StegDetector: Automated Steganalysis Report

Mihai Bontea

December 7, 2025

Abstract

Steganalysis is the science of detecting the presence of hidden messages in media. There are multiple steganalysis techniques, which usually look for the artifacts created by the message hiding process. This tool combines metadata analysis, statistical analysis and neural networks, for the purpose of identifying whether or not an image contains embedded messages.

Input Image

Filename: C:/Users/Mihai/OneDrive/Desktop/example.jpg

File Metadata Anomalies

Sometimes the process of steganography leaves other, more noticeable traces besides the noise in the image itself. The image has been scanned for such traces: structural, metadata, and container anomalies that might indicate tampering or hidden data.

Examples include unusual file size, invalid file headers, missing or strange EXIF metadata, and corrupted or suspicious PNG chunks. While these are not a direct proof of steganography, they are heuristic red flags.

- No EXIF metadata found (possible stripping or modification), but low priority.

RS Analysis

RS Analysis is a statistical steganalysis technique designed to detect data hidden by LSB (least-significant bit) embedding in images. It is based on the fact that in a non-stego image, there are certain regular statistical relationships between the different bit-planes and among neighboring pixels. By embedding a message (modifying LSBs), these correlations are disturbed, and RS Analysis can detect these disturbances. It is effective when the cover image is a "natural image" (e.g. photos with natural textures and smooth areas) because those have the spatial correlations that RS exploits. In addition, the amount of data hidden needs to be sufficiently high, since small payloads may produce a disturbance that is too subtle to reliably detect.

Estimated Stego Confidence: 0.6065

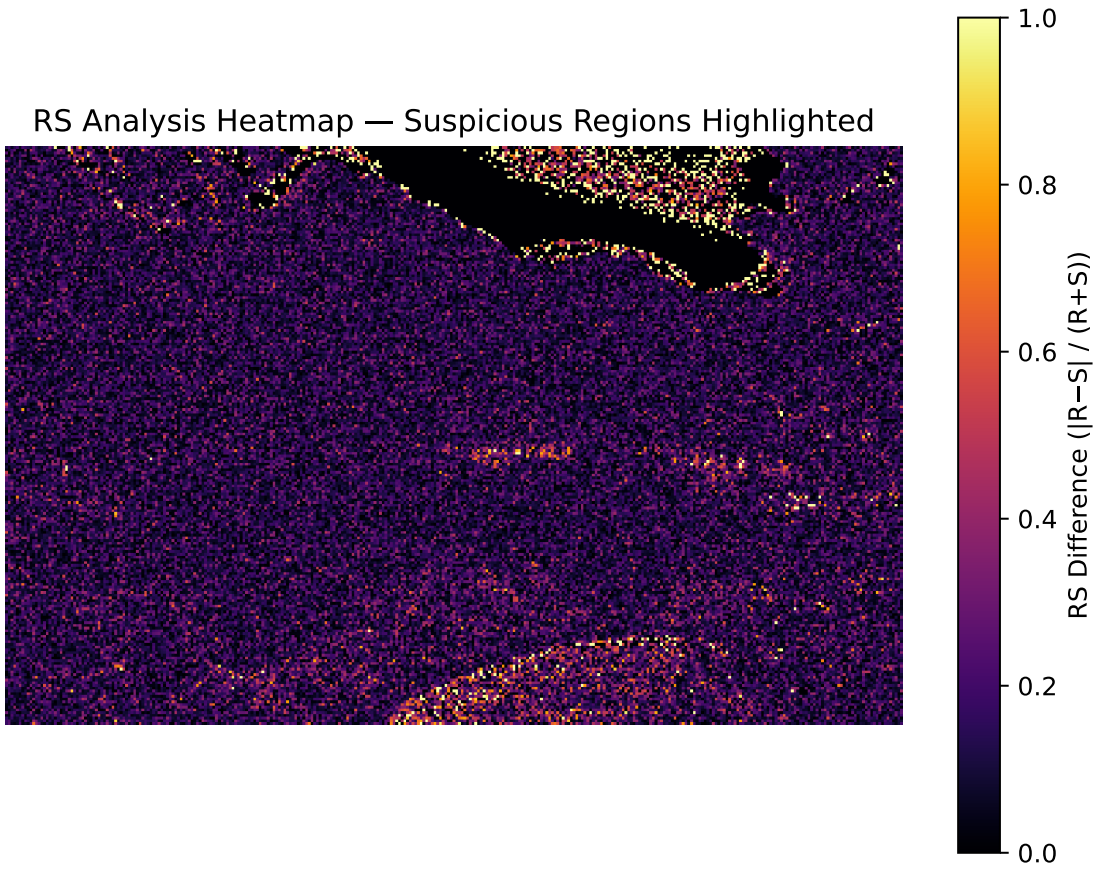

Figure 1: RS Analysis Heatmap

Figure 2: RS Overlay Visualization

High-Pass Residual Analysis

High-pass residual analysis refers to a class of statistical steganalysis techniques that apply high-pass filters(or residual filters) to an image, in order to suppress the "normal content" (smooth, large-scale textures) and amplify the small perturbations or "noise" caused by steganography use. This results in a residual image: a two-dimensional array showingdifferences, edges, and high-frequency detail. Hidden data often shows up in these residuals more clearly than in the original image. For this reason, these residuals are used by many modern steganalysis methods.

Estimated Stego Confidence: 0.5900

Figure 3: High-Pass Residual Steganalysis

CNN-Based Steganography Detection

A convolutional neural network (CNN) was used to provide an additional learned-based estimate of steganographic embedding probability. The model has been trained on the BOSSBase(Break Our Steganographic System Base), which is one of the most important and widely used benchmark datasets in steganography and steganalysis research. It consists of 10,000 grayscale images.

Estimated Stego Confidence: 0.4939

Dataset Preparation

Training data is organized into two classes: *cover* (non-stego) and *stego*.

Network Architecture

The CNN architecture consists of three convolutional blocks followed by a fully connected classifier:

- Convolutional layer with 32 filters, 3×3 kernel, ReLU activation
- Batch normalization and 2×2 max pooling
- Convolutional layer with 64 filters, 3×3 kernel, ReLU activation
- Batch normalization and 2×2 max pooling

- Convolutional layer with 128 filters, 3×3 kernel, ReLU activation
- Batch normalization and global average pooling
- Fully connected layer with 64 units and ReLU activation
- Dropout regularization with rate 0.4
- Final sigmoid output neuron for binary classification

The model is optimized using the Adam optimizer with binary cross-entropy as the loss function and accuracy as the primary training metric.

Training Procedure

Training is performed using mini-batch stochastic gradient descent with a batch size of 32. Data augmentation is applied during training in the form of random horizontal flips and small rotations (up to 10 degrees) in order to improve generalization and reduce overfitting. Validation performance is monitored on a held-out 20% subset of the training data.

Conclusion: Detection Result

The final estimated stego confidence score is obtained from a weighted mean of the methods previously presented.

Estimated Stego Confidence: 0.5549 – *Likely* to be a stego image.