

Light-cone feature selection for quantum machine learning

Yudai Suzuki^{1*} · Rei Sakuma^{2,4} · Hideaki Kawaguchi^{3,4}

Received: date / Accepted: date

Abstract Feature selection plays an essential role in improving the predictive performance and interpretability of trained models in classical machine learning. On the other hand, the usability of conventional feature selection could be limited for quantum machine learning tasks; the technique might not provide a clear interpretation on embedding quantum circuits for classical data tasks and, more importantly, is not applicable to quantum data tasks. In this work, we propose a feature selection method with a specific focus on quantum machine learning. Our scheme treats the light-cones (i.e., subspace) of quantum models as features and then select relevant ones through training of the corresponding local quantum kernels. We numerically demonstrate its versatility for four different applications using toy tasks: (1) feature selection of classical inputs, (2) circuit architecture search for data embedding, (3) compression of quantum machine learning models and (4) subspace selection for quantum data. The proposed framework paves the way towards applications of quantum machine learning to practical tasks. Also, this technique could be used to practically test if the quantum machine learning tasks really need quantumness, while it is beyond the scope of this work.

1 Introduction

Quantum enhancement of machine learning attracts much attentions due to provable advantages of certain quantum machine learning (QML) methods over classical counterparts [1, 2, 3, 4, 5]. Of particular interest are QML approaches for data analysis. A family of fruitful results have theoretically shown that tailored quantum models can efficiently learn patterns of synthetic datasets that are hard for classical ones to find [6, 7, 8]. Therefore, the next step is to quest whether such advantages can be translated into practical machine learning tasks.

A primary candidate of supervised QML models in the near-term is the so-called quantum neural networks [9]. The key idea of this approach is to utilize the Hilbert space as the feature space for machine learning tasks by exploiting the power of parameterized quantum circuits (PQCs) and data-embedding circuits [10, 11]. Due to the mathematical equivalence, learning models based on quantum kernels can also be in the same class of models as the quantum neural networks [12]; quantum neural networks and quantum kernel methods are categorized as explicit and implicit models within the same concept of QML, respectively [13, 14, 15, 16, 17]. Focus of this paper is on these supervised QML models.

Thus far, a number of attempts have been made to understand these quantum models analytically. In contrast to the original expectation, some literatures have elucidated that the high expressivity of the models causes trainability problems such as barren plateaus [18, 19, 20, 21, 22] and the exponential concentration [23, 24, 25, 26]. This immediately suggests the need to carefully construct quantum models for real-world applications. A remedy for the issue in clas-

1. Department of Mechanical Engineering, Keio University, Hiyoshi 3-14-1, Kohoku, Yokohama, 223-8522, Japan

2. Materials Informatics Initiative, RD Technology & Digital Transformation Center, JSR Corporation, 3-103-9, Tonomachi, Kawasaki-ku, Kawasaki, Kanagawa, 210-0821, Japan

3. Human Biology-Microbiome-Quantum Research Center (WPI-Bio2Q), Keio University, Mita 2-15-45, Minato-ku, Tokyo, 108-8345, Japan

4. Quantum Computing Center, Keio University, Hiyoshi 3-14-1, Kohoku, Yokohama, 223-8522, Japan

*E-mail: yudai.suzuki.sh@gmail.com

sical machine learning is feature selection [27, 28, 29, 30]. With the technique, redundant features can be removed and hence the predictive performance improves. Moreover, the model’s interpretability can be enhanced thanks to the reduced features.

We note that feature selection techniques have been studied in the quantum computing community. For example, Quadratic Unconstrained Binary Optimization problems corresponding to the classical feature selection have been introduced, so that quantum computing can deal with it to boost the performance [31, 32, 33]. Other works use quantum support vector machines with multi-objective genetic algorithms [34], quantum circuit evolution algorithms [35] and graph-theoretic approach using quantum computers [36] for the performance improvement. However, the usability of classical feature selection could be limited for QML tasks. While classical feature selection methods still play a critical role in embedding data on quantum circuits, it does not help understand why some data embedding quantum circuits work better than others. More importantly, when dealing with quantum data tasks such as quantum phase recognition, classical input features do not explicitly appear in input quantum data and thus one cannot employ the classical techniques straightforwardly.

In this work, we propose a feature selection method with a specific focus on QML problems. More specifically, we treat the light-cones structures (i.e., subspaces) of local quantum kernels as quantum features and then select a subset of beneficial ones through training of parameters in the kernels. Note that, by local quantum kernels, we mean a family of quantum kernels that measure local similarity between a pair of data; examples are projected quantum kernels (PQKs) [23] and quantum Fisher kernels (QFKs) [26]. Thanks to the extension, the proposed framework is applicable not only to the feature selection of classical inputs, but also to the quantum data tasks that conventional feature selection methods cannot handle. To the best of our knowledge, this is the first work to introduce a QML-oriented feature selection method. Numerical simulations using toy tasks demonstrate the effectiveness of our scheme for four applications: (1) classical feature selection, (2) circuit architecture search for data embedding, (3) the compression of QML models and (4) relevant subspace selection for quantum data tasks. These results indicate that our proposal will encourage practitioners to use QML models for real-world applications. In addition, as elaborated in Conclusion & Discussion section, this scheme is linked to the argument about classical simulability and trainability [37]. Therefore, our scheme could work as a practical test to see if quantumness is

really needed for the QML tasks at hand, while this is left for future work.

The rest of this paper is organized as follows. In Section 2, we provide the settings of QML framework and elaborate on our proposal. Then, Section 3 demonstrates numerical simulations to see the effectiveness and versatility of our proposal. Lastly, we discuss the potential of our methods in practical situations and then conclude this paper in Section 4.

2 Method

2.1 Framework of Quantum Machine Learning

In this work, we focus on supervised quantum machine learning (QML) tasks; given a training dataset that consists of input data and the corresponding targets, our goal is to train a QML model so that its output accurately predicts the targets of new unseen data. Here, we consider the case where the dataset is either classical or quantum.

In case of classical data, the dataset is composed of pairs of classical input $\mathbf{x}_i$ and its target value or label y_i ; an example of the input is a vectorized d -dimensional image $\mathbf{x} \in \mathbb{R}^d$. For QML models to deal with the classical data, we have to encode the input to quantum states via embedding quantum circuits. Namely, we apply an embedding circuit $V(\mathbf{x}_i)$ to an initial state ρ_0 to construct data-dependent quantum states $\rho_i = V(\mathbf{x}_i)\rho_0V^\dagger(\mathbf{x}_i)$. We can also utilize parameterized quantum circuits (PQCs) $U(\boldsymbol{\theta})$ with tunable parameters $\boldsymbol{\theta}$ to find better features in the space. Typically, we employ data re-uploading technique [38] to create quantum states from classical inputs;

$$\rho_i(\boldsymbol{\theta}) = U(\mathbf{x}_i, \boldsymbol{\theta})\rho_0U^\dagger(\mathbf{x}_i, \boldsymbol{\theta}) \quad (1)$$

with L -layer data re-uploading quantum circuits

$$U(\mathbf{x}, \boldsymbol{\theta}) = \prod_{l=1}^L U(\boldsymbol{\theta}_l)V(\mathbf{x}). \quad (2)$$

In the numerical experiments in Section 3, we use the re-uploading quantum circuits with a brick-like structure [19]; an example for $n = 6$ and $L = 3$ is shown in Figure 1.

As for quantum data, the dataset constitutes quantum states and their targets, i.e., $\{\rho_i, y_i\}$ [39, 40]. In this case, we employ PQCs $U(\boldsymbol{\theta})$ to the quantum inputs to find better representation for the tasks;

$$\rho_i(\boldsymbol{\theta}) = U(\boldsymbol{\theta})\rho_iU^\dagger(\boldsymbol{\theta}). \quad (3)$$

Previous work defined two approaches of constructing quantum models [13]: explicit models where model

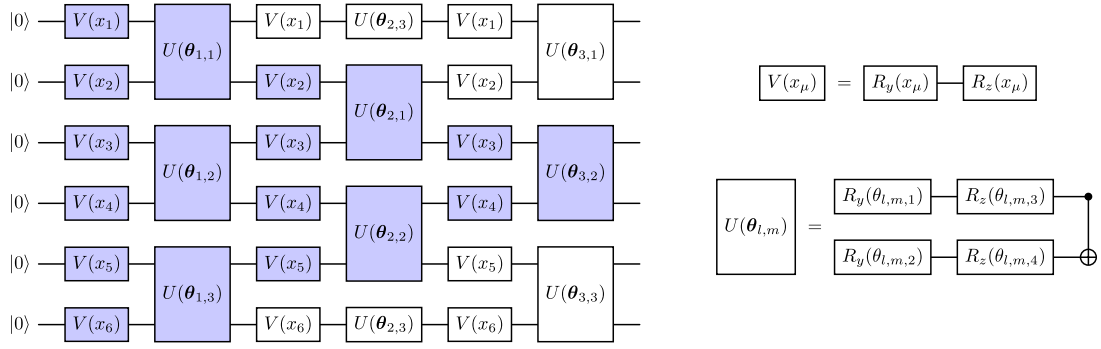


Fig. 1 An example of the embedding quantum circuit with $L = 3$ for six features. The colored gates represent the light-cone of measurement on l -th qubit with $l = 3, 4$ to construct the PQKs in Eq. (7).

predictions are given by measuring a certain Hermitian operator O on quantum states, and implicit models where the inner products of quantum states called quantum kernels are used for predictions. Then, with a training dataset $S = \{\mathbf{x}_i, y_i\}_{i=1, \dots, N}$ ($S = \{\rho_i, y_i\}_{i=1, \dots, N}$) of size N , we perform optimization to obtain performant quantum models. For explicit models, we optimize θ to minimize a loss function $\mathcal{L}(\theta)$ represented as

$$\mathcal{L}(\theta) = \sum_{i=1}^N l(f_i(\theta), y_i) \quad (4)$$

where $f_i(\theta) = \text{Tr}[\rho_i O]$ is the output of explicit model and $l(f_i(\theta), y_i)$ is the loss such as squared loss, i.e., $l(f_i(\theta), y_i) = (f_i(\theta) - y_i)^2$. On the other hand, the representer theorem [41, 42, 43] guarantees that an optimal solution of implicit models can be in the form of $f_i^{\text{opt}}(\mathbf{x}) = \sum_i \alpha_i^{\text{opt}} k(\mathbf{x}_i, \mathbf{x})$ with a kernel function $k(\mathbf{x}, \mathbf{x}')$. A common choice of the kernel for QML is the fidelity quantum kernel [10] expressed as

$$k_Q(\mathbf{x}_i, \mathbf{x}_j) = \text{Tr}[\rho_i(\theta) \rho_j(\theta)]. \quad (5)$$

For instance, the optimization problem for binary classification is reduced to minimizing the following function over the parameter α ;

$$\mathcal{L}(\alpha) = - \sum_i \alpha_i + \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j). \quad (6)$$

Note that the kernel function can include the parameters θ as shown in Eq. (5); the parameters θ can also be optimized separately [44] or simultaneously [45]. Basically, classical optimizers are employed for both models in the optimization.

2.2 Light-Cone Feature Selection

We here extend the idea of classical feature selection [27] to propose its counterpart for QML. The classical feature selection methods can be categorized into

three types: the filter, wrapper, and embedded methods. The filter approach performs selections based on the relationship between features and the targets. The wrapper approach evaluates the performance of the model for different subsets of features and then select features based on a performance metric. Lastly, the embedded approach provides important features via training of models such as LASSO. See, e.g., Ref. [28, 29, 30] for more comprehensive review of classical feature selections.

Our proposal is similar in spirit to the embedded approach; we select relevant local light-cone of quantum models through training of local quantum kernels. By local quantum kernels, we mean quantum kernels which measure local similarity between a pair of data. Note that fidelity quantum kernels in Eq. (5) measure global similarity and thus do not fall into the category. The examples are projected quantum kernels (PQKs) [23] and simplified quantum Fisher kernels (QFKs) [26] defined as

$$k_{PQ}(\mathbf{x}_i, \mathbf{x}_j) = \sum_l \lambda_l \text{Tr}[\rho_i^{(l)}(\theta) \rho_j^{(l)}(\theta)] \quad (7)$$

and

$$k_{QF}(\mathbf{x}_i, \mathbf{x}_j) = \sum_l \lambda_l \text{Tr}[\rho_0 \{\tilde{B}_{i, \theta_l}, \tilde{B}_{j, \theta_l}\}] \quad (8)$$

respectively. Here, $\rho_i^{(l)} = \text{Tr}_l[\rho_i(\theta)]$ is the partial trace of the quantum state $\rho_i(\theta)$ over all qubits except for the l -th qubit. Also, $\tilde{B}_{i, \theta_l} = U_{1:l}^\dagger(\mathbf{x}_i, \theta) B_{\theta_l} U_{1:l}(\mathbf{x}_i, \theta)$ with $B_{\mathbf{x}_i, \theta_l} = 2i(\partial U(\mathbf{x}_i, \theta)/\partial \theta_l) U^\dagger(\mathbf{x}_i, \theta)$, where $U_{k:l}(\mathbf{x}, \theta)$ denotes a sequence of unitary gates from $U_k(\mathbf{x}, \theta_k)$ to $U_l(\mathbf{x}, \theta_l)$ in the representation $U(\mathbf{x}, \theta) = U_{N_p}(\mathbf{x}, \theta_{N_p}) \cdots U_2(\mathbf{x}, \theta_2) U_1(\mathbf{x}, \theta_1)$ for the total number of parameters N_p .

These local quantum kernels can be expressed as summation of local terms, $k(\mathbf{x}, \mathbf{x}') = \sum_l \lambda_l k^{(l)}(\mathbf{x}, \mathbf{x}')$, where each local kernel $k^{(l)}(\mathbf{x}, \mathbf{x}')$ has the corresponding light-cone, i.e, the subspace of quantum circuits (or

quantum states) that are relevant to a certain operation like a local gate operation and measurement. For example, as illustrated in Figure 1, the light-cone obtained by measurement on l -th qubit ($l = 3, 4$) for the PQK is denoted by the colored gates. We regard the light-cones induced by local quantum kernels as features and then select important ones through training of parameters $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ in quantum kernels.

As for the training of quantum kernels, we maximize the kernel target alignment (KTA) [46] defined as

$$\text{KTA} = \frac{\sum_{pq} y_p y_q k(\mathbf{x}_p, \mathbf{x}_q)}{\sqrt{\sum_{pq} k(\mathbf{x}_p, \mathbf{x}_q)^2} \sqrt{\sum_{pq} y_p y_q}}. \quad (9)$$

We note that the KTA is a common choice as an objective function for training kernels in both classical [44, 46] and quantum cases [24, 25, 47]. In this work, the optimization of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is done in two steps: (1) with $\boldsymbol{\lambda}$ fixed, $\boldsymbol{\theta}$ are optimized via the Adam optimizer by computing the gradients of Eq. (9) with respect to $\boldsymbol{\theta}$, and then, (2) $\boldsymbol{\lambda}$ are optimized following the approach proposed in Ref. [44]:

$$\boldsymbol{\lambda}^{\text{opt}} = \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|_1}, \quad (10)$$

$$\mathbf{v}^* = \arg \min_{\mathbf{v} \geq \mathbf{0}} \left[\sum_{ij} v_i M_{ij} v_j - 2 \sum_i v_i a_i \right], \quad (11)$$

$$M_{ij} = \sum_{pq} k^{(i)}(\mathbf{x}_p, \mathbf{x}_q) k^{(j)}(\mathbf{x}_p, \mathbf{x}_q), \quad (12)$$

$$a_i = \sum_{pq} k^{(i)}(\mathbf{x}_p, \mathbf{x}_q) y_p y_q. \quad (13)$$

This two-step process is iteratively carried out until the value of the KTA stops changing significantly. The optimization can be done efficiently, as the first step relies on gradient-based algorithms and the solution of the second step can be obtained by solving a quadratic program. In our numerical experiments, Eq. (11) in the second step is solved using CVXOPT library [48, 49]. After the optimization, we select relevant light-cones based on the amplitude of parameters $\boldsymbol{\lambda}$; that is, we interpret the local kernel $k^{(l)}(\mathbf{x}, \mathbf{x}')$ whose parameter λ_l has a large absolute value as important.

We remark that our proposal is more versatile than classical feature selection methods in QML tasks. This scheme can be used not only for feature selection of classical features as the conventional methods do, but also for seeking relevant subspace of embedding quantum circuits. More importantly, the proposed method is applicable to quantum data tasks where classical features do not appear and hence conventional ones cannot work. In addition, while some previous works proposed methods for feature selection using quantum computing

devices, all of them focus on improvement of the classical feature selection [31, 32, 33, 34, 36, 35]. Therefore, to the best of our knowledge, this work is the first to propose a QML-oriented feature selection scheme.

3 Numerical Demonstration

In what follows, we perform numerical simulations to demonstrate efficacy and versatility of our proposal. More specifically, we deal with four application tasks: (1) feature selection of classical inputs, (2) circuit architecture search for data embedding, (3) compression of QML models and (4) subspace feature selection for quantum data. For all the simulations shown below, PennyLane library [50] is used for quantum circuit simulation. Also, in the learning tasks, we used 80 data points for both training and test, where models' performance are checked for different 25 settings of data samples and initial parameters $\boldsymbol{\theta}$; five sets of data samples and five different parameters are prepared and then we tried all combinations of these settings. As for local quantum kernels, we set a uniform weights to the initial parameters $\boldsymbol{\lambda}$, i.e., $\lambda_l = 1/(\text{the number of terms in the local quantum kernel})$ for all l . The learning rate and the number of iterations in the Adam optimization step are set to 0.01 and 500, respectively.

3.1 Feature Selection of Classical Inputs

To begin with, we check the performance of our scheme for classical feature selection tasks. Here, we utilize the importance of light-cones determined by our proposal to seek out relevant classical features; that is, we evaluate the importance of the features based on a quantity determined by the optimized parameters $\boldsymbol{\lambda}$. We call the quantity the importance score. In this study, we define the importance score $P(\mu)$ of a classical feature μ as

$$P(\mu) = \frac{1}{\mathcal{N}} \sum_l w_{\lambda_l}(\mu) \lambda_l \quad (14)$$

where $w_{\lambda_l}(\mu)$ is the frequency for the classical feature μ to appear in the l -th local quantum kernel $k^{(l)}(\mathbf{x}, \mathbf{x}')$, and $\mathcal{N} = \sum_{\mu} \tilde{P}(\mu)$ with $\tilde{P}(\mu) = \sum_l w_{\lambda_l}(\mu) \lambda_l$ is the normalization constant. Recall that $\boldsymbol{\lambda}$ satisfy $\sum_l \lambda_l = 1$ with $\lambda_l \geq 0$. In case of the light-cone structure shown in Figure 1 (i.e., λ_l with $l = 3$ or 4 for PQKs), $w_{\lambda_l}(\mu) = 1$ for $\mu = x_1, x_6$, $w_{\lambda_l}(\mu) = 2$ for $\mu = x_2, x_5$, and $w_{\lambda_l}(\mu) = 3$ for $\mu = x_3, x_4$. We notice that one can take another quantity as the importance score, so long as it tells the relevance of features in the tasks.

As a toy example, we consider a so-called parity-based binary classification task introduced in Ref. [13]

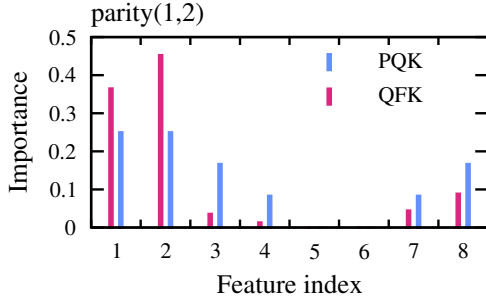


Fig. 2 Importance scores of features for the parity(1,2) dataset.

The goal of this task is to classify d -dimensional binary input $\mathbf{x} \in \{-1, 1\}^d$ whose label is determined based on its parity for a certain subset $A_p \subset \{1, \dots, d\}$; namely, the input's label is assigned according to $y(\mathbf{x}) = \prod_{i \in A_p} x_i$. We chose this task because important classical features are clearly A_p and hence we can interpret the performance of our proposal easily. In the following, we consider a dataset with $A_p = \{1, 2\}$, which we denote as parity(1,2). Here, the number of qubits n is equal to the dimension of inputs d and we set $d = 8$. Also, the data re-uploading quantum circuits with the depth $L = 3$ is used.

Figure 2 shows the importance scores of the local quantum kernels for the parity(1,2) dataset. Features whose importance scores are higher than the threshold $P_{th} = 0.1$ for QFKs are x_1 and x_2 , suggesting our scheme with QFKs can pinpoint key features for the task. As for PQKs, features x_1, x_2, x_3 and x_8 witness the scores larger than P_{th} . Hence, the PQKs can also select relevant features on the dataset. The features chosen by the PQKs include more features than the minimum required. This is because the number of classical features in the light-cone induced by a local measurement gets larger as the depth increases [51]. We note that the light-cone of l -th term of QFKs is determined by the position of gates with the local parameter θ_l [26], and thus only necessary features can be selected in this case. Actually, the performance difference between these two kernels can be seen by looking at the KTAs; the values of the KTAs for PQKs and QFKs are 0.605 and 0.987, respectively. Recall that the KTA of the best kernel outputs one, and the larger value of the quantity indicates the better performance. Hence, the performance gap between the PQK and the QFK would be enlarged as the depth increases. Yet, we could resolve the problem in the PQKs, e.g., by using different feature scores.

3.2 Circuit Architecture Search for Data Embedding

Next, we check if our scheme can be exploited to find better embedding quantum circuits. For classical data tasks, embedding quantum circuits play a crucial role in the performance, since the data distribution in the Hilbert space is determined by them. Previous works introduced techniques to increase the expressivity in terms of Fourier analysis [52, 53]. Still, there is room for investigation in guidelines for constructing performant embedding quantum circuits. Hence, we here propose a new embedding strategy based on our light-cone feature selection technique.

Specifically, this work focuses on the order in which classical features are encoded onto qubits. A straightforward and the most popular strategy is to assign a feature x_i to the i -th qubit for $i = 1, \dots, n$. However, due to the restricted connectivity of entangling gates, underlying patterns of data in the Hilbert space might not be found with the standard encoding approach in some cases; for example, the features assigned in the farthest qubits in the quantum circuit could be less correlated, leading to poor expressivity of the model for certain tasks. With this in mind, we introduce a problem-specific ordering of data encoding. Here, we consider the following scheme: we start with the standard encoding described above and, after the light-cone feature selection optimization, the order of features encoded onto the qubits are sorted in descending order of their importance scores in Eq. (14). Let us note that our technique is applicable to more complicated settings; one could also use this idea to find a better gate set among some candidates.

For the proof-of-concept demonstration of this scheme, we again work on the parity-based binary classification task. In the numerical experiments, we consider the dataset with $A_p = \{1, 5\}$ denoted as parity(1,5). Table 1 shows the KTA values for the cases with and without the re-ordering technique. Here, the optimization of local kernels are done three times for both cases. It turns out that the KTAs increase from -0.033 to 0.987 (-0.015 to 0.560) for QFKs (PQKs). The performance improvement can be interpreted in terms of feature selection. Figure 3 shows the importance scores for these cases. When the technique is not employed, every feature is evenly regarded as important for QFKs and PQKs. Together with the KTA values, this indicates that the optimization works poorly because the restriction on entangling gates limits the spread of light-cones. Note that a few local terms in PQKs can cover features x_1 and x_5 , but they are so far away that optimization could not lead to good prediction. On the other hand, QFKs and PQKs can success-

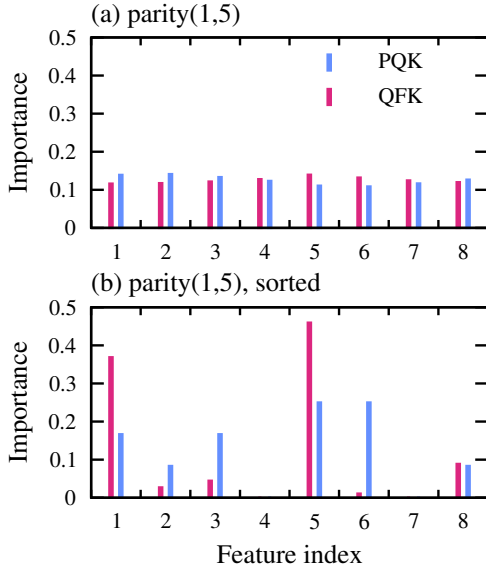


Fig. 3 Importance scores of features for the parity(1,5) datasets (a) without and (b) with re-ordering of input encoding.

Table 1 The KTA values for the parity(1,5) and the breast-cancer datasets.

parity(1,5)		PQK	QFK
without sorting	train	0.037	0.077
	test	-0.015	-0.033
with sorting	train	0.562	0.987
	test	0.560	0.987

breast-cancer		PQK	QFK
without sorting	train	0.397	0.749
	test	0.366	0.719
with sorting	train	0.389	0.752
	test	0.377	0.726

fully select the relevant features, suggesting that the re-ordering facilitates the optimization and results in the better performance.

Moreover, we work on the benchmark dataset, the breast-cancer dataset, to see the performance for a more practical task. This is a classical dataset for binary classification where 212 data points are labeled as “malignant” and 357 as “benign”. In the numerical experiments, we choose $d(=12)$ features out of 30 and then rescale the data so that it has a mean zero and unit variance. The distribution of each feature is shown in Figure 4 (a), in which we show the L1 norm of “malignant” and “benign” distributions δ to show the discrepancy (i.e., separability) of them. The dataset is taken from scikit-learn [54]. Also, we employ the same quantum circuits as the parity-based classification tasks.

As shown in Table 1, the re-ordering technique increases the KTA values of QFKs (PQKs) from

0.719 to 0.726 (from 0.366 to 0.377) for test, while the KTA of PQKs for training slightly decreases. The improvement in the test data is reasonable because more compact models are constructed by our scheme. Figure 4 (b) illustrates importance scores with and without re-ordering. We observed that four features ($x_5, x_{10}, x_{11}, x_{12}$) for QFKs with re-ordering show smaller importance scores than the smallest value for the case without the technique (0.04 for x_6), implying that the model gets sparse thanks to our scheme. Similarly, the number of important features selected by PQKs with the threshold $P_{th} = 0.1$ is reduced by this scheme; six features (x_3, x_4, x_5, x_6, x_7 and x_8) are higher than the threshold for the case without the method, whereas five (x_1, x_4, x_7, x_8 and x_{10}) are selected for re-ordering. From these results, we can confirm the effectiveness of the new encoding scheme based on the importance score.

3.3 Compression of Quantum Machine Learning Models

Another application where our proposal can be utilized is the compression of quantum circuit models. The model compression aims to reduce the size of machine learning models, e.g., the number of parameters in neural networks [55]; as a result, resource efficiency, faster inference and better generalization can be realized.

Especially, pruning is a powerful tool to reduce the model size in classical machine learning. With the pruning technique, one can construct a compressed model that requires less resources to execute inference and is comparable to the original model in the performance. For example, parameters of neural networks are pruned based on their magnitudes [56] (i.e., the parameters with small values are removed). Indeed, the magnitude-based pruning method has been explored in variational quantum algorithms [57, 58]. On the other hand, careful consideration is needed when applying the magnitude-based pruning to data re-uploading quantum circuits. More concretely, this technique fails to prune the embedding layer, suggesting its inability to reduce the model width (i.e., the number of qubits). This also means that the reduction in the depth might not be huge as the embedding layers will never be removed.

With a focus on the problem, we propose a new pruning method using the light-cone feature selection. Here, we evaluate the magnitudes of parameters λ to determine which part of quantum circuits are redundant. Recall that λ are the weights for local terms and thus their amplitudes tell the importance of each light-cone. Consequently, this scheme can prune even the embedding layers.

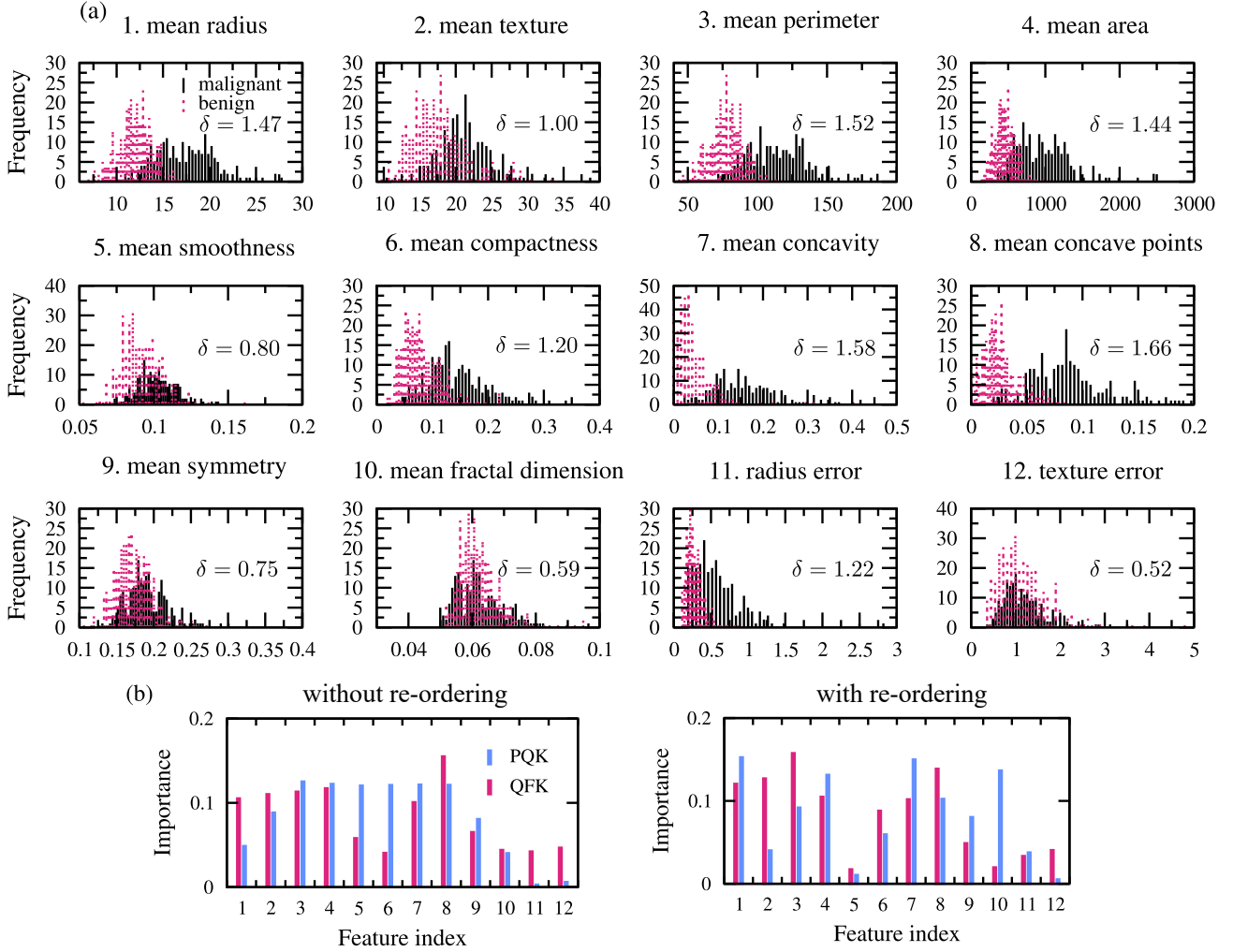


Fig. 4 Results on the breast-cancer dataset. (a) Histograms of the 12 features of the breast-cancer dataset used in this work are illustrated. In each panel, δ denotes the L1 norm of the distributions computed with the fixed number of bins (100). (b) Importance scores are calculated for the cases with (right) and without (left) re-ordering technique.

To see how the pruning scheme works, we consider the parity(1,2) dataset. To be precise, we prune all the light-cones of models except for the one with the largest parameter λ and check the relationship to the performance of the corresponding compressed model. Figure 5 depicts the quantum circuits constructed by removing all but the light-cone associated with the largest λ for the PJK and QFK, respectively. The light-cone chosen by this scheme is reasonable as the important subspace in this task is the part of the circuit where features x_1 and x_2 are included. Moreover, the performances of the pruned models are comparable to the original models; the KTA value of PJKs (QFKs) just decreases from 0.605 to 0.600 (0.987 to 0.983) by pruning. Thus, our numerical experiment suggests the validity of our proposal.

3.4 Subspace Feature Selection for Quantum Data

Lastly, we work on a task in which the goal is to find relevant light-cones (i.e., subspaces) of quantum data. We recall that conventional feature selection methods cannot deal with the tasks because classical features do not appear. The advantage of the approach would be that one can construct an efficient and better-performing QML model by removing irrelevant subspaces of quantum states.

To see the effectiveness of our proposal, we consider a toy binary classification task with a synthetic quantum dataset $\{\rho_i, y_i\}$. Here, the input quantum state is represented as $\rho_i = U(\theta^*)\rho_{0,i}U^\dagger(\theta^*)$, where $\rho_{0,i} = \rho_{0,i}^{(1)} \otimes \rho_{0,i}^{(2)} \otimes \dots \otimes \rho_{0,i}^{(n)}$ is the tensor product of single-qubit Haar random states $\{\rho_{0,i}^{(l)}\}_{l=1,\dots,n}$ and $U(\theta^*)$ denotes the unitary operator with certain fixed parameters θ^* .

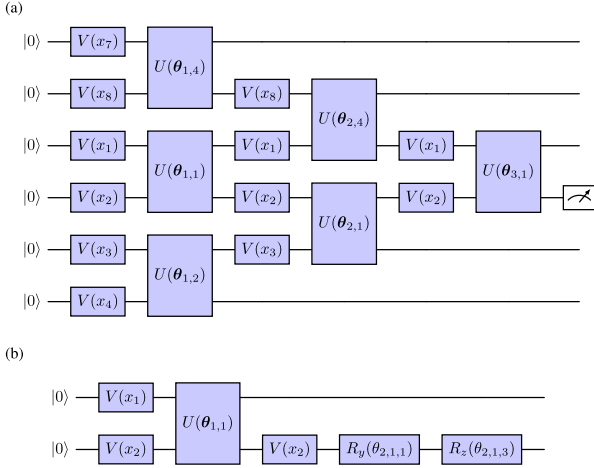


Fig. 5 Pruned quantum circuits. Here, we consider the light-cone of the local term of (a) the PQK and (b) the QFK whose parameter λ is largest for the parity(1,2) dataset.

Then, the label is determined by $y_i = \text{sign}(\text{Tr}[Z_m \rho_{0,i}])$ with the Pauli Z operator acting on the m -th qubit Z_m ; see Figure 6 (a) for the details of this numerical experiment. We construct the dataset inspired by actual quantum data tasks; like the quantum phase recognition tasks in Ref. [59], we assume that there exist a unitary operator that can transform input quantum state ρ_i to another state in the form of $\tilde{\rho}_{i,r} \otimes \rho_{irr}$ with a relevant quantum state $\tilde{\rho}_{i,r}$ and an irrelevant state ρ_{irr} . By a relevant quantum state, we mean that a simple measurement on that state is enough to identify the label or the property of the original state. In this work, we focus on PQKs using alternating layered ansatzes [19] with the number of qubit $n = 8, 12$ and the circuit depth $L = 1, 2, 3$. Note that we do not take into account QFKs since the kernels are not defined via input quantum states ρ_i , but via the unitary representation of data $U(\mathbf{x}_i, \boldsymbol{\theta})$. Also, we set $m = n/2 + 1$.

Figure 6 (b) shows the amplitudes of $\boldsymbol{\lambda}$ after optimization of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for each setting. We clearly see that PQKs for the depth $L = 1, 2$ can point out the relevant quantum subspace. For $L = 3$, however, this method selects the $n/2$ -th qubit as the most important. This would be attributed to how the parameters $\boldsymbol{\theta}$ are optimized; the light-cone induced by the measurement of $n/2$ -th qubit also covers the key qubit and the optimization in the light-cone is better than others. To see if the understanding is correct, we compute the KTA values of all the possible PQKs where only a single-qubit reduced density matrix is used. As shown in the insets of Figure 6 (b), we found that the PQK with $\lambda_{n/2} = 1$ and 0 otherwise outputs the best KTA for $L = 3$, indicating the reduced density matrix $\rho_i^{(n/2)}$ is the most relevant

subspace. In summary, we can numerically validate the potential of our scheme using a toy quantum data task.

4 Conclusion & Discussion

In this work, we propose a QML-oriented feature selection method where important light-cone features (subspaces) of quantum models are selected through training of local quantum kernels such as PQKs and QFKs. Numerical simulations demonstrate that our scheme works well in (1) classical feature selection, (2) circuit architecture search for data embedding, (3) model compression and (4) subspace feature selection for quantum data. These results suggest its versatility and encourage practitioners to use this technique for the application of QML models to real-world tasks.

As described in Section 2, our framework relies on the training of parameters $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ in quantum kernels; this means that the classical computational cost for the quantum kernel estimation scales at least quadratically with respect to the number of training data points. Hence, there is room for investigation in improving its scalability. As the remedy to the similar issue in quantum kernel methods, previous works used techniques such as classical shadow and surrogate models based on random Fourier features [23, 60, 61]. Thus, it would be interesting to explore the improvement of its scalability in this direction.

According to Ref. [37], the absence of trainability issues such as barren plateaus and exponential concentrations is strongly tied with classical simulability. This indicates that, for classical data tasks, our technique might not work in the classically-hard regime; even if it works, the compressed model obtained through our method could be classically simulable. On the other hand, this also means that our method can be used as a practical test to see if the QML tasks really requires quantumness. More significantly, our proposal is also applicable to quantum data tasks such as quantum phase recognition [62, 59, 63] and entanglement detection [64] (i.e., quantum physics-related learning tasks). Therefore, our proposal could also be used for constructing efficient settings of such quantum tasks and the investigation of practical quantum advantages.

Acknowledgements This work was supported by MEXT Quantum Leap Flagship Program Grant Number JPMXS0118067285 and JPMXS0120319794. Y.S. was supported by Grant-in-Aid for JSPS Fellows 22KJ2709.

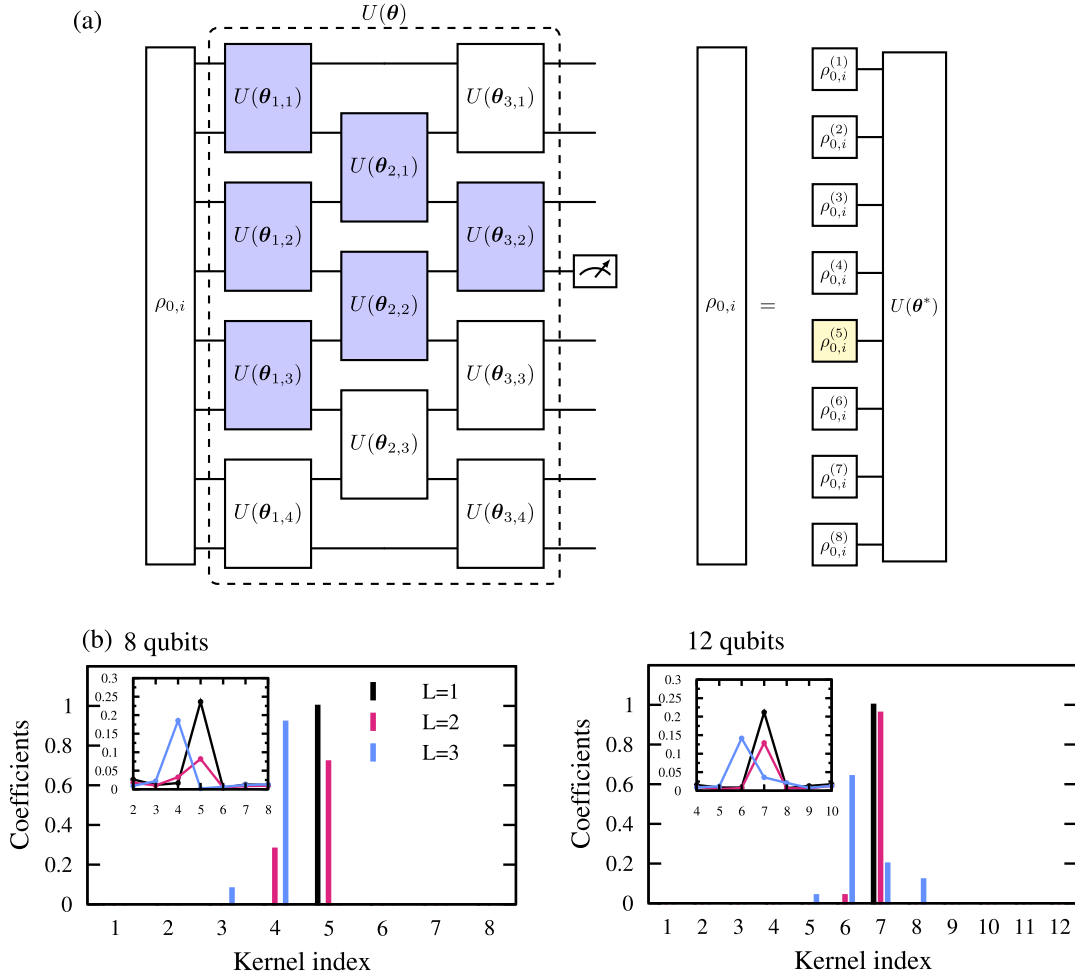


Fig. 6 Results for a toy quantum data task. (a) The setting of the QML model with the PQC's depth $L = 3$ and the quantum data considered in this experiment are illustrated. The colored quantum state $\rho_{0,i}^{(l)}$ with $l = n/2 + 1$ (in this case, $l = 5$) is a key state used to assign its label y_i . (b) Optimized parameters λ of the PQKs for quantum data tasks with $n = 8, 12$ are shown. Here, quantum circuits with $L = 1, 2, 3$ are used. The insets of figures illustrate the averaged KTA of PQKs calculated using a single reduced density matrix $\rho_i^{(l)}(\theta)$; that is, l in the horizontal axis denotes the PQK in Eq. (7) with $\lambda_l = 1$ and 0 otherwise.

References

1. N. Wiebe, D. Braun, S. Lloyd, Quantum Algorithm for Data Fitting, *Physical Review Letters* **109**(5), 050505 (2012)
2. E. Aïmeur, G. Brassard, S. Gambs, Quantum speed-up for unsupervised learning, *Machine Learning* **90**, 261 (2013)
3. P. Rebentrost, M. Mohseni, S. Lloyd, Quantum support vector machine for big data classification, *Physical Review Letters* **113**(13), 130503 (2014)
4. S. Lloyd, M. Mohseni, P. Rebentrost, Quantum principal component analysis, *Nature Physics* **10**(9), 631 (2014)
5. J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, S. Lloyd, Quantum machine learning, *Nature* **549**(7671), 195 (2017)
6. Y. Liu, S. Arunachalam, K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, *Nature Physics* **17**(9), 1013 (2021)
7. J. Jäger, R.V. Krems, Universal expressiveness of variational quantum classifiers and quantum kernels for support vector machines, *Nature Communications* **14**(1), 576 (2023)
8. T. Muser, E. Zapusek, V. Belis, F. Reiter, Provable advantages of kernel-based quantum learners and quantum preprocessing based on Grover's algorithm, *arXiv preprint arXiv:2309.14406* (2023)
9. M. Schuld, I. Sinayskiy, F. Petruccione, The quest for a quantum neural network, *Quantum Information Processing* **13**, 2567 (2014)
10. V. Havlíček, A.D. Córcoles, K. Temme, A.W. Harrow, A. Kandala, J.M. Chow, J.M. Gambetta, Supervised learning with quantum-enhanced feature spaces, *Nature* **567**(7747), 209 (2019). DOI <https://doi.org/10.1038/s41586-019-0980-2>
11. M. Schuld, N. Killoran, Quantum machine learning in feature Hilbert spaces, *Physical Review Letters* **122**(4), 040504 (2019). DOI <https://doi.org/10.1103/PhysRevLett.122.040504>
12. M. Schuld, Supervised quantum machine learning models are kernel methods, *arXiv preprint arXiv:2101.11020* (2021). DOI <https://doi.org/10.48550/arXiv.2101.11020>

13. S. Jerbi, L.J. Fiderer, H. Poulsen Nautrup, J.M. Kübler, H.J. Briegel, V. Dunjko, Quantum machine learning beyond kernel methods, *Nature Communications* **14**(1), 517 (2023)
14. H. Yano, Y. Suzuki, K.M. Itoh, R. Raymond, N. Yamamoto, Efficient Discrete Feature Encoding for Variational Quantum Classifier, *IEEE Transactions on Quantum Engineering* **2**, 1 (2021)
15. N.A. Nghiem, S.Y.C. Chen, T.C. Wei, Unified framework for quantum classification, *Physical Review Research* **3**(3), 033056 (2021)
16. D. Herman, R. Raymond, M. Li, N. Robles, A. Mezzacapo, M. Pistoia, Expressivity of Variational Quantum Machine Learning on the Boolean Cube, *IEEE Transactions on Quantum Engineering* **4**, 1 (2023)
17. R. Sweke, E. Recio, S. Jerbi, E. Gil-Fuster, B. Fuller, J. Eisert, J.J. Meyer, Potential and limitations of random fourier features for dequantizing quantum machine learning, *arXiv preprint arXiv:2309.11647* (2023)
18. J.R. McClean, S. Boixo, V.N. Smelyanskiy, R. Babbush, H. Neven, Barren plateaus in quantum neural network training landscapes, *Nature Communications* **9**(1), 4812 (2018). DOI <https://doi.org/10.1038/s41467-018-07090-4>
19. M. Cerezo, A. Sone, T. Volkoff, L. Cincio, P.J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, *Nature Communications* **12**(1), 1791 (2021). DOI <https://doi.org/10.1038/s41467-021-21728-w>
20. Z. Holmes, A. Arrasmith, B. Yan, P.J. Coles, A. Albrecht, A.T. Sornborger, Barren Plateaus Preclude Learning Scramblers, *Physical Review Letters* **126**(19), 190501 (2021). DOI <https://doi.org/10.1103/PhysRevLett.126.190501>
21. Z. Holmes, K. Sharma, M. Cerezo, P.J. Coles, Connecting Ansatz Expressibility to Gradient Magnitudes and Barren Plateaus, *PRX Quantum* **3**(1), 010313 (2022). DOI [10.1103/PRXQuantum.3.010313](https://doi.org/10.1103/PRXQuantum.3.010313)
22. L. Leone, S.F. Oliviero, L. Cincio, M. Cerezo, On the practical usefulness of the Hardware Efficient Ansatz, *arXiv preprint arXiv:2211.01477* (2022). DOI <https://doi.org/10.48550/arXiv.2211.01477>
23. H.Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, J.R. McClean, Power of data in quantum machine learning, *Nature Communications* **12**, 2631 (2021)
24. J. Kübler, S. Buchholz, B. Schölkopf, The Inductive Bias of Quantum Kernels, *Advances in Neural Information Processing Systems* **34**, 12661 (2021). URL <https://proceedings.neurips.cc/paper/2021/file/69adc1e107f7f7d035d7baf04342e1ca-Paper.pdf>
25. S. Thanasilp, S. Wang, M. Cerezo, Z. Holmes, Exponential concentration and untrainability in quantum kernel methods, *arXiv preprint arXiv:2208.11060* (2022). DOI <https://doi.org/10.48550/arXiv.2208.11060>
26. Y. Suzuki, H. Kawaguchi, N. Yamamoto, Quantum Fisher kernel for mitigating the vanishing similarity issue, *arXiv preprint arXiv:2210.16581* (2022). DOI <https://doi.org/10.48550/arXiv.2210.16581>
27. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**, 1157 (2003)
28. Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, *Bioinformatics* **23**(19), 2507 (2007)
29. G. Chandrashekar, F. Sahin, A survey on feature selection methods, *Computers & Electrical Engineering* **40**(1), 16 (2014)
30. N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr, J.M. O'Sullivan, A review of feature selection methods for machine learning-based disease risk prediction, *Frontiers in Bioinformatics* **2**, 927312 (2022)
31. C. Zoufal, R.V. Mishmash, N. Sharma, N. Kumar, A. Sheshadri, A. Deshmukh, N. Ibrahim, J. Gacon, S. Woerner, Variational quantum algorithm for unconstrained black box binary optimization: Application to feature selection, *Quantum* **7**, 909 (2023)
32. S. Mücke, R. Heese, S. Müller, M. Wolter, N. Piatkowski, Feature selection on quantum computers, *Quantum Machine Intelligence* **5**(1), 11 (2023)
33. G. Hellstern, V. Dehn, M. Zaefferer, Quantum computer based feature selection in machine learning, *IET Quantum Communication* (2023)
34. H. Wang, A novel feature selection method based on quantum support vector machine, *arXiv preprint arXiv:2311.17646* (2023)
35. A.S. Albino, O.M. Pires, M.Q. Nooblath, E.G. Nascimento, Evolutionary quantum feature selection, *arXiv preprint arXiv:2303.07131* (2023)
36. S. Chakraborty, S.H. Shaikh, A. Chakrabarti, R. Ghosh, A hybrid quantum feature selection algorithm using a quantum inspired graph theoretic approach, *Applied Intelligence* **50**(6), 1775 (2020)
37. M. Cerezo, M. Larocca, D. García-Martín, N. Diaz, P. Braccia, E. Fontana, M.S. Rudolph, P. Bermejo, A. Ijaz, S. Thanasilp, et al., Does provable absence of barren plateaus imply classical simulability? or, why we need to rethink variational quantum computing, *arXiv preprint arXiv:2312.09121* (2023)
38. A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, J.I. Latorre, Data re-uploading for a universal quantum classifier, *Quantum* **4**, 226 (2020). DOI <https://doi.org/10.22331/q-2020-02-06-226>
39. L. Schatzki, A. Arrasmith, P.J. Coles, M. Cerezo, Entangled Datasets for Quantum Machine Learning, *arXiv preprint arXiv:2109.03400* (2021)
40. E. Perrier, A. Youssry, C. Ferrie, QDataSet, quantum datasets for machine learning, *Scientific Data* **9**(1), 582 (2022)
41. G. Kimeldorf, G. Wahba, Some results on Tchebycheffian spline functions, *Journal of Mathematical Analysis and Applications* **33**(1), 82 (1971)
42. A.J. Smola, B. Schölkopf, *Learning with Kernels*, vol. 4 (Berlin, Germany: GMD-Forschungszentrum Informationstechnik, 1998)
43. T. Hofmann, B. Schölkopf, A.J. Smola, Kernel methods in machine learning, *The Annals of Statistics* **36**(3), 1171 (2008)
44. C. Cortes, M. Mohri, A. Rostamizadeh, Algorithms for learning kernels based on centered alignment, *Journal of Machine Learning Research* **13**(28), 795 (2012). URL <http://jmlr.org/papers/v13/cortes12a.html>
45. V. Sydorov, M. Sakurada, C.H. Lampert, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1402–1409
46. N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. Kandola, On kernel-target alignment, *Advances in Neural Information Processing Systems* **14** (2001)
47. S. Miyabe, B. Quanz, N. Shimada, A. Mitra, T. Yamamoto, V. Rastunkov, D. Alevras, M. Metcalf, D.J. King, M. Mamouei, et al., Quantum multiple kernel learning in financial classification tasks, *arXiv preprint arXiv:2312.00260* (2023)
48. M. Andersen, J. Dahl, Z. Liu, L. Vandenberghe, Interior-point methods for large-scale cone programming, *Optimization for machine learning* pp. 55–83 (2011)

49. M.S. Andersen, J. Dahl, L. Vandenberghe. CVXOPT: A python package for convex optimization, version 1.3.2. URL <http://www.cvxopt.org>
50. V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M.S. Alam, G. Alonso-Linaje, B. Akash-Narayanan, A. Asadi, J.M. Arrazola, U. Azad, S. Banning, C. Blank, T.R. Bromley, B.A. Cordier, J. Ceroni, A. Delgado, O.D. Matteo, A. Dusko, T. Garg, D. Guala, A. Hayes, R. Hill, A. Ijaz, T. Isaacsson, D. Ittah, S. Jangiri, P. Jain, E. Jiang, A. Khandelwal, K. Kottmann, R.A. Lang, C. Lee, T. Loke, A. Lowe, K. McKiernan, J.J. Meyer, J.A. Montañez-Barrera, R. Moyard, Z. Niu, L.J. O’Riordan, S. Oud, A. Panigrahi, C.Y. Park, D. Polatajko, N. Quesada, C. Roberts, N. Sá, I. Schoch, B. Shi, S. Shu, S. Sim, A. Singh, I. Strandberg, J. Soni, A. Száva, S. Thabet, R.A. Vargas-Hernández, T. Vincent, N. Vitucci, M. Weber, D. Wierichs, R. Wiersema, M. Willmann, V. Wong, S. Zhang, N. Killoran. PennyLane: Automatic differentiation of hybrid quantum-classical computations (2022)
51. Y. Suzuki, M. Li, Effect of alternating layered ansatzes on trainability of projected quantum kernel, arXiv preprint arXiv:2310.00361 (2023)
52. M. Schuld, R. Sweke, J.J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, *Physical Review A* **103**(3), 032430 (2021)
53. S. Shin, Y.S. Teo, H. Jeong, Exponential data encoding for quantum supervised learning, *Physical Review A* **107**(1), 012422 (2023)
54. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825 (2011)
55. Y. Cheng, D. Wang, P. Zhou, T. Zhang, A survey of model compression and acceleration for deep neural networks, arXiv preprint arXiv:1710.09282 (2017)
56. J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks, arXiv preprint arXiv:1803.03635 (2018)
57. S. Sim, J. Romero, J.F. Gonthier, A.A. Kunitsa, Adaptive pruning-based optimization of parameterized quantum circuits, *Quantum Science and Technology* **6**(2), 025019 (2021)
58. H. Wang, Y. Ding, J. Gu, Y. Lin, D.Z. Pan, F.T. Chong, S. Han, in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2022), pp. 692–708
59. E. Lake, S. Balasubramanian, S. Choi, Exact quantum algorithms for quantum phase recognition: Renormalization group and error correction, arXiv preprint arXiv:2211.09803 (2022)
60. K. Nakaji, H. Tezuka, N. Yamamoto, Deterministic and random features for large-scale quantum kernel machine, arXiv preprint arXiv:2209.01958 (2022)
61. S. Shin, Y.S. Teo, H. Jeong, Analyzing quantum machine learning using tensor network, arXiv preprint arXiv:2307.06937 (2023)
62. I. Cong, S. Choi, M.D. Lukin, Quantum convolutional neural networks, *Nature Physics* **15**(12), 1273 (2019)
63. Y. Wu, B. Wu, J. Wang, X. Yuan, Quantum phase recognition via quantum kernel methods, *Quantum* **7**, 981 (2023)
64. O. Gühne, G. Tóth, Entanglement detection, *Physics Reports* **474**(1-6), 1 (2009)