UDC 004.056.55

A Review of Modern Methods for Steganalysis and Localization of Embedded Data in Digital Images

Pavlo Yatsura[1], Dmytro Progonov[1]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Prospect Beresteiskyi 37, Kyiv, 03056, Ukraine*

Abstract

The article provides a systematic review of modern steganalysis methods for digital images based on artificial neural networks. The primary stages of development of advanced cover-image models, from widely used artificial neural networks to contemporary hybrid models, are considered. Advantages and limitations of various types of neural networks for constructing stegodetectors for digital images are investigated. Based on comparative analysis of steganalysis accuracy, it is established that the use of advanced artificial neural networks achieves a detection accuracy of steganograms exceeding 90%, even at low embedding rates (less than 20%). Additionally, applying complex methods of processing both examined images, and feature vectors in multidimensional spaces with studied neural networks allows reducing the computational complexity of configuring stegodetectors without significant losses in stego images detection accuracy.

Keywords: steganography, steganalysis, artificial neural networks, digital images, cybersecurity

Introduction

Given the rapid growth in volumes of digital data circulating, processed, and stored in information and communication systems, there is an increasing need for effective protection of sensitive information (SI). Special attention of developers of restricted access information protection systems is devoted to early detection of covert communication channels, which are widely used by intruders for unauthorized SI transmission. In particular, this relates to counteraction to military, industrial, political espionage, and prevention of terrorist attempts to name a few.

The novel adaptive steganographic methods, such as WOW, HILL, or S-UNIWARD [24, 23, 1], have become widely used in order to help increasing the robustness of the resulting steganograms against widespread methods of steganalysis of digital cover files, especially digital images (DI). This is due to the embedding of data (stegodata) into cover images (CI) while considering the values of their local characteristics, such as statistical and spectral parameters of each block of CI partitioning [19].

A significant number of modern steganalysis methods have been proposed to reveal the created stego images. Among these statistical methods based on analyzing changes in statistical parameters of the examined image, caused by message embedding, are most widely used [25]. This increases the probability of detecting messages embedded with classical steganographic methods (such as the LSB group). However, the effectiveness of these statistical detectors is significantly reduced in the case of data embedding in adaptive way and low embedding levels (payload) of steganographic data into the CI (less than 10%).

One of the promising directions in developing novel steganalysis methods is the application of artificial neural networks (ANN) [10]. Significant advantage of this approach is the capability of neural networks to detect and generalize weak (insignificant) deviations from CI in coefficient values or pixel brightness values of DI [10, 9]. This allows detection and analysis of unmasking features of stego images without usage of pre-configured statistical models of CI. This feature of ANN-based stegodetectors (SD) is of particularly interest for cases of revealing of stego images formed according to prior unknown embedding methods.

Comparative analysis results [10] of such SD confirmed the expediency of transitioning from traditional statistical detectors to machine learning based methods, which do not require forming a large set of pixel brightness values that are set prior the analysis. Furthermore, applying of novel approaches for designing of ANN allows for additional improve the accuracy of such SD [10, 9].

To overcome the limitations of SD based on neural networks, a number of approaches have been proposed, such as multilevel processing of CI [45], global averaging of statistical parameters of DI [21], focusing the ANN "attention" on regions of the analyzed image where small brightness or color variations are visually less noticeable [22], and employing reinforcement learning methods [13]. However, open-source literature does not include information regarding achievable accuracy of SD based on the mentioned approaches. This complicates the selection of an appropriate type of ANN for design detectors due to the necessity to examine several ANN architectures. It leads to increased duration in SD construction that may be inapplicable for real usage.

Thus, we may conclude that comparative analysis of modern approaches to constructing ANN-based SD is topical and important task today. In particular, it is of interest to investigate the influence of modern ANN characteristics (such as ResNet blocks, attention mechanisms, reinforcement learning, and methods of transforming feature vectors in multidimensional spaces) on detection accuracy of stego images, formed according to adaptive steganographic methods. The paper explores the achievable accuracy of ANN-based SD when embedding steganographic data in both spatial and frequency domains of the CI.

The structure of the paper is as follows: the review of used terms and abbreviation in the domain of DI steganalysis is presented in the Section 1. The Section 2 is devoted to a literature review for the current overall state of the ANN based SD. The Section 3 follows with the review of the most popular and latest SD models. The Section 4 is dedicated to a comparative analysis of the reviewed models. Next, the Section 5 contains discussions based on finding of the previous section. The Section 6 presents conclusions of the paper.

1. Preliminaries

1.1. Digital images steganography

Steganography is a branch of science that studies models, methods, and means of embedding messages (stegodata) into physical and digital information carriers while maintaining minimal visual changes. One of the most common types of cover files is DI, primarily due to their widespread usage in global and local information systems, as well as the availability of numerous processing methods. This significantly simplifies the masking of minor pixel brightness changes caused by embedding stegodata into the CI [17, 40].

In the field of digital steganography, numerous methods for embedding messages into CI have been proposed. These methods are based on embedding stegobits through modifications of pixel brightness values of the CI or changes in the coefficients obtained from image transformations (e.g., when using a two-dimensional discrete cosine transform, 2D DCT). Depending on the specifics of the stegodata embedding process into DI, particularly the magnitude of brightness changes in the pixels of the CI, known steganographic methods can be divided into two groups [17]:

- non-adaptive methods, where changes in pixel brightness occur with approximately equal probability regardless of their position in the DI, typically resulting in a uniform distribution of changes throughout the CI. Examples include the group of LSB methods, embedding stegobits using pseudo-random sequences, etc.
- adaptive methods, which account for the specifics of the distribution of brightness values within the CI to minimize changes in its statistical characteristics during the creation of steganograms. This allows for masking message embedding by altering pixel brightness in textured regions of the image. Examples include WOW, HILL, S-UNIWARD methods, and others [16, 26, 36].

To minimize distortions of the statistical and spectral parameters of the CI when forming steganograms, additional processing stages of stegodata can be used, such as message encoding using syndrome-trellis codes. This allows reducing the message bit-size while ensuring robustness against possible changes in individual

bit values, thus decreasing the number of CI pixels used to embed the message [16, 2].

To quantitatively assess the degree of CI changes caused by message embedding, the embedding rate indicator ($\Delta\alpha$) is used. This indicator reflects the number of bits per pixel (Bits Per Pixel, BPP) that can be hidden in a single pixel of the CI at a given level of changes to statistical and spectral parameters. It is generally accepted in the literature that message embedding at values $\Delta\alpha \leq 20\%$ provides a compromise between the message size and the robustness of the resulting steganograms to detection methods. With increasing embedding rates ($\Delta\alpha \geq 40\%$), a steganographer can embed more stegobits into the CI at the expense of reduced resistance to steganalysis methods.

Depending on the approach used for embedding stegobits into the CI, known steganographic methods can be divided into two groups: embedding in the spatial domain of the CI and embedding in the transform domain. When embedding messages in the spatial domain, individual stegobits are hidden by changing the brightness values of a selected group of pixels from CI [32, 15]. It should be noted that the majority of methods for detecting steganograms embedded in the spatial domain are based on detecting weak changes in the degree of correlation between the brightness values of neighboring pixels. Detection is carried out by analyzing deviations in statistical parameters of pixel distributions caused by embedding hidden data. For instance, the Subtractive Pixel Adjacency Matrix (SPAM) method is based on assessing changes in the distribution of brightness differences between pairs of neighboring pixels. SD are configured by comparing these characteristics with corresponding parameters obtained from processing original (unmodified) CI [43, 39].

In the case of message embedding in the transform domain of DI, the transform coefficients of the CI in the chosen transform basis undergo changes. An example of methods based on this approach is JPEG steganography. These methods involve adaptive or nonadaptive modifications of block-based DCT coefficients to minimize visual distortions of the CI and ensure robustness of the created steganograms against possible transformations, especially lossy compression [44, 41, 53].

1.2. Digital images steganalysis

Steganalysis methods are used to detect formed steganograms. Currently, a wide range of DI steganalysis methods has been proposed, including signature-based detection, statistical and spectral analysis methods [17, 31, 3]. One of the widespread steganalysis directions involves building SD based on complex statistical models ("rich models"). These methods are based on combining several "simple" statistical models into a single (complex) model to enhance detection accuracy of subtle distortions introduced by steganographic embedding. The selection of these individual "simple" models is aimed at capturing as many parameters of the processed DI as possible. For instance, intra-block differences are computed as deviations between neighboring pixels (or their coefficients) within one DI partition block (most commonly 8 × 8), while inter-block differences encompass similar deviations between adjacent DI blocks. Resulting statistical parameters (both intra- and inter-block) can be combined and used to train a classifier that identifies differences caused by steganographic embedding [18, 34, 27]. To further increase detection accuracy, calibration methods for DIs are widely applied [48]. Calibration involves identifying differences between processing results of original CIs and steganograms using standard image processing techniques [33, 34, 41].

The continuous improvement of steganographic methods, particularly involving generative adversarial networks (GAN) [34, 47], necessitates corresponding enhancements in SDs to ensure high detection accuracy (exceeding 95%). One modern direction in the development of SDs involves using an ensemble of classifiers capable of incorporating multiple feature types and adapting to changes in stego-bit distributions within CIs (e.g., via domain adaptation or ensemble processing) [34, 28, 36, 54].

As a preventive measure, active steganalysis methods aimed at damaging or destroying (destructing) hidden stegodata may also be employed. One contemporary direction for improving stegodata destruction methods is determining pixel positions in the CI used for embedding specific stego-bits. Utilizing this information enables targeted modifications of processed DI, effectively masking interference in the steganographic transmission channel [41].

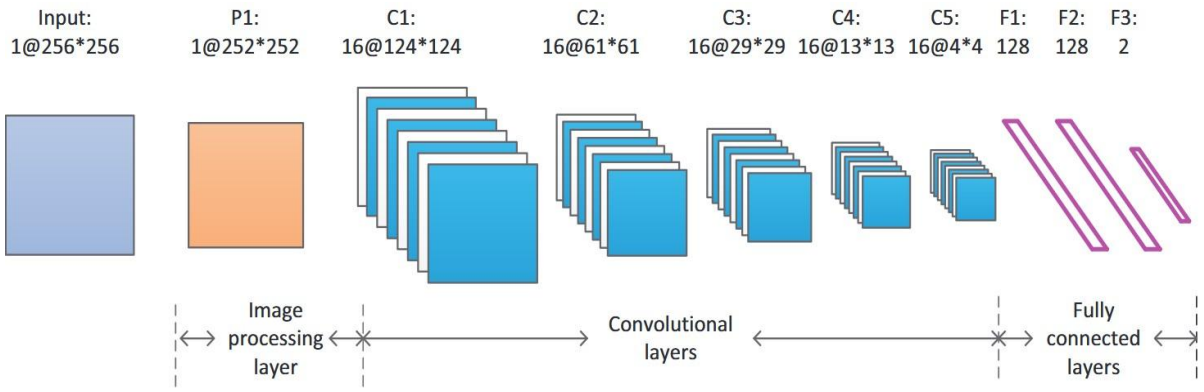

Figure 1: Structural architecture of the QianNet artificial neural network according to paper [11]

2. Literature Review

A significant portion of existing steganalysis methods is based on constructing SD by identifying differences between statistical, spectral, and structural parameters of CI and steganograms [17, 34, 27]. As an example, modern steganalysis methods utilizing complex statistical image models can be mentioned, such as Markov models and Gabor-filter residuals [43, 18, 33]. Due to the large number of image features used in various statistical models (which can reach approximately 40,000), ensemble classifiers are employed to configure SDs, achieving high steganogram detection accuracy (above 90%) even at low embedding rates ($\Delta\alpha \in [10\%;20\%]$) [16, 36, 2]. When embedding stegodata in the transform domain, a common approach to detecting formed steganograms involves statistical analysis of transform coefficients and calibration methods of the processed DI [7, 33, 41]. This allows for more precise detection of weak anomalous shifts in the transform coefficients of DI, particularly violations of typical statistics [26, 39, 44].

Given the difficulty of reliably detecting formed steganograms and the proliferation of steganographic methods based on generative models (e.g., GAN, CycleGAN), active steganalysis methods are increasingly being employed as a preventive measure.

The need to counteract these methods motivates the development of new steganalysis approaches that not only detect but also remove or distort hidden data by combining calibration procedures with ensemble classifiers [7, 26, 30]. One of the contemporary approaches to constructing these methods is the application of ANN [7]. However, the information available in open sources about such ANNs is limited, complicating their comparison.

Therefore, the purpose of this paper is to provide an analytical review of modern ANN-based SDs to identify the advantages and limitations of their practical application. The study focuses on the formation of steganograms using adaptive embedding methods (namely WOW, HILL, S-UNIWARD, etc.) and variations in embedding rates $\Delta\alpha$ across a wide range.

Subsequent sections review common ANN architectures (e.g., QianNet, XuNet, YeNet, YedroudjNet, and others), present results from comparative analyses of steganogram detection accuracy using these models, and discuss the advantages and limitations of their practical applications. Particular attention is given to the impact of embedding rate ($\Delta\alpha$) on detection accuracy, computational resource requirements, and the ability of various models to detect data hidden using adaptive methods in different embedding domains.

3. Steganalysis Methods Based on Convolutional Neural Networks

This section provides a comparative analysis of ANN construction features for solving steganogram detection tasks. A significant advantage of using ANN to build SD, compared to traditional statistical SD, is the automatic identification of revealing features during SD training. This eliminates the need for prolonged manual feature identification by the analyst, thus accelerating the SD training process.

The following subsections examine the most commonly used ANN architectures employed in contemporary SD development. The characteristics of each neural network type are analyzed, along with the advantages and limitations of their practical application in

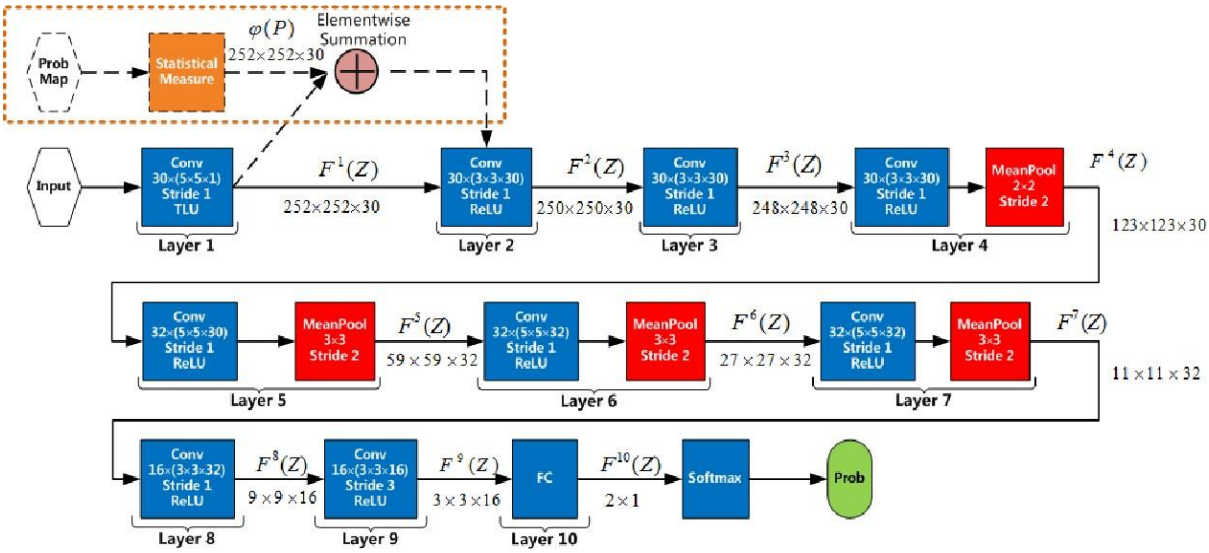

Figure 2: Structural architecture of the YeNet artificial neural network according to paper [29]

detecting steganograms and localizing stegobit embedding positions within CI pixels.

3.1. QianNet Model

The QianNet model, proposed in [11], was among the first ANN adapted for solving image steganalysis tasks. Specifically, this model enables all stages of steganogram detection (including extraction of statistical parameters from the processed DI and subsequent classification) without prior manual identification of the features revealing hidden messages.

The structural scheme (architecture) of the QianNet model is shown on figure 1.

The architecture of the QianNet network comprises five convolutional layers (fig. 1). A distinctive feature of these layers is the use of a special activation function type, specifically based on the Gaussian function:

$$f(x) = exp(-\alpha x^2), \quad (1)$$

where α is frequently set at 1 to moderately "compress" large values and enhance small deviations (brightness changes of 1−2 units in the range from 0 to 255). The use of activation function (1) amplifies minor brightness variations in CI pixels arising from stegobit embedding.

Data obtained at each convolutional layer's output in the QianNet network is processed using average pooling with a sliding window (kernel) size of 2 × 2 or 3 × 3 pixels.

At the last convolutional layer's output, data is forwarded to the classification block (1). This block consists of several fully connected layers, whose outputs utilize the softmax function:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}, \quad (2)$$

where zi represents outputs of the i-th neuron. Values of the softmax function (2) correspond to the probabilities of classifying the analyzed DI as either a CI or a steganogram.

The increased detection accuracy of SDs based on QianNet, combined with relatively low computational complexity, spurred further research in this direction. Researchers particularly focused on adapting this model for localizing stegobit embedding positions by modifying preprocessing layers.

3.2. XuNet Model

The XuNet artificial neural network [20] was proposed by Xu, Shi, and collaborators. A distinctive feature of this network is the use of an ABS layer of artificial neurons at the first stage of processing the investigated DI, in combination with convolutional layers. The ABS layer is applied to improve the accuracy of detecting deviations in the statistics of pixel values (or transform coefficients) by employing the operation $y = |X|$ on the feature vectors obtained at the outputs of the network's convolutional layers.

Unlike the QianNet network, the data processing at the output of the fully connected

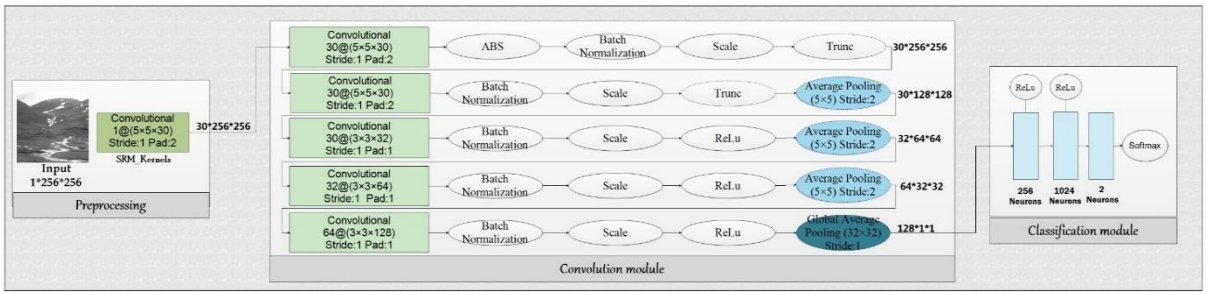

Figure 3: Structural architecture of the YedroudjNet artificial neural network according to paper [37]

layers in XuNet is performed using the *tanh* activation function in the first two layers of artificial neurons, and the ReLU activation function [38] in the subsequent layers. To reduce the variability of the values of the elements in the feature vectors at the output of each convolutional layer and, accordingly, to increase the robustness of the model tuning procedure to changes in the distribution of statistical parameters of the DI, the batch normalization (BN) procedure is applied during the configuration of the XuNet model:

$$BN(\hat{x}) = \gamma\hat{x} + \beta, \hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}},$$

where μ_B and σ_B^2 are the mean and variance of the values of the feature vector elements for the processed batch of feature vectors, and γ and β are parameters used to normalize the output vectors at each layer of the network. This ensures robustness to potential perturbations of the network's parameters during the configuration of the SD. The parameter ε is a small constant (on the order of 10^{-5} or 10^{-8}) introduced to ensure numerical stability during the computation of the normalized values.

3.3. YeNet Model

A distinctive feature of the YeNet network [29] is the use of a group of filters proposed for the statistical SRM model, instead of a single high-pass filter in the input layers of the network, which is characteristic of earlier ANN-based SD such as QianNet. This approach made it possible to improve the accuracy of detecting subtle changes in pixel intensities of the CI caused by message embedding by approximately 5–7% compared to the QianNet network, assuming the use of S-UNIWARD at $\Delta_\alpha \approx 20\%$ [49].

The configuration of the YeNet network Is performed in several stages. In the first step, the parameters of the input convolutional layers are initialized by employing a group of 30 SRM filters as convolution kernels of size 5 × 5 pixels. In the second step, the output of each convolutional layer is processed using the *TLU* activation function:

$$TLU(x) = \begin{cases} T, & x \geq T, \\ x, & -T < x < T, \\ -T, & x \leq -T \end{cases}$$

where T is a threshold value in the truncated linear unit (TLU) function, which limits the output amplitude and typically equals 3. In total, the YeNet network consists of ten convolutional layers (fig. 2), which employ either the *ReLU* or *TLU* activation functions. The output of the last convolutional layer is passed to fully connected artificial neural layers for assigning the processed image to either the cover or stego class.

3.4. YedroudjNet Model

In the study [37], the authors proposed combining the previously discussed neural networks XuNet [20] and YeNet [29] in order to merge their advantages. This resulted in the construction of the artificial neural network YedroudjNet, whose architectural diagram is shown in fig. 3.

The first two convolutional layers of the network use the *TLU* activation function (fig. 3), while the next three layers apply the *ReLU* function. To reduce the dimensionality of the feature vectors at the output of intermediate layers, the YedroudjNet network employs an averaging procedure. At the output, classification of the analyzed image is performed using two fully connected layers of artificial neurons. The use of a double fully connected structure enables a reduction in the false decision rate by

approximately 0.5–1% compared to the single-layer design used in the YeNet model.

3.5. ZhuNet Model

The ZhuNet network [14] is a further development of the previously discussed neural networks (XuNet, YeNet, and YedroudjNet). In designing this architecture, the authors proposed the use of not only a group of SRM filters within the convolutional layers but also a specialized convolutional procedure known as separable convolution, to further enhance the detection of subtle changes in cover image parameters caused by steganogram embedding [8].

A key feature of separable convolution is the use of convolution kernels of size $3 \times 3 \times B$ (where B is the number of images in the input batch), followed by averaging the resulting outputs using a sliding window of size $1 \times 1 \times B$ pixels. This approach enables more efficient extraction of both spatial and inter-channel features, while maintaining fixed computational complexity.

To further improve the accuracy of SD based on the ZhuNet model, the authors proposed using a feature averaging method based on the pyramid decomposition of the CI [42]. This method involves multi-level partitioning of the extracted features into groups of elements at different scales, followed by averaging the values within each cell. As a result, a final matrix is formed that captures characteristics of the CI across multiple levels of detail simultaneously and adapts to variations in its size or aspect ratio. This contributes to improved robustness of the network to input image scale or proportion changes, thereby enhancing the overall detection accuracy of steganograms.

3.6. SRNet Model

The SRNet (Spatial Residual Network) architecture [6] was proposed as one of the first neural network models capable of detecting changes in CI parameters caused by message embedding in both the spatial and frequency domains. A key feature of this network is the use of residual blocks, which allow the depth of the network to be increased without encountering the adverse effects of gradient vanishing.

When applying the SRNet model to an image, the following sequence of transformations is performed:

$$F^{(l)} = \sigma\left(W^{(l)} * F^{(l-1)} + b^{(l)}\right),$$

where $l = 1,\ldots,L$ is the layer index, $W^{(l)}$ is the convolution kernel, $b^{(l)}$ is the bias vector, $\sigma(\cdot)$ is the activation function (e.g., ReLU), and * denotes the convolution operation.

The distinguishing feature of residual blocks lies in the use of skip connections between layers k and $k + 1$ [51]:

$$F^{(k+1)} = \sigma\left(W^{(k+1)} * F^{(k)} + b^{(k+1)}\right) + F^{(k)}.$$

This structure enables improved gradient flow during the training of deep networks with a large number of layers (25–30) [18]. In the SRNet model, the depth can reach up to 25–30 layers (incorporating both *T2* and *T3* block types), which enhances its ability to detect even weak steganographic signals at low payload rates (e.g., $\Delta_\alpha \approx 20\%$).

3.7. MRS-Net Model

The MRS-Net (Multi-Resolution Steganalysis Network) was proposed in [45] to improve the detection accuracy of stego images at low embedding rates (Δ_α). To address the issue of diminishing feature values in consecutive convolutional layers, the authors suggested using parallel processing branches with convolutional kernels of varying sizes.

At the first stage of MRS-Net operation, the input image X is processed by a set of SRM filters [45]:

$$X_{HPF}(i,j) = \sum_{p,q} h_{p,q}\, X(i+p,\, j+q),$$

where $h_{p,q}$ are the elements of one of the 30 SRM kernels.

The resulting features X_{HPF} are then distributed among m parallel branches (subnetworks):

$$(F_1, F_2, \ldots, F_m) = \mathrm{Split} X_{HPF},$$

each of which performs convolutions at different scales.

At level *k*, the feature representation is computed as:

$$F_k^{(l+1)} = \sigma(W_k^l * F_k^l)$$

where W_k is the convolutional kernel for the *l*-th layer in the *k*-th subnetwork and $\sigma(\cdot)$ is the activation function (typically ReLU).

This architecture allows MRS-Net to capture multiscale spatial information, thereby improving the detection of subtle changes in cover images caused by steganographic embedding under low payload conditions.

3.8. ResFormer Model

The ResFormer model is based on a hybrid architecture that combines residual blocks with a projection method for feature vectors into a multidimensional space [35, 50, 55], aiming to reduce the number of parameters required for model training while maintaining high steganogram detection accuracy [35].

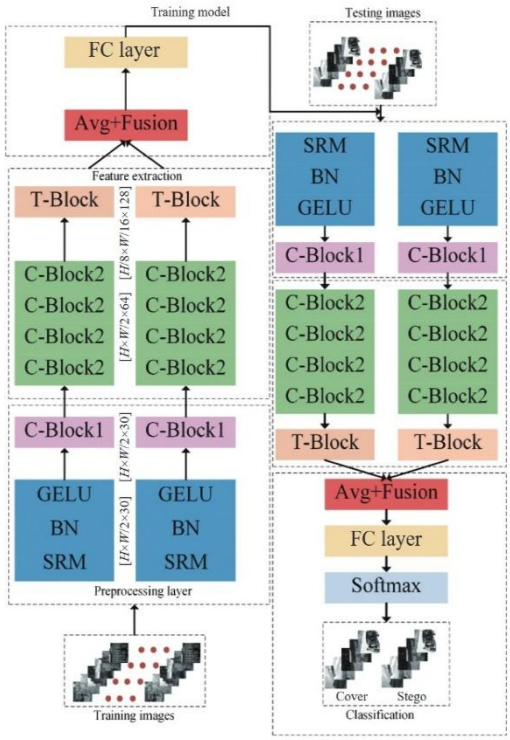

Figure 4: Structural architecture of the ResFormer artificial neural network according to paper [35]

The ResNet architecture (fig. 4) consists of a sequence of convolutional layers, each computing the activation $F^{(l)}$ using the equation:

$$F^{(l)} = \sigma\left(W^{(l)} * F^{(l-1)} + b^{(l)}\right),$$

where $\sigma(\cdot)$ denotes the activation function, $W^{(l)}$ and $b^{(l)}$ are the convolutional kernel and bias vector respectively, and * denotes the convolution operation with appropriate stride and padding.

A key role in the operation of ResFormer is played by the skip connection mechanism, which enables signals to bypass the main path of the network. Specifically, if $F^{(k)}$ is the output of layer *k*, then layer *k* + 1 receives both the transformed and the original signal:

$$F^{(k+1)} = \sigma\left(W^{(k+1)} * F^{(k)} + b^{(k+1)}\right) + F^{(k)}$$

These features of the ResFormer model help mitigate the vanishing gradient problem and make it feasible to train deep networks with a large number of layers.

4. Comparative Analysis of Detection Accuracy for Stegodetectors Based on Artificial Neural Networks

Given the application of the considered types of ANN in the design of steganalyzers, it is of practical interest to conduct a comparative analysis of their detection accuracy in the case of steganograms formed using modern adaptive steganographic methods.

It is worth noting that the literature lacks standardized results for evaluating these networks under identical test image datasets and embedding techniques. Therefore, this study presents a comparative performance assessment of ANN-based steganalyzers under harmonized testing conditions — using the same dataset of test images and identical steganographic embedding methods. The results of the analysis are presented in Table 1.

Based on testing using BOSSBase and BOWS2 datasets [4], the majority of ANNs (QianNet, XuNet, YeNet, YedroudjNet) demonstrate detection accuracies in the range of 70–80% for low embedding rates ($\Delta\alpha \approx 20\%$), depending on the specific steganographic method (WOW, S-UNIWARD, or HILL) and the test set used.

According to recommendations in [11], QianNet is typically trained on BOSSBase using

a 60/20/20 split for training, validation, and testing. Compared to traditional SPAM-based detectors [43], QianNet improves detection by approximately 3–4%. However, under more complex scenarios (e.g., SRM+EC [18, 27]), QianNet may underperform by 5–6%.

YeNet offers a 2–3% improvement over XuNet while YedroudjNet surpasses YeNet by an additional 1.6–2% at Δ_α = 20%, nearly matching the SRM+EC baseline at moderate embedding rates of $\Delta_\alpha \in [30\%;40\%]$.

ZhuNet further outperforms YedroudjNet by 2–3% at Δ_α = 40%, with differences up to 7–8% relative to simpler ANNs.

SRNet, which does not require SRM filter initialization, achieves over 90% accuracy even at low embedding levels ($\Delta_\alpha \approx 20\%$), albeit with higher computational demands.

MRS-Net maintains high-resolution feature extraction via multi-branch processing and yields 2–3% higher accuracy than SRNet, particularly for $\Delta_\alpha \in [20\%;40\%]$, though it requires more memory and training time.

Finally, ResFormer combines residual blocks with high-dimensional projection layers, reducing the total number of parameters by approximately 90% compared to SRNet and improving detection accuracy by 2–5%, depending on the method (WOW, S-UNIWARD, HILL) and the chosen Δα.

5. Discussions

Based on the comparative analysis of detection accuracy for steganalyzers built using ANN such as QianNet, XuNet, YeNet, and YedroudjNet, it was found that the achievable detection accuracy on standard digital image datasets (e.g., BOSSBase) ranges from 68.7% to 77.4% for medium levels of cover image payload ($\Delta_\alpha \in [20\%;40\%]$) and modern types of steganographic methods [11, 20]. The use of the ZhuNet architecture enables an increase in detection accuracy up to 84.5% at Δ_α = 40%, which outperforms XuNet and YeNet, although it requires additional time for steganalyzer training and configuration [12].

In contrast, applying deep ANN models — such as SRNet, which includes up to 25 layers — makes it possible to achieve high detection accuracy (above 80%) even for low embedding levels ($\Delta_\alpha \in [10\%;20\%]$) [6, 22]. A significant practical advantage of SRNet is its independence from high-pass filtering operations when processing digital images. However, the practical deployment of SRNet-based steganalyzers requires longer training times and large datasets due to the depth and complexity of the architecture [52].

The use of parallel feature extraction in the MRS-Net model reduces computational complexity during training while maintaining accuracy in the evaluation of image characteristics, particularly high frequency components [45]. Experimental studies have confirmed that the detection accuracy of MRS-Net based steganalyzers improves by approximately 2–3% compared to SRNet in scenarios with $\Delta_\alpha \in [20\%;40\%]$ [45, 10].

The hybrid model ResFormer combines residual blocks with feature transformation layers. The use of attention mechanisms enables the network to emphasize the influence of elements in adjacent layers and compactly represent their interdependencies, while the residual blocks allow the network to effectively process residual signals. This approach reduces the number of parameters by more than 90% compared to SRNet while maintaining the desired level of detection accuracy [35].

Thus, the ResFormer architecture integrates the advantages of SRNet — notably, its ability to model dependencies in brightness variations between adjacent pixel groups — and those of MRS-Net, by enabling the network to focus on textured regions in the image that are more likely to be targeted for steganographic embedding [5].

6. Conclusion

This paper presents results of comparative analysis of state-of-the-art stegodetectors based on artificial neural networks. The state-of-the-art ANN architectures (such as QianNet, XuNet, YeNet, YedroudjNet, ZhuNet, SRNet, MRSNet, and ResFormer) were reviewed, and their detection accuracy was analyzed under various evaluation conditions (for example, by changing of steganographic payload from 20% to 40%). It was shown that first convolutional neural network models (e.g., QianNet, XuNet, YeNet models) improve detection accuracy by approximately 2–3% compared to statistical detection methods (such as those based on SRM filters, SPAM/CC-PEV statistical models, or rich cover models). However, performance of ANN-based SD in case of usage of adaptive embedding methods (like HILL or WOW) at low payload

rates (up to 20%) remains insufficient for practical steganalysis applications (the detection accuracy is about 68–79%).

The development of steganalyzers based on YedroudjNet and ZhuNet models made it possible to further improve detection accuracy (up to 80-85%) due to the application of specialized convolution techniques, the use of SRM filters for preprocessing, and spatial-pyramid averaging of compared to SRNet, while maintaining the similar detection accuracy. This makes such ANN architectures promising candidates for practical deployment in real-time steganogram detection systems that maintain high levels of accuracy (above 90%).

Table 1. Comparison of steganogram detection accuracy for steganalyzers based on various artificial neural networks, tested on common datasets with different embedding rates and steganographic methods

ANN Model	Test Image Dataset	Embedding Rate	Steganographic Method	Detection Accuracy
QianNet	BOSSBase	10–30%	S-UNIWARD, HUGO, WOW	≈ 69–73%
	BOWS2	40%	S-UNIWARD, HUGO, WOW	≈ 78–79%
XuNet	BOSSBase	20%	S-UNIWARD, HILL	70–76%
	BOWS2	40%	S-UNIWARD, HILL	≈ 78–79%
YeNet	BOSSBase	20–40%	S-UNIWARD, WOW, HILL	≈ 68%
YedroudjNet	BOSSBase	20–40%	S-UNIWARD, WOW	≈ 77–78%
ZhuNet	BOSSBase	20–40%	S-UNIWARD, WOW	80–84.5%
SRNet	BOSSBase	10–40%	S-UNIWARD, WOW, HILL	80–90%
	BOWS2	30–40%	S-UNIWARD, WOW, HILL	≈ 92%
MRSNet	BOSSBase	20–40%	S-UNIWARD, WOW, MiPOD	83–92%
ResFormer	BOSSBase	10–40%	S-UNIWARD, WOW, HILL	85–95%
	BOWS2	20%	S-UNIWARD, WOW, HILL	≈ 90%

extracted feature vectors. Meanwhile, SRNet and MRS-Net models achieved even higher accuracy levels (up to 90–93%), although they require greater computational resources.

Recent advances in digital image steganalysis research [35] focus on applying of hybrid approaches that incorporate ResNet blocks and feature transformation in multidimensional spaces.

Specifically, the application of the ResFormer model in steganalyzer design reduces the total number of parameters by more than 90%

References

[1] A New Cost Function for Spatial Image Steganography / B. Li, M. Wang, J. Huang, X. Li // Proceedings of IEEE International Conference on Image Processing (ICIP). — 2014. — C. 4206 — 4210. — DOI: 10.1109/ICIP.2014.7025854.

[2] A. Sarkar K. S., Manjunath B. Obtaining higher rates for steganographic schemes

while maintaining the same detectability // Information Hiding, 12th International Workshop. Vol. 6387. — Calgary, Canada: Springer-Verlag, 2010. — P. 178–192.

[3] An improved approach to steganalysis of JPEG images / Q. Liu, A. H. Sung, M. Qiao, Z. Chen, B. Ribeiro // Information Sciences. — 2010. — Vol. 180, no. 9. — P. 1643–1655.

[4] Bas P., Filler T., Pevn´y T. "Break our steganographic system": The ins and outs of organizing BOSS // Lecture Notes in Computer Science: Information Hiding. T. 6958 / за ред. T. Filler, T. Pevny´, S. Craver, A. Ker. — Berlin, Germany: Springer, 2011. — C. 59—70.

[5] Bas P., Filler T., Pevný T. Break our steganographic system¨: The ins and outs of organizing BOSS // Proceedings of the 13th International Workshop on Information Hiding. — 2011.—P. 59–70.

[6] Boroumand M., Chen M., Fridrich J. Deep residual network for steganalysis // IEEE Transactions on Information Forensics and Security. — 2019. — Vol. 14, no. 5. — P. 1181–1193.

[7] Chen C., Shi Y. Q. JPEG image steganalysis utilizing both intrablock and interblock correlations // Circuits and Systems, ISCAS 2008. IEEE International Symposium on. — 05/2008. — P. 3029–3032.

[8] Chollet F. Xception: Deep learning with depthwise separable convolutions // Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — Honolulu, HI, USA, 2017. — P. 1800–1807.

[9] CNN-Based Adversarial Embedding for Image Steganography / W. Tang, B. Li, S. Tan, M. Barni, J. Huang // IEEE Transactions on Information Forensics and Security. — 2019.

[10] Deep learning applied to steganalysis of digital images: A systematic review / T.-S. Reinel, R.-S. Raul, I.-S. Gustavo, S.-H. Alexandra // IEEE Access. — 2019. — Vol. 7. — P. 68970–68990.

[11] Deep learning for steganalysis via convolutional neural networks / Y. Qian, J. Dong, W. Wang, T. Tan // Proceedings of SPIE 9409, Media Watermarking, Security, and Forensics. — 2015. — P. 1–10.

[12] Depth-wise separable convolutions for spatial CNNbased steganalysis / R. Zhang, F. Zhu, J. Liu, G. Liu // IEEE Transactions on Information Forensics and Security. — 2020. — Vol. 15. — P. 1138–1150.

[13] Digital Image Steganalysis Based on Visual Attention and Deep Reinforcement Learning / D. Hu, S. Zhou, Q. Shen, S. Zheng, Z. Zhao, Y. Fan // IEEE Access. — 2019.

[14] Efficient feature learning and multi-size image steganalysis based on CNN / R. Zhang, F. Zhu, J. Liu, G. Liu // arXiv preprint. — 2018. — July. — arXiv: 1807.11428[cs.CV]. — URL: https://arxiv.org/abs/1807.11428.

[15] Embedded methods / T. N. Lal, O. Chapelle, J. Weston, A. Elisseeff // Feature Extraction: Foundations and Applications. — Physica-Verlag, Springer, 2006. — P. 137–165. — (Studies in Fuzziness and Soft Computing).

[16] Filler T., Judas J., Fridrich J. Minimizing additive distortion in steganography using syndrome-trellis codes // IEEE Transactions on Information Forensics and Security. — 2011. — Vol. 6. — P. 920–935.

[17] Fridrich J. Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes // Information Hiding, 6th International Workshop. — Springer-Verlag, 2004. — P. 67–81.

[18] Fridrich J., Kodovský J. Rich models for steganalysis of digital images // IEEE Transactions on Information Forensics and Security. — 2012. — Under review.

[19] Fuzzy Localization of Steganographic Flipped Bits via Modification Map / Q. Liu, T. Qiao, M. Xu, N. Zheng // IEEE Transactions on Information Forensics and Security. — 2019.

[20] G. Xu H.-Z. W., Shi Y.-Q. Structural design of convolutional neural networks for steganalysis // IEEE Signal Processing Letters. — 2016. — Vol. 23, no. 5. — P. 708–712.

[21] GBRAS-Net: A Convolutional Neural Network Architecture for Spatial Image Steganalysis / T. Reinel, A.-A. H. Brayan, B.-0. M. Alejandro, M.-R. Alejandro, A.-G. Daniel, A. A.-G. J // IEEE Access. — 2020.

[22] Han X., Zhang T. Spatial steganalysis based on non-local block and multi-channel convolutional networks // IEEE Access. — 2022.

[23] Holub V., Fridrich J. WOW: A novel distortion function for spatial domain steganography // Proceedings of SPIE, Media Watermarking, Security, and Forensics. T. 8665. — 2013. — C. 86650V. — DOI: 10.1117/12.2009974.

[24] Holub V., Fridrich J., Denemark T. Designing steganographic distortion using directional filters // Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS). — 2012. — C. 234—239. — DOI: 10.1109/WIFS.2012.6412655.

[25] Image steganography techniques for resisting statistical steganalysis attacks: A systematic literature review / R. Apau, M. Asante, F. Twum, J. Ben Hayfron-Acquah, K. O. Peasah // PLOS ONE. — 2024. — T. 19, № 9.

[26] J. Fridrich T. P., Kodovský J. Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities // Proceedings of the 9th ACM Multimedia Security Workshop. — Dallas, TX, 2007. — P. 3–14.

[27] J. Kodovský J. F., Holub V. Ensemble classifiers for steganalysis of digital media // IEEE Transactions on Information Forensics and Security. — 2012. — To appear.

[28] J. Kodovský T. P., Fridrich J. Modern steganalysis can detect YASS // Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XII. Vol. 7541. — San Jose, CA, 2010. — P. 02–01–02–11.

[29] J. Ni J. Y., Yi Y. Deep learning hierarchical representations for image steganalysis // IEEE Transactions on Information Forensics and Security. — 2017. — Vol. 12, no. 11. — P. 2545–2557.

[30] K. Solanki A. S., Manjunath B. S. YASS: Yet another steganographic scheme that resists blind steganalysis // Information Hiding, 9th International Workshop. Vol. 4567. — Saint Malo, France: Springer-Verlag, 2007. — P. 16–31.

[31] Kim Y., Duric Z., Richards D. Modified matrix encoding technique for minimal distortion steganography // Information Hiding, 8th International Workshop. Vol. 4437. — Alexandria, VA: SpringerVerlag, 2006. — P. 314–327.

[32] Kodovský J. Steganalysis of Digital Images Using Rich Image Representations and Ensemble Classifiers: PhD thesis / Kodovský J. — Binghamton University, NY, 2012.

[33] Kodovský J., Fridrich J. Calibration revisited // Proceedings of the 11th ACM Multimedia Security Workshop. — Princeton, NJ, 2009. — P. 63–74.

[34] Kodovský J., Fridrich J. Steganalysis in high dimensions: Fusing classifiers built on random subspaces // Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XIII. Vol. 7880. — San Francisco, CA, 2011. — OL 1–13.

[35] Lightweight steganography detection method based on multiple residual structures and transformer / H. Li, Y. Zhang, J. Wang, W. Zhang, X. Luo // Chinese Journal of Electronics. — 2024. — Vol. 33, no. 4. — P. 965–978.

[36] Liu Q. Steganalysis of DCT-embedding based adaptive steganography and YASS // Proceedings of the 13th ACM Multimedia Security Workshop. — Niagara Falls, NY, 2011. — P. 77–86.

[37] M. Yedroudj F. C., Chaumont M. Yedroudj-Net: An efficient CNN for spatial steganalysis // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. — 2018. — P. 2092–2096.

[38] Nair V., Hinton G. E. Rectified linear units improve restricted Boltzmann machines // Proceedings of the 27th International Conference on Machine Learning. — Haifa, Israel, 2010. — P. 807–814.

[39] Pevný T., Fridrich J. Merging Markov and DCT features for multi-class JPEG steganalysis // Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX. Vol. 6505. — San Jose, CA, 2007. — P. 31–314.

[40] Sallee P. Model-based steganography // Digital Watermarking, 2nd International Workshop. — Springer-Verlag, 2003. — P. 154–167.

[41] Shi Y. Q., Chen C., Chen W. A Markov process-based approach to effective attacking JPEG steganography // Information Hiding, 8th International Workshop. Vol. 4437. — Alexandria, VA: Springer-Verlag, 2006. — P. 249–264.

[42] Spatial pyramid pooling in deep convolutional networks for visual recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the European Conference on Computer Vision. — Zurich, Switzerland, 2014. — P. 346–361.

[43] T. Pevný P. B., Fridrich J. Steganalysis by subtractive pixel adjacency matrix // Proceedings of the 11th ACM Multimedia Security Workshop. — Princeton, NJ, 2009. — P. 75–84.

[44] V. Sachnev H. J. K., Zhang R. Less detectable JPEG steganography method based on heuristic optimization and BCH syndrome coding // Proceedings of the 11th ACM Multimedia Security Workshop. — Princeton, NJ, 2009. — P. 131–140.

[45] Wang Z., Wu J. Multi-resolution network-based image steganalysis model // IEEE Transactions on Information Forensics and Security. — 2023.

[46] Westfeld A. High capacity despite better steganalysis (F5 – a steganographic algorithm) // Information Hiding, 4th International Workshop. Vol. 2137. — Pittsburgh, PA: Springer-Verlag, 2001. — P. 289–302.

[47] Luo Y., Zhang X. GAN-based cover image synthesis and its steganalysis // Proceedings of the 9-th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'21). — New York, NY, 2021. — P. 81–92. — DOI: 10.1145/3466752.3467058.

[48] Deng L., Li B. Self-supervised calibration for robust image steganalysis // IEEE Access. — Vol. 10, 2022. — P. 75132–75144. — DOI: 10.1109/ACCESS.2022.3185021.

[49] Su H., Zhang R. Channel-attention YeNet for improved spatial steganalysis // Signal Processing: Image Communication. — Vol. 105, 2022. — Art. 103451 (14 p.). — DOI: 10.1016/j.image.2022.103451.

[50] Li K., Zhang W., Luo X. Vision-Transformer steganalysis with hierarchical feature tokens // IEEE Transactions on Information Forensics and Security. — Early Access, 2024. — 13 p. — DOI: 10.1109/TIFS.2024.3367890.

[51] Boroumand M., Fridrich J. Deep residual network for JPEG steganalysis // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). — Seattle, WA, 2020. — P. 10590–10599. — DOI: 10.1109/CVPR42600.2020.01059.

[52] Cozzolino R., Thies J., Rössler A., Riess C., Nießner M., Verdoliva L. ID-Reveal: Identity disclosure attacks on deep-fake videos // IEEE Journal of Selected Topics in Signal Processing. — Vol. 14, No. 5, 2020. — P. 188–205. — DOI: 10.1109/JSTSP.2020.3035434.

[53] Jia Y., Chen C. Pixel-wise steganographic map prediction via U-Net for active steganalysis // IEEE Transactions on Information Forensics and Security. — Vol. 17, 2022. — P. 2148–2160. — DOI: 10.1109/TIFS.2022.3141234.

[54] Dong J., Li B., Tan S. CSM-ResNet: Cover-source-mismatch aware residual network for universal image steganalysis // Pattern Recognition. — Vol. 136, 2023. — Art. 109338 (13 p.). — DOI: 10.1016/j.patcog.2023.109338.

[55] He K., Zhang Q., Sun J. Dual-domain Transformer for JPEG steganography detection // Proceedings of the 31-st IEEE International Conference on Image Processing (ICIP 2024). — Singapore, 2024. — P. 1299–1303. — DOI: 10.1109/ICIP47073.2024.10235679.