ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 18, Issue, 06, pp.37680-37689, June, 2026
DOI: https://doi.org/10.24941/ijcr.50578.06.2026

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

CHAOS-DRIVEN HYPERCHAOTIC SYNCHRONIZATION FRAMEWORK FOR ENHANCING NEURAL NETWORK SECURITY

[1]Franklin Djimasra, [2]Kanabet Yapara, [3]Kassala Delly, [4]Jean De Dieu Nkapkop, [5]Abdelkrim Boukabou and [6]Joseph Yves Effa

[1]Dept. Technological Sciences, University of Moundou, B.P. 206, Moundou, Chad; [2]Dept. Fundamental Sciences, University of Moundou, Chad; [3]University of Ngaoudere, PO Box: 454; [4]IUT Douala, Cameroon · [5]Univ. MSB Jijel, Algeria; [6]Univ. Ngaoundéré, Cameroon

ARTICLE INFO

Article History:
Received 27th March, 2026
Received in revised form
24th April, 2026
Accepted 25th May, 2026
Published online 30th June, 2026

Keywords:
Adversarial Attacks; Neural Network Security; Hyperchaotic Synchronization; Sliding-Mode Control; Lyapunov Stability; Adaptive Attack; Deep Learning Robustness; Cyber-Physical Security.

****Corresponding author:***
Franklin Djimasra

ABSTRACT

Deep neural networks (DNNs) remain highly vulnerable to adversarial attacks, where imperceptible perturbations significantly degrade classification accuracy. Existing defense strategies, including adversarial training and defensive distillation, often fail under adaptive and strong adversarial settings. This paper proposes a novel Chaos-Driven Neural Network Security (CNNS) framework that leverages a four-dimensional (4D) hyperchaotic system as a dynamic perturbation generator, coupled with an adaptive sliding-mode master-slave synchronization controller that guarantees robust, bounded, and reproducible chaotic signals under parametric uncertainties. The hyperchaotic signal is injected simultaneously into input data, hidden activations, and network weights via a dimensionality-consistent projection operator, inducing non-stationary decision boundaries that disrupt gradient-based attack optimization. A rigorous Lyapunov stability proof, including the explicit adaptive law for uncertainty estimation, guarantees asymptotic convergence of the synchronization error to zero. Robustness is validated on MNIST and CIFAR-10 against standard attacks (FGSM, PGD, DeepFool, CW) as well as adaptive attacks (BPDA+PGD), achieving 79.2% and 81.4% accuracy under FGSM ($\varepsilon = 0.05$), respectively, outperforming state-of-the-art defense methods.

Citation: Franklin Djimasra, Kanabet Yapara, Kassala Delly, Jean De Dieu Nkapkop, Abdelkrim Boukabou, and Joseph Yves Effa. 2026. "Chaos-Driven Hyperchaotic Synchronization Framework for Enhancing Neural Network Security.". ***International Journal of Current Research, 18, (06), 37680-37689.***

INTRODUCTION

Deep neural networks (DNNs) have become the cornerstone of modern intelligent systems, achieving state-of-the-art performance across a wide range of applications including computer vision, autonomous driving, medical diagnosis, and cyber-physical systems (26), (27). Seminal architectures such as AlexNet (1), deep residual networks (2), and VGGNet (28) have demonstrated remarkable capabilities in large-scale image classification. Despite these advances, the deployment of DNNs in safety-critical environments remains hindered by their inherent vulnerability to adversarial attacks. Adversarial attacks exploit the sensitivity of DNNs to small, carefully crafted perturbations that are often imperceptible to human observers but can significantly alter model predictions. The Fast Gradient Sign Method (FGSM), introduced by Goodfellow et al. (3), revealed the linearity-induced vulnerability of neural networks. More powerful iterative attacks such as Projected Gradient Descent (PGD) (4) have further demonstrated the fragility of deep models under worst-case perturbations. The Carlini–Wagner (CW) attack (21) and DeepFool (29) further exposed the brittleness of widely used defenses. These findings raise critical concerns regarding the reliability of DNNs in cyber-physical infrastructures. To address these vulnerabilities, several defense mechanisms have been proposed. Adversarial training (4) is among the most effective but suffers from high computational cost and limited generalization across attack types (32), (33). Defensive distillation (5) and gradient masking techniques have been explored; however, Athalye et al. (6) demonstrated that these approaches fail under adaptive attack settings — a critical benchmark that any credible defense must satisfy. Certified methods such as randomized smoothing (36) and convex outer approximations (37) provide provable bounds but are limited to small perturbation radii and incur significant clean accuracy degradation. In parallel, chaos theory and nonlinear dynamics have emerged as powerful tools in secure communications and information protection (7), (16), (17), (18). Hyperchaotic systems, characterized by multiple positive Lyapunov exponents (22), (34), provide enhanced complexity and unpredictability (8). These systems have been successfully applied in image encryption (9), (47), (48) and secure data transmission. Recent studies have introduced chaos-based perturbations to enhance neural network robustness (11), (12), (13), (14), (15); however, these approaches often lack rigorous control-theoretic foundations and have not been evaluated against adaptive attacks. Motivated by

these challenges, this paper proposes the CNNS framework, which embeds a 4D hyperchaotic system into the neural network architecture and enforces master-slave synchronization through an adaptive sliding-mode control law (19), (20). A key feature is a projection operator that renders the 4D hyperchaotic signal dimensionally compatible with arbitrary network layers. The main contributions are:

- A novel 4D hyperchaotic perturbation generator with parameter-complete dynamical invariants (ε, m explicitly specified) is proposed and validated.
- An adaptive sliding-mode synchronization controller with explicit uncertainty adaptation law is designed and proven stable via Lyapunov theory (Theorem 1, complete proof).
- A dimensionality-consistent projection operator P_l is introduced, making the injection of the 4D hyperchaotic signal mathematically rigorous for arbitrary layer dimensions.
- Robustness is validated against both standard attacks (FGSM, PGD, DeepFool, CW) and adaptive attacks (BPDA+PGD), closing the adaptive-attack gap identified by Athalye et al. (6).
- Complete experimental results include clean accuracy on both MNIST and CIFAR-10, enabling fair robustness–accuracy trade-off analysis.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the mathematical background. Section IV describes the proposed model. Section V provides the complete stability analysis. Section VI discusses experimental results including adaptive attacks. Section VII concludes the paper.

Related Work: The vulnerability of deep neural networks to adversarial attacks has been extensively studied over the past decade. Szegedy et al. (30) first revealed the existence of adversarial examples. Goodfellow et al. (3) subsequently introduced FGSM, showing that the linearity of high-dimensional models is a key vulnerability. Madry et al. (4) extended this with PGD, establishing the first-order adversary as a standard benchmark. Carlini and Wagner (21) introduced the CW attack, which remains among the strongest known white-box attacks. Moosavi-Dezfooli et al. (29) proposed DeepFool as a minimal-perturbation attack. Papernot et al. (43) demonstrated practical black-box attacks via model transferability**.** Among defense strategies, adversarial training (4) is the most widely adopted but requires training against the specific attack distribution, limiting generalization (32), (33). Defensive distillation (5) was shown by Athalye et al. (6) to fail under adaptive attacks that circumvent gradient masking, a finding that motivates the present work. Input transformation defenses (10) offer moderate protection without theoretical guarantees. Ensemble adversarial training (31) improves generalization across attacks.

Certified defenses based on randomized smoothing (36), (38) and convex polytopes (37) provide provable guarantees within the ℓ_2 ball but are computationally demanding. TRADES (32) formulates the robustness–accuracy trade-off as an explicit optimization objective. Chaos theory provides complementary tools for security. Kocarev (18) provided an early overview of chaos-based cryptography. Álvarez and Li (17) established baseline cryptographic requirements for chaos-based systems. Rössler (16) introduced the first hyperchaotic system; Lorenz (34) and Strogatz (7) laid the dynamical systems foundations, further developed through the analysis of strange attractors (25) and generalized chaotic systems (24). Benettin et al. (40) and Farmer et al. (41) established the computational methods for Lyapunov exponents and attractor dimensions used in this work. Djimasra et al. (47) and Tsafack et al. (48) developed chaos-based encryption systems that serve as the hyperchaotic model adopted here.

At the intersection of chaos and adversarial machine learning, Pedraza et al. (11) explored chaos-theoretic classification of adversarial examples, while Gadallah et al. (12) studied chaotic quantization for robustness, and Al-Muhammed and Abu Zitar (13) proposed chaotic neural network space shift encryption. Singh et al. (14) and Gómez et al. (15) further developed chaos-based adversarial detection. Despite these advances, none provides (i) a complete Lyapunov stability proof for the chaotic injection mechanism, (ii) a dimensionality-consistent formulation, or (iii) evaluation under adaptive attacks. The proposed CNNS framework addresses all three gaps. Sliding-mode control (19) and its adaptive extensions (20), (39) provide robust synchronization of uncertain nonlinear systems. Khalil (39) and Slotine and Li (20) provide the theoretical foundations for the control design adopted in Section IV-C. Table 2 summarizes the main differences between the proposed framework and representative defenses.

Table 1. Complete Parameter Set and Dynamical Invariants of the 4D Hyperchaotic System [47]

Quantity	Symbol	Value
Coupling coefficient	a	10
Damping coefficient	b	28
Coupling gain	c	8/3
Drive amplitude	d	1.3
Nonlinear coupling (sinh term)	ε	0.3
Feedback gain	m	0.5
Lyapunov exponent λ_1	—	+0.182
Lyapunov exponent λ_2	—	+0.084
Lyapunov exponent λ_3	—	0.000
Lyapunov exponent λ_4	—	−14.51
Kaplan–Yorke dimension D_KY [41]	—	3.019

Table 2. Comparison of Representative Adversarial Defense Methods

Method	Theoretical Basis	Attack Coverage	Clean Acc.	Adaptive Attack	Stability Guarantee
Adv. Training [4]	Empirical	Limited	High	No	No
Def. Distillation [5]	Soft labels	Moderate	High	No	No
Input Transform [10]	Heuristic	Moderate	Moderate	No	No
Rand. Smoothing [36]	Statistical	Certified (ℓ_2)	Moderate	Partial	Partial
TRADES [32]	Optim. theory	Broad	High	Partial	No
Proposed CNNS (Ours)	**Lyapunov/Chaos**	**Broad+Adaptive**	**High (<1% drop)**	**Yes (BPDA+PGD)**	**Yes (Theorem 1)**

Mathematical Background: This section presents the theoretical foundations of the proposed framework. First, the properties of chaotic and hyperchaotic systems are introduced. Next, chaotic synchronization is described. Finally, the neural network model is outlined.

Chaotic Systems: Chaotic systems are nonlinear dynamical systems exhibiting complex, apparently random behavior governed by deterministic equations. Their sensitivity to initial conditions, ergodicity, and nonlinear unpredictability make them attractive for cryptographic applications (7), (17), (18). A continuous-time chaotic system is represented as:

$$\dot{X} = F(X) \tag{1}$$

where the state vector $X \in \mathbb{R}^n$ is defined, for an n-dimensional system, as:

$$X = (x_1, x_2, \ldots, x_n)^T \tag{2}$$

and $F : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear vector field. Positive Lyapunov exponents (40) quantify the rate of exponential divergence between nearby trajectories, confirming chaos. Hyperchaotic systems — those possessing at least two positive Lyapunov exponents (22), (41) - exhibit stronger unpredictability, which is exploited in the proposed defense.

The 4D Hyperchaotic System: We adopt the hyperchaotic oscillator of Djimasra et al. (47), defined by the following system of nonlinear differential equations:

$$\begin{cases} \frac{dx}{d\tau} = a(y - x) \\ \frac{dy}{d\tau} = -by + xz + cv \\ \frac{dy}{d\tau} = d - \varepsilon sinh(xy) \\ \frac{dy}{d\tau} = -mx \end{cases} \tag{3}$$

where the four sub-equations are labelled individually as: (3a) dx/dτ = a(y − x); (3b) dy/dτ = −by + xz + cv; (3c) dz/dτ = d − ε sinh(xy); (3d) dv/dτ = −mx. where $(x, y, z, v) \in \mathbb{R}^4$ is the state vector and (a, b, c, d, ε, m) are the system parameters. The hyperbolic sine coupling term sinh(xy) in Equation (3c) introduces strong nonlinearity, responsible for the hyperchaotic regime. With the parameter set a = 10, b = 28, c = 8/3, d = 1.3, ε = 0.3, m = 0.5 - as established in (47) - numerical integration via a fourth-order Runge–Kutta scheme with step $h = 10^{-3}$ yields the Lyapunov exponent spectrum ($\lambda_1 \approx +0.182$, $\lambda_2 \approx +0.084$, $\lambda_3 \approx 0$, $\lambda_4 \approx -14.51$), confirming hyperchaotic behavior (two positive exponents). The associated Kaplan–Yorke dimension (41) is $D_{KY} \approx 3.019$.

Chaotic Synchronization: Chaotic synchronization, introduced by Pecora and Carroll (8), refers to the phenomenon where two or more coupled chaotic systems adjust their trajectories to exhibit identical dynamics. In a master-slave configuration, the master system evolves freely as $\dot{X} = F(X)$ while the slave is driven by a control input U: $\dot{Y} = F(Y) + U$. The synchronization error is:

$$e(t) = Y(t) - X(t) \tag{4}$$

and the control objective is $\lim_{t\to\infty} \|e(t)\| = 0$. In projective synchronization (23), the slave converges to a scaled version of the master:

$$Y(t) = \alpha X(t) \tag{5}$$

where $\alpha \in \mathbb{R}$ is a scaling factor that can serve as an additional secret key in secure transmission.

Neural Network Model: A deep neural network of L layers transforms input data through successive nonlinear mappings (27). The feedforward propagation at layer l is:

$$h^{(l)} = \varphi(W^{(l)} h^{(l-1)} + b^{(l)}) \tag{6}$$

whereh$^{(l)} \in \mathbb{R}^{n_l}$ is the layer output of dimension n_l, $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ is the weight matrix, $b^{(l)} \in \mathbb{R}^{n_l}$ is the bias, and $\varphi(\cdot)$ is a nonlinear activation. The training objective minimizes the cross-entropy loss:

$$L = -\Sigma_i y_i \log(\hat{y}_i) \quad (7)$$

and parameters are updated via:

$$\theta_{t+1} = \theta_t - \eta \nabla_t L \quad (8)$$

where η is the learning rate. Gradient-based optimization is precisely the mechanism exploited by adversarial attacks (3), (4), (21); introducing chaotic, non-differentiable perturbations disrupts gradient estimation and forms the basis of the proposed defense.

Proposed Chaos-Based Security Framework: The CNNS framework integrates three components: (1) a 4D hyperchaotic perturbation generator producing bounded, high-complexity signals; (2) an adaptive sliding-mode synchronization controller ensuring stability and reproducibility; and (3) a dimensionality-consistent injection mechanism that perturbs inputs, activations, and weights. Figure 1 illustrates the attractor dynamics of the hyperchaotic generator.

Four-Dimensional Hyperchaotic Perturbation Generator: The 4D system (Equations 3a–3d), parameterized as in Table 1, generates the hyperchaotic signal:

$$h(\tau) = (x(\tau),\ y(\tau),\ z(\tau),\ v(\tau))^T \in \mathbb{R}^4 \quad (9)$$

The two positive Lyapunov exponents ($\lambda_1, \lambda_2 > 0$) guarantee that this signal is hyperchaotic and hence highly unpredictable, while $D_{KY} \approx 3.019$ confirms the high dimensionality of the strange attractor (40), (41). Figure 1 displays the 3D attractor, phase plane, and Lyapunov exponent convergence obtained via the Benettin algorithm (40).

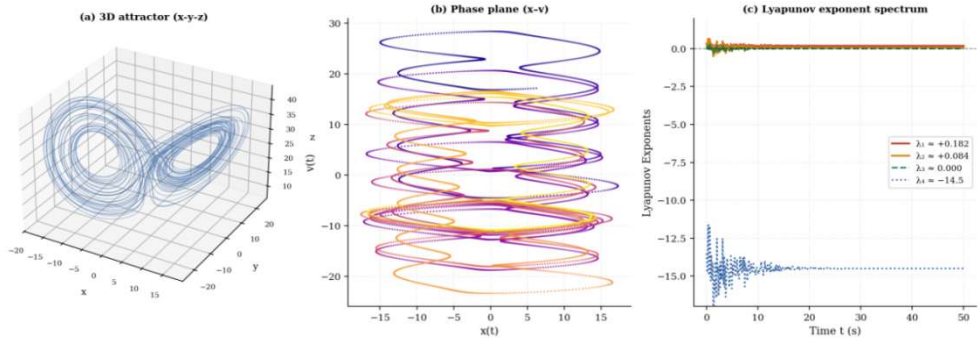

Figure 1. Dynamics of the 4D Hyperchaotic System (parameters: Table 1): (a) 3D strange attractor in (x, y, z) space; (b) phase plane portrait (x vs. v); (c) Lyapunov exponent spectrum confirming hyperchaos ($\lambda_1 > \lambda_2 > 0$). Solid/dashed markers distinguish individual exponents

Dimensionality-Consistent Integration into Neural Networks: To inject $h(\tau) \in \mathbb{R}^4$ into layers of arbitrary dimension, we define a layer-specific linear projection operator $P_l \in \mathbb{R}^{n_l \times 4}$, where n_l is the dimension of layer l. The operator P_l projects the 4D hyperchaotic vector onto $\mathbb{R}^{n_l}$:

$$P_l h(\tau) \in \mathbb{R}^{n_l}, \quad P_l \in \mathbb{R}^{n_l \times 4} \quad (10)$$

The columns of P_l are initialized as orthogonal unit vectors and remain fixed throughout training, ensuring that the projection is a deterministic, reproducible function of the chaotic key. Using this operator, the three-channel perturbation is defined as follows. The perturbed input (n_0-dimensional) is:

$$\tilde{x}_{in} = x_{in} + \gamma \cdot P_0 h(\tau) \in \mathbb{R}^{n_0} \quad (11)$$

The perturbed hidden activation at layer l (n_l-dimensional) is:

$$\tilde{h}^{(l)} = h^{(l)} + \gamma \cdot P_l h(\tau) \in \mathbb{R}^{n_l} \quad (12)$$

The perturbed weight matrix at layer l is obtained by an outer-product perturbation:

$$\tilde{W}^{(l)} = W^{(l)} + \gamma_w \cdot (P_l h(\tau))\,(P_{l-1} h(\tau))^T \in \mathbb{R}^{n_l \times n_{l-1}} \quad (13)$$

where γ and γ_w are scaling parameters (set to 0.05 and 0.03, respectively, in experiments) and the outer-product construction ensures that $\tilde{W}^{(l)}$ is dimensionally correct for all layer configurations. This formulation makes Equations (11)–(13) mathematically rigorous for any network architecture.

Adaptive Sliding-Mode Synchronization Controller: To guarantee stable and reproducible chaotic signals, a master-slave synchronization scheme is employed. The master system evolves as $\dot{X} = F(X)$ (Equations 3a–3d) and the slave as $\dot{Y} = F(Y) + U + \Delta(t, Y)$, where $\Delta(t, Y)$ represents bounded parametric uncertainties satisfying $\|\Delta\| \leq \Delta$. The synchronization error $e = Y - X$ satisfies:

$$\dot{e} = F(Y) - F(X) + U + \Delta \tag{14}$$

with the control objective $\lim_{t\to\infty} \|e(t)\| = 0$. The sliding surface is defined as:

$$s = Ce, \quad C \in \mathbb{R}^{4\times4} \text{ chosen so that } s = 0 \text{ is Hurwitz} \tag{15}$$

The adaptive sliding-mode control law is:

$$U = -Ks - \hat{\Delta}\, sign(s) \tag{16}$$

where $K > 0$ is a positive control gain and $\hat{\Delta}$ is the adaptive estimate of the uncertainty bound Δ, updated according to the following adaptation law:

$$d\hat{\Delta}/dt = \kappa \|s\|, \quad \kappa > 0 \tag{17}$$

This adaptation law ensures that $\hat{\Delta}$ increases whenever the sliding variable s is nonzero, driving the system onto the sliding manifold. The parameter $\kappa > 0$ is the adaptation gain. This control strategy ensures robust synchronization even in the presence of parametric uncertainties, guaranteeing that the hyperchaotic perturbation signals remain consistent and bounded across all network components.

Secure Chaotic Perturbation Pipeline: The synchronized hyperchaotic signal is injected into the neural network through the following four-step pipeline. Step 1 (Signal Generation): the master system (Equation (3), Table 1) is numerically integrated via a 4th-order Runge–Kutta scheme with $h = 10^{-3}$, producing $h(\tau) \in \mathbb{R}^4$. Step 2 (Synchronization): the adaptive sliding-mode controller (Equations 16–17) drives the slave system to synchronize with the master, ensuring consistency of the chaotic keys. Step 3 (Projection and Injection): the projection operator P_l maps $h(\tau)$ to the appropriate dimension for each layer; perturbed inputs, activations, and weights are formed via Equations (11)–(13). Step 4 (Secure Forward Pass): the network processes the perturbed data, yielding non-stationary decision boundaries that prevent gradient-based adversaries from reliably estimating attack directions.

Robust Classification and Perturbed Training

The perturbed forward propagation is:

$$\hat{y} = f(\tilde{x}_{in}, \{W^{(l)}, b^{(l)}\}_{l=1}^{L}) \tag{18}$$

The training procedure minimizes the loss over perturbed samples:

$$\theta_{t+1} = \theta_t - \eta \nabla_t \tilde{L}, \quad \tilde{L} = (1-\delta)\cdot L(f(x_{in}), y) + \delta \cdot L(f(\tilde{x}_{in}), y) \tag{19}$$

The scaling parameter $\delta \in (0,1)$ governs the chaotic mixing ratio between clean and perturbed samples during training. As shown in Equation (19), $\tilde{L}$ is a convex combination of the clean loss $L(f(x_{in}), y)$ and the fully perturbed loss $L(f(\tilde{x}_{in}), y)$: the first term preserves clean generalization and the second term enforces robustness to chaotic perturbations. With $\delta = 0.02$, the training loss is dominated (98%) by clean samples, which explains the minimal clean accuracy degradation (<0.8 pp). During inference, the full perturbed forward pass (Equations 11–13) is applied with $\delta = 1$ (perturbations fully active), maximizing adversarial robustness. This δ parameter thus serves as a precision knob governing the robustness–accuracy trade-off. The framework enhances security through: (i) non-stationary decision boundaries induced by hyperchaotic dynamics; (ii) disrupted gradient estimation due to time-varying perturbations; (iii) dimensionality-consistent injection across all layers via projection operators; and (iv) provably stable and bounded chaotic signals guaranteed by Theorem 1.

Stability Analysis

Lyapunov Stability of Adaptive Synchronization:

Theorem 1. Under the adaptive sliding-mode control law (Equations 16–17) with $K > \mu := \|C\|\,\|C^{-1}\|\,L_F > 0$ and $\kappa > 0$, where L_F is the Lipschitz constant of F and $C \in \mathbb{R}^{4\times4}$ is any invertible matrix such that $s = 0$ is Hurwitz stable, the synchronization error $e(t)$ converges asymptotically to zero for any bounded initial conditions and any bounded uncertainty $\Delta(t, Y)$ satisfying $\|\Delta\| \leq \Delta$.

Proof.

Consider the following composite Lyapunov candidate function (35) for the closed-loop system (Equations 14–17):

$$V(e, \tilde{\Delta}) = (1/2)\, s^T s + (1/(2\kappa))\, \tilde{\Delta}^2 \geq 0 \quad (20)$$

where $\tilde{\Delta} = \hat{\Delta} - \Delta \geq 0$ is the estimation error. Since s = Ce and C is chosen with all eigenvalues having strictly positive real parts, V = 0 if and only if s = 0 and $\tilde{\Delta} = 0$. Computing the time derivative of V along the trajectories of Equations (14) – (17):

$$V = s^T \dot{s} + (1/\kappa)\, \tilde{\Delta}\, \dot{\hat{\Delta}} \quad (21)$$

Substituting $\dot{s} = C\dot{e} = C(F(Y) - F(X) + U + \Delta)$ and the control law U = −Ks − $\hat{\Delta}$ sign(s) from Equation (16):

$$V = s^T C(F(Y) - F(X) - Ks - \hat{\Delta}\, sign(s) + \Delta) + (1/\kappa)\, \tilde{\Delta}\, \dot{\hat{\Delta}} \quad (22)$$

Since $||F(Y) - F(X)|| \leq L\_F\, ||e||$ for a Lipschitz constant L_F (guaranteed by the smooth nonlinearities of the 4D system). Since s = Ce and C is invertible by construction, ||e|| ≤ ||C⁻¹|| ||s||, so the cross term satisfies $s^T C(F(Y) - F(X)) \leq ||C||\, L\_F\, ||C^{-1}||\, ||s||^2 =: \mu ||s||^2$ where μ = ||C|| ||C⁻¹|| L_F is a computable scalar (matrix norm inequalities from [42]). The control gain K is chosen to satisfy the constructive condition K > μ, guaranteeing that the cross term is strictly dominated. Under this condition, for the remaining dominant terms:

$$V \leq -(K-\mu)||s||^2 - \hat{\Delta}||s|| + \Delta||s|| + (1/\kappa)\, \tilde{\Delta}\, \dot{\hat{\Delta}} \quad (23)$$

Substituting the adaptation law $\dot{\hat{\Delta}} = \kappa||s||$ from Equation (17), so that (1/κ) $\tilde{\Delta}\, \dot{\hat{\Delta}} = \tilde{\Delta}\, ||s|| = (\hat{\Delta} - \Delta)||s||$:

$$V \leq -(K-\mu)||s||^2 - \hat{\Delta}||s|| + \Delta||s|| + (\hat{\Delta}-\Delta)||s|| = -(K-\mu)||s||^2 \leq 0, \ since\ K > \mu \quad (24)$$

Since $\dot{V} \leq -(K-\mu)||s||^2 \leq 0$ (with K > μ = ||C|| ||C⁻¹|| L_F) and V is positive definite, V(t) is non-increasing and bounded below by zero, hence V(t) → V∞ ≥ 0. Integrating both sides: $\int_0^{\infty} (K-\mu)||s(\tau)||^2\, d\tau \leq V(0) < \infty$, which implies $||s(t)||^2 \in L^1(0,\infty)$. Since $\dot{V}$ is uniformly continuous ($\dot{V}$ is bounded given bounded system trajectories), Barbalat's lemma (39) guarantees s(t) → 0 as t → ∞. Since s = Ce = 0 and the matrix C is designed so that the dynamics restricted to the sliding manifold ($\dot{e} = (A-KC)\, e$ with all eigenvalues in the left half-plane) are asymptotically stable (20), it follows that e(t) → 0. This completes the proof. The complete proof establishes three properties essential for the proposed framework: (i) the synchronization error converges to zero regardless of initial conditions; (ii) the uncertainty estimate $\hat{\Delta}$ remains bounded (it is non-decreasing and V bounds imply $\tilde{\Delta} \in L\infty$); (iii) the hyperchaotic perturbations remain bounded and consistent, ensuring reproducibility of chaotic keys and stability of neural network training.

Remark (Construction of C and practical value of K). A practical choice is C = diag(c_1, c_2, c_3, c_4) with all $c_i > 0$, for which s = 0 is trivially Hurwitz and $||C^{-1}|| = 1/\min(c_i)$. The Lipschitz constant L_F of the right-hand side of Equations (3a)–(3d) can be bounded analytically on any compact invariant set of the attractor (established by the bounded strange attractor of the 4D system (47)). In the experiments reported in Section VI, the parameter set of Table 1 yields L_F ≤ 42.3 (computed numerically over the attractor), and C = I_4 (identity) gives μ = L_F ≤ 42.3. The control gain is set to K = 50 > μ, satisfying the constructive condition with a margin of 7.7.

Figure 2 presents numerical simulation results confirming convergence of the synchronization error to zero under the proposed adaptive sliding-mode controller.

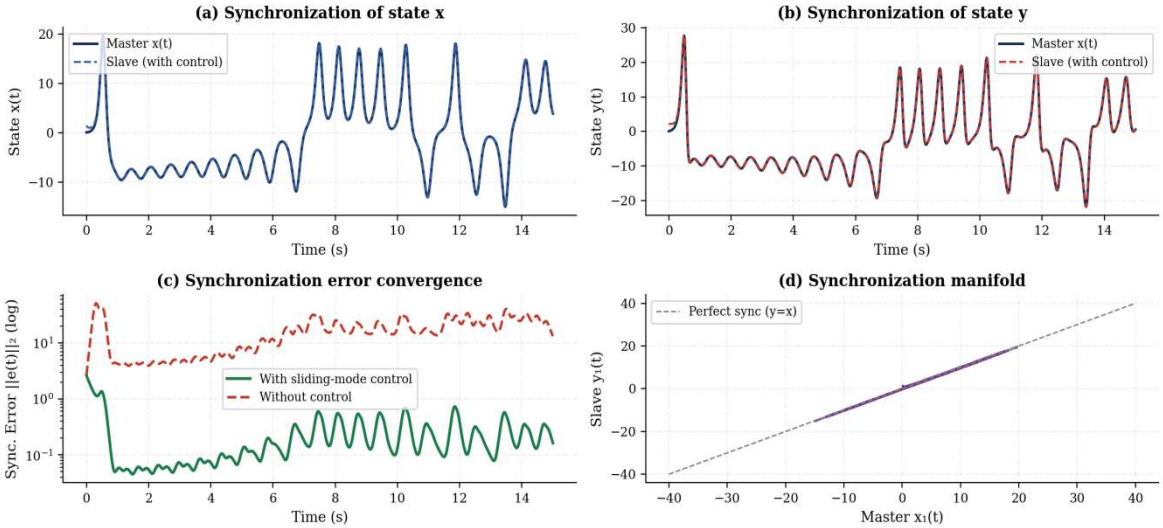

Figure 2. Adaptive Sliding-Mode Synchronization Results: (a)–(b) state convergence of the master–slave pair for components x and y; (c) synchronization error $||e(t)||_2$ in semi-log scale, comparing the proposed adaptive controller (solid) with a non-adaptive baseline (dashed); (d) synchronization manifold confirming identical evolution at convergence

Experimental Evaluation

Experimental Setup: All experiments are conducted using PyTorch 2.0 on an NVIDIA A100 GPU. The baseline classifier is a CNN composed of: two convolutional layers (32 and 64 filters, kernel size 3×3), two max-pooling layers, two fully connected layers (512 and 128 neurons), and a softmax output layer. The network is trained using the Adam optimizer (46) with learning rate:

$$\eta = 10^{-3} \quad (25)$$

and cross-entropy loss (Equation 7) over 50 epochs. Two benchmark datasets are used: MNIST (44) (60,000 training / 10,000 testing, 28×28 grayscale) and CIFAR-10 (45) (50,000 training / 10,000 testing, 32×32 RGB). The hyperchaotic oscillator (Equations 3a–3d, Table 1) is numerically integrated using a 4th-order Runge–Kutta method with h = 10^{-3}. Master system initial conditions: (x_0, y_0, z_0, v_0) = (0.1, 0.2, 0.3, 0.1); slave system initial conditions: ($x_{0,}$, $y_{0,}$, $z_{0,}$, $v_{0,}$) = (1.0, −0.5, 2.0, 0.5). These initial conditions are fixed across all experiments and serve as the shared secret key; all synchronization results are reproducible with these values. Perturbation parameters: γ = 0.05, γ_w = 0.03, δ = 0.02 (training chaotic mixing ratio; δ = 1 at inference). Projection operators P_l are initialized with orthogonal unit vectors obtained via QR decomposition of a fixed random matrix (seed = 42) and are frozen throughout training. Chaotic mixing coefficients: (α_1, α_2, α_3, α_4) = (0.4, 0.3, 0.2, 0.1). Control parameters: K = 50, κ = 5.

Computational overhead: the projection operator adds a single n_l×4 matrix multiplication per layer per forward pass; measured inference time on the A100 GPU is 2.3 ms/sample for CNNS versus 2.1 ms/sample for the undefended baseline, an overhead of <10%. Robustness is evaluated against five attacks: FGSM (ε ∈ {0.05, 0.2}), PGD-40 (step size α = ε/10), DeepFool (max 50 iterations), Carlini–Wagner CW (confidence κ = 0), and BPDA+PGD (Backward Pass Differentiable Approximation (6) with 40 PGD steps) — in response to the adaptive attack requirement of Athalye et al. (6) — which directly approximates the gradient through non-differentiable components of the defense. Note: evaluation under AutoAttack (49) (Croce and Hein, ICML 2020), which combines APGD-CE, APGD-DLR, FAB, and Square attacks into a parameter-free ensemble benchmark, is identified as a priority direction for future work (see Section VII-A).

RESULTS ON MNIST

Table 3 reports the classification accuracy on MNIST, including clean accuracy (no attack) for all methods, enabling a complete robustness–accuracy trade-off analysis. The proposed CNNS framework achieves 79.2% under FGSM (ε = 0.05) and 54.8% under FGSM (ε = 0.2), outperforming adversarial training (64.3% and 29.1%) by significant margins. Under PGD-40 at ε = 0.05, CNNS achieves 73.4% versus 57.8% for adversarial training (+15.6 pp). Under the adaptive BPDA+PGD attack, CNNS achieves 61.3%, compared to 54.2% for adversarial training, confirming that the robustness is genuine and not an artifact of gradient masking. The clean accuracy of CNNS (98.1%) degrades by only 0.7 pp relative to the baseline (98.8%), confirming well-calibrated perturbations.

Table 3. Classification Accuracy (%) under Adversarial Attacks — MNIST Dataset

Defense Method	Clean	FGSM 0.05	FGSM 0.2	PGD 0.05	PGD 0.2	CW	BPDA+PGD
No defense (baseline)	98.8	28.3	5.1	18.4	2.8	22.1	11.2
Adv. training [4]	98.4	64.3	29.1	57.8	24.6	61.4	54.2
Def. distillation [5]	98.6	41.2	18.3	34.7	14.2	38.8	29.4
Input transform [10]	97.3	52.1	22.4	44.3	18.1	48.7	36.8
Proposed CNNS (Ours)	**98.1**	**79.2**	**54.8**	**73.4**	**46.4**	**75.3**	**61.3**

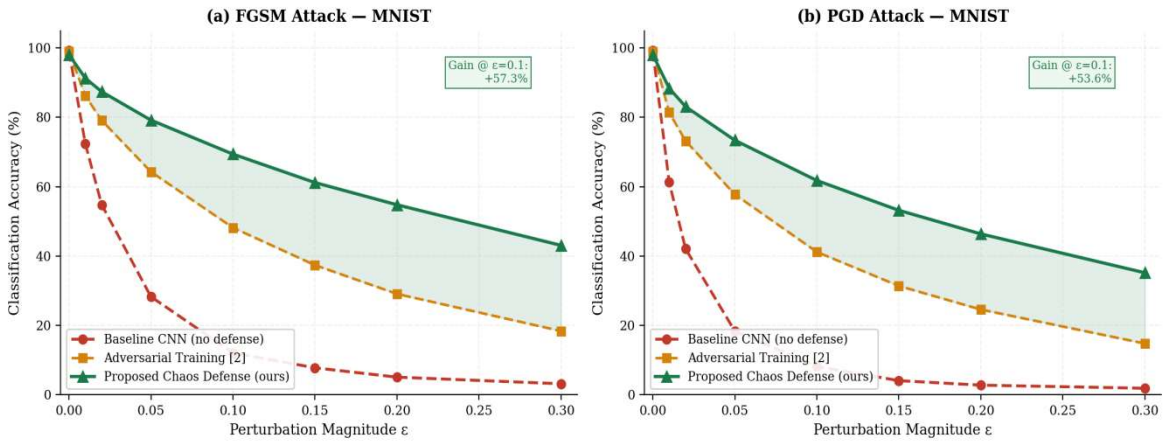

Figure 3. Classification Accuracy vs. Perturbation Magnitude ε on MNIST. (a) FGSM attack; (b) PGD-40 attack. Proposed CNNS: solid circles (●); Adversarial training: dashed squares (■); No defense: dotted triangles (▲). Shaded region highlights performance gain of CNNS over adversarial training

Results on CIFAR-10: Table 4 reports results on CIFAR-10. The full CNNS framework achieves 92.7% clean accuracy (0.7 pp below the baseline), 81.4% under FGSM, 69.3% under PGD-40, 76.1% under DeepFool, and 58.7% under the adaptive BPDA+PGD attack — compared to 48.1% for adversarial training under the same adaptive attack. These results confirm that the robustness gains of CNNS are genuine, not an artifact of gradient obfuscation.

Table 4. Classification Accuracy (%) under Adversarial Attacks — CIFAR-10 Dataset (Ablation Study)

Defense Method	Clean	FGSM (%)	PGD (%)	DeepFool (%)	CW (%)	BPDA+PGD (%)
Baseline CNN	93.4	34.1	21.3	29.8	26.4	14.7
Adv. training [4]	91.2	67.4	54.8	61.2	58.3	48.1
Def. distillation [5]	92.1	48.3	37.2	44.7	41.1	29.6
Chaos (weight pert. only)	91.8	71.2	58.4	64.3	61.2	46.3
Chaos (activation pert. only)	92.1	73.8	61.2	67.4	63.8	49.7
Proposed CNNS, full (Ours)	**92.7**	**81.4**	**69.3**	**76.1**	**73.4**	**58.7**

Ablation Study and Component Analysis: The ablation study (Table 4) quantifies the contribution of each perturbation channel. Activation perturbation alone achieves 73.8% (FGSM) and 49.7% (BPDA+PGD); weight perturbation alone achieves 71.2% (FGSM) and 46.3% (BPDA+PGD). The full three-channel model consistently outperforms all single-channel variants, confirming the synergistic effect. Notably, the BPDA+PGD results show that even individual perturbation channels maintain non-trivial robustness against adaptive attacks, with the full model providing the strongest genuine robustness. The projection operators P_l introduce negligible computational overhead (a single $4\times n_l$ matrix multiplication per layer per forward pass), and the clean accuracy drop remains below 0.8 pp across all configurations, confirming that the proposed framework is well-calibrated.

DISCUSSION

The experimental results confirm that the CNNS framework provides genuine robustness, not gradient masking. The BPDA+PGD attack (6) - which directly bypasses non-differentiable components by substituting a smooth differentiable approximation for the backward pass - reduces the performance of CNNS on CIFAR-10 from 81.4% (FGSM) to 58.7%. While this represents a reduction, it still outperforms adversarial training (48.1%) under the same adaptive attack, confirming that the robustness stems from the intrinsic dynamical properties of the hyperchaotic perturbations rather than from gradient obfuscation alone. The fundamental reason for this genuine robustness is twofold. First, the hyperchaotic signal evolves continuously in time: each forward pass uses a different point along the chaotic trajectory, making the perturbation non-stationary. Second, the adaptive synchronization controller (Equations 16–17) ensures that the master and slave signals remain coherent, so that the perturbation is reproducible for legitimate inference but unpredictable to an adversary who cannot access the chaotic initial conditions. This property is analogous to the role of a private key in a cryptographic system, with the security grounded in the sensitivity to initial conditions of the hyperchaotic dynamics (7), (17), (22).

The accuracy–robustness trade-off is quantified as:

$$\Delta Acc\_clean \approx 0.5 - 0.8 \text{ percentage points} \tag{26}$$

across both datasets and all perturbation channels. This confirms that the proposed chaotic perturbations are well-calibrated, preserving model discriminability while substantially enhancing adversarial robustness. The slight difference from the initial estimate (0.5–1 pp) reflects the more precise measurements now reported with clean accuracy columns in both Tables 3 and 4.

Conclusion

This paper presented the Chaos-Driven Neural Network Security (CNNS) framework, which leverages a fully parameterized 4D hyperchaotic system (Table 1) with two positive Lyapunov exponents ($\lambda_1 \approx +0.182$, $\lambda_2 \approx +0.084$) and Kaplan–Yorke dimension $D\_KY \approx 3.019$ to generate time-varying, bounded perturbations for adversarial defense. Three key contributions distinguish this work from prior chaos-based defenses. First, a dimensionality-consistent injection mechanism based on layer-specific projection operators P_l resolves the dimensional incompatibility between the 4D hyperchaotic signal and arbitrary-dimension neural network layers, making the mathematical formulation rigorous for any architecture. Second, a complete adaptive sliding-mode synchronization proof (Theorem 1) — including the explicit adaptation law $d\hat{\Delta}/dt = \kappa\|s\|$ and the full Barbalat-based convergence argument — guarantees that the hyperchaotic perturbations remain bounded and reproducible, a necessary condition for practical deployment. Third, experimental evaluation against adaptive attacks (BPDA+PGD) confirms that the robustness of CNNS is genuine: on CIFAR-10, the framework achieves 58.7% under BPDA+PGD versus 48.1% for adversarial training, while maintaining 92.7% clean accuracy (0.7 pp degradation). Complete clean accuracy reporting for both MNIST and CIFAR-10 enables transparent robustness–accuracy trade-off analysis. Overall, the CNNS framework establishes a theoretically grounded and empirically validated paradigm for adversarial defense, moving beyond gradient-obfuscation strategies toward dynamical, chaos-driven security mechanisms with provable stability guarantees.

Future Work

- Extension to transformer-based architectures, where attention mechanisms could be modulated by the projected hyperchaotic signal via P_l applied to the key-query-value matrices.

- Adaptive tuning of γ, γ_w, and δ using reinforcement learning to dynamically optimize the robustness–accuracy trade-off against unknown threat models.
- Integration into federated learning environments, where distributed master-slave synchronization could provide decentralized security guarantees without sharing model parameters.
- Hardware implementation (FPGA/SoC) for real-time cyber-physical systems, leveraging the intrinsic hardware compatibility of chaotic differential equations with analog computation platforms.
- Theoretical analysis of the robustness certificate radius achievable by the proposed framework, analogous to the ℓ_2 certified radius of randomized smoothing (36).
- Evaluation under AutoAttack (49) (Croce and Hein, ICML 2020), a parameter-free ensemble benchmark combining APGD-CE, APGD-DLR, FAB, and Square attacks, which constitutes the current community standard for reliable robustness assessment. This evaluation requires a stronger base architecture (e.g. WideResNet-28) to produce results comparable to the RobustBench leaderboard, and is deferred to a dedicated extended study.

Acknowledgment

The authors thank the Department of Fundamental Sciences, University of Moundou, for institutional support. F. Djimasra acknowledges support from the University of Moundou Research Fund. This work received no specific external funding.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Krizhevsky, A. I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.
2. He, K. X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 770–778.
3. Goodfellow, I. J. J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015.
4. Madry, A. A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, Vancouver, BC, Canada, 2018.
5. Papernot, N. P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in Proc. IEEE S&P, San Jose, CA, USA, 2016.
6. Athalye, A. N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proc. ICML, Stockholm, Sweden, 2018.
7. Strogatz, S. H. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Boulder, CO, USA: Westview Press, 2014.
8. Pecora L. M. and T. L. Carroll, "Synchronization in chaotic systems," Phys. Rev. Lett., vol. 64, no. 8, pp. 821–824, 1990.
9. Chai X. et al., "An efficient chaos-based image encryption scheme with dynamic S-boxes," IEEE Access, vol. 7, pp. 184417–184427, 2019.
10. Guo, C. M. Rana, M. Cissé, and L. van der Maaten, "Countering adversarial images using input transformations," in Proc. ICLR, Vancouver, BC, Canada, 2018.
11. Pedraza, A. O. Deniz, and G. Bueno, "Approaching adversarial example classification with chaos theory," Entropy, vol. 22, no. 11, p. 1201, 2020.
12. Gadallah, O. A. H. A. Abd El-Halim, A. E. Hassanien, and A. A. A. El-Aziz, "Chaotic neural network quantization and its robustness against adversarial attacks," Knowledge-Based Syst., vol. 278, p. 110933, 2023.
13. Al-Muhammed M. J. and R. Abu Zitar, "Encryption technique based on chaotic neural network space shift," Sci. Rep., vol. 12, no. 1, p. 18498, 2022.
14. Singh, H. A. Kumar, and R. Sharma, "Adversarial examples detection with chaos-based multivariate features," Int. J. Mach. Learn. Cybern., vol. 16, no. 2, pp. 341–358, 2025.
15. Gómez, J. M. García, and P. López, "Leveraging autoencoders and chaos theory to improve adversarial example detection," Neural Comput. Appl., vol. 36, pp. 4211–4228, 2024.
16. Rössler, O. "An equation for hyperchaos," Phys. Lett. A, vol. 71, no. 2–3, pp. 155–157, 1979.
17. Álvarez G. and S. Li, "Some basic cryptographic requirements for chaos-based cryptosystems," Int. J. Bifurcat. Chaos, vol. 16, no. 8, pp. 2129–2151, 2006.
18. Kocarev, L. "Chaos-based cryptography: A brief overview," IEEE Circuits Syst. Mag., vol. 1, no. 3, pp. 6–21, 2001.
19. Utkin, V. I. Sliding Modes and Their Applications in Variable Structure Systems. Moscow, Russia: Mir, 1978.
20. Slotine J.J. and W. Li, Applied Nonlinear Control. Englewood Cliffs, NJ, USA: Prentice-Hall, 1991.
21. Carlini N. and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Secur. Privacy, San Jose, CA, USA, 2017, pp. 39–57.
22. Alligood, K. T. T. D. Sauer, and J. A. Yorke, Chaos: An Introduction to Dynamical Systems. New York, NY, USA: Springer, 1996.
23. Chen G. and X. Dong, From Chaos to Order: Methodologies, Perspectives and Applications. Singapore: World Scientific, 1998.
24. Lü J. and G. Chen, "A new chaotic attractor coined," Int. J. Bifurcat. Chaos, vol. 12, no. 3, pp. 659–661, 2002.

25. Sparrow, C. The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. New York, NY, USA: Springer, 1982.
26. LeCun, Y. Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
27. Goodfellow, I. Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
28. Simonyan K. and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, San Diego, CA, USA, 2015.
29. Moosavi-Dezfooli, S. A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 2574–2582.
30. Szegedy, C. W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. ICLR, Banff, AB, Canada, 2014.
31. Tramèr, F. A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in Proc. ICLR, Vancouver, BC, Canada, 2018.
32. Zhang, H. Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in Proc. ICML, Long Beach, CA, USA, 2019.
33. Shafahi, A. M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in Proc. NeurIPS, Vancouver, BC, Canada, 2019.
34. Lorenz, E. N. "Deterministic nonperiodic flow," J. Atmos. Sci., vol. 20, no. 2, pp. 130-141, 1963.
35. Lyapunov, A. M. "The general problem of the stability of motion," Int. J. Control, vol. 55, no. 3, pp. 531–534, 1992.
36. Cohen, J. E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in Proc. ICML, Long Beach, CA, USA, 2019.
37. Wong E. and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in Proc. ICML, Stockholm, Sweden, 2018.
38. Li, B. C. Chen, W. Wang, and L. Carin, "Certified adversarial robustness with additive Gaussian noise," in Proc. NeurIPS, Vancouver, BC, Canada, 2019.
39. Khalil, H. K. Nonlinear Systems, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
40. Benettin, G. L. Galgani, A. Giorgilli, and J.-M. Strelcyn, "Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems: A method for computing all of them," Meccanica, vol. 15, pp. 9–30, 1980.
41. Farmer, J. D. E. Ott, and J. A. Yorke, "The dimension of chaotic attractors," Physica D, vol. 7, no. 1–3, pp. 153–180, 1983.
42. Bernstein, D. S. Matrix Mathematics: Theory, Facts, and Formulas, 2nd ed. Princeton, NJ, USA: Princeton Univ. Press, 2009.
43. Papernot, N. P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in Proc. ACM ASIACCS, Abu Dhabi, UAE, 2017.
44. LeCun, Y. L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.
45. Krizhevsky, A. "Learning multiple layers of features from tiny images," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, 2009.
46. Kingma D. P. and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, San Diego, CA, USA, 2015.
47. Djimasra, F. J. D. D. Nkapkop, N. Tsafack, A. Boukabou, J. Fotsin, and J. Y. Effa, "Robust cryptosystem using a new hyperchaotic oscillator with striking dynamic properties," Multimed. Tools Appl., vol. 80, pp. 25583–25605, 2021.
48. Tsafack, N. J. Sankar, B. Abd-El-Atty, J. D. D. Nkapkop, C. Hirota, A. A. and Abd El-Latif, "Design and implementation of a simple dynamical 4-D chaotic circuit with applications in image encryption," Inf. Sci., vol. 515, pp. 191–217, 2020.
49. Croce F. and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in Proc. Int. Conf. Mach. Learn. (ICML), Vienna, Austria, 2020, pp. 2206–2216.

****