ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 7, No. 1, 2022

PERFORMANCE ANALYSIS OF STEGO IMAGE CALIBRATION WITH THE USAGE OF DENOISING AUTOENCODERS

Dmytro Progonov

Igor Sikorsky Kyiv Polytechnic Institute, 37, Prosp. Peremohy, Kyiv, 03056, Ukraine.
Authors' e-mail*: d.progonov@kpi.ua*

https://doi.org/10.23939/acps2022.__.___

Submitted on XX.XX.2022

***Abstract*: Methods for early detection of sensitive information leakage by data transmission in open (public) communication systems have been of special interest. Reliable detection of modified (stego) cover files, like digital images, requires usage of computation-intensive methods of statistical steganalysis, namely covering rich models and deep convolutional neural networks. Necessity of fine-tuning parameters of such methods to minimize detection accuracy for each embedding methods has made fast re-train of stegdetectors in real cases impossible. Therefore, development of low-complexity methods for detection of weak alterations of cover image parameters under limited prior information about used embedding methods has been required. For solving this task, we have proposed to use special architectures of artificial neural networks, such as denoising autoencoder. Ability of such networks to estimate parameters of original (cover) image from the noisy ones under limited prior information about introduced alterations has made them an attractive alternative to state-of-the-art solutions. The results of performance evaluation for shallow denoising autoencoders showed increasing of detection accuracy (up to 0.1 for Matthews correlation coefficient) in comparison with the state-of-the-art stegdetectors by preserving low-computation complexity of network retraining.**

***Index Terms*: adaptive embedding methods, digital image steganalysis, denoising autoencoders.**

INTRODUCTION

Feature of modern critical information infrastructures (CII) of state as well as private organization is tight integration with local and global high-speed communication systems [1]. This considerably improves performance of CII in scenarios related to remote control, load balancing, fast resources re-allocation in case of failure to name a few. On the other hand, connection of CII elements to local and/or global communication systems increases risk of sensitive information leakage. Therefore, reliable protection of such information during message transmission between elements of CII is topical task today.

In most cases, unauthorized transmission of sensitive information related to CII elements is performed with the usage of advanced steganographic systems [2, 3]. They allow hiding even the fact of sensitive data transmission by its embedding into innocuous files, like digital images (DI). Appearance of adaptive embedding methods (AEM) in the last years makes it possible to considerably decrease alterations of cover image statistical parameters. Detection of formed stego images required usage of computation-intensive methods of statistical steganalysis, such as cover rich models [4] and convolutional neural networks (CNN) [5]. In spite of high detection accuracy, such stegdetectors (SD) requires time-consuming adjusting of parameters that makes fast re-training of used SD for detection of unknown embedding methods (zero-day problem) impossible.

The modern approach to improve detection accuracy for stego images formed by AEM is image pre-processing (calibration). Proposed methods allow to detect and extract weak alterations of CI parameters caused by message hiding [5]. However, this is achieved by corresponding increase of method's computation complexity as well as limited ability to fast re-tune unknown embedding methods. Therefore, of special interest there are low-complexity methods for detecting weak perturbations of pixels brightness for DI, such as image denoising algorithms.

To solve this task, we propose to apply advanced types of artificial neural networks, namely denoising autoencoder (DAE), for improving performance of image calibration by preserving low-complexity as well as ability to fast re-tune new embedding methods. The paper is aimed at performance analysis of shallow DAE for calibration of stego images formed by adaptive embedding methods.

RELATED WORKS

State-of-the-art approach for stego image detection is based on applying extensive set of high-pass filters (HPF), and further analysis of statistical parameters for obtained residuals [6, 7]. The approach was widely used for development of modern SD for reliable detection of stego images formed by widespread steganographic methods. Nevertheless, practical usage of such SD requires time-consuming estimation of optimal parameters for HPF for minimization of detection error P_E. This makes SD impractical for fast re-tuning by appearance of new steganographic methods that is a common situation for modern intrusion detection systems.

One of the proposed approaches for overcoming mentioned limitations of modern SD is based on applying additional image pre-processing (calibration) methods [8]. Proposed solutions for fast estimation of optimal parameters under criteria of minimization the P_E are based on applying artificial neural networks, namely CNN [5, 9]. Ability to fast adjustment of network's parameters to new inputted samples (cover and stego images) by backpropagation procedure makes CNN a promising candidate for advanced SD. Still, providing high detection accuracy (more than 95%) with the usage of CNN requires utilization of deep networks that complicates their training and increases requirements to volume of the used datasets. The proposed solutions for decreasing computation-complexity of CNN-based SD re-tuning are based on applying pre-trained networks and theirs tuning to target dataset [10]. However, limited quantity of publicly available pre-trained CNN for steganalysis-related tasks requires further time-consuming optimization of CNN for target dataset.

Alternative approach for improving performance of SD is based on applying special types of image calibration methods, such as those based on cover and stego estimations [8, 11, 12]. A distinct feature of these methods is estimation parameters of either cover image from current noisy DI, or expected distortions by message re-embedding. Results of performance evaluation of such methods proved their effectiveness in the case of stego image formation by AEM [11-13]. However, their practical usage requires analysis of prior information about used AEM for selection of appropriate calibration methods. In most cases, steganalytics has limited opportunities to give this information. This causes necessity of development of calibration methods that can be adjusted to minimize P_E values under available prior information about AEM.

Mentioned task can be reformulated as an optimization problem of anisotropic image denoising under limited prior information about statistical and spectral parameters of noising. Therefore, we may apply of advanced architectures of artificial neural networks, namely DAE, for solving this problem. Despite wide range of the proposed SD based on CNN and DAE, their performance for AEM remains unclear. We aimed at filling this gap by performance analysis of shallow DAE for calibration of stego images formed by novel S-UNIWARD and MG embedding methods.

ADAPTIVE EMBEDDING METHODS FOR DIGITAL IMAGES

The state-of-the-art paradigm of DI steganography is based on minimization of CI alteration during message hiding [14]. This leads to considerable decrease of stego images unmasking features (e.g., changes of statistical features) that decrease performance of modern stegdetectors.

Mentioned breakthrough of novel steganography methods is achieved by representation of message hiding procedure as the optimization problem with constraints [15]:

$$D(\mathbf{X},\mathbf{Y}) = \sum_{i,j} \rho_{i,j}(\mathbf{X},\mathbf{Y}) \xrightarrow{|\mathbf{M}|=\mathrm{const}} \min, \qquad (1)$$

where $\mathbf{X},\mathbf{Y} \in \Im = \{0,1,\ldots,2^k - 1\}^{M \times N}$ are cover and stego images of size MxN pixels correspondingly; $k \in \mathbb{N}$ is color bit-depth; $D(\cdot,\cdot)$ is empirical function for estimation of CI distortion during stego data hiding; $\rho_{i,i}(\cdot,\cdot)$ is empirical function for estimating cover image's statistical feature alteration by changes of $(i,j)^{\text{th}}$ pixel; $\mathbf{M}$ is a binary representation of stego data m-bits.

In the general case, the function $\rho(\cdot)$ in (1) allows to estimate changes of CI statistical parameters caused by a single pixel alteration as well as non-linear dependencies between these changes by embedding series of bits [15]. The former alteration can be performed using common statistical models of DI [4, 16]. The latter highly depends on mutual influence of altered pixels that requires utilization of computationally intensive methods for such dependency estimation. In most cases, mentioned dependencies may be estimated only for small (short) message (up to 100 bits) [15]. Therefore, majority of modern embedding methods includes "simplified" functions $\rho(\cdot)$ that provide tractable estimation on single pixel alterations only.

The selection of CI pixels to be used for stego bits hiding is usually made by analysis of statistical parameters of current pixel neighborhood (clique) [15]. This allows to provide low cover image alteration during message hiding by preserving tractable complexity of the embedding algorithm.

The advance adaptive embedding methods S-UNIWARD [17] and MG [18] were considered in the work. The S-UNIWARD embedding method is based on spectral transformation for estimation CI distortions caused by embedding of individual stegobits. The S-UNIWARD method takes additive empirical distortion estimation function [17]:

$$D(\mathbf{X},\mathbf{Y}) = \sum_{k,u,v} \frac{\left|\mathbf{W}_{uv}(\mathbf{X},k) - \mathbf{W}_{uv}(\mathbf{Y},k)\right|}{\sigma + \left|\mathbf{W}_{uv}(\mathbf{X},k)\right|}, \qquad (2)$$

where $\mathbf{W}_{uv}(\mathbf{X},\ k)$, $\mathbf{W}_{uv}(\mathbf{Y},\ k)$ – coefficients of two-dimensional discrete wavelet transform (2D-DWT) of the cover $\mathbf{X}$ and stego $\mathbf{Y}$ images with coordinates $(u,\ v)$ in the k^{th} sub-band; $\sigma>0$ – stabilizing constant.

Variation of 2D-DWT basis functions in (2) allows to analyze specific distortions of CI caused by message hiding. Also, usage of empirical distortion estimation function (6) makes it possible to message hiding in spatial (alteration of CI pixels brightness) and transformation (by changing of a CI decomposition coefficients) domains in the uniform way.

The feature of MG method is minimization of both the CI distortion, and SD performance (detection accuracy) during stego data embedding [18, 19]. This is achieved through the usage of Gaussian mixture models (GMM) for estimation cover image's noises parameters [19].

The cover image processing pipeline is similar for MG [18] method. At the first stage, the CI is pre-processed (filtered) for suppressing the influence of cover image context using a filter F_{dn}:

$$\mathbf{r} = \mathbf{X} - F_{dn}\left(\mathbf{X}\right).$$

Then, variance σ_l^2 of pixels brightness for computed residuals r is calculated using next linear model:

$$\mathbf{r}_l = \mathbf{G}\mathbf{a}_l + \xi, l \in \left[1; M \cdot N\right]. \qquad (3)$$

Sedighi et al proposed to use Maximum Likelihood for estimation of the mentioned model parameters [19]:

$$\sigma_l^2 = \frac{\left\|\hat{\mathbf{P}}_{\mathbf{G}}\mathbf{r}_l\right\|_F}{p^2 - q}, q \in \mathrm{N},$$

where $\hat{\mathbf{P}}_{\mathbf{G}}$ is a projection operator for residuals $\mathbf{r}_l$ (3) on sub-space with (p^2-q) dimensionality, created from eigenvectors of matrix **G**; $\left\|\cdot\right\|_F$ is Frobenius norm. Residuals $\mathbf{r}_l$ are computed within neighborhood of pxp pixels for current lth pixel.

At the third stage, the magnitude β_l, $1 \le l \le M \cdot N$, of pixel brightness changes that minimizes the deflection coefficient ς^2 between cover and stego image distributions is estimated:

$$\varsigma^2\left(\beta_l\right) = 2\sum_{l=1}^{M \cdot N} \beta_l^2 \sigma_l^4 \xrightarrow{\sum_{l=1}^{M \cdot N} \mathrm{H}_4\left(\beta_l\right) = const} \min, \qquad (4)$$

$$\mathrm{H}_4\left(z\right) = -2z\log\left(z\right) - \left(1 - 2z\right)\log\left(1 - 2z\right),$$

where $\mathrm{H}_4(z)$ is ternary entropy function. The deflection coefficient ς^2 (4) provides statistical measurement of divergence between cover and stego images distribution that reflects expected performance of statistical SD [18, 19].

The mentioned optimization task for coefficient ς^2 (4) can be solved using Lagrange multipliers method [19]. Then, optimal values of β_l and Lagrange multipliers λ_L can be calculated by numerical solving of next equations:

$$\beta_l^2 \sigma_l^4 = \frac{1}{2\lambda_L} \ln\left(\frac{1 - 2\beta_l}{\beta_l}\right), l \in \left[1; M \cdot N\right].$$

Estimated optimal values of β_l are used for calculating corresponding values of ρ_l function during embedding stegobit into lth pixel of CI:

$$\rho_l = -\ln\left(\beta_l - 2\right), l \in \left[1; M \cdot N\right]. \qquad (5)$$

At the last stage, message **M** bits are embedded into CI using trellis-code along with magnitudes of pixel brightness alteration estimated with ρ_l (5).

It should be noted that the GMM used in MG method allows to accurately estimate local alterations of pixel brightness during stego image formation [19]. This provides high robustness of formed stego images to known statistical steganalsis methods without involving of computationally intensive methods for image modeling, such as Random Markov Fields [20].

DIGITAL IMAGE CALIBRATION USING DENOISING AUTOENCODERS

The feature of modern DI denoising methods is adjustment of method's parameters by taking into account image's local statistical parameters. One of widespread approaches to solving this task is usage of anisotropic filtering methods [21] that takes into account parameters of textures and objects within current position of sliding window (SW). Examples of such denoising methods are bilateral filtering (BF) and non-local means (NLM) algorithms [21].

The bilateral filter is a non-linear, edge-preserving, and noise-reducing smoothing filter [22]. It is based on applying composition of two filters to reduce impact of additive noise on image quality [22]:

$$F_{BF}\left(\mathbf{U}_{x,y}\right) = \frac{1}{N_{BF}\left(i,j\right)} \sum_{k=-\left(h_k-1\right)/2}^{\left(h_k-1\right)/2} \sum_{n=-\left(h_n-1\right)/2}^{\left(h_n-1\right)/2} \mathbf{U}_{x+k,y+n} \cdot \qquad (6)$$
$$\cdot h\left(k,n\right) \cdot g\left(\mathbf{U}_{x+k,y+n} - \mathbf{U}_{x,y}\right),$$

$$N_{BF}\left(i,j\right) = \sum_{k=-\left(h_k-1\right)/2}^{\left(h_k-1\right)/2} \sum_{n=-\left(h_n-1\right)/2}^{\left(h_n-1\right)/2} h\left(k,n\right) \cdot$$
$$\cdot g\left(\mathbf{U}_{x+k,y+n} - \mathbf{U}_{x,y}\right),$$

where $h(\cdot,\cdot)$ – smoothing filter with size of h_kxh_n (pixels) that is used for suppression of additive noises; $g(\cdot)$ – stop-functions that limits impact of filter $h(k, n)$ by processing image areas near contours; $N_{BF}(i, j)$ – normalizing factor for the current position of SW.

The stop-function $g(\cdot)$ is adjusted to preserve fixed impact of smoothing filter $h(k, n)$ in the image's areas where variation of pixel brightness is below of predefined threshold, for example, textured objects, sand, grass etc. On the other hand, values of function $g(\cdot)$ tends to zero in case of processing areas with sharp changes of pixel brightness (for example, near contours) that suppress influence of smoothing filter $h(k, n)$. In most cases, the considered functions $h(k, n)$ and $g(\cdot)$ are based on Gaussian function that decreases computation complexity of their usage in image processing tools [18]:

$$G\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(x-\mu\right)^2}{2\sigma^2}}, \qquad (7)$$

where μ, σ – respectively, the mean and variance of the pixel brightness for a current position of SW. Estimation of these parameters can be performed in a manner similar to Wiener filter [23]:

$$\sigma_\eta^2 = \frac{1}{MN} \sum_{n,m \in \eta} \mathbf{U}_{n,m}^2 - \mu_\eta^2, \qquad (8)$$

$$\mu_{\eta} = \frac{1}{MN}\sum_{n,m\in\eta} \mathbf{U}_{n,m}, \qquad (9)$$

where **U** – a grayscale image with size *N*x*M* (pixels); $\eta -$ the current position of SW with size w_Wxw_W (pixels); $\mu_{\eta}, \sigma_{\eta}^{2} -$ estimations of the mean and variance value of the pixels brightness for current position of SW respectively. The value σ^2 in eq. (7) is estimated by averaging of σ_{η}^{2} values obtained for all positions of the SW for Wiener filter.

The bilateral filter allows to adaptively decrease impact of additive noise including influence of alterations caused by message hiding. This makes BF a promising candidate for stego image calibration tasks in steganalysis. On the other hand, usage of smoothing filters in the bilateral filter (6) may negatively impact on image calibration performance due to influence on the whole image region instead of processing of an individual pixels only. Therefore, of special interest there are methods of image denoising that takes into account variability of local pixel brightness like NLM-filter [24]. The filter is based on minimizing the variance of pixel brightness by analyzing deviation of current pixel brightness from the mean brightness for current position of SW [24]:

$$F_{NLM}\left(\mathbf{U}_{x,y}\right) = \frac{1}{N_{NLM}\left(x,y\right)}\sum_{(x,y)\in w_n} \mathbf{U}_{x,y} \cdot w\left(x,y\right),$$
$$N_{NLM}\left(x,y\right) = \sum_{(x,y)\in w_n} w\left(x,y\right),$$

where w_n – current position of SW; $w(x, y)$ – is a function that scales the brightness value of the current pixel depending on its deviation from the mean brightness for current position of the SW; $N_{NLM}(i, j)$ – normalization factor for current position of SW. Similar to BF, the Gaussian function (7) is widely used as function $w(x, y)$ for NLM-filter to reduce the impact of small objects on an image of processing results.

Performance of considered bilateral and NLM filters is theoretically established for additive noises with predefined statistical or spectral parameters, like color noise, shot noise, speckle noise to name a few. Therefore, effectiveness of their usage for stego image calibration tasks may reduce due to non-local character of CI pixel brightness modifications. This is proved by performance evaluation of image calibration with the usage of novel image denoising techniques for stego images formed according to AEM [11, 12]. The reason of this is "aggressively" suppression of noises (interferences) that lead to significantly decrease of differences between cover and stego images processing results. Therefore, application of specific image denoising methods for detection weak perturbation of pixels brightness is of special interest

One of promising methods for solving mentioned task is usage of CNN. The important feature of CNN is ability to adjust their parameters during training to minimize predefined objective function, for example, total variation of image's pixel brightness [25]. Results of performance evaluation of state-of-the-art architectures for CNN proved effectiveness of this approach for improving SD detection accuracy in case of processing widespread embedding methods [26, 27]. However, performance of CNN-based image calibration highly depends on either prior information of used embedding methods, or samples of stego images that can be used for CNN tuning [27]. This requires often retraining of CNN-based SD to preserve fixed detection accuracy for new set of images that is computation-intensive operations even by usage of pre-trained models [10]. This limits practical usage of CNN in real cases.

One of approaches for overcoming mentioned limitations is usage of special types of artificial neural networks such as autoencoder networks (AEN) [25]. The feature of AEN is estimation of image's features that can be sensitive to changes of image parameters. This is achieved by usage of bottleneck-like architecture with encoder and decoder parts (Fig. 1).

The first part of the AEN is encoder network that is aimed at projection of a given multidimensional signal (feature vector of an image) into lower-dimensional space, while maintaining its statistical features (Fig. 1). Restoration of initial image is performed by the decoder network according to the obtained representation **h**.

Imposing additional restrictions on parameters of encoder/decoder networks (Fig. 1) allows to obtain specific properties of AEN, for example, adaptive image denoising, image inpainting [25]. This feature makes AEN an attractive candidate for image calibration related tasks when parameters of image alterations cannot be estimated in advance. Therefore, image calibration task can be solved with the usage of the DAE by solving the following optimization task [13, 25]:

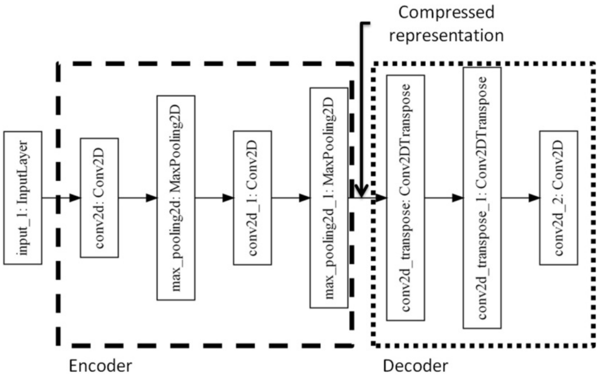

Compressed representation
input_1: InputLayer
conv2d: Conv2D
max_pooling2d: MaxPooling2D
conv2d_1: Conv2D
max_pooling2d_1: MaxPooling2D
conv2d_transpose: Conv2DTranspose
conv2d_transpose_1: Conv2DTranspose
conv2d_2: Conv2D
Encoder
Decoder

Fig. 1. General architecture of an autoencoder network for digital image processing

$$-\mathbb{E}_{\mathbf{U}\sim\hat{p}_{data}(\mathbf{U})}\mathbb{E}_{\tilde{\mathbf{U}}\sim C(\tilde{\mathbf{U}}|\mathbf{U})}\log\left(p_{decoder}\left(\mathbf{U}\,|\,\mathbf{h}=f\left(\tilde{\mathbf{U}}\right)\right)\right)\to\min,$$

where $\mathbf{U},\tilde{\mathbf{U}} -$ are initial and noised images respectively; $C\left(\tilde{\mathbf{U}}\,|\,\mathbf{U}\right) -$ function of introducing specified distortions to the image; $\hat{p}_{data}\left(\mathbf{U}\right) -$

probabilistic distribution of DI used to DAE training; $p_{decoder}(\cdot)$ – probabilistic distribution of images at the output of the DAE's decoder network; **h** –encoder network output (latent representation of the feature vector of inputted image **U** in a lower-dimension space).

In most cases, the shallow DAE (up to a dozen of hidden layers) is considered due to high computation complexity of its training procedure. The autoencoder part of ASSAF model [13] can be mentioned as an example of DAE used for image steganalysis. The architecture of this autoencoder is represented in Fig. 2.

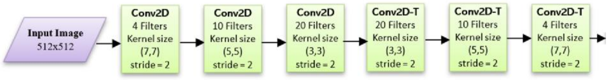


Fig. 2. The architecture of the denoising autoencoder for the ASSAF model. According to [13]

The feature of considered DAE is usage of convolutional layers to obtain an intermediate (latent) representation of the given image feature vector in a lower-dimension space. This allows to significantly reduce the requirements for the size of the test dataset, as well as the complexity of DAE training in comparison with the usage of fully connected layers [25].

Note that DAE is used as a key element part of stego image processing pipelines for CNN-based stegdetectors [13]. Therefore, the autoencoder is trained with either a whole model, or pre-trained on predefined (standard) dataset of stego images formed by widespread embedding methods. This makes its performance for image calibration tasks unclear. We aimed at filling this gap by performance evaluation of DAE pre-trained with the usage of advanced embedding methods, such as MG and UNIWARD group of methods.

EXPERIMENTS

The performance analysis of DAE and adaptive image denoising methods was done on a subset of 10,000 randomly sampled from a standard data package ALASKA [28]. These images were converted to grayscale mode and resized to fixed size of 512x512 pixels

Stego images were formed according to advance adaptive methods S-UNIWARD and MG. The CI payload was varied in the following range – 3%, 5%, 10%, 20%, 30%, 40% and 50%.

The cover and stego image pre-processing with bilateral and NLM filter was performed with recommended parameters for image denoising applications. Namely, size of sliding windows was chosen equal to 5x5 pixels, the mean and standard deviation values for pixel brightness were estimated according to eq. (8)-(9).

The DAE (Fig. 2) was trained with the usage of additional subset of 10,000 images from the standard ALASKA dataset [28]. The new subset was used for forming of stego images according to S-UNIWARD and MG methods by uniform sampling of CI payload from 5% to 50%. The size of the input image of used autoencoder (Fig. 2) is chosen equal to 512x512 pixels, which allows to process test images from standard image datasets without having to rescale these images. The DAE was trained for 200 epochs with Adam optimizer, by decreasing of learning rate from 0.01 on each 25 epochs with scaler 0.1.

The stegdetector was based on using the mentioned image calibration methods as well as standard SPAM model [29] for estimating statistical parameters of the processed images. The classification of extracted features to the classes of cover or stego images was performed with the usage of ensemble classifier, namely Random Forest [30]. The classifier was tuned by minimization of total error P_E [30]:

$$P_E = \left(P_{FA} + P_{MD}\left(P_{FA}\right)\right) / 2,$$

where P_{FA} and P_{MD} are probabilities of false alarm (type I error) and missed detection (type II error) correspondingly. Validation of SD was performed 10 times by pseudo random splitting of image dataset into train (70%) and test (30%) samples.

Note that image pre-processing (calibration) leads to extension of the number of features that can be used for stego image detection. According to the results of research [11, 12], the following features were used for SD tuning:

Linearly transformed features of the calibrated image – correspond to the difference between the features of calibrated and original images:

$$\mathbf{F}_{DF} = \mathbf{F}_{calib} - \mathbf{F}_{nc}. \tag{10}$$

Cartesian product of the features for calibrated and original images:

$$\mathbf{F}_{CC} = \left\{\mathbf{F}_{calib}; \mathbf{F}_{nc}\right\}.$$

where $\mathbf{F}_{nc}$ and $\mathbf{F}_{calib}$ are features for initial (non-processed) and calibrated images respectively.

Also, performance of SD considerably depends on prior information about used embedding methods. Estimation of stegdetectors accuracy in this case can be done with the usage of the following index F_α [31]:

$$F_\alpha = \left|\left\{\left(\mathbf{X}, \mathbf{Y}\right) : \left(\mathbf{X}_i, \mathbf{Y}_{\mathbf{X}_i}\right), i \in S_{train}\right\}\right| \Big/ \left|S_{train}\right| \times 100\%$$

where S_{train} is a set of digital images used during training of stegdetector; $\mathbf{Y}_{\mathbf{X}_i}$ is stego images formed from cover $\mathbf{X}_i$. The F_α parameter varies from 0% (absent from cover-stego images pairs in training set) to 100% (training set consists only from cover-stego image pairs). The former case corresponds to a situation when steganalytics can use only captured stego images. The latter relate to the situation when steganalytics can cover ones for stego images from arbitrary, but analytics limited in knowledge about features of embedding process. The most interesting case of SD evaluation under absence of prior information about used embedding method (F_α =0%) was considered in the research.

The stegdetectors based on standard SPAM [29] (without image calibration) as well as state-of-the-art maxSRMd2 [16] models were considered for comparison. The maxSRMd2 model is based on image calibration by usage of extensive set of HPF. This allows to considerably reduce impact of image content on image calibration results. On the other hand, this leads to enormous set of 12,753 features that complicates fast tuning of SD for detection of new embedding methods.

The detection accuracy of trained SD was evaluated with the usage of Matthews Correlation Coefficient *MCC*. The coefficient is used to estimate the degree of correlation of the (true) labels of the classes of the studied images with the output of SD [32]:

$$MCC = \left(P_{TP} \times P_{TN} - P_{FP} \times P_{FN}\right) \Big/ \sqrt{N_{MCC}}\ ,$$

$$N_{MCC} = \left(P_{TP} + P_{FP}\right) \times \left(P_{TP} + P_{FN}\right) \times$$

$$\times \left(P_{TN} + P_{FP}\right) \times \left(P_{TN} + P_{FN}\right),$$

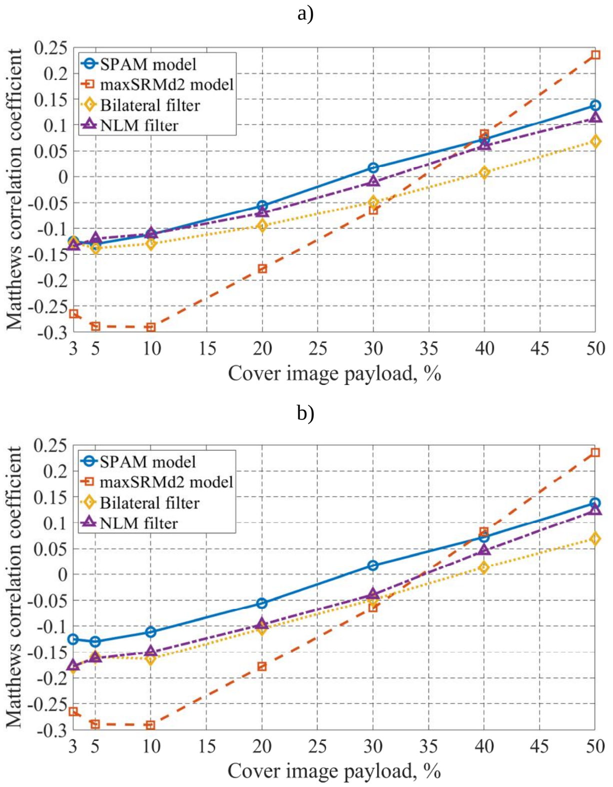
where P_{TP}, P_{TN} – the probabilities of correct classification of stego and cover images respectively; P_{FP} – the probability of incorrect classification of cover images as stego ones; P_{FN} – the probability of incorrect classification of stego images as cover ones.

The value of the *MCC* varies from (-1) that corresponds to the case of classification of stego images as cover ones and vice versa, to (+1) that relates to correct classification of both cover and stego images. The special case is *MCC*=0 that corresponds to the case of assigning analyzed image to the classes of cover or stego images randomly (P_{FN}=P_{FP}).

Performance evaluation of considered image denoising methods and DAE was done in two stages. Firstly, the detection accuracy was estimated for the case of applying the state-of-the-art image denoising methods to stego images formed according to S-UNIWARD and MG embedding methods. The dependencies of Matthews Correlation Coefficient *MCC* on the cover image payload by usage of maxSRMd2 and SPAM models as well as considered image denoising methods for the S-UNIWARD steganographic method by F_{α}=0% are shown in Fig. 3.

Note that usage of standard SPAM model allows to considerably (Δ*MCC*=0.15) improve values of MCC in comparison with advanced maxSRMd2 model (Fig. 3). This is achieved for most difficult cases of low (less than 10%) and medium (less than 20%) cover image payload. Obtained results can be explained by negative impact of huge ensemble of HPF that complicates tuning of SD.

On the other hand, application of considered denoising methods led to decrease of *MCC* values in comparison with corresponding results for SPAM model (Fig. 3). This can be explained by insufficient “selectivity” of the mentioned methods to detect and suppress local perturbation of pixel brightness caused by stego bits embedding. Note that usage of $\mathbf{F}_{\mathrm{DF}}$ features (Fig. 3a) leads to improvement of *MCC* values in comparison with widely used $\mathbf{F}_{\mathrm{CC}}$ features (Fig. 3b). This is caused by doubling $\mathbf{F}_{\mathrm{CC}}$ feature dimensionality that decreases SD performance by training on fixed size training set. In comparison, dependencies of Matthews Correlation Coefficient *MCC* on the cover image payload by usage of maxSRMd2 and SPAM models as well as considered image denoising methods for the MG steganographic method by F_{α}=0% are shown in Fig. 4.


a)

b)

Fig. 3. Dependencies of Matthews correlation coefficient on the cover image payload by usage of $\mathbf{F}_{DF}$ *(a) and* $\mathbf{F}_{CC}$ *(b) features for the adaptive steganographic methods S-UNIWARD by applying of advanced methods for image denoising and* F_{α}*=0%*

In contrast to S-UNIWARD embedding method (Fig. 3), usage of SPAM model for stego images formed by MG method (Fig. 4) allows to improve *MCC* values only for low CI payload range (less than 10%). Further increase of CI payload leads to drastic increase in detection accuracy for SD based on maxSRMd2 model (Fig. 4). This can be explained by message embedding with the usage of Gaussian like noises for MG method that can be effectively suppressed by ensemble of HPF for maxSRMd2 model. The S-UNIWARD method is based on usage of two-dimensional discrete wavelet transformation [17]. Therefore, embedding of a single stego bit for this method is achieved by alteration of CI pixels without perturbance of intrinsic noises parameters.

Usage of $\mathbf{F}_{\mathrm{DF}}$ (Fig. 4a) and $\mathbf{F}_{\mathrm{CC}}$ (Fig. 4b) features leads to similar values of *MCC* for MG embedding method. Therefore, we may conclude that difference between results of application of considered denoising methods are also similar. Thus, of interest there is usage of DAE-based image calibration methods for improving detection accuracy in comparison with novel denoising methods.

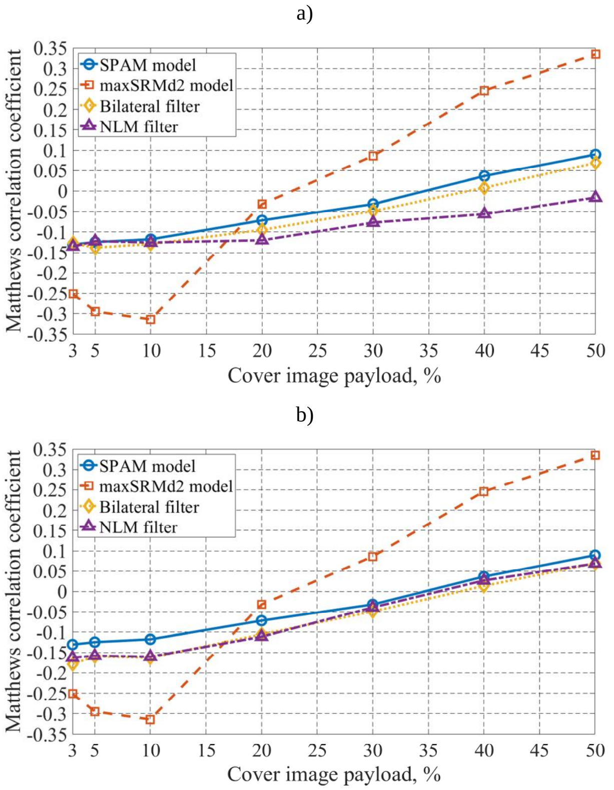


Fig. 4. Dependencies of Matthews correlation coefficient on the cover image payload by usage of $\mathbf{F}_{DF}$ *(a) and* $\mathbf{F}_{CC}$ *(b) features for the adaptive steganographic methods MG by application of advanced methods for image denoising and* F_α*=0%.*

The analysis of SD performance by the usage of DAE-based image calibration was node on the second stage of research. The dependencies of Matthews Correlation Coefficient *MCC* on the cover image payload by the usage of SPAM models and trained DAE network for the S-UNIWARD steganographic method by F_α=0% are shown in Fig. 5.

Applying DAE for stego image calibration leads to considerable improvement of MCC values (Δ*MCC*=0.10) in comparison with the case of SPAM model usage (Fig. 5). Still, the improvement is achieved for $\mathbf{F}_{\mathrm{DF}}$ features (Fig. 5a) and high CI payload (more than 20%), where SD based on maxSRMd2 model achieves even better results (Fig. 3). On the other hand, application of $\mathbf{F}_{\mathrm{CC}}$ features (Fig. 5b) makes detection accuracy improvement possible in the whole cover image payload range. However, an increase of MCC in this case is much smaller (Δ*MCC*=0.5) in comparison of the usage of $\mathbf{F}_{\mathrm{DF}}$ features (Fig. 5a). This can be explained by negligible differences between features of initial (non-processed) and calibrated images for low CI payload (less than 10%) that negatively impact SD training with the usage of $\mathbf{F}_{\mathrm{DF}}$ features (10).

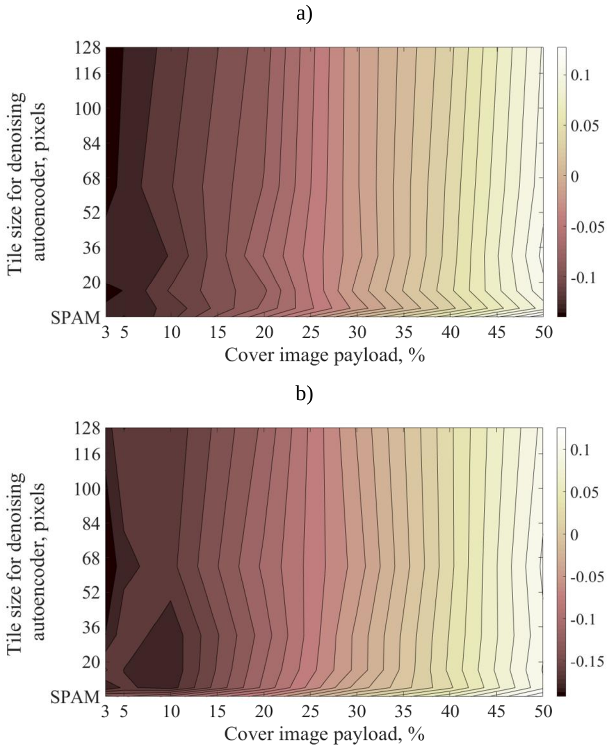


Fig. 5. Dependencies of Matthews correlation coefficient on the cover image payload by usage of $\mathbf{F}_{DF}$ *(a) and* $\mathbf{F}_{CC}$ *(b) features for the adaptive steganographic methods S-UNIWARD by application of DAE and* F_α*=0%.*

We also note that increasing image's tiles used for DAE does not lead to increase of MCC values (Fig. 5). Therefore, we may conclude that training of denoising autoencoders with the size of inputted image bigger than 32x32 (pixels) is redundant due to unnecessary overhead of computation complexity.

For comparison, dependencies of Matthews Correlation Coefficient *MCC* on the cover image payload by the usage of SPAM models and trained DAE network for the MG steganographic method by F_α=0% are shown in Fig. 6.

Note that application of DAE to stego images formed according to MG method leads to changes of *MCC* values (Fig. 6) similar to those obtained earlier for S-UNIWARD method (Fig. 5). The usage of $\mathbf{F}_{\mathrm{CC}}$ features (Fig. 6b) allows to improve Matthews correlation coefficient even higher – up to Δ*MCC*=0.10 for the whole range of cover image payload. This makes DAE-based image calibration an attractive candidate for improving performance of modern SD in the most difficult case of low cover image payload (less than 10%).

DISCUSSIONS

Estimated values of Matthews correlation coefficient for modern image denoising methods proved obtained earlier conclusion about low effectiveness of such methods for image calibration related task. Application of novel DAE allows to improve detection accuracy in comparison with the mentioned case (Table 1).

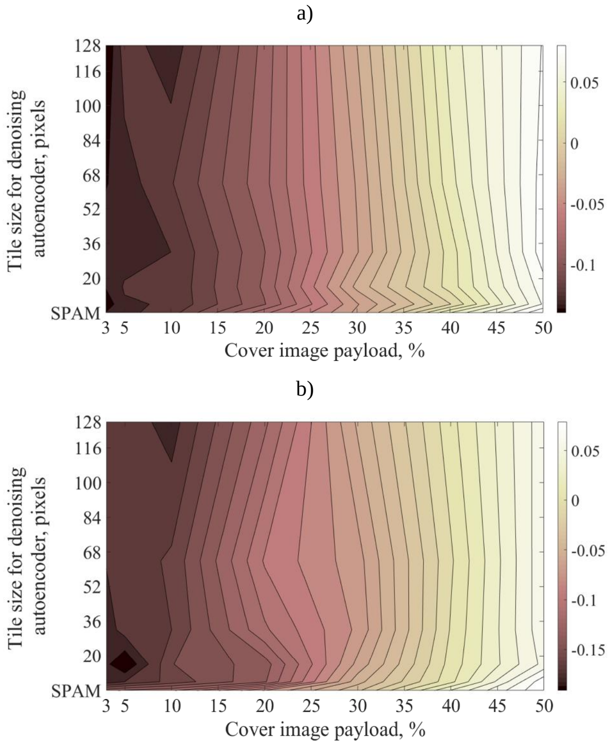


Fig. 6. Dependencies of Matthews correlation coefficient on the cover image payload by usage of **F***$_{DF}$ (a) and* **F***$_{CC}$ (b) features for the adaptive steganographic methods MG by application of DAE and* F$_{\alpha}$*=0%.*

Table 1

Mean and standard deviation of the Matthews correlation coefficients by the usage of bilateral filter (BF), non-local mean filter (NLM) and considered denoising autoencoder (DAE) with the image's tile size of 32x32 (pixels) by variation of cover image payload Δ$_P$. The case of F$_{DF}$ and F$_{CC}$ features and ratio F_A=0% is considered.

Stego images detection method		Cover image payload					
		Δ$_P$=5%		Δ$_P$=20%		Δ$_P$=50%	
		mean	std	mean	std	mean	std
S-UNIWARD embedding method							
SPAM model		-0.130	0.021	-0.056	0.013	0.138	0.013
maxSRMd2 model		-0.289	0.088	-0.178	0.109	0.235	0.021
BF	**F**$_{DF}$	-0.138	0.015	-0.095	0.018	0.069	0.022
	F$_{CC}$	-0.160	0.031	-0.105	0.026	0.069	0.016
NLM	**F**$_{DF}$	-0.120	0.018	-0.070	0.018	0.113	0.017
	F$_{CC}$	-0.162	0.028	-0.097	0.031	0.123	0.013
DAE	**F**$_{DF}$	-0.128	0.009	-0.064	0.013	0.120	0.018
	F$_{CC}$	-0.158	0.043	-0.109	0.013	0.110	0.016
MG embedding method							
SPAM model		-0.125	0.024	-0.072	0.018	0.089	0.015
maxSRMd2 model		-0.295	0.151	-0.314	0.052	0.334	0.028
BF	**F**$_{DF}$	-0.138	0.015	-0.095	0.018	0.069	0.022
	F$_{CC}$	-0.160	0.031	-0.105	0.026	0.069	0.013
NLM	**F**$_{DF}$	-0.122	0.019	-0.120	0.011	-0.016	0.014
	F$_{CC}$	-0.158	0.035	-0.111	0.033	0.069	0.027
DAE	**F**$_{DF}$	-0.126	0.021	-0.088	0.011	0.086	0.018
	F$_{CC}$	-0.165	0.037	-0.107	0.025	0.065	0.023

Image calibration with considered BF and NLM filters leads to decrease of *MCC* values in the whole range of cover image calibration. The decrease achieves up to Δ*MCC*=0.05 for both considered embedding methods in comparison with the standard SPAM model (Tab. 1) that makes these image denoising methods inappropriate for steganalysis related task. On the other hand, usage of DAE allows to improve *MCC* values in comparison with the case of SPAM model usage. Still, achieved improvement may be insufficient for practical usage (Δ*MCC*=0.10) that requires usage of deeper denoising autoencoder model.

CONCLUSION

Development of the advanced methods for image calibration is topical task in digital image steganalysis domain today. The proposed methods for solving this task were based on the application of the novel methods for the advanced image denoising methods. Based on the results of performance analysis for bilateral and non-local mean filtering, we may conclude limitations of practical usage of these methods for stego image calibrations. This can be explained by insufficient "selectivity" of the mentioned methods to detect and suppress local perturbation of pixel brightness caused by stego bits embedding.

We proposed to use denoising autoencoder for overcoming the mentioned limitations. Feature of DAE is ability to learn an appropriate transformation of inputted (noisy) image for restoration of initial (cover) image for a wide range of distortions. However, results of performance evaluation of SD with the usage of such networks showed limited increase of detection accuracy (up to 0.1 for Matthews correlation coefficient). This can be explained by the usage of shallow autoencoder (only 4 hidden layers), so application of deeper network may improve detection accuracy.

Also, it was revealed that an increase was obtained in the whole range of cover image payload that allowed to improve performance of stegdetector without necessity to use a set of calibration methods for each CI payload range.

REFERENCES

[1] J.-P. A. Yaacoub, O. Salman, H. N. Noura, N. Kaaniche, A. Chehab, M. Malli. "Cyber-physical systems security: Limitations, issues and future trends", *Microprocessors and Microsystems*, vol. 77, 2020. [Online]. DOI: 10.1016/j.micpro.2020.103201

[2] D. Legezo. "MontysThree: Industrial espionage with steganography and a Russian accent on both sides", *SecureList*. Available at:: https://securelist.com/montysthree-industrial-espionage/98972/ (Accessed 2022-March-30).

[3] V. Kopeytsev. "Steganograph in attacks on industrial enterprises". Kaspersky Inc., Tech. Rep, 2020. [Online] Available at:https://ics-cert.kaspersky.com/media/KASPER-SKY_Steganography_in_targeted_attacks_EN.pdf (Accessed: 10 November 2021)

[4] J. Fridrich, J. Kodovsky. "Rich models for steganalysis of digital images", *IEEE Transactions on Information Forensics and Security*, vol. 7, iss. 3, 2012, pp. 868-882, DOI 10.1109/TIFS.2012.2190402.

[5] M. Boroumand, M. Chen, J. Fridrich. "Deep Residual Network for Steganalysis of Digital Images", *IEEE Transactions on*

Information Forensics and Security, vol. 14, iss. 5, 2018, pp. 1181-1193. DOI: 10.1109/TIFS.2018.2871749.

[6] J. Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge: Cambridge University Press, 2009, 437 pages, ISBN 978-0-521-19019-0, DOI: 10.1017/CBO9781139192903.

[7] G. Konachovych, D. Progonov, O. Puzyrenko. *Digital steganography processing and analysis of multimedia files.* Kyiv, 'Tsentr uchbovoi literatury' publishing, 2018, 558 pages, ISBN 978-617-673-741-4, Available at: http://pdf.lib.vn-tu.edu.ua/books/2019/Konahovich_2018_558.pdf (Accessed: 17 November 2021).

[8] J. Kodovsky, J. Fridrich. "Calibration revisited", in *Multimedia and security: 11th ACM workshop*, Princeton, 2009, pp. 63-74, DOI: 10.1145/1597817.1597830.

[9] R. Zhang, F. Zhu, J. Liu, and G. Liu, ''Efficient feature learning and multisize image steganalysis based on CNN,'' Jul. 2018, arXiv:1807.11428. [Online]. Available: http://arxiv.org/abs/1807.11428 (Accessed: 10 November 2021)

[10] J. Butora, Y. Yousfi, J. Fridrich. "How to Pretrain for Steganalysis", in *ACM Workshop on Information Hiding and Multimedia Security,* Brussels, Belgium, 2021, pp. 143-148, DOI: 10.1145/3437880.3460395.

[11] D. Progonov. "Influence of digital images preliminary noising on statistical stegdetectors performance", *Radio Electronics, Computer Science, Control*, vol. 1(56), pp. 184-193, 2021, DOI: 10.15588/1607-3274-2021-1-18.

[12] D. Progonov. "Detection Of Stego Images With Adaptively Embedded Data By Component Analysis Methods", *Advances in Cyber-Physical Systems*, Vol. 6, Number 2, pp. 146-154, 2021, DOI: 10.23939/acps2021.02.146.

[13] A. Cohenab, A. Cohena, N. Nissim. "ASSAF: Advanced and Slim StegAnalysis Detection Framework for JPEG images based on deep convolutional denoising autoencoder and Siamese networks", *Neural Networks*, vol. 131, pp. 64-77, Nov. 2020. [Online]. DOI: 10.1016/j.neunet.2020.07.022

[14] T. Filler, J. Fridrich. "Gibbs construction in steganography", *IEEE Transactions on Information Forensics Security*, vol. 5, 2010, pp. 705-720, DOI: 10.1109/TIFS.2010.2077629.

[15] T. Filler, J. Fridrich. "Design of adaptive steganographic schemes for digital images", in *Electronic Imaging, Media Watermarking, Security, and Forensics: The International Society for Optical Engineering,* San Francisco, CA, 2011, DOI: 10.1117/12.872192.

[16] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, J. Fridrich. "Selection-Channel-Aware Rich Model for Steganalysis of Digital Images", in *IEEE Workshop on Information Forensic and Security,* Atlanta, USA, 2014, DOI 10.1109/WIFS.2014.7084302.

[17] V. Holub, J. Fridrich, T. Denemark. "Universal Distortion Function for Steganography in an Arbitrary Domain", *EURASIP Journal on Information Security*, Vol. 1, 2014, DOI: 10.1186/1687-417X-2014-1.

[18] V. Sedighi, J. Fridrich, R. Cogranne. "Content-adaptive pentary steganography using the multivariate generalized gaussian cover model", in *Electronic Imaging, Media Watermarking, Security, and Forensics: The International Society for Optical Engineering,* San Francisco, CA, 2015, DOI: 10.1117/12.2080272.

[19] V. Sedighi, R. Cogranne, J. Fridrich. "Content adaptive steganography by minimizing statistical detectability", *IEEE Transactions on Information Forensics Security,* vol. 11, 2015, pp. 221-234, DOI: 10.1109/TIFS.2015.2486744.

[20] Stan Z. Li. *Markov Random Field Modeling in Image Analysis.* In Advances in Computer Vision and Pattern Recognition Series, Springer, 2009, 362 pages, ISBN 978-1-84800-278-4, Available at: https://link.springer.com/book/10.1007/978-1-84800-279-1 (Accessed: 17 November 2021).

[21] R. Gonzalez, R. Woods. *Digital Image Processing.* 4th ed. Pearson Press, 2017. 1192 pages, ISBN 978-0133356724, Available at: http://sdeuoc.ac.in/sites/default/files/sde_videos/Digital%20 Image%20Processing%2 03rd% 20ed.%20-%20R.%20Gonzalez %2C% 20R.% 20Woods-ilovepdf-compressed.pdf (Accessed: 17 November 2021).

[22] C. Tomasi, R. Manduchi. "Bilateral Filtering for Gray and Color Images." *IEEE International Conference on Computer Vision*, 1998, pp. 839-846. DOI:10.1109/ICCV.1998.710815

[23] Jae S. Lim. *Two-Dimensional Signal and Image Processing*, Englewood Cliffs, NJ, Prentice Hall, 1990, p. 548. ISBN: 978-0139353222

[24] A. Buades. A non-local algorithm for image denoising. *Computer Vision and Pattern Recognition*, 2005. 2. pp. 60–65. DOI:10.1109/CVPR.2005.38.

[25] I. Goodfellow, Y. Bengio, A. Courville. *Deep Learning*, Cambridge: The MIT Press, 2016, p. 800. ISBN: 978-0262035613.

[26] D. Progonov. "Multi-Datasets Evaluation Of GB-Ras Network Based Stegdetectors Robustness To Domain Adaptation Problem". *Information Theories & Applications*. Volume 28, Number 4, 2021. pp. 372-396.

[27] D. Progonov, M. Yarysh. "Analyzing The Accuracy Of Detecting Steganograms Formed By Adaptive Steganographic Methods When Using Artificial Neural Networks", *Eastern-European Journal of Enterprise Technologies*, Vol. 1, Issue 9 (115), pp.45-55, 2022, DOI: 10.15587/1729-4061.2022.251350.

[28] R. Cogranne, Q. Gilboulot, P. Bas. "The alaska steganalysis challenge: A first step towards steganalysis", in *Information Hiding and Multimedia Security*, Paris, 2019, ACM Press, pp. 125-137, DOI: 10.1145/3335203.3335726.

[29] T. Pevny, P. Bas, J. Fridrich. "Steganalysis by subtractive pixel adjacency matrix", *IEEE Transactions on Information Forensics Security*, vol. 5, 2010, pp. 215-224, DOI: 10.1109/TIFS.2010.2045842.

[30] J. Kodovsky, J. Fridrich. "Ensemble classifiers for steganalysis of digital media", *IEEE Transactions on Information Forensics Security,* vol. 7, 2012, p. 432-444, DOI: 10.1109/TIFS.2011.2175919.

[31] D. Progonov. "Performance of Statistical Stegdetectors in Case of Small Number of Stego Images in Training Set", in *IEEE Problems of Infocommunications Science and Technology*, Kharkiv, 2020, DOI:10.1109/PICST51311.2020.9467901.

[32] D. Chicco, G. Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", *BMC Genomics*, vol. 21, 2020. DOI: https://doi.org/10.1186/s12864-019-6413-7.

Dmytro Progonov was born in Kyiv, Ukraine, in 1991. He received the B.S. and M.S. degrees in information protection systems from Kyiv Polytechnic Institute, Kyiv, in 2011 and 2013 respectively. He received the Ph.D. degree in information security from Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, in 2016.

From 2013 to 2017, he was an Assistant with the Physics and Information Security Systems Department, Igor Sikorsky Kyiv Polytechnic Institute. Since 2017, Mr. Progonov has been an Associate Professor at the Physics and Information Security Systems Department, Igor Sikorsky Kyiv Polytechnic Institute, Kyiv. From 2021, he joined the Information Security Department, Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, as Associate Professor.

He is the author of the book in the domain of digital image steganalysis, and more than 15 papers related to digital image forensics. His research interests include digital media forensics, behavior-based person authentication, machine learning and advanced signal processing. He is an Associate Editor of the journal *Information Models & Analyses*, and holds five patents.

Mr. Progonov was a recipient of the President of Ukraine Young Scientist Award in 2018.