LSB-based Fragile Watermarking for Image Integrity Verification Under Morphing Attacks

Davide Ghiani[1], Sergio Soggiu[1], Jefferson David Rodriguez Chivata[1], Simone Maurizio La Cava[1], Marco Micheletto[1], Giulia Orrù[1,*], Federico Lama[2] and Gian Luca Marcialis[1]

[1]Department of Electrical and Electronic Engineering (DIEE), University of Cagliari, Cagliari, Italy
[2]Dedem S.p.A., Via Cancelleria 59 - 00072 Ariccia (Italy)

Abstract

Ensuring the integrity of biometric images is becoming increasingly critical as modern generative models enable sophisticated manipulations that can bypass identity verification systems. Recent studies have investigated watermarking through deep steganographic embedding as a proactive defense, but such approaches rely on complex architectures and provide limited interpretability. In this work, we revisit the classical Least Significant Bit (LSB) technique and assess its suitability as a deliberately fragile and inherently explainable steganographic embedding method for integrity verification in ICAO-compliant facial images. We conduct an extensive evaluation of LSB fragility under a wide range of manipulations, including compression, resizing, noise injection, blurring, sharpening, and two categories of morphing attacks, using both traditional landmark-based and diffusion-based generation approaches. Our results highlight the distinctive degradation patterns induced by different alterations, supporting LSB as a lightweight and transparent mechanism for detecting tampering in biometric imagery.

Keywords

watermarking, image certification, morphing

1. Introduction

Biometric identity verification increasingly relies on facial images embedded in electronic documents and remote authentication systems. Ensuring the integrity of these images is therefore essential: any modification introduced after acquisition, whether accidental or malicious, can compromise the reliability of automated facial recognition pipelines [1, 2, 3]. In this context, proactive integrity mechanisms capable of detecting post-issuance alterations are gaining significant attention [4]. A growing research direction explores the use of steganographic embedding as a form of fragile watermarking [5]. The underlying principle is straightforward: an integrity marker is hidden into the image at enrollment time, and any subsequent manipulation corrupts the embedded payload. Therefore, the resulting degradation provides a direct signal of tampering. Most existing approaches rely on deep neural networks to achieve high-capacity embedding and recovery. However, such models introduce substantial complexity, offer limited interpretability, and are typically optimized for robustness, an undesirable property when the goal is to detect even the slightest modification.

In contrast, this work investigates a simple, deliberately fragile, and fully explainable alternative: the Least Significant Bit (LSB) method [6] for biometric image certification. Although LSB is among the earliest and simplest steganographic techniques, its inherent sensitivity to minimal pixel perturbations makes it particularly suitable for integrity verification. Moreover, its deterministic behavior enables transparent analysis of how specific manipulations affect the hidden marker, providing clear forensic cues in the recovered signal.

Beyond commonly studied transformations, namely compression, resizing, noise, blurring, and sharpening, this study evaluates the behavior of LSB under two categories of morphing attacks. The

Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09-13, 2026, Cagliari, IT
*Corresponding author.
davide.ghiani@unica.it (D. Ghiani); s.soggiu1@studenti.unica.it (S. Soggiu); jeffersond.rodriguez@unica.it (J. D. R. Chivata); simonem.lac@unica.it (S. M. L. Cava); marco.micheletto@unica.it (M. Micheletto); giulia.orru@unica.it (G. Orrù); federico.lama@dedem.it (F. Lama); marcialis@unica.it (G. L. Marcialis)

first consists of classical landmark-based morphing, a well-established technique that blends facial identities through geometric alignment, preserving the biometric imprint of the contributing individuals [7]. The second involves diffusion-based morphing, a modern generative approach capable of producing high-fidelity synthetic blends that preserve identity traits more convincingly [8]. Specifically, we introduced these malicious attacks due to their capability to generate realistic images that can be falsely considered to belong to different individuals, hence posing a new challenge for biometric security. Assessing LSB's fragility across both morphing families provides insight into its ability to detect manipulations ranging from traditional methods to state-of-the-art generative techniques.

Accordingly, we present a systematic study of LSB-based fragile watermarking for ICAO-compliant facial images, which constitute the international standard for passports and electronic identity documents and therefore represent the most relevant and security-critical setting for integrity verification [9, 10]. We analyze degradation patterns induced by various manipulations, quantify their impact using standard image-quality metrics, and evaluate the ability of a classifier to infer the type of alteration solely from the recovered marker. To our knowledge, no prior work has conducted such an extensive analysis of LSB-based fragile watermarking within the ICAO compliance framework. To summarize, the main contributions of this work are the following: (i) A comprehensive evaluation of LSB fragility across a wide range of image manipulations, from classical distortions to advanced diffusion-based morphing; and (ii) a classification framework that exploits the extracted LSB marker to identify the applied manipulation.

The rest of this paper is organized as follows. Section 2 reviews the current literature on steganography and watermarking for digital images. Section 3 describes the proposed framework. Section 4 reports the experimental protocol employed to conduct our evaluation, while Section 5 presents the obtained results. Finally, conclusions are drawn in Section 6.

2. Related Work

Steganography and digital watermarking are two closely related data hiding paradigms that differ primarily in their tolerance to image modifications [11]. Fragile methods react strongly to even minimal perturbations, making them suitable for integrity verification, whereas robust approaches are designed to withstand post-processing operations and attacks, including modern synthetic content generation such as deepfakes [12]. In the context of ICAO-compliant facial images, fragile techniques are preferable, as their universal sensitivity supports the detection of any unauthorized post-acquisition alteration. This is consistent with the proactive security requirements enforced in biometric identity documents, where even subtle inconsistencies must be identifiable.

A second distinction involves how information is embedded and interpreted. Classical digital watermarking techniques typically encode compact signatures or bitstrings, often in predefined spatial or frequency-domain locations. These approaches have proven effective for copyright protection [13] and tamper localization [14], but they provide limited interpretability when the watermark is degraded or corrupted. On the other hand, modern steganographic techniques enable the embedding of richer content, such as full images, into the host structure. Initially intended for covert communication, these methods have recently garnered interest in forensic applications: any manipulation applied to the host image inevitably corrupts the hidden payload, and the degradation visible in the recovered content can serve as an actionable integrity indicator. Advances in deep learning have been a driving force behind this trend. Architectures based on convolutional neural networks (CNNs) [15], generative adversarial networks (GANs) [16], and transformer-based autoencoders [17] have significantly improved both capacity and perceptual invisibility, making them compelling candidates for fragile watermarking.

An example of this perspective is the work by Ghiani et al. [4], which explored deep steganographic embedding for ICAO-compliant image certification. Their study demonstrated that deep models can generate distinctive degradation patterns in the recovered marker when the host image is subjected to various manipulations, including compression, noise, blur, sharpening, resizing, or morphing. The recovered marker then becomes a forensic signal useful for both tamper detection and manipulation

classification. This line of work aligns with a growing research direction that investigates steganographic embedding as a form of fragile watermarking [5]. However, most existing approaches achieve this through deep neural networks, thus inheriting their high capacity but also their complexity and limited interpretability, characteristics that are not always desirable in application scenarios involving identity verification from certified biometric data [18, 19, 1].

Building on these considerations, our work investigates a complementary viewpoint. Rather than employing complex deep architectures, we investigate the potential of the classical Least Significant Bit (LSB) method as an intentionally fragile and fully explainable embedding strategy. Despite its simplicity, LSB offers deterministic behaviour and extreme sensitivity to perturbations, enabling transparent analysis of manipulation effects and providing a baseline for evaluating fragile watermarking mechanisms under standard and generative morphing attacks.

3. Proposed Framework

As discussed in the previous sections, we adopt the LSB strategy as our steganographic embedding mechanism, chosen for its simplicity, interpretability, and inherent fragility to manipulation. To enhance concealment, the method is further combined with a scattering procedure, which distributes the embedded payload across the image, improving imperceptibility and reducing localized artefacts.

In an 8-bit grayscale image, each pixel assumes an intensity ranging from 0 to 255. The least significant bit (i.e., the rightmost bit in the binary representation) contributes only a value of 1 to the pixel intensity: modifying this bit, therefore, produces a minimal change that is typically imperceptible to the human eye. Extending this logic, modifying the two least significant bits alters pixel values by at most 3, which remains visually indistinguishable in most cases. This property enables the LSB domain to be exploited for information embedding: the more LSBs are used, the higher the embedding capacity. However, increasing the number of manipulated bits also leads to a proportional rise in distortion, which may eventually compromise the invisibility requirement and break the imperceptibility principle that steganography relies upon.

In standard LSB embedding, the least significant bits of one or more color channels (e.g., R, G, B) are directly replaced with the same number of bits of the secret payload. The insertion typically proceeds in a sequential, pixel-by-pixel manner. Although straightforward, this deterministic process may concentrate bit modifications within contiguous spatial areas, making them more susceptible to localized detection.

In contrast, scattering randomizes the embedding positions by distributing the payload across pseudo-random pixel coordinates derived from a secret key. This dispersive strategy mitigates clustered alterations and produces a modification pattern that more closely resembles natural image noise. As a result, perceptual quality measures such as PSNR and SSIM generally show higher values [20], although the magnitude of improvement decreases as the number of substituted LSBs grows.

It is important to emphasize that no robustness mechanisms are applied to protect the hidden data. Consequently, any manipulation applied to the stego image, whether intentional or incidental, directly affects the extracted message. The core hypothesis of this work is that such degradations leave distinctive, manipulation-dependent signatures in the recovered payload, enabling us to analyze and classify the type of transformation performed on the image. To verify the validity of this hypothesis, we created a three-stage methodology:

1. A steganographic process E based on LSB is applied to the ICAO-compliant facial image (cover image) I_C to embed a secret marker image I_S within it, producing a certified watermarked image I_{stego}:
$$I_{stego} = E(I_C, I_S) \tag{1}$$
The hidden image I_S, imperceptibly hidden within the cover image, acts as a fragile integrity marker, ensuring that any future modification to the stego-image affects the embedded content.
2. A set of controlled transformations T are applied to I_{stego} to simulate real-world manipulations:
$$I_t = T(I_{stego}) \tag{2}$$

The considered manipulations, namely resizing, compression, noise, blur, sharpening, and morphing, reflect both common post-processing operations and intentional biometric attacks. These transformations simulate real-world conditions in which an image might be altered after issuance, such as during storage, allowing us to assess whether the hidden integrity marker can serve as a forensic signal.

3. A decoder D is used to extract the revealed image I_r from the potentially modified image I_t:

$$I_r = D(I_t) \tag{3}$$

If the stego image is not subjected to any transformations (i.e., no transformation-induced artefacts are introduced), the recovered secret image preserves its structural integrity. The fidelity of the extracted secret depends on the number of least significant bits n used for embedding: higher values of n yield reconstructions that more closely resemble the original. However, when the stego image undergoes manipulations, characteristic artefacts emerge in the revealed secret, reflecting both the type and the magnitude of the applied transformation. To analyze these introduced distortions, we propose a classification model that identifies the specific manipulation performed by examining the artefactual patterns present in the recovered secret.
The operation of the proposed classification system is strictly dependent on the presence of the secret at the time of verification. In this context, if the secret is missing or irretrievable, then the classification model will provide unreliable output.

4. Experimental protocol

4.1. Dataset

Our goal is to assess the feasibility of our fragile watermarking approach applied to facial images that comply with ICAO guidelines. For this purpose, we employed the Chicago Face Database (CFD) [21], which provides high-quality images of 827 individuals, including both men and women, from diverse ethnic groups and aged between 17 and 65. For each subject, the database provides a single image compliant with ICAO standards [9, 10] with a resolution of 2444×1718 pixels.

To prepare the data for the embedding phase, each image was cropped to a square format of 1718×1718 pixels, removing peripheral areas of the background but keeping the facial region intact. Subsequently, the cropped images were resized to 224×224 pixels to allow comparison with the results obtained in Ghiani et al. [4]. The ICAO logo (224×224) was selected as the integrity marker to be embedded in the facial image during the certification phase.

4.2. Watermarking method

As explained in Section 2, LSB-based watermarking produces an encoded image containing a fragile watermark that cannot withstand even minimal alterations applied to the cover. This fragility is an essential property in integrity verification scenarios: the more sensitive the embedded watermark is to pixel-level modifications, the more easily any attempt at alteration or counterfeiting can be detected. However, information in an image is not uniformly distributed, and a naïve pixel-to-pixel embedding often concentrates the hidden content in a predictable region.

To formalize the embedding process, we consider two 8-bit images: the cover image C and the secret image S, both defined on spatial coordinates (i, j) and color channels $k \in \{R, G, B\}$. For any pixel intensity $x \in \{0, \dots, 255\}$, we introduce the following operators:

$$\mathrm{LSB}_b(x) = x \bmod 2^b \tag{4}$$

$$\mathrm{MSB}_b(x) = \left\lfloor \frac{x}{2^{8-b}} \right\rfloor \tag{5}$$

which extract, respectively, the lowest and highest b bits of x, with $b \in \{1, \dots, 7\}$ being the number of bits to be embedded in each cover pixel.

LSB classical embedding: In the classical formulation of b-LSB substitution, the stego-pixel C'_{ijk} is obtained by removing the b least significant bits of C_{ijk} and inserting in their place the b most significant bits of the corresponding secret pixel S_{ijk}. Formally,

$$C'_{ijk} = C_{ijk} - \mathrm{LSB}_b(C_{ijk}) + \mathrm{MSB}_b(S_{ijk}) \tag{6}$$

The subtraction term clears the b LSBs of the cover pixel, while the final term inserts the b MSBs of the secret pixel, preserving all higher-order bits of the cover.

LSB embedding with scattering: To avoid a spatially predictable embedding pattern and to distribute the hidden information across the image, we adopt a scattering mechanism. Let the secret image S have spatial dimensions w x h, its pixels are first linearized into indices:

$$p \in \{0, \dots, S_p - 1\} \qquad S_p = wh$$

, and a pseudorandom permutation of the cover pixel indices is generated using a fixed seed. Let q_p denote the p-th element of this permutation, i.e., $q_p = \pi(p)$, where π is a bijection over $\{0, \dots, WH - 1\}$. The p-th secret pixel is then embedded in the cover pixel indexed by q_p. The embedding rule becomes the following.

$$C'_{q_p,k} = C_{q_p,k} - \mathrm{LSB}_b(C_{q_p,k}) + \mathrm{MSB}_b(S_{p,k}), \tag{7}$$

This equation is identical to Eq. (6), except that the location of the cover is determined by permutation rather than by spatial alignment of C and S. This ensures that the embedded information is spatially dispersed and not confined to a contiguous region of the image. The use of a permutation π introduces a key-like behavior: decoding requires knowledge of the same seed to reproduce the same pixel ordering, and any alteration in the stego image affects the recovered secret in a non-local and easily detectable manner. The combination of LSB substitution and scattering, therefore, yields a highly fragile watermarking scheme well suited for integrity verification tasks.

In our experiments, we considered three configurations of the parameter b, namely **1**, **3**, and **5** bits. As discussed earlier, increasing b introduces the classical trade-off inherent in LSB-based methods: on the one hand, a larger number of substituted LSBs allows more information from the secret image to be embedded, therefore improving the fidelity of the recovered watermark; on the other hand, it reduces the imperceptibility of the embedding, since a greater portion of the cover pixel values is modified.

4.3. Image Manipulations

Identity-Preserving (Benign) Manipulations: To assess the robustness of the proposed integrity-verification framework, we subject the certified stego-images to a series of controlled manipulations that emulate realistic post-acquisition modifications. These perturbations encompass both unintentional degradations and routine post-processing operations, enabling us to assess their impact on the embedded integrity marker.

Compression: Digital images frequently undergo re-encoding during storage, transmission, or automated verification. We apply JPEG and WebP compression with a quality factor $QF \in [80, 100]$, where lower values introduce noticeable artifacts and higher values preserve most visual content.

Resizing: ICAO-compliant images are often rescaled for different document formats or submission pipelines. Each image is downsampled according to a resizing factor $RF \in [50\%, 99.9\%]$ and subsequently restored to its original dimensions. Strong downscaling (low RF) results in significant detail loss, while factors close to 100% preserve most structural information.

Noise Addition: Noise may arise from low-bitrate encoding, repeated compression cycles, or scanning processes. We simulate:

- *Gaussian noise*, with standard deviation $\sigma_G \in [2, 32]$;

- *Salt-and-pepper noise*, parameterized by $P_{SP} = (P_{\text{Salt}}, P_{\text{Pepper}})$, defining the probability that pixels are replaced by extreme intensity values.

Blurring: Smoothing operators reduce noise but may disrupt the steganographic structure. We apply:

- *Gaussian blur*, with kernel sizes $K_G \in \{3, 5, 7, 9\}$;
- *Median blur*, with kernel sizes $K_M \in \{3, 5, 7, 9\}$.

Sharpening: To assess the effect of enhanced edge contrast, we adjust the sharpening intensity within the range $S_F \in [0, 1]$, where stronger values increase high-frequency components that may affect the embedded marker.

Identity-Altering Manipulation (Morphing): In contrast with previous transformations, which do not modify the underlying identity, morphing represents a deliberate biometric attack aimed at blending facial characteristics from multiple subjects. We consider two morphing techniques to assess the vulnerability of the integrity marker under identity-altering manipulations.

FaceMorpher: We employ *FaceMorpher*[1], an open-source landmark-based morphing tool, to blend the stego-identity with a secondary individual. A blending coefficient of $\alpha_M = 0.9$ biases the morph towards the primary identity while incorporating subtle geometric and textural elements from the secondary contributor, preserving plausibility while maximizing the likelihood of successful verification.

DiffMorpher: We further adopt *DiffMorpher*[2] [22], a recent diffusion-based morphing method that enables smooth and semantically coherent transitions between two images by leveraging the prior knowledge of a pre-trained text-to-image diffusion model. Instead of relying on geometric alignment, DiffMorpher fits a separate Low-Rank Adaptation (LoRA) module to each input image and performs morphing by interpolating both the LoRA parameters and the corresponding latent noise vectors. This dual interpolation captures high-level semantic identity while ensuring gradual variation in spatial details. The resulting morphs exhibit high perceptual realism and smooth identity transitions, offering a substantially more advanced identity-altering manipulation compared to traditional landmark-based approaches.

4.4. Image quality assessment

To assess image quality after watermarking and study the impact of manipulations on marker integrity, we employ three full-reference image quality metrics: Peak Signal-to-Noise Ratio (*PSNR*), Structural Similarity Index Measure (*SSIM*), and Mean Squared Error (*MSE*).

PSNR quantifies the ratio between the power of the signal and the power of the introduced noise, thus measuring the fidelity between two compared images. It is computed as:

$$\text{PSNR}(x, y) = 10 \cdot \log_{10} \left(\frac{\text{MAX}_x^2}{\text{MSE}(x, y)} \right), \tag{8}$$

where x and y denote the original and reconstructed images, respectively, both of size $m \times n$; MAX_x is the maximum possible pixel value (255 for 8-bit grayscale images). The MSE is defined as:

$$\text{MSE}(x, y) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [x(i, j) - y(i, j)]^2, \tag{9}$$

[1] https://github.com/alyssaq/face_morpher
[2] https://github.com/Kevin-thu/DiffMorpher

and measures the average squared difference between corresponding pixels of the two images. Higher PSNR values correspond to greater similarity, whereas increasing MSE values indicate stronger distortion. Typically, a PSNR above 40 dB corresponds to negligible degradation [23, 24].

In addition to pixel-wise error metrics, SSIM provides a perceptual similarity measure by comparing local patterns of luminance, contrast, and structure. It is defined as:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{10}$$

where μ_x and μ_y are the local means, σ_x^2 and σ_y^2 the variances, and σ_{xy} the covariance between the two images. The constants C_1 and C_2 are used to stabilize the expression in the presence of weak denominators. An SSIM value close to 1 indicates a high degree of structural similarity.

Overall, these metrics provide complementary insights into the degradation experienced by the hidden marker following image manipulations, and enable us to assess the extent to which an image has been altered.

4.5. Classification Protocol

To train a model capable of identifying the type of manipulation applied to an image based on the revealed secret, we employed ResNet-50 [25], pre-trained on ImageNet [26], as the backbone for feature extraction. We first addressed the classification problem without considering morphing, focusing on six manipulation types: compression, resizing, blurring, Gaussian noise, salt-and-pepper noise, and sharpening. In a second step, we expanded the task to include two additional classes, FaceMorpher and DiffMorpher, corresponding to the morphing methods described above.

To perform classification from the extracted embeddings, we appended a sequence of fully connected layers. First, a linear layer reduces the feature dimensionality from 2048 to 512 units. A ReLU activation function is then applied to introduce non-linearity, followed by a dropout layer with a probability of 0.5. Finally, a second linear layer maps the resulting vector to the space of the target manipulation classes. The model was trained using the Adam optimiser with a learning rate of 1×10^{-3} and the cross-entropy loss function. Each image was resized to a resolution of 224×224 pixels before being fed to the network. Training was performed for 20 epochs using a mini-batch size of 32. To assess the reliability of the proposed classification model, 70% of the user identities in the dataset (578 identities) were employed for fine-tuning, while the remaining 30% (249 identities) were used for testing. This ensured balanced manipulation classes across training and test sets and simulated a realistic application scenario in which the facial identities present during training do not overlap with those encountered at inference time.

The performance of the model was evaluated using standard pattern-recognition metrics: accuracy, precision, recall, and F1-score. To investigate the model's ability to generalize under previously unseen manipulation conditions and different embedding payloads, we defined four evaluation protocols.

- **P8-8**: training and testing are performed on the same eight variations of each manipulation type (e.g., different levels of noise or compression).
- **P6-8**: training is conducted on six variations per manipulation type, whereas testing includes all eight, thus introducing two previously unseen variations for each manipulation during evaluation.

Furthermore, because the model was trained independently on images recovered after manipulations from steganographic embeddings with different LSB configurations (1, 3, and 5), classification could be evaluated under two operational scenarios:

- **Intra-stega scenario**: both training and test images are embedded using the same steganographic configuration and are subjected to identical manipulation types and strengths.
- **Cross-stega scenario**: training and test images are embedded using different steganographic configurations, while the manipulation types remain consistent.

Table 1
Differences between the original image and the certified image (certification task), and between the original secret and the recovered secret (recovery task), reported in terms of SSIM, MSE, and PSNR for different numbers of embedded bits. Values are expressed as mean ± standard deviation.

Task	# Bit	SSIM	MSE	PSNR
Certifyng	1 Bit	0.9977 ± 0.0003	0.37 ± 0.03	52.46 ± 0.38
	3 Bit	0.9447 ± 0.0049	10.91 ± 0.78	37.76 ± 0.29
	5 Bit	0.5752 ± 0.0216	194.77 ± 12.85	25.24 ± 0.27
Recovery	1 Bit	0.5794 ± 0.0000	10986.43 ± 0.00	7.72 ± 0.00
	3 Bit	0.9280 ± 0.0000	593.06 ± 0.00	20.40 ± 0.00
	5 Bit	0.9800 ± 0.0000	32.30 ± 0.00	33.04 ± 0.00

Although a practical certification deployment would typically rely on a fixed embedding configuration throughout the system's life cycle, the cross-stega analysis provides valuable insights into the generalizability and robustness of the proposed manipulation detection framework. In particular, it simulates realistic inconsistencies, such as re-certification using a different embedding scheme. Additionally, it sheds light on the transferability of learned features across embedding strategies, an essential aspect for scalable and future-proof implementations.

5. Results

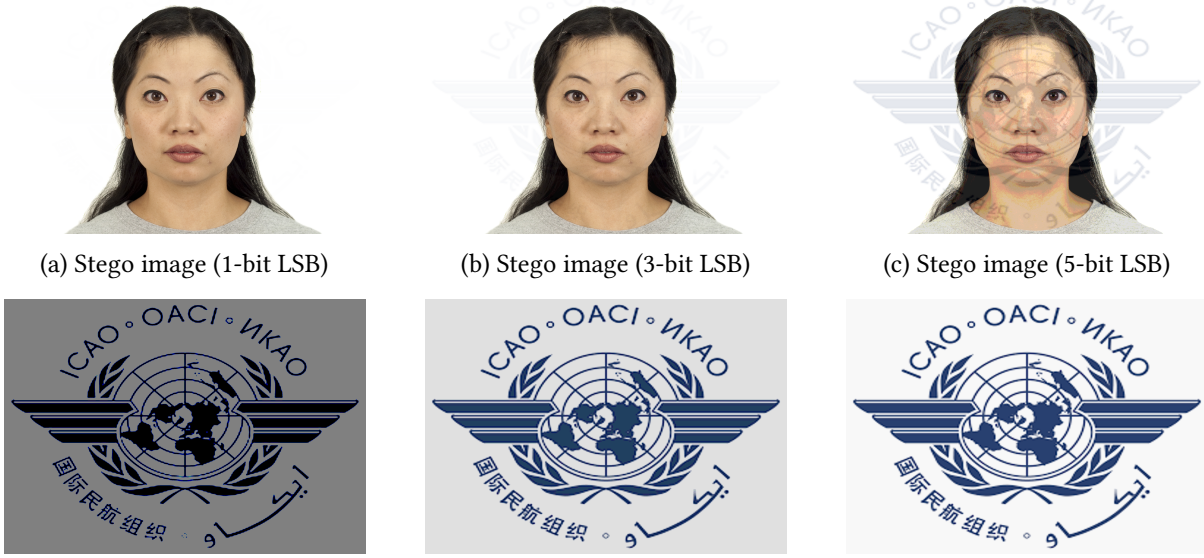
In this Section, we present the results obtained from the previously described experimental protocol. In Section 5.1, we reported the results obtained by analyzing the impact of the modifications in the images employed on the integrity verification process. In Section 5.2, we discussed the capabilities of the classification model in distinguishing identity-preserving manipulations. Finally, in Section 5.3 we conducted a specific analysis regarding morphing manipulations.

5.1. Image quality and manipulation assessment results

5.1.1. Image quality

First, to evaluate the feasibility of using secret image hiding and recovery for manipulation detection, we approach the quality of the images after applying the LSB algorithm at different bit levels. Table 1 presents the average quality of the certified images and the revealed images, as quantified by SSIM, MSE, and PSNR for each bit variation using the LSB method. These results differ significantly from deep steganographic techniques, such as those presented in [4]. Such models, named Steguz and StegFormer, in fact, showed almost negligible degradation in both the certification and recovery phases. Specifically, for the certification task, Stegformer achieved SSIM = 0.9696, MSE = 7.23, and PSNR = 39.56 dB, while Steguz settled at slightly lower values (SSIM = 0.9266, MSE = 123.27, PSNR = 27.37 dB). Performance also remained high in the recovery task, with SSIM consistently above 0.93 and PSNR between 28 dB and 32 dB.

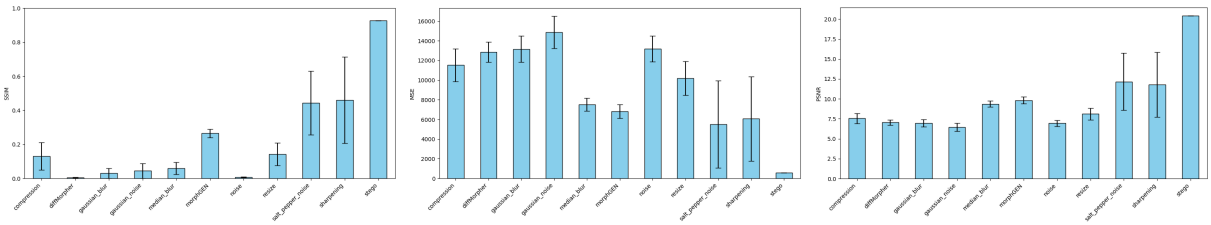
The behavior of the LSB method is radically different. With 1 bit of modification, the certified image maintains extremely high quality (SSIM $\approx$ 0.997, PSNR $\approx$ 52 dB). However, the quality of the recovered secret is extremely low (SSIM $\approx$ 0.58, PSNR $\approx$ 7.8 dB), indicating that a 1-bit modification has a very low capacity for hiding an RGB image as a secret. If bit capacity is increased to 3 and 5 bits, an opposite trade-off is observed: the quality of the stego-image progressively decreases, up to 0.94, 0.57 for SSIM and 37 dB, 25 dB for PSNR respectively, while the quality of the recovered secret improves but remains inferior to those reported by deep models (approximately 0.93, 0.98 for SSIM and 20 dB, 33 dB for PSNR respectively). The visual differences for these configurations are shown in Figure 1, where it is clearly evident that the 3-bit setting provides a more balanced trade-off between the invisibility of the watermark and the usability of the recovered image, and is therefore adopted as the reference configuration for the manipulation impact assessment and the morph impact analysis.



(a) Stego image (1-bit LSB) (b) Stego image (3-bit LSB) (c) Stego image (5-bit LSB)

(d) Recovered secret (1-bit LSB) (e) Recovered secret (3-bit LSB) (f) Recovered secret (5-bit LSB)

Figure 1: Visual comparison of the certification (top row) and recovery (bottom row) stages for different LSB depths.



(a) SSIM (b) MSE (c) PSNR

Figure 2: Comparison between the original secret image and the watermark recovered after applying different manipulations to the cover image. The metrics (MSE, PSNR, SSIM) quantify how each manipulation corrupts the hidden image.

5.1.2. Image manipulation assessment

Having the fixed operational point of **3-LSB**, we next investigate how different image manipulations impact the integrity of the embedded watermark. While the previous analysis focused on the distortion introduced by the embedding and recovery processes alone, this analysis explicitly considers the effect of applying a benign manipulation to the watermarked cover image before the recovery step. This allows us to quantify the degree to which the hidden secret degrades under a variety of distortions that are representative of both benign post-processing operations and potential attack strategies.

As described in the experimental protocol, each manipulation was applied over a range of values for its key parameters (for example, the standard deviation in the case of Gaussian noise), and the quality metrics were computed on all recovered images. Figure 2 reports how these manipulations applied to the container image affect the quality of the recovered secret image. For each manipulation type, the MSE, PSNR and SSIM are computed between the original watermark and the recovered one, thus quantifying how severely the LSB embedding is corrupted by the alteration of the container image.

Overall, the results confirm the fragile nature of the 3-LSB watermark: most manipulations introduce a significant degradation of the recovered secret. Noise-based perturbations (Gaussian noise, Gaussian blur, Noise, Compression) produce the highest MSE and lowest PSNR/SSIM, indicating that local pixel-level disturbances are particularly harmful for LSB substitution, which relies on the integrity of the low-order bits. Resize also introduces noticeable degradation, likely due to resampling artifacts that

propagate across the embedded payload.

Other manipulations produce more moderate effects. Median blur, Sharpening, and Salt–and–Pepper noise yield lower distortion on average, and in particular, the last two, preserve a large part of the structural information of the watermark, as reflected by higher SSIM values. This suggests that these distortions, which preserve global luminance patterns while altering local variability, are less destructive to the embedded watermark.

In Figure 3, it is possible to observe the different recovery patterns that are evident when manipulations are applied to the container. It is clear to see that sharpening (Figure 3-g) presents the cleanest recovery pattern, while blurring (Figure 3-b and 3-d) seems to be the manipulation that most affects the secret image recovery. On the other hand, it is interesting to note that each manipulation, even variations within the same manipulation type, such as blurring or noise, generates different recovery patterns. In forensic analysis, it is essential not only to identify that a manipulation has been applied, but also to determine which manipulation was used, in order to create more appropriate and tailored response protocols. Therefore, we have created a manipulation classifier (post-recovery) that, based on the recovered embedding, predicts which manipulation was used to attack the container. The results are presented in the following sections.

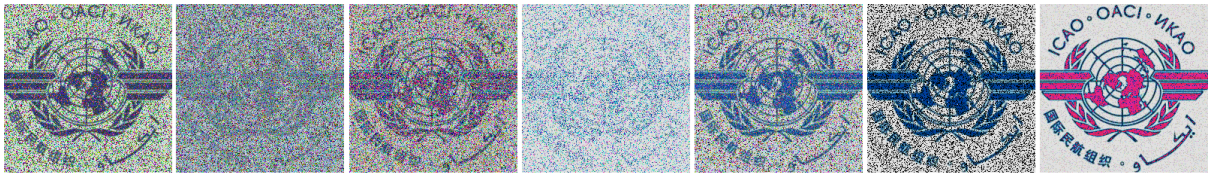


Figure 3: Image recovery patterns under benign manipulations for the 3-LSB method: a) JPEG compression ($Q_F = 80$), b) Gaussian Blur ($K_G = 7$), c) Gaussian Noise ($\sigma = 8$), d) Median Blur ($K_M = 7$), e) Resize ($R_F = 85\%$), f) Salt Pepper Noise ($P_{SP} = (0.3, 0.01)$) and g) Sharpening ($S_F = 0.5$).

5.2. Identity-Preserving manipulation classification results

In this Section, we evaluate the classifier's ability to discriminate between distinct identity-preserving manipulations. Our analysis focuses on two key dimensions: the robustness of the model against unseen manipulation parameters (comparing protocols P8-8 and P6-8) and the impact of mismatched steganographic embedding depths (comparing intra-stega and cross-stega scenarios).

Robustness within consistent embedding depths (intra-stega): When the classifier is trained and tested on images generated with the same LSB depth, the proposed framework demonstrates remarkable efficacy. As detailed in Tables 2 and 3, the accuracy remains consistently high across all configurations. Specifically, for the 3-LSB configuration, which we identified as the optimal trade-off in Section 5.1, the model achieves an accuracy of 97.60% in the P8-8 protocol, increasing to 99.77% for the 5-LSB configuration.

Crucially, this performance does not degrade significantly when identifying manipulations with previously unseen types (P6-8 protocol). The 3-LSB accuracy shifts only slightly to 94.98%, while the 5-LSB configuration maintains 94.93%. This stability suggests that the model successfully learns the fundamental artefactual features of each manipulation type (e.g., the specific blurring patterns or noise distributions in the recovered image) rather than overfitting to specific intensity values encountered during training. This behavior is visually corroborated by the confusion matrices in Figure 4, which show clear diagonal dominance. However, closer inspection reveals specific misclassification patterns rooted in the physical nature of the manipulations. In the P8-8 protocol, the Resize class exhibits a 9% confusion rate with the Blur category (aggregating Gaussian and Median blur). This is likely attributable to the interpolation algorithms used during resizing, which introduce smoothing effects analogous to those of a low-pass filter. A similar reciprocal confusion appears in the P6-8 protocol, where Blur is misclassified as Resize in 9.3% of cases. Additionally, the P6-8 protocol reveals ambiguities between

Table 2
Classification performance on identity preserving manipulations in (P8-8 sub-protocol).

Training set	Test Set	Accuracy	Precision	Recall	F1 Score
1-LSB	1-LSB	90.86%	91.77%	90.86%	90.32%
	3-LSB	20.47%	6.07%	20.47%	9.15%
	5-LSB	16.67%	2.81%	16.67%	4.80%
3-LSB	1-LSB	32.20%	29.18%	32.20%	24.04%
	3-LSB	97.60%	97.72%	97.60%	97.58%
	5-LSB	81.82%	85.59%	81.82%	81.95%
5-LSB	1-LSB	19.09%	21.00%	19.09%	8.43%
	3-LSB	80.80%	87.24%	80.80%	79.93%
	5-LSB	99.77%	99.77%	99.77%	99.77%

Table 3
Classification performance on identity preserving manipulations in (P6-8 sub-protocol).

Training set	Test Set	Accuracy	Precision	Recall	F1 Score
1-LSB	1-LSB	90.43%	90.33%	90.43%	90.12%
	3-LSB	25.00%	8.40%	25.00%	12.47%
	5-LSB	25.00%	11.22%	25.00%	13.70%
3-LSB	1-LSB	26.29%	28.08%	26.29%	19.07%
	3-LSB	94.98%	95.46%	94.98%	94.98%
	5-LSB	81.28%	86.30%	81.28%	81.43%
5-LSB	1-LSB	18.75%	7.92%	18.75%	8.00%
	3-LSB	74.74%	83.61%	74.74%	74.55%
	5-LSB	94.93%	95.37%	94.93%	94.79%

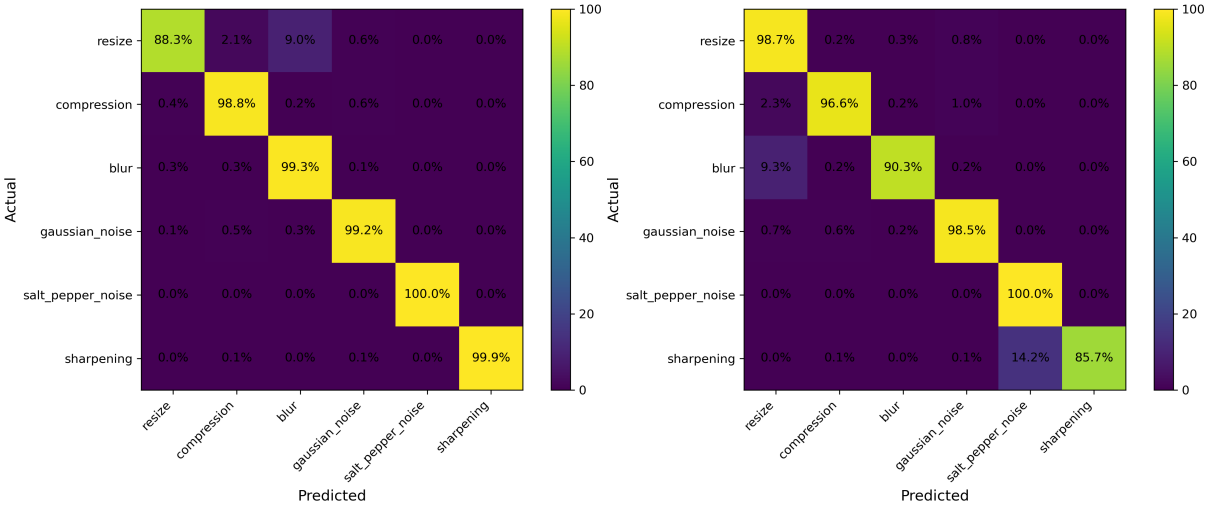
Sharpening and Salt-and-Pepper noise (14.2% of error). This aligns with our earlier quality assessment, where these two manipulations were found to induce the least structural degradation on the recovered secret, resulting in subtle and less distinct artefactual signatures.

Generalization across embedding depths (cross-stega): Conversely, the results highlight a significant dependency on the embedding depth. As evidenced by the results for mismatched configurations in Tables 2 and 3, cross-depth generalization is poor. For instance, a model trained on 1-LSB samples fails to classify manipulations on 3-LSB or 5-LSB images, with accuracy dropping to the 16%–25% range. This sharp decline indicates that LSB steganography introduces highly depth-specific statistical footprints. A manipulation applied to a 1-bit payload results in a fundamentally different degradation pattern compared to the same manipulation applied to a 5-bit payload. Consequently, features learned from one depth do not transfer linearly to another. However, among the tested depths, the 3-LSB configuration proves to be the most balanced set point. While it struggles to generalize to 1-LSB (likely due to the scarcity of signal in 1-bit embeddings), it maintains a reasonable transferability to 5-LSB in the P8-8 and P6-8 protocols (≈81% accuracy for both). This reinforces our selection of 3-LSB as the reference configuration for the subsequent morphing analysis, as it provides a robust signal that is sufficiently representative of the structural degradation caused by manipulations.

5.3. Morph impact analysis

While identity-preserving manipulations introduce noise or smoothing, morphing attacks fundamentally restructure the image to blend two identities. This section evaluates whether the proposed LSB watermarking framework can detect and classify these malicious alterations, specifically comparing traditional landmark-based method (FaceMorpher) and modern diffusion-based generation (DiffMorpher).

Distinctive Degradation Patterns: Unlike global filters (e.g., blurring) that affect pixels uniformly, morphing introduces complex, non-linear artifacts into the LSB plane. In the case of FaceMorpher,



(a) (b)

Figure 4: Confusion matrices for identity-preserving manipulation classification using the 3-LSB configuration. (a) P8-8 protocol (with seen parameters only) and (b) P6-8 protocol (including unseen parameters).

the warping process used to align facial landmarks disrupts the spatial coherence of the embedded bits. Consequently, these attacks induce significantly higher distortion levels compared to the least aggressive benign manipulations, such as sharpening. Moreover, the resulting degradation pattern diverges fundamentally from that observed in deep steganographic methods [4]. While deep models tend to localize watermark corruption strictly to the facial regions where landmarks are manipulated (e.g., eyes and mouth), our proposed LSB approach, enhanced by the scattering mechanism, disperses these local alterations across the entire spatial extent of the recovered secret (Figure 5). This global propagation amplifies the evidence of tampering, ensuring that even subtle, localized blending operations result in widespread and easily detectable corruption of the integrity marker, thereby increasing the security of the detection mechanism.

In contrast, DiffMorpher relies on a generative process that synthesizes new pixel values based on latent noise vectors. This effectively overwrites the original LSB content with generated high-frequency noise. Although the resulting facial image appears hyper-realistic to the human eye, the recovered secret reveals a complete destruction of the original signal (Figure 5), replaced by a unique pseudo-random noise pattern distinct from standard Gaussian noise.

Classification Performance: These distinctive degradation signatures allow the classifier to identify morphing attacks with high precision. As presented in Table 4, the model achieves an overall accuracy of 98.74% in the P8-8 protocol using the reference 3-LSB configuration. This result is particularly significant because it demonstrates the model's ability to distinguish malicious identity alterations from the six benign manipulations analyzed previously.

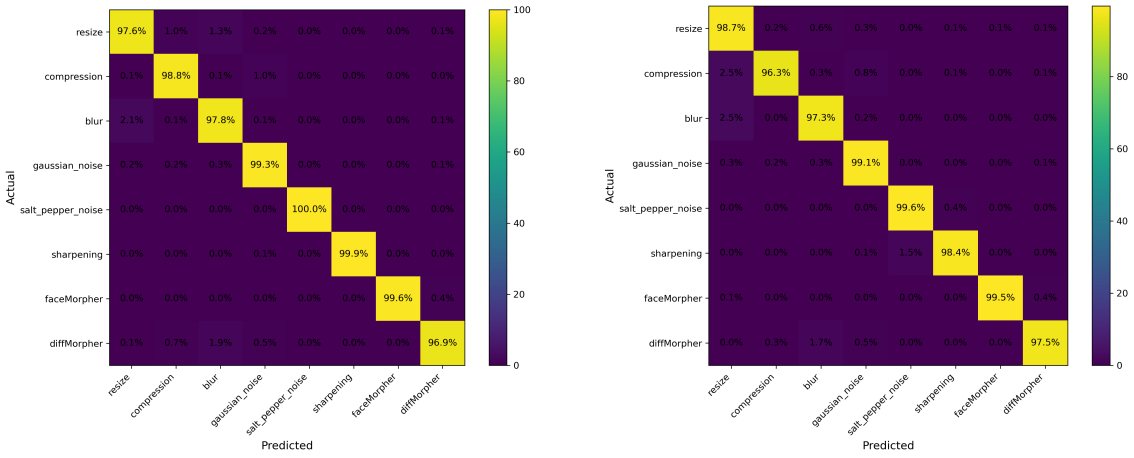
Furthermore, the system demonstrates exceptional robustness in the P6-8 protocol, maintaining an accuracy of 98.33%. The minimal drop in performance (less than 0.5%) indicates that the model has learned the structural nature of morphing artifacts such as the specific warping traces or generative diffusion patterns, rather than memorizing specific attack instances. This suggests that the fragile LSB watermark serves as a powerful forensic tool since it forces morphing algorithms to leave a visible trace in the hidden domain, even when the visible domain is perceptually flawless.

Finally, inspection of the confusion matrices reveals that the classifier distinguishes between the two specific morphing techniques with near-perfect precision: FaceMorpher is detected with 99.6% and 99.5% accuracy in the P8-8 and P6-8 protocols, respectively, while DiffMorpher achieves 96.9% and 97.5% (Figure 6). This confirms that, despite both being identity-altering attacks, each mechanism leaves a unique and clearly separable footprint in the recovered secret.

Figure 5: Morphing and watermark recovery for DiffMorpher (top row) and FaceMorpher (bottom row). Columns show, from left to right, the original watermarked image, the second identity, the morphed result, and the recovered watermark pattern.

Table 4
Classification performance on identity-preserving and identity-altering manipulations (intra 3-LSB), under the P8-8 and P6-8 sub-protocols.

Protocol	Accuracy	Precision	Recall	F1 Score
P8-8	98.74%	98.75%	98.74%	98.74%
P6-8	98.33%	98.36%	98.33%	98.33%



(a) (b)

Figure 6: Confusion matrices of the 8-class manipulation classifier for the P8-8 (left) and P6-8 (right) protocols. The six non-identity-altering operations and the two identity-altering morphing attacks are accurately distinguished.

Interestingly, when comparing the results before and after the introduction of malicious manipulations, the overall classification performance improves in the latter setting. This enhancement cannot be attributed solely to a better discrimination of the newly added classes: the improvement is also evident within the initial benign classes. For example, under the P6–8 protocol, the confusion between salt-and-pepper noise and sharpening is significantly reduced once morphed samples are included in the training and evaluation pipeline (Figures 4-b and 6-b). This is noteworthy because a visual inspection reveals FaceMorpher produces artefacts that partially resemble those introduced by the addition of salt-and-pepper noise (Figures 3-f and 5). Our hypothesis is that the inclusion of multiple morphing attacks substantially increases the variability and richness of the degradation patterns present in the extracted LSB payloads. Although some morphing artifacts overlap with those produced by specific benign manipulations, the broader manipulation space forces the classifier to organize its feature

representation more effectively. In particular, the increased diversity encourages the model to learn more discriminative and stable features, thereby refining the decision boundaries among benign perturbations and mitigating confusions, such as the one between salt-and-pepper noise and sharpening, that arise in the more constrained benign-only scenario.

6. Conclusions

The proposed work presents a fragile watermarking framework for verifying the integrity of ICAO-compliant biometric images. The method relies on traditional LSB steganography and embeds a known marker within the facial image, which is later extracted to detect post-acquisition manipulations by analysing its degradation. To assess the diagnostic potential of the retrieved secret, a set of manipulations was considered, including two morphing techniques and identity-preserving transformations such as compression and noise addition. Beyond detecting possible alterations, the study also evaluated the effectiveness of a classifier in identifying the type of manipulation by exploiting the degradation patterns observed in the recovered secret. The results demonstrated that fragile LSB-based watermarking allows a binary distinction between altered and unaltered images through the analysis of the extracted secret image. Furthermore, the classifier appears to be able to discriminate between secrets subjected to different manipulations, suggesting that fragile LSB watermarking may be a valuable tool for integrity verification in the context of ICAO-compliant biometric images. Future research will extend this analysis to additional manipulation strategies and explore the steganalysis resistance of LSB-based fragile watermarking, aiming to evaluate its practical security and robustness in real-world biometric pipelines.

Acknowledgments

This work was partially supported by Project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU. Davide Ghiani's PhD grant is partly funded by Dedem SpA under the PNRR program.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT5.2 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. M. La Cava, G. Orrù, M. Drahansky, G. L. Marcialis, F. Roli, 3d face reconstruction: the road to forensics, ACM Computing Surveys 56 (2023) 1–38.

[2] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, C. Busch, Face image quality assessment: A literature survey, ACM Computing Surveys (CSUR) 54 (2022) 1–49.

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, An Introduction to Digital Face Manipulation, Springer International Publishing, Cham, 2022, pp. 3–26. URL: https://doi.org/10.1007/978-3-030-87664-7_1. doi:10.1007/978-3-030-87664-7_1.

[4] D. Ghiani, J. D. R. Chivata, S. Lilliu, S. M. La Cava, M. Micheletto, G. Orrù, F. Lama, G. L. Marcialis, Fragile watermarking for image certification using deep steganographic embedding, arXiv preprint arXiv:2504.13759 (2025).

[5] J. D. R. Chivata, D. Ghiani, S. M. La Cava, M. Micheletto, G. Orrú, F. Lama, G. L. Marcialis, Deep data hiding for icao-compliant face images: A survey, arXiv preprint arXiv:2508.19324 (2025).

[6] S. Gupta, A. Goyal, B. Bhushan, Information hiding using least significant bit steganography and cryptography, International Journal of Modern Education and Computer Science 4 (2012) 27.
[7] O. Çeliktutan, S. Ulukaya, B. Sankur, A comparative study of face landmarking techniques, EURASIP Journal on Image and Video Processing 2013 (2013) 13.
[8] Z. W. Blasingame, C. Liu, Leveraging diffusion for strong and high quality face morphing attacks, IEEE Transactions on Biometrics, Behavior, and Identity Science 6 (2024) 118–131.
[9] International Civil Aviation Organization, Machine Readable Travel Documents: Part 9 – Deployment of Biometric Identification and Electronic Storage of Data in MRTDs, Technical Report Doc 9303, Part 9, International Civil Aviation Organization (ICAO), Montréal, Canada, 2021.
[10] A. Wolf, Icao: Portrait quality (reference facial images for mrtd), version 1.0. standard, International Civil Aviation Organization (2018).
[11] Z. Wang, O. Byrnes, H. Wang, R. Sun, C. Ma, H. Chen, Q. Wu, M. Xue, Data hiding with deep learning: A survey unifying digital watermarking and steganography, IEEE Transactions on Computational Social Systems 10 (2023) 2985–2999.
[12] Y. Zhao, B. Liu, M. Ding, B. Liu, T. Zhu, X. Yu, Proactive deepfake defence via identity watermarking, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 4602–4611.
[13] J. Wang, H. Wang, J. Zhang, H. Wu, X. Luo, B. Ma, Invisible adversarial watermarking: A novel security mechanism for enhancing copyright protection, ACM Transactions on Multimedia Computing, Communications and Applications 21 (2024) 1–22.
[14] X. Zhang, R. Li, J. Yu, Y. Xu, W. Li, J. Zhang, Editguard: Versatile image watermarking for tamper localization and copyright protection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 11964–11974.
[15] J. Zhu, R. Kaplan, J. Johnson, L. Fei-Fei, Hidden: Hiding data with deep networks, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 657–672.
[16] K. A. Zhang, A. Cuesta-Infante, L. Xu, K. Veeramachaneni, Steganogan: High capacity image steganography with gans, arXiv preprint arXiv:1901.03892 (2019).
[17] X. Ke, H. Wu, W. Guo, Stegformer: rebuilding the glory of autoencoder-based steganography, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 2723–2731.
[18] T. Bollé, E. Casey, M. Jacquet, The role of evaluations in reaching decisions using automated systems supporting forensic analysis, Forensic Science International: Digital Investigation 34 (2020) 301016.
[19] A. K. Jain, D. Deb, J. J. Engelsma, Biometrics: Trust, but verify, IEEE Transactions on Biometrics, Behavior, and Identity Science 4 (2021) 303–323.
[20] M. M. Emam, A. A. Aly, F. A. Omara, An improved image steganography method based on lsb technique with random pixel selection, International Journal of Advanced Computer Science and Applications 7 (2016).
[21] D. S. Ma, J. Correll, B. Wittenbrink, The chicago face database: A free stimulus set of faces and norming data, Behavior research methods 47 (2015) 1122–1135.
[22] K. Zhang, Y. Zhou, X. Xu, B. Dai, X. Pan, Diffmorpher: Unleashing the capability of diffusion models for image morphing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7912–7921.
[23] M. M. Sadek, A. S. Khalifa, M. G. Mostafa, Robust video steganography algorithm using adaptive skin-tone detection, Multimedia Tools and Applications 76 (2017) 3065–3085.
[24] J. Sun, Z. Yang, Y. Zhang, T. Li, S. Wang, High-capacity data hiding method based on two subgroup pixels-value adjustment using encoding function, Security and Communication Networks 2022 (2022) 4336526.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.