Quantum Machine Intelligence (2024) 6:52
https://doi.org/10.1007/s42484-024-00191-y

RESEARCH ARTICLE

On the interpretability of quantum neural networks

Lirandë Pira[1,2] · **Chris Ferrie**[1]

Received: 8 January 2024 / Accepted: 13 August 2024 / Published online: 28 August 2024

Abstract
Interpretability of artificial intelligence (AI) methods, particularly deep neural networks, is of great interest. This heightened focus stems from the widespread use of AI-backed systems. These systems, often relying on intricate neural architectures, can exhibit behavior that is challenging to explain and comprehend. The interpretability of such models is a crucial component of building trusted systems. Many methods exist to approach this problem, but they do not apply straightforwardly to the quantum setting. Here, we explore the interpretability of quantum neural networks using local model-agnostic interpretability measures commonly utilized for classical neural networks. Following this analysis, we generalize a classical technique called LIME, introducing Q-LIME, which produces explanations of quantum neural networks. A feature of our explanations is the delineation of the region in which data samples have been given a random label, likely subjects of inherently random quantum measurements. We view this as a step toward understanding how to build responsible and accountable quantum AI models.

Keywords Quantum neural networks · Interpretable machine learning · Q-LIME · Explainable AI

1 Introduction

Artificial intelligence (AI) has become ubiquitous. Often manifested in machine learning algorithms, AI systems promise to be evermore present in everyday high-stakes tasks (Russell and Norvig 2010; Mitchell 1997). This is why building fair, responsible, and ethical systems is crucial to the design process of AI algorithms. Central to the topic of *trusting* AI-generated results is the notion of *interpretability*, also known as *explainability* (Russell and Norvig 2022; Rudin et al. 2022; Du et al. 2022; Doshi-Velez and Kim 2017; Roscher et al. 2020). The interpretability of AI models is a pivotal concern in contemporary AI research, particularly with the ubiquity of deep neural networks. This has given rise to research topics under the umbrella of interpretable machine learning (or IML) and explainable AI (or XAI), noting that the terms *interpretable* and *explainable* are used synonymously throughout the corresponding literature. Generically, interpretability is understood as the extent to which humans comprehend the output of an AI model that leads to decision-making (Russell and Norvig 2022). Humans strive to understand the "thought process" behind the decisions of the AI model — otherwise, the system is referred to as a "black box". Even though the terms "interpretation" and "explanation" are used colloquially in the literature with varying degrees, throughout this paper, we use them as follows. *Interpretable* entails a model that humans can understand and comprehend through direct observation of its internal workings or outputs. It implies that the model is intuitive to the human observer and that no further tools are required. On the other hand, *explanation* refers to the output of a tool used to articulate the behavior of a model. In our case, and indeed many others, an explanation is a model itself, which is justifiably called such when it is interpretable. Complex models and their simpler explanations are often called black-box and white-box models, respectively.

The precise definition of a model's interpretability has been the subject of much debate (Molnar et al. 2020; Lipton 2018). However, its definition becomes mathematically precise when a specific interpretability technique is introduced. Naturally, there exist learning models which are more interpretable than others, such as simple decision trees. On the other hand, the models we prefer best for solving complex tasks, such as deep neural networks (DNNs), happen to be highly non-interpretable, which is due to their inherent nonlinear layered architecture Goodfellow et al. 2016. We note that DNNs are one of the most widely used techniques in machine learning (Krizhevsky et al. 2012). Thus, the interpretability of neural networks is an essential topic within the field of interpretable ML research (Zhang et al. 2021;

✉ Lirandë Pira
lirande.pira@student.uts.edu.au

[1] Centre for Quantum Software and Information, University of Technology Sydney, Sydney, Australia

[2] Present Address: Centre for Quantum Technologies, National University of Singapore, Singapore, Singapore

Guidotti et al. 2018). In this work, we focus on precisely the topic of interpretability as we consider the quantum side of neural networks.

In parallel, recent years have witnessed a surge of research efforts in *quantum* machine learning (QML) (Biamonte et al. 2016; Schuld and Petruccione 2021). This research area sits at the intersection of machine learning and quantum computing. The development of QML has undergone different stages. Initially, the field started with the quest for speedups or quantum advantages. More recently, the target has morphed into further pursuits in expressivity and generalization power of quantum models. Nowadays, rather than "competing" with classical models, quantum models are further being enhanced on their own, which could, in turn, improve classical machine learning techniques. One of the key techniques currently used in QML research is the variational quantum algorithm, which acts as the quantum equivalent to classical neural networks (Cerezo et al. 2022). To clarify the analogy, we will refer to such models as quantum neural networks (QNNs) (Cerezo et al. 2021).

Given the close conceptual correspondence to classical neural networks, it is natural to analyze their interpretability, which is important for several reasons. Firstly, QNNs may complement classical AI algorithm design, making their interpretability at least as important as classical DNNs. Secondly, the quantum paradigms embedded into QNNs deserve to be understood and explained in their own right. The unique non-intuitive characteristics of quantum nature can make QNNs more complicated to interpret from the point of view of human understandability. Finally, with the growing interest and capabilities of quantum technologies, it is crucial to identify and mitigate potential sources of errors that plague conventional AI due to a lack of transparency.

In this work, we define and establish key notions pertaining to the interpretability of QNNs. Through this formalization, we aim to contribute a robust foundation for advancing the discourse on the interpretability of QNNs within the broader context of quantum machine learning research. In doing so, we generalize some well-known interpretable techniques to the quantum domain. Consider the standard relationship diagram in QML between data and algorithm (or device) type, where either can be classical (C) or quantum (Q). This entails the following combinations (CC, CQ, QC, and QQ), shown in Fig. 1. Classical interpretable techniques are the apparent domain of CC. We will discuss, but not dwell on, the potential need for entirely new techniques when the data is quantum (QC and QQ). In CQ, the domain that covers the so-called quantum-enhanced machine learning techniques, although the data is classical, the output of the quantum devices is irreversibly probabilistic. CQ can be seen as a specialization of the quantum data techniques above. Generalizing classical notions of interpretability to this domain is the subject of our work.

The question of interpretability in quantum machine learning models more broadly, as well as of QNNs more specifically, has already started to receive attention (Heese et al. 2023; Mercaldo et al. 2022; Steinmuller et al. 2022), particularly involving the concept of Shapley values (Lundberg and Lee 2017), which attempt to quantify the importance of features in making predictions. In Heese et al. (2023), interpretability is explored using Shapley values for quantum models by quantifying the importance of each gate to a given prediction. The complexity of computing Shapley values for generalized quantum scenarios is analyzed in Burge et al. (2023). In Steinmuller et al. (2022), Shapley values are computed for a particular class of QNNs. Our work complements these efforts using an alternative notion of explainability to be discussed in detail next.

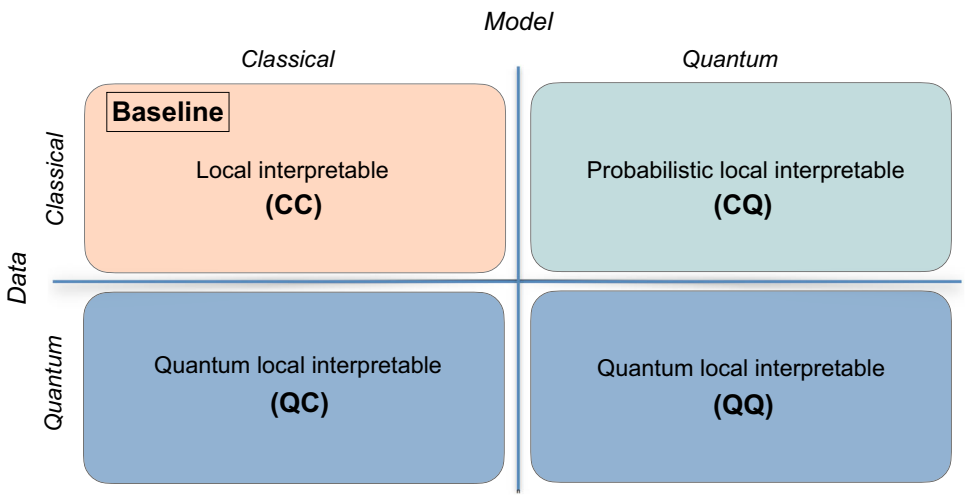

Fig. 1 Categorization of interpretability techniques as they apply to classical and quantum resources. Here, the well-known QML diagram represents data and a learning model, which can be classical (C) or quantum (Q) in four different scenarios. We consider a reformulation of interpretable techniques to be required in the CQ scenario. In the QC and QQ quadrants, the design of explicitly quantum interpretable methods may be required. The scope of this paper covers CQ approaches

2 Interpretability in AI

2.1 Taxonomy

There are several layers to the design of interpretability techniques. To start, they can be *model-specific* or *model-agnostic*. As the name suggests, model-specific methods are more restrictive in terms of which models they can be used to explain. As such, they are designed to explain one single type of model. In contrast, model-agnostic methods allow for more flexibility in usage as they can be used on a wide range of model types. At large, model-agnostic methods can have a *global* or *local* interpretability dimension. Locality determines the scope of explanations with respect to the model. Interpretability at a global level explains the average predictions of the model as a whole. At the same time, local interpretability gives explanations at the level of each sample. In another axis, these techniques can be *active* (inherently interpretable) or *passive* (post hoc). The state of these interpretable paradigms implies the level of involvement of interpretable techniques in the outcome of the other parameters. Active techniques change the structure of the model itself by leaning toward making it more interpretable. In contrast, passive methods explain the model outcome once the training has finished. In comparison to model-agnostic methods, which work with samples at large, there also exist example-based explanations that explain selected data samples from a dataset. An example of this method is the k-nearest neighbors models, which average the outcome of k-nearest selected points.

Other than the idea of building interpretable techniques, or more precisely, techniques that interpret various models, there exist models that are inherently interpretable. Such models include linear regression, logistic regression, naive Bayes classifiers, decision trees, and more. This feature makes them good candidates as surrogate models for interpretability. Based on this paradigm, the concept of surrogate models exists, which uses interpretable techniques as a building block for designing other interpretable methods. Such important techniques are, for example, local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016) and Shapley additive explanations (known as SHAP) (Lundberg and Lee 2017). The idea of surrogate models will be extensively explored in the next sections, as we extend this paradigm to quantum models.

2.2 Interpretability of neural networks

The interpretability of neural networks remains a challenge on its own. This tends to amplify in complex models with many layers and parameters, such as DNNs. Nevertheless, there is active research in the field and several proposed interpretable techniques (Zhang et al. 2021; Ghorbani et al. 2018). Such techniques that aim to gain insights into the decision-making of a neural network include saliency maps (Simonyan et al. 2013), feature visualization (Olah et al. 2017; Yosinski et al. 2015), perturbation or occlusion-based methods (Ribeiro et al. 2016; Lundberg and Lee 2017), and layer-wise relevance propagation (also known by its acronym LRP) (Bach et al. 2015).

To expand further on the above-mentioned techniques, saliency maps use backpropagation and gradient information to identify the most influential regions contributing to the output result. This technique is also called pixel attribution (Molnar et al. 2020). Feature visualization, particularly useful for convolutional natural networks, is a technique that analyses the importance of particular features in a dataset by visualizing the patterns that activate the output. In the same remark, in terms of network visualizations, Ref. Zeiler and Fergus (2014) goes deeper into the layers of a convolutional neural network to gain an understanding of the features. This result, in particular, shows the intricacies and the rather intuitive process involved in the decision-making procedure of a network as it goes through deeper layers. Occlusion-based methods aim to perturb or manipulate certain parts of the data samples to observe how the explanations change. These methods are important in highlighting deeper issues in neural networks. Similarly, layerwise relevance propagation techniques re-assign importance weight to the input data by analyzing the output. This helps the understanding by providing a hypothesis over the output decision. Finally, the class of surrogate-based methods mentioned above is certainly applicable in neural networks as well.

The importance of these techniques goes beyond the interpretability measures for human understanding. In addition, they can be seen as methods of debugging and thus improving the result of a neural network as in Ref. Zeiler and Fergus (2014). Below, we take a closer look at surrogate model-agnostic local interpretable techniques, which are applicable to DNNs as well.

2.3 Local interpretable methods

Local interpretable methods tend to focus on individual data samples of interest (Goldstein et al. 2014). One of these methods relies on explaining a black-box model using inherently interpretable models, also known as surrogate methods. These methods act as a bridge between the two model types. The prototype of these techniques is the so-called local interpretable model-agnostic explanations (LIME), which has received much attention since its invention in 2016 (Ribeiro et al. 2016). Local surrogate methods work by training an interpretable surrogate model that approximates the result of the black-box model to be explained. LIME, for instance, is categorized as a perturbation-based technique that perturbs the input dataset. Locality in LIME refers to the fact that

the surrogate model is trained on the data point of interest, as opposed to the whole dataset (which would be the idea behind *global* surrogate methods). Equation (1) represents the explanation ξ of a sample x via its two main terms, namely the term $L(f, g, \pi_x)$ representing the loss, which is the variable to be minimized, and $\Omega(g)$ which is the complexity measure, which encodes the degree of interpretability. Here f is the black-box model, g is the surrogate model, and π_x defines the region in data space local to x (See Alg. 1 pseudocode). In broader terms, LIME is a trade-off between interpretability and accuracy,

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g). \tag{1}$$

In the following, we make use of the concept of local surrogacy to understand the interpretability of quantum models using LIME as a starting point. Much like LIME, we develop a *framework* to provide explanations of black-box models in the quantum domain. The class of surrogate models, the locality measure, and the complexity measure are free parameters that must be specified and justified in each application of the framework.

Algorithm 1 LIME Algorithm (Ribeiro et al. 2016).

function LIME(f, x, D, K)
  Input:
    f: Classifier model
    x: Data point to be explained
    D: Distribution from which to sample perturbations
    K: Number of samples to generate
  Output: Surrogate model $\xi(x)$
  $Z \leftarrow$ Sample K points from D
  for $i = 1$ to K **do**
    Sample z_i from D
    $g_i = f(z_i)$ ▷ g_i is the prediction of the classifier on the perturbed sample
  end for
  $\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x)$ ▷ L is a loss function, π_x is a locality measure to the data point, and G is the set of surrogate models
  return $\xi(x)$
end function
function LOSS(f, g, π_x)
  $L(f, g, \pi_x) = \sum_{z_i \in D} \pi_x(z_i)[f(z_i) \neq g(z_i)] + \Omega(g)$ ▷ Penalize the difference between predictions
  return $L(f, g, \pi_x)$
end function
function LOCALITY(π_x, z, x)
  $\pi_x(z) = \exp\left(-\frac{d(x,z)^2}{\sigma^2}\right)$ ▷ d is a distance metric, σ is a bandwidth parameter
  return $\pi_x(z)$
end function

3 The case for quantum AI interpretability

As mentioned in Sect. 1, interpretability in the quantum literature in the context of machine learning can take different directions. We consider the case when data is classical and encoded into a quantum state, which is manipulated by a variational quantum circuit before outputting a classical decision via quantum measurement. Our focus is on interpreting the classical output of the quantum model.

A quantum machine learning model f takes as input data x and first produces quantum data $|\psi(x)\rangle$. A trained quantum algorithm, such as the QNN, subsequently processes the quantum data and generates a classification decision based on the outcome of a quantum measurement. This is not conceptually different from a classical neural network beyond the fact that the weights and biases have been replaced by parameters of quantum processes, except for one crucial difference — quantum measurements are unavoidably probabilistic.

Probabilities, or quantities interpreted as such, often arise in conventional neural networks. However, these numbers are encoded in bits and directly accessible, so they are typically used to generate deterministic results (through thresholding, for example). Qubits, on the other hand, are not directly accessible. While procedures exist to reconstruct qubits from repeated measurements (generally called *tomography*), these are inefficient — defeating any purpose of encoding information into qubits in the first place. Hence, QML uniquely forces us to deal with uncertainty in interpreting its decisions.

In the case of probabilistic decisions, the notion of a *decision boundary* is undefined. A reasonable alternative might be to define the boundary as those locations in data space where the classification is purely random (probability $\frac{1}{2}$). A data point here is randomly assigned a label. For such a point, any *explanation* for its label in a particular realization of the random decision process would be arbitrary and prone to error. It would be more accurate to admit that the algorithm is *indecisive* at such a point. This rationale is equally valid for data points near such locations. Thus, we define the *region of indecision* as follows.

Definition of R (Region of Indecision): Let $f : X \to [0, 1]$ be a probabilistic classifier mapping an input space X to a probability score. The region of indecision, R, is defined as the subset of X where the classifier's probability is within an ϵ-margin of $\frac{1}{2}$. Formally,

$$R = \left\{ x' \in X : \left| P(f(x') = 1) - \frac{1}{2} \right| < \epsilon \right\}, \tag{2}$$

where $\epsilon > 0$ is a small constant representing the threshold of uncertainty tolerated in the classification decision. In a certain context, points situated within this rationale lack a discernible *rationale*, hence explanation, for their specific outcomes, relying instead on random circumstances.

At present, although certain data points possess labels that were randomly assigned, one may pose the question as to the underlying rationale for such assignments. In other words, even data points lying within the region of indecision demand an explanation. Next, we will show how the ideas of local interpretability can be extended to apply to the probabilistic or quantum settings. In doing so, we aim to provide a nuanced and comprehensive understanding of model predictions.

3.1 Quantum LIME: probabilistic local interpretability

Here we define Quantum LIME (or Q-LIME) as an extension of the classical counterpart for quantum models, namely quantum local interpretable model-agnostic explanations. A specialization of the definition of Q-LIME is the hybrid version in the CQ, such as in Fig. 1.

In the context of the LIME algorithms, the loss function is typically chosen to compare models and their potential surrogates on a per-sample basis. However, if the model's output is random, the loss function will also be a random variable. An obvious strategy would be to define loss via expectation:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, \mathbb{E}[L(f, g, \pi_x)] + \Omega(g). \tag{3}$$

Nevertheless, even under these circumstances, a definitive assertion regarding ξ as an explanation remains elusive. This stems from the recognition that its predictions only capture the average behavior of the underlying model's randomness. In fact, the label provided by ξ may be the opposite of that assigned to x by the model in any particular instance, which clearly does not capture a reasonable notion of *explanation*.

To mitigate this, we generalize the classical definition and call an *explanation* the distribution Ξ of trained surrogate models g obtained through the application of LIME to the random output of any probabilistic classifier. Note again that g is random, trained on synthetic local data with random labels assigned by the underlying model. Thus, the explanation inherits any randomness from the underlying model. It is not the case that the explanation provides an interpretation of the randomness *per se* — however, we can utilize the distribution of surrogate models to simplify the region of indecision, hence providing an interpretation of it.

The symbol Ξ represents the Q-LIME explanation, emphasizing the interpretability framework tailored for QNNs. Additionally, Alg. 2 refers to a Monte Carlo approximation method employed within the Q-LIME process. This algorithm plays a pivotal role in facilitating the computation of interpretable insights by utilizing a Monte Carlo sampling approach, thereby enhancing the efficiency of the interpretation process.

Algorithm 2 Quantum LIME (Q-LIME).

function QLIME(U, x, D, K)
  Input:
    U: Quantum classifier model
    x: Data point to be explained
    D: Distribution from which to sample perturbations
    K: Number of samples to generate
  Output: List Ξ of quantum surrogate models
  $\Xi \leftarrow$ Empty list to store quantum surrogate models
  for $i = 1$ to K **do**
    $D_i \leftarrow$ Generate synthetic quantum data locally around x
    $\gamma_i \leftarrow$ LIME(U, x, D_i, K) ▷ Apply LIME to generate a quantum surrogate model (see Alg. 1)
    Add γ_i to Ξ
  end for
  return Ξ
end function

3.2 Local region of indecision

In this section, we define the *local region of indecision*. In a broad sense, this is the region of indecision interpreted locally through a distribution of surrogate models. Suppose a particular data point lies within its own local region of indecision. The explanation for its label is thus the data was given a random label, even though its explanation may not be a satisfying one. Moreover, this is a strong statement because — in principle — all possible interpretable surrogate models have been considered in the optimization.

The local region of indecision can be defined as the region of the input space where the classification decision of the quantum model is uncertain.

Definition of B (Local Region of Indecision): *Consider a set of surrogate models $g \in \Xi$ derived from a local interpretability method applied to the probabilistic classifier f. Let π_x be a weighting function that defines the locality around a point x. The local region of indecision B for a data point x in a dataset X as,*

$$B = \left\{ x' \in X : \left| P(g(x') = 1 | f, \Xi, \pi_x) - \frac{1}{2} \right| < \epsilon \right\}, \tag{4}$$

where $\epsilon > 0$ is a small constant representing a threshold of uncertainty tolerated in the classification decision — this time with reference to the *explanation* rather than the underlying model. Probability $P(g(x') = 1 \mid f, \Xi, \pi_x)$ is taken over the distribution of the surrogate models Ξ.

The exact relationship between B and R is complicated, depending on the details of the loss function. However, for a $0-1$ loss function and π_x defining a closed boundary around x, $B = R$ within π_x. In a more general case, B approximates R up to the freedom that elements of Ξ are allowed to deviate from f. Note that the distribution in Eq. (4) is over Ξ as each g provides deterministic labels. This has been formalized in Algorithm 3.

In much the same way that an interpretable model approximates decision boundaries locally in the classical context, a local region of indecision approximates the proper region of indecision in the quantum (or probabilistic) context. The size and shape of this region will depend on several factors, such as the choice of the interpretability technique, the complexity of the surrogate model, and the number of features in the dataset. This region is depicted in Fig. 2.

In the examples presented in this paper, we chose the interquartile range (IQR) as a summary measure for the set of quantum surrogate models Ξ. However, other options are undoubtedly possible. It is important to note that in higher-dimensional spaces, the complexity of quantum models and the diversity of decision boundaries may necessitate alternative summary measures.

4 Numerical experiments for interpreting QNNs

4.1 Dataset and encoding

We use the well-known Iris dataset (Fisher 1936) for our numerical experiments. It consists of a matrix $X \in \mathbb{R}^{150\times 4}$, where each row represents an observation of an iris flower, and each column represents a feature: sepal length, sepal width, petal length, and petal width, all measured in centimeters. The corresponding label vector $y \in \mathbb{R}^{150}$ categorizes each observation into one of three classes: Iris-setosa, Iris-versicolor, and Iris-virginica, with $y_i \in \{0, 1, 2\}$.

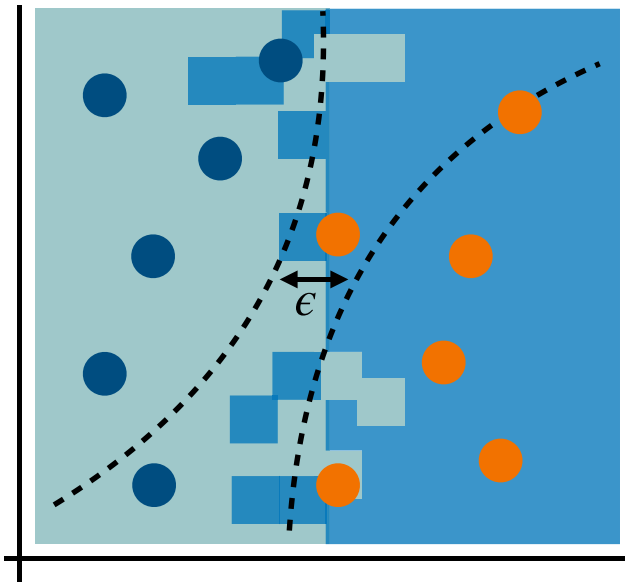

Fig. 2 **Depiction of the concept of the *local region of indecision*.** The space within the dashed lines represents the region in the decision space where data samples exhibit ambiguous classification due to randomness. The figure showcases a two-class classification task in a two-dimensional space with two features represented along the horizontal and vertical axes. Here, ϵ is the pre-defined threshold. Likely, data samples, either inside or close to the region, are interpreted as randomly assigned

Algorithm 3 Range.

Input:
U: Quantum classifier model
x: Data point to be explained
D: Distribution from which to sample perturbations
K: Number of samples to generate in QLIME
M: Number of iterations for sampling quantum surrogate models
Output: Interquartile range (IQR) of quantum decision boundaries
$B \leftarrow$ Empty list to store quantum decision boundaries
$\Xi \leftarrow$ QLIME(U, x, D, K) ▷ Generate quantum surrogate models using Quantum LIME (see Alg. 2)
for $i = 1$ to M **do** ▷ Repeat for M iterations
  $\xi \leftarrow$ Sample from Ξ
  $B_i \leftarrow$ BOUNDARY(ξ) ▷ Obtain the quantum decision boundary from the sampled quantum surrogate model
  Add B_i to B
end for
$IQR \leftarrow$ INTERQUARTILERANGE(B) ▷ Compute the interquartile range of quantum decision boundaries
return IQR
function INTERQUARTILERANGE(B)
  Sort the quantum decision boundaries in B
  $Q_1 \leftarrow$ The median of the lower half of B
  $Q_3 \leftarrow$ The median of the upper half of B
  $IQR \leftarrow Q_3 - Q_1$ ▷ Interquartile range
  return IQR
end function

To enhance the explicability of our methodology, we opt to simplify the analysis by framing it as a binary classification problem. This entails the exclusive utilization of the first two classes, Iris-setosa and Iris-versicolor, thus defining a reduced label vector $y' \in \mathbb{R}^{100}$ where $y'_i \in \{0, 1\}$. Additionally, we restrict our analysis to only two of the four available features: sepal length and sepal width. Hence, we define a reduced feature matrix $X' \in \mathbb{R}^{100\times 2}$ where

$$X' = \begin{pmatrix} \text{sepal length}_1 & \text{sepal width}_1 \\ \text{sepal length}_2 & \text{sepal width}_2 \\ \vdots & \vdots \\ \text{sepal length}_{100} & \text{sepal width}_{100} \end{pmatrix}.$$

which we refer to as the reduced version of the Iris dataset.

In consideration of our indifference toward classifying the two categories of irises, we abstract the names of these classes and labels below. Each data point is encoded into a quantum state with the angle encoding (Schuld and Petruccione 2018), which acts by mapping numerical values to specific angles, facilitating their integration into quantum models. More concretely, the quantum state $|\psi(\boldsymbol{\theta})\rangle$ is parameterized by a vector $\boldsymbol{\theta}$ of angles. These angles can be encoded into the quantum state by applying parameterized quantum gates. From here, the formula for angle encoding is expressed as:

$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|0\rangle^{\otimes n}$, where $U(\boldsymbol{\theta})$ represents the parameterized quantum circuit, and $\boldsymbol{\theta}$ is the vector of angles encoding the classical data. The angle encoding method we employ in our numerical analysis utilizes a default setting of a two-layer scheme.

More concretely, the reduced version of the Iris dataset has been angle encoded into quantum states. Each feature vector $\mathbf{x}_i' = (\text{sepal length}_i, \text{sepal width}_i)$ is encoded into a quantum state $|\psi_i\rangle$ using the following mapping:

$$\mathbf{x}_i' \mapsto |\psi_i\rangle = \cos(\theta_i)|0\rangle + \sin(\theta_i)|1\rangle, \tag{5}$$

where the angle θ_i is defined as:

$$\theta_i = \frac{\pi}{2}\frac{\text{sepal length}_i}{\max(\text{sepal length})}. \tag{6}$$

Similarly, the second feature (sepal width) can be encoded into another quantum state or combined into a multi-qubit system, but for simplicity, we illustrate the encoding of the first feature.

4.2 Model optimization

Following encoding, a variational quantum circuit is applied, parameterized by a set of trainable weights θ. This circuit involves applying Y-rotation gates (R_Y) on each qubit, entangling adjacent qubits via controlled-X CX gates, and applying additional Y-rotations based on the same set of parameters. Formally, for a single qubit state $|\psi_i\rangle$, the initial rotation can be represented as

$$|\psi_i\rangle = R_Y(\theta_i)|0\rangle, \tag{7}$$

where $R_Y(\theta_i) = \exp\left(-i\frac{\theta_i}{2}Y\right)$. The entangling operation between qubits i and j is given by the CX gate:

$$\text{CX}_{ij} = |0\rangle\langle 0|_i \otimes I_j + |1\rangle\langle 1|_i \otimes X_j. \tag{8}$$

The trained quantum model to be explained is a hybrid QNN trained using simultaneous perturbation stochastic approximation (better known by its acronym SPSA) (Spall 1998), which is an optimization algorithm that efficiently estimates the gradient of an objective function. The stopping criterion encapsulated in the number of iterations is kept at 100-iterations. This model has been built and simulated using the Qiskit framework (IBM 2021). The QNN model in question is an autoencoder with alternating layers of single qubit rotations and entangling gates (Farhi and Neven 2018). The encoding and the variational models are depicted in Fig. 3. Since our goal here is to illustrate the local region of indecision, as in Eq. (4), we do not optimize over the complexity of surrogate models and instead fix our search to the class of logistic regression models with two features.

4.3 Approximating the local region

Approximating the local region of indecision requires a trained QNN, a sample data point, a function defining "local", and some meta-parameters defining the quality of the numerical approximation, such as the amount of synthetic data to be used and the number of Monte Carlo samples to take. Once these are defined, the general procedure is as follows, which is formalized in Alg. 2 and Alg. 3. The main steps of this process are sketched below.

1. Create synthetic data locally around the chosen data point. This involves creating a set of perturbed classical data points around x by adding small random variations. These perturbed points are then encoded into quantum states using the same encoding scheme as the original data point.
2. Apply LIME to generate a surrogate model. As sketched throughout the manuscript, LIME is used to explain the predictions of the QNN by approximating them with a simpler model locally around a specific data point.
3. Repeat 1 and 2 several times to produce an ensemble of decision boundaries. To capture the variability of the QNN's decision boundaries, it is essential to repeat the LIME process multiple times. Each repetition results in a new surrogate model and, thus, a new decision boundary that approximates the QNN's decision function.
4. Take the interquartile range (for example) of the generated boundaries. These boundaries represent the regions

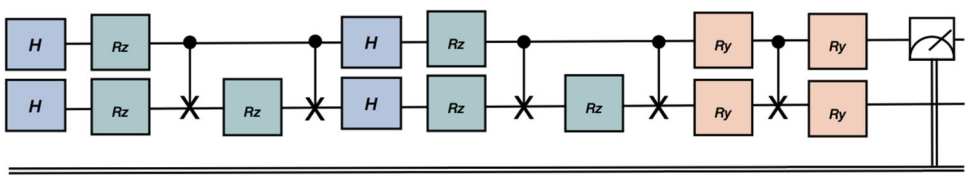

Fig. 3 Quantum circuit schema illustrating the encoding and variational layers. The encoding layer applies Hadamard and RZ gates to each qubit, followed by CX gates to entangle the qubits. The variational layer consists of parameterized RY gates, CX gates for entanglement, and another set of parameterized RY gates, followed by measurement on the first qubit

where the QNN's predictions change in response to the perturbed data points. Statistical measures such as the interquartile range can be used to analyze the set of samples.

To make evident, Fig. 4 is produced through a single application of steps 1 and 2 while Fig. 5 is produced after the final step. It is important to recognize that, in a practical or real-world application, it would be infeasible to plot the entire decision region to any reasonable level of accuracy. The computational demands and complexities involved in plotting such extensive regions can be prohibitive. This makes the region of indecision both valuable and convenient in interpreting QNNs.

The shaded background in each plot of Figs. 4 and 5 shows the decision region of the trained QNN. Upon inspection, it is clear that these decision regions, and the implied boundary, change with each execution of the QNN. By each execution, we refer to repeated runs of the QNN under the same experimental setup, including identical input data and parameter initialization, but resulting in variations due to sampling noise and the stochastic nature of the SPSA optimization algorithm. These variations highlight the inherent probabilistic nature of quantum measurements and the optimization process, leading to an ill-defined decision boundary. In Fig. 4, we naively apply the LIME methodology to two data points — one in the ambiguous region and one deep within the region corresponding to one of the labels. In the latter case, the output of the QNN is nearly deterministic in the local neighborhood of the chosen data point, and LIME works as expected — it produces a consistent explanation.

Nonetheless, in the first example, the data point receives a random label. It is clearly within the *region of indecision* for a reasonably small choice of ϵ. The "explanation" provided

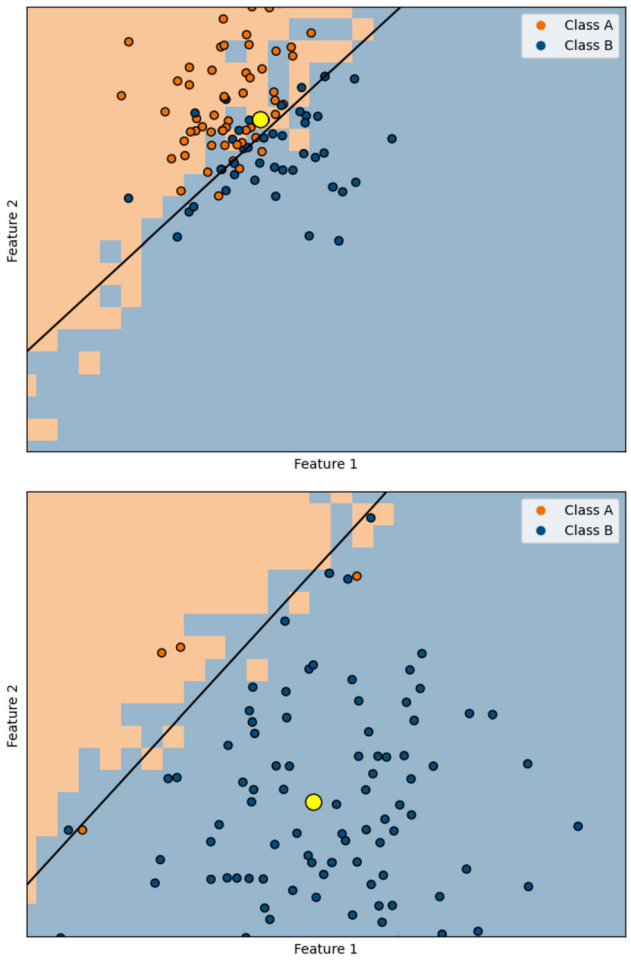

Fig. 4 The classification uncertainty for a chosen data sample (yellow). The two shadings represent the decision boundary of the QNN, which is clearly randomly defined. The synthetic data and the produced surrogate linear decision boundary are shown for a single instance of QNN labels. (top) A clear example of an undefined classification result for the selected data point. (bottom) Decision boundary unambiguously separates the who classes with respect to the selected data point

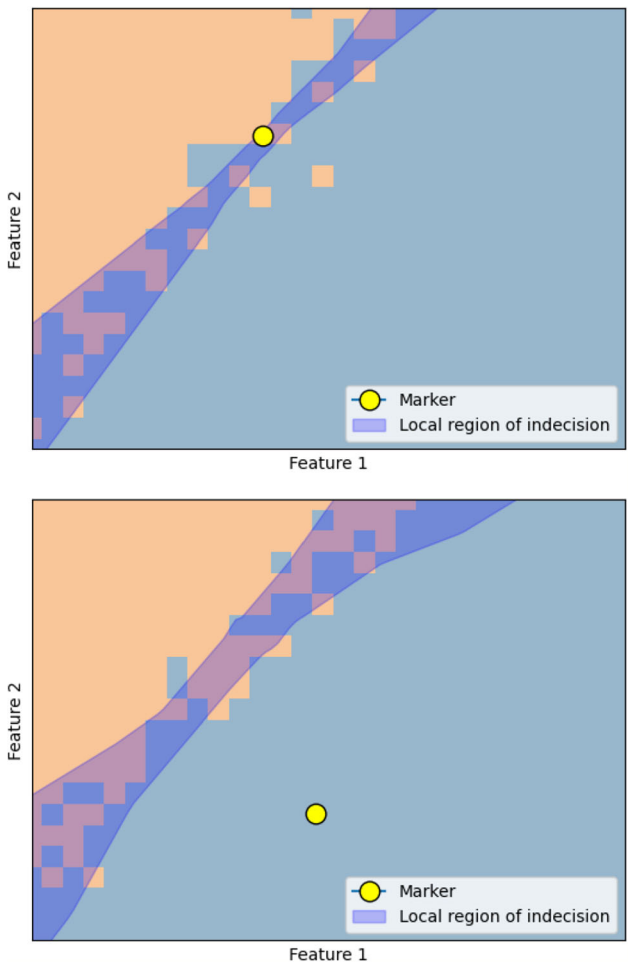

Fig. 5 The approximated local region of indecision. (top) An example of a marked data point that lies in the local region of indecision. (bottom) A data point that is outside of this region can be assessed for interpretability using interpretable techniques

by LIME (summarized by its decision boundary shown as the solid line in Fig. 4) is random. In other words, each application of LIME will produce a different explanation. For the chosen data point, the explanation itself produces opposite interpretations roughly half the time, and the predictions it makes are counter to the actual label provided by the QNN model to be explained roughly half the time. Clearly, this is an inappropriate situation in which such interpretability techniques are applied. Heuristically, if a data point lies near the decision boundary of the surrogate model for QNN, we should not expect that it provides a satisfactory explanation for its label. Q-LIME and the *local region of indecision* rectify this.

For the same sample data points, the local region of indecision is shown in Fig. 5. Data points within their own region should be regarded as having been classified due purely to chance — and hence have no "explanation" in the conventional sense for the label they have been assigned. Note that the region itself is unlikely to yield an analytic form. Hence, some numerical approximation is required to calculate it in practice. Our approach, conceptually, was to repeat what was done to produce Fig. 4 many times and summarize the statistics to produce Fig. 5. A more detailed description follows, and the implementation to reproduce the results presented here can be found at Pira and Ferrie (2023).

5 Discussion and future directions

In summary, in interpreting the classification output of a QNN, we have defined Q-LIME and the local region of indecision, a simple description of a region where the QNN is interpreted to be indecisive. The data samples within their associated region should not be expected to have an "explanation" in the deterministic classical sense. While directly useful for hybrid quantum-classical models, we hope this stimulates further research on the fundamental differences between quantum and classical model interpretability. In the remainder of this section, we discuss possible future research directions.

Our results are pointed squarely at the randomness of quantum measurements, which might suggest that they are "backward compatible" with classical models. Indeed, randomness is present in training classical DNNs due to random weight initialization or the optimization techniques used. However, this type of randomness can be "controlled" via the concept of a *seed*. Moreover, the penultimate output of DNNs (just before label assignment) is often a probability distribution. Nevertheless, these are produced through arbitrary choices of the activation function (i.e., the Softmax function), which force the output of a vector that merely *resembles* a probability distribution. Each case of randomness in classical DNNs is very different from the innate and unavoidable randomness of quantum models. While our techniques *could* be applied in a classical setting, the conclusions drawn from them may ironically be more complicated to justifiably action.

In this work, we have provided concrete definitions for QNNs applied to binary classification problems. Using a probability of $\frac{1}{2}$ would not be a suitable reference point in multi-class decision problems. There are many avenues to generalize our definitions, which would mirror standard generalizations in the move from binary to multi-class decision problems. One such example would be defining the region of indecision as that with nearly maximum entropy.

We took as an act of brevity the omission of the word *local* in many places where it may play a pivotal role. For example, the strongest conclusion our technique can reach is that a QNN is merely *locally* inexplicable in the classical sense. In such cases, we could concede that (for some regions of data space) the behavior of QNN is inherently stochastic, lacking any discernible patterns or deterministic explanations. Alternatively, we can use this conclusion to signal that an explanation at a higher level of abstraction is required. This shift toward higher-level abstraction becomes imperative for gaining more insights into QNNs, acknowledging their inherent quantum nature and the limitations imposed by classical interpretability paradigms. Classically, should a data point inquire, "Why has this label been assigned to me?", within a quantum framework, an initial response may invoke the inherent stochastic nature of quantum systems. Nonetheless, persistent interrogation from the data point could necessitate a quantum extension of *global* interpretability methodologies to clarify the specific features or attributes contributing to the assigned label.

Referring back to Fig. 1, we have focused here on CQ quantum machine learning models. Nevertheless, the core idea behind local surrogate models remains applicable in the context of quantum data — use interpretable *quantum* models as surrogate models to explain black box models producing quantum data. Of course, one of our assumptions is the parallels between the classical interpretable models we mentioned above, with their quantum equivalents. This can be a line for future work. The ideas here encapsulate inherently quantum models such as matrix product states or tensor network states, which can act as surrogate models for quantum models as they may be considered more interpretable. These surrogate models can serve as interpretable proxies, facilitating a comprehension of the underlying quantum processes, thus paving the way for enhanced interpretability of quantum models.

Furthermore, the idea behind *interpreting* or "opening up" black-box models may be of interest in control theory (Youssry et al. 2023, 2020, 2023). In this scenario, the concept of "grey-box" models — portions of which encode specific physical models — give insights into how to engineer certain parameters in a system. These grey-box models can

thus be considered *partially explainable* models. The proposed algorithm in Weitz et al. (2023) may also be of interest in terms of creating intrinsically quantum interpretable models, which would act as surrogates for other more complex quantum models.

Additionally, advanced ideas for future work entail exploring and developing specific metrics tailored for assessing interpretability in QML models. This may involve defining measures that capture the degree of coherence or entanglement contributing to model predictions.

An obvious open question that inspires future research remains to investigate the difference in computational tractability of interpretability methods in quantum versus classical. This will lead to understanding whether it is more difficult to interpret quantum models as opposed to classical models. We hope such results shed light on more philosophical questions as well, such as *is inexplicability, viz. complexity, necessary for learning?*. This exploration holds the potential to advance our understanding not only of the interpretability challenges in quantum models, but also of the fundamental relationship between complexity and the learning process itself.

For completeness, the case for the interpretability of machine learning models does not go without critique. Certain viewpoints contend that the pursuit of insights into a model's decision-making process should not come at the expense of sacrificing performance, and, in practical terms, this prioritization might not always be feasible or deemed necessary (Sarkar 2022; McCoy et al. 2022). It is often suggested that simpler models inherently possess a greater degree of explainability. Nonetheless, it is the more complex models that require explanations, as they may be more likely employed in critical applications.

Regardless of the two distinct camps of beliefs, the niche field of interpretable machine learning keeps growing in volume. An argument is that having a more complete picture of the model's performance can help improve the performance of the model overall. As QML becomes increasingly relevant to AI research, we expect the need for quantum interpretability will also be in demand. This research contributes to a broader understanding of explicability in quantum AI models, paving the way for the development of accountable and transparent systems in the quantum computing and AI era.

More broadly, trusting high-risk, untested, and unproven technologies requires much caution and careful consideration, especially when potential outcomes are uncertain. A thorough analysis via interpretable machine learning techniques will lead to increased social acceptance of AI technology, including those eventual applications powered by quantum models. Our framework helps identify regions where the quantum model's decisions are less stable and unreliable, thereby contributing to a more robust interpretation of quantum models. By providing clearer insights into how and why quantum models make decisions, this framework enhances transparency and, subsequently, social trust in deploying these advanced technologies.

Author Contributions Both authors developed the idea. LP prepared the figures, and wrote the pseudocode snippets. Both authors wrote and reviewed the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. LP was supported by the Sydney Quantum Academy, Sydney, NSW, Australia.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):e0130140

Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S (2016) Quantum Machine Learning. Nature 549:195–20. https://doi.org/10.1038/nature23474

Burge I, Barbeau M, Garcia-Alfaro J (2023) A quantum algorithm for shapley value estimation. arXiv preprint arXiv:2301.04727

Cerezo M, Arrasmith A, Babbush R, Benjamin SC, Endo S, Fujii K, McClean JR, Mitarai K, Yuan X, Cincio L, Coles PJ (2021) Variational quantum algorithms. Nature Rev Phys 3(9):625–64. https://doi.org/10.1038/s42254-021-00348-9

Cerezo M, Verdon G, Huang HY, Cincio L, Coles P (2022) Challenges and opportunities in quantum machine learning. Nature Comput Sci. https://doi.org/10.1038/s43588-022-00311-3

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77

Farhi E, Neven H (2018) Classification with quantum neural networks on near term processors. arXiv preprint https://doi.org/10.48550/arXiv.1802.06002

Fisher RA (1936) The use of multiple measurements in taxonomic problems. https://archive.ics.uci.edu/ml/datasets/iris

Ghorbani A, Abid A, Zou J (2018) Interpretation of neural networks is fragile

Goldstein A, Kapelner A, Bleich J, Pitkin E (2014) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning MIT Press,. http://www.deeplearningbook.org

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42

Heese R, Gerlach T, Mücke S, Müller S, Jakobs M, Piatkowski N (2023) Explaining quantum circuits with shapley values: Towards explainable quantum machine learning. arXiv preprint https://doi.org/10.48550/arXiv.2301.09138

IBM (2021) Qiskit: An Open-Source Framework for Quantum Computing. https://qiskit.org/. Accessed on 1 May 2023

Krizhevsky A, Sutskever I, Hinton GE (2012) In: Advances in Neural Information Processing Systems, vol 25, (eds). by Pereira F, Burges C, Bottou L, Weinberger K. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31–57

Lundberg SM, Lee SI (2017) In: Advances in Neural Information Processing Systems, vol. 30 Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

McCoy L, Brenna C, Chen S, Vold K, Das S (2022) Believing in black boxes: Must machine learning in healthcare be explainable to be evidence-based? Journal of Clinical Epidemiolog. https://doi.org/10.1016/j.jclinepi.2021.11.001

Mercaldo F, Ciaramella G, Iadarola G, Storto M, Martinelli F, Santone A (2022) Towards explainable quantum machine learning for mobile malware detection and classification. Appl Sci 12.https://doi.org/10.3390/app122312025

Mitchell TM (1997) Machine Learning, 1st edn. McGraw-Hill Inc, USA

Molnar C (2022) Interpretable Machine Learning, 2nd edn. https://christophm.github.io/interpretable-ml-book

Molnar C, Casalicchio G, Bischl B (2020) In: ECML PKDD 2020 Workshops Springer International Publishing, pp 417–431. https://doi.org/10.1007/978-3-030-65965-3_28

Olah C, Mordvintsev A, Tyka M (2017) Feature visualization: How neural networks build up their understanding of images. Distill 2(11)

Pira L, Ferrie C (2023) Interpret QNN: Explicability and Inexplicability in the Interpretation of Quantum Neural Networks. https://github.com/lirandepira/interpret-qnn. Accessed on 24 July 2023

Ribeiro MT, Singh S, Guestrin C (2016) In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining association for computing machinery, New York, NY, USA, KDD '16, pp 1135–114

Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable machine learning for scientific insights and discoveries. Ieee Access 8:42,200–42,216

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2021) Interpretable machine learning: Fundamental principles and 10 grand challenges

Russell S, Norvig P (2010) Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall

Sarkar A (2022) Is explainable AI a race against model complexity? arXiv preprint https://doi.org/10.48550/arXiv.2205.10119

Schuld M, Petruccione F (2018) Supervised Learning with Quantum Computers. Quantum Science and Technology Springer International Publishin. https://doi.org/10.1007/978-3-319-96424-9

Schuld M, Petruccione F (2021) Machine Learning with Quantum Computers Springer International Publishing

Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint https://doi.org/10.48550/arXiv.1312.6034

Spall JC (1998) An overview of the simultaneous perturbation method for efficient optimization. J Hopkins APL Tech Dig 19(4):482–492

Steinmüller P, Schulz T, Graf F, Herr D (2022) explainable ai for quantum machine learning. arXiv preprint https://doi.org/10.48550/arXiv.2211.01441

Weitz G, Pira L, Ferrie C, Combes J (2023) Sub-universal variational circuits for combinatorial optimization problem

Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) In: Proceedings of the 32nd international conference on machine learning (ICML), pp 1582–1591

Youssry A, Paz-Silva GA, Ferrie C (2020) Characterization and control of open quantum systems beyond quantum noise spectroscopy. npj Quantum Information 6:95

Youssry A, Paz-Silva GA, Ferrie C (2023) Noise detection with spectator qubits and quantum feature engineering. New J Phys 25:073004

Youssry A, Yang Y, Chapman RJ, Haylock B, Lenzini F, Lobino M, Peruzzo A (2023) Experimental graybox quantum system identification and control. arXiv preprint https://doi.org/10.48550/arXiv.2206.12201

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. European conference on computer vision pp 818–833

Zhang Y, Tino P, Leonardis A, Tang K (2021) A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence 5(5):726–7. https://doi.org/10.1109/tetci.2021.3100641