RESEARCH ARTICLE | JULY 25 2025

# Adversarial attacks on hybrid classical-quantum deep learning models for histopathological cancer detection

Biswaraj Baral ; Bhavika Bhalgamiya ; Reek Majumder ; Divya Dutta Roy ; Taposh Dutta Roy



*APL Mach. Learn.* 3, 036106 (2025)
https://doi.org/10.1063/5.0270673

# Adversarial attacks on hybrid classical-quantum deep learning models for histopathological cancer detection

Cite as: APL Mach. Learn. 3, 036106 (2025); doi: 10.1063/5.0270673
Submitted: 12 March 2025 • Accepted: 8 July 2025 •
Published Online: 25 July 2025



Biswaraj Baral,[1,a] Bhavika Bhalgamiya,[2,b] Reek Majumder,[3,c] Divya Dutta Roy,[4,d] and Taposh Dutta Roy[4,e]

## AFFILIATIONS

[1] Quantum Computing Group, Qausal AI, San Ramon, California 94583, USA
[2] Enterprise Computing Solutions, Unisys, Blue Bell, Pennsylvania 19422, USA
[3] Civil Engineering Department, Clemson University, Clemson, South Carolina 29634, USA
[4] Quantum Computing Group, Silicon Valley Quantum Computing Group, San Ramon, California 94583, USA

[a] **Author to whom correspondence should be addressed:** biswa@qausal.ai
[b] **Electronic mail:** Bhavika.bhalgamiya@unisys.com
[c] **Electronic mail:** rmajumd@clemson.edu
[d] **Electronic mail:** divya@qausal.ai
[e] **Electronic mail:** taposh.dr@gmail.com

## ABSTRACT

We analyzed the application of quantum machine learning in histopathological cancer detection under adversarial attacks, demonstrating its potential to enhance diagnostic performance in adverse circumstances. Adversarial attacks are one of the major concerns in any image classification machine learning model and are responsible for perturbing the original input images, which, therefore, results in misclassification. To encounter this problem, we first developed the hybrid quantum transfer learning model by incorporating multiple transfer learning architectures such as ResNet-18, VGG-16, Inception-v3, and AlexNet with variational quantum circuits for histopathological cancer detection. Second, we introduced white-box adversarial attacks using the Fast Gradient Sign Method (FGSM) and DeepFool and projected gradient descent (PGD) methods in this model and evaluated each model's performance against these adversarial attacks. We analyzed that the Hybrid Classical Quantum Deep Learning model (HCQ-DL) with ResNet-18 provides 78.05% accuracy compared to the Classical ResNet-18 model with the highest accuracy of 50.84% for FGSM attacks. Similarly, for DeepFool attacks, HCQ-DL with ResNet-18 performs 52.12% accurately compared to the Classical ResNet-18 model with the highest accuracy of 37.87% and for PGD attacks, HCQ-DL with ResNet-18 has 52.94% performance accuracy compared to the Classical Inception-V3 model with the highest accuracy of 32.05%. As a result, we observed that HCQ-DL models are more resilient to these adversarial attacks compared to classical deep learning models and show potential to achieve greater robustness when combined with additional defense techniques.



## I. INTRODUCTION

Predictive models face the reality of encountering various attacks from vindictive entities. Adversarial attacks are one of these attacks, which mainly target AI models such as deep learning (DL) or machine learning (ML) models. These attacks involve deliberately perturbing original input images with a carefully crafted noisy image, resulting in incorrect image classification by the model. Perturbed images are imperceptible to the human eye, but it confuses the model, leading to misclassification. As per a recent study,[1] adversarial machine learning is a critical aspect of the ML field, pressing the need for practitioners and researchers to acknowledge and address the potential threats posed by adversarial attacks to the effectiveness and trustworthiness[2] of machine learning models.

The use of machine learning in the healthcare system is increasing to make the diagnosis and decision system robust. Due to the widespread use of machine learning models in healthcare systems, such systems are at a high risk of adversarial attacks. One common impact of adversarial attacks that could be experienced in the healthcare system is misleading the insurance approval system. Insurance companies use predictive models to confirm the approval of insurance reimbursement. Fraudsters may integrate the insurance data with perturbed data and lead to false insurance claims.[3] It is crucial for next-generation DL models to mitigate these attacks to solve the image misclassification problem.

In this study, we aim to investigate the impact of adversarial attacks[4] on classical deep learning (C-DL) and hybrid classical-quantum deep learning (HCQ-DL) models. The primary goal here is to present more resilient HCQ-DL models compared to C-DL models, which can perform better during adversarial attacks in order to obtain better performance accuracy. Our HCQ-DL models are trained with quantum simulators. As a result, we provide a comparative study of C-DL and HCQ-DL models for histopathological adversarial images.

The structure of this paper is outlined as follows: some notable studies on adversarial attacks are mentioned in Sec. II; in Sec. III, we delve into the methodology of model construction, integration of QNN layers, and the generation of adversarial images employing various adversarial attack algorithms; the results obtained from our experiment on different classical and hybrid classical-quantum models are included in Sec. IV; and finally, the experiment performed in this study is summarized in conclusion Sec. V.

## II. LITERATURE REVIEW

The recent development in computing technology and the availability of superior computing power and GPU leads to the extensive use of machine learning models in different sectors. Due to the availability of a large collection of health datasets, machine learning models are widely used in health sectors such as gastroenterology, ophthalmology, pathology, and dermatology for appropriate diagnosis and decision-making. Research performed on adversaries by different researchers in various applications of machine learning proved that almost all deployed machine learning models are extremely vulnerable to adversarial attacks. Formally, the term "adversarial input" was first described in 2004 by Dalvi *et al.* when they designed the framework to defend different adversarial manipulations by spammers on spam classifiers.[5]

Finlayson *et al.*[6] used different white-box and block-box attacks to generate adversarial perturbations. The experiment was performed on three use cases of medical images classification: fundoscopy, chest x-ray, and dermoscopy. The attack success rate of up to 100% with a confidence score of 100% is achieved from the experiment. The adversarial experiment performed on a real-time smart healthcare system deteriorated the performance of the system.[7] The experiment used four different black-box and white-box attack methods to generate adversarial perturbations. There was a significant drop in classification accuracy under both targeted and untargeted attacks. The highest success rate achieved under adversarial attacks is 15.68%. The adversarial experiment performed on the ISIC dataset shows that there is a huge difference between the classification accuracy with and without adversarial perturbations.[8] Selvakkumar *et al.* used a pre-trained VGG-19 transfer learning model for binary image classification. The study used the Fast Gradient Sign Method (FGSM) algorithm for adversarial image generation, which dropped the accuracy of classification from 88% to 11%.

There are several techniques that can be used for **adversarial attacks** on machine learning models. These threat models are categorized into black-box and white-box adversarial attack methods. In white-box attack, the attacker has the information of the deployed model such as inputs, the architecture of the model, internal gradients and weights, and other parameters, while in a black-box, the attacker has no access to such parameters. Some of the most common techniques of adversarial attacks are listed in Table I.

In this experiment, we attempt to generate adversarial perturbations using some of the methods of gradient-based attacks and boundary attacks only.

**TABLE I.** Examples of commonly used adversarial attack techniques.

| Adversarial technique | Examples of attacks |
|---|---|
| Gradient-based attacks | Auto projected gradient-descent (Auto-PGD)[9]<br>Fast gradient method[10]<br>Shadow attack[11] |
| Genetic algorithms | Wasserstein attack[12]<br>Targeted universal adversarial-perturbations[13] |
| Boundary attack | Brendel–Bethge attack[14]<br>Threshold attack[15]<br>DeepFool attack[16] |
| Zeroth-order optimization | Zeroth order optimization (ZOO)[17] |
| Transferability attacks | Functionally equivalent extraction[18]<br>Copycat CNN[19]<br>KnockoffNets[20] |

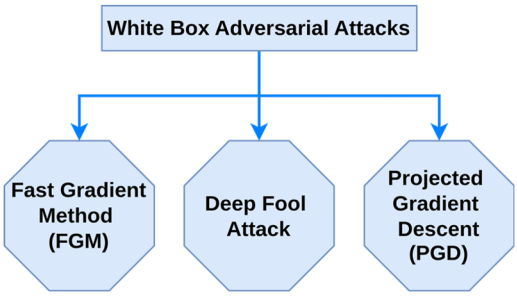


**FIG. 1.** Three different types of white box adversarial attacks analyzed in this study.

### A. Gradient-based attacks

In gradient-based attacks, an attacker computes the gradients of the model with respect to the input data and then modifies the input data to maximize the loss function. This can be achieved using techniques such as the fast gradient sign method, the projected gradient descent method, or the momentum iterative method.

### B. Boundary attack

In boundary attack-based attacks, an attacker generates a series of inputs that lie near the decision boundary of the model and then perturbs these inputs in such a way that they are misclassified by the model. This technique can be more effective than other techniques because it does not require knowledge of the model's parameters or the gradients of the model. DeepFool attack is an example of such an attack. ***"Fig. 1" represents the three white-box attacks used in this study.***

## III. METHOD/FRAMEWORK

### A. Original images and dataset preparation

A lot of research has been done on the automated classification of histopathological cancer using different datasets. For our experiment, we used the benchmark dataset known as PatchCamelyon(PCam).[21] This is a large-scale patch-level dataset derived from Camelyon16[22] data.

The aggregate of the patches makes up the slide-level image, which can be used to predict the likelihood of metastases, stage cancer. Example of patch data samples showing the likelihood of cancer is shown in "Fig. 2." The dataset contains a total of 327 680 RGB images of with resolution of 96 × 96 pixels. However, we used up to 20 000 images in this work model due to computational limitations, such as longer time required to train on quantum simulators and to generate adversarial perturbations using different algorithms. Each image in the dataset is labeled with a binary tag indicating whether there is a presence of metastatic tissue or not. To balance the dataset between both labeled classes, 10 000 images are taken from each target label. Furthermore, the dataset is split into train and test sets in the ratio of 80%–20%.

### B. Classical and hybrid classical-quantum binary image classification models

Convolutional Neural Networks (CNNs) are widely used in image-related operations due to their formidable performance. Instead of designing and training neural networks from scratch, different pre-trained transfer learning models[23] are used to enhance image classification performance. As the primary objective of this research is to evaluate the behavior of hybrid classical-quantum models within adversarial perturbations, we used simple and well-known transfer learning models such as VGG-16,[24] Inception-V3,[25] ResNet-18,[26] and AlexNet.[27] These models are trained on an ImageNet dataset with 1000 target categories. However, initial layers of pre-trained models can act as feature extraction layers for customizing image classification tasks for the newer datasets. In our experiment, we have designed the classical and hybrid classical-quantum neural network using the pre-trained transfer learning models mentioned above. These neural networks are fine-tuned by replacing the final fully connected layer with the classical (for C-DL models) or quantum layer (for HCQ-DL models) while keeping the weights of initial layers constant for feature extraction.

#### *1. Classical image classifiers*

For the classical model, "N" features are extracted from the input image using initial layers of a specific pre-trained transfer learning model. These features are inputted into the hidden layer,

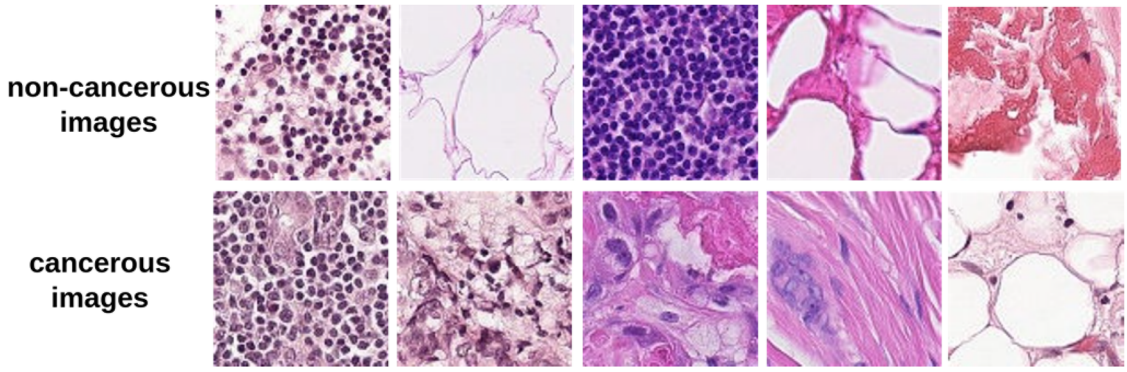


**FIG. 2.** Data samples: non-cancerous and cancerous images.

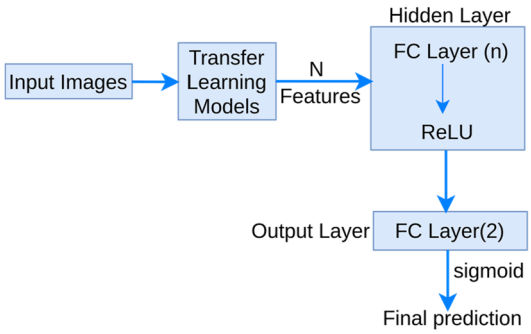


**FIG. 3.** Model architecture: classical, where "N" is the number of features extracted from transfer learning models and "n" is the number of layers in the hidden layer.

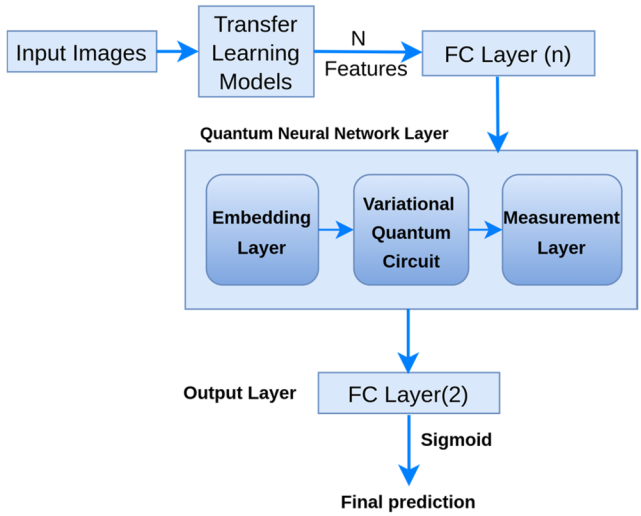


**FIG. 4.** Model architecture: hybrid classical-quantum, where "N" is the number of features extracted from transfer learning models and "n" is the number of qubits used in quantum circuit.

which consists of a fully connected layer of "n" neurons with activation functions such as ReLU, or sigmoid. "n" is chosen the same as the no. of qubits used in similar architecture of the classical-quantum hybrid model. Finally, an output layer is introduced with neurons equal to the number of target classes in our study, i.e., 2. This provides us with comparable architecture with classical-quantum models since we introduce our quantum layers between the fully connected layers of the hidden layer and output layer. The illustration in "Fig. 3" depicts the model architecture for the classical transfer learning-based classification model. To ensure a valid comparison between the performance of hybrid models and classical models, we also constructed a classical bottle-neck model with fewer parameters than the transfer learning architectures considered in this work. The classical bottle-neck model is implemented using custom CNN-based architecture consisting of three convolutional layers to perform feature extraction, followed by a flattening operation and a fully connected bottleneck consisting two layers (256 and 64 units respectively) prior to the output layer. This architecture constrains the model's parameter count and features representation, reflecting the dimensionality-reducing behavior of quantum layers in the hybrid models.

### 2. Hybrid classical-quantum image classifiers

For the hybrid classical-quantum model, a quantum neural network (QNN) layer based on variational quantum circuits (VQC) is sandwiched between two classical neural network layers.[28] Features extracted from the initial layers are thresholded between 2 and 8 since an equivalent number of n-qubit systems is initialized for the quantum layer and features are embedded in the quantum systems. This is done keeping in consideration of various quantum hardware constraints. These features are the same in number as the number of qubits used in VQC and are the inputs to the QNN layer. For the hybrid classical-quantum model, a QNN layer based on VQC is sandwiched between two classical neural network layers. The outputs from the QNN layer are input to the final output layer.

From the model architecture presented in "Fig. 4," we can observe that the quantum operation is performed on three different layers of the QNN layer. The first layer is the embedding layer, which is responsible for mapping the data in the classical vector into a quantum state. In order to map classical data into the quantum state, different single qubit gates such as Hadamard gate, U1, U2, U3 gates, Rotational X, Rotational Y, and Rotational Z gates are used. Next layer is the variational quantum circuit layer, which is the concatenation of quantum layers of depth "d." Two qubit gates such as controlled Z, controlled-NOT, controlled-RX are used with parameterized single qubit gates to design parameterized circuits. Various variational quantum circuits (VQCs) are created and tested as a part of our study and the model with best-performing VQCs are selected.

The next step is to map the obtained outputs from the quantum circuit to the classical domain. For this, the expectation values from the quantum circuit are measured on one of X, Y, or Z basis. The result from this layer is the input to the next classical layer, which is the output layer. In our case, the output layer is the fully connected layer with two neurons for binary classification with sigmoid activation.

One of the key advantages of hybrid classical-quantum models over purely classical models lies in their resilience to gradient-based adversarial attacks. Many such attacks rely on calculating precise gradients to introduce adversarial perturbations. In hybrid models, the inclusion of a VQC introduces complex, non-classical optimization landscapes. These quantum-powered attributes can disrupt gradient information, making it harder for traditional attack algorithms to identify impactful perturbation paths. As a result, it contributes to robustness, much like how gradient masking operates.

In our experiment, we have used pre-trained transfer learning models from torchvision[29] and designed a classification model using Pytorch.[30] For hybrid classical-quantum model, the circuit is designed using PennyLane[31] and the integration of the quantum node with the classical PyTorch layer is done using the TorchLayer class of the QNN module from PennyLane. The quantum circuits are executed on the PennyLane default simulator.

**Selection of optimal VQCs:** for each of the transfer learning model, different VQCs are designed and trained on a small subset of data (i.e., 1000 images) through multiple trials, by tuning parameters such as the following.

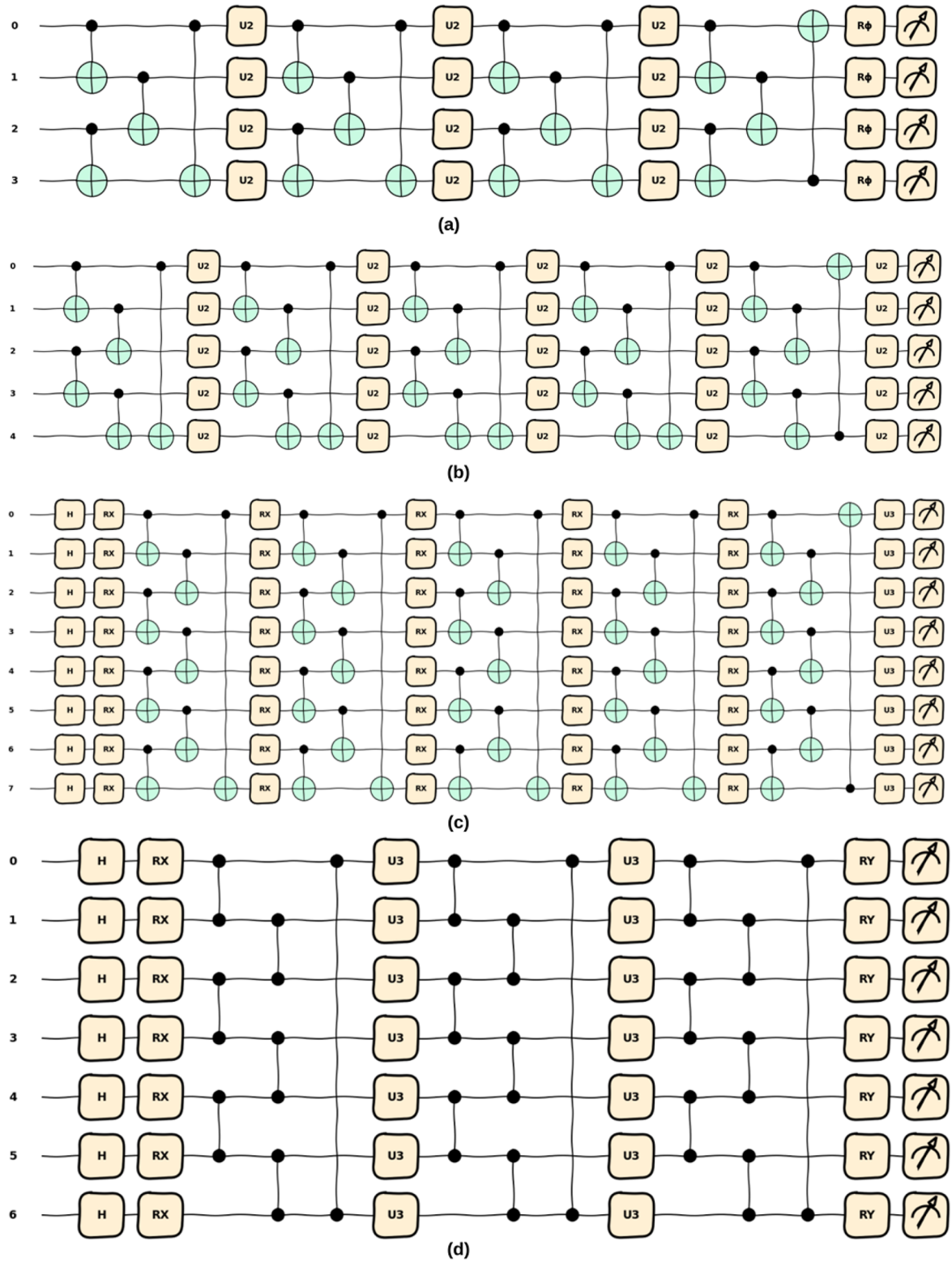


**FIG. 5.** VQCs used in our hybrid classical-quantum model, which achieved the highest classification accuracy on (a): VQC-1: ResNet-18, (b): VQC-2: AlexNet, (c): VQC-3: VGG-16, and (d): VQC-4: Inception-V3.

- n_qubits: no. of qubits to be used in quantum circuit
- input: parameter "input" is either 0 or 1. If the input is 1, the circuit consists of the Hadamard layer, followed by the selected quantum gate specified in the parameter "front_layer."
- front_layer: this parameter includes gates such as Rx, Ry, Rz, U1, U2, U3, and R$\varphi$
- entanglement_layer: specifies the type of entanglement layers to be included in quantum circuit such as C-Rx, C-Z, and C-Not.
- middle_layer: defines the quantum gate to be applied after each entanglement layer in order to achieve the desired circuit depth
- quantum_depth: specifies the number of repetitions of the combination of entanglement layers and corresponding quantum gate layers to control the overall depth of the quantum circuit
- output: a binary variable (0 or 1) indicating whether to include an additional quantum layer before the final measurement layer.
- final_layer: specifies the quantum gate to be used in the final layer before the measurement step, applicable only if output is set to 1
- measurement: specifies the basis for measurement, which can be set to either the X, Y, or Z basis.

VQCs are created through the combination of these parameters. The quantum circuit configurations are optimized through the choices of quantum gates, circuit depth, and entanglement patterns to achieve the high classification accuracy and adversarial robustness within the practical feasibility of quantum hardware. Different rotational gates are used to introduce sufficient parameter flexibility, and entanglement layers are selected for their ease of implementation and hardware compatibility. Initially, hundreds of models are created and tested using combinations of hundreds of VQCs to obtain the optimal-performing classification model through a trial-and-error method on a small subset of data. The VQC with optimal performance corresponding to each TL model is then trained with a larger amount of data, i.e., 20 000 images from the dataset. VQCs shown in "Fig. 5" are the circuits used with hybrid models in our study that achieved the highest classification accuracy on ResNet-18, AlexNet, VGG-16, and Inception-V3 models, respectively.

The VQC associated with the best-performing hybrid model (Hybrid ResNet-18), as depicted in "Fig. 5(a)," is designed using the following configuration: input = 0, entangle-ment_layer = C-Not, middle_layer = U2, quantum_depth = 4, output = 1, final_layer = R$\varphi$, and measurement = X-basis. VQCs for the remaining transfer learning models were similarly constructed by varying these parameters, based on experimental trials to identify optimal configurations.

### C. Preparing adversarial images

Deep learning models based on CNN have achieved higher accuracy in histopathological cancer detection.[21,22] However, these classification models are highly vulnerable to different kinds of adversarial attacks, which leads to unexpected predictions with higher confidence scores. The analysis of adversarial attacks on such models helps better estimate the reliability of the classifier model and design a method to defend against such attacks. The general scenario of an adversarial attack on the image classification model is depicted in "Fig. 6."

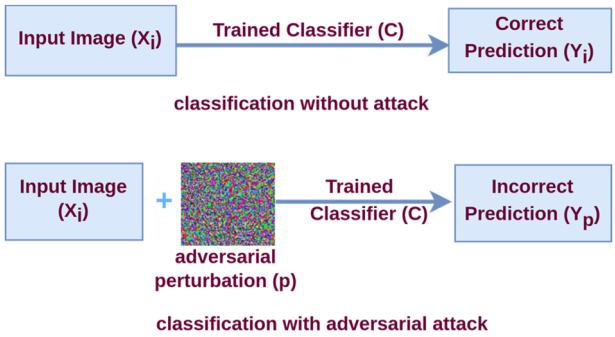


**FIG. 6.** Adversarial attack scenario: $X_i$ is the input image; $Y_i$ is the prediction without attack; and $Y_p$ is the prediction with adversarial perturbation.

Under normal conditions, the input image $X_i$ is fed into the classifier $C$, which gives output $Y_i$ that is the predicted target class corresponding to the input sample $X_i$. Under an adversarial attack, the input sample is intentionally adulterated with subtle, targeted noise, which is commonly known as adversarial perturbations. The integration of noise with the images is human-unobtrusive but they lead the classifier to misclassify the input sample with a high confidence score.

As our initial step toward exploring adversarial robustness in hybrid classical-quantum models, we focused on three widely recognized white-box attacks: Fast Gradient Sign Method (FGSM) attack, DeepFool attack, and projected gradient descent (PGD) attack. These attack methods are well-established perturbation strategies in the fields of adversarial machine learning and provide various strategies to develop robust models against these attacks. For the generation of adversarial images, we have used an untargeted attack method. In FGSM attack, adversarial images are generated using the sign of the gradient. Input images are fed into the classifier to generate the prediction and loss. Then, the gradient of loss is calculated with respect to the input. The gradient undergoes a sign function to calculate its sign value. The process of generating adversarial perturbation using FGSM is expressed in

$$X_{adv} = X + \varepsilon * \text{sign}\left(\nabla_x L(C, X, Y)\right) \quad (1)$$

where $X_{adv}$ = the generated adversarial image, $\varepsilon$ = perturbation coefficient that is lower enough to detect from human eye and higher enough to fool the classifier, and $L$ = loss function for classifier $C$ with input $X$ and target Y.

PGD attack is another variant of gradient-based attack. Perturbations are generated by running FGSM multiple times with small step size, and the adversarial values are clipped after each step to the perturbation constraint, which are already defined.[32,33] The DeepFool attack algorithm works on the basis of decision boundaries to generate perturbations.[16] The algorithm iteratively computes the gradient of the classification model's output with respect to the input image and then determines the direction of the gradient that leads to the smallest change in the image classification. This process is

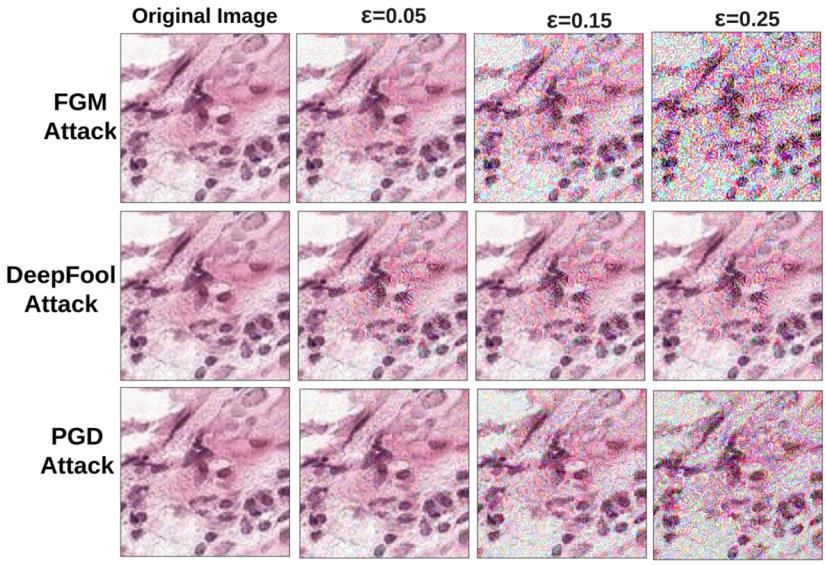


**FIG. 7.** Adversarial images: each row contains different types of attacks and each column contains perturbed images under different values of epsilon ($\varepsilon$).

repeated until there is a change in the prediction label of the input or the current iteration is the maximum.

"Figure 7" shows a sample of adversarial images under different values of epsilon ($\varepsilon$). "$\varepsilon$" is the perturbation coefficient used in each of the adversarial attack algorithms.

An example source code for this adversarial study is available on GitHub at https://github.com/tqausal/qadversarial/.

## IV. RESULTS

We evaluated the performance of various classical and hybrid classical-quantum binary image classification models under both the standard condition and when subjected to various adversarial perturbations. For classical computation, we created and trained four distinct models, which are based on four prominent transfer learning strategies, namely, ResNet-18, VGG-16, AlexNet, and Inception-v3. Similarly, we employed these transfer learning models as feature extractors to integrate them with different VQCs to create hybrid classical-quantum models and generate adversarial perturbations using different methods outlined in this study. We created and trained four different hybrid classical-quantum models with integration of four different VQCs. These classical and hybrid models are trained and tested on histopathological data images to evaluate their efficacy and robustness. Table II outlines the performance of different classical and hybrid classical-quantum models from our experiment. Column I represents the names of different models evaluated in our experiment. The computation types of different model architectures are mentioned in column II. The computations are either classical or hybrid classical-quantum. The test accuracies achieved from each of these models when trained on a quantum simulator without adversarial perturbations are included in column III. The VQCs used in each of the TL mode and the number of qubits used for their construction are listed in column IV and V, respectively. Different values of perturbation coefficients ($\varepsilon$) used in each of the attack models are included in column VI. For our experiment, we evaluated the performance of each of the classical and hybrid classical-quantum models under three perturbation coefficients: 0.05, 0.15, and 0.25. The accuracies of each model tested on adversarial samples generated using FGSM, DeepFool, and PGD with perturbation coefficient ($\varepsilon$) are listed in columns VII, VIII, and IX, respectively.

### A. Experimental results of classical models

From the performance in Table II, it is observed that in classical models without adversarial perturbation, the classification accuracies of CNN models with transfer learning strategies are admirable. The highest classification accuracy achieved by the classical models is 90.90%, obtained from the ResNet-18 transfer learning-based model trained and tested on a dataset of 20 000 images. On the other hand, when evaluating the performance of these classical transfer learning models, under the influence of various adversarial perturbations, their performance is lacking. Under the adversarial perturbation generated using FGSM attack, the highest classification accuracy is 50.84% ($\varepsilon$ = 0.25), which is obtained from the classical ResNet-18 based transfer learning model. Using the DeepFool method to generate perturbations, the highest classification accuracy of 37.87% ($\varepsilon$ = 0.25) is achieved from the classical ResNet-18 based model. Out of these classical models, the classical Inception-V3 transfer learning-based model achieved the highest classification accuracy of 32.05% ($\varepsilon$ = 0.05) under the influence of PGD attack.

### B. Experimental results of hybrid classical-quantum models

We analyzed the performance of each transfer learning model within hybrid classical-quantum models using a smaller subset of data consisting of 1000 images. The transfer learning model with admirable classification performance was then selected for further evaluation with a larger number of images. In our experiment,

**TABLE II.** Model performance with each of the adversarial attacks used in this study.

| Model | Computation type | Accuracy on simulator | VQC | No. of qubits | Epsilon ($\varepsilon$) | Accuracy under FGSM attack | Accuracy under DeepFool attack | Accuracy under PGD attack |
|---|---|---|---|---|---|---|---|---|
| Classical ResNet-18 | Classical | 90.90 | N/A | N/A | 0.05 | 18.82 | 35.17 | 17.75 |
| | | | | | 0.15 | 43.15 | 36.97 | 10.10 |
| | | | | | 0.25 | 50.84 | 37.87 | 10.10 |
| Classical AlexNet | Classical | 88.90 | N/A | N/A | 0.05 | 23.90 | 22.25 | 14.35 |
| | | | | | 0.15 | 39.72 | 24.17 | 11.07 |
| | | | | | 0.25 | 49.45 | 25.07 | 11.07 |
| Classical VGG-16 | Classical | 85.39 | N/A | N/A | 0.05 | 25.75 | 14.60 | 15.42 |
| | | | | | 0.15 | 31.67 | 14.60 | 14.60 |
| | | | | | 0.25 | 47.69 | 14.60 | 14.60 |
| Classical Inception-V3 | Classical | 82.52 | N/A | N/A | 0.05 | 40.77 | 28.12 | 32.05 |
| | | | | | 0.15 | 49.95 | 31.92 | 17.95 |
| | | | | | 0.25 | 50.02 | 33.55 | 18.25 |
| Hybrid ResNet-18 | Hybrid Classical Quantum | 84.77 | 1 | 4 | 0.05 | 78.05 | 48.62 | 52.94 |
| | | | | | 0.15 | 69.12 | 50.34 | 50.12 |
| | | | | | 0.25 | 50.37 | 52.12 | 50.00 |
| Hybrid AlexNet | Hybrid Classical Quantum | 80.90 | 2 | 5 | 0.05 | 56.59 | 48.19 | 50.00 |
| | | | | | 0.15 | 50.60 | 49.85 | 50.00 |
| | | | | | 0.25 | 50.00 | 50.92 | 50.00 |
| Hybrid VGG-16 | Hybrid Classical Quantum | 83.37 | 3 | 8 | 0.05 | 52.00 | 28.97 | 52.90 |
| | | | | | 0.15 | 49.62 | 32.47 | 48.80 |
| | | | | | 0.25 | 50.00 | 34.62 | 45.85 |
| Hybrid Inception-V3 | Hybrid Classical Quantum | 80.05 | 4 | 7 | 0.05 | 50.04 | 25.12 | 50.80 |
| | | | | | 0.15 | 50.00 | 28.24 | 48.87 |
| | | | | | 0.25 | 50.00 | 31.80 | 49.85 |
| Classical Bottleneck | Classical | 84.27 | N/A | N/A | 0.05 | 26.02 | 15.80 | 46.40 |
| | | | | | 0.15 | 49.07 | 15.95 | 22.00 |
| | | | | | 0.25 | 49.97 | 16.10 | 15.75 |

the hybrid ResNet-18 transfer learning model outperformed the hybrid models based on the other three transfer learning architectures. The Hybrid ResNet-18 model, trained and tested with a dataset consisting 20 000 images, achieved the highest classification accuracy, which is found to be 84.77% without adversarial perturbations. However, the success rate of up to 78.05% is achieved when these hybrid classical-quantum transfer learning models are subjected to various adversarial attacks. The hybrid classical-quantum model based on the ResNet-18 transfer learning model with VQC-1 outperformed other hybrid models under FGSM, DeepFool, and PGD adversarial attacks. Under FGSM, DeepFool, and PGD attacks, classification accuracies achieved are 78.05% ($\varepsilon$ = 0.05), 52.12% ($\varepsilon$ = 0.25), and 52.94% ($\varepsilon$ = 0.05), respectively. To ensure statistical significance, the best-performing model was evaluated using three different random seed values. The results remained stable across these runs, demonstrating the consistency of the model's performance.

## V. CONCLUSION

Machine learning models have achieved state-of-the-art performance on different medical image-related operations such as image generation, image classification, object detection, image segmentation, and anomaly detection. However, the use of these models in medical sectors is extremely vulnerable to different kinds of malicious attacks, commonly known as adversarial attacks. In this experiment, we explored the impact of adversarial attacks on different classical and hybrid classical-quantum image classification models for histopathological cancer detection. For the evaluation of classical models, we chose five classical models: four based on transfer learning architectures and one low-parameter classical bottleneck model. Similarly, we chose four hybrid classical-quantum models to evaluate their performance under different adversarial attacks. The experiment we performed shows that both the classical and hybrid classical-quantum models deployed are highly vulnerable

to adversarial attacks. However, the success rate of defense against such adversarial perturbations of hybrid classical-quantum models is higher than that of classical TL-based classification models and a parameter-reduced bottleneck model. This shows that integrating quantum principles with classical image classifiers alone is not sufficient to tackle adversarial attacks. However, the positive results obtained from our experiments suggest the potential for further development of robustness against adversarial attacks through additional image processing and the utilization of defense mechanisms.

As our initial step toward exploring adversarial attacks, we have utilized these four simple transfer learning models with three popular white-box threat methods. As a consequent work, we plan to extend this research by using different advanced state-of-the-art image classification models and incorporating additional attack strategies, including black-box attacks. Our goal is to develop robust and resilient models capable of with-standing various malicious attacks by incorporating the architecture of hybrid classical-quantum models along with other adversarial defense mechanisms. Furthermore, this work can be tested on real quantum hardware to evaluate how well it performs in practical world.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

**Biswaraj Baral**: Methodology (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Bhavika Bhalgamiya**: Methodology (equal); Validation (equal). **Reek Majumder**: Conceptualization (equal); Formal analysis (equal); Methodology (equal). **Divya Dutta Roy**: Methodology (equal); Writing – review & editing (equal). **Taposh Dutta Roy**: Conceptualization (equal); Methodology (equal); Resources (equal).

## DATA AVAILABILITY

The sample data and code that support the findings of this study are openly available on GitHub at: https://github.com/tqausal/qadversarial/tree/main. The Python open-source library QausalML is currently under documentation and will be made available soon.

## REFERENCES

[1]H. Dong, J. Dong, S. Yuan, and Z. Guan, "Adversarial attack and defense on natural language processing in deep learning: A survey and perspective," in *International Conference on Machine Learning for Cyber Security* (Springer, 2023), pp. 409–424.

[2]M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security* (Association for Computing Machinery (ACM, 2006), pp. 16–25.

[3]S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," Science **363**, 1287–1289 (2019).

[4]H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," BMC Med. Imaging **21**, 9 (2021).

[5]N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[6]S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," arXiv:1804.05296 (2018).

[7]A. I. Newaz, N. I. Haque, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "Adversarial attacks to machine learning-based smart healthcare systems," in *GLOBECOM 2020-2020 IEEE Global Communications Conference* (IEEE, 2020), pp. 1–6.

[8]A. Selvakkumar, S. Pal, and Z. Jadidi, "Addressing adversarial machine learning attacks in smart healthcare perspectives," in *Sensing Technology: Proceedings of the ICST 2022* (Springer, 2022), pp. 269–282.

[9]F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International Conference on Machine Learning* (PMLR, 2020), pp. 2206–2216.

[10]I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," CoRR abs/1412.6572 (2014).

[11]A. Ghiasi, A. Shafahi, and T. Goldstein, "Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates," International Conference on Learning Representations (ICLR) (2020).

[12]E. Wong, F. Schmidt, and Z. Kolter, "Wasserstein adversarial examples via projected Sinkhorn iterations," in *International Conference on Machine Learning* (PMLR, 2019), pp. 6808–6817.

[13]H. Hirano and K. Takemoto, "Simple iterative method for generating targeted universal adversarial perturbations," Algorithms **13**, 268 (2020).

[14]W. Brendel, J. Rauber, M. Kümmerer, I. Ustyuzhaninov, and M. Bethge, "Accurate, reliable and fast robustness evaluation," in *Advances in Neural Information Processing Systems*, 32 (Curran Associates, Inc., 2019).

[15]S. Kotyan and D. V. Vargas, "Adversarial robustness assessment: Why in evaluation both $L_0$ and $L_\infty$ attacks are necessary," PLOS ONE **17**, e0265723 (2022).

[16]S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 2574–2582.

[17]P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on artificial Intelligence and Security* (Association for Computing Machinery, 2017), pp. 15–26.

[18]M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High accuracy and high fidelity extraction of neural networks," in *Proceedings of the 29th USENIX Conference on Security Symposium* (USENIX Association, 2020), pp. 1345–1362.

[19]J. R. Correia-Silva, R. F. Berriel, C. Badue, A. F. de Souza, and T. Oliveira-Santos, "Copycat CNN: Stealing knowledge by persuading confession with random non-labeled data," in *2018 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2018), pp. 1–8.

[20]T. Orekondy, B. Schiele, and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 4954–4963.

[21]B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11* (Springer, 2018), pp. 210–218.

[22]B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," JAMA **318**, 2199–2210 (2017).

[23]K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," J. Big Data **3**, 9 (2016).

[24]K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
[25]C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), pp. 2818–2826.
[26]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos, CA, 2016), pp. 770–778.
[27]A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc., 2012), Vol. 25.
[28]A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, "Transfer learning in hybrid classical-quantum neural networks," Quantum **4**, 340 (2020).
[29]S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of torch," J. Assoc. Comput. Mach. 1485–1488 (2010).
[30]A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vol. 32, pp. 8024–8035.
[31]V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi *et al.*, "PennyLane: Automatic differentiation of hybrid quantum-classical computations," arXiv:1811.04968 (2018).
[32]A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv:1706.06083 [stat.ML] (2019).
[33]Y. Jiang, G. Yin, Y. Yuan, and Q. Da, "Project gradient descent adversarial attack against multisource remote sensing image scene classification," Secur. Commun. Networks **2021**, 6663028.