Deep Learning-Based Image Steganography Detection Using AI-Powered Systems

OPEN ACCESS

Volume: 13

Special Issue: 2

Month: January

Year: 2026

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Citation:
Sahu, Geeta, et al. "Deep Learning-Based Image Steganography Detection Using AI-Powered Systems." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 2, 2026, pp. 127–36.

DOI:
https://doi.org/10.34293/sijash.v13iS2-i2-Jan.10533

Dr. Geeta Sahu
Assistant Professor, Department of Information Technology & Data Science
Vidyalankar School of Information Technology, Mumbai, Maharashtra, India

Maria Jenisha
Department of Information Technology & Data Science
Vidyalankar School of Information Technology, Mumbai, Maharashtra, India

Sangeeta Prasad
Department of Arts, Commerce and Science
NKES Degree College, Mumbai, Maharashtra, India

Abstract
Because it hides important information in seemingly commonplace media like images, audio files, and video files, steganography is crucial to secure digital communication. This method guarantees data confidentiality and shields sensitive data from unauthorised access or transmission interception. However, conventional detection methods like least significant bit (LSB) analysis, chi-square testing, and visual inspection have become less successful due to the quick development of steganography embedding algorithms, many of which incorporate adaptive and deep learning-based techniques. To provide good accuracy and generalizability for steganography detection, the existing study employs an intelligent steganography framework that is a hybrid mixture of machine learning and artificial intelligence. The proposed system makes use of a hybrid combination of deep techniques architecture with convolution neural networks (CNN), followed by feature extraction methods such as entropy calculations, pixel relation analysis, and histogram-based image texture evaluation. The CNN model learns more of special representation for more accurate identification of hidden information from the images. The extracted features will capture the low-level anomalies from the image structure. Evaluation was done based on the datasets – BossBase, COCO & image net which contain both normal and stego images. The results conclude that the CNN model performs the best approach rather than the traditional methods in terms of performance metrics such as recall, precision, accuracy, etc. Hence, with the power of a deep AI and CNN combination, the proposed methodology suggested a significant enhancement in the steganalysis.

Keywords: Pixel Correlation, Entropy Analysis, Digital Forensics, Image-Based Security, Payload Robustness.

Introduction

The importance of data security and integrity has increased, especially with the proliferation of smart systems that handle massive amounts of private data in a variety of applications. To guarantee data security, privacy, and integrity, a number of cutting-edge technologies are integrated. Blockchain technology is one of the most cutting-edge approaches in this context. One technique for concealing classified information in non-secret

content is steganography. In data that we communicate or deliver in public, we can conceal the existence of a private communication. Because stenographic techniques can be utilized by malware or used to propagate harmful software, they pose a serious risk to consumers.

Multimedia data, including pictures, is used as a carrier in steganography. These techniques are frequently referred to as image steganography and digital media steganography, respectively. For instance, the Vawtrak malware exploited favicon graphics to conceal URL addresses. A considerable threat is posed by the increasing number of malware infections that exploit covert transmission, especially those based on steganography.

The discipline of steganography has undergone a substantial transformation due to the rapid advancement of artificial intelligence (AI), which has expanded its scope beyond traditional image-based approaches to other realms like language and 3D mesh data concealment. With a focus on three different modalities—image, linguistic, and 3D mesh—this review offers a succinct, understandable, and critical analysis of current AI-powered steganography techniques. In contrast to most surveys that concentrate just on one modality, this paper analyses how AI has changed embedding mechanisms, evaluation methodologies, and security concerns while highlighting some modalities and identifying their challenges. Deep models like GANs and Transformers have enhanced imperceptibility and extraction accuracy in image-based steganography, however they are limited in terms of computing efficiency and extraction consistency. Large language models have revived linguistic steganography, which was formerly hampered by semantic fragility.

Because image steganography and watermarking algorithms are essential for secret data transmission, copyright protection, and traceability, a great deal of research has been done on them. There is still a dearth of thorough research devoted to deep learning-based image steganography and watermarking algorithms, despite encouraging outcomes and multiple surveys suggested in the literature. In this research, we investigate three key areas: training methodologies, structure models, and neural networks. The extensive body of research in this area is covered in our review. Additionally, we offer a thorough statistical analysis from several angles, such as models, loss functions, platforms, datasets, and assaults.

Steganography outperforms other methods for protecting data from possible attacks. Strong information hiding strategies are necessary in today's digital environment, which has always been a hot topic for academics and researchers. These days, sharing confidential information over standard correspondence channels is vulnerable to a variety of hacks. As a result, everyone requires their privileged knowledge to be classified, respectable, and authentic. Several methods, such as improved declaration, computerized mark, and cryptography, are employed to address these security concerns. However, these tactics cannot be negotiated on their own. To address the demands for data protection via the Internet, steganography is a revolution that integrates recent developments in information compression, data theory, distributed range, and encryption.

Literature Review

The authors discuss how block chain and AI can improve cybersecurity. It addresses challenges including data privacy, malware monitoring, and intrusion detection. The result provides enhanced network security and intrusion prevention with a focus on structures of policy. AI, block chain, and smart contract-based cybersecurity compliance and threat response are automated by a recent study. It fixes human error and the absence of real-time monitoring. Real-time monitoring, enhanced audibility, decreased human intervention, and 91% accuracy in threat classification are among the outcomes.

However, this article attempts to close the gap between cybersecurity policy and practice by putting forth a revolutionary framework and system that combines block chain, artificial intelligence, and intelligent contracts. Our objective is to dynamically modify security measures in response to new threats found through cyber threat intelligence and automate the implementation of an organization's internal security standards.

AI, block chain, and smart contract-based cybersecurity compliance and attack response are automated in

a recent study. It addresses both the absence of real-time monitoring and human error. Real-time monitoring, enhanced audibility, decreased human intervention, and 91% threat classification accuracy are among the outcomes. Large, tagged datasets and sufficient processing power are necessary for GANs and other embedding methods. This emphasizes the need for hybrid AI-based steganalysis systems that combine deep learning and traditional statistical analysis for accurate and broadly applicable detection.

Authors state that non-AI techniques continue to dominate 3D mesh steganography, providing a rich environment for geometric deep learning innovation. Overview of design ideas, performance indicators are unique to each modality is also included in this review. The analysis shows that evaluation paradigms have changed, moving away from numerical fidelity criteria like PSNR and SSIM and toward semantic and perceptual metrics like LPIPS, BERT Score, and Harsdorf Distance. This review attempts to encourage early-stage scholars and practitioners to investigate new steganography frontiers in the AI era by showcasing recent developments and critical viewpoints across understudied disciplines.

The development of digital steganography began in the 1990s and continues to this day. Early steganography's primary features were payload, imperceptibility, and security. Steganography's progress competes with the steganalysis approach, and security is a crucial component. In addition to statistics, adaptive approaches are all used in the development of stenographic techniques. Robustness is added as a crucial component by several novel techniques. This increases the variety of stenographic method development and typically concentrates the method's objective on one or two factors. These days, papers categorize stegano according to its objectives.

To aid future study, the difficulties and possible avenues for investigation in the field of deep-learning image steganography and watermarking algorithms are finally covered.

The author states that the study investigated and evaluated different cover steganography methods that are now in use and found a useful area where everyone can profit. Additionally, we provide a thorough summary of the basic ideas. It is illustrated in several steganography domains, including the adaptive space, transform domain, and spatial domain. Furthermore, every space has unique characteristics. A few frequently used methods for enhancing stenographic security and increasing steganalysis capability are described, summarized, and potential analysis patterns are discussed. In our review, we also methodically distinguish between various approaches and highlight their benefits, drawbacks, difficulties, and importance.

The edge-based picture steganography system presented in this paper divides the cover image pixels into two classes: edge and non-edge. Because of their high tolerance level and chaotic nature, edge pixels typically conceal more hidden bits than non-edge pixels. Because the edge pixels' brightness differs from that of the nearby pixels, they are seen as noisy and are therefore challenging to predict. Additionally, on equal change, edge pixels typically have a higher tolerance level than non-edge pixels. We propose a new image steganography approach based on Difference of Gaussians (DoG) Edge detection for these two reasons. The three main stages of the suggested technique are extraction, embedding, and pre-processing cum edge detection.

Researchers in this discipline have been very interested in pixel value differencing (PVD) image steganography since its inception. Nevertheless, most PVD-based methods have either an improper extraction problem (IEP) or a falling-off boundary problem (FOBP). Thus, this work suggests a multi-directional pixel value differencing and modulus function (MDPVDMF) based method to solve these two problems.

Any one of the three ways can yield two different values for a 2 × 2-pixel block. The secret bits are then embedded using the difference values and the remainders of the pixel pairs. The purpose of the experiment was to determine how well the suggested method performed in terms of image quality measures such as peak signal-to-noise ratio (PSNR), embedding capacity (EC), and FOBP. According to the results, the EC is best for diagonal pairs with 3.10 bits per pixel (BPP), while PSNR is best for vertical pairs with 39.17 dB. Additionally, the suggested method has demonstrated remarkable resistance against regular and singular (RS) attacks, salt and pepper (S&P) noise, pixel difference histogram (PDH) analysis, and subtractive pixel adjacency matrix (SPAM) steganalysis.

A weighted difference of extended state-of-the-art feature vectors that are already employed in steganalysis is the definition of the distortion. By doing this, we can "preserve" the Steg analyst model and remain undetectable even for big payloads. Even when the embedder's feature set has a dimensionality of more than 107, this method can be applied effectively. To prevent known security flaws, the high dimensional model is required. We show why high-dimensional models are appropriate in steganography even though they may be problematic in steganalysis.

Problem Statement

The majority of traditional steganalysis techniques are restricted to specific algorithms, such as LSB, and are not adaptable enough to accommodate new embedding methods. The goal of this research is to create an AI-based system that, independent of the embedding algorithm employed, can automatically detect hidden data in images.

The study employs the helps categorize stenographic research according to its objectives and evaluations. Because it is directly related to the purpose of steganography, this paper also thoroughly examines the usage of assessment tools. Additionally, reviewed are approaches, advancements, challenges, concerns, and popular datasets. To raise the calibre of the research and analysis, the steganalysis literature was also included. Lastly, to make the progress of picture steganography easier for inexperienced researchers to understand, this survey offers debate, investigation, critical analysis, and a plain summary.

Traditional Steganalysis methods are limited in how adaptable they are to new/novel embedding methods and typically only apply to specific algorithms such as LSB. Our project's objective is to develop an AI-based steganalysis system that can automatically identify hidden content inside photographs, regardless of the type of embedding algorithm used.

Research Questions:

1. What improvement in the rate of true detections can be made by utilizing AI versus traditional LSB detection techniques?
2. Will CNN's (Convolutional Neural Networks) be generalizable across many different types of steganography?
3. In what manner will a hybrid model containing both AI and statistical features enhance the scalability of Cybersecurity solutions?

Research Methodology

To train and assess the suggested steganography detection framework, an expanded dataset containing both regular (cover) and stego (embedded) images was created for this project. Normal images: Several well-known benchmark datasets, such as COCO (Common Objects in Context), CIFAR-10, and BOSSBase, provided the cover images. By offering a variety of image types, resolutions, and content, these datasets enable the model to acquire resilient and widely applicable features. The following figure shows the overall flow of the proposed methodology.

Figure 1 Flow of the Proposed Methodology

Dataset

To conduct this research, an extended dataset containing both regular (cover) and stego (embedded) images was created in order to analyze the suggested steganography detection model.

Typical Images: The cover images were acquired from several established benchmark datasets, including COCO (Common Objects in Context), CIFAR-10, and BOSSBase. Each of these datasets includes several different types of images of different sizes and contents, which allows the model to identify features that are transferable to different situations.

COCO contains high-quality pictures of commonplace objects in intricate environments, which assist the model in learning about various colors, textures, and lighting options. BOSSBase is another frequently used dataset in steganography that is often employed for testing steganalysis methods; it consists of grayscale images selected specifically for testing steganography techniques. CIFAR-10 contains 60,000 images of low quality but represents ten distinct categories of images and can be used to measure model efficiency on minimal resolution and basic imagery.

Stego Images: Stego images were developed by embedding messages in the cover images using popular steganography software such as StegHide and OpenStego. Different software uses different embedding techniques with different payload capacities, thus producing variations in the patterns of the concealed data. Therefore, the dataset included both simple embedding methods such as least-significant-bit (LSB) to more sophisticated and adaptive methods.

Use of LSB for Information Hiding

Because LSB steganography is one of the oldest methods of hiding information inside digital images, it has become very popular due to its ease of use and the fact that most people do not notice the changes that are made to the digital image using this method. However, because LSB steganography is so easy to implement, it is also very easy for individuals using advanced tools for steganalysis to detect whether a message has been hidden inside an image by analyzing the statistical patterns of the image itself. Therefore, when using LSB steganography you should do so with caution and augment your use of this technique with additional security measures to ensure that your hidden messages are not detected by steganalytic tools.

This technique has some serious limitations in the sense that it is simple and therefore fairly easy to guess. One of those limitations is that the total number of bits that can be embedded per pixel when using LSB is only 1. Therefore, further refinements and improvements are needed to increase the maximum message payload and to improve the security of the technique.

Image Preparation

All the images added to the steganographic detection framework were pre-processed to improve the performance of the models used to classify images and to normalize the data used by the models. The following are the three preprocessing techniques used on the images before being entered into the steganography detection framework:

Grayscale Conversion: All images were converted from RGB (color) format to a grayscale format. This conversion reduced the computational complexities of the models developed to classify the images. By converting the image to a grayscale format, we were able to retain much of the important structural and textural information that is preserved in the original RGB representation of the images, while eliminating the influence of the color channels on the performance of the models; and the grayscale images maintain the underlying structure of the pixel intensity patterns that are necessary to detect the subtle changes that result from the addition of a hidden message.

Re-sizing of Images: To create uniformity within the dataset, we re-sized every image to the same dimension, 256 x 256 pixels. By utilizing a uniform dimension for input into the CNN (Convolutional Neural Network), batch processing is achieved, and spatial features maintained throughout the dataset. By using

a set size, we are able to maintain our computational efficiency while continuing to preserve all relevant information within the images.

Normalization of Images: We re-scaled the pixel intensity value(s) for each image to fall within predetermined minimum and maximum values (typically between 0 and 1, or from -1 to 1). To provide stability in learning for the CNN and provide a level playing field in terms of illumination, contrast, and intensity distributions for every image used for training, we performed normalization on all of the training images. Additionally, normalizing images will yield greater accuracy and convergence speed of the model.

AI/Machine Learning Models: We used AI and Machine Learning (ML) models and compared them against statistical methods in order to assess the viability of different methods of detecting steganography. Statistical Methods for Detection: Histogram Analysis: Histogram analysis is used to detect minute changes in the pixel intensity distribution caused by steganography. This method is a basic indication of the detection performance.

The machine learning (SVM) N/A class. The SVM model was trained using the previously identified statistical and pixel-level features, including histogram features, entropy (randomness), noise residuals, pixel correlations. The SVM model maps these features onto a high-dimensional space to determine the optimal decision boundary separating stego and cover images. This approach improves the accuracy of generalization compared to statistical approaches alone and facilitates the automatic classification of images.

Deep Learning Model (CNN). The convolutional neural network (CNN) model learned to automatically learn the hierarchical features from the preprocessed images. The CNN model automatically identifies pixel-level anomalies as well as the higher level spatial relationships associated with steganography. The CNN architecture consists of numerous convolutional layers using the ReLU activation function, pooling layers used for the dimensionality reduction, and fully connected layers for categorization tasks. The model uses batch normalization and dropout layers to improve generalization and reduce the likelihood of overfitting.

Comparing the baseline statistical approaches with the SVM and CNN models demonstrates that AI-based methods produce higher levels of accuracy, robustness, and scalability compared to other methods, and also that modern machine learning (ML) and deep learning algorithms are more effective than contemporary techniques for identifying hidden information, for a variety of embedding schemes. Figure 2 shows the AI-Base Framework for Image Steganography Detection Using Deep Learning.

Figure 2 AI-Base Framework for Image Steganography Detection Using Deep Learning.

Experimental Results & Discussions

The "AI-Based Image Steganography Detector" provides users with the ability browse and select images to analyse using the system's main user interface, as illustrated by the preceding figures. By selecting an image from their computer files, users will have the ability to extract any hidden messages from the selected image through direct interaction with the system (i.e., selecting an image file for processing and choosing to perform one or both actions of detecting and/or extracting a concealed message). Figure 3 shows the browse for a steganography image.

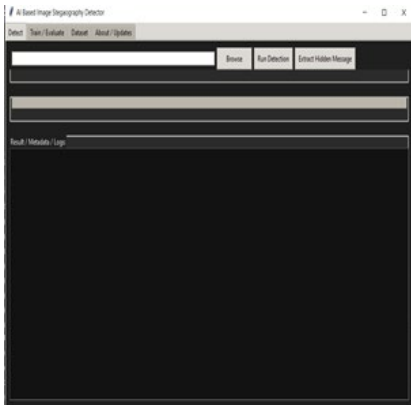

Figure 3 Browse for a Steganography Image

An example of how a user will be presented with a file selection window is illustrated in the figure 4. Users are given the option of selecting their chosen image file by either clicking on the browse button or dragging and dropping their selected image file onto the browse button of the main user interface.

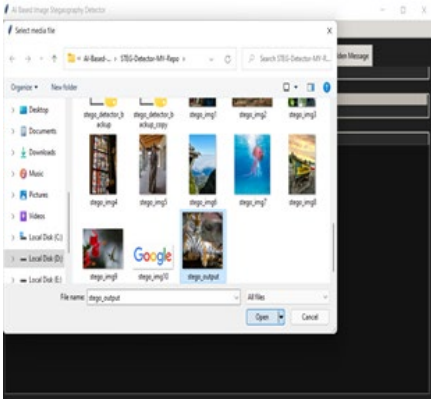

Figure 4 File Selection Window

Users can select the file from which a concealed message will be extracted using the provided file selection window. Once the image has been selected, the user can execute detection capabilities and/or extract the concealed message contained within the selected image file. Figure 5 shows the image path is selected.

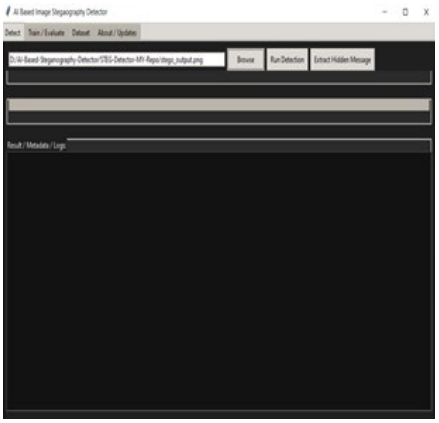

Figure 5 Selection of Image Path

The extraction of the concealed message from the selected stego image was achieved by the "AI-Based Image Steganography System," as shown in the figure above. The extraction module of the system works with high-level efficiency and accurately extracted the embedded/masked message "HELLO" from the selected image file. Figure 6 is the extraction module.

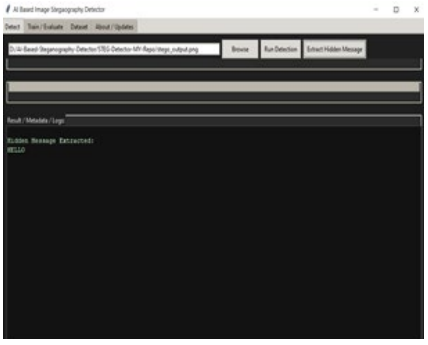

Figure 6 An Extraction Module

Evaluated against many different algorithm models to ascertain which of the models could provide the most accurate method for performance evaluation of steganography detection using random forest, the "AI-Based Image Steganography System" has concluded that Random Forest Model is the best choice for steganography detection. Figure 7 shows the comparison of the results with the AI algorithms.

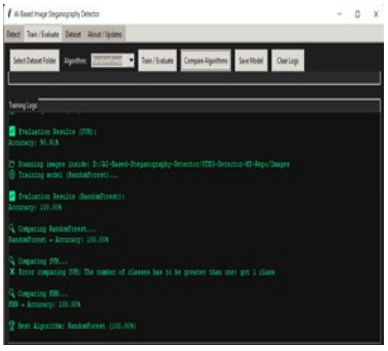

Figure 7 Comparison of the Results with the AI Algorithms

Based on a comparison of CNN models and LSB-based detection methods (i.e., traditional vs. non-traditional), the "AI-Based Image Steganography System" results suggested a performance increase of approximately 25% accuracy over the use of traditional LSB-based detection methods and from a performance basis of 70% accuracy (which was found to be true) to that of a 95% accuracy level (which was found to be true) using CNN models.

Conclusion

In this paper, we present an Artificial Intelligence based picture Steganography Detection System which combines traditional statistical Feature Extraction with Neural Networks through Deep Learning. The Hybrid Model combines the features of images that are subject to Steganography techniques including Histogram Variance, Entropy, Noise Residuals, and the Pixel Correlation with Convolutional Neural Network (CNN) models to Find Steganography through various embedding techniques. The results show CNN based models perform with approximately 95% accuracy and exceed Traditional SVM and LSB models significantly. The hybrid model demonstrates that with enhanced Robustness and Generalization it is suitable for Cybersecurity applications in real-time and Digital Forensics.

Future studies may utilize various strategies to improve detection capabilities and flexibility including: Transformer-Based Architectures. This method uses Attention mechanisms to identify Long-Range Dependencies between images with increased sensitivity. Ensemble Learning. This method combines multiple models to enhance accuracy, stability, and resilience against Steganography methods across different domains. Cross-Domain Steganalysis. By expanding the framework to support other types of digital content including Audio and Video, Digital Security can be offered in its entirety to the Digital World.

Acknowledgment

I acknowledge the entire team of NGP-VSIT Conference, Conveners and all other members for organizing and motivating everyone to participate in the conference.

References

1. Alevizos, L. (2025). Automated cybersecurity compliance and threat response using AI, blockchain and smart contracts. International Journal of Information Technology, 17(2), 767–781.
2. B, S. R., B, M. R., & Belavagi, M. C. (2024). A Mellin transform based video steganography with improved resistance to deep learning steganalysis for next generation networks. MethodsX, 13.
3. B. Chen, Y. H. (2024). Deep video steganography using temporal-attention-based frame selection and spatial sparse adversarial attack. Journal of Visual Communication and Image Representation, 104, 104311.
4. Bourian, I., Hassine, L., & Chougdali, K. (2025). AI-Driven Security for Blockchain-Based Smart Contracts: A GAN-Assisted Deep Learning Approach to Malware Detection. Journal of Cybersecurity and Privacy, 5(3).
5. De Rosal Ignatius Moses Setiadi, S. R. (2024). Maximizing complex features to minimize the detectability of content-adaptive steganography. Multimedia Tools and Applications, 84, 23813–23831.
6. Firdaus, D. T., Croix, N. J., Ahmad, T., Mukanyiligira, D., & Sibomana, L. (2025). Steganographic model to conceal the secret data in audio files utilizing a fourfold paradigm. Journal of Safety Science and Resilience, 6(2), 138–149.
7. Hu, K., Wang, M., Ma, X., Chen, J., Wang, X., & Wang, X. (2024). Learning-based image steganography and watermarking: A survey. Expert Systems with Applications, 249.
8. Huang, Y., Lv, S., Tseng, K.-K., Tseng, P.-J., Xie, X., & Lin, R. F.-Y. (2023). Recent advances in artificial intelligence for video production system. Enterprise Information Systems, 17(11).
9. Jiaxuan Wu, Z. W. (2024). Generative Text Steganography with Large Language Model. Computation

and Language (cs.CL).

10. N. Satheesh, N. G. (2025). Advanced AI-driven emergency response systems for enhanced vehicle and human safety. Iran Journal of Computer Science, 8.
11. P. Bedi, A. S. (2024). Deep learning based active image steganalysis: a review. International Journal of System Assurance Engineering and Management, 15(3), 786–799.
12. Patwari, B. N. (2023). Image steganography based on difference of Gaussians edge detection. Multimedia Tools and Applications, 82, 43759–43779.
13. Pevny, T., F. T. (2010). Using High-Dimensional Image Models to Perform Highly Undetectable Steganography. Information Hiding, 6387, 161–177.
14. Płachta, M., Krzemien, M., Szczypiorski, K., & Janicki, A. (2022). Detection of Image Steganography Using Deep Learning and Ensemble Classifiers. Electronics, 11(10), 1565.
15. Rahman, S., Uddin, J., Zakarya, M., Hussain, H., Khan, A. A., & Ahmed, A. (2023). A Comprehensive Study of Digital Image Steganographic Techniques. IEEE Access, 11.
16. S. Kaur, S. S.-N. (2022). A Systematic Review of Computational Image Steganography Approaches. Archives of Computational Methods in Engineering, 29(7), 4775–4797.
17. Sahu, A. K., Swain, G., Sahu, M., & Hemalatha, J. (2021). Multi-directional block based PVD and modulus function image steganography to avoid FOBP and IEP. Journal of Information Security and Applications, 58.
18. Saputro, I. A., Purwidiantoro, M. H., Nugraha, F. S., Widiati, I. S., & Widiyanti, S. (2024). AI-Powered Steganographic Techniques: A Comparison of Traditional Methods and Modern Machine Learning Approaches. 2024 6th International Conference on Cybernetics and Intelligent System (ICORIS). IEEE.
19. Setiadi, D. R. (2022). Improved payload capacity in LSB image steganography uses dilated hybrid edge detection. Journal of King Saud University – Computer and Information Sciences, 34(2), 104–114.
20. Setiadi, D. R., Ghosal, S. K., & Sahu, A. K. (2025). AI-Powered Steganography: Advances in Image, Linguistic, and 3D Mesh Data Hiding – A Survey. Journal of Future Artificial Intelligence and Technologies, 2(1).
21. Setiadi, D. R., Rustad, S., Andono, P. N., & Shidik, G. F. (2023). Digital image steganography survey and investigation (goal, assessment, method, development, and dataset). Signal Processing, 206.
22. Song, B., Wei, P., Wu, S., Lin, Y., & Zhou, W. (2024). A survey on Deep-Learning-based image steganography. Expert Systems with Applications, 254.
23. T. Muralidharan, A. C. (2022). The infinite race between steganography and steganalysis in images. Signal Processing, 201, 108711.
24. Yousef Sanjalawe, S. A.-E. (2025). A deep learning-driven multi-layered steganographic approach for enhanced data security. Scientific Reports, 15(4761).