

Review Article

Encryption and Privacy-Preserving Strategies in Deep Learning: A Review

Siham A. Al-Athori^{1*}, Najib Ahmed Mohammed²¹ Computer Science Program, Faculty of Graduate Studies, National University, Yemen² Independent Researcher, London, United Kingdom* Corresponding Author: Siham A. Al-Athori (siham.alathori@nu.edu.ye)

Article Info

ISSN (online): 2979-2967

Volume: 1 (2026)

Issue: 2

Received: 29 June 2026

Accepted: 25 July 2026

Published: 10 August 2026

Page No: 74–85

DOI:

<https://doi.org/10.67373/mirj.ea.2026.001>

Cite this article:

Al-Athori, S. A., & Mohammed, N. A. (2026). Encryption and privacy-preserving strategies in deep learning: A review. *MIRJ*, 1(2), 74–85.**Keywords:** deep learning, image encryption, convolutional neural networks, GAN, CycleGAN, privacy preservation, federated learning, homomorphic encryption

Abstract

Deep neural networks have recently been applied to image encryption and privacy preservation, exploiting their pattern-analysis and feature-extraction capacity. This review surveys the field along three families — feature-based models (autoencoders, CNNs), generative models (GANs, CycleGAN), and weight-based or hybrid schemes — and consolidates the reported cryptographic performance of representative works published between 2015 and 2024. We define the evaluation metrics used throughout, including information entropy, NPCR, UACI, PSNR and SSIM, and apply them consistently in a comparative analysis (Table 2). Our findings indicate that CNN-and-chaos schemes lead on classical encryption metrics (entropy ≈ 8 , NPCR $\approx 99.6\%$, UACI $\approx 33.4\%$), that CycleGAN-based methods trade pixel-level fidelity for strong anonymisation, and that weight-as-key approaches offer lossless recovery suited to resource-constrained deployment. We distinguish encryption, anonymisation and steganography, note that high entropy is necessary but not sufficient for security, and outline open challenges in generalisability, interpretability and quantum-era robustness. We further situate these encryption families against the complementary privacy-preserving learning strategies — homomorphic inference, federated learning and differential privacy — and map the security and efficiency analyses actually disclosed across the surveyed schemes, finding that key-space, key-sensitivity, randomness-suite and runtime results are reported only sporadically.

1. INTRODUCTION

Data protection is increasingly important in the digital age, as sensitive and personal information becomes vulnerable to escalating cyber-threats. Image encryption is a key means of protecting privacy and ensuring information integrity. With the advance of artificial intelligence, deep learning has emerged as an effective tool for image encryption, enabling new ways of processing encrypted data. Deep-learning techniques rely on deep neural networks

with a strong capacity to extract patterns and features from large volumes of data.

A turning point came in 2012, when the AlexNet model of Krizhevsky et al. (2012) won the ImageNet competition, drawing wide interest from research and industry. Deep learning has since progressed rapidly with models such as VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), Inception (Szegedy et al., 2017) and YOLO (Redmon et al., 2016), improving accuracy in tasks such as image recognition and natural-language processing. Deep networks are trained by optimisation and backprop-

agation over multiple layers (Higham and Higham, 2019), and encompass deep neural networks, recurrent networks, long short-term memory networks, generative adversarial networks (GANs) (Goodfellow et al., 2014), and attention/transformer mechanisms.

This paper reviews the use of deep learning for image encryption and privacy preservation by surveying the current literature. Relative to prior surveys, our contribution is four-fold: (i) a single, consistently defined metric set applied across all surveyed works; (ii) an explicit separation of encryption, anonymisation and steganography, which are frequently conflated; (iii) a reconciled citation base and a corrected comparative table; and (iv) a positioning of image-encryption techniques against the complementary privacy-preserving learning strategies — homomorphic inference, federated learning and differential privacy — together with a systematic map of which security and efficiency analyses the surveyed schemes actually disclose. The remainder of the paper is organised as follows. Section 2 states the review methodology and positions this review against prior surveys. Section 3 presents encryption techniques. Section 4 situates the complementary privacy-preserving learning strategies. Section 5 defines the performance measures. Section 6 outlines a threat model. Section 7 analyses previous works. Sections 8–10 cover applications, challenges and future directions, and Section 11 concludes.

2. REVIEW METHODOLOGY

This review follows a structured, narrative protocol. We searched IEEE Xplore, Scopus, ACM Digital Library, ScienceDirect and arXiv using combinations of the terms “deep learning”, “image encryption”, “privacy preserving”, “autoencoder”, “GAN”, “CycleGAN”, “chaotic map” and “network weights”. The window 2015–2024 was chosen because 2015 marks the first works applying distributed/deep learning to privacy-preserving image tasks, while 2024 captures the most recent hybrid GAN-and-chaos schemes. Inclusion criteria were: peer-reviewed or archival works that (a) apply a deep model to image encryption, anonymisation or steganography and (b) report at least one quantitative security or fidelity metric. We excluded works without empirical evaluation and those not addressing image data.

The search returned 216 records; after de-duplication and title/abstract screening, 44 full-text articles were assessed for eligibility, of which 13 empirical schemes met all inclusion criteria and are compared in Table 2. The 31 full-text exclusions were for lack of empirical evaluation, not addressing image data, or reporting no quantitative security or fidelity metric; per-reason counts were not recorded at screening time. All screening was performed by the author. The selection process is summarised in the flow diagram of Fig. 1. Additional foundational and background works (architectures, metrics and privacy techniques) are cited where relevant. Where a work reported a metric under a non-standard definition, we normalised it or marked it “not

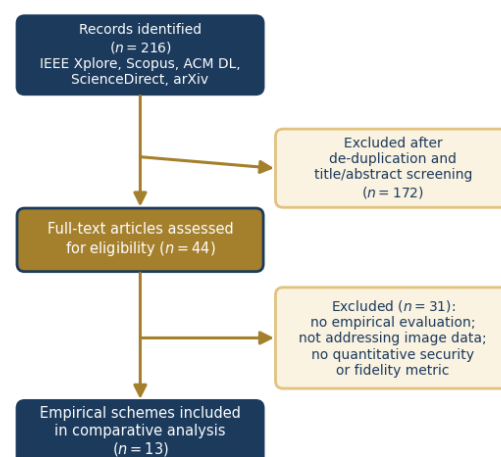


Figure 1. Study selection flow: 216 records identified, 44 full-text articles assessed, 13 empirical schemes included in the comparative analysis.

Table 1. Comparison with adjacent prior surveys.

Survey	Families	Metrics	Task sep.	Discl. map
Mhawi et al. (2024)	GANs only	×	×	×
Zhu et al. (2019)	CycleGAN only	×	×	×
Mienye and Swart (2025)	AEs only	×	×	×
This review	Three + compl.	✓	✓	✓

reported” (n/r). A taxonomy of the field is shown in Fig. 2.

A supplementary screening pass covering January 2025 to June 2026 was carried out during revision. The most recent developments combine *classical* chaos-based ciphers with post-quantum key-encapsulation mechanisms rather than deep models, while generative-diffusion-model approaches to image protection — for example, key-controlled inversion of public diffusion models (Zhang et al., 2026) — remain at the preprint stage. No additional peer-reviewed deep-learning scheme meeting all inclusion criteria was identified, so Table 2 is unchanged by the supplement; we return to these emerging directions in Section 10.

2.1 Relation to prior surveys

Existing surveys adjacent to this review are each confined to a single technique family: Mhawi et al. (2024) review GANs for image-security applications, Zhu et al. (2019) briefly review CycleGAN itself, and Mienye and Swart (2025) review autoencoders across application domains without a security focus. None applies one consistently defined metric set across families, separates encryption from anonymisation and steganography, or maps the security analyses disclosed by the primary studies. Table 1 summarises the comparison; these three properties are the contributions (i), (ii) and (iv) claimed in Section 1.

3. IMAGE ENCRYPTION TECHNIQUES USING DEEP LEARNING

Traditional encryption performs rounds of substitution and diffusion on image pixels using secret keys. Deep-learning methods offer complementary approaches that enhance security and efficiency. The most prominent families are feature-based models, generative models, and weight-based or hybrid schemes.

3.1 Deep Autoencoders in Encryption

An autoencoder is a neural network that learns a compressed representation of data without labels (Bank et al., 2023; Mienye and Swart, 2025). It consists of an **encoder**, which maps the input image to a lower-dimensional representation, and a **decoder**, which reconstructs the original image. We note that a vanilla autoencoder is reconstructive rather than generative; the capacity to synthesise new images belongs to variational or adversarial variants and is not assumed here.

In the encryption setting, the original image X is passed through the encoder to produce a compressed representation (the ciphertext), and the decoder reconstructs the image. The encoder computes a linear map followed by a non-linear activation such as ReLU or Sigmoid (Datta, 2020):

$$Z = WX + b, \quad A = f(Z) \quad (1)$$

where W and b are the layer weights and bias and f is the activation. Training minimises the mean-squared reconstruction loss

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{X}_i)^2 \quad (2)$$

with the weights updated by gradient descent and back-propagation until convergence.

3.2 Generative Adversarial Networks in Image Encryption

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are generative models trained by an adversarial game between a **generator** G , which produces images, and a **discriminator** D , which distinguishes real from generated images; their use in image security has been reviewed recently (Mhawi et al., 2024). In cryptographic use, G is trained to produce encrypted images that are difficult to interpret without the model or an appropriate key. The generator maps a latent variable z to an image, and the discriminator outputs the probability that an image is real:

$$G : z \mapsto x_g, \quad D : x \mapsto [0, 1] \quad (3)$$

The defining objective is the two-player min–max game introduced by Goodfellow et al. (2014):

$$\min_G \max_D \mathbb{E}_x [\log D(x)] + \mathbb{E}_z [\log(1 - D(G(z)))] \quad (4)$$

Most studies train the generator to produce data hard to distinguish from the original, creating an implicit encryption without traditional keys. Reported extensions include

convolutional encryption layers, conditional/ID-based keys, and hybrids with chaotic maps. Practical challenges include imperfect reconstruction when a decoder is included, sensitivity to adversarial perturbation, and vulnerability if model weights are compromised.

3.3 CycleGAN Networks in Image Encryption

CycleGAN, introduced by Zhu et al. (2017), transforms images between two domains without paired training data (Zhu et al., 2019). It uses two generators (G_{AB} and G_{BA}) and two discriminators. For an image x in domain A , the forward and inverse mappings are

$$y = G_{AB}(x), \quad \hat{x} = G_{BA}(y) = G_{BA}(G_{AB}(x)). \quad (5)$$

The component that defines CycleGAN is the cycle-consistency loss, which we state explicitly:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_x \|G_{BA}(G_{AB}(x)) - x\|_1 + \mathbb{E}_y \|G_{AB}(G_{BA}(y)) - y\|_1. \quad (6)$$

CycleGAN is not an autoencoder and does not guarantee exact inversion; cycle-consistency is a soft constraint, so we describe it as approximately reversible rather than invertible. In the literature it has been used less as a conventional cipher and more as a framework for near-reversible domain translation supporting image hiding (steganography), semantic-level anonymisation, and encryption via domain translation, sometimes combined with chaotic maps for key generation. Reported limitations include the possibility of training an approximate inverse, loss of fine detail, and poor cross-dataset generalisability.

3.4 Weight-Based and Hybrid Schemes

The third family in Fig. 2 treats the trained network itself — specifically its weight tensors — as key material. A trained deep network embodies millions of high-precision parameters whose values are unique to the training run; deep models are, moreover, highly sensitive to targeted weight perturbations (Grosse et al., 2019), which is precisely the property a key must have. In the general form of these schemes, a keystream or permutation is derived from selected weight tensors W_ℓ of a secret model,

$$K = \phi(W_\ell), \quad C = \pi_K(P) \oplus K', \quad (7)$$

where ϕ maps weights to a pseudorandom sequence, π_K is a key-driven pixel permutation and $\oplus$ denotes XOR diffusion with a weight-derived sequence K' . Decryption inverts the permutation and XOR exactly, so recovery is lossless by construction — no reconstruction network is involved.

Representative instantiations include the chaotic-weights approach of Lin et al. (2021), in which a chaotic map drives a selective permutation of weight positions so that the network functions correctly only for a key-holder — conceived for intellectual-property protection but equally a template for weight-as-key encryption — and the artificial-neural-network permutation–diffusion cipher of Panigrahy et al. (2024), which reports near-ideal differential metrics (NPCR $\approx 99.6\%$, UACI $\approx 33.4\%$) with SSIM ≈ 0.002

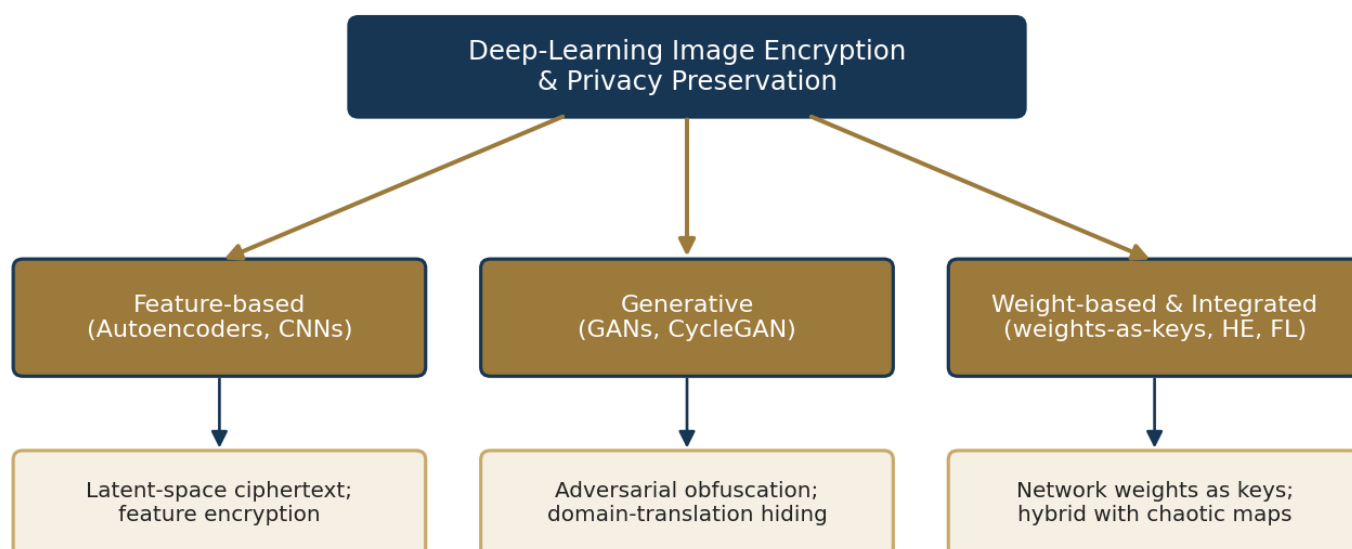


Figure 2. Taxonomy of deep-learning approaches to image encryption and privacy preservation (author's own illustration).

$$\text{Loss} = (1/N) \sum_i (X_i - \hat{X}_i)^2 \rightarrow \text{backprop / gradient descent}$$

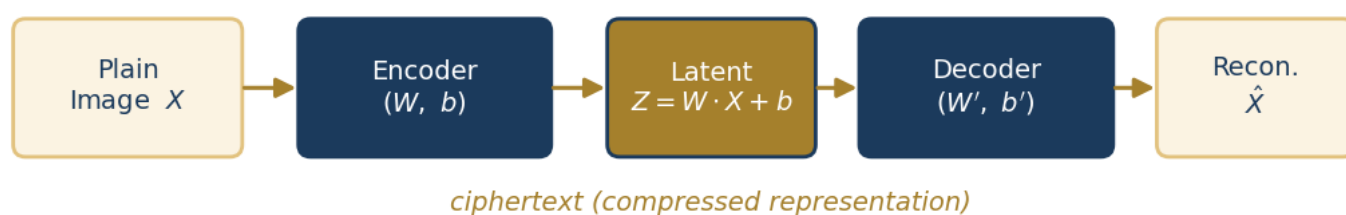


Figure 3. Autoencoder encryption pipeline: the latent representation serves as ciphertext (author's own illustration).

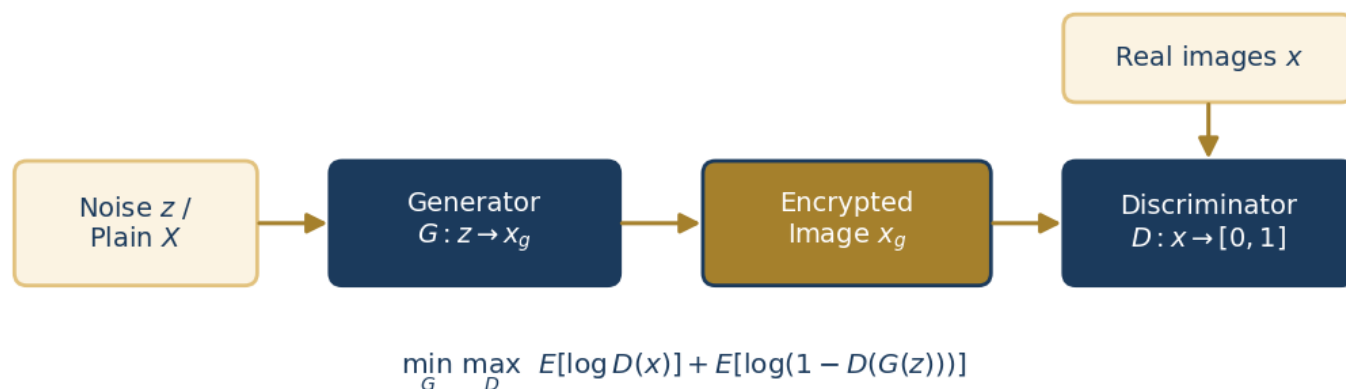


Figure 4. GAN-based encryption: the generator produces the encrypted image while the discriminator drives adversarial training (author's own illustration).

between plain and encrypted images and ≈ 1 after decryption. In collaborative settings, combined learning of weights across parties (Ioste et al., 2022) extends the idea from protecting data to protecting the model jointly.

The strengths of this family are practical: exact (loss-less) decryption, low computational cost at encryption time since no forward pass through a deep generator is required, and simultaneous protection of the model as intellectual property and of the data it processes — attractive

for IoT and embedded deployment. Its limitations mirror its premise. Because the weights *are* the key, any weight exposure — a leaked checkpoint, a model-extraction attack — is a total key compromise (Section 6); the effective key space is bounded by weight precision and shrinks under quantisation; keys change whenever the model is retrained; and, as with the other families, no formal security proofs comparable to those of classical ciphers are yet available.

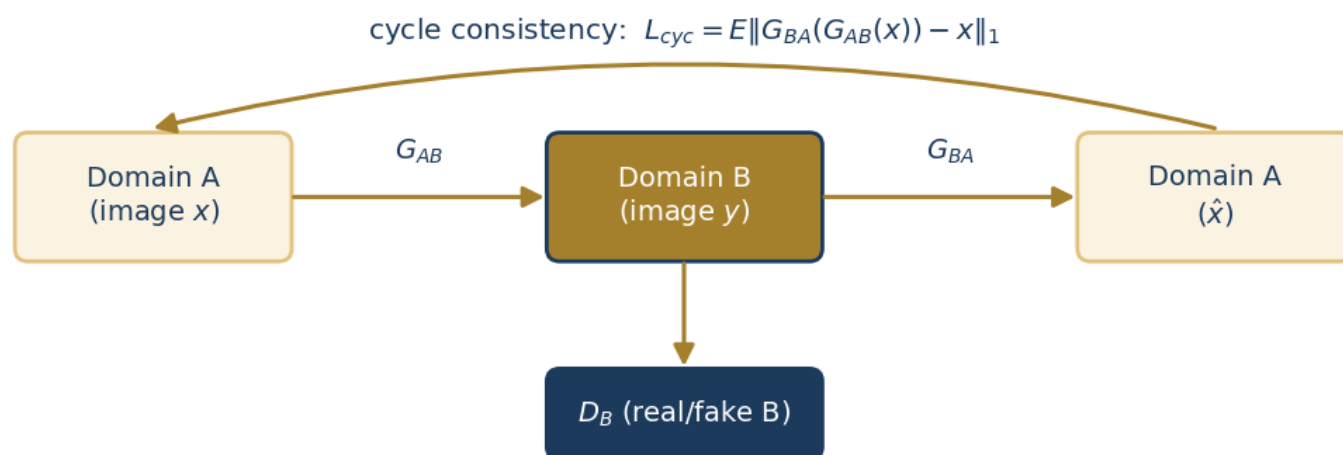


Figure 5. CycleGAN domain translation with cycle-consistency (author's own illustration); used chiefly for anonymisation and steganography rather than keyed encryption.

4. COMPLEMENTARY PRIVACY-PRESERVING LEARNING STRATEGIES

The families of Section 3 protect *images* — data at rest or in transit. The title of this review, however, spans privacy-preserving strategies in deep learning more broadly, and three further families protect data *during computation*: homomorphic inference, federated learning and differential privacy. They appear in the third branch of Fig. 2 and recur throughout the surveyed literature as companions to image encryption, so we position them explicitly here.

4.1 Homomorphic encryption and encrypted inference

Fully homomorphic encryption (FHE), established by Gentry (2009), permits computation directly on ciphertexts. Applied to deep learning, it allows a model to classify an image that is never decrypted: CryptoNets (Gilad-Bachrach et al., 2016) demonstrated neural-network inference on homomorphically encrypted MNIST digits with throughput of tens of thousands of predictions per hour. The costs are substantial — ciphertext expansion, orders-of-magnitude runtime overhead, and the need to replace non-polynomial activations (ReLU, Sigmoid) with low-degree polynomial approximations, which constrains network depth and accuracy. A complementary line avoids homomorphic overhead by operating learned models directly in the encrypted domain: deep convolutional networks can classify and retrieve encrypted images without decrypting them (Al-Athori et al., 2026), bringing encrypted-domain *search* within reach of the same infrastructure that stores the ciphertexts. Homomorphic and encrypted-domain inference therefore complement rather than replace image encryption: they protect data during third-party *computation*, whereas the schemes of Section 3 protect storage and transmission.

4.2 Federated learning and secure aggregation

Federated learning (McMahan et al., 2017) keeps training data on the devices or institutions that own it: participants train locally and share only model updates, which a server averages into a global model. This removes the need

to centralise sensitive images — the setting anticipated by the earliest work in our corpus, the distributed privacy-preserving training of Shokri and Shmatikov (2015). Model updates themselves can leak information about the underlying data, however, so practical deployments add secure aggregation (Bonawitz et al., 2017), in which cryptographic masking ensures the server learns only the *sum* of updates, never an individual contribution. Weight-sharing schemes such as Iosif et al. (2022) sit at the boundary of this family and the weight-based encryption family of Section 3.4.

4.3 Differential privacy

Differential privacy (DP) (Dwork et al., 2006) gives a formal guarantee: the distribution of a mechanism's output changes by at most a factor e^ϵ when any single record is added to or removed from the data. In deep learning it is realised chiefly by DP-SGD (Abadi et al., 2016), which clips per-example gradients and adds calibrated Gaussian noise during training, with a moments accountant tracking the cumulative privacy budget (ϵ, δ) . Shokri and Shmatikov (2015) applied the same principle to selective parameter sharing. DP protects against inference *about training data* from a released model; it does not conceal the images themselves, and its utility–privacy trade-off (accuracy degrades as ϵ tightens) is the direct analogue of the fidelity–security trade-off in Section 3.

In combination, the strategies compose naturally: images can be stored and transmitted under the encryption families of Section 3, contributed to model training federatively with secure aggregation, and released as model outputs under differential privacy — a layered architecture that several of the medical applications in Section 8 already approximate.

5. PERFORMANCE MEASURES

Deep-learning image-encryption schemes are assessed by reconstruction-quality, computational, and cryptographic-strength metrics.

5.1 Reconstruction quality

Mean squared error (MSE) and peak signal-to-noise ratio (PSNR = $10 \log_{10}(\text{MAX}^2/\text{MSE})$, in dB) measure fidelity between original and decrypted images, with higher PSNR indicating better recovery. The structural similarity index (SSIM) of Wang et al. (2004) captures perceptual similarity via local means, variances and covariance.

5.2 Cryptographic strength

Information entropy $H(X) = -\sum_i p(x_i) \log_2 p(x_i)$ (Shannon, 1948); for 8-bit images, values near 8 indicate strong randomness. Histogram uniformity is tested by the chi-square statistic.

Because Section 7 relies on them, we define the differential-attack metrics explicitly, following Wu et al. (2011). The number of pixels change rate (NPCR) and unified average changing intensity (UACI) compare two ciphertexts C_1 and C_2 whose plaintexts differ in a single pixel:

$$\text{NPCR} = \frac{1}{MN} \sum_{i,j} D(i,j) \times 100\% \quad (8)$$

where

$$D(i,j) = \begin{cases} 0, & C_1(i,j) = C_2(i,j) \\ 1, & \text{otherwise} \end{cases}$$

$$\text{UACI} = \frac{1}{MN} \sum_{i,j} \frac{|C_1(i,j) - C_2(i,j)|}{255} \times 100\% \quad (9)$$

For 8-bit images the ideal values are $\text{NPCR} \approx 99.6094\%$ and $\text{UACI} \approx 33.4635\%$ (Wu et al., 2011). We also report the adjacent-pixel correlation coefficient $r = \text{cov}(x, y) / (\sigma_x \sigma_y)$, which should approach 0 for a secure cipher, and recommend key-sensitivity analysis — the NPCR/UACI between ciphertexts produced under keys differing by a single bit — together with key-space size as primary security indicators. We stress that high global entropy is necessary but not sufficient; local entropy and standard randomness tests (e.g. NIST SP 800-22 (Rukhin et al., 2010)) should accompany any strong-security claim.

5.3 Computational cost

Deployment feasibility is governed by a third metric group: encryption and decryption latency (milliseconds per image at a stated resolution), throughput (images or megabytes per second), and model complexity (parameter count or on-disk size, which determines memory footprint on embedded hardware). For learned schemes the one-off *training* cost should be separated from the recurring *inference* cost of encrypting or decrypting a single image: a scheme with expensive training but cheap inference can still suit IoT deployment, whereas a scheme that requires a full generator forward-and-inverse pass per image may not. Disclosure of these metrics across the surveyed works is markedly poorer than for the security metrics above — several works claim “real-time” or “lightweight” operation without reporting hardware, resolution or timings. We therefore record computational disclosure alongside security disclosure in Table 3 rather than tabulating incom-

parable numbers, and we treat this reporting gap as a finding in its own right (Section 9).

6. THREAT MODEL AND SECURITY CONSIDERATIONS

To make the comparison in Section 7 meaningful, we state the adversary model assumed by most surveyed schemes. The adversary is typically assumed to observe ciphertext (ciphertext-only) and, in stronger analyses, to mount known- or chosen-plaintext attacks; the objective is either plaintext recovery or key recovery. Learned schemes face two additional vectors absent from classical ciphers: model-inversion or approximate-inverse attacks that reconstruct content from a compromised or surrogate network, and adversarial perturbations that corrupt decryption. Schemes that expose or reuse model weights must therefore treat those weights as key material. Anonymisation and steganography methods target a different notion — unlinkability or undetectability — and should not be judged against keyed-cipher benchmarks; we keep these categories separate in Table 2.

7. ENCRYPTION PERFORMANCE ANALYSIS OF PREVIOUS WORKS

We review representative studies on deep-learning image encryption and summarise them in Table 2; Fig. 6 places the 13 schemes on a timeline by technique family. Development between 2015 and 2024 spans convolutional networks, autoencoders, GAN variants (including CycleGAN), and the recent use of network weights as keys, alongside the complementary strategies of Section 4. We discuss the schemes family by family, and consolidate in Table 3 which security and efficiency analyses each primary source actually discloses.

Training-privacy origins. The earliest work in the corpus predates image encryption proper: Shokri and Shmatikov (2015) protected privacy during training via distributed selective parameter sharing with differential privacy, reporting model accuracy rather than ciphertext metrics. Ioste et al. (2022) continue this line by learning combined neural-network weights across collaborating parties, reporting accuracies of 77–95% on breast-cancer, COVID-19 and CIFAR-10 tasks. Neither produces a ciphertext, so neither can be judged on entropy or differential metrics — they are tabulated as training-privacy schemes and are better read alongside Section 4 than against keyed ciphers.

CNN and autoencoder schemes with chaotic maps. This family holds the strongest classical results. Alhakak et al. (2025) combine an encoder–decoder with a Logistic map for chest X-ray encryption, reporting entropy 7.99, NPCR/UACI at the ideal values, PSNR above 40 dB and SSIM 0.99 — although evaluation on a single imaging modality leaves cross-dataset generalisability untested. Long et al. (2024) pair deep feature encoding/decoding with AES across MNIST, CIFAR-10 and ImageNet, reaching entropy 7.999; the cryptographic strength here is substantially inherited from the embedded AES layer, with

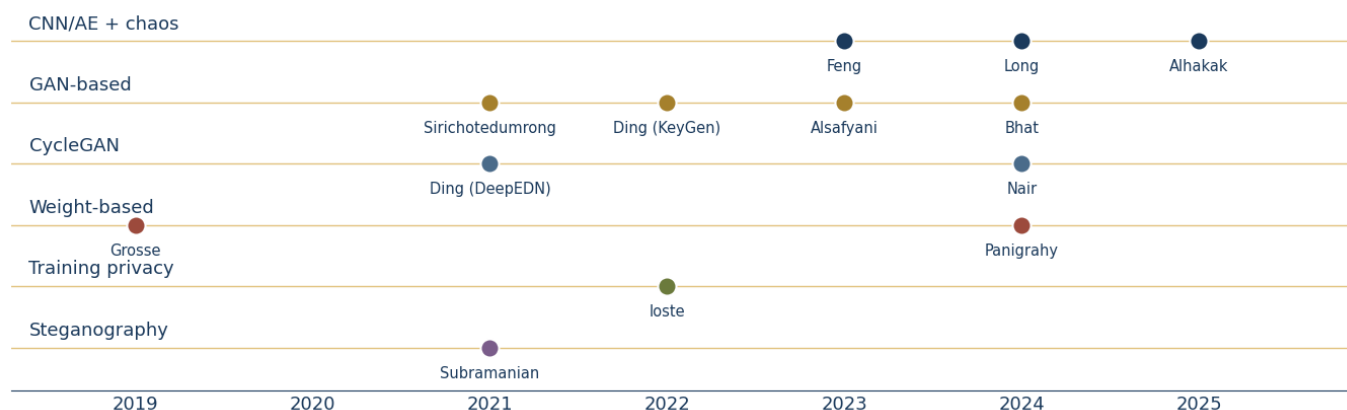


Figure 6. The 13 compared schemes by publication year and technique family (data from Table 2).

the network contributing representation rather than security, which makes the scheme robust but less informative about deep learning as a cipher. Feng et al. (2023) interleave chaotic scrambling with a CNN encoder–decoder on COCO and standard test images (entropy ≈ 7.99 , PSNR 35–40 dB, SSIM ≈ 1), but do not report NPCR/UACI, leaving their differential resistance unquantified. Earlier CNN-and-chaos work using Logistic, Tent and Chebyshev maps (Panwar et al., 2023) established the pattern: entropy near 8, NPCR above 99.6%, UACI near 33%, PSNR of 40–43 dB.

GAN-based key generation, transformation and analysis. DeepKeyGen (Ding et al., 2022) uses a GAN as a private-key stream generator for XOR encryption of medical images; its UACI of 20–24% falls well short of the 33.46% ideal, indicating residual statistical structure between related ciphertexts. The GAN transformation of Sirichotedumrong and Kiya (2021) conceals identity while preserving classification utility (accuracy 90.7%, SSIM 0.09 against the plain image), but — consistent with Section 6 — it is keyless and therefore anonymisation, not encryption; it cannot resist an adversary who trains an approximate inverse. Bhat and Nanjundegowda (2024) provide a metric-focused analysis of CryptoGAN-style encryption on MNIST, CIFAR-10 and face data, with fidelity values that are dataset-specific rather than summarisable in a single number, and Alsafyani et al. (2023) combine a secure CryptoGAN with an optical chaotic map for face images, reporting entropy ≈ 8 with fidelity on a normalised per-dataset scale — in both cases the non-standard reporting impedes direct comparison, which is precisely the motivation for Table 3.

CycleGAN medical encryption. DeepEDN (Ding et al., 2021) integrates CycleGAN with a ResNet-based region-of-interest mechanism for Internet of Medical Things (IoMT) images, reporting entropy 7.94–7.97 and SSIM ≈ 0.01 between original and encrypted images (effective anonymisation) while exceeding 0.90 after decryption (high-quality recovery). Its NPCR of 94.2–94.8%, however, is far below the 99.61% ideal of Section 5: roughly one pixel in twenty remains unchanged when a single plaintext

pixel flips, a differential weakness that is acceptable for anonymisation-oriented medical transport but does not meet keyed-cipher standards. Nair et al. (2024) address exactly this by adding chaotic key generation to the CycleGAN pipeline, restoring NPCR/UACI to the ideal values on chest X-rays while retaining the anonymisation profile (SSIM 0.01 encrypted, ≥ 0.90 decrypted).

Weight-based schemes. Following the intellectual-property protection of Lin et al. (2021) (Section 3.4), the permutation–diffusion–XOR cipher with weights as keys (Grosse et al., 2019) and the ANN-based cipher of Panigrahy et al. (2024) achieve lossless recovery (PSNR $\rightarrow \infty$, SSIM ≈ 1 after decryption; SSIM ≈ 0.002 between plain and encrypted) with NPCR/UACI at the ideal values, and both claim real-time operation — though neither reports full hardware-qualified timings. The open question for this family is key management: the key *is* the model, so its distribution and revocation inherit every difficulty of secure model distribution.

Steganography. The convolutional-autoencoder steganography of Subramanian et al. (2021) hides images end-to-end in cover images (COCO, CelebA, ImageNet) with stego–cover SSIM ≈ 1 and PSNR of 37–40 dB. Its security goal is undetectability, which none of the metrics of Section 5 measures; resistance to modern steganalysis is not evaluated, and we tabulate it separately for that reason.

Overall, CNN-and-chaos hybrids lead on classical metrics, CycleGAN adds anonymity at some fidelity cost (with Nair et al. (2024) showing the differential gap is closable by explicit keys), and weight-based schemes offer practical lossless solutions for IoT and embedded deployment. Table 3 shows the other side of the picture: key-space and key-sensitivity analyses are disclosed only partially and chiefly by the chaos-based schemes, complete randomness-suite results are absent from the corpus, and runtime is claimed more often than it is measured.

8. ENCRYPTION APPLICATIONS USING DEEP LEARNING

Table 2. Cryptographic performance of representative works (consistent units, task-type column, reconciled citations).

Study	Model / method	Task type	Dataset	Entropy	NPCR / UACI	PSNR (dB)	SSIM (enc / dec)
Alhakak et al. (2025)	Encoder–Decoder + Logistic map	Encryption (keyed)	Chest X-ray	7.99	99.6% / 33%	>40	0.99
Long et al. (2024)	DL feature encode/decode + AES	Encryption	MNIST, CIFAR-10, ImageNet	7.999	99.6% / 33%	40	0.99
Nair et al. (2024)	CycleGAN + chaotic maps	Encryption	Chest X-ray	7.94–7.97	99.6% / 33.4%	35–38	0.01 / ≥0.90
Ding et al. (2021)	DeepEDN (CycleGAN + ROI)	Encryption (medical)	IoMT medical	7.94–7.97	94.2–94.8% / ~33%	35–38	0.01 / ≥0.90
Ding et al. (2022)	DeepKeyGen (GAN key + XOR)	Key generation	Medical images	≈7.99	99.6% / 20–24%	n/r	n/r
Sirichotedumrong and Kiya (2021)	GAN transformation (no key)	Anonymisation	CIFAR-10/100, STL-10	n/r	n/r	n/r	0.09 (acc 90.7%)
Ioste et al. (2022)	Combined NN-weight learning	Privacy (training)	Breast cancer, COVID-19, CIFAR-10	n/r	n/r	n/r	acc 77–95%
Panigrahy et al. (2024)	ANN permutation + XOR + weights	Encryption	Standard test images	≈7.99	99.6% / 33.4%	n/r	0.002 / ≈1
Subramanian et al. (2021)	Deep conv. autoencoder + stego	Steganography	COCO, CelebA, ImageNet	n/r	n/r	37–40	≈1 (stego/cover)
Feng et al. (2023)	Scrambling + CNN encoder–decoder	Encryption	COCO + standard	≈7.99	n/r	35–40	≈1
Bhat and Nanjundegowda (2024)	CryptoGAN (metric analysis)	Encryption (analysis)	MNIST, CIFAR-10, faces	≈8	n/r	n/r	dataset-specific*
Alsafyani et al. (2023)	Secure CryptoGAN + optical chaos	Encryption	LFW, ImageNet faces	≈8	n/r	n/r	dataset-specific*
Grosse et al. (2019)	Permutation + Diffusion + XOR (weights as keys)	Encryption	Standard + real-time	≈7.99	99.6% / 33.4%	∞ (lossless)	0.002 / ≈1

Notes: metric values are reproduced as reported by the cited primary sources (rounded); PSNR in dB; adjacent-pixel correlation ≈ 0 for all keyed schemes; SSIM shown as (encrypted vs plain / decrypted vs plain) where two values apply; n/r = not reported in the source; acc = classification accuracy; ROI = region of interest; IoMT = Internet of Medical Things; LFW = Labeled Faces in the Wild dataset; COCO, CelebA, STL-10, MNIST, CIFAR and ImageNet are standard public image datasets. *Bhat and Nanjundegowda (2024) and Alsafyani et al. (2023) report fidelity per dataset (the latter on a normalised per-dataset scale); no single summary value is given in the sources.

Table 3. Security and efficiency analyses disclosed by the 13 compared schemes. ✓ = reported; ~ = partially reported or claimed without a complete quantitative analysis; × = not reported; n.a. = not applicable to the task type.

Study	Entropy	NPCR/UACI	PSNR	SSIM	Key space/sens.	Randomness suite	Runtime
Alhakak et al. (2025)	✓	✓	✓	✓	~	×	×
Long et al. (2024)	✓	✓	✓	✓	~	×	×
Nair et al. (2024)	✓	✓	✓	✓	~	~	×
Ding et al. (2021)	✓	✓	✓	✓	~	×	×
Ding et al. (2022)	✓	✓	×	×	~	×	×
Sirichotedumrong and Kiya (2021)	×	×	×	✓	n.a.	×	×
Ioste et al. (2022)	×	×	×	×	n.a.	×	×
Panigrahy et al. (2024)	✓	✓	×	✓	~	×	~
Subramanian et al. (2021)	×	×	✓	✓	n.a.	×	×
Feng et al. (2023)	✓	×	✓	✓	~	×	×
Bhat and Nanjundegowda (2024)	✓	×	×	~	×	×	×
Alsafyani et al. (2023)	✓	×	×	~	~	×	×
Grosse et al. (2019)	✓	✓	✓	✓	~	×	~

Coverage reflects the metric disclosure consolidated in Table 2 and the analyses published in the primary sources. “Randomness suite” refers to complete statistical batteries such as NIST SP 800-22; “Runtime” requires hardware-qualified timings, not qualitative real-time claims.

The surveyed techniques are already aligned with concrete application domains, and the choice of technique family follows from each domain’s threat model (Sec-

tion 6).

Medical imaging and the IoMT. The largest application cluster in the corpus is medical: DeepEDN encrypts In-

ternet of Medical Things imagery in transit between acquisition devices and hospital systems (Ding et al., 2021), DeepKeyGen generates per-image stream-cipher keys for medical image exchange (Ding et al., 2022), and chaos-hybrid schemes target chest X-rays specifically (Alhakak et al., 2025; Nair et al., 2024). Keyed feature-based schemes fit this domain because diagnostic use demands faithful (ideally lossless) recovery after decryption. For privacy-preserving research collaboration *between* hospitals — where the goal is joint model training rather than image exchange — federated learning with secure aggregation (McMahan et al., 2017; Bonawitz et al., 2017) and combined weight learning (Ioste et al., 2022) are the natural complements.

Identity and biometric data. Face imagery calls for anonymisation as often as for encryption: GAN transformation preserves classification utility while concealing identity (Sirichotedumrong and Kiya, 2021), and CryptoGAN schemes with optical chaos target face datasets directly (Alsafyani et al., 2023). Here generative methods are preferred precisely because pixel-exact recovery is not the goal — unlinkability is.

Cloud storage and outsourced computation. Encrypted cloud storage benefits from schemes in which breached ciphertexts are unusable without the model or key — the weight-based family's lossless ciphers (Panigrahy et al., 2024; Grosse et al., 2019) suit resource-limited client devices, while homomorphic inference (Gilad-Bachrach et al., 2016) and encrypted-domain classification and retrieval (Al-Athori et al., 2026) cover the further step of allowing an untrusted cloud to *compute on and search* the encrypted data without ever decrypting it.

Covert transmission and data provision. Steganographic autoencoders (Subramanian et al., 2021) serve applications where the existence of the payload must itself be hidden. Finally, encrypted or anonymised corpora can be released for education and research without exposing the original content, with differential privacy (Abadi et al., 2016) bounding what any released model can reveal about individual training images.

9. CHALLENGES

Attack robustness. Learned ciphers must resist not only the classical differential attacks that NPCR/UACI quantify (Wu et al., 2011), but the model-specific vectors of Section 6: an adversary who obtains or approximates the network can mount inversion attacks, and CycleGAN-style translation is in principle re-learnable from example pairs. Because deep models are sensitive to small, targeted perturbations of inputs and weights (Grosse et al., 2019), adversarially corrupted ciphertexts can also disrupt decryption in ways classical ciphers do not exhibit.

Security versus efficiency. The strongest schemes in Table 2 rely on generator networks whose forward (and sometimes inverse) passes dominate encryption cost, while the devices that most need image encryption — IoMT sensors, embedded cameras — are the least able

to afford them. Weight-based ciphers (Panigrahy et al., 2024; Grosse et al., 2019) show that lossless, low-cost operation is achievable, but as Table 3 records, hardware-qualified runtime figures are almost never published, so the efficiency frontier of the field is effectively unmeasured.

Generalisability. Most schemes are tuned to one dataset or modality — chest X-rays (Alhakak et al., 2025; Nair et al., 2024), faces (Alsafyani et al., 2023) — and CycleGAN-based methods in particular are known to lose fine detail and transfer poorly across domains (Section 3). Whether a cipher trained on one corpus retains its security properties on another is rarely tested.

Interpretability and provable security. Many deep schemes behave as black boxes whose security is hard to prove mathematically relative to classical ciphers, an instance of the broader explainability problem in deep learning (Dhar et al., 2023). No surveyed scheme offers a reduction to a hard problem in the sense of modern cryptography; empirical metrics are currently the only evidence.

The reporting gap. Finally, as Table 3 makes explicit, key-space size and formal key-sensitivity analysis are disclosed only partially and chiefly by the chaos-based schemes; complete randomness-suite results (Rukhin et al., 2010) are absent from the corpus. This gap — rather than any single method — is currently the clearest obstacle to comparing these approaches on strictly cryptographic terms, and we recommend that future studies report key space, key sensitivity, a full SP 800-22 battery and hardware-qualified timings as standard.

10. FUTURE DIRECTIONS

Broader evaluation. Future work should evaluate models on larger, higher-resolution and more diverse datasets than the single-modality corpora that dominate Table 2, and should report the cross-dataset transfer experiments that generalisability claims require.

Lightweight deployment. The IoT and IoMT settings that motivate much of this literature (Ding et al., 2021) need lightweight, energy-efficient models; the lossless weight-based ciphers (Panigrahy et al., 2024; Grosse et al., 2019) are a promising base, but progress cannot be demonstrated until hardware-qualified timings become standard reporting (Section 9).

Hybrid frameworks. The best differential results in the corpus come from hybrids — CycleGAN with chaotic keys (Nair et al., 2024), deep features with AES (Long et al., 2024) — suggesting that deep learning is most effective when composed with, rather than substituted for, established cryptographic primitives, including chaos theory, steganography (Subramanian et al., 2021) and homomorphic encryption (Gentry, 2009; Gilad-Bachrach et al., 2016). Emerging generative directions point the same way: key-controlled inversion of public diffusion models (Zhang et al., 2026), still at the preprint stage, extends the generative-family template of Section 3 to diffusion architectures.

Explainability. Incorporating explainable-AI techniques (Dhar et al., 2023) would build trust in sensitive domains and is a first step toward security arguments that go beyond empirical metrics.

Post-quantum robustness. Quantum algorithms threaten the foundations these schemes implicitly rely on: Grover's search (Grover, 1996) halves the effective strength of symmetric key material — directly relevant to weight-derived keys, whose effective key space is already bounded by weight precision — and Shor's algorithm (Shor, 1997) breaks the public-key primitives used to distribute keys or models. Learned ciphers should therefore be sized and analysed against quantum-capable adversaries, and combining them with post-quantum key-encapsulation mechanisms (as the most recent classical-chaos work already does, Section 2) is a concrete near-term direction. The overarching aim remains a balance of security, efficiency, scalability and interpretability.

11. CONCLUSION

This review examined deep-learning image encryption and privacy preservation, organised consistently with Fig. 2 into feature-based, generative, and weight-based/integrated approaches. Feature-based encryption is preferred when confidentiality is paramount; weight-based encryption when the privacy of both data and model matters; and generative methods when anonymisation is the goal. CNN-and-chaos schemes excel on classical metrics, CycleGAN contributes anonymity at some fidelity cost, and weight-as-key methods offer practical, lossless solutions for embedded settings. These encryption families are, moreover, complementary to the privacy-preserving learning strategies of Section 4: homomorphic inference protects computation, federated learning with secure aggregation protects training, and differential privacy bounds what released models reveal — so layered architectures combining them are both feasible and, in medical settings, already emerging. Combining deep learning with feature- and weight-based strategies strengthens protection of sensitive data, pointing to a substantial role for these techniques in future cybersecurity — provided the open challenges of generalisability, interpretability, security-reporting completeness (Table 3) and quantum-era robustness are addressed.

DECLARATIONS

Competing interests. The author declares no competing financial or non-financial interests.

Funding. This review received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability. No new datasets were generated. All data discussed are drawn from the cited primary sources; the corpus and the citation-to-reference mapping are available from the author on reasonable request.

Ethics. This is a literature review of previously published work and did not involve human participants or animals.

REFERENCES

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 308–318.
- Al-Athori, S. A., Al-Tamimi, A.-G. T., and Alqobaty, A. (2026). Toward efficient privacy-preserving classification and retrieval of encrypted images using deep convolutional neural networks. *Multidisciplinary International Research Journal*, 1(1):15–25.
- Alhakak, Z., Salman, T., and Miry, A. (2025). Medical image encryption using chaotic algorithms and deep learning. *International Journal of Intelligent Engineering and Systems*, 18.
- Alsafyani, M., Alhomayani, F., Alsuwat, H., and Alsuwat, E. (2023). Face image encryption based on feature with optimization using secure crypto general adversarial neural network and optical chaotic map. *Sensors*, 23(3):1415.
- Bank, D., Koenigstein, N., and Giryas, R. (2023). Autoencoders. In *Machine Learning for Data Science Handbook*, pages 353–374. Springer.
- Bhat, R. and Nanjundegowda, R. (2024). Comparative analysis of CryptoGAN: Evaluating quality metrics and security in GAN-based image encryption. *Journal of Robotics and Control (JRC)*, 5(5):1557–1569.
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1175–1191.
- Datta, L. (2020). A survey on activation functions and their relation with Xavier and He normal initialization. *arXiv preprint arXiv:2004.06632*.
- Dhar, T., Dey, N., Borra, S., and Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society*, 4(1):68–75.
- Ding, Y., Tan, F., Qin, Z., Cao, M., Choo, K.-K. R., and Qin, Z. (2022). DeepKeyGen: A deep-learning-based stream-cipher generator for medical image encryption and decryption. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4915–4929.
- Ding, Y., Wu, G., Chen, D., Zhang, N., Gong, L., Cao, M., and Qin, Z. (2021). DeepEDN: A deep-learning-based image encryption and decryption network for Internet of Medical Things. *IEEE Internet of Things Journal*, 8(3):1504–1518.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006).

- Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference (TCC)*, pages 265–284.
- Feng, L., Du, J., Fu, C., and Song, W. (2023). Image encryption algorithm combining chaotic image encryption and convolutional neural network. *Electronics*, 12(16):3455.
- Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 169–178.
- Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., and Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, PMLR vol. 48, pages 201–210.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 2672–2680.
- Grosse, K., Trost, T. A., Mosbach, M., Backes, M., and Klakow, D. (2019). On the security relevance of weights in deep learning. *arXiv preprint arXiv:1902.03020*.
- Grover, L. K. (1996). A fast quantum mechanical algorithm for database search. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing (STOC)*, pages 212–219.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Higham, C. F. and Higham, D. J. (2019). Deep learning: An introduction for applied mathematicians. *SIAM Review*, 61(4):860–891.
- Ioste, A. R., Durham, A. M., and Finger, M. (2022). Combined learning of neural network weights for privacy in collaborative tasks. *arXiv preprint arXiv:2205.00361*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.
- Lin, N., Chen, X., Lu, H., and Li, X. (2021). Chaotic weights: A novel approach to protect intellectual property of deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(7):1327–1339.
- Long, B., Chen, Z., Liu, T., Wu, X., He, C., and Wang, L. (2024). A novel medical image encryption scheme based on deep-learning feature encoding and decoding. *IEEE Access*, 12:38382–38398.
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR vol. 54, pages 1273–1282.
- Mhawi, M. Y., Abdullah, H. N., and Sikora, A. (2024). GANs for image security applications: A literature review. *Iraqi Journal of Information and Communication Technology*, 7(2):89–102.
- Mienye, I. D. and Swart, T. G. (2025). Deep autoencoder neural networks: A comprehensive review and new perspectives. *Archives of Computational Methods in Engineering*.
- Nair, B. V., Muni, S. S., Durdu, A., et al. (2024). Deep learning and chaos: A combined approach to image encryption and decryption. *arXiv preprint arXiv:2406.16792*.
- Panigrahy, A. K. et al. (2024). A faster and robust artificial neural network based image encryption technique with improved SSIM. *IEEE Access*, 12:10818–10833.
- Panwar, K., Kukreja, S., Singh, A., and Singh, K. K. (2023). Towards deep learning for efficient image encryption. *Procedia Computer Science*, 218:644–650.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.
- Rukhin, A. et al. (2010). A statistical test suite for random and pseudorandom number generators for cryptographic applications. Technical Report NIST Special Publication 800-22 Rev. 1a, National Institute of Standards and Technology.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.
- Shokri, R. and Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1310–1321.
- Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sirichotedumrong, W. and Kiya, H. (2021). A GAN-based image transformation scheme for privacy-preserving deep neural networks. In *Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, pages 745–749.
- Subramanian, N., Cheheb, I., Elharrouss, O., Al-Maadeed, S., and Bouridane, A. (2021). End-to-end image steganography using deep convolutional autoencoders. *IEEE Access*, 9:135585–135593.
- Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). Inception-v4, Inception-ResNet and the impact of resid-

- ual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.
- Wu, Y., Noonan, J. P., and Agaian, S. (2011). NPCR and UACI randomness tests for image encryption. *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT)*, 1(2):31–38.
- Zhang, L., Li, W., Liu, B., and Xu, Z. (2026). Public diffusion models, private images: Key-controlled inversion for conditional reconstruction. *arXiv preprint arXiv:2606.22988*.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232.
- Zhu, M., Gong, S., Qian, Z., and Zhang, L. (2019). A brief review on cycle generative adversarial networks. In *Proceedings of the 7th IIAE International Conference on Intelligent Systems and Image Processing (ICISIP)*, pages 235–242.