steadily increases from approximately 10% in the beginning (random guessing) to approximately 90% after 100 iterations. The accuracy of individual classifiers varies from 100% for "six versus others" to 92% for "eight verus others," which decreases the full accuracy to 90%. To analyze the errors, we construct a confusion matrix, which is shown in Fig. 3(c), and find that most confusion is caused by the "8" classifier, with most misclassifications between "8" and "3," which looks reasonable. In all likelihood, the accuracy could be further improved, however, as the training for Fig. 3 had taken around 100 h, we did not continue it further. We found in numerical simulations that suggested PQC is not optimal, and classification accuracy might be increased to approximately 93% with nearly the same number of circuit layers. The detailed analysis of possible circuit architectures is presented in Appendix B. We also tested the model on the FASHION MNIST dataset [51]. As for the classical models [51], the classification accuracy for the FASHION MNIST dataset turned out to be worse than for the MNIST dataset. We achieve only 85% accuracy for four different types of clothes, while for four digits "0–3" we report 98% accuracy.

IV. CONCLUSION

We experimentally implement a supervised quantum learning algorithm in a chain of superconducting qubits to solve multilabel classification and image recognition problems. We first realize loading image datasets onto PQC via

convolutions, present a suitable gate sequence and a training algorithm. This algorithm allows us to achieve classification accuracy 90% for the MNIST dataset. Our classification algorithm is convenient for implementation on near-term quantum devices without error correction because it is sufficiently resistant to imperfections and requires measurement of only one qubit. We note that the presented model does not yet outperform even the simplest classical model, such as the linear classifier, which achieves an accuracy of 95% with only 570 trainable parameters. This means though that it is possible to obtain 95% accuracy using a PQC with only one layer and one qubit by training ten one-versus-other classifiers if a linear combination of all features is recorded in the angle of single-qubit operations. However, the main work in that case will be performed by a classical computer, while to achieve the advantage in quantum machine learning the right balance between the classical and quantum parts of the model should be found. For example, replacing the last low-dimensional layers of a convolutional network with a PQC could be a direction of further study.

We also address the issue of the low speed of inference that we observe in practice with a real device. Despite the fact that we use state-of-the-art gate durations (10-s of ns), and the fact that the superconducting quantum computing platform currently features the fastest known gates (compared with silicon, with 100-s of ns [52], and trapped atom or ion or diamond platforms, with 100-s of μs [53–55]), the training process is currently orders of magnitude slower than for the classical machine learning methods. We note, however, that the training time of the hybrid model presented here could be significantly reduced by implementing an unconditional reset instead of simple waiting [39,40] (about 50 times faster) and training only one multilabel classifier using multiplexed readout instead of ten binary classifiers. Thus, the training of our model for the MNIST dataset could be reduced from 100 hours to approx. 12 min., which is more competitive. It is also possible to reduce the total number of layers in the PQC by performing several two-qubit operations in parallel.

To solve more complex tasks in the domain of quantum machine learning it is necessary to realize very fast gates on a supremacy-scale register of qubits. While in this work we do not notice a significant impact of gradient decay on the performance of quantum classifiers, increasing the number of qubits will require an ingenious circuit architecture to cope with that problem. In the classical machine learning, a similar problem was overcome by using skip connections [56] and batch normalization [57], but at the moment it is not known whether any analogs of these techniques could be reasonably implemented in the quantum case and further research is necessary.

Experimental data are available upon reasonable request from the corresponding authors.

valuable technical support, and the anonymous referees for the valuable comments. All samples were fabricated in the Shared Facility Center of MIPT.

All authors declare no competing interests.

APPENDIX A: SINGLE-QUBIT AND TWO-QUBIT OPERATIONS

During the training process, we use parameterized single-qubit operations and fixed two-qubit operations. Single-qubit gates with variational parameters were implemented by using fixed duration (40 ns) microwave pulses with a Gaussian envelope. To change the rotation angle θ, we vary the amplitude of pulses, as θ is proportional to the area of their envelope. To change the rotation axis (X or Y) we vary the phase of pulses (choosing 0 or $\pi/2$, respectively).

For the realization of two-qubit operations, the transmon frequencies are tuned into resonance by fast flux pulses, as shown in Fig. 4(a); the resonance regions are highlighted with ellipses. Taking into account the capacitive interaction between the transmons, this procedure leads to the realization of a so-called "f-Sim"-like two-qubit gate [42]. It consists of single-qubit rotations due to changes in the frequency of the transmons and of a two-qubit part. As we include single-qubit rotations before and after the two-qubit gates in the PQC, the single-qubit part can be automatically compensated during training process. Thus we can limit ourselves to considering only the two-qubit part of the gate

$$\text{f-Sim}(\theta, \phi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -i\sin\theta & 0 \\ 0 & -i\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & e^{-i\phi} \end{pmatrix}. \quad (A1)$$

To calibrate the flux pulses, we measure two-qubit chevron-type oscillations. In this experiment, one of the transmons in the pair is excited by a π pulse and then brought close to resonance with its pair transmon. The population of the latter is recorded versus amplitude and duration of the flux pulse. The resulting dependencies for the three pairs of adjacent transmons are shown in Figs. 4(b) to 4(d). We maximize the population transfer, so that $\theta \approx \pi/2$. We also directly measure the nonzero phase ϕ due to the ZZ interaction to be $\approx 0.1\pi$ in a separate Ramsey-type experiment.

To address the robustness of the PQC performance to inaccuracies of the parameters θ and ϕ, we numerically study

TABLE III. Comparison of different PQC architectures for image recognition problem. In the first column the corresponding circuit label from Fig. 6 is shown.

PQC	LRF size	Stride	Parameters	Layers	Accuracy
a	2×2	1	248	27	0.95
c	2×2	2	92	15	0.925
b	2×2	2	164	17	0.93
c	3×3	2	244	15	0.9
a	3×3	1	376	27	0.77

the dependence of the accuracy of the classification for the CANCER dataset on θ and ϕ. This result is shown in Fig. 5(f). It can be seen that the classification accuracy remains almost unchanged and the convergence is preserved for $\theta \in [0.2\pi, 0.8\pi]$ and $\phi \in [-0.5\pi, 0.5\pi]$.

APPENDIX B: ALTERNATIVE CIRCUIT ARCHITECTURES

During our investigation we were focused on only one PQC architecture for simple datasets and one architecture for the more complex MNIST dataset. The chosen architecture provides a sufficient degree of entanglement using a relatively small number of layers. We can also consider PQC in which two-qubit operations between qubits 1,2 and 3,4 are done simultaneously. Possible structures of these PQCs are shown in Figs. 5(a) to 5(d). We numerically tested the ability of these PQCs to solve binary classification problems using the CANCER dataset. The dependence of cross-validation accuracy on learning rate for these circuits is shown in Fig. 5(e). As it can be seen, the best accuracy does not depend significantly on the choice of PQC. For experimental realization it is convenient to use PQC without simultaneous two-qubit operations. This is the reason why we choose PQC (a) for our experiments.

The PQC for image recognition has many control parameters, for example, the size of the LRF or stride. These parameters may strongly affect the final classification accuracy. Possible architectures of PQCs are shown in Fig. 6. To find the optimal circuit, we conducted numerical simulations for various PQCs and control parameters. The results of our numerical modeling are summarized in Table III. It can be noticed that the choice for experimental realization architecture of PQC [Fig. 6(c)] with size of LRF 3×3 and stride 2 does not allow to achieve the highest accuracy of classification. Using a circuit with nearly the same number of layers, the total accuracy might be improved to approximately 93%.

[1] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Phys. Rev. A **98**, 032309 (2018).

[2] M. Schuld and N. Killoran, Phys. Rev. Lett. **122**, 040504 (2019).

[3] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Nature (London) **567**, 209 (2019).

[4] S. Jerbi, L. J. Fiderer, H. Poulsen Nautrup, J. M. Kübler, H. J. Briegel, and V. Dunjko, Nat. Commun. **14**, 517 (2023).

[5] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio et al., Nat. Rev. Phys. **3**, 625 (2021).

[6] I. Cong, S. Choi, and M. D. Lukin, Nat. Phys. **15**, 1273 (2019).

[7] P.-L. Dallaire-Demers and N. Killoran, Phys. Rev. A **98**, 012324 (2018).

[8] S. Lloyd and C. Weedbrook, Phys. Rev. Lett. **121**, 040502 (2018).

[9] J. Bausch, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin (2020), https://researchr.org/publication/Bausch20.

[10] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Nat. Commun. **9**, 4812 (2018).

[11] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Nat. Commun. **12**, 6961 (2021).

[12] E. R. Anschuetz and B. T. Kiani, Nat. Commun. **13**, 7760 (2022).

[13] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, PRX Quantum **3**, 010313 (2022).

[14] T. Volkoff and P. J. Coles, Quantum Sci. Technol. **6**, 025008 (2021).

[15] A. A. Mele, G. B. Mbeng, G. E. Santoro, M. Collura, and P. Torta, Phys. Rev. A **106**, L060401 (2022).

[16] Y. Du, T. Huang, S. You, M.-H. Hsieh, and D. Tao, npj Quantum Inf. **8**, 62 (2022).

[17] Y. Zeng, H. Wang, J. He, Q. Huang, and S. Chang, Entropy **24**, 394 (2022).

[18] T. Hur, L. Kim, and D. Park, Quantum Mach. Intell. **4**, 3 (2022).

[19] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Quantum **4**, 226 (2020).

[20] A. Pérez-Salinas, D. López-Núñez, A. García-Sáez, P. Forn-Díaz, and J. I. Latorre, Phys. Rev. A **104**, 012405 (2021).

[21] Y. Liu, S. Arunachalam, and K. Temme, Nat. Phys. **17**, 1013 (2021).

[22] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Nat. Commun. **12**, 2631 (2021).

[23] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, 2017), Vol. 70, pp. 2847–2854.

[24] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Phys. Rev. A **101**, 032308 (2020).

[25] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, Nat. Comput. Sci. **1**, 403 (2021).

[26] Y. Du, Z. Tu, X. Yuan, and D. Tao, Phys. Rev. Lett. **128**, 080506 (2022).

[27] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Nat. Commun. **13**, 4919 (2022).

[28] D. Zhu, N. M. Linke, M. Benedetti, K. A. Landsman, N. H. Nguyen, C. H. Alderete, A. Perdomo-Ortiz, N. Korda, A. Garfoot, C. Brecque *et al.*, Sci. Adv. **5**, eaaw9918 (2019).

[29] H.-L. Huang, Y. Du, M. Gong, Y. Zhao, Y. Wu, C. Wang, S. Li, F. Liang, J. Lin, Y. Xu *et al.*, Phys. Rev. Appl. **16**, 024051 (2021).

[30] W. Ren, W. Li, S. Xu, K. Wang, W. Jiang, F. Jin, X. Zhu, J. Chen, Z. Song, P. Zhang, H. Dong, X. Zhang, J. Deng, Y. Gao, C. Zhang, Y. Wu, B. Zhang, Q. Guo, H. Li, Z. Wang *et al.*, Nat. Comput. Sci. **2**, 711 (2022).

[31] T. Dutta, A. Pérez-Salinas, J. P. S. Cheng, J. I. Latorre, and M. Mukherjee, Phys. Rev. A **106**, 012411 (2022).

[32] M. Schuld and F. Petruccione, *Machine Learning with Quantum Computers* (Springer International Publishing, Cham., 2021), pp. 217–245.

[33] L. van der Maaten and G. Hinton, J. Mach. Learn. Res. **9**, 2579 (2008).

[34] L. McInnes, J. Healy, N. Saul, and L. Großberger, J. Open Source Software **3**, 861 (2018).

[35] W. Li, P.-C. Chu, G.-Z. Liu, Y.-B. Tian, T.-H. Qiu, and S.-M. Wang, Quantum Eng. **2022**, 5701479 (2022).

[36] UCI, Machine learning repository, https://archive.ics.uci.edu/ml/index.php.

[37] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, J. Majer, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, Phys. Rev. A **76**, 042319 (2007).

[38] R. Barends, J. Kelly, A. Megrant, D. Sank, E. Jeffrey, Y. Chen, Y. Yin, B. Chiaro, J. Mutus, C. Neill *et al.*, Phys. Rev. Lett. **111**, 080502 (2013).

[39] K. Geerlings, Z. Leghtas, I. M. Pop, S. Shankar, L. Frunzio, R. J. Schoelkopf, M. Mirrahimi, and M. H. Devoret, Phys. Rev. Lett. **110**, 120501 (2013).

[40] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff, Phys. Rev. Lett. **121**, 060502 (2018).

[41] A. Dewes, Ph.D. thesis, Université Pierre et Marie Curie-Paris VI, 2012.

[42] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, K. Satzinger, R. Barends *et al.*, Phys. Rev. Lett. **125**, 120504 (2020).

[43] B. Kraus and J. I. Cirac, Phys. Rev. A **63**, 062309 (2001).

[44] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi *et al.*, arXiv:1811.04968

[45] K. K. Dobbin and R. M. Simon, BMC Med. Genomics **4**, 31 (2011).

[46] Yu. E. Nesterov, Dokl. Akad. Nauk SSSR **269**, 543 (1983).

[47] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.

[48] G. De Palma, B. Kiani, and S. Lloyd, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32.

[49] D. Riste, M. P. Da Silva, C. A. Ryan, A. W. Cross, A. D. Córcoles, J. A. Smolin, J. M. Gambetta, J. M. Chow, and B. R. Johnson, npj Quantum Inf. **3**, 16 (2017).

[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Proc. IEEE **86**, 2278 (1998).

[51] H. Xiao, K. Rasul, and R. Vollgraf, arXiv:1708.07747.

[52] S. G. J. Philips, M. T. Mądzik, S. V. Amitonov, S. L. de Snoo, M. Russ, N. Kalhor, C. Volk, W. I. Lawrie, D. Brousse, L. Tryputen *et al.*, Nature (London) **609**, 919 (2022).

[53] L. Postler, S. Heuβen, I. Pogorelov, M. Rispler, T. Feldker, M. Meth, C. D. Marciniak, R. Stricker, M. Ringbauer, R. Blatt *et al.*, Nature (London) **605**, 675 (2022).

[54] D. Bluvstein, H. Levine, G. Semeghini, T. T. Wang, S. Ebadi, M. Kalinowski, A. Keesling, N. Maskara, H. Pichler, M. Greiner *et al.*, Nature (London) **604**, 451 (2022).

[55] M. H. Abobeih, Y. Wang, J. Randall, S. Loenen, C. Bradley, M. Markham, D. Twitchen, B. Terhal, and T. Taminiau, Nature (London) **606**, 884 (2022).

[56] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, 2016), pp. 770–778.

[57] S. Ioffe and C. Szegedy, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei (PMLR, 2015), Vol. 37, pp. 448–456.