

Towards Quantum Enhanced Adversarial Robustness with Rydberg Reservoir Learning

Shehbaz Tariq^{1*†}, Muhammad Talha^{2†}, Symeon Chatzinotas¹,
Hyundong Shin²

^{1*}Interdisciplinary Centre for Security, Reliability, and Trust (SnT),
University of Luxembourg, L-1855, Luxembourg City, Luxembourg.

²Department of Electronics and Information Convergence Engineering,
Kyung Hee University, 1732, Deogyeong-daero, Giheung-gu., Yongin-Si,
17104, Korea.

*Corresponding author(s). E-mail(s): shehbaz.tariq@uni.lu;
Contributing authors: talhazakir@khu.ac.kr;
symeon.chatzinotas@uni.lu; hshin@khu.ac.kr;
†These authors contributed equally to this work.

Abstract

Quantum reservoir computing (QRC) leverages the high-dimensional, nonlinear dynamics inherent in quantum many-body systems for extracting spatiotemporal patterns in sequential and time-series data with minimal training overhead. Although QRC inherits the expressive capabilities associated with quantum encodings, recent studies indicate that quantum classifiers based on variational circuits remain susceptible to adversarial perturbations. In this perspective, we investigate the first systematic evaluation of adversarial robustness in a QRC-based learning model. Our reservoir comprises an array of strongly interacting Rydberg atoms governed by a fixed Hamiltonian, which naturally evolves under complex quantum dynamics, producing high-dimensional embeddings. A lightweight multilayer perceptron serves as the trainable readout layer. We utilize the balanced datasets, namely MNIST, FASHION-MNIST, and KUZUSHIJI-MNIST as a benchmark for rigorously evaluating the impact of augmenting the quantum reservoir with an Multilayer perceptron (MLP) in white-box adversarial attacks to assess its robustness. We demonstrate that this approach yields significantly higher accuracy than purely classical models across all perturbation strengths tested. This hybrid approach reveals a new source of quantum advantage and provides practical guidance for the secure deployment of machine learning models on quantum-centric supercomputing with near-term hardware.

Keywords: quantum reservoir computing, quantum machine learning, adversarial robustness, Rydberg Hamiltonian

Introduction

Classical computing architectures are increasingly constrained by the saturation of Moore’s Law, reaching fundamental limits in terms of transistor density, energy efficiency, and linear scalability [1]. Quantum computing has emerged as a promising alternative, offering to overcome these constraints through quantum mechanical principles, including superposition and interference, as well as non-classical correlations such as entanglement [2]. Consequently, quantum algorithms can outperform their classical counterparts in problems such as quantum simulation and specific sampling tasks, with rigorous demonstrations of the quantum advantage achieved in programmable superconducting processors [3, 4] and photonic quantum computers [5]. However, quantum processors remain limited by qubit count, error rates, and the challenges associated with quantum error correction and error mitigation [6]. These challenges limit the scalability of quantum machine learning (QML) models on near-term devices, underscoring the need for architectures that strike a balance between expressiveness and hardware feasibility.

Within QML, variational approaches such as variational quantum circuits (VQCs) have been widely studied across classification, generative modeling, and kernel methods, but suffer from limitations including vanishing gradients, barren plateaus, and the significant overhead of classical optimization loops [7–10]. These limitations motivate exploration of alternative models beyond variational approaches, with quantum reservoir computing (QRC) offering a particularly promising route.

QRC is inspired by classical extreme learning machines such as echo state networks and recurrent neural networks, where data is encoded into the parametrized Hamiltonian of a quantum system and evolves under quantum dynamics [11, 12]. Observables, such as Pauli operators, extract transformed data representations, which are then processed using a classical trainable readout layer. Conceptually, this approach bypasses the optimization bottlenecks of VQCs and avoids barren plateaus. Beyond efficiency, QRC leverages intrinsic physical processes, such as noise and dissipation, as computational resources, thereby enhancing temporal memory and forecasting while preserving universality in approximating fading memory maps. Noise-mapping techniques also enable the precise characterization of circuit dynamics under realistic hardware conditions, ensuring that QRC implementations remain effective on near-term devices [13, 14]. Crucially, the parameters that govern the non-linear interactions and dynamical regime of the reservoir remain fixed during training, bypassing the limitations inherent in variational approaches. This approach excels in dynamical-system-level expressiveness, particularly for tasks based on temporal and sequential data such as time-series prediction, forecasting, and anomaly detection. Recent large-scale implementations on hardware architectures, including Gaussian boson sampling and neutral

atom-based quantum computing platforms, have demonstrated the broad applicability and competitive performance of QRC across machine learning tasks [15, 16].

As AI becomes integral to critical safety applications, from healthcare and autonomous driving to security systems, ensuring robustness against adversarial attacks is paramount [17–19]. In classical machine learning, adversarial examples reliably expose vulnerabilities, sparking extensive defense research [20]. Parallel studies show that quantum classifiers, particularly VQC-based models, also exhibit adversarial fragility, with required perturbation strengths decreasing as system size grows [21–23]. These vulnerabilities persist under both white-box and black-box adversarial attacks, with quantum noise and decoherence offering only limited protection [24, 25]. Furthermore, recent benchmark studies have revealed asymmetric attack transferability, where perturbations crafted on quantum models tend to fool classical networks more effectively than the reverse [26, 27].

Parallel studies have explored defense strategies such as randomized quantum encodings to suppress adversarial gradients [28] and data augmentation to harden quantum kernel methods against input perturbations [29]. However, while the adversarial robustness of QML classifiers has been extensively investigated, the robustness characteristics of QRC models—particularly those employing analog Rydberg-atom implementations—remain largely unexplored. Here, we present the first systematic study of adversarial robustness in a Rydberg-based QRC framework. Specifically, we examine whether augmenting lightweight MLPs with quantum reservoir embeddings enhances resilience against gradient-based perturbations. Under a white-box threat model, where the adversary has full access to model parameters and gradients, robustness is evaluated using three canonical attacks: the Fast Gradient Sign Method (FGSM) [30–32], Projected Gradient Descent (PGD) [32–35], and DeepFool [21, 36, 37]. Empirical evaluations across the MNIST, FASHION-MNIST, and KUZUSHIJI-MNIST benchmarks demonstrate that integrating a Rydberg quantum reservoir with a classical readout consistently enhances adversarial robustness, highlighting a scalable and hardware-realistic pathway for robust quantum learning.

In this context, it is increasingly important to emphasize practical robustness and hardware realizability rather than purely asymptotic arguments. Reservoir-style models offer low-overhead readout and intrinsic echo-state properties that remain stable even under noise, highlighting their suitability for near-term devices. At the same time, theoretical perspectives caution that avoiding issues such as the curse of dimensionality or barren plateaus may confine models to effective subspaces that are classically simulable [10, 38, 39]. These insights strengthen the motivation to frame QRC not around abstract speedups, but as a pathway to robust, tunable, and hardware-realistic models—qualities that our Rydberg-based Hamiltonian implementation seeks to exploit in adversarial settings.

This work addresses this gap by investigating the adversarial robustness of QRC implemented via Rydberg Hamiltonians. We evaluate whether augmenting multilayer perceptrons multilayer perceptron (MLP) with quantum reservoir embeddings enhances resilience to adversarial perturbations, providing the first systematic study

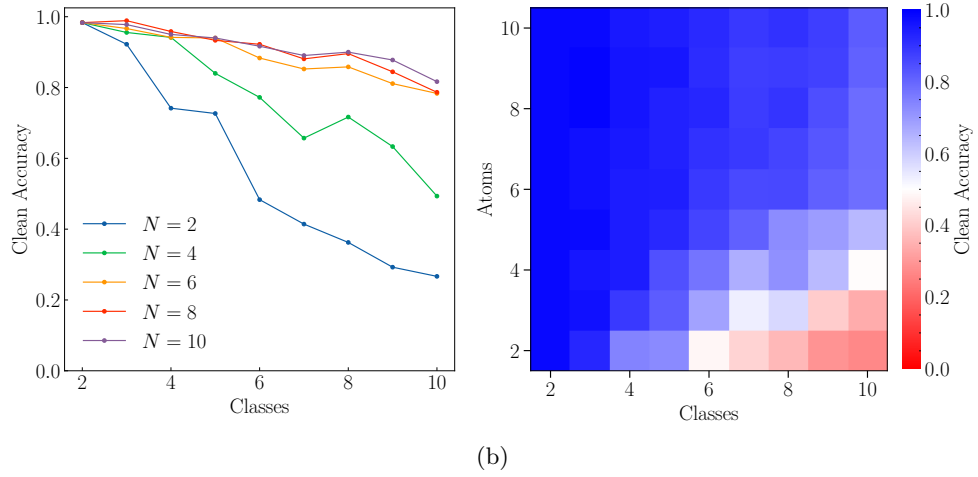
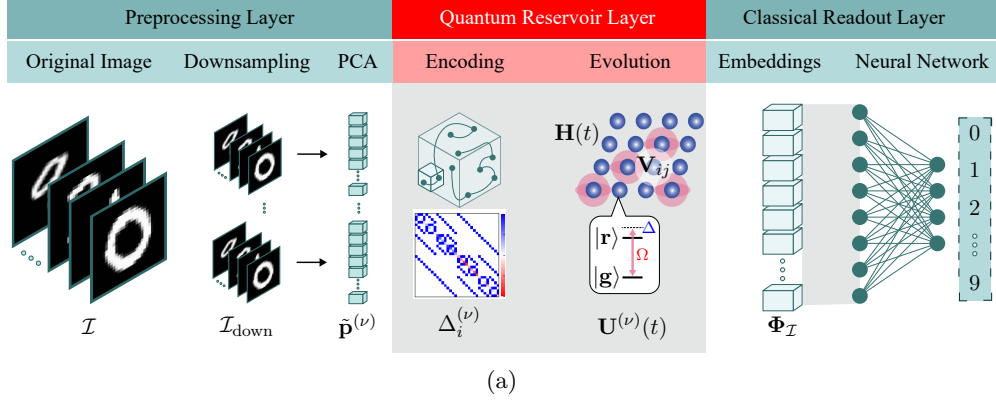


Fig. 1: (a) QRC based learning Framework. (b) Classification performance and robustness of the quantum reservoir on a balanced MNIST subset. Here, we use a balanced MNIST dataset comprising 100 handwritten-digit samples per class (10 classes in total), with a 70%/30% train-test split. We keep $\delta = N$ here. (Left) Clean-test accuracies achieved by a fixed classical readout layer when driven by quantum reservoirs under different configurations (atoms) for N . (Right) Dependence of classification robustness on the dimensionality N of the reservoir’s learning space.

of QRC robustness in adversarial settings [27–29]. Our findings suggest that Rydberg-based QRC provides a hardware-realistic and robust pathway for QML, advancing both theoretical understanding and practical deployment in adversarial settings.

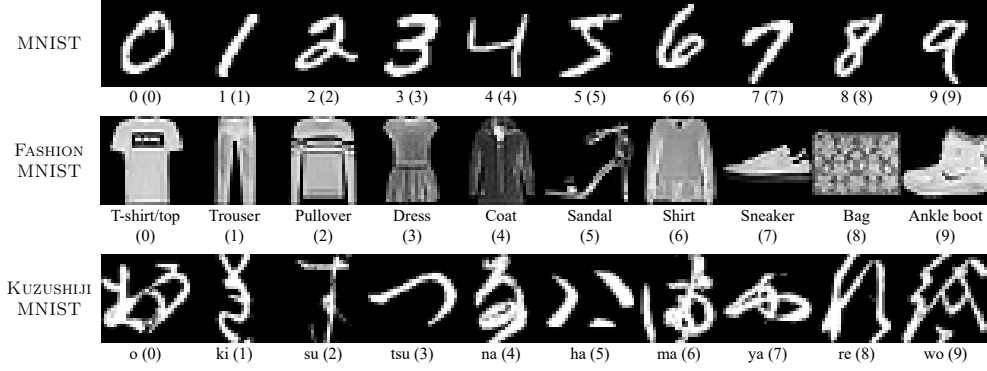


Fig. 2: Sample images from the three benchmark datasets employed in this study: MNIST, Fashion-MNIST, and Kuzushiji-MNIST. Each dataset consists of 10 balanced classes used to rigorously evaluate the adversarial robustness of the quantum reservoir learning model.

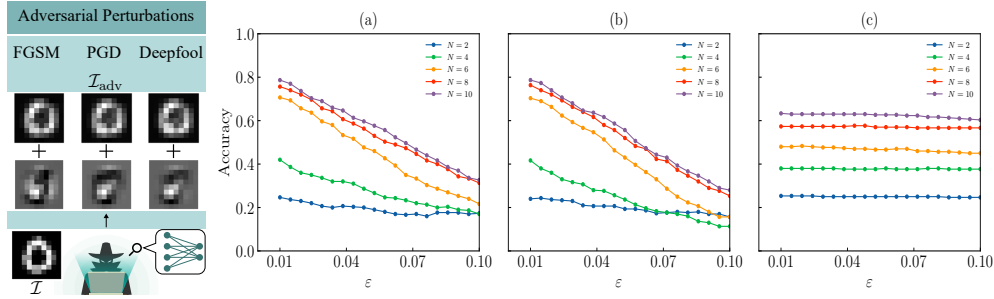


Fig. 3: Classification performance and adversarial robustness of the quantum reservoir on a balanced MNIST subset of all the 10 classes under three attacks. (a) FGSM (b) PGD (c) Deepfool. For all the attacks, we vary the budget $\epsilon \in [0.0, 0.1]$, with 100 gradient steps and perturbation rate of 10^{-3} .

Results

Simulation Setup

The dynamics of the quantum reservoir-based learning are simulated using NVIDIA's CUDA-Q platform [40], which utilizes graphics processing unit (GPU) acceleration for efficient quantum evolution of many-body systems through built-in numerical integrators. The Rydberg atom array is configured with experimentally validated parameters from [15] as shown in Table 1, where we choose a uniform modulation $\alpha_i = 0.15$ for all sites, effectively controlling the strength of the local detuning relative to the global drive. Moreover, the time-dependent Hamiltonian was constructed as a sparse operator with CUDA-optimized evaluation of interaction terms, $\mathbf{V}_{ij}$, parallelized over the

GPU threads. The embedding vector $\Phi(x^{(k)}) \in \mathbb{R}^D$ is passed into a three-layer MLP readout. The network has an input dimension D , a first hidden layer of size 64, a second hidden layer of size 32, and an output layer of size C , which equals the number of classes. Each hidden layer applies a ReLU activation followed by a dropout rate of 10^{-3} . We train only these MLP parameters using the Adam optimizer with learning rate 10^{-3} , batch size 64, and for up to 500 epochs, minimizing the softmax cross-entropy loss. Since the quantum reservoir itself remains fixed, all learnable parameters reside in this lightweight classical readout, keeping training efficient. In this study, we use the balanced MNIST, FASHION-MNIST, and KUZUSHIJI-MNIST datasets comprising 100 samples per class (700 samples for 10 classes total), with a 70%/30% train-test split ratio.

Table 1: Physical parameters for the Rydberg reservoir

Parameter	Symbol	Value	Unit
Number of atoms	N	8	–
Lattice spacing	d	10.0	μm
Interaction coefficient	C_6	$2\pi \times 2000$	$\text{MHz} \cdot \mu\text{m}^6$
Rabi frequency	Ω	$2\pi \times 5$	MHz
Detuning range	$[\Delta_{\min}, \Delta_{\max}]$	$[0, 2\pi \times 10]$	MHz
Local detuning modulation	α	0.15	–
Total evolution time	T	3.0	μs
Time steps	M	6	–
Initial state	$ \psi_0\rangle$	$ +\rangle^{\otimes N}$	–

Encoding Images

Classical image data requires specialized encoding to interface with quantum reservoirs due to dimensionality mismatches between pixel spaces and qubit resources. In this paper, we employ a multi-stage approach:

Image Preprocessing

Original images $\mathcal{I} \in \mathbb{R}^{L \times L}$ are first downsampled to $\mathcal{I}_{\text{down}} \in \mathbb{R}^{S \times S}$ using area interpolation (OpenCV) or Lanczos resampling (PIL). This reduces computational complexity and filters high-frequency noise while preserving structural information:

$$\mathcal{I}_{\text{down}} = \text{resize}(\mathcal{I}, (S, S)) \quad (1)$$

Each downsampled image is decomposed into κ non-overlapping patches $\{\mathbf{p}^{(\nu)}\}_{\nu=1}^{\kappa}$ with $\mathbf{p}^{(\nu)} \in \mathbb{R}^{P^2}$, where P is the patch width; hence $\kappa = (S/P)^2$. Patch extraction enables localized feature analysis, capturing spatial redundancies inherent in natural images:

$$\mathbf{p}^{(\nu)} = \text{extract}_{\nu}(\mathcal{I}_{\text{down}}) \quad (2)$$

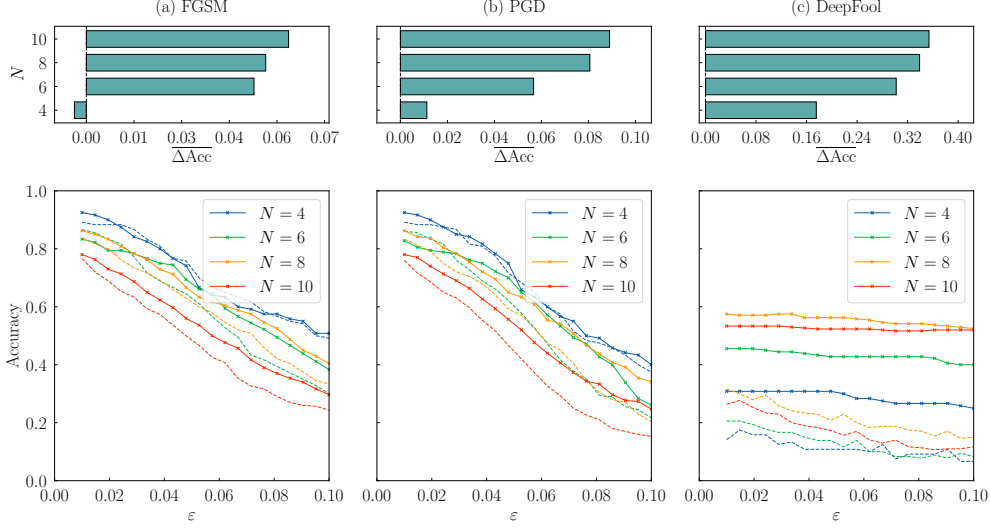


Fig. 4: Adversarial robustness with $\delta = N$ for (a) FGSM, (b) PGD, (c) DeepFool, evaluated on classes $N \in \{4, 6, 8, 10\}$ (atoms = N). Top row (horizontal bars): mean Δ Accuracy per class, $\overline{\Delta \text{Acc}}_N = \frac{1}{|\mathcal{E}|} \sum_{\varepsilon \in \mathcal{E}} (\text{Acc}_N^{\text{QRC+MLP}}(\varepsilon) - \text{Acc}_N^{\text{MLP}}(\varepsilon))$, with $\mathcal{E} \subset [0.0, 0.1]$. Bottom row (line plots): Accuracy vs. ε (solid: QRC+MLP; dashed: MLP). Positive bars indicate enhancement from QRC (larger Δ Accuracy), while negative bars indicate degradation. **Dataset:** MNIST.

Each patch is projected to $\tilde{\mathbf{p}}^{(\nu)} \in \mathbb{R}^\delta$, with $\delta \leq N$ matching the number of atoms. Dimensionality reduction via principal component analysis (PCA) ($\mathbf{W}$) decorrelates patch features and compresses data by retaining maximal variance directions :

$$\tilde{\mathbf{p}}^{(\nu)} = \mathbf{W}^\top (\mathbf{p}^{(\nu)} - \mu) \quad (3)$$

Here $\mathbf{W} \in \mathbb{R}^{P^2 \times \delta}$ contains the principal components and μ is the global patch mean. The projection matrix solves the eigenproblem

$$\sigma \mathbf{W} = \mathbf{W} \mathbf{\Lambda}, \quad \sigma = \frac{1}{\kappa} \sum_{\nu=1}^{\kappa} (\mathbf{p}^{(\nu)} - \mu) (\mathbf{p}^{(\nu)} - \mu)^\top \quad (4)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_\delta)$ contains the largest δ eigenvalues chosen such that $\frac{\sum_{i=1}^{\delta} \lambda_i}{\sum_{i=1}^{P^2} \lambda_i} > \mu$, thus preserves at least a fraction μ of the total variance. This variance-retention criterion, standard in dimensionality reduction, ensures that the compressed representation remains information-rich while discarding redundant components.

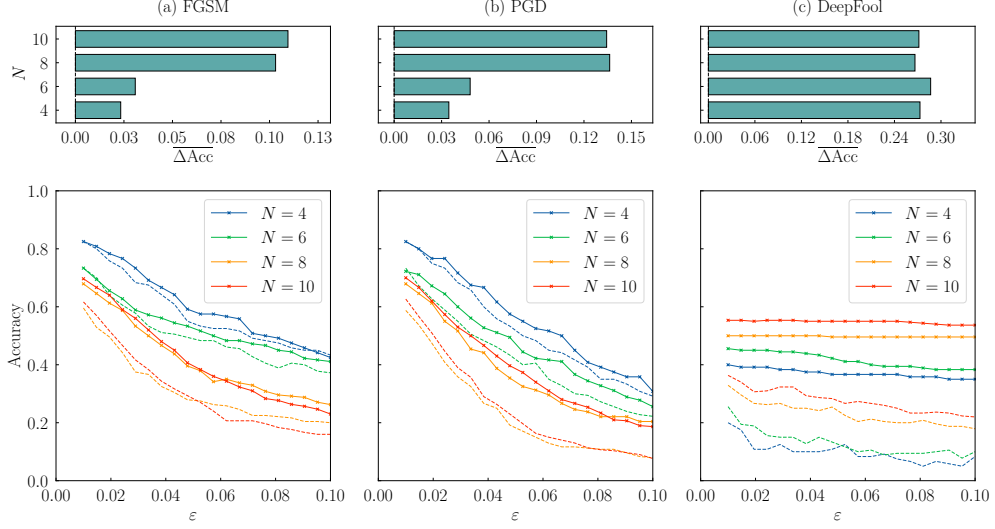


Fig. 5: Adversarial robustness with $\delta = N$ for (a) FGSM, (b) PGD, (c) DeepFool, evaluated on classes $N \in \{4, 6, 8, 10\}$ (atoms = N). Top row (horizontal bars): mean Δ Accuracy per class, $\overline{\Delta \text{Acc}}_N = \frac{1}{|\mathcal{E}|} \sum_{\varepsilon \in \mathcal{E}} (\text{Acc}_N^{\text{QRC+MLP}}(\varepsilon) - \text{Acc}_N^{\text{MLP}}(\varepsilon))$, with $\mathcal{E} \subset [0.0, 0.1]$. Bottom row (line plots): Accuracy vs. ε (solid: QRC+MLP; dashed: MLP). Positive bars indicate enhancement from QRC (larger Δ Accuracy), while negative bars indicate degradation. **Dataset:** FASHION-MNIST.

Quantum Encoding

Compressed features are mapped to detunings:

$$\Delta_i^{(\nu)} = \Delta_{\min} + \alpha_i^{(\nu)} (\Delta_{\max} - \Delta_{\min}), \quad (5)$$

$$\alpha_i^{(\nu)} = \frac{\tilde{p}_i^{(\nu)} - \min_{\nu'} \tilde{p}_i^{(\nu')}}{\max_{\nu'} \tilde{p}_i^{(\nu')} - \min_{\nu'} \tilde{p}_i^{(\nu')}}. \quad (6)$$

The Hamiltonian for patch ν is

$$\mathbf{H}^{(\nu)}(t) = \sum_{i=1}^N \frac{\Omega(t)}{2} \sigma_i^x - \sum_{i=1}^N \Delta_i^{(\nu)} \hat{\mathbf{n}}_i + \sum_{i < j} \mathbf{V}_{ij} \hat{\mathbf{n}}_i \hat{\mathbf{n}}_j, \quad (7)$$

where $\hat{\mathbf{n}}_i = \frac{1}{2}(\mathbf{I} - \sigma_i^z)$. Each patch undergoes independent reservoir evolution. Observables are sampled at M time points $\{t_m\}_{m=1}^M$, yielding

$$\Phi^{(\nu)}(t_m) = \left(\langle \sigma_1^z \rangle_{t_m}^{(\nu)}, \dots, \langle \sigma_N^z \rangle_{t_m}^{(\nu)}, \langle \sigma_1^z \sigma_2^z \rangle_{t_m}^{(\nu)}, \dots, \langle \sigma_{N-1}^z \sigma_N^z \rangle_{t_m}^{(\nu)} \right)^\top. \quad (8)$$

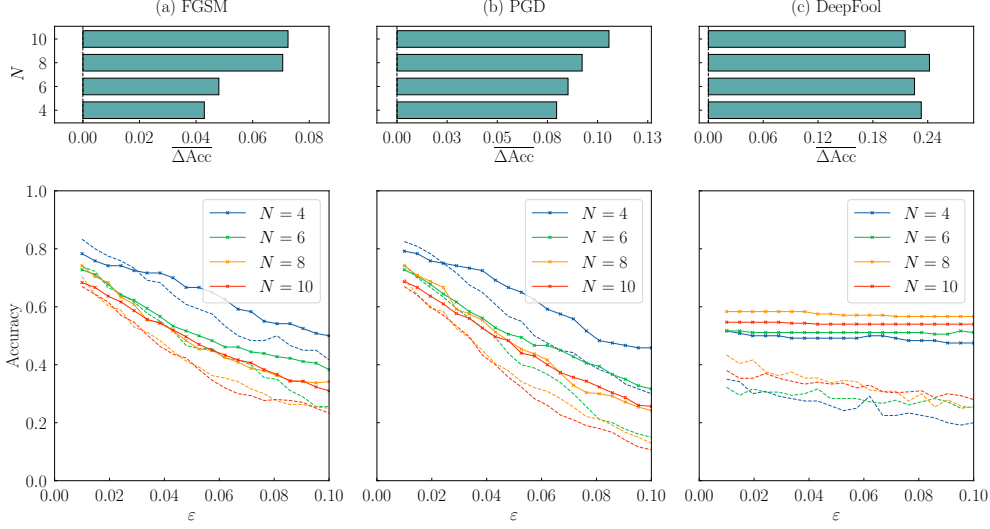


Fig. 6: Adversarial robustness with $\delta = N$ for (a) FGSM, (b) PGD, (c) DeepFool, evaluated on classes $N \in \{4, 6, 8, 10\}$ (atoms = N). Top row (horizontal bars): mean Δ Accuracy per class, $\overline{\Delta\text{Acc}}_N = \frac{1}{|\mathcal{E}|} \sum_{\varepsilon \in \mathcal{E}} (\text{Acc}_N^{\text{QRC+MLP}}(\varepsilon) - \text{Acc}_N^{\text{MLP}}(\varepsilon))$, with $\mathcal{E} \subset [0.0, 0.1]$. Bottom row (line plots): Accuracy vs. ε (solid: QRC+MLP; dashed: MLP). Positive bars indicate enhancement from QRC (larger Δ Accuracy), while negative bars indicate degradation. **Dataset:** KUZUSHIJI-MNIST.

Concatenating all M snapshots gives the patch embedding

$$\Phi^{(\nu)} = (\Phi^{(\nu)}(t_1)^\top | \dots | \Phi^{(\nu)}(t_M)^\top)^\top \in \mathbb{R}^\phi, \quad \phi = M \left(N + \binom{N}{2} \right). \quad (9)$$

The image-level representation is obtained by patch averaging:

$$\Phi_{\mathcal{I}} = \frac{1}{\kappa} \sum_{\nu=1}^{\kappa} \Phi^{(\nu)}. \quad (10)$$

This hierarchical encoding leverages quantum dynamics for intra-patch feature extraction, while utilizing classical aggregation for inter-patch dimensionality reduction.

Adversarial Attacks

Adversarial attacks craft worst-case inputs that force a trained model to misclassify. In this paper, we focus on three canonical white-box attacks, namely, FGSM, PGD, and DeepFool, that assume full access to network gradients. Each method returns an adversarial image $\mathcal{I}_{\text{adv}}$ whose perceptual distance from the original image $\mathcal{I}$ does not exceed a chosen attack *budget* ε . In this paper, we vary the $\varepsilon \in [0.0, 0.1]$, with 100 gradient steps and a learning rate of 10^{-3} .

FGSM

FGSM is a foundational adversarial technique that generates perturbed inputs by leveraging the direction of steepest ascent in the model’s loss landscape. It operates by applying a single targeted distortion to the input in the direction that most increases prediction error, thus inducing misclassification with minimal computational effort. This method exploits local linearity in neural networks and serves as an efficient diagnostic of model vulnerability. In the context of this study, FGSM serves as a benchmark for baseline robustness, where quantum reservoir models consistently enhance their classical counterparts by resisting shallow gradient-based manipulations, indicating a fundamental mismatch between classical perturbation geometry and quantum feature encoding. FGSM linearises the loss $\mathcal{L}(\mathcal{I}, y)$ around the clean image and moves ε in the direction that maximally increases the loss in ℓ_∞ geometry:

$$\mathcal{I}_{\text{adv}} = \mathcal{I} + \varepsilon \text{sign}(\nabla_{\mathcal{I}} \mathcal{L}(\mathcal{I}, y)), \quad \|\mathcal{I}_{\text{adv}} - \mathcal{I}\|_\infty \leq \varepsilon. \quad (11)$$

A single forward-backward pass suffices, making FGSM a lightweight “stress test”. Budgets in image classification typically span $\varepsilon \in \{8/255, 16/255\}$ for inputs scaled to $[0, 1]$.

PGD

PGD extends the principle of FGSM into a more aggressive and iterative regime. It applies repeated, controlled perturbations that stay within a bounded region, systematically exploring the model’s decision boundaries to expose deeper vulnerabilities. Each iteration applies a limited displacement in the gradient direction, maintaining control over the perturbation’s size at every step. It is widely regarded as a rigorous adversarial test due to its ability to converge on high-impact inputs. Within this study, PGD is used to assess the stability of quantum and classical models under sustained adversarial perturbation. PGD refines FGSM into an iterative constrained optimization that maximizes the loss over the ℓ_∞ ball of budget ε . From an initial point $\mathcal{I}_0$ (often random inside the ball), it performs T steps of size ζ :

$$\mathcal{I}_{t+1} = \text{Proj}_{\|\eta\|_\infty \leq \varepsilon} \left(\mathcal{I}_t + \zeta \text{sign}(\nabla_{\mathcal{I}_t} \mathcal{L}(\mathcal{I}_t, y)) \right), \quad t = 0, \dots, T-1. \quad (12)$$

For step sizes $\zeta \approx 2/255$ and $T \approx 40$, PGD closely approximates the worst-case ℓ_∞ perturbation and is the de-facto benchmark for robustness studies.

DeepFool

DeepFool is a geometry-driven attack that seeks to identify the closest point at which an input crosses the decision boundary of a classifier, producing adversarial examples with minimal perceptual distortion. It operates by incrementally adjusting the input until it reaches the region of misclassification, effectively modeling the local topology of the decision boundary. This attack provides insight into the structural vulnerability of models. More specifically, it seeks the smallest ℓ_2 perturbation that crosses the

current decision boundary. At each iteration, the classifier is locally linearized into a hyperplane, and the input is nudged orthogonally toward it:

$$\boldsymbol{\eta}^* = -\frac{f(\mathcal{I})}{\|\nabla f(\mathcal{I})\|_2^2} \nabla f(\mathcal{I}), \quad \mathcal{I} \leftarrow \mathcal{I} + \boldsymbol{\eta}^*, \quad (13)$$

where f is the signed decision function. Iterations stop once the predicted label changes. To enforce the attack budget ε , the final perturbation is rescaled if necessary:

$$\boldsymbol{\eta} = \min\left\{1, \frac{\varepsilon}{\|\boldsymbol{\eta}^*\|_2}\right\} \boldsymbol{\eta}^*, \quad \|\boldsymbol{\eta}\|_2 \leq \varepsilon, \quad \mathcal{I}_{\text{adv}} = \mathcal{I} + \boldsymbol{\eta}. \quad (14)$$

DeepFool typically converges in fewer than ten iterations and yields quasi-imperceptible ℓ_2 perturbations.

Discussion

This work provides a systematic investigation of adversarial robustness in quantum reservoir learning. The results show that augmenting a classical MLP with a Rydberg-based quantum reservoir improves both clean and adversarial accuracies. By demonstrating enhanced performance without requiring variational training, the proposed framework establishes Rydberg reservoirs as practical and scalable components for robust QML on near-term quantum processors. In particular, we investigated the robustness of the proposed QRC–MLP architecture across varying reservoir dimensions N . For each configuration, PCA was applied to project input images into $\delta = N$ principal components, ensuring feature–reservoir dimensional consistency. As shown in Fig. 1b, larger reservoirs consistently achieved higher clean accuracies, indicating that the Rydberg-based reservoir enriches the nonlinear feature space accessible to the classical readout. Under the constant detuning (CD) encoding scheme—where each feature corresponds to one atom—the condition $N \geq \delta$ ensures sufficient expressiveness and provides a clear baseline for assessing scalability.

Adversarial evaluations under FGSM, PGD, and DeepFool attacks further confirm the benefit of reservoir augmentation. In Figs. 4–6, the hybrid QRC–MLP model maintains higher accuracy than the purely classical MLP across all perturbation budgets. The robustness improvement, expressed as the mean accuracy enhancement

$$\overline{\Delta \text{Acc}}_N = \frac{1}{|\mathcal{E}|} \sum_{\varepsilon \in \mathcal{E}} \left(\text{Acc}_N^{\text{QRC+MLP}}(\varepsilon) - \text{Acc}_N^{\text{MLP}}(\varepsilon) \right), \quad (15)$$

$$\text{Acc}_N^{(\cdot)}(\varepsilon) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \mathbb{I}\left\{\hat{y}_{N,\varepsilon}^{(\cdot)}(x) = y\right\}, \quad (16)$$

increases with N , demonstrating that the high-dimensional embeddings generated by the reservoir are less susceptible to gradient-based perturbations. Across all attack methods, configurations with $N \geq \delta$ sustain higher accuracies throughout the entire perturbation range $\varepsilon \in [0, 0.1]$, confirming that Rydberg interactions play a decisive role in shaping adversarial robustness.

Although reservoir augmentation provides a clear advantage, the overall robustness depends strongly on the design of the quantum reservoir. The initialization state ψ_0 , the choice of observables $\hat{\mathbf{O}}$, and the encoding function that maps classical data to quantum detunings collectively determine the expressive capacity of the system. These factors directly influence how information is distributed within the reservoir and how effectively temporal correlations are captured. Moreover, realistic evaluations must include the impact of decoherence, parameter drift, and cross-talk in actual Rydberg hardware. Incorporating these noise effects in future studies will be essential to accurately reflect performance on near-term devices.

This work also highlights that increasing the reservoir dimension N enhances robustness but introduces practical trade-offs in qubit resources and measurement overhead. Despite these challenges, the proposed Rydberg-based QRC architecture remains attractive because it achieves robustness without requiring gradient-based optimization. This property makes it especially suited for noisy intermediate-scale quantum (NISQ) systems, where stability and resource efficiency are critical for implementation feasibility.

In future studies we aim to extend the current framework toward more diverse data modalities, such as spatiotemporal data and time series, while exploring systematic methods for tuning Hamiltonian parameters to further improve performance. Additionally, selecting appropriate observables and refining encoding strategies will help maximize information extraction and minimize redundancy. Experimental validation on real quantum hardware, such as neutral-atom or superconducting platforms under realistic noise and sampling constraints, will be a crucial next step in verifying the robustness advantages observed in simulation. Moreover, incorporating realistic noise models and extending the reservoir-augmentation framework to advanced architectures such as transformers will be vital for translating these quantum advantages into practical, robust machine learning systems.

Methods

Quantum Reservoir Computing

QRC builds upon the principles of classical reservoir computing by exploiting the high-dimensional dynamics and quantum parallelism inherent in quantum systems for spatiotemporal representation learning [41, 42]. The QRC based learning framework is illustrated in Fig. 1a. Unlike quantum neural networks, which require extensive training across multiple layers of adjustable parameters, QRC utilizes a fixed, untrained quantum reservoir. The reservoir functions as a natural feature map, where the input data is encoded in the Hamiltonian of the system, and the resulting quantum dynamics transforms the input into a non-linear high-dimensional representation [15]. QRC requires no iterative training within the reservoir, restricting the learning complexity to optimizing only the readout layer and greatly simplifying implementation on NISQ devices, where resource constraints and noise hinder large-scale parameter tuning. Adaptation to a specific learning task is effected through an offline configuration of the reservoir’s dynamics to tune Hamiltonian parameters by using genetic or other meta-optimization algorithms[43]. This approach is particularly well-suited

for extracting temporal patterns and sequential data, akin to classical recurrent neural networks (RNN)-based echo-state networks or long short-term memory (LSTM). Various physical implementations have been explored, including coherently coupled quantum oscillators [44], Rydberg atom arrays or neutral-atom quantum computing systems [45], superconducting quantum devices [46], and Gaussian boson samplers [16]. In general, the system Hamiltonian is modeled as

$$\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_{\text{int}} + \mathbf{H}_{\text{drive}}(t), \quad (17)$$

where $\mathbf{H}_0$ represents the internal energy, $\mathbf{H}_{\text{int}}$ encodes interactions between constituents, and $\mathbf{H}_{\text{drive}}(t)$ incorporates the data as time-dependent driving fields. The temporal evolution of the reservoir provides a rich, non-linear mapping from the input data space to a high-dimensional Hilbert space. The classical features for downstream tasks are extracted through local observables (e.g., Pauli operators). The reservoir Hamiltonian is initially designed to meet the constraints of quantum hardware, such as the neural atom-based quantum computer [15], and can be optionally tuned with offline meta-optimization; Once configured, it remains unchanged, and all subsequent batched or online learning is applied exclusively to the classical readout layer, preserving the original dynamics of the reservoir. In the subsequent section, we model the end-to-end framework used in this study, based on the Rydberg Hamiltonian described in [15].

Rydberg Hamiltonian

Consider a system of N neutral atoms arranged in a one-dimensional lattice, each modeled as a two-level system with ground state $|g\rangle$ and Rydberg excited state $|r\rangle$. The many-body dynamics under Rydberg blockade are governed by the time-dependent Hamiltonian (with $\hbar = 1$):

$$\mathbf{H}^{(k)}(t) = \sum_{i=1}^N \frac{\Omega(t)}{2} \sigma_i^x - \sum_{i=1}^N \alpha_i \Delta_i^{(k)} \hat{\mathbf{n}}_i + \sum_{i<j} \mathbf{V}_{ij} \hat{\mathbf{n}}_i \hat{\mathbf{n}}_j, \quad (18)$$

where $\sigma_i^x = |g_i\rangle\langle r_i| + |r_i\rangle\langle g_i|$ is the Pauli- x operator for atom i , and $\hat{\mathbf{n}}_i = \frac{1}{2}(\mathbf{I} - \sigma_i^z)$ is the projector onto the Rydberg state. Here, $\alpha_i \in [0, 1]$ is a site-dependent modulation factor that scales the feature-encoded detuning $\Delta_i^{(k)}$ at atom i . The interaction coefficients $\mathbf{V}_{ij} = C_6/|\mathbf{r}_i - \mathbf{r}_j|^6$ represent van der Waals interactions between atoms i and j . Given a data set of n samples with feature vectors $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_N^{(k)})^\top \in \mathbb{R}^N$, we employ CD encoding by mapping each feature to a static detuning via min-max normalization:

$$\Delta_i^{(k)} = \Delta_{\min} + \left(\frac{x_i^{(k)} - x_{\min}}{x_{\max} - x_{\min}} \right) (\Delta_{\max} - \Delta_{\min}), \quad (19)$$

where $x_{\min} = \min_{i,k} x_i^{(k)}$ and $x_{\max} = \max_{i,k} x_i^{(k)}$ are computed across the data set. The range $[\Delta_{\min}, \Delta_{\max}]$ defines the bounds of the applied detunings. This encoding

results in sample-specific, time-independent Hamiltonians where the input features are embedded as local energy shifts. The Rabi frequency $\Omega(t)$ quantifies the rate of coherent population oscillations between the ground state $|g\rangle$ and the Rydberg state $|r\rangle$, which control the quantum dynamics of the reservoir through external laser fields that set both the global drive strength and the site-dependent detunings. Both the Rabi frequency $\Omega(t)$ and the site-dependent detunings $\Delta_i^{(k)}$ are controlled by laser parameters: $\Omega(t)$ corresponds to the time-dependent laser amplitude driving coherent Rabi oscillations, while each $\Delta_i^{(k)}$ corresponds to the local detuning of the laser from resonance at atom i . The drive $\Omega(t)$ is applied globally and kept fixed across samples, whereas the detunings $\Delta_i^{(k)}$ encode input-dependent information and remain static during evolution.

Constructing Embeddings

The reservoir dynamics for each encoded sample $\mathbf{x}^{(k)}$ are generated by evolving the system under the sample-dependent Hamiltonian $\mathbf{H}^{(k)}(t)$ from t_0 to t_{end} . The evolution is governed by the unitary

$$\mathbf{U}^{(k)}(t) = \mathcal{T} \exp \left[-i \int_{t_0}^t \mathbf{H}^{(k)}(\tau) d\tau \right], \quad (20)$$

where $\mathcal{T}$ denotes time-ordering, and the time-evolved state is $|\psi^{(k)}(t)\rangle = \mathbf{U}^{(k)}(t) |\psi_0\rangle$, with $|\psi_0\rangle$ the fixed initial reservoir state. Selecting the initial state $|\psi_0\rangle$ —whether a ground state, a superposition state, or with tailored pre-evolution optimization adapted to the given task—serves as a hyperparameter that can markedly influence the reservoir’s memory capacity and nonlinearity [47]. To extract classical features we measure, at discrete times $\{t_m\}_{m=1}^M$ spanning $[t_0, t_{\text{end}}]$, the set of local Pauli- z operators and their pairwise correlations:

$$\hat{\mathbf{O}} \in \left\{ \sigma_i^z \mid 1 \leq i \leq N \right\} \cup \left\{ \sigma_i^z \sigma_j^z \mid 1 \leq i < j \leq N \right\}, \quad (21)$$

where $\sigma_i^z = |g_i\rangle\langle g_i| - |r_i\rangle\langle r_i|$ acts on atom i . These operators relate directly to the Rydberg occupations via $\hat{\mathbf{n}}_i = \frac{1}{2}(\mathbf{I} - \sigma_i^z)$ and $\hat{\mathbf{n}}_i \hat{\mathbf{n}}_j = \frac{1}{4}(\mathbf{I} - \sigma_i^z - \sigma_j^z + \sigma_i^z \sigma_j^z)$. The measurement instants are uniformly spaced with step size Δt :

$$t_m = t_0 + m \Delta t, \quad m = 1, \dots, M, \quad \Delta t = \frac{t_{\text{end}} - t_0}{M}. \quad (22)$$

At each t_m we evaluate all $R = N + \binom{N}{2}$ observables, yielding the expectation values

$$\langle \sigma_i^z \rangle_{t_m}^{(k)} = \left\langle \psi^{(k)}(t_m) \left| \sigma_i^z \right| \psi^{(k)}(t_m) \right\rangle, \quad (23)$$

$$\langle \sigma_i^z \sigma_j^z \rangle_{t_m}^{(k)} = \left\langle \psi^{(k)}(t_m) \left| \sigma_i^z \sigma_j^z \right| \psi^{(k)}(t_m) \right\rangle. \quad (24)$$

Concatenating all measurements forms the embedding vector

$$\mathbf{\Phi}(\mathbf{x}^{(k)}) = \left(\langle \sigma_1^z \rangle_{t_1}^{(k)}, \dots, \langle \sigma_N^z \rangle_{t_1}^{(k)}, \langle \sigma_1^z \sigma_2^z \rangle_{t_1}^{(k)}, \dots, \langle \sigma_{N-1}^z \sigma_N^z \rangle_{t_1}^{(k)}, \right. \\ \left. \langle \sigma_1^z \rangle_{t_2}^{(k)}, \dots, \langle \sigma_{N-1}^z \sigma_N^z \rangle_{t_2}^{(k)}, \dots, \langle \sigma_1^z \rangle_{t_M}^{(k)}, \dots, \langle \sigma_{N-1}^z \sigma_N^z \rangle_{t_M}^{(k)} \right)^\top, \quad (25)$$

whose dimension is $D = M(N + \binom{N}{2})$. This spatiotemporal feature vector serves as input to the classical readout layer.

Data availability

The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

Code availability

The source code used to support the findings of this study is available from the corresponding author upon reasonable request.

References

- [1] Theis, T.N., Wong, H.-S.P.: The end of Moore’s law: A new beginning for information technology. *IEEE Comput. Sci. Eng.* **19**(2), 41–50 (2017)
- [2] Boixo, S., Isakov, S.V., Smelyanskiy, V.N., Babbush, R., Ding, N., Jiang, Z., Bremner, M.J., Martinis, J.M., Neven, H.: Characterizing quantum supremacy in near-term devices. *Nat. Phys.* **14**(6), 595–600 (2018)
- [3] Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J.C., Barends, R., Biswas, R., Boixo, S., Brandao, F.G.S.L., Buell, D.A.: Quantum supremacy using a programmable superconducting processor. *Nature* **574**(7779), 505–510 (2019)
- [4] Kim, Y., Eddins, A., Anand, S., Wei, K.X., Berg, E.V.D., Rosenblatt, S., Nayfeh, H., Wu, Y., Zaletel, M., Temme, K.: Evidence for the utility of quantum computing before fault tolerance. *Nature* **618**(7965), 500–505 (2023)
- [5] Zhong, H.-S., Wang, H., Deng, Y.-H., Chen, M.-C., Peng, L.-C., Luo, Y.-H., Qin, J., Wu, D., Ding, X., Hu, Y.: Quantum computational advantage using photons. *Science* **370**(6523), 1460–1463 (2020)
- [6] Preskill, J.: Quantum computing in the NISQ era and beyond. *Quantum* **2**(0), 79 (2018)
- [7] McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**(1), 4812 (2018)
- [8] Symons, B.C., Galvin, D., Sahin, E., Alexandrov, V., Mensa, S.: A practitioner’s guide to quantum algorithms for optimisation problems. *J. Phys. A: Math. and Theor.* **56**(45), 453001 (2023)
- [9] Schuld, M., Killoran, N.: Is quantum advantage the right goal for quantum machine learning? *PRX Quantum* **3**(3), 030101 (2022)
- [10] Cerezo, M., Larocca, M., García-Martín, D., Diaz, N.L., Braccia, P., Fontana, E., Rudolph, M.S., Bermejo, P., Ijaz, A., Thanasilp, S.: Does provable absence of barren plateaus imply classical simulability? *Nat. Commun.* **16**(1), 7907 (2025)
- [11] Nokkala, J., Martínez-Peña, R., Giorgi, G.L., Parigi, V., Soriano, M.C., Zambrini, R.: Gaussian states of continuous-variable quantum systems provide universal and versatile reservoir computing. *Commun. Phys.* **4**(1), 53 (2021)
- [12] Innocenti, L., Lorenzo, S., Palmisano, I., Ferraro, A., Paternostro, M., Palma, G.M.: Potential and limitations of quantum extreme learning machines. *Commun. Phys.* **6**(1), 118 (2023)

- [13] Sannia, A., Martínez-Peña, R., Soriano, M.C., Giorgi, G.L., Zambrini, R.: Dissipation as a resource for quantum reservoir computing. *Quantum* **8**, 1291 (2024)
- [14] Gelman, M.: A survey of methods for mitigating barren plateaus for parameterized quantum circuits. *arXiv:2406.14285* (2024)
- [15] Kornjača, M., Hu, H.-Y., Zhao, C., Wurtz, J., Weinberg, P., Hamdan, M., Zhdanov, A., Cantu, S.H., Zhou, H., Bravo, R.A., et al.: Large-scale quantum reservoir learning with an analog quantum computer. *arXiv:2407.02553* (2024)
- [16] Cimini, V., Sohoni, M.M., Presutti, F., Malia, B.K., Ma, S.-Y., Yanagimoto, R., Wang, T., Onodera, T., Wright, L.G., McMahon, P.L.: Large-scale quantum reservoir computing using a gaussian boson sampler. *arXiv:2505.13695* (2025)
- [17] Yu, K.-H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**(10), 719–731 (2018)
- [18] Muhammad, K., Ullah, A., Lloret, J., Ser, J.D., Albuquerque, V.H.C.D.: Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Trans. Intell. Transp. Syst.* **22**(7), 4316–4336 (2020)
- [19] Sarker, I.H., Furhad, M.H., Nowrozy, R.: AI-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Comput. Sci.* **2**(3), 173 (2021)
- [20] Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. *arXiv:1611.01236* (2016)
- [21] West, M.T., Erfani, S.M., Leckie, C., Sevier, M., Hollenberg, L.C., Usman, M.: Benchmarking adversarially robust quantum machine learning at scale. *Phys. Rev. Res.* **5**(2), 023186 (2023)
- [22] Lu, S., Duan, L.-M., Deng, D.-L.: Quantum adversarial machine learning. *Phys. Rev. Res.* **2**(3), 033212 (2020)
- [23] West, M.T., Tsang, S.-L., Low, J.S., Hill, C.D., Leckie, C., Hollenberg, L.C., Erfani, S.M., Usman, M.: Towards quantum enhanced adversarial robustness in machine learning. *Nat. Mach. Intell.* **5**(6), 581–589 (2023)
- [24] Wendlinger, M., Tschärke, K., Debus, P.: A comparative analysis of adversarial robustness for quantum and classical machine learning models. In: *Proceedings of 2024 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2024)
- [25] Yocam, E., Rizi, A., Kamepalli, M., Vaidyan, V., Wang, Y., Comert, G.: Quantum Adversarial Machine Learning and Defense Strategies: Challenges and

Opportunities. arXiv:2412.12373 (2024)

- [26] West, M.T., Erfani, S.M., Leckie, C., Sevier, M., Hollenberg, L.C.L., Usman, M.: Benchmarking Adversarially Robust Quantum Machine Learning at Scale. *Phys. Rev. Res.* **5**(2), 023186 (2023)
- [27] Maouaki, W.E., Marchisio, A., Said, T., Shafique, M., Bennai, M.: Robqunns: A methodology for robust quanvolutional neural networks against adversarial attacks. In: *Proceedings of 2024 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)* (2024)
- [28] Gong, W., Yuan, D., Li, W., Deng, D.-L.: Enhancing quantum adversarial robustness by randomized encodings. *Phys. Rev. Res.* **6**(2), 023020 (2024)
- [29] Montalbano, G., Banchi, L.: Quantum adversarial learning for kernel methods. *Quantum Mach. Intell.* **7**(1), 15 (2025)
- [30] Maouaki, W.E., Marchisio, A., Said, T., Shafique, M., Bennai, M.: Robqunns: A methodology for robust quanvolutional neural networks against adversarial attacks. In: *Proceedings of 2024 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, pp. 4090–4095 (2024)
- [31] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv:1412.6572 (2015)
- [32] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 (2019)
- [33] Yocam, E., Rizi, A., Kamepalli, M., Vaidyan, V., Wang, Y., Comert, G.: Quantum adversarial machine learning and defense strategies: Challenges and opportunities. arXiv:2412.12373 (2024)
- [34] West, M.T., Nakhl, A.C., Heredge, J., Creevey, F.M., Hollenberg, L.C., Sevier, M., Usman, M.: Drastic circuit depth reductions with preserved adversarial robustness by approximate encoding for quantum machine learning. *Intell. Comput.* **3**, 0100 (2024)
- [35] Majumder, R., Chowdhury, M., Khan, S.M., Khan, Z., Ahmad, F., Ngeni, F., Comert, G., Mwakalonge, J., Michalaka, D.: Quantum computing supported adversarial attack-resilient autonomous vehicle perception module for traffic sign classification. arXiv:2504.12644 (2025)
- [36] Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2574–2582 (2016)

- [37] Tursynbek, N., Petiushko, A., Oseledets, I.: Geometry-inspired top-k adversarial perturbations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3398–3407 (2022)
- [38] Jing, Y., Wang, P., Zhang, S., Zeng, Z., Liang, S.-J., Chen, W.: Quantum transport reservoir computing. arXiv:2509.07778 (2025)
- [39] Lorenzis, A.D., Casado, M.P., Gullo, N.L., Lux, T., Plastina, F., Riera, A.: Behind the scenes of the quantum extreme learning machines. arXiv:2509.06873 (2025)
- [40] Bayraktar, H., Charara, A., Clark, D., Cohen, S., Costa, T., Fang, Y.-L.L., Gao, Y., Guan, J., Gunnels, J., Haidar, A., *et al.*: cuquantum sdk: A high-performance library for accelerating quantum science. In: Proceedings of 2023 IEEE International Conference on Quantum Computing and Engineering (QCE), IEEE (2023)
- [41] Zhu, C., Ehlers, P.J., Nurdin, H.I., Soh, D.: Practical and scalable quantum reservoir computing. arXiv:2405.04799 (2024)
- [42] Jang, Y.H., Lee, S.H., Han, J., Kim, W., Shim, S.K., Cheong, S., Woo, K.S., Han, J.-K., Hwang, C.S.: Spatiotemporal data processing with memristor crossbar-array-based graph reservoir. Adv. Mater. **36**(7), 2309314 (2024)
- [43] Xia, W., Zou, J., Qiu, X., Chen, F., Zhu, B., Li, C., Deng, D.-L., Li, X.: Configured quantum reservoir computing for multi-task machine learning. Sci. Bull. **68**(20), 2321–2329 (2023)
- [44] Dudas, J., Carles, B., Plouet, E., Mizrahi, F.A., Grollier, J., Marković, D.: Quantum reservoir computing implementation on coherently coupled quantum oscillators. npj Quantum Inf. **9**(1), 64 (2023)
- [45] Bravo, R.A., Najafi, K., Gao, X., Yelin, S.F.: Quantum reservoir computing using arrays of rydberg atoms. PRX Quantum **3**(3), 030325 (2022)
- [46] Yasuda, T., Suzuki, Y., Kubota, T., Nakajima, K., Gao, Q., Zhang, W., Shimono, S., Nurdin, H.I., Yamamoto, N.: Quantum reservoir computing with repeated measurements on superconducting devices. arXiv:2310.06706 (2023)
- [47] Čindrak, S., Donvil, B., Lüdge, K., Jaurigue, L.: Enhancing the performance of quantum reservoir computing and solving the time-complexity problem by artificial memory restriction. Phys. Rev. Res. **6**(1), 013051 (2024)

Acknowledgements

We acknowledge the use of CUDA Quantum for this work. The views expressed are those of the authors and do not reflect the official policy or position of NVIDIA or the CUDA Quantum team..

Funding

The work of Shehbaz Tariq and Symeon Chatzinotas was supported by the project Lux4QCI (GA 101091508) funded by the Digital Europe Program, and the project LUQCIA Funded by the European Union – Next Generation EU, with the collaboration of the Department of Media, Connectivity and Digital Policy of the Luxembourgish Government in the framework of the RRF program. The work of Muhammad Talha and Hyundong Shin is supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under RS-2025-00556064 and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-2021-0-02046) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

Authors’ contribution

ST and MT contributed the idea. ST and MT developed the theory and wrote the manuscript. SC and HS improved the manuscript and supervised the research. All authors contributed to the analysis and discussion of the results and improved the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.