

Detecting Hidden Data in Images Using Deep Learning

Rusul Fadhil Yaser^{1*}

1. Ministry of Education

*Correspondence: rusulfadhil77@gmail.com

Article information:

Manuscript received: 04 July 2026; **Accepted:** 10 August 2026; **Published:** 28 September 2026

Abstract: Data cloaking and digital concealment techniques pose a growing security challenge in modern communication environments, where the spatial dimension of images is exploited to transmit secret messages or malicious data. In contrast, steganography techniques are designed to detect these hidden modifications. This paper proposes a model based on deep learning, specifically convolutional neural networks (CNNs), supported by Spatial Rich Models (SRM) filters, to detect LSB data embedding. We review the latest studies on the development of cloaking and detection techniques and then present our experimental model, which was evaluated on a locally constructed dataset. The results demonstrate the model's ability to extract accurate statistical features and achieve a 91% recall rate for cloaked images. Data security technologies require the development of faster and more automated systems. This necessitates the development of a fast and automated cloaking analysis system that operates differently from traditional manual verification methods. Deep learning-based data hiding analysis systems employ a convoluted neural network (CNN) that classifies images into easy-to-lead and hidden images. Easy-to-lead images do not contain hidden data, while hidden images do. The methodology relies on three main phases: Data preparation: Data hiding analysis is performed using digitally combined random text images modified using Least Significant Bit (LSB) technology. The generated images form a balanced dataset appropriately classified for model training. Model building and training: The model is built and trained in layers. Each layer is responsible for amplifying the minor statistical variations resulting from the data hiding analysis process. The model is then trained on the dataset, attempting to associate underlying features that may not be visible to the naked eye. Evaluation and analysis: The model is evaluated based on criteria of accuracy, precision, and recall. The model has been shown to perform excellently in image classification, as a result of the weighted criteria. This also demonstrates the impact of deep learning in the practice of data hiding. The results of this research confirm that artificial intelligence techniques improve data security systems, and also confirm that deep learning models can be a powerful and effective tool for identifying taught activities in media content.

Keywords: CNN, LSB, Data Encryption, Artificial Intelligence.

1. Introduction

In today's digital landscape, ensuring the secure transmission of data across public networks for confidential communications is paramount. Secure data transmission over public networks is essential to protect sensitive information from unauthorized access and manipulation[1]. Digital signatures are primarily used to authenticate digital assets, but they are also employed in licensing and identity verification, digital forensics, and data protection in smart cities. Digital signatures are not limited to images and documents; they also encompass real-time audio and video, languages, and the protection of chips and electronic devices[2]. Data escaping is the process of concealing a type of information within another medium[3]. In the field of image escaping, the original image before the inclusion of confidential information is called the original image, and the image after inclusion is called the escaping image[4]. In today's digital age, ensuring the security and freedom of data transmission over public networks has become critical. Data hiding techniques offer means to protect intellectual property through watermarking or to secure communications through

steganography. However, the malicious use of these techniques to spread malware necessitates the development of highly accurate automated steganography systems.

Hiding techniques have evolved from simple spatial-scale algorithms (such as LSB) to include the use of deep neural networks, leading to an ongoing arms race between "hiding" and "detection" techniques. This paper aims to review the latest research in this field and present a practical design for a CNN model capable of detecting spatial hiding[5].

1.1 Previous Studies

The field of data steganography and analysis has undergone a true revolution thanks to deep learning. Previous studies can be divided into the following sections based on the attached sources:

1.2 Advances in Steganography

Research has moved from simply hiding text to hiding entire images within other images. Baluja introduced a pioneering framework for hiding full-color images (image-in-image) using autoencoders, demonstrating that the hidden data is distributed across color channels and spatial dimensions.

To overcome the limitations of cover distortion, "coverless steganography" emerged. Liu et al. proposed an algorithm based on extracting features (DenseNet) and converting them into hash sequences, thus preventing their detection by traditional analysis tools.

The field has also expanded to include "cross-modal cloaking," where Han et al. used implicit neural representations (INRs) to hide complex data (video, audio, 3D shapes) within flat images. In another innovative approach, Chen et al. proposed hiding images within "deep probabilistic models" such as SinGAN, where the hidden image is generated using a specific extraction key, thus enhancing security. Techniques have also been extended to include hiding data within neural compressed latent representations to accelerate the process and increase secrecy[6][7][8].

1.3 Steganalysis and Detection Using Deep Learning

To counter sophisticated cloaking techniques, the need for deep learning-based detection models has emerged. Yaser and Jawad pointed to the necessity of moving from manual inspection to automated systems based on CNNs to classify images into Docile and Stego.

To enhance the ability of networks to detect hiding noise, Lwowski et al. proposed combining Spatial Rich Models (SRM) with Principal Component Analysis (PCA) and DNNs. In a related vein, Agarwal and Jung presented a CNN guided by 31 filters (30 high-pass filters and one neutral filter), using a Clipped ReLU function and a SVM classifier to improve accuracy against algorithms such as HILL and S-UNIWARD. Delenia et al. also presented a novel model integrating Mix Pooling and multi-scale Inception, outperforming previous models such as Yedroudj-Net[9][10].

1.4 Active Steganalysis and Sanitization

Some research has moved beyond detection to "sanitization" (removing hidden data). Jung et al. proposed a PixelSteganalysis system that scans for hidden information while preserving visual image quality by restoring the original statistical distribution of pixels using PixelCNN++ networks. Robinette et al., on the other hand, used diffusion models within a framework called DM-SUDS to sanitize images from global and dependent cloaking (UDH & DDH), achieving a significant improvement in the recovered image quality[11][12].

1.5 Advanced Challenges: Adversary Attacks and Network Cloaking

Some studies have demonstrated the possibility of deceiving analysis systems. Yang et al. proposed the MRAG framework to cloak JPEG images by generating adversarial perturbations that combine local (CNN) and global (Transformers) features to mislead steganalysis detectors. Conversely, in another security dimension, Tang et al. proposed a technique for detecting hidden data within the neural network weights themselves by relying

on the symmetric distribution of the output histogram, without needing to access the network's internal structure[13][14][15].

2. Materials and Methods

Based on recommendations from the literature, specifically the use of SRM filters as a preprocessing step, a custom CNN model was built and trained for LSB-based caching detection.

2.1 Dataset Generation

A standard flower image set was used, converted to grayscale, and resized to 256x256 times. The hidden images were generated using the LSB embedding function at a payload rate of 0.2 bpp. The data was then split into 80% for training and 20% for verification, with a batch size of 16.



2.2 Proposed CNN Architecture

- Preprocessing Layer (SRM Filter): A basic 3x3 pixel SRM filter is included as a non-trainable layer ('trainable=False') to extract high frequencies (masking noise).
- Feature Extraction Layers: Traditional Convolutional Layers (Conv2D) followed by Separable Convolutional Layers (SeparableConv2D) are used to reduce the computational load.
- Batch Normalization and Clipped ReLU activation are implemented to enhance control over statistical feature extraction.
- Pooling and Classification Layers: Global Average Pooling is used to prevent loss of statistical information, and a 30% Dropout technique is used to prevent overfitting, followed by a Dense layer with a Sigmoid function for binary classification.

3. Results and Discussion

The model was trained over 10 epochs using the Adam optimizer and a 1e-3 learning rate.

3.1 Metrics

By evaluating the model on validation data, the following figures were obtained:

Training Accuracy: 0.5000

Validation Accuracy: 0.4450

Training Loss: 0.7030

Validation Loss: 0.7001

Precision: 0.4341

Recall Rate: 0.9080



3.2. Classification Report

(Class)	(Precision)	(Recall)	F1-Score	(Support)
Cover (0)	0.56	0.09	0.15	113
Stego (1)	0.43	0.91	0.59	87
(Accuracy)	-	-	0.45	200

The results reveal a significant superiority in the recall rate for hidden images (0.91), indicating that the SRM filter and convolutional layers were highly successful in detecting minor distortions and identifying potentially hidden images.

However, the model exhibited a decrease in overall accuracy and positive resolution for intact images. This is attributed (as described in the literature) to the difficulty in separating data with a low payload (0.2 bpp) using a simple network architecture, leading to a high false alarm rate (predicting intact images as hidden). These results demonstrate the success of the fundamental principle (modulation detection) and recommend increasing network depth and training data size in future work.

4. Conclusion

This paper has reviewed the rapid development in the fields of steganography and steganalysis, from classical models to deep learning and diffusion models. Building upon previous studies, we constructed a CNN model supported by SRM filters for spatial caching detection. The results demonstrated the effectiveness of the approach in achieving high sensitivity for detecting modifications (Recall=91%). As future work, we suggest leveraging complex probabilistic models and expanding the range of extracted features to reduce false alarms and improve overall classification accuracy.

REFERENCES

- [1] R. F. Yaser and S. Jawad, "Detecting hidden data in images using deep learning," 2025.
- [2] X. Tang, Z. Wang, and X. Zhang, "Steganalysis of neural networks based on symmetric histogram distribution," *Symmetry*, vol. 15, no. 5, p. 1079, 2023.
- [3] S. Baluja, "Hiding images within images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1685–1697, 2020.
- [4] S. Agarwal and K.-H. Jung, "Identification of content-adaptive image steganography using convolutional neural network guided by high-pass kernel," *Applied Sciences*, vol. 12, no. 22, p. 11869, 2022.
- [5] P. K. Robinette et al., "SUDS: Sanitizing universal and dependent steganography," *arXiv preprint arXiv:2310.06951*, 2023.

- [6] E. Delenia et al., "Detecting hidden data in images using convolutional neural networks," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 5, pp. 1227–1235, 2024.
- [7] G. Han et al., "Deep cross-modal steganography using neural representations," *arXiv preprint arXiv:2307.08671*, 2023.
- [8] H. Chen et al., "Hiding images in deep probabilistic models," in *Proc. 36th Conf. Neural Information Processing Systems (NeurIPS)*, 2022.
- [9] J. Yang et al., "Towards imperceptible JPEG image hiding: Multi-range representations-driven adversarial stego generation," *arXiv preprint arXiv:2507.08343*, 2025.
- [10] M. Yedroudj, A. Kouider, and M. Chaumont, "Yedroudj-Net: An efficient CNN for spatial image steganalysis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2092–2096.
- [11] Q. Liu et al., "Coverless image steganography based on DenseNet feature mapping," *EURASIP Journal on Image and Video Processing*, vol. 2020, 2020.
- [12] C.-H. Huang and J.-L. Wu, "Image data hiding in neural compressed latent representations," *arXiv preprint arXiv:2310.00568*, 2023.
- [13] D. Jung et al., "PixelSteganalysis: Pixel-wise hidden information removal with low visual degradation," *IEEE Transactions on Dependable and Secure Computing*, 2021.
- [14] O. Byrnes et al., "Data hiding with deep learning," *arXiv preprint arXiv:2107.09287*, 2021.
- [15] J. Lwowski, I. A. Corley, and J. Hoffman, "Neural steganalysis with spatial rich models for image steganography detection," 2020.