ADVANCED QUANTUM TECHNOLOGIES
www.advquantumtech.com

A Multi-Class Quantum Kernel-Based Classifier

Shivani Mahashakti Pillay, Ilya Sinayskiy,* Edgar Jembere, and Francesco Petruccione*

Multi-class classification problems are fundamental in many varied domains in research and industry. A popular strategy for solving multi-class classification problems involves first transforming the problem into many binary classification problems. However, this requires the number of binary classification models that need to be developed to grow with the number of classes. Recent work in quantum machine learning has seen the development of multi-class quantum classifiers that circumvent this growth by learning a mapping between the data and a set of label states. This work presents the first multi-class SWAP-Test classifier inspired by its binary predecessor and the use of label states in recent work. With this classifier, the cost of developing multiple models is avoided. In contrast to previous work, the number of qubits required, the measurement strategy, and the topology of the circuits used is invariant to the number of classes. In addition, unlike other architectures for multi-class quantum classifiers, the state reconstruction of a single qubit yields sufficient information for multi-class classification tasks. Both analytical results and numerical simulations show that this classifier is not only effective when applied to diverse classification problems but also robust to certain conditions of noise.

1. Introduction

Quantum machine learning is an emergent field that seeks to leverage the unique properties of quantum computers to solve problems in machine learning. One class of problems that has received considerable attention in quantum machine learning is classification. Quantum kernel methods, inspired by the link between natural operations that are performed on quantum computers and kernel methods, have been developed to solve these problems.[1–4] At the core of these methods is the estimation of a kernel as the squared state overlaps between two quantum states encoding the classical data.

The SWAP-Test classifier is a quantum kernel method that estimates a weighted sum of kernel values for a given test datum and all the training data in parallel.[4] This sum is the result of applying the SWAP-Test, involving only a single qubit measurement, to a quantum state encoding the test datum, training data and their respective labels in a specific format.[5] In contrast to other quantum kernel methods, the need to prepare the test datum for each estimation of a kernel value is avoided. In addition, a decision rule that makes use of this sum is obtained without the aid of a classical subroutine for training.

The SWAP-Test classifier is a binary classifier that can categorize data into only two categories. However, many real-world problems in image processing, natural language processing, and other domains require classifiers that can perform multi-class classification. The binary SWAP-Test classifier can be applied to a multi-class classification problem if the problem is first reduced to many binary classification problems. This can be achieved using one of two strategies: One-versus-All or One-versus-One.[6] But, the number of binary classification models that must be developed will grow with the number of classes and ambiguities[7,8] can arise when combining the results of these models.

Some of the recent work in quantum machine learning has seen the development of multi-class quantum and quantum-inspired classifiers that avoid these heuristic strategies.[9–16] Most recently, the quantum-inspired methods[15,16] use techniques from quantum state discrimination for multi-class classification. Other work has seen the development of quantum convolutional neural networks (QCNNs) for multi-class classification.[9,12] In these methods, a QCNN is trained to reduce the cross entropy loss between its output, interpreted as a probability distribution over all the classes, and the one-hot encodings of the classes.

S. M. Pillay, E. Jembere
School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Durban 4001, South Africa
E-mail: shivani.pillay@nithecs.ac.za

S. M. Pillay, I. Sinayskiy, F. Petruccione
National Institute for Theoretical and Computational Sciences (NITheCS)
Stellenbosch 7604, South Africa
E-mail: sinayskiy@ukzn.ac.za

I. Sinayskiy, F. Petruccione
School of Chemistry and Physics
University of KwaZulu-Natal
Durban 4001, South Africa

F. Petruccione
School of Data Science and Computational Thinking and Department of Physics
Stellenbosch University
Stellenbosch 7604, South Africa

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/qute.202300249

DOI: 10.1002/qute.202300249

Other methods avoid the heuristic strategies by training parametrized circuits, with the aid of a classical subroutine, to map the classical data to separate regions in Hilbert space. In some approaches, these regions are centred around label states; a set of quantum states corresponding to the set of classes.[11,13] In other approaches, these regions or clusters are constructed, through training, to be as far apart as possible in Hilbert space.[10] In both these approaches, the choice of label states is inherently tied to the data encoding strategy and vice versa. If, as proposed, orthogonal states in Hilbert space are chosen as the label states then the data encoding strategy will have to utilize a sufficient number of qubits to accommodate the orthogonal states. Alternatively, if many qubits are used for the embedding, then clusters need to be formed around label states in large Hilbert spaces. In this case, the task of finding similarities between the data in these vast spaces may not be easy.[17]

This work presents the first multi-class SWAP-Test classifier. Our work is inspired by the binary SWAP-Test classifier and the use of label states in previous work. With this multi-class quantum classifier, we are able to avoid the cost of constructing multiple binary classifiers to perform multi-class classification. The use of only single qubit label states, regardless of the data encoding strategy and the number of classes, is novel to this work. Furthermore, the label states are stored separately from the data, ensuring that the choice of label states remains independent from the data encoding strategy. This has the advantage that the number of qubits required and the topology of the circuits used need not change with the number of classes. Importantly, the state reconstruction of only a single qubit is performed regardless of the number of classes or the data encoding strategy. This ensures that the measurement strategy is also independent from the number of classes. In contrast to other architectures for multi-class quantum classifiers,[14] the state reconstruction of a single qubit yields all the necessary information for multi-class classification tasks.

Given some multi-class classification problem, the classifier is realised by first preparing a quantum state encoding the test datum, the training data and their respective labels in a specific format. The label states are chosen such that their corresponding Bloch vectors, which we refer to as label vectors, are maximally separate on the Bloch sphere. A modified SWAP-Test, involving a state reconstruction of the qubit storing the label states, is performed on the prepared state. This effectively yields a linear combination of label vectors. The contribution of each label vector is a weighted sum of kernel values between the test data and all the training data with that label. The kernel values in this sum are computed in parallel, just as with the binary SWAP-Test classifier. The type of kernel that is evaluated can be tailored to suit the classification problem by changing the data encoding strategy. Finally, the overlap between the resulting vector and each label vector is evaluated classically. The test datum is assigned the label whose label vector achieves the highest overlap. Like the binary SWAP-Test classifier, this multi-class SWAP-Test classifier does not inherently rely on a classical subroutine for training.

The effectiveness of the multi-class SWAP-Test classifier is demonstrated by applying it to a number of different classification problems, each with a different dimension of features and a different number of classes. The datasets used are generated XOR datasets as well as real-world datasets including Iris (3 classes), Wine (3 classes), Digits (10 classes) and Letter Recognition (12 classes) datasets.[19] We show analytically that the multi-class classifier achieves high accuracies on these real-world datasets. Even without training and with only standard data encoding strategies, these experiments show that the multi-class SWAP-Test classifier is remarkably powerful.

The performance of the multi-class SWAP-Test classifier is then considered under realistic conditions. To do this, we demonstrate the robustness of the classifier to finite sampling and noise. Through variance analysis, we show that the number of label states that can be accurately distinguished on a single qubit grows linearly with the number of repetitions of the required measurements. We also show that, under certain depolarizing noise conditions, the classification process remains unaffected. These theoretical results are demonstrated by numerical experiments that incorporate depolarizing noise and finite sampling.

The paper is organized as follows: Section 2 outlines the steps that constitute the proposed multi-class SWAP-Test classifier, discusses its robustness to noise, and assesses the number of label states that can be stored with a single qubit with this classifier. Section 3 presents the experimental set-up and results of experiments conducted on various datasets. Lastly, Section 4 draws concluding remarks, highlighting possible areas for future work.

2. Experimental Section

2.1. Classification with the Multi-Class SWAP-Test Classifier

Classification is a fundamental problem in machine learning. Given a dataset

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M} \tag{1}$$

consisting of training data $\mathbf{x}_i \in \mathbb{R}^N$ and their respective labels $y_i \in \{1, .., L\}$, the goal of supervised classification is to develop a model for classifying unlabelled data. The algorithms for developing these models are called classifiers. This section describes the steps that constitute the multi-class SWAP-Test classifier. These steps are also outlined in **Figure 1**.

For some unlabelled test datum $\tilde{\mathbf{x}}$, the multi-class SWAP-Test classifier requires the test datum, the training data, and their respective labels to be encoded in a quantum state as

$$|\Psi_i\rangle = \sum_{m=1}^{M} \sqrt{w_m}|0\rangle|\tilde{\mathbf{x}}\rangle|\mathbf{x}_m\rangle|y_m\rangle|m\rangle \tag{2}$$

The first qubit is initialized in the ground state $|0\rangle$ and will later be used as the ancilla in a modified SWAP-Test. The index register $\sum_{m=1}^{M} \sqrt{w_m}\,|m\rangle$ with $\sum_{m=1}^{M} w_m = 1$ is used to link the training data to their respective labels. The test and training data are encoded in separate registers. Some unitary operator $\mathcal{U}_{\Phi(\mathbf{x})}$ encodes the test and training data into quantum states as $|\tilde{\mathbf{x}}\rangle = \mathcal{U}_{\Phi(\tilde{\mathbf{x}})}|0\rangle$ and $|\mathbf{x}_m\rangle = \mathcal{U}_{\phi(\mathbf{x}_m)}|0\rangle$, respectively. This operation can be understood as applying a feature map to the classical data.[2] The choice of operator, corresponding to a choice in encoding method, decides the type of kernel that will be evaluated by the classifier.[20,21] Each y_m is mapped to a unique single qubit state $|y_m\rangle$, which we

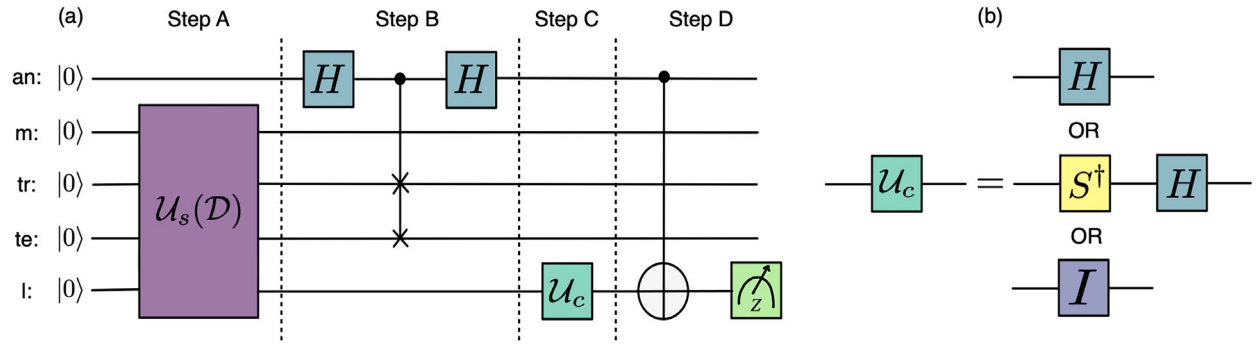

Figure 1. a) The circuit required for the multi-class SWAP-Test classifier. The first register (an) stores the ancilla. The second register stores the index register (m) which links the training data in the training register (tr) to their respective label states on the label qubit (l). The test data is stored in (te). To perform a state tomography of (l) at the end of the circuit, three circuits performing Steps A and B will be prepared. In each of these circuits, Step A applies $\mathcal{U}_s(\mathcal{D})$, which prepares the test data, training data, and training labels in a quantum state $|\Psi_i\rangle$, given in Equation (2). Step B then swaps the registers containing the test and training data. In each circuit, Step C applies one of the gate sequences in (b) to perform a change of basis to the X-basis, Y-basis or maintain the Z-basis. The three circuits evaluate the predicted vector $\mathbf{y}_{pred}$, which is then used in an assignment function to classify the test data.

refer to as the label state of γ_m. Each label state

$$|\gamma_m\rangle = \cos\left(\frac{\theta_{\gamma_m}}{2}\right)|0\rangle + e^{i\phi_{\gamma_m}}\sin\left(\frac{\theta_{\gamma_m}}{2}\right)|1\rangle \tag{3}$$

with $0 \leq \theta_{\gamma_m} \leq \pi$ and $0 \leq \phi_{\gamma_m} \leq 2\pi$ can be represented as a Bloch vector

$$\mathbf{y}_m = \begin{pmatrix} \cos\phi_{\gamma_m}\sin\theta_{\gamma_m} \\ \sin\phi_{\gamma_m}\sin\theta_{\gamma_m} \\ \cos\theta_{\gamma_m} \end{pmatrix} \tag{4}$$

which will be referred to as the label vector of γ_m. The label states are chosen such that their Bloch vectors are as far apart as possible on the Bloch sphere. For this classifier, we propose that the optimal placement of Bloch vectors be identified as solutions to the Tammes problem: the problem of placing T points on a unit sphere so that the two closest points are as far apart as possible.[22] This allows maximal separation of vectors around the Bloch sphere. Some of these placements, for $T = 2, 3$ and 4, are shown in **Figure 2**. For an L-class classification problem, we use the co-ordinates of the points defined as the solutions to the Tammes problem for L points. These co-ordinates define the label vectors, which are then used to obtain the label states.

Following the required state preparation, a modified SWAP-Test is performed on the test and training data registers. This modified SWAP-Test is shown in Steps B-D of Figure 1. A `C-SWAP` gate conditioned on the ancilla, in between two Hadamard gates applied to the ancilla, swaps the two registers. The result is

$$\begin{aligned} |\Psi_f\rangle &= H_a.\texttt{C-SWAP}.H_a|\Psi_i\rangle \\ &= \sum_{m=1}^{M}\frac{\sqrt{w_m}}{2}(|0\rangle|\psi_+\rangle + |1\rangle|\psi_-\rangle)|\gamma_m\rangle|m\rangle \end{aligned} \tag{5}$$

with $|\psi_\pm\rangle = |\tilde{\mathbf{x}}\rangle|\mathbf{x}_m\rangle \pm |\mathbf{x}_m\rangle|\tilde{\mathbf{x}}\rangle$. The subscript a denotes that the Hadamard gate H is applied to the ancilla.

In order to perform the necessary state reconstruction, the above state needs to be prepared three times. Then, the required change of basis on each of the three states may be performed:

$$\begin{aligned} |\Psi_{fx}\rangle &= H_l|\Psi_f\rangle \\ |\Psi_{fy}\rangle &= H_l S_l^\dagger|\Psi_f\rangle \\ |\Psi_{fz}\rangle &= |\Psi_f\rangle \end{aligned} \tag{6}$$

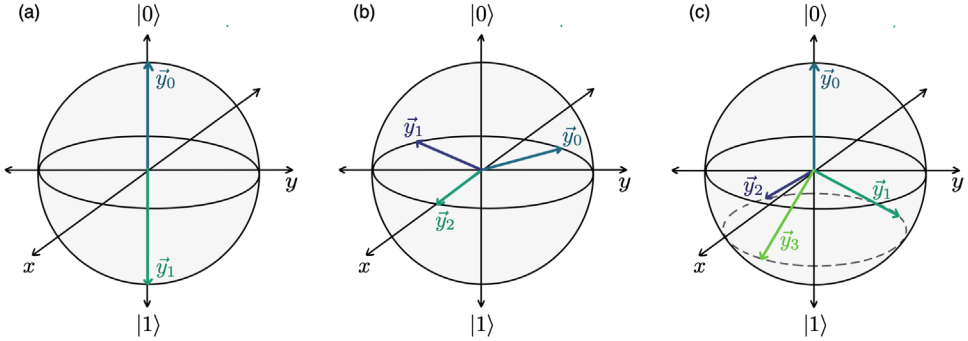

Figure 2. The optimal choice of label vectors for a) 2 classes $\{\gamma_0 : [0, 0, 1], \gamma_1 : [0, 0, -1]\}$, b) 3 classes $\{\gamma_0 : [-0.5, 0.866, 0], \gamma_1 : [-0.5, -0.866, 0], \gamma_2 : [1, 0, 0][-0.5, -0.866, 0]\}$ and c) 4 classes $\{\gamma_0 : [0, 0, 1], \gamma_1 : [-0.471, 0.861, -0.333], \gamma_2 : [-0.471, -0.861, -0.333], \gamma_3 : [0.943, 0, -0.333]\}$. These vectors point to solutions of the Tammes problem.

here, the subscript l denotes that the gates H and $S^\dagger$ are applied to the label qubit.

Before we perform any measurement on the states, a C-NOT operation controlled on the ancilla and targeted on the label qubit is applied to each state:

$$|\tilde{\Psi}_{fx}\rangle = \texttt{c-not}_{a,l}|\Psi_{fx}\rangle$$
$$|\tilde{\Psi}_{fy}\rangle = \texttt{c-not}_{a,l}|\Psi_{fy}\rangle \quad (7)$$
$$|\tilde{\Psi}_{fz}\rangle = \texttt{c-not}_{a,l}|\Psi_{fz}\rangle$$

This converts what would be a two qubit measurement to a single qubit measurement in each case.[23] Finally, the measurement of a single qubit observable $\langle\sigma_z^l\rangle$, where the superscript l indicates that the operator acts only on the label qubit, is performed on each state $\rho_{fs} = |\tilde{\Psi}_{fs}\rangle\langle\tilde{\Psi}_{fs}|$ for $s \in \{x, y, z\}$. Each measurement will be an estimation of

$$\begin{aligned}\langle\sigma_z^l\rangle &= \mathrm{Tr}(\sigma_z^l \rho_{fs}) \\ &= \langle\tilde{\Psi}_{fs}|\sigma_z^l|\tilde{\Psi}_{fs}\rangle\end{aligned} \quad (8)$$

The results of these measurements are used to construct a vector, which we refer to as the predicted vector $\mathbf{y}_{pred}$

$$\mathbf{y}_{pred} = \begin{pmatrix} \mathrm{Tr}\left(\sigma_z^{(l)}\rho_{fx}\right) \\ \mathrm{Tr}\left(\sigma_z^{(l)}\rho_{fy}\right) \\ \mathrm{Tr}\left(\sigma_z^{(l)}\rho_{fz}\right) \end{pmatrix} \quad (9)$$

For each $s \in \{x, y, z\}$, we can use the final state $|\tilde{\Psi}_{fs}\rangle$ to derive an analytical expression for $\mathrm{Tr}(\sigma_z^l \rho_{fs})$

$$\mathbf{y}_{pred} = \begin{pmatrix} \sum_m \mathrm{w}_m |\langle\tilde{\mathbf{x}}|\mathbf{x}_m\rangle|^2 \cos\phi_{y_m} \sin\theta_{y_m} \\ \sum_m \mathrm{w}_m |\langle\tilde{\mathbf{x}}|\mathbf{x}_m\rangle|^2 \sin\phi_{y_m} \sin\theta_{y_m} \\ \sum_m \mathrm{w}_m |\langle\tilde{\mathbf{x}}|\mathbf{x}_m\rangle|^2 \cos\theta_{y_m} \end{pmatrix} \quad (10)$$

At first, the significance of $\mathbf{y}_{pred}$ may not seem clear. However, we can see that in each element of $\mathbf{y}_{pred}$, w_m weights the contribution of the fidelities $|\langle\tilde{\mathbf{x}}|\mathbf{x}_m\rangle|^2$. These fidelities actually represent a valid kernel $k(\tilde{\mathbf{x}}, \mathbf{x}_m) = |\langle\tilde{\mathbf{x}}|\mathbf{x}_m\rangle|^2$. We can restrict the sums in each element of $\mathbf{y}_{pred}$ to reveal the contribution of the kernel values in each class.

$$\mathbf{y}_{pred} = \begin{pmatrix} \sum_{i=1}^{L} \sum_{m|y_m=i} \mathrm{w}_m k(\tilde{\mathbf{x}}, \mathbf{x}_m) \cos\phi_{y_m} \sin\theta_{y_m} \\ \sum_{i=1}^{L} \sum_{m|y_m=i} \mathrm{w}_m k(\tilde{\mathbf{x}}, \mathbf{x}_m) \sin\phi_{y_m} \sin\theta_{y_m} \\ \sum_{i=1}^{L} \sum_{m|y_m=i} \mathrm{w}_m k(\tilde{\mathbf{x}}, \mathbf{x}_m) \cos\theta_{y_m} \end{pmatrix} \quad (11)$$

Then, if we let $\alpha_i = \sum_{m|y_m=i} \mathrm{w}_m k(\tilde{\mathbf{x}}, \mathbf{x}_m)$ the predicted vector may be expressed as

$$\mathbf{y}_{pred} = \sum_{i=1}^{L} \alpha_i \begin{pmatrix} \cos\phi_{y_i} \sin\theta_{y_i} \\ \sin\phi_{y_i} \sin\theta_{y_i} \\ \cos\theta_{y_i} \end{pmatrix} \quad (12)$$

or equivalently,

$$\mathbf{y}_{pred} = \sum_{i=1}^{L} \alpha_i \mathbf{y}_i \quad (13)$$

where each $\mathbf{y}_i$ is the label vector of y_i.

Now, it is apparent that the predicted vector is a linear combination of the label vectors. The weight of each label vector, α_i, is the sum of the kernel values between the test data and the training data that have that label. A high α_i increases the overlap between the $\mathbf{y}_{pred}$ and $\mathbf{y}_i$ and indicates a high similarity between the test datum and the training data belonging to class y_i.

The predicted vector that has been estimated by the quantum circuits is then used in the following assignment function

$$\tilde{y} = \max_{y_i}\{\mathbf{y_i} \cdot \mathbf{y}_{pred}\} \quad (14)$$

Effectively, the test datum is assigned the class y_i when the inner product between the label vector $\mathbf{y_i}$ and the predicted vector $\mathbf{y}_{pred}$ is the highest. This indicates that $\mathbf{y_i}$ and $\mathbf{y}_{pred}$ have the highest overlap. **Figure 3** shows a few examples of predicted vectors that could be obtained by the classifier and how the test datum that generated the predicted vector would be classified.

Interestingly, this multi-class SWAP-Test classifier reduces to the binary SWAP-Test classifier when $|0\rangle$ and $|1\rangle$ are chosen as label states for binary classification. In fact, this result applies when any two label states $|y_0\rangle$, $|y_1\rangle$ corresponding to label vectors $\mathbf{y_0}$, $\mathbf{y_1}$ that satisfy $\mathbf{y_0} \cdot \mathbf{y_1} = -1$ and $\mathbf{y_i} \cdot \mathbf{y_i} = 1$, $i = 0, 1$ are chosen. A proof of this can be found in Note SI (Supporting Information).

2.2. Analysis of Multi-Class SWAP-Test Classifier

Even though great strides have been made in the development of quantum hardware, two sources of error are inevitably encountered when executing quantum algorithms on real devices: noise and finite sampling. In this section, we discuss the robustness of the multi-class SWAP-Test classifier to noise. We also describe the relationship between the number of label states that can be accurately distinguished on a single qubit as well as how this number is affected by noise.

2.2.1. Robustness to Noise

Depolarizing noise is a simple yet important noise model for quantum systems.[24] It describes the loss of information about a quantum system. Depolarizing noise is commonly encountered on currently available quantum devices, making it a practically relevant noise model. It also subsumes other noise models, namely the bit-flip and phase-flip noise models. For these reasons, depolarizing noise is a popular choice for analysis of the effects of noise on quantum algorithms.

Noise models in quantum computing are described by quantum channels. The depolarising noise model is given by a quantum channel

$$\mathcal{E}_d(\rho) = \frac{pI}{2} + (1 - p)\rho \quad (15)$$

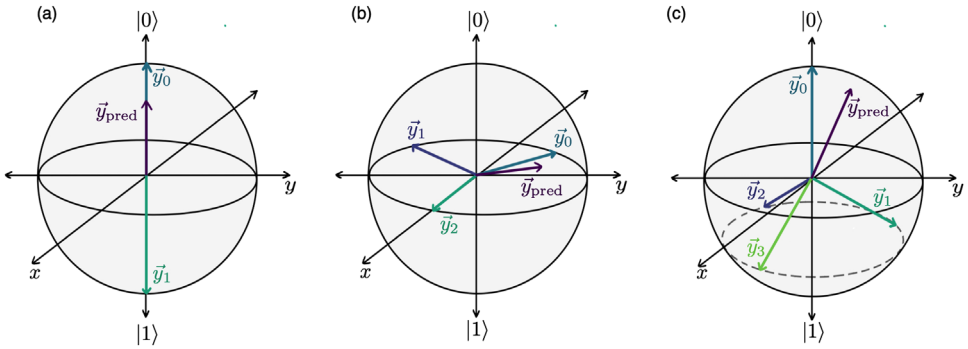

Figure 3. Illustrative predicted vectors that could be obtained by this multi-class SWAP-Test classifier for problems with a) 2 classes, b) 3 classes, and c) 4 classes. According to the assignment function given in Equation (14), the test point will be assigned to the class y_i when the inner product between its label vector y_i and the obtained predicted vector is highest. According to the definition of the inner product, the inner product will be its highest when the angle between the vectors is the smallest. In each of these diagrams, the test point will be assigned to class y_0. *Note*: these predicted vectors are not obtained from any concrete classification problem.

where $p \in [0, 1]$ is the depolarization parameter p corresponding to the 'degree' of depolarizing noise. A depolarization parameter of $p = 0$ indicates no depolarizing noise and while $p = 1$ indicates a complete loss of information about the system.

When analysing the effect of depolarizing noise on the outcome of the classifier, we consider the same conditions used in ref. [5] to demonstrate the robustness of the binary SWAP-Test classifier to depolarizing noise. We consider the single qubit depolarising channel in Equation (15) but in Kraus form,[25]

$$\mathcal{E}_d(\rho) = \sum_{k=1}^{4} E_k^l \; \rho \; \left(E_k^l\right)^\dagger \tag{16}$$

where the set of Kraus operators are

$$\left\{ E_1 = \sqrt{1 - \frac{3p}{4}}\, I,\; E_2 = \sqrt{\frac{p}{4}}\sigma_x,\; E_3 = \sqrt{\frac{p}{4}}\sigma_y,\; E_4 = \sqrt{\frac{p}{4}}\sigma_z \right\} \tag{17}$$

and the superscript l indicates that the Kraus operator E_k acts only on the label qubit. The Kraus operators also satisfy the completeness relation $\sum_{k=1}^{4} (E_k^l)^\dagger E_k^l = \mathbb{I}_2$.

We then evaluate the effect of this channel acting only on the label qubit right before the measurement, that is, on label qubit of the states ρ_{fx}, ρ_{fy} and ρ_{fx}. The effect of the channel on ρ_{fz} is

$$\mathcal{E}_d(\rho_{fx}) = \sum_{k=1}^{4} E_k^l \rho_{fx} (E_k^l)^\dagger \tag{18}$$

The total effect on the outcome of the measurement is

$$\begin{aligned} \text{Tr}\left(\sigma_z^{(l)} \mathcal{E}_d(\rho_{fx})\right) &= \text{Tr}\left(\mathcal{E}_d(\rho_{fx})\right)\sigma_z^{(l)}) \\ &= \text{Tr}\left(\sum_{k=1}^{4} E_k^l \rho_{fx} \left(E_k^l\right)^\dagger \sigma_z^{(l)}\right) \\ &= \sum_{k=1}^{4} \text{Tr}\left(E_k^l \rho_{fx} \left(E_k^l\right)^\dagger \sigma_z^{(l)}\right) \end{aligned} \tag{19}$$

where we have used the linearity of the trace. Then, we can use the cyclic property of the trace, and the knowledge that the Kraus operators are Hermitian $E_k^l = (E_k^l)^\dagger$ to obtain

$$\sum_{k=1}^{4} \text{Tr}\left(E_k^l \rho_{fx} \left(E_k^l\right)^\dagger \sigma_z^{(l)}\right) = \sum_{k=1}^{4} \text{Tr}\left(\rho_{fx} E_k^l \sigma_z^{(l)} E_k^l\right)$$

Once again using the linearity of the trace, we obtain

$$\begin{aligned} \sum_{k=1}^{4} \text{Tr}\left(\rho_{fx} E_k^l \sigma_z^{(l)} E_k^l\right) &= \text{Tr}\left(\rho_{fx} \sum_{k=1}^{4} E_k^l \sigma_z^{(l)} E_k^l\right) \\ &= (1-p)\text{Tr}\left(\sigma_z^{(l)} \rho_{fx}\right) \end{aligned} \tag{21}$$

since $\sum_{k=1}^{4} E_k^l \sigma_z^{(l)} E_k^l = (1-p)$.

Similarly, the effect of the channel on ρ_{fy} and ρ_{fz} is

$$\begin{aligned} \text{Tr}\left(\sigma_z^{(l)} \mathcal{E}_d(\rho_{fy})\right) &= (1-p)\text{Tr}\left(\sigma_z^{(l)} \rho_{fy}\right) \\ \text{Tr}\left(\sigma_z^{(l)} \mathcal{E}_d(\rho_{fz})\right) &= (1-p)\text{Tr}\left(\sigma_z^{(l)} \rho_{fz}\right) \end{aligned} \tag{22}$$

These results can be used to construct a vector

$$(1-p)\begin{pmatrix} \text{Tr}\left(\sigma_z^{(l)} \rho_{fx}\right) \\ \text{Tr}\left(\sigma_z^{(l)} \rho_{fy}\right) \\ \text{Tr}\left(\sigma_z^{(l)} \rho_{fz}\right) \end{pmatrix} = (1-p)\mathbf{y}_{pred} \tag{23}$$

This shows that, under the depolarizing noise conditions considered, the predicted vector is only scaled by a factor of $(1-p)$. An example of the effect of scaling the predicted is demonstrated in **Figure 4**.

We can rewrite the assignment function in Equation (14) as

$$\begin{aligned} \tilde{y} &= \max_{y_i}\{\|\mathbf{y}_i\|\|\mathbf{y}_{pred}\|\cos\lambda_i\} \\ &= \max_{y_i}\{\cos\lambda_i\} \end{aligned} \tag{24}$$

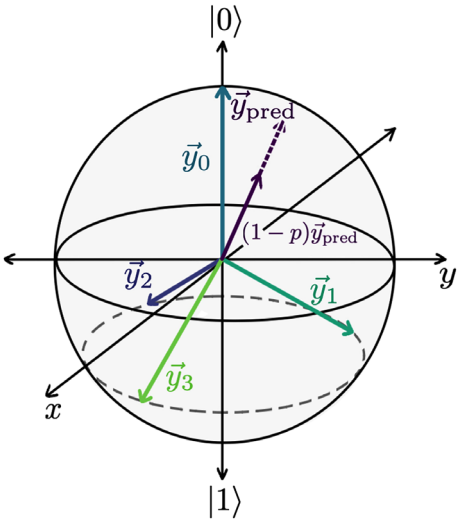

Figure 4. An illustrative example of how the outlined conditions of depolarizing noise would affect the predicted vector. The predicted vector will be scaled by a factor of $(1-p)$. This has no impact on the angle between the predicted vector and the label vectors so this will have no impact on the classification outcome. *Note*: this example is not related to any concrete classification problem.

where we can remove $\|\mathbf{y}_i\|$ since $\|\mathbf{y}_i\| = 1$ for all i and we can remove $\|\mathbf{y}_{pred}\|$ since it is constant for each y_i. Then, we can see that the assignment function depends only on the angle λ_i between the i^{th} label vector $\mathbf{y}_i$ and the predicted vector $\mathbf{y}_{pred}$. It then becomes apparent that scaling the predicted vector by a factor of $(1-p)$ has no effect on the assignment function. In this way, the outcome of the classifier is unaffected by these conditions of depolarizing noise.

In this paper, we have restricted the analysis to the effects of depolarizing noise. In Note SII (Supporting Information), we perform a similar analysis with a single qubit Pauli channel; the most general single qubit noise model. We show that this channel also only scales the predicted vector. In this way, the outcome of the classifier is unaffected by the single qubit Pauli channel, making it robust to special cases of the single qubit Pauli channel including bit-flip and phase-flip noise.

2.2.2. Number of Label States

Our next consideration involves the number of label states that can be stored on a single qubit. Due to the design of the classifier, this number depends on our ability to accurately measure the predicted vector. The number of predicted vectors we can distinguish on the Bloch sphere will be the number of label states that we should store on the single label qubit.

In order to determine the number of predicted vectors that we can distinguish, we first estimate the standard error Δ in the required measurements. For brevity, we make the following assignments: $x = \text{Tr}(\sigma_z^{(l)}\rho_{fx})$, $y = \text{Tr}(\sigma_z^{(l)}\rho_{fy})$ and $z = \text{Tr}(\sigma_z^{(l)}\rho_{fz})$. Then, the standard error for $s \in \{x, y, z\}$ is

$$\Delta s = \sqrt{\frac{4\nu_s(1-\nu_s)}{R}} \tag{25}$$

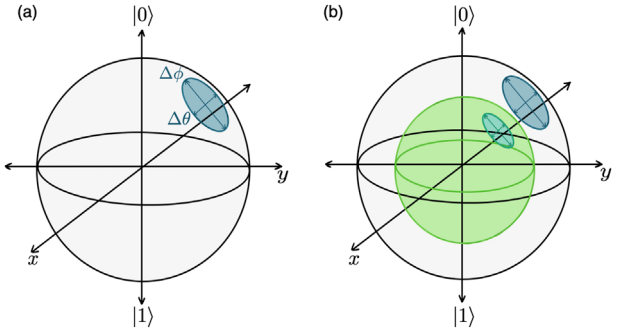

Figure 5. Ellipsoids representing the standard error in the measurement of a predicted vector. a) shows the ellipsoid on the Bloch sphere while b) shows the ellipsoid on a shrunken Bloch sphere that can be expected under certain depolarizing noise conditions. In each case, the number of label states that should be stored on the label qubit can be calculated by dividing the area of the sphere by the area of the ellipsoid.

where we have defined $\nu_s = \frac{1}{2}(\text{Tr}(\sigma_z^{(l)}\rho_{fs}) + 1)$ for simplicity and R is the number of repetitions of the required measurements.

Using the standard error propagation formula, shown in Note SIII (Supporting Information), we estimate $\Delta\theta$ and $\Delta\phi$. This allows us to estimate the area of the ellipse that represents the uncertainty in the measurement of the predicted vector. An illustration of this ellipse is in **Figure 5**. By dividing the surface area of the Bloch sphere by the area of this ellipse, we find the number of label states (N_s) that can be stored:

$$N_s = \frac{\text{Area of Bloch Sphere}}{\text{Area of Ellipsoid}} = \frac{4\pi||\vec{r}||^2}{\pi\Delta\theta\Delta\phi} \tag{26}$$

where $\vec{r} = (x, y, z)$.

After making the necessary substitutions, we find that

$$N_s = \mathcal{O}(R) \tag{27}$$

This means that the number of label states that can be stored grows linearly with the number of repetitions of the required measurements.

If we consider the effect of depolarising noise on this number, then the number of label states that can be stored under these conditions ($\tilde{N}_s$) is:

$$\tilde{N}_s = \frac{4\pi||(1-p)\vec{r}||^2}{\text{Area of Ellipsoid}} = \frac{4\pi(1-p)^2||\vec{r}||^2}{A(p)} \tag{28}$$

where the Area of Ellipsoid is a function $A(p)$ that is now dependent on the depolarising parameter p.

To perform further analysis, we perform a Taylor expansion of $A(p)$ about $(p = 0)$, up to second order, reveals:

$$A(p) = A_0 + A_1 p + A_2 p^2 + \mathcal{O}(p^3) \tag{29}$$

where $A_1 = \frac{\partial A(p)}{\partial p}|_{p=0}$ and $A_2 = \frac{1}{2}\frac{\partial^2 A(p)}{\partial p^2}|_{p=0}$.

By substituting this expansion into Equation (28), we obtain an expression for the number of label states that accounts for depolarizing noise:

$$\tilde{N}_s = \frac{4\pi(1-p)^2||\vec{r}||^2}{A(p)} \simeq \frac{4\pi(1-p)^2||\vec{r}||^2}{\frac{1}{A_0}(1+\frac{A_1}{A_0}p+\frac{A_2}{A_0}p^2)} \tag{30}$$

Then, if we make use of the fractional binomial theorem to approximate $(1+\frac{A_1}{A_0}p+\frac{A_2}{A_0}p^2)^{-1}$ we obtain

$$\frac{4\pi(1-p)^2||\vec{r}||^2}{\frac{1}{A_0}(1+\frac{A_1}{A_0}p+\frac{A_2}{A_0}p^2)} \simeq \frac{4\pi||\vec{r}||^2}{A_0}(1-2p+p^2)\left[1-\frac{A_1}{A_0}p-(\frac{A_2}{A_0}-\frac{A_1^2}{A_0^2}p^2)\right] \simeq N_s\left[1-\left(\frac{A_1}{A_0}+2\right)p-\left(\frac{A_2}{A_0}-\frac{A_1^2}{A_0^2}-2\frac{A_1}{A_0}-1\right)p^2\right] \tag{31}$$

We perform a numerical search on the domain $-1 \leq x, y, z \leq 1$ and $\sqrt{x^2+y^2+z^2} \leq 1$ for the maximum values of $(\frac{A_1}{A_0}+2)$ and $(\frac{A_2}{A_0}-\frac{A_1^2}{A_0^2}-2\frac{A_1}{A_0}-1)$. This search yields

$$\left(\frac{A_1}{A_0}+2\right) \leq 5 \tag{32}$$

$$\left(\frac{A_2}{A_0}-\frac{A_1^2}{A_0^2}-2\frac{A_1}{A_0}-1\right) \leq -6 \tag{33}$$

This reveals that in the worst case:

$$\tilde{N}_s \approx N_s(1-5p+6p^2) \tag{34}$$

These conditions of depolarizing noise decrease the number of label states by at most $(1-5p+6p^2)$. The depolarizing parameter that we expect in real experiments is $p << 0.1$. According to Equation (34), this will have a negligible effect on the number of label states that can be stored.

3. Results and Discussion

We demonstrate the effectiveness of the multi-class SWAP-Test classifier by applying it to several diverse multi-class classification problems.

For the numerical simulations, we apply the classifier to a 3D generated dataset with 4 classes; 16 data points in each class. This dataset, shown in **Figure 6**, will be denoted by `4-XOR` since it is inspired by the 2D XOR dataset. Like the 2D XOR dataset, `4-XOR` is not linearly separable.

To evaluate the performance of the multi-class SWAP-Test classifier on this dataset, we perform leave-one-out cross validation. For each data point, we numerically simulate the circuits that construct the predicted vector using Qiskit's Statevector Simulator.[26] Note SIV (Supporing Information) provides more details on the state preparation required. The predicted vector is used in the assignment function given in Equation (14) to classify the data point. We first numerically simulate the circuits ideally, with no noise or finite sampling. We then numerically simulate the circuits with finite sampling and under the depolarizing noise conditions outlined in Section 2.2.1. For the simulations with finite sampling, the circuits were executed with 8192 shots. **Figure 7** depicts the circuits that are simulated with depolarizing noise, with **Figure 8** providing a more details on the simulation of the depolarizing noise.[27]

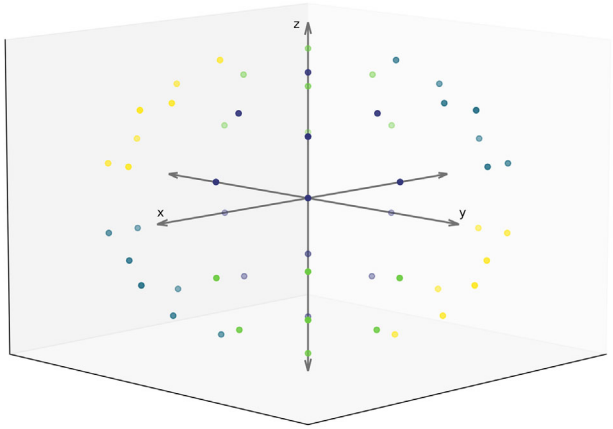

Figure 6. The `4-XOR` dataset (3 features, 4 classes and 64 points). The dataset was generated such that data points from the same class would lie directly opposite each other on a sphere so that the kernel in Equation (36) would be most effective.

When encoding one of the test or training data, we first normalize the datum so that $\|\mathbf{x}\|^2 = \sum_i |x_i|^2 = 1$. Then, we encode each component of the normalized datum in a different computational basis state of the appropriate register:

$$\mathbf{x} \rightarrow \sum_{i=1}^{2^n} x_i|i\rangle \tag{35}$$

This strategy is commonly known as amplitude encoding and it gives rise to the following quantum kernel:

$$k(\mathbf{x}, \mathbf{z}) = |\langle \mathbf{x}|\mathbf{z}\rangle|^2 \tag{36}$$

which is simply the absolute square of the linear kernel.[2] It should also be noted that uniform training weights are used, that is: $\sum_{m=1}^{M} w_m = 1,\ w_m = 1/M\ \forall\ M$.

Our simulations show that high accuracies (100%) are obtained by the classifier when executed under ideal conditions and realistic conditions with depolarizing noise and finite sampling. The results of these experiments are shown in **Table 1**. Upon analysis of the predicted vectors obtained from simulations with depolarizing noise, the result from Section 2.2.2 is confirmed. We see that the obtained predicted vectors are only scaled under the depolarising noise conditions and that this has no effect on the classification process. This is illustrated in **Figure 9**. This is also illustrated in Table 1 where we see that the average norms of the predicted vectors decreases by a factor of $(1-p)$.

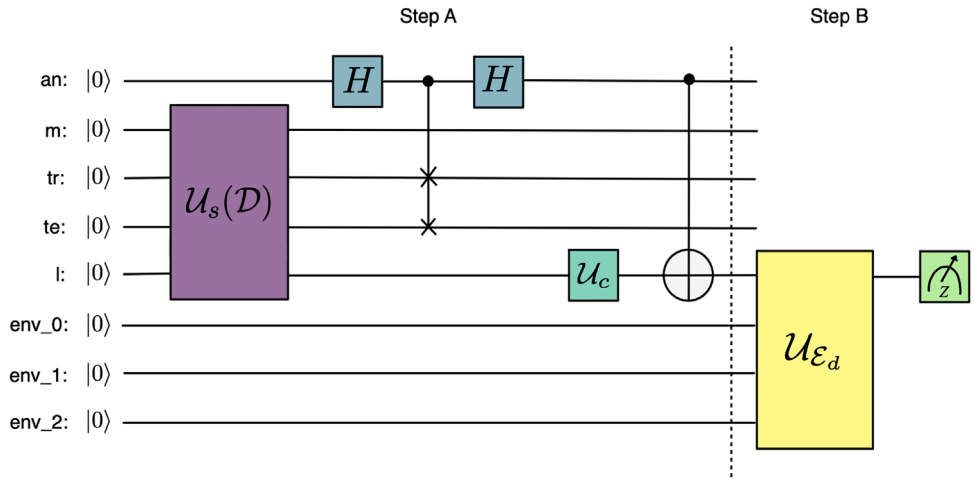

Figure 7. The circuit required to simulate the multi-class SWAP-Test classifier with depolarising noise. Step A involves the state preparation, modified SWAP-Test and change of basis. Step B involves applying $\mathcal{U}_{\mathcal{E}d}$ to simulate the depolarising noise.

We further evaluate the performance of the multi-class classifier on several other datasets using fivefold cross validation. These datasets include three generated XOR datasets, the Iris, Wine, and Digits datasets provided by scikit-learn as well as the Letter Recognition dataset provided by the UCI Machine Learning Repository.[19] It should be noted that the Wine, Digits and Letter Recognition datasets were artificially balanced. For the Wine dataset, this was done by uniformly sampling 48 data points from each class in each dataset. For the Digits dataset, 174 points from each class were uniformly sampled. Lastly, for the Letter Recognition dataset, 734 points were uniformly sampled from classes for the letters A-L.

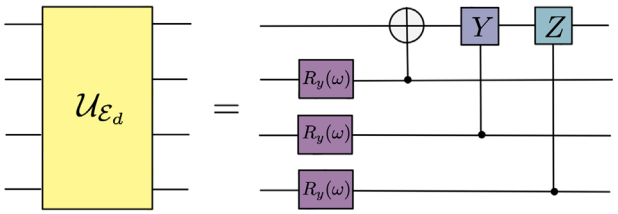

Figure 8. A more detailed look at the operator $\mathcal{U}_{\mathcal{E}d}$ that simulates the depolarising noise in the circuits. Here, R_y is a Y-rotation with $\omega = \frac{1}{2}\arccos(1-2p)$.

Table 1. The accuracies obtained by the multi-class SWAP-Test classifier when applied to the generated 4-XOR dataset. The accuracies presented are from numerical simulations with finite sampling and depolarizing noise. The average norms of the predicted vectors are from numerical simulations with just depolarizing noise, to illustrate the result from Section 2.2.1

Depolarisation Rate (p)	Accuracy (%)	Av. Norm of Predicted Vector
0	100	0.1356
0.02	100	0.1331
0.04	100	0.1302
0.06	100	0.1275
0.08	100	0.1248
0.1	100	0.1223

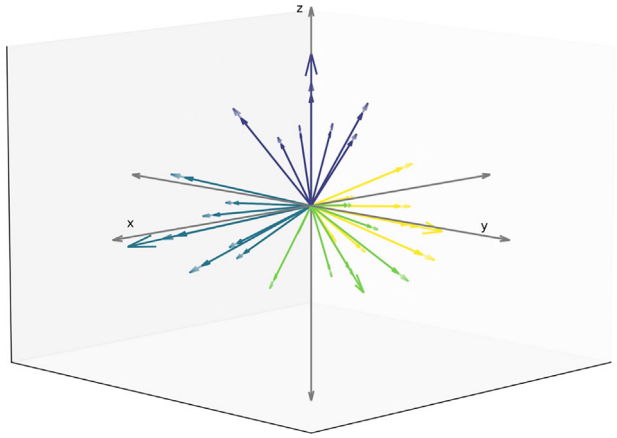

Figure 9. The predicted vectors produced through the classification of the 4-XOR dataset. Here, the darker vectors are the predicted vectors that have been evaluated with depolarizing noise ($p = 0.1$) while the lighter vectors are the predicted vectors evaluated without depolarizing noise. We can see that the predicted vectors are only scaled down but the angles between the vectors do not change.

For each test point in these datasets, we evaluate the predicted vector classically using Equation (13). To do this, we evaluate each $\alpha_i = \sum_{m|y_m=i} w_m |\langle \tilde{\mathbf{x}}|\mathbf{x}_m\rangle|^2$ directly as $\alpha_i = \sum_{m|y_m=i} w_m k(\tilde{\mathbf{x}}, \mathbf{x}_m)$. This is only possible because the methods that we would use to encode the test and training data in quantum states give rise to kernels $k(\tilde{\mathbf{x}}, \mathbf{x}_m)$ that can be evaluated classically. In some cases, we use the kernel $k(\mathbf{x}, \mathbf{z}) = |\langle \mathbf{x}|\mathbf{z}\rangle|^2$ which would arise from amplitude encoding. In other cases, we use the kernel $k(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^{n} |\cos(x_k - z_k)|^2$ which would arise from angle encoding. Once the predicted vector is constructed, it is then used in the assignment function given in Equation (14) to classify the test point. The results of these experiments can be seen in **Table 2**. It can be seen that the accuracies are high, with accuracies greater than or equal to 90% being obtained for each dataset.

Table 2. The accuracies obtained by the multi-class SWAP-Test classifier when applied to various datasets.

Dataset	# Classes	# Features	# Points	Encoding	Accuracy (%)
XOR	2	2	100	Amplitude	100
	4	3	200	Amplitude	99
	8	4	400	Amplitude	99
Iris	3	4	150	Angle	95
Wine	3	13	144	Angle	92
Digits	10	64	1740	Angle	92
Letter Recognition	12	16	8808	Angle	90

4. Conclusion

We present a multi-class quantum classifier comprising of a set of quantum circuits and an assignment function that is evaluated classically. The quantum circuits estimate a weighted sum of kernel values between the test and training data and the assignment function allows us to meaningfully interpret this sum. Analytically and through experiments, the multi-class SWAP-Test classifier has been shown to be powerful and robust to noise. It has also been shown that the classifier can handle $\mathcal{O}(R)$ classes, where R is the number of executions of the required circuits. In our numerical experiment, a standard encoding strategy has been employed and has proven effective on the generated dataset. It should be noted that for future work, encoding strategies that give rise to kernels that are classically intractable to simulate[3] can be chosen. Like the binary SWAP-Test classifier, this classifier can also raise the evaluated kernels to the power of n at the cost of n copies of the test and training data. Our numerical experiments also utilized uniform training weights but future work could develop a strategy that utilizes trainable or non-uniform training weights. Also of interest for future work would be the use of real quantum devices to solve multi-class classification problems with this classifier.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work is based upon research supported by the National Research Foundation of the Republic of South Africa. Support from the NICIS (National Integrated Cyber Infrastructure System) e-research grant QICSA is kindly acknowledged. Support from the CSIR DSI-Interbursary Support (IBS) Programme is gratefully acknowledged. Support from the Center of Artificial Intelligence Research is appreciated. The authors acknowledged the use of IBM Quantum for the use of their simulators. The views expressed were those of the authors and did not reflect the official policy or position of IBM or the IBM Quantum team. The authors would like to thank Mr A. W. Pillay and Mr I. J. David for their assistance in proof reading the manuscript.

Conflict of Interest

Francesco Petruccione the Chair of the Scientific Board and Co-Founder of QUNOVA computing. The authors declare no other conflict of interest.

Author Contributions

S.M.P. designed the multi-class SWAP-Test classifier, performed the analysis on the robustness of the classifier to noise and conducted the analytical and numerical experiments. S.M.P. and I.S. performed the analysis on the number of label states that can be stored on a single qubit with this classifier. I.S., E.J., and F.P. supervised the research. All authors reviewed and discussed the results and contributed toward writing the manuscript.

Data Availability Statement

The XOR datasets generated during the current study are available in the Multi_Class_SWAP_Test_Classifier repository, github.com/Shivani-M-Pillay/Multi_Class_SWAP_Test_Classifier. The other datasets (Iris, Wine, Digits) are provided by scikitlearn.

Keywords

multi-class classification, quantum machine learning, quantum kernel, SWAP-test

Received: August 2, 2023
Revised: September 27, 2023
Published online: November 22, 2023

[1] P. Rebentrost, M. Mohseni, S. Lloyd, *Phys. Rev. Lett.* **2014**, *113*, 130503.
[2] M. Schuld, N. Killoran, *Phys. Rev. Lett.* **2019**, *122*, 040504.
[3] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, J. M. Gambetta, *Nature* **2019**, *567*, 209.
[4] C. Blank, D. K. Park, J.-K. K. Rhee, F. Petruccione, *Npj Quantum Inf.* **2020**, *6*, 1.
[5] D. K. Park, C. Blank, F. Petruccione, *Phys. Lett. A* **2020**, *384*, 126422.
[6] E. Alpaydin, *Introduction to machine learning*, MIT press, Cambridge **2020**.
[7] B. Liu, Z. Hao, X. Yang, *Soft Comput.* **2007**, *11*, 383.
[8] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, *Pattern Recognit.* **2011**, *44*, 1761.
[9] Y. Li, R.-G. Zhou, R. Xu, J. Luo, W. Hu, *Quantum Sci. Technol.* **2020**, *5*, 044003.
[10] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, J. I. Latorre, *Quantum* **2020**, *4*, 226.
[11] N. A. Nghiem, S. Y.-C. Chen, T.-C. Wei, *Phys. Rev. Phys.* **2021**, *3*, 033056.
[12] D. Bokhan, A. S. Mastiukova, A. S. Boev, D. N. Trubnikov, A. K. Fedorov, *Front. Phys.* **2022**, *10*, 1069985.
[13] A. Zhang, X. He, S. Zhao, *Quantum Inf. Process.* **2022**, *21*, 358.
[14] Y. Zeng, H. Wang, J. He, Q. Huang, S. Chang, *Entropy* **2022**, *24*, 394.
[15] R. Giuntini, A. C. G. Arango, H. Freytes, F. H. Holik, G. Sergioli, *Fuzzy Sets Syst.* **2023**, 108509.
[16] R. Giuntini, F. Holik, D. K. Park, H. Freytes, C. Blank, G. Sergioli, *Appl. Soft Comput.* **2023**, *134*, 109956.
[17] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, J. R. McClean, *Nat. Commun.* **2021**, *12*, 1.
[18] E. Alpaydin, C. Kaynak, *Optical Recognition of Handwritten Digits, UCI Machine Learning Repository*, **1998**, https://doi.org/10.24432/C50P49.
[19] D. Slate, *Letter Recognition*, UCI Machine Learning Repository, **1991**, https://doi.org/10.24432/C5ZP40.

[20] M. Schuld, F. Petruccione, *Supervised learning with quantum computers*, Vol. *17*, Springer, Berlin **2018**.
[21] M. Schuld, *arXiv preprint arXiv:2101.11020* **2021**.
[22] P. M. L. Tammes, *Recueil des travaux botaniques néerlandais* **1930**, *27*, 1.
[23] D. K. Park, C. Blank, F. Petruccione, in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Piscataway, NJ **2021**, pp. 1–7.
[24] M. A. Nielsen, I. Chuang, *Quantum computation and quantum information*, American Association of Physics Teachers, Maryland **2002**.
[25] K. Kraus, *Ann. Phys.* **1971**, *64*, 311.
[26] M. Treinish, J. Gambetta, S. Thomas, P. Nation, qiskit bot, P. Kassebaum, D. M. Rodríguez, S. de la Puente González, J. Lishman, S. Hu, L. Bello, J. Garrison, K. Krsulich, J. Huang, J. Yu, M. Marques, E. Arellano, J. Gacon, D. McKay, J. Gomez, L. Capelluto, Travis-S-IBM, A. Mitchell, A. Panigrahi, lerongil, R. I. Rahman, S. Wood, T. Itoko, A. Pozas-Kerstjens, C. J. Wood, Qiskit/qiskit: Qiskit 0.42.1, **2023**. https://doi.org/10.5281/zenodo.7757946
[27] G. García-Pérez, M. A. Rossi, S. Maniscalco, *Npj Quantum Inf.* **2020**, *6*, 1.