Adversarial Attack-Resilient Perception Module for Traffic Sign Classification

Reek Majumder[1], Mashrur Chowdhury[1], Sakib Mahmud Khan[1], Zadid Khan [2], Fahim Ahmed[3], Frank Ngeni[4], Gurcan Comert[5], Judith Mwakalonge[4], and Dimitra Michalaka[6]

[1] Glenn Department of Civil Engineering, Clemson, SC 29631

[2] Walmart Supply Chain (Transportation), Bentonville, AR 72712

[3] Department of Civil and Environmental Engineering, University of South Carolina, Columbia, SC 29208

[4] Department of Engineering, South Carolina State University, Orangeburg, SC 29117

[5] Computer Science., Physics., and Engineering Department, Benedict College, Columbia, SC 29204

[6] Department of Civil and Environmental Engineering, The Citadel, Charleston, SC 29409

Corresponding author: Reek Majumder (e-mail: rmajumd@g.clemson.edu).

ABSTRACT: Deep Learning (DL)-based image classification models are essential for autonomous vehicle (AV) perception modules since incorrect categorization might have severe repercussions. Adversarial attacks are widely studied cyberattacks that can lead DL models to predict inaccurate output, such as incorrectly classified traffic signs by the perception module of an autonomous vehicle. In this study, we create and compare Hybrid Classical-Quantum Deep Learning (HCQ-DL) models with Classical Deep Learning (C-DL) models to demonstrate robustness against adversarial attacks for perception modules. Before feeding them into the quantum system, we used transfer learning models like AlexNet and VGG-16 as feature extractors. We tested over 1000 quantum circuits in our HCQ-DL models for Projected Gradient Descent (PGD), Fast Gradient Sign Attack (FGSA), and Gradient Attack (GA), which are three well-known untargeted adversarial approaches. We evaluated the performance of all models during adversarial attack and no-attack scenarios. Our HCQ-DL models maintain accuracy above 95% during a no-attack scenario and above 91% for GA and FGSA attacks, which is higher than C-DL models. During the PGD attack, our AlexNet-based HCQ-DL model maintained an accuracy of 85% compared to C-DL models that achieved accuracies below 21%.

Keywords: Quantum Machine Learning, Quantum-circuits, Deep Learning, Adversarial Attacks

I. INTRODUCTION

A. BACKGROUND

AUTONOMOUS Vehicles (AVs) widely use Deep Learning (DL) models to implement object detection [1], [2] and classification [3] tasks for their perception [4] module tasks to detect lanes, obstacles, and traffic signs. Nevertheless, these models are susceptible to adversarial attacks, and their performance deteriorates significantly when a carefully crafted perturbation/noise has been injected with the input image [5]–[9]. The severity of these adversarial attacks on DL models depends mainly on an attacker's goals and knowledge of the model. The attacker's intent can be to perform a targeted or non-targeted attack. An attacker influences the model in a targeted attack to predict a particular output. While in non-targeted attacks, an attacker does not care about a prediction of a particular class. These attacks have been categorized into three groups based on the attacker's knowledge of DL models: white-box, gray-box, and black-box attacks. An attacker that uses a white-box attack is aware of the trained DL model and creates carefully considered perturbations to the input data to deceive it. In a gray-box attack, an attacker is aware of the model's design but is uninformed of its training weights. In a black-box attack, the attacker creates random disruption, with no knowledge of the model, to trick a trained model.

Various defense strategies have been suggested, including image transformation or model re-training techniques. The image transformation technique involves input reconstruction [10]–[12] and inputs denoising [10] methods, which aim to pre-process the images before entering the DL models using techniques like smoothing, filtering (e.g., JPEG Filter [13], [14], Binary Filter [15], Random filtering [16],), and feature squeezing [15]. Model re-training involves adversarial attack detection [17] and training [18] methods, where adversarial samples are generated using vigorous attacks, and the DL models are re-trained on adversarial samples. Another method called the defensive distillation technique [19] is another strategy that combines detection and training networks. In this method, the detection network creates the probability vectors to label the original dataset, and the training network is used to re-train the model using the labeled adversarial samples dataset generated by detection network. However, these techniques can perform well for known adversarial attacks used to create adversarial samples for re-training, but they can perform poorly for unknown attacks in the future; therefore, we need resilient DL models during adversarial attacks. Based on [20] definition of cyber-resilience, we define DL resilience as the capability to correctly categorize the image, although malicious parties perturb the input image.

Since quantum computing is becoming more mainstream, quantum computers [21] have recently joined the race for high-performing computing systems due to their computational advantages of using quantum mechanical properties like superposition and entanglement. Theoretically, Hilbert space for quantum systems [22] increases exponentially to system size, which makes it harder to simulate on classical computers. For example, a quantum system with tens and hundreds of qubits is classically intractable and proposed to demonstrate quantum supremacy over classical supercomputers [23]. Companies like Google have recently claimed quantum supremacy with its 53-qubit system named *Sycamore*, which will take classical supercomputers almost 10,000 years to solve [24]. Our research aims to use classical and quantum computers in parallel, to develop Quantum Machine Learning (QML) based hybrid classical-quantum deep learning (HCQ-DL) architecture. An experiment for classifying traffic signs with DL models during an adversarial attack in [25] incorrectly detected a *stop sign* as a *speed sign* which can lead to severe collisions if AVs operate with the help of these DL models. The goal is to test the performance and resiliency of HCQ-DL models against adversarial attacks without using image transformation or model re-training techniques. We use Transfer Learning (TL) [26] to extract features from pre-trained DL models like AlexNet and VGGNet before inputting the data to the quantum systems because currently available quantum processors, also known as Noisy Intermediate-Scale Quantum (NISQ) systems, cannot embed image data in a quantum system directly. These pre-trained models are developed using the 1.2 million images for 1000 categories from the ImageNet dataset [27]**.** For image classification, the initial convolution layer of these models is frozen and acts as a feature extractor. Finally, the last layers are replaced with custom layers of Artificial Neural Networks (ANNs) or Quantum Neural Networks (QNN) and tuned for our LISA traffic sign dataset [28] during our analysis. LISA dataset consists of traffic signs taken from video shots from driving vehicles. We designed our QNN model consisting of low-depth Variational Quantum Circuits (VQC) [29]–[32], which can learn based on the quantum circuit learning framework [30] for currently available NISQ hardware from IBM [33], Xanadu [34] and Google [35].

B. CONTRIBUTION

To our knowledge, the performance and resilience comparison of the QML- based HCQ-DL with C-DL models has not been studied. In this research, we tested over 1000 quantum circuit-based QNN layers for HCQ-DL models. We compared them against C-DL models for traffic sign classification, primarily for the AV sign perception module. Both our HCQ-DL and C-DL models are developed using AlexNet and VGG-16 as feature extractors. Traditionally, VGG-16 models aimed to improve performance by using a deeper network with smaller filters (sixteen layers) than other models like AlexNet (eight layers). C-DL models with AlexNet and VGG-16 as feature extractors are vulnerable to adversarial attacks like Gradient attacks (GA), Fast gradient Sign attacks (FGSA), and Projected Gradient Descent attacks (PGD). Among the HCQ-DL models, our analysis shows that

AlexNet-based HCQ-DL models (8% decrease in accuracy) outperform VGG-16-based HCQ-DL models (89% decrease in accuracy) in traffic sign classification resiliency during highly effective adversarial attacks, such as PGD attacks, with shorter training time. This study will also motivate the development of quantum-enabled HCQ-DL models for other use cases.

C. OUTLINE

Section II discusses datasets describing the dataset's origin and attack dataset's generation. The creation of the C-DL and HCQ-DL models and a description of the performance metrics applied in our study are covered in Section III's discussion of research method. Section IV summarizes the findings of our investigation and shows how the performance of the model changes as cyberattack intensity rises. The conclusion based on the findings is discussed in Section V. Section VI provides suggestions for additional analysis for future studies.

II. DATASET

A. IMAGE DATASET

We used a portion of the extended LISA [28] traffic sign dataset, which contains around 7,855 annotations from 6,610 video frames identifying 47 different traffic signs. To examine the performance of the HCQ-DL and C-DL models we focused on stop signs and combination of other signs to design a balanced dataset. Image frames vary from 640 x 480 to 1024 x 522 pixels. The size of annotation boxes for traffic signs ranges from 6 x 6 to 167 x 168 pixels. The number of samples for each type of traffic sign differs significantly. Initially, we grouped the traffic sign dataset into 18 traffic signs and cropped the images to reduce the noise in their surroundings.

In the binary classification models [36], both HCQ-DL and C-DL models are trained to classify *stop signs* and *other signs* by creating a balanced dataset. On the extracted balanced dataset, which included 231 samples with an 80-20 split between training and testing data, both HCQ-DL and C-DL models were trained and tested for twenty-five epochs (number of training samples:182, number of testing samples: 49).

B. ATTACK MODELS FOR ADVERSARIAL DATASET

In this study, three types of white-box attacks have been chosen based on severity (elementary, intermediate, and advanced), and attacks generated with Fast Gradient Sign Attack (FGSA) [9], Gradient Attack (GA)[9], and Projected Gradient descent (PGD) attack [37] for both HCQ-DL and C-DL models. Before the image is entered into the classification model, these attacks create perturbations to the input data based on the epsilon coefficient. These attacks differ in how the perturbation is applied to the original input images to generate misclassified outputs. Equation 1 illustrates how the gradient attack modifies the input image considering the gradient of the loss function for the DL model to produce an adversarial image. Moreover, the FGSA is a single-step gradient ascent strategy that uses a sign of gradient with a fixed epsilon coefficient to generate an adversarial image, as shown in equation 2. However, PGD obtains adversarial samples by iteratively using the fast gradient method, and the iteration starts uniformly at a randomly chosen data point. It projects (*Proj*) the adversarial samples from each iteration into the next using the product of epsilon and gradient of loss function, as represented in equation 3 [16].

Adversarial image = input image + [Epsilon * Gradient of Loss function] (1)

Adversarial image = input image + [Epsilon * Sign of Gradient of Loss function] (2)

Iterative Adversarial Image = Projection of [Adversarial image at previous instance + Epsilon * Sign of gradient of Loss Function] (3)

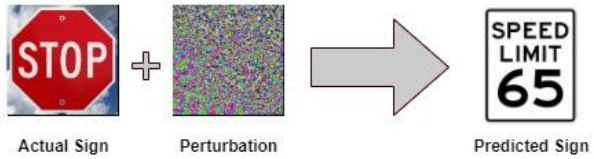


FIGURE 1. Example of adversarial attack on stop signs.

All attacks studied during this experiment modify input images to deceive DL models, as shown in figure 1, and attack intensity has varied from perturbation (Epsilon) coefficients ranging from 0.05 to 0.5, as discussed in Section IV.

III. RESEARCH METHOD

In this section, we discuss transfer learning and quantum neural networks as building blocks for C-DL and HCQ-DL models.

A. TRANSFER LEARNING

A well-known machine learning method for feature extraction and handling situations when we lack training data is transfer learning (TL) [38], [39]. Since we have only 182 samples in our training set and want to encode image data to a quantum system, we use known TL models as feature extractors (AlexNet and VGG-16). For image classification tasks, initial convolution layers of these TL models are believed to learn similar features, so their weights are fixed and extended to fine-tune on newer tasks by replacing final layers with ANNs.

These TL models are trained on ImageNet [27] Dataset with 1.2 million images for over 1000 categories. The objective behind employing TL models is to utilize the initial convolution layer of these pre-trained models and swap out the final linear layer with layers of custom linear layers according to the given use case.

B. CLASSICAL DEEP LEARNING (C-DL) and HYBRID CLASSICAL-QUANTUM DEEP LEARNING (HCQ-DL) MODELS

This section describes the development of the C-DL and HCQ-DL model architecture used in our study for training and testing traffic sign classifiers. We also represent the final hyperparameters of our HCQ-DL and C-DL models with VGG-16 and AlexNet pre-trained models as feature extractors.

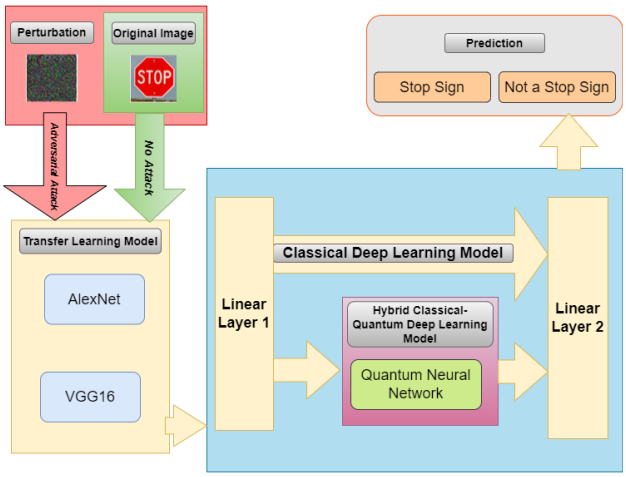


FIGURE 2. Architecture for C-DL and HCQ-DL models

1) Classical Deep Learning (C-DL) Models

As shown in figure 2, our experiment uses two state-of-the-art Convolutional Neural Network-based deep learning models, VGG-16[40] and AlexNet [41], as our feature extractors for our C-DL and HCQ-DL models. Due to the lack of training samples, we use transfer learning principles for the image classification task and freeze the initial layers of the pre-trained model. Later we introduced two linear layers for our C-DL model (figure 2) with a rectilinear unit (ReLU) as an activation function for linear layer one and SoftMax for linear layer two. We later trained these C-DL models on stop signs and other sign classes.

2) Hybrid Classical-Quantum Deep Learning (HCQ-DL) Models

We developed and tested over 1000 quantum circuits-based QNN models with pre-trained DL- models to develop our HCQ-DL models. The QNN models were composed of various single qubits and multiple qubits gates. These gates are fundamental building blocks for circuit-based quantum algorithms. Usually, these gates are evaluated based on the impact of each gate on 3-dimensional space for each qubit, often referred to as the Bloch sphere. The gates used in our study are explained below.

a. Hadamard gate: Fundamental quantum gate. It helps us to create a superposition of states $|0\rangle$ and $|1\rangle$ by moving away from the poles of the Bloch Sphere.[42]

b. Rotational Gates: There are three Pauli rotational gates Rotational-X(RX), Rotational-Y(RY), and Rotational-Z(RZ). These gates rotate the state vector about the corresponding axis. Often generated by taking the exponential of Pauli operators. Where Pauli-X(X) is a bit-flip gate, Pauli-Y(Y) is a bit, and phase-flip gate, and Pauli-Z(Z) is a phase-flip date.[43]

c. Universal gates: These advanced single-qubit quantum gates combine rotational and phase shift gates to represent different states in the Bloch sphere. The three most common universal gates are Universal gate 1(U1), Universal gate 2(U2), and Universal gate 3 (U3). The U1 gate is equivalent to a phase shift gate, the U2 gate is a combination of rotational gates around the Y and Z- axis in the Bloch sphere with a phase shift. The U3 gate combines the rotation around all three axes of the Bloch Sphere.[43], [44]

d. Controlled gates: In our study, we have used 2-qubit (control and target) controlled gates with Pauli operators (X, Y, Z) and rotational gates (RX, RY, RZ). Furthermore, these gates are implemented on the target qubit only when the control qubit is in the $|1\rangle$ state. It is one of the unique features of our study, which induces a correlated state (entanglement) among various qubits.[43]

e. Measurement gate: Finally, all 1000 quantum circuits were introduced with a measurement gate, which enables the quantum states to be mapped to classical bits. Each of the 1000 circuits was run over 1000 times for each sample to provide the probability of each possible outcome. [44] .The output state with the highest probability was used as an input

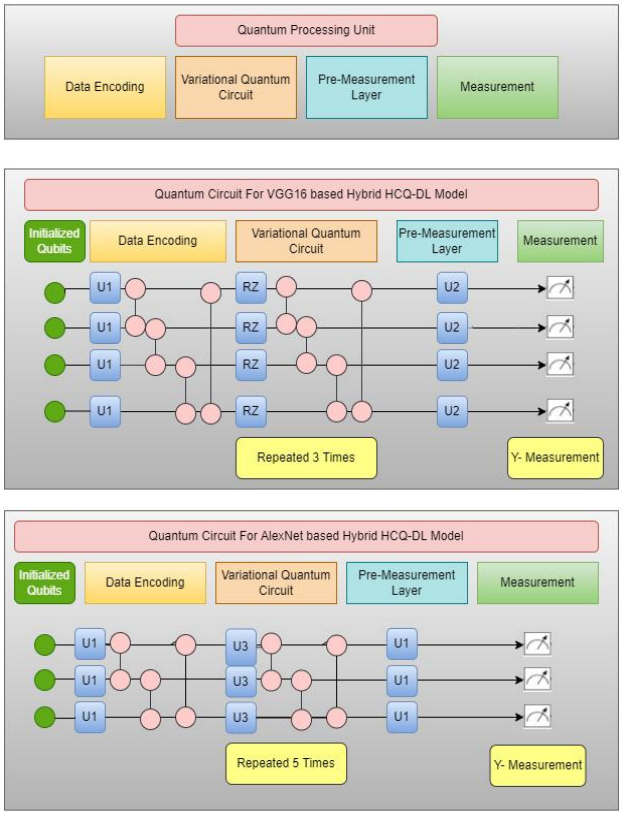


FIGURE 3. Quantum Circuit Architecture with chosen Circuits for our VGG16-based HCQ-DL and AlexNet-based HCQ-DL models with better resiliency

for our next phase, which is linear layer 2 as shown in Figure 2. Figure 3 represents our architecture for QNN models and best quantum circuit for our HCQ-DL model with VGG-16 and AlexNet as the feature extractor. As we cannot map image data directly to quantum computers [30], [45], we use the HCQ-DL models to pre-process large input images classically using CNN-based pre-trained models (VGG-16 and AlexNet) and replace its final layer with a quantum layer sandwiched between two linear layers as shown in figure 2.

Furthermore, we divide the QNN layer into four broad categories. The Data Encoding layer embeds data from the classical system into the quantum system. The Variational Quantum Circuit combines repetitive single and multi-qubit gates, a pre-measurement layer consisting of single-qubit gates, and a measurement layer that reads data from the quantum system to the classical system. We have optimized the first linear layer's number of neurons, ranging from 2 to 8, for the C-DL, and it refers to the number of qubits needed to initialize for quantum layers of HCQ-DL models. The ranges of other DL parameters are schedular step size ranging from 5 to 10 and learning rate ranging from 0.01 to 0.0001. We consider batch sizes 2, 4, 8, 16, 32, and 64. We tested our models for Adam and Stochastic Gradient Descent (SGD) optimizer and reported the tuned model parameters for our best models in Table I.

TABLE I
MODEL HYPERPARAMETERS AND TRAINING TIME

	Classical Model		Hybrid Classical-Quantum Model	
Transfer Learning Model	VGG-16	AlexNet	VGG-16	AlexNet
Schedular Step Size	8	9	8	9
Batch Size	32	64	2	8
Learning Rate	0.000697	0.000269	0.000194	0.002906
Optimizer	Adam	Adam	Adam	Adam
Classical Model (Neuron) / Hybrid Model (N-Qubits)	2	6	4	3
Training Time	14 min 11 seconds	2 min 27 seconds	47 min 38 seconds	28 min 11 seconds
Number of parameters	14,764,872	2,525,012	14,815,066	2,497,370

Comparing C-DL AlexNet vs. VGG-16 in Table I, we found that AlexNet models have relatively lower training time due to fewer parameters. Lower training time due to a lower number of parameters can be seen as consistent in the HCQ-DL AlexNet model vs. the HCQ-DL VGG-16 model. However, HCQ-DL models have considerably higher training time even after having fewer parameters because we use quantum simulators from PennyLane [34] and cannot exclude the wait time for each batch during training. With the improvement in performance and availability of quantum hardware, the computation time will likely go down.

3) Performance Matrix

The C-DL and HCQ-DL models are evaluated using accuracy for training and testing, where *accuracy* is defined as the proportion of correct predictions to all predictions for samples in train and test set. For our investigation of adversarial attacks, we selected the models with higher accuracy

We tested our models with other performance metrics like sensitivity, specificity, false positive rate, precision, and F1 score. Specificity refers to the proportion of other sign classes that are correctly classified. At the same time, sensitivity, also known as recall, relates to the fraction of stop sign classes that are correctly classified. We also calculated the False Positive Rate, which refers to the fraction of *other sign class* samples misclassified as a *stop sign*. We also calculated the precision for each model to study how well the models are classifying *stop signs* and *other sign classes* and tested the balance between precision and recall by evaluating the F1 score of our model.

$$MCC = \left(\frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TN + FN)}}\right) * \frac{1}{\sqrt{(TP + FN)(TN + FP)}} \quad (4)$$

Where
TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

Finally, we tested C-DL and HCQ-DL-based classifier models on statistical measure phi-coefficient, also referred to as Matthew's correlation coefficient (MCC), which measures the correlation between the actual classes of the dataset and the predicted classes by the models and calculated as shown in equation 4.

IV. ANALYSIS

This section discusses the performance results for our C-DL and HCQ-DL models. And QNN models architecture for best HCQ-DL models.

Our best VGG-16-based HCQ-DL model, we initialize a 4-qubit system, where during the data encoding phase, we use a U1 gate with a controlled-Z (CZ) entanglement gate, while in a variational quantum circuit includes RZ and CZ gate repeated for three times and U2 gate in Pre-Measurement. In the AlexNet-based HCQ-DL model, we initialize a 3-qubit system, where the data encoding phase uses a U1 gate with a CZ entanglement gate, while our variational quantum circuit includes U3 and CZ gates repeated five times and a U1 gate in Pre-Measurement layer. Finally, we perform quantum Y-measurement for both models to map quantum states to classical output. The circuits are executed for 1000 shots, and the state with maximum probability is considered quantum measurement output. The measurement output is mapped to the ANN layer with neurons equivalent to the qubit system initialized to each model, as shown in Figure 2. Finally, an ANN layer of two neurons with SoftMax activation function is used to

provide a final output. Table II.

Table II represents the complete report regarding the accuracy, sensitivity, specificity, false positive rate, precision, F1 score, and Matthew's correlation coefficient for each of our best C-DL and HCQ-DL models. We analyzed these parameters and found that all our models satisfy the criteria of higher sensitivity, specificity, precision, F1-score, and Matthew's correlation coefficient while having a lower false positive rate, which is an essential criterion for a good classifier.

TABLE II
MODEL PERFORMANCE METRICS

	Classical Model		Hybrid Classical-Quantum Model	
Transfer Learning Model	VGG-16	AlexNet	VGG-16	AlexNet
Accuracy	96%	96%	98%	96%
Specificity	1	1	1	1
Sensitivity	0.90	0.90	0.95	0.90
False Positive Rate	0.10	0.10	0.05	0.10
Precision Score	0.93	0.93	0.97	0.93
F1-Score	0.97	0.97	0.98	0.97
Mathew Correlation Coefficient (MCC)	0.92	0.92	0.96	0.92

Table II lists our top C-DL and HCQ-DL model's respective accuracy, sensitivity, specificity, false positive rate, precision, F1 score, and Matthew's correlation coefficient. As a result of our analysis of these variables, we discovered that all of our models meet the crucial criteria of a strong classifier: —higher sensitivity, specificity, accuracy, precision F1-score, and Matthew's correlation coefficient—

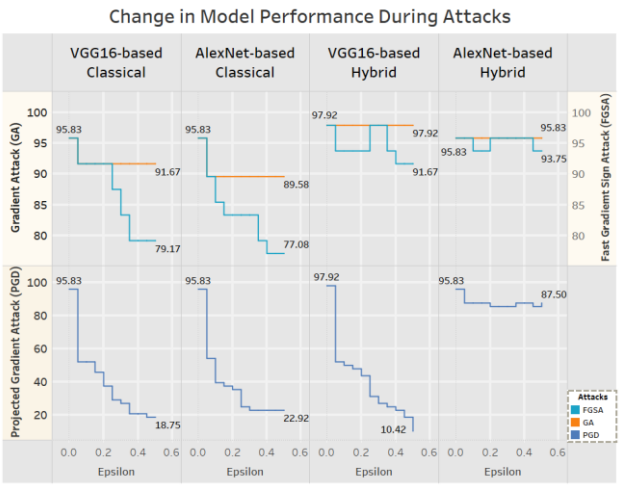


FIGURE 4. Change in Model Accuracy with the change in intensity of Attacks

while having lower false positive rates.

Figure 4 shows the performance change for HCQ-DL and C-DL models during GA, FGSA, and PGD attacks with a perturbation coefficient ranging from 0.05 to 0.5 with an interval of 0.05. We can see from figure 4 shows that VGG-16 and AlexNet-based HCQ-DL models are more resilient and maintain higher accuracy even after increasing the intensity of GA and FGSA attacks.

The most aggressive attack taken into consideration in our analysis is projected gradient descent (PGD), and Figure 4 shows the change in performance of our model with increasing intensity of PGD attack. We can see AlexNet-based HCQ-DL model exponentially outperforms all our models by continuously maintaining higher accuracy.

TABLE III
MODEL WORST PERFORMANCE DURING ADVERSARIAL ATTACK WITH PERTURBATION COEFFICIENT (PC.)

	Classical Model		Hybrid Classical-Quantum Model	
Transfer Learning Model	VGG-16	AlexNet	VGG-16	AlexNet
Accuracy (PC.)	96% (0.0)	96% (0.0)	98% (0.0)	96% (0.0)
Gradient Attack (PC.)	92% (0.05)	90% (0.05)	98% (0.5)	96% (0.5)
Fast Gradient Sign (PC.)	79% (0.35)	77% (0.4)	92% (0.4)	94% (0.45)
Projected Gradient Descent Attack (PC.)	**19% (0.45)**	**23% (0.3)**	**10% (0.5)**	**85% (0.2)**

The worst results of our C-DL and HCQ-DL models are shown in Table III during adversarial attacks with the perturbation coefficient.

V. CONCLUSIONS

From AlexNet to VGG-16, CNN-based DL models have historically evolved, intending to add more convolutional layers for better data mapping. In this investigation, we discovered that the performance and resistance of DL models to adversarial attacks could be enhanced using HCQ-DL models built by combining quantum layers with the current C-DL model. We obtained better performance accuracy and resilience outcomes under adversarial attacks while not using well-known defenses such as image modification or model re-training on adversarial samples in our study. Image modification techniques are pre-processing steps emphasizing smoothing, and filtering introduces additional implementation steps during the training and testing phase without considering the type of attack. While model re-training methods emphasize re-training on adversarial images, they do not account for unknown adversarial attacks, which could be more malicious than the known attack models based on which these are re-trained. Our HCQ-DL method presented in this study offers a performance accuracy boost during a known or unknown adversarial attack by introducing a quantum network within C-DL models without

requiring pre-processing procedures like Image modification or post-processing procedures like model-re-training. This occurs because quantum systems are known for mapping various counter-intuitive patterns, which is leveraged in our HCQ-DL model development.

Higher accuracy, precision, recall, F1 score, specificity, Mathew's correlation coefficient, and a lower false positive rate are typically required for a classifier to function well. Our study shows that both our C-DL and HCQ-DL satisfy these criteria; however, they are vulnerable to adversarial attacks. Our study showed that compared to C-DL models, HCQ-DL models maintain a higher accuracy (above 95%) during gradient attacks and above 90% during Fast Gradient sign attacks. For the projected gradient descent attack, we found that the AlexNet-based hybrid model outperforms other HCQ-DL and C-DL models by constantly having an accuracy above 85%.

Our AlexNet-based HCQ-DL shows better performance and resiliency during adversarial attacks than VGG-16-based HCQ-DL models. This opens the possibility of looking into shorter networks with quantum layers, resulting in fewer parameters but still maintaining a higher level of feature mapping, thus improving the performance accuracy and resilience of these next-generation models against adversarial attacks.

VI. FUTURE WORK

Our research demonstrates that the C-DL architecture may be considerably strengthened by adding a single quantum layer to increase the robustness of deep learning models against adversarial attacks. HCQ-DL models can maintain relatively higher accuracy for highly effective adversarial attacks like L-infinite Projected gradient descent attacks. In the future, we will evaluate the effect of the perturbation on the internal layers of DL models, which are used as feature extractors for our HCQ-DL models. We would also test these models for different lighting and environmental conditions to evaluate real-world performance. We have used error-free quantum simulators from pennylane for this study. In the future, we would also like to train these models on physical quantum computers and handle quantum errors, usually present with physical quantum computers.

VII. ACKNOWLEDGEMENTS

This work was supported by the Centre for Connected Multimodal Mobility(C2M2) (the US Department of Transportation Tier 1 University Transportation Centre) headquartered at Clemson University, Clemson, SC, USA. Any Opinions, findings, conclusion, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of C2M2, and the US Government assumes no liability for the contents and use thereof.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[2] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125. Accessed: Aug. 01, 2021. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html

[3] J. Liu and F.-P. An, "Image Classification Algorithm Based on Deep Learning-Kernel Function," *Scientific Programming*, vol. 2020, p. e7607612, Jan. 2020, doi: 10.1155/2020/7607612.

[4] S. Pendleton *et al.*, "Perception, Planning, Control, and Coordination for Autonomous Vehicles," *Machines*, vol. 5, p. 6, Feb. 2017, doi: 10.3390/machines5010006.

[5] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv:1312.6199 [cs]*, Feb. 2014, Accessed: Aug. 01, 2021. [Online]. Available: http://arxiv.org/abs/1312.6199

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582. Accessed: Aug. 01, 2021. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html

[7] X. Wei, S. Liang, N. Chen, and X. Cao, "Transferable Adversarial Attacks for Image and Video Object Detection," *arXiv:1811.12641 [cs]*, May 2019, Accessed: Aug. 01, 2021. [Online]. Available: http://arxiv.org/abs/1811.12641

[8] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018, doi: 10.1109/ACCESS.2018.2807385.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv:1412.6572 [cs, stat]*, Mar. 2015, Accessed: Aug. 01, 2021. [Online]. Available: http://arxiv.org/abs/1412.6572

[10] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial Attacks and Defenses in Deep Learning," *Engineering*, vol. 6, no. 3, pp. 346–360, Mar. 2020, doi: 10.1016/j.eng.2019.12.012.

[11] M. Aprilpyone, Y. Kinoshita, and H. Kiya, "Adversarial Robustness by One Bit Double Quantization for Visual Classification," *IEEE Access*, vol. 7, pp. 177932–177943, 2019, doi: 10.1109/ACCESS.2019.2958358.

[12] P. Panda, I. Chakraborty, and K. Roy, "Discretization Based Solutions for Secure Machine Learning Against Adversarial Attacks," *IEEE Access*, vol. 7, pp. 70157–70168, 2019, doi: 10.1109/ACCESS.2019.2919463.

[13] Z. Liu *et al.*, "Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples." arXiv, Apr. 16, 2019. Accessed: May 21, 2022. [Online]. Available: http://arxiv.org/abs/1803.05787

[14] N. Das *et al.*, "Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression." arXiv, May 08, 2017. Accessed: May 21, 2022. [Online]. Available: http://arxiv.org/abs/1705.02900

[15] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," in *Proceedings 2018 Network and Distributed System Security Symposium*, San Diego, CA, 2018. doi: 10.14722/ndss.2018.23198.

[16] Z. Khan, M. Chowdhury, and S. M. Khan, "A Hybrid Defense Method against Adversarial Attacks on Traffic Sign Classifiers in Autonomous Vehicles," arXiv, arXiv:2205.01225, Apr. 2022. doi: 10.48550/arXiv.2205.01225.

[17] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting Adversarial Samples from Artifacts," arXiv, arXiv:1703.00410, Nov. 2017. doi: 10.48550/arXiv.1703.00410.

[18] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019, doi: 10.1109/TNNLS.2018.2886017.

[19] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks," arXiv, arXiv:1511.04508, Mar. 2016. doi: 10.48550/arXiv.1511.04508.

[20] F. Björck, M. Henkel, J. Stirna, and J. Zdravkovic, "Cyber Resilience – Fundamentals for a Definition," in *New Contributions in Information Systems and Technologies*, Cham, 2015, pp. 311–316. doi: 10.1007/978-3-319-16486-1_31.

[21] "The Future of Quantum Computing," *The Business Standard*, Nov. 25, 2021. https://www.tbsnews.net/tech/future-quantum-computing-334501 (accessed May 21, 2022).
[22] R. B. Griffiths, "Hilbert Space Quantum Mechanics," p. 13.
[23] J. Preskill, "Quantum computing and the entanglement frontier," arXiv, arXiv:1203.5813, Nov. 2012. doi: 10.48550/arXiv.1203.5813.
[24] "Quantum supremacy using a programmable superconducting processor | Nature." https://www.nature.com/articles/s41586-019-1666-5 (accessed May 21, 2022).
[25] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi United Arab Emirates, Apr. 2017, pp. 506–519. doi: 10.1145/3052973.3053009.
[26] O. Soria Emilio, G. Mart¡n Jos, David, M.-S. Marcelino, M.-B. Rafael Jose, and S. L. Jos Antonio, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. IGI Global, 2009.
[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[28] "LISA Traffic Sign Dataset." https://git-disl.github.io/GTDLBench/datasets/lisa_traffic_sign_dataset/ (accessed May 21, 2022).
[29] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, "Hybrid quantum-classical algorithms and quantum error mitigation," *J. Phys. Soc. Jpn.*, vol. 90, no. 3, p. 032001, Mar. 2021, doi: 10.7566/JPSJ.90.032001.
[30] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Phys. Rev. A*, vol. 98, no. 3, p. 032309, Sep. 2018, doi: 10.1103/PhysRevA.98.032309.
[31] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, "A generative modeling approach for benchmarking and training shallow quantum circuits," *npj Quantum Inf*, vol. 5, no. 1, p. 45, Dec. 2019, doi: 10.1038/s41534-019-0157-8.
[32] Z. Khan, S. M. Khan, J. M. Tine, and A. T. Comert, "Hybrid Quantum-Classical Neural Network for Incident Detection," p. 14.
[33] C. Fisher, "IBM | Quantum Computing," 02019-04-02. https://www.ibm.com/quantum-computing// (accessed Aug. 01, 2021).
[34] V. Bergholm *et al.*, "PennyLane: Automatic differentiation of hybrid quantum-classical computations," *arXiv:1811.04968 [physics, physics:quant-ph]*, Feb. 2020, Accessed: Aug. 01, 2021. [Online]. Available: http://arxiv.org/abs/1811.04968
[35] "Cirq," *Google Quantum AI* https://quantumai.google/cirq (accessed Aug. 01, 2021).
[36] R. Majumder *et al.*, "Hybrid Classical-Quantum Deep Learning Models for Autonomous Vehicle Traffic Image Classification Under Adversarial Attack," arXiv, arXiv:2108.01125, Aug. 2021. doi: 10.48550/arXiv.2108.01125.
[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv:1706.06083 [cs, stat]*, Sep. 2019, Accessed: Aug. 01, 2021. [Online]. Available: http://arxiv.org/abs/1706.06083
[38] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.
[39] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, "Transfer learning in hybrid classical-quantum neural networks," *Quantum*, vol. 4, p. 340, Oct. 2020, doi: 10.22331/q-2020-10-09-340.
[40] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, Accessed: Aug. 01, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556
[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[42] G. E. Crooks, "Gates, States, and Circuits," p. 79.
[43] "Operations glossary," *IBM Quantum*. https://quantum-computing.ibm.com/services/programs/docs/runtime/operations_glossary (accessed Oct. 27, 2022).
[44] "qml.operation — PennyLane 0.26.0 documentation." https://docs.pennylane.ai/en/stable/code/qml_operation.html (accessed Oct. 27, 2022).
[45] M. Schuld, A. Bocharov, K. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, Apr. 2018, doi: 10.1103/PhysRevA.101.032308.

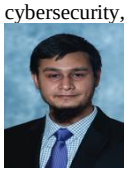
Reek Majumder received his Bachelor's in Technology degree in Computer Science from Kalinga Institute of Industrial Technology (KIIT), Bhubaneshwar, Orissa in 2015. Then he served technology sales and Marketing Intern at a mobility start-up HyHop, Bangalore from 2015 to 2016. Followed by software developer at Cognizant Information Technology from 2016 to 2019. After that, he received his M.Sc. degree in Computer Science (Data Science and Informatics major) from Clemson University, Clemson, U.S.A., in 2021. Currently, pursuing his PhD in civil engineering (transportation major) under the supervision of Dr. Mashrur Chowdhury, Professor, Dept. of Civil Engineering. His primary research focus is connected and autonomous vehicles (CAVs) and Unmanned Aerial Vehicles (UAVs). Within the CAVs and UAVs domain, his research interests are machine/deep learning, quantum machine learning, cloud computing and cybersecurity.

Mashrur "Ronnie" Chowdhury (SM'14) is the Eugene Douglas Mays Professor and Chair of Transportation at Clemson University. He is the director of the USDOT Center for Connected Multimodal Mobility (C2M2) ((http://cecas.clenson.edu/c2m2). He is the co-director of the Complex Systems, Analytics and Visualization Institute (CSAVI) (http://clemson-csavi.org) at Clemson University. He is the director of the Transportation Cyber-Physical Systems Laboratory at Clemson University. He previously served as an elected member of the IEEE ITS Society Board of Governors and is currently a senior member of the IEEE. He is a Fellow of the American Society of Civil Engineers (ASCE) and an alumnus of the National Academy of Engineering (NAE) Frontiers of Engineering program. Dr. Chowdhury is a member of the Transportation Research Board (TRB) Committee on Intelligent Transportation Systems. He is a registered professional engineer in Ohio.

Sakib Mahmud Khan (Member'20) is the Assistant Research Professor at Glenn Department of Civil Engineering, Clemson University, and Assistant Director of the Center for Connected Multimodal Mobility (C2M2). Before joining the center, he was a post-doctoral research scholar working at California Partners for Advanced Transportation Technology (PATH), University of California Berkeley. He received his Ph.D. and M.Sc. in Civil Engineering from Clemson University in 2019 and 2015, respectively.

Zadid Khan received the B.Sc. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2014, and the M.Sc. and Ph.D. degrees in civil engineering (transportation major) from Clemson University, Clemson, SC, USA, in 2018 and 2021, respectively. From 2014 to 2016, he served as a Petroleum Engineer for Chevron in Bangladesh. He is currently working as a Senior Data Analyst in supply chain (transportation) with Walmart Inc. During his M.Sc. and Ph.D. degrees, he worked with the Cyber Physical Systems (CPS.) Laboratory as a Graduate Research Assistant, under the supervision of Dr. Mashrur Chowdhury, Professor, Department of Civil Engineering. His primary research interests include transportation cyber physical systems, connected and autonomous vehicles, and data science and analytics for transportation. Within these domains, his research interests include data science (machine/deep learning), computer communication and networking, data analytics (data

mining, data fusion, big data, and data visualization), cloud computing, cybersecurity, and optimization.



Fahim Ahmed received the M.Sc. degree in Civil engineering and Master of Applied Statistics (MAS) from the University of South Carolina (UofSC) where he is currently a Doctoral Researcher in the SC. Transportation Group. His research focuses on developing mathematical models and solution algorithm for goods movement in a freight network under disruption. His research interest include freight logistics, traffic safety, and pavement systems.

Frank C. Ngeni is a master's degree student in Transportation (MST) at South Carolina State University (SCSU), Orangeburg, USA. He graduated his bachelor's degree in Civil Engineering (2016) at the University of Dar-es-salaam, Tanzania and worked for the Tanzanian National Roads Agency (TANROADS) between 2016 and 2020. Currently, he serves as a graduate research assistant at SCSU with interests in Multimodal mobility, Connected and Automated Vehicles(CAVs), Quantum computing, Transportation planning, Travel demand modeling, Transportation systems analysis, Transportation economics, and Traffic safety.

Gurcan Comert received the B.Sc. and M.Sc. degrees in industrial engineering from Fatih University, Istanbul, Turkey, in 2003 and 2005, respectively, and the Ph.D. degree in civil engineering from the University of South Carolina, Columbia, SC, USA, in 2008. He is a data scientist with Vericast and an associate professor of practice with the Physics and Engineering Department, Benedict College, Columbia. He is also an associate director with the Centre for Connected Multimodal Mobility led by Clemson University and a researcher with the Information Trust Institute, University of Illinois Urbana-Champaign. His research interests include applications of statistical models to problems in different fields, real-time parameter prediction, and stochastic models.

Judith Mwakalonge is an associate professor in the Department of Civil Engineering & Mechanical Engineering Technology and Nuclear Engineering at South Carolina State University. Her research interests are mainly focused on developing solutions for improving safety and efficiency of transportation systems. Specifically, her main research focus includes transportation safety and operations, travel demand modelling and simulation, the impact of distracted biking/walking on transportation safety, and smart mobile applications in transportation. She received her Ph.D. in civil engineering from Tennessee Technological University.

Dr. Dimitra Michalaka is an Associate Professor at the department of civil and environmental engineering at The Citadel and the Associate Director for the Centre for Connected Multimodal Mobility (C2M2). Dr. Michalaka received her undergraduate diploma in civil engineering from the National Technical University of Athens (NTUA), after which she entered the transportation engineering graduate program at University of Florida (UF). She graduated with a Master of Science (M.S) in 2009 and with a Ph.D. in 2012. Her research is primarily focused on Traffic Operations, Traffic Congestion/Relief, Managed Lanes, Microsimulation, Connected and Autonomous Vehicles, and Engineering Education. Dr. Michalaka is a registered Professional Engineer in the state of South Carolina. Also, in December 2020, she graduated with a MS in Project Management from The Citadel.