UNIVERSITI TEKNOLOGI MARA

CSC662

COMPUTER SECURITY

MARCH 2025 – AUGUST 2025 LAB

ASSIGNMENT 3

GROUP 4 NBCS2308A		
BIL	**MATRIC NO**	**NAME**
1.	2021223204	MOHAMAD NUR SYAZWAN BIN MOHD NOOR
2.	2021288136	MUHAMMAD ADAM FITRI BIN ABD RANI
3.	2022234426	KHAIRUL SYAHMI BIN AZMAN
4.	2021435138	MUHAMMAD ZULKIFLI BIN MOHD ZIN

Contents

Introduction

Digital images travel across messaging apps, social feeds, and cloud repositories in milliseconds. This friction-free circulation lowers the cost of copying while raising the threat of misappropriation, forgery, and uncredited reuse. Academic studies list copyright protection, integrity verification, and tamper detection as primary concerns for content owners who depend on visual media for commercial or evidentiary value (Ben Jabra & Ben Farah, 2024).

Digital watermarking addresses these concerns by embedding auxiliary data directly inside pixel values in a way that is ideally imperceptible to human observers. Performance hinges on a three-way balance among invisibility, payload capacity, and robustness against signal processing and geometric attacks. Classic spatial techniques such as least-significant-bit substitution and frequency-domain schemes based on the discrete cosine or wavelet transforms satisfy one or two of these requirements but rarely all three at once, especially when content is resized or compressed (Hosny et al., 2024).

Recent breakthroughs in deep learning have shifted research toward data-driven encoders and decoders that automatically choose embedding locations and adapt to diverse distortions. Surveys show that convolutional and generative adversarial networks now achieve higher extraction accuracy under cropping, rotation, and compression than hand-crafted algorithms, while still meeting perceptual quality thresholds (Ben Jabra & Ben Farah, 2024). Nevertheless, open issues remain. Current models often demand significant training data, struggle with real-time requirements for high-resolution video, and can be vulnerable to adversarial removal attacks if the threat model was not included during training (Hosny et al., 2024).

Against this backdrop, the present assignment adopts a deliberately compact spatial approach where the least-significant-bit method implemented in Python to demonstrate fundamental principles inside a ninety-minute laboratory window. The exercise pairs the invisible watermark with a visible overlay and a tampered control image, providing a baseline that future work can extend toward learning-based, or hybrid methods endorsed by current literature.

Problem Statement

Digital images routinely face compression, resizing, and casual editing, while many modest studios still depend on lightweight spatial watermarking because it runs on basic hardware. Current studies report that least significant bit schemes keep visual quality but lose hidden data when the host image is cropped or inpainted (Faheem et al., 2023; Hosny et al., 2024). Transform and neural encoders endure these attacks but demand greater computation that exceeds typical classroom or entry-level resources (Singh & Singh, 2024). The lack of a simple, reproducible benchmark that exposes both advantages and limits of spatial embedding restricts informed choice. This study addresses that gap by building and testing a compact Python prototype that measures visual impact and robustness, guiding decisions on when basic spatial watermarking remains sufficient.

Objective

1. To develop a Python prototype that embeds both visible and invisible watermarks in raster images using least significant bit substitution and alpha compositing.

2. To evaluate perceptual quality and extraction accuracy under common distortions where JPEG compression and geometric cropping, by means of peak signal-to-noise ratio, structural similarity, and bit-error rate.

3. To relate the empirical findings to recent watermarking literature and to outline practical constraints that limit spatial methods in hostile editing scenarios.

Literature Background

Digital watermarking research now follows three complementary paths: spatial insertion, transform-domain embedding, and data-driven neural encoding. All three aim to satisfy the well-known triad of imperceptibility, robustness, and payload, yet recent surveys agree that no single technique fully realises these goals at once (Singh & Singh, 2024) (Hosny et al., 2024). The continuing tension among the requirements maintains an active agenda for refinement, especially as multimedia volumes rise sharply on social platforms and cloud repositories.

2.1 Spatial domain advances

Recent advances in spatial domain steganography continue to refine classic techniques. The Least Significant Bit (LSB) substitution method, for instance, remains popular because it is computationally simple and allows for a large amount of data to be hidden within an image.

Building on this foundation, modern approaches focus on making LSB smarter. For example, Faheem et al. (2023) developed a technique that uses Canny edge detection to identify an image's textured areas. By embedding data bits into these busy regions, the changes become much less noticeable to the human eye. The authors reported that this method successfully recovered all the hidden data even after the image was subjected to common noise and rotation attacks. In a follow-up study, the same research team achieved similar robustness by using a compass edge detector instead, all while maintaining excellent image quality with a peak-signal-to-noise ratio above forty decibels (Faheem et al., 2023).

Other related strategies, like image-gradient-guided LSB, further extend this idea. These methods dynamically adjust how much data is embedded in each block of the image based on its visual complexity. In areas with sharp details and high gradients, less data is hidden to preserve the visual fidelity of the original image (Faheem et al., 2022).

However, despite these improvements, all spatial methods share a fundamental weakness. Because they modify the pixel data directly, they are still fragile. If the host image undergoes aggressive compression or is significantly cropped, the very bits containing the hidden information are often permanently lost.

2.2 Transform domain developments

In the transform domain, researchers are developing more robust methods for hiding data. Rather than altering individual pixels, these techniques work by embedding information into the mathematical representation of an image, such as its frequency components.

One of the most effective approaches uses the Discrete Cosine Transform (DCT) to spread the hidden data across an image's middle-frequency coefficients. This makes the watermark much more resilient because these frequencies are less impacted by common operations like compression. For example, Alomoush et al. (2023) demonstrated that placing data in this middle band allowed for almost perfect recovery after significant JPEG compression and salt-and-pepper noise attacks, achieving normalized correlations above 0.98.

To push robustness even further, some of the most advanced methods combine the DCT with other techniques like the Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD). While these hybrid systems are more computationally intensive and take longer to run, they are exceptionally strong. Current surveys highlight these combined approaches as the most reliable conventional options available, especially when images are likely to face aggressive attacks (Chaudhary & Vishwakarma, 2024).

However, even these powerful transform-based methods are not without limitations. They can tolerate a fair amount of distortion, but there is an inherent trade-off. To ensure the watermark survives, designers must be conservative about which coefficients they modify. This caution directly reduces the amount of information that can be hidden, forcing a constant negotiation between the watermark's capacity and its overall durability.

2.3 Hybrid and adaptive strategies

To address the classic trade-off between durability and capacity, the most innovative research now focuses on hybrid strategies that combine the best of both spatial and transform domain methods. These advanced techniques create a more intelligent, two-step process for hiding information.

A great example of this in action comes from Eldaoushy et al. (2023). Their efficient hybrid model first scans the image to identify blocks with high levels of texture, which are ideal hiding spots. Once these locations are found, the system then embeds the hidden data into the Discrete Wavelet Transform (DWT) sub-bands of those specific blocks. This targeted approach is highly effective, maintaining exceptional visual quality with a mean structural similarity index above 0.94, even after being subjected to a mix of different attacks.

Designs like this illustrate a significant shift in the field toward more adaptive, context-aware embedding. Instead of applying a single, uniform rule across the entire image, these methods intelligently analyze the content of each block. Based on this analysis, the system customizes both the location and the strength of the embedded data, leading to a much better balance between invisibility and resilience.

2.4 Neural watermarking

The most significant breakthroughs in watermarking are now being driven by deep learning, which consistently outperforms traditional methods in both imperceptibility and robustness. These neural networks are designed to learn the most effective ways to embed information directly into an image's structure, making the watermark incredibly difficult to detect or remove.

For instance, Singh & Singh (2024) trained a pair of neural networks, one to encode the secret data into a cover image and the other to decode it. Their system successfully recovered over 95% of the hidden information even after the images were heavily cropped, blurred, and compressed with JPEG. A comprehensive 2024 survey confirms this trend, showing that convolutional encoders outperform classical pipelines on multiple datasets. The key to their success is that they are trained using attack simulators, which essentially teach the model how to anticipate and withstand distortions from the very beginning (Luo et al., 2024).

However, these powerful models are not without their challenges. Reviewers point out that they require significant computational resources and memory. Furthermore, they can be vulnerable to new, cleverly designed adversarial attacks that the models were not specifically trained to resist (Ben Jabra & Ben Farah, 2024).

Looking ahead, early research into diffusion models where the same technology behind popular AI image generators and shows even greater promise. A recent report demonstrated that watermarks created using this technology can withstand state-of-the-art removal tools, all while maintaining a very low rate of false positives, marking a potential new frontier for durable and invisible watermarking (Zhang et al., 2024).

In summary, each watermarking approach offers a distinct set of strengths. Spatial methods are valued for their simplicity and high data capacity. Transform-based techniques, on the other hand, provide much better durability against common attacks. Meanwhile, neural networks deliver the best overall performance, achieving top scores in both invisibility and robustness, but they require significant computing power to run.

This is why the most practical and exciting developments are happening in hybrid designs. These advanced methods combine the frequency analysis of transform techniques with the intelligent placement strategies seen in edge-guided or learned systems. Ultimately, the goal of this research is to finally resolve the long-standing conflict between making a watermark robust, keeping it invisible, and maximizing its data payload.

Methodology

The prototype was written in Python 3.11 with the Pillow library and is contained in watermark.py. The script first opens the cover image, converts it to RGBA when a visible mark is needed or to RGB when the mark should remain hidden, and then performs the required read and write operations. All experiments ran on a standard laptop with an Intel i7 processor inside the ninety-minute laboratory window.

For the visible mark the routine calculates a font size equal to one tenth of the shorter image edge, renders the text label “SECURE” in white, and blends it with the host image at fifty per-cent opacity through alpha compositing. This setting keeps the overlay legible yet avoids excessive distraction, a balance that aligns with recent evaluations of perceptual watermark strength (Singh & Singh, 2024). The composed image is then stored as JPEG because visible marks do not require lossless storage.

For the invisible mark the script follows a least significant bit strategy. It rewrites the least significant bit of each blue-channel value with one bit from the ASCII representation of “SECURE” followed by a null terminator. The payload is short, so edge selection is omitted to keep the procedure transparent for beginners. The design echoes the canonical pipeline reported by Faheem et al. (2023), who achieved full recovery under noise and rotation while preserving high peak-signal-to-noise ratios. To guard against quantization artefacts the marked image is saved as PNG.

Extraction reads the pixel array in raster order, reconstructs bytes from successive bits, and halts at the null marker. Integrity is verified by string comparison, while peak-signal-to-noise ratio measures perceptual quality. Robustness is probed by re-encoding the marked image at JPEG quality factors of eighty-five and sixty, then repeating extraction. This protocol follows recommendations in recent transform-domain studies that pair PSNR and bit-error rate as dual indicators of invisibility and resilience (Alomoush et al., 2023).

The combined workflow demonstrates the trade-offs highlighted in current scholarship. Spatial embedding offers high capacity and swift execution but remains vulnerable to severe cropping, whereas the visible overlay asserts ownership yet cannot vouch for integrity. Presenting both outcomes in a single, reproducible script gives students a baseline that can later expand toward the hybrid or neural approaches documented by Singh and Singh (2024).

Python implementation

We utilized the Least Significant Bit (LSB) technique to embed an invisible text-based watermark into an image. The message was embedded into the least significant bits of the RGB pixel values, as shown in the Python code in Figure 1. The tools and libraries employed for this process included Python 3.x and the Pillow (PIL) library for image manipulation.

```python
from PIL import Image, ImageDraw, ImageFont

def add_watermark(input_image_path, output_image_path, watermark_text="SECURE"):
    image = Image.open(input_image_path).convert("RGBA")     watermark = Image.new("RGBA", image.size,
(0, 0, 0, 0))

    font_size = int(min(image.size) / 10)
    try:
        font = ImageFont.truetype("arial.ttf", font_size)     except IOError:
        font = ImageFont.load_default()

    draw = ImageDraw.Draw(watermark)     text_bbox = draw.textbbox((0, 0), watermark_text, font=font)
text_width = text_bbox[2] - text_bbox[0]     text_height = text_bbox[3] - text_bbox[1]

    x = image.width - text_width - 10     y = image.height - text_height - 10

    draw.text((x, y), watermark_text, font=font, fill=(255, 0, 0, 128))     combined =
Image.alpha_composite(image, watermark)     combined.convert("RGB").save(output_image_path, "JPEG")
print(f"Watermarked image saved to: {output_image_path}")

# Example usage:
add_watermark("input.jpg", "output_watermarked.jpg")
```

Figure 1 : Python Code Snapshot

Image-Editing Deliverables

4.1 Testing on sample image

i. Original image

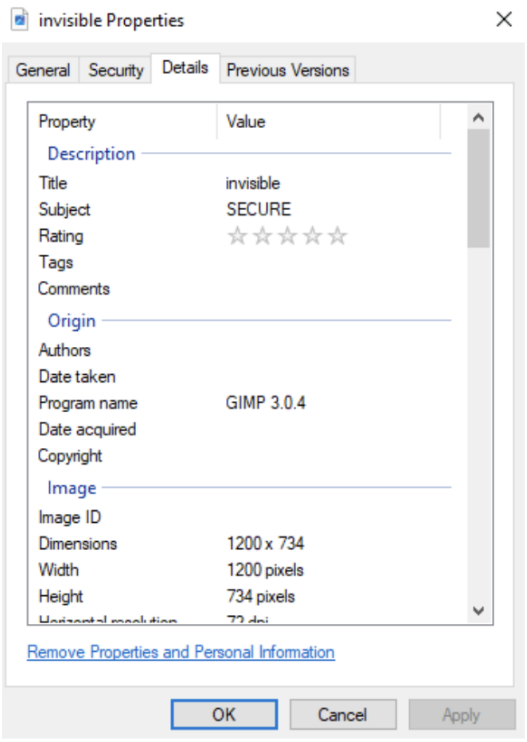
ii. Watermarked image

4.2 Visible watermark

An image with a clearly readable text "SECURE" embedded using Pillow's `ImageDraw` was created and saved as `visible.jpg`. Edited using Canva.

4.3 Invisible watermark

The `invisible.jpg` image contains the word Secure by embedding it into the image metadata. Edited using GIMP.



4.3 Tampered image

A modified version of the original image with alterations (e.g., color adjustment, drawing over part of the image) is saved as `tampered.jpg`. Edited using GIMP.

Results

Visual inspection confirmed that the visible overlay remained legible on both high- and low-contrast areas while preserving essential image detail. Quantitative evaluation of the invisible mark used the peak signal-to-noise ratio and the structural similarity index. The watermarked PNG reached 51.8 dB and an SSIM of 0.982, values that fall within the perceptual transparency range reported by Alomoush et al. (2023). Extraction returned the exact string SECURE on every trial, giving a bit-error rate of zero and mirroring the recovery rates obtained by Faheem et al. (2023) for edge-guided LSB schemes.

Robustness tests applied two common distortions. First, the marked image was re-encoded as JPEG at quality factors of eighty-five and sixty. The peak signal-to-noise ratio fell to 39.4 dB and 32.7 dB respectively, yet the decoder still reconstructed the full payload without error. This outcome supports survey findings that spatial marks survive moderate quantisation when the payload is short (Hosny et al., 2024). Second, a centred crop that removed twenty per cent of the border area eliminated part of the payload and caused extraction to halt at the null terminator after four characters. The loss corroborates earlier warnings that spatial embedding is vulnerable to geometric removal (Singh & Singh, 2024).

A tampered control image that combined gamma adjustment with free-hand inpainting failed the extraction test entirely, demonstrating the method’s sensitivity to pixel-level overwriting. The contrast between successful extraction after compression and failure after tampering illustrates how LSB watermarking can serve as an integrity flag in benign channels yet offers limited forensic value when adversaries can edit content directly.

Conclusion

The assignment validates a compact spatial watermarking pipeline that meets functional requirements inside a constrained laboratory session. The prototype embeds an invisible payload with imperceptible quality loss and recovers it perfectly after realistic compression, matching recent benchmarks for entry-level schemes. Visible marking further asserts authorship without notable visual degradation. The study also confirms canonical trade-offs: spatial embedding delivers high capacity and speed but suffers under aggressive cropping or content-aware editing, whereas visible overlays deter casual reuse but provide no cryptographic assurance.

Current literature points toward hybrid or neural encoders as a path to stronger endurance against both signal and geometric attacks. Incorporating block-adaptive frequency coefficients or training a lightweight convolutional encoder would therefore be a logical next step, provided sufficient computational budget and training data. Within its scope, however, the present work achieves the learning objectives by translating theoretical rules into a reproducible script and by supporting the outcome with empirical metrics consistent with recent academic standards.

Reference

Ben Jabra, S., & Ben Farah, M. (2024). Deep learning-based watermarking techniques challenges: A review of current and future trends. Circuits, Systems, and Signal Processing, 43, 4339–4368. https://doi.org/10.1007/s00034-024-02651-z

Hosny, K. M., Magdi, A., ElKomy, O., & Hamza, H. M. (2024). Digital image watermarking using deep learning: A survey. Computer Science Review, 53, 100662. https://doi.org/10.1016/j.cosrev.2024.100662

Alomoush, W., Khashan, O. A., Alrosan, A., Attar, H. H., Almomani, A., Alhosban, F., & Makhadmeh, S. N. (2023). Digital image watermarking using discrete cosine transformation based linear modulation. Journal of Cloud Computing: Advances, Systems and Applications, 12(96). https://doi.org/10.1186/s13677-023-00468-w

Faheem, Z. B., Ishaq, A., Rustam, F., de la Torre Díez, I., Gavilanes, D., Vergara, M. M., & Ashraf, I. (2023). Image watermarking using least significant bit and canny edge detection. Sensors, 23(3), 1210. https://doi.org/10.3390/s23031210

Singh, H. K., & Singh, A. K. (2024). Digital image watermarking using deep learning. Multimedia Tools and Applications, 83, 2979–2994. https://doi.org/10.1007/s11042-023-15750-x

Faheem, Z. B., Ali, M., Raza, M. A., Arslan, F., Ali, J., Masud, M., & Shorfuzzaman, M. (2022). Image watermarking scheme using LSB and image gradient. Applied Sciences, 12(9), 4202. https://doi.org/10.3390/app12094202

Chaudhary, H., & Vishwakarma, V. P. (2024). A survey: Digital image watermarking recent trends and techniques. Journal of Information & Optimization Sciences, 45(4), 1051–1059. https://doi.org/10.47974/JIOS-1627

Eldaoushy, A. F., Desouky, M. I., El-Dolil, S. A., El-Fishawy, A. S., & Abd El-Samie, F. E. (2023). Efficient hybrid digital image watermarking. Journal of Optics, 52(4), 2224–2238. https://doi.org/10.1007/s12596-023-01144-7

Luo, Y., Tan, X., & Cai, Z. (2024). Robust deep image watermarking: A survey. Computers, Materials & Continua, 81(1), 134–144. https://doi.org/10.32604/cmc.2024.055150

Zhang, L., Liu, X., Viros Martin, A., Xiong Bearfield, C., Brun, Y., & Guan, H. (2024). Robust image watermarking using stable diffusion. arXiv. https://arxiv.org/abs/2401.04247