DOI 10.20535/2411-1031.2025.13.2.344716
UDC 004.056.54

OLEKSANDR USPENSKYI,
YURII BONDARCHUK

AI-BASED IMAGE STEGANALYSIS UNDER LIMITED COMPUTATIONAL RESOURCES

This study addresses the challenges of modern steganalysis, which lies in the dichotomy between highly effective yet computationally expensive State-of-the-Art (SOTA) artificial intelligence models [1], [2] and lightweight architectures that are fast but incapable of independently detecting weak steganographic signals [3], [4]. The hypothesis proposed in this research suggests that combining classical feature engineering techniques – particularly the use of Spatial Rich Model (SRM) filters to enhance residual noise [5], [6] – with a modern self-supervised learning (SSL) approach for regularization and improved generalization capability [7], [8], can endow a lightweight convolutional neural network with the necessary properties for effective performance.

To verify this hypothesis, a comprehensive comparative experiment was conducted involving four models: a baseline lightweight architecture [3], a model employing SRM filters [6], a heavy SOTA SRNet (Residual Network) model [1], and the proposed hybrid model [9], [10].

The experiment was carried out on a complex heterogeneous dataset comprising images processed by three distinct steganographic algorithms with two embedding rates [11]. Performance evaluation was conducted on two datasets: a test sample from the same data domain (in-distribution) and a completely new, external dataset to assess generalization capability (out-of-distribution) [11], [12].

The experimental results fully confirmed the main hypothesis. The hybrid model achieved the highest detection accuracy among lightweight approaches (AUC – Area Under the ROC Curve of 0.636) and, most importantly, demonstrated the greatest robustness to domain shift (AUC of 0.539 on the external dataset), showing the smallest degradation in performance [10], [13]. The study also revealed a counterintuitive effect: the heavy SOTA SRNet architecture exhibited a significant failure (AUC of 0.348) under heterogeneous data conditions, indicating its tendency to overfit to specific artifacts [1], [2].

Keywords: computer vision, steganographic algorithm, neural network model, performance evaluation, comprehensive experiment.

Problem Statement. Progress in the field of steganography, particularly the development of adaptive algorithms aimed at minimizing statistical distortions in digital images, presents researchers with increasingly complex challenges [8], [12], [14]. In recent years, the dominant approach to addressing these challenges has been the use of deep convolutional neural networks (CNNs), which have demonstrated an unprecedented ability to automatically extract and analyze subtle, almost imperceptible patterns that remain within an image's structure after steganographic embedding [15].

The pinnacle of this approach is represented by specialized State-of-the-Art (SOTA) architectures such as SRNet [3], which achieve exceptionally high detection accuracy under controlled laboratory conditions. However, their outstanding performance comes at the cost of significant computational expense. These models are characterized by tens of millions of parameters and demand substantial computational resources, including powerful server-grade graphics processing units (GPUs) and extensive training times, often lasting several days. This fundamental limitation creates a substantial practical gap between academic achievements and real-world deployment, rendering such models unsuitable for many common scenarios: data analysis on mobile devices, integration into embedded security systems with limited power consumption, or rapid largescale scanning of data archives where speed and computational efficiency are paramount [6]. An alternative approach involves the direct use of lightweight architectures, such as MobileNetV3 [4], which were specifically designed for efficient operation under constrained computational conditions. However, this approach proves to be entirely ineffective for steganalysis. The reason lies in the intrinsic nature of the steganographic signal – it is an extremely weak, high-frequency noise distributed across the entire image, masked by the much stronger content signal. Standard lightweight CNNs, architecturally optimized for recognizing semantically significant features, are incapable of independently isolating and analyzing such subtle statistical anomalies when operating on raw pixel data.

This, a fundamental scientific and practical dilemma arises: finding an optimal compromise between highly accurate but resource-intensive SOTA models and fast yet inefficient lightweight alternatives. This leads to a key research question: is it possible to design an architecture and training methodology for a lightweight steganalysis model that would be computationally efficient while maintaining high accuracy and robustness, particularly the ability to generalize to data from unknown distributions?

The present study hypothesizes that the solution to this problem lies in the hybridization of approaches. It is assumed that the synergy between classical and modern methods can compensate for the inherent limitations of lightweight models. First, the application of classical feature engineering in the form of Spatial Rich Model (SRM) filters [5] can serve as a preprocessing stage, enhancing weak steganographic noise and transforming it into more prominent residual maps that a lightweight CNN can effectively process. Second, the use of modern self-supervised learning (SSL) techniques for pretraining the model on a large corpus of "clean" images can provide it with a fundamental understanding of the statistical structure of natural images. This approach is expected to serve as an effective regularizer, encouraging the model to learn more general anomaly patterns rather than overfitting to specific noise artifacts, thereby significantly improving its reliability and robustness to domain shifts [2], [7].

The objective of this work is to provide a comprehensive evaluation of the effectiveness of the proposed hybrid approach. To achieve this goal, a complex experiment was conducted, encompassing the implementation of four architectures, training on a diverse and challenging dataset, and a two-level performance evaluation, which included a critical generalization test [11]. The entire process was deliberately constrained to the computational capabilities of a typical workstation.

1. Architectures of the examined models

Within the framework of the experiment, a comparative analysis of four architectural approaches was carried out. The first, *Baseline*, represented an unmodified lightweight MobileNetV3-Small architecture, which received raw pixel data as input. This model served as a control group to assess the baseline performance of standard lightweight CNNs. The second approach, *SRM*, utilized the same MobileNetV3-Small backbone [4], but was preceded by a fixed (non-trainable) convolutional layer composed of a set of SRM filters [5]. This model was designed to isolate and evaluate the impact of classical feature engineering. The third approach, *SRNet*, was a full implementation of the State-of-the-Art architecture of the same name [3], [15], serving as a reference benchmark to determine the upper performance limit for this task. Finally, the fourth approach, *srm_ssl*, represented the proposed hybrid architecture, similar to the SRM model, but with its convolutional backbone pretrained using self-supervised learning (SSL) [2], [7] on 10,000 "clean" images.

The mathematical essence of the SRM layer lies in the application of the two-dimensional convolution operation. Each SRM filter is a small kernel matrix f, which is sequentially "slid" across the input image I. At each position (x, y), element-wise multiplication of the kernel values and the corresponding image pixels is performed, followed by summation of the results to form a new value in the feature map. This process is formally expressed by Equation (1):

$$(f * I)(x, y) = \sum_{i,j} f(i, j) \cdot I(x-i, y-j). \tag{1}$$

SRM filter kernels are specifically designed to function as high-pass filters [5]. They mathematically nullify or substantially suppress low-frequency signals corresponding to the main content of the image, while simultaneously enhancing subtle high-frequency noise introduced by steganographic algorithms. Thus, this stage represents a mathematical transformation aimed at improving the signal-to-noise ratio (SNR).

To ensure an unbiased comparison, all models were trained under a unified protocol using advanced process control techniques. Model parameters were optimized using the AdamW optimizer [13], [14] with an initial learning rate of $le-4$. The learning rate was adaptively reduced using a ReduceLROnPlateau scheduler in the event of stagnation in the validation AUC metric over three consecutive epochs. The batch size was set to 32, an empirically determined optimal value for a hardware configuration with 6 GB of VRAM. To address the significant class imbalance $(\sim 1:6)$, a weighted Cross-Entropy Loss function was applied. Training was limited to a maximum of 100 epochs and incorporated early stopping if the validation AUC did not improve over ten consecutive epochs. The key performance metrics were AUC and EER (Equal Error Rate).

The training process consisted of two stages: self-supervised learning (SSL) for the *srm_ssl* model, and supervised learning for all four models. The SSL stage aimed to teach the model fundamental visual patterns without using class labels, employing a contrastive approach with the NT-Xent loss function [8,12]. For each image, two randomly augmented versions were generated and passed through the model to obtain vector representations Z_1 and Z_2. A key operation involved computing the cosine similarity between these vectors and vectors from other images within the batch, representing the angle between vectors u and v in a high-dimensional space (Equation 2):

$$sim(u, v) = \frac{u \cdot v}{|u||v|}. \tag{2}$$

The NT-Xent loss function, a variant of cross-entropy, mathematically encouraged the model to maximize the cosine similarity between vectors of the same image (positive pair) while simultaneously minimizing the similarity with all other vectors in the batch (negative pairs) [7]. The supervised learning stage faced a significant class imbalance $(\sim 1:6)$. To address this issue in a mathematically sound manner, a weighted cross-entropy loss function was applied. The standard binary cross-entropy loss is defined as follows (Equation 3):

$$L = -\left[y \log(p) + (1-y)\log(1-p)\right]. \tag{3}$$

It was modified by introducing weights w, inversely proportional to the frequency of each class (Equation 4):

$$L_{weighted} = -\left[w_1 \cdot y \log(p) + w_0 \cdot (1-y)\log(1-p)\right]. \tag{4}$$

This approach compensated for the imbalance, as the weight w_0 for the rare class was significantly higher. The minimization of this loss function was performed using the AdamW (Adaptive Moment Estimation) optimizer [9,14]. The learning rate was adaptively reduced using the ReduceLROnPlateau scheduler in the event of stagnation in the validation AUC metric over three epochs, and the entire process was subject to early stopping if no improvement was observed over ten consecutive epochs.

To obtain an objective evaluation of model performance, metrics based on Receiver Operating Characteristic (ROC) analysis were employed. AUC (Area Under the Curve) represents the area under the ROC curve. Mathematically, it is defined as the integral of this curve, reflecting the model's overall ability to rank positive instances higher than negative ones. EER (Equal Error Rate) is the point on the ROC curve where the proportion of false negatives equals the proportion of false positives ($FPR = 1 - TPR$).

2. Hardware and Software

The experimental platform was deliberately limited to the resources of a typical personal computer to validate the approaches under conditions approximating real-world scenarios.

All computations were performed on a LENOVO 81Y6 laptop equipped with an Intel Core processor operating at approximately 2.5 GHz, 16 GB of RAM, and an NVIDIA GPU with 6 GB of VRAM.

The software environment was deployed on Microsoft Windows 11 Pro (x64). All models and algorithms were implemented in Python using the PyTorch framework [10] with CUDA support enabled. Data analysis and result preparation were performed using the scikit-learn and pandas libraries.

As the primary dataset for training, validation, and testing, the standard BOSSbase 1.01 dataset [1] was employed. The “clean” subset representing the cover class consisted of 10,000 images. The stego subset, representing the stego class, was expanded to 60,000 images generated using three contemporary spatial steganographic algorithms: WOW, S-UNIWARD, and HILL [8,12,14]. Two embedding payloads were used for each algorithm: low (0.2 bits per pixel) and high (0.4 bits per pixel). This combination of algorithms and payloads created a complex heterogeneous dataset, forcing the models to learn generalizable features. For the final assessment of robustness and generalization capability, the BOWS2 dataset [1], which was entirely unseen during training, was employed. Prior to the experiments, all images from both datasets were converted to grayscale and standardized to a uniform size of 256×256 pixels.

Given the hardware constraints, several software-level optimizations were implemented to improve computational efficiency. To mitigate input/output delays, data caching was performed by pre-processing the entire dataset once and storing it in fast binary files. For parallelism, a DataLoader with four worker processes was employed for asynchronous data preparation on the CPU. On the GPU level, a set of measures was applied, including Mixed Precision Training, enabling TensorFloat-32 mode to accelerate matrix operations, and using cuDNN Benchmark mode for dynamically selecting the fastest convolution algorithms [10].

3. Experimental Procedure

The results obtained from the conducted experiment constitute a comprehensive and multifaceted set of empirical data, enabling an in-depth analysis of the effectiveness of the investigated approaches. The data not only provide clear quantitative confirmation of the proposed hypothesis regarding the advantages of the hybrid method but also reveal several important, sometimes counterintuitive, patterns in the behavior of steganalyzers of varying architectural complexity, particularly under heterogeneous datasets and limited computational resources.

The analysis of these results is carried out along two key directions, corresponding to the two-tier validation system embedded in the experimental design.

First, the final performance of the models on the test sets is considered, allowing for a definitive quantitative assessment of their discriminative capability on both familiar and previously unseen data domains. Second, an in-depth analysis of the learning dynamics of each model is conducted, providing insight not only into the final outcomes but also into the factors leading to them, such as convergence speed, stability, and susceptibility to overfitting. This dual approach enables the formation of a holistic picture, where quantitative metrics are complemented by a qualitative understanding of the behavior of each investigated architecture.

In the following subsections, a sequential presentation of these results is provided. The final evaluation of the models was performed on two separate test datasets. The first test, conducted on a subset of the BOSSbase dataset [1], assessed the models' performance on data from the same domain as the training set (in-distribution performance). The second, critical test on the BOWS2 dataset, evaluated their generalization capability on entirely new, unseen data (out-of-distribution generalization). The results of both tests are presented in Tables 1 and 2.

A deeper understanding of each model's behavior can be obtained by analyzing the dynamics of their training. The graphs below illustrate the changes in the loss function and key metrics on the validation set over the epochs. The experiment was conducted using four models:

- *baseline*;
- *SRNet*;
- *SRM*;
- *srm_ssl*.

Table 1. Final Testing Results on the BOSSbase Dataset (In-Distribution)

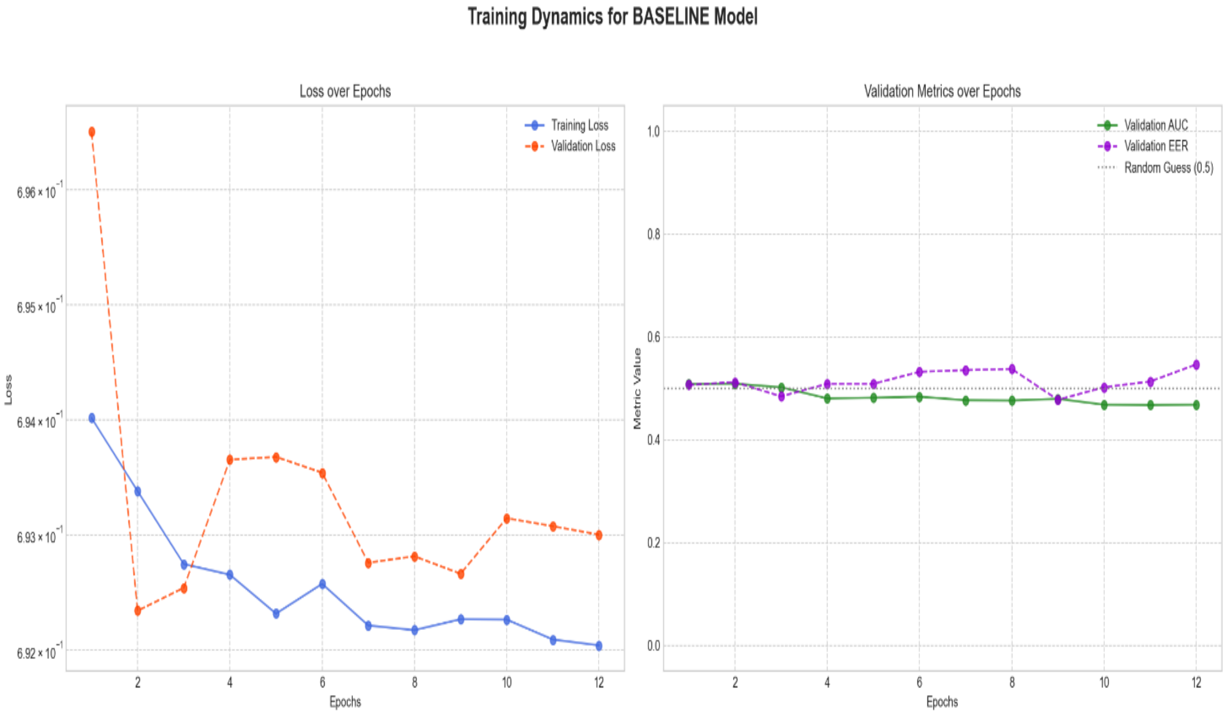
Model	Accuracy	AUC	EER
baseline	0.8554	0.4925	0.5000
SRM	0.7193	0.6300	0.4230
SRNet	0.8509	0.3484	0.6120
srm_ssl	0.6763	0.6355	0.4220

Table 2. Generalization Testing Results on the BOWS2 Dataset (Out-of-Distribution)

Model	Accuracy	AUC	EER
baseline	0.6643	0.5001	0.4891
SRM	0.6098	0.5327	0.4848
SRNet	0.6663	0.5001	0.4817
srm_ssl	0.6048	0.5387	0.4747

The training dynamics of the baseline model (Fig.1) confirm the complete inability of this architecture to extract relevant features from raw data. The validation AUC fluctuates stochastically around 0.5 throughout the entire training process, which is equivalent to random guessing. The model exhibits no signs of convergence toward a solution, and the early stopping mechanism correctly terminates training after 12 epochs. This visually corroborates the initial assumption that standard lightweight CNNs are not suited for directly processing raw pixels for steganography detection.

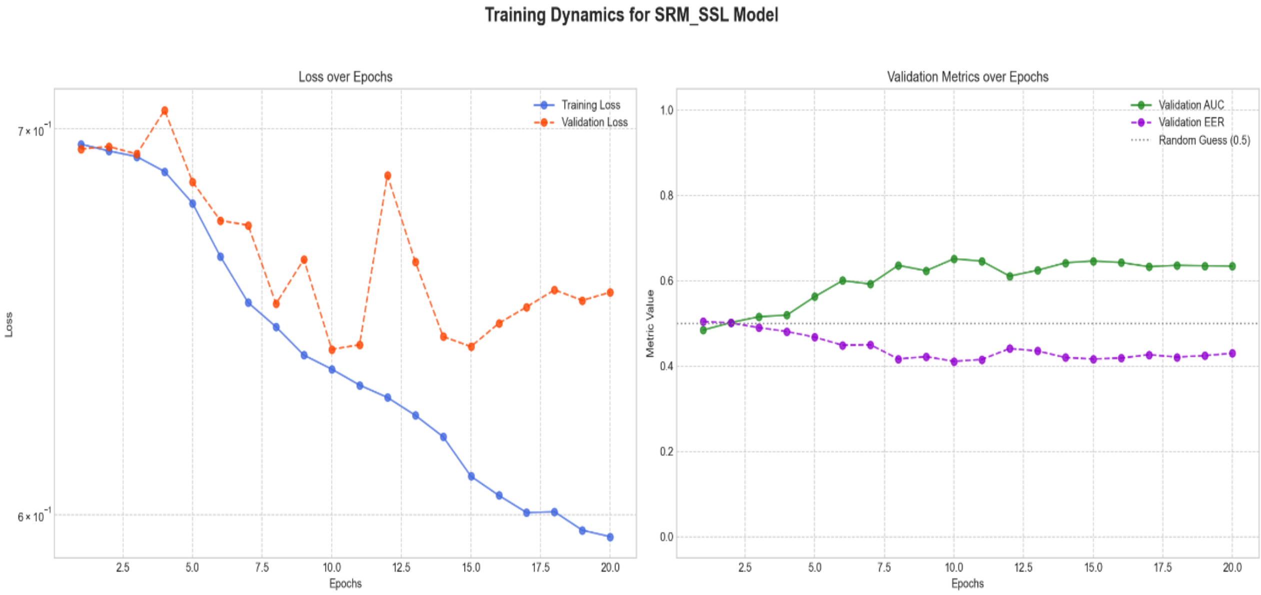
The training dynamics of the *SRNet* model (Fig.2) demonstrate anomalous and counterintuitive behavior. Although the training loss decreases as expected, the validation AUC exhibits persistent degradation: after the first epoch, it rapidly drops well below 0.5. This is a classic indication of catastrophic overfitting, where a highly expressive model learns highly specific and spurious correlations in the training data, causing its predictions on new data to become systematically worse than random. The early stopping mechanism successfully prevented further deterioration of the validation metric.



а)

б)

Figure 1 – Training dynamics of the *baseline* model:
a – loss over epochs; b – validation metrics over epochs

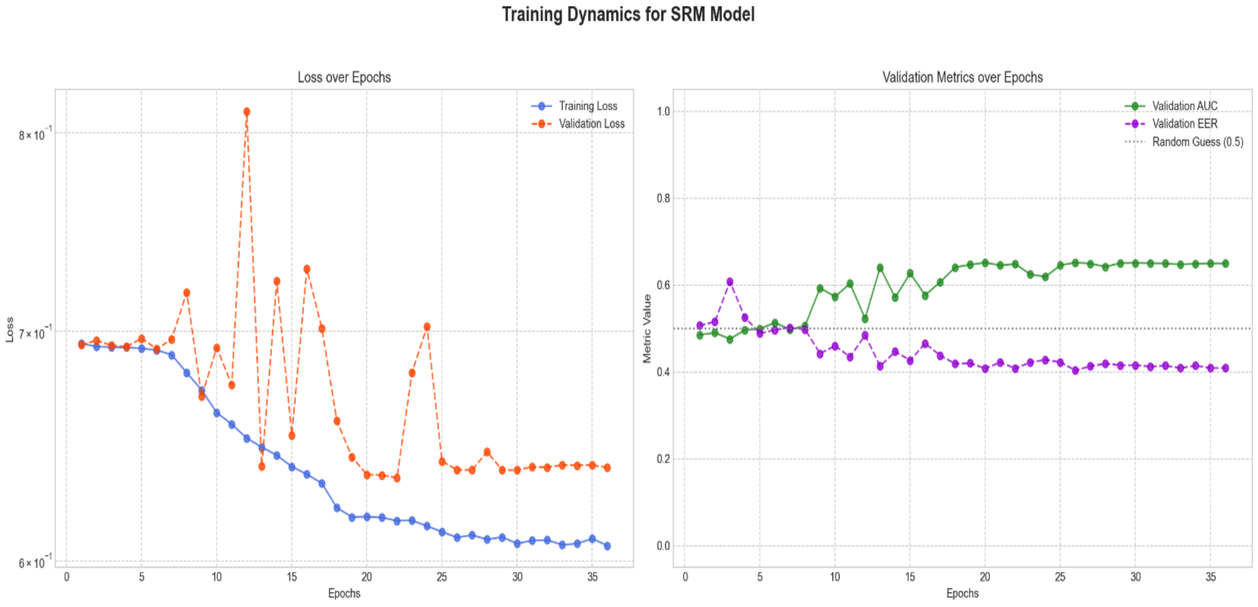


а)

б)

Figure 2 – Training dynamics of the *SRNet* model:
a – loss over epochs; b – validation metrics over epochs

The *SRM* model (Fig. 3), unlike the previous ones, demonstrates a pronounced learning capability. The validation AUC steadily increases during the first 20-25 epochs, reaching a performance plateau with a peak value of approximately 0.65. This indicates that the prior feature extraction using *SRM* filters successfully transformed the input signal, making it suitable for analysis by a lightweight network.



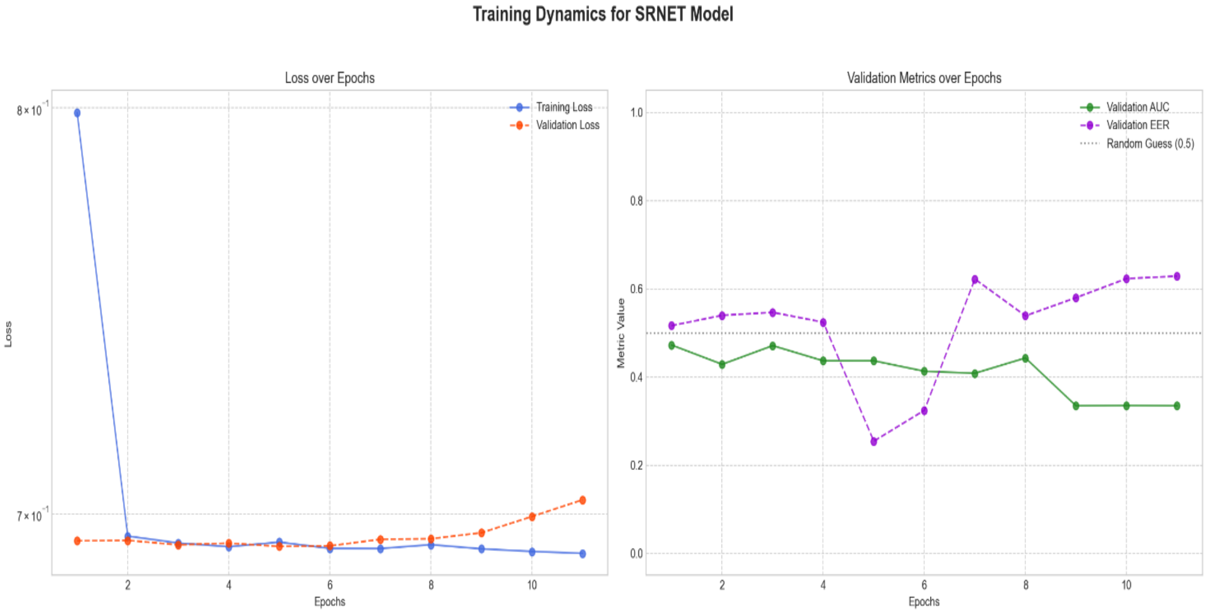
а)

б)

Figure 3 – Training dynamics of the SRM model:
a – loss over epochs; b – validation metrics over epochs

However, after reaching the peak, a phase of stagnation and slight degradation begins, and the validation loss curve becomes highly unstable. This suggests that, although the model has learned, it quickly reached its ceiling and started overfitting to the peculiarities of the training dataset.

The training dynamics of the hybrid *srm_ssl* model initially appear similar to the *SRM* model but exhibit key differences (Fig. 4).



а)

б)

Figure 4 – Training dynamics of the *srm_ssl* model:
a – loss over epochs; b – validation metrics over epochs

First, the validation AUC curve achieves a slightly higher peak value. Second, and most importantly, the training process is significantly more stable. The validation loss curve does not show the sharp stochastic fluctuations observed in the *SRM* model, visually confirming the regularizing effect introduced by prior self-supervised learning [2], [7]. The model demonstrates smoother and more stable convergence, indicating a higher quality of learned features.

4. Analysis of Experimental Results

The conducted experiment provided results that not only quantitatively confirm the proposed hypothesis but also offer a deep qualitative understanding of the processes occurring during the training of steganalysis models of varying complexity.

The hybrid *srm_ssl* model demonstrated the highest overall performance, outperforming all other models in key metrics, namely AUC and EER, across both test datasets. On the “in-domain” BOSSbase dataset, it achieved the best discriminative capability, albeit with a marginal advantage over the *SRM* model. However, its primary strength was revealed during the generalization test. When evaluated on the unseen BOWS2 dataset, the drop in performance (difference in AUC) for the *srm_ssl* model was the smallest among all models capable of learning. This serves as direct experimental evidence that prior self-supervised learning (SSL) acts as a powerful regularizer [2], [13]. It provides the model with fundamental knowledge of the structure of natural images, forcing it to focus on more general, invariant anomalies rather than “memorizing” the specific artifacts of six different types of noise present in the training set. Consequently, this results in higher model robustness against domain shifts, a critical property for any practical security tool, especially one intended for deployment on mid-range hardware.

A notable observation concerns the unsatisfactory performance of the SOTA *SRNet* architecture. Its AUC of 0.348, significantly worse than random guessing, requires careful explanation. The likely cause lies in the combination of two factors: the high expressive capacity of the architecture itself and the heterogeneity of the training data. *SRNet* is specifically designed to capture the smallest anomalies. In our experiment, where the training set consisted of a mixture of six different types of steganographic noise, the model likely overfitted to the unique, specific artifacts of each noise type instead of extracting a generalizable steganography feature. When these noises were mixed in the validation set, the narrowly specialized patterns learned by the model began to conflict, resulting in chaotic and systematically incorrect predictions. Unlike *srm_ssl*, *SRNet* was trained “from scratch” without any prior knowledge of what natural images look like, making it fully vulnerable to this type of “noise overfitting”. This outcome strongly supports the notion that, in complex and realistic conditions, blindly applying the most powerful architectures can be counterproductive.

Conclusions. This work has demonstrated and experimentally confirmed the high effectiveness of a hybrid approach for developing lightweight steganalysis models, specifically adapted to operate under constrained computational resources. It was shown that while feature engineering using *SRM* filters is a necessary step for extracting the steganographic signal, the key element ensuring model reliability and robustness is prior self-supervised learning (SSL). SSL provides the model with fundamental knowledge of natural images, acting as a powerful regularizer and significantly enhancing its ability to generalize to new, unseen data.

From a practical standpoint, this study offers a ready-to-use and validated methodology for creating efficient lightweight steganalysis tools suitable for deployment across a wide range of hardware, not just specialized servers. In the scientific context, the research emphasizes that generalization capability is as critical a performance criterion as peak accuracy on ideal datasets. It also highlights the crucial role of SSL pretraining as a method to improve neural network robustness against domain shifts, a relevant issue not only in steganalysis but in numerous other areas of computer vision.

REFERENCES

[1] G. Xu et al, "SFRNet: Feature Extraction-Fusion Steganalysis Network", *Security and Communication Networks*, vol. 2021, art. 3676720, 11 p. 2021. doi: https://doi.org/10.1155/2021/3676720.

[2] H. Kheddar, M. Hemis, Y. Himeur, D. Megías, and A. Amira, "Deep learning for steganalysis of diverse data types: review of methods, taxonomy, challenges and future directions", *Neurocomputing*, vol. 581, iss. C, art. 127528, 2024. doi: https://doi.org/10.1016/j.neucom.2024.127528.

[3] E. Hong, K. Lim, T.-W. Oh, and H. Jang, "Lightweight image steganalysis with block-wise pruning", *Scientific Reports*, vol. 13, art. 16148, 2023. doi: https://doi.org/10.1038/s41598-023-43386-2.

[4] S. Hong, et al. "Author Correction: Lightweight image steganalysis with block-wise pruning", *Scientific Reports*, vol. 13, art. 17300, 2023. doi: https://doi.org/10.1038/s41598-023-44614-5.

[5] S. Liu, C. Zhang, L. Wang, P. Yang, S. Hua, and T. Zhang, "Image Steganalysis of Low Embedding Rate Based on the Attention Mechanism and Transfer Learning", *Electronics*, vol. 12 (4), art. 0969, 2023. doi: https://doi.org/10.3390/electronics12040969.

[6] F. Liu, X. Zhou, X. Yan, Y. Lu, and S. Wang, "Image Steganalysis via Diverse Filters and Squeeze-and-Excitation Convolutional Neural Network", *Mathematics*, vol. 9 (2), art. 189, 2021. doi: https://doi.org/10.3390/math9020189.

[7] J. Liu, F. Xu, Y. Zhao, X. Xin, K. Liu, and Y. Ma, "Sterilization of image steganography using self-supervised convolutional neural network (SS-Net)", *PeerJ Computer Science*, vol. 10, art. e23302024. doi: https://doi.org/10.7717/peerj-cs.2330.

[8] W. Guo, "Dilated Separable Convolution Network for Image Steganalysis", in *Proc. 2024 Int. Conf. on Image Proc., Mult. Tech. and ML (IPMML'24)*, Dali Henan, China, 2024, pp. 43-46. doi: https://doi.org/10.1145/3722405.3722413.

[9] S. Mekruksavanich, and A. Jitpattanakul Hybrid convolutional architectures and channel attention studies for detection tasks. *Scientific Reports*, vol. 13, art. 12067, 2023. doi: https://doi.org/10.1038/s41598-023-39080-y.

[10] L. Qiushi, L. Shenghai, T. Shunquan, and L. Zhenjun, "SEAP: Squeeze-and-Excitation Attention Guided Pruning for Image Steganalysis Networks", *EURASIP Journal on Information Security*, art. 25, 2025. doi: https://doi.org/10.1186/s13635-025-00212-8.

[11] N.J. De La Croix, T. Ahmad, and F. Han, "Comprehensive survey on image steganalysis using deep learning", *Array*, art. 100353, 2024. doi: https://doi.org/10.1016/j.array.2024.100353.

[12] S. Agarwal, and K.-H. Jung, "Digital image steganalysis using entropy driven deep neural network", *Jour. of Inf. Se. and App.*, vol. 84, art. 103799, 2024. doi: https://doi.org/10.1016/j.jisa.2024.103799.

[13] T. Fu, L. Chen, Z. Fu, K. Yu, and Y. Wang, "CCNet: CNN model with channel attention and convolutional pooling mechanism for spatial image steganalysis", *Jour. of Vis. Comm. and Image Repres.*, vol. 88, art. 103633, 2022. doi: https://doi.org/10.1016/j.jvcir.2022.103633.

[14] L. Bohang et al. "Image steganalysis using active learning and hyperparameter optimization", *Scientific Reports*, vol. 15, art. 7340, 2025. doi: https://doi.org/10.1038/s41598-025-92082-w.

[15] Z. Fu et al. “Adaptive, Dilated and Hybrid Techniques for JPEG and Domain-Aware Steganalysis”, *Signal Processing*, vol. 216, art. 109299, 2024. doi: https://doi.org/10.1016/j.sigpro.2023.109299.

The article was received 10.09.2025.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] G. Xu et al, “SFRNet: Feature Extraction-Fusion Steganalysis Network”, *Security and Communication Networks*, vol. 2021, art. 3676720, 11 p. 2021. doi: https://doi.org/10.1155/2021/3676720.

[2] H. Kheddar, M. Hemis, Y. Himeur, D. Megías, and A. Amira, “Deep learning for steganalysis of diverse data types: review of methods, taxonomy, challenges and future directions”, *Neurocomputing*, vol. 581, iss. C, art. 127528, 2024. doi: https://doi.org/10.1016/j.neucom.2024.127528.

[3] E. Hong, K. Lim, T.-W. Oh, and H. Jang, “Lightweight image steganalysis with block-wise pruning”, *Scientific Reports*, vol. 13, art. 16148, 2023. doi: https://doi.org/10.1038/s41598-023-43386-2.

[4] S. Hong, et al. “Author Correction: Lightweight image steganalysis with block-wise pruning”, *Scientific Reports*, vol. 13, art. 17300, 2023. doi: https://doi.org/10.1038/s41598-023-44614-5.

[5] S. Liu, C. Zhang, L. Wang, P. Yang, S. Hua, and T. Zhang, “Image Steganalysis of Low Embedding Rate Based on the Attention Mechanism and Transfer Learning”, *Electronics*, vol. 12 (4), art. 0969, 2023. doi: https://doi.org/10.3390/electronics12040969.

[6] F. Liu, X. Zhou, X. Yan, Y. Lu, and S. Wang, “Image Steganalysis via Diverse Filters and Squeeze-and-Excitation Convolutional Neural Network”, *Mathematics*, vol. 9 (2), art. 189, 2021. doi: https://doi.org/10.3390/math9020189.

[7] J. Liu, F. Xu, Y. Zhao, X. Xin, K. Liu, and Y. Ma, “Sterilization of image steganography using self-supervised convolutional neural network (SS-Net)”, *PeerJ Computer Science*, vol. 10, art. e23302024. doi: https://doi.org/10.7717/peerj-cs.2330.

[8] W. Guo, “Dilated Separable Convolution Network for Image Steganalysis”, in *Proc. 2024 Int. Conf. on Image Proc., Mult. Tech. and ML (IPMML'24)*, Dali Henan, China, 2024, pp. 43-46. doi: https://doi.org/10.1145/3722405.3722413.

[9] S. Mekruksavanich, and A. Jitpattanakul Hybrid convolutional architectures and channel attention studies for detection tasks. *Scientific Reports*, vol. 13, art. 12067, 2023. doi: https://doi.org/10.1038/s41598-023-39080-y.

[10] L. Qiushi, L. Shenghai, T. Shunquan, and L. Zhenjun, “SEAP: Squeeze-and-Excitation Attention Guided Pruning for Image Steganalysis Networks”, *EURASIP Journal on Information Security*, art. 25, 2025. doi: https://doi.org/10.1186/s13635-025-00212-8.

[11] N.J. De La Croix, T. Ahmad, and F. Han, “Comprehensive survey on image steganalysis using deep learning”, *Array*, art. 100353, 2024. doi: https://doi.org/10.1016/j.array.2024.100353.

[12] S. Agarwal, and K.-H. Jung, “Digital image steganalysis using entropy driven deep neural network”, *Jour. of Inf. Se. and App.*, vol. 84, art. 103799, 2024. doi: https://doi.org/10.1016/j.jisa.2024.103799.

[13] T. Fu, L. Chen, Z. Fu, K. Yu, and Y. Wang, “CCNet: CNN model with channel attention and convolutional pooling mechanism for spatial image steganalysis”, *Jour. of Vis. Comm. and Image Repres.*, vol. 88, art. 103633, 2022. doi: https://doi.org/10.1016/j.jvcir.2022.103633.

[14] L. Bohang et al. “Image steganalysis using active learning and hyperparameter optimization”, *Scientific Reports*, vol. 15, art. 7340, 2025. doi: https://doi.org/10.1038/s41598-025-92082-w.

[15] Z. Fu et al. “Adaptive, Dilated and Hybrid Techniques for JPEG and Domain-Aware Steganalysis”, *Signal Processing*, vol. 216, art. 109299, 2024. doi: https://doi.org/10.1016/j.sigpro.2023.109299.

ОЛЕКСАНДР УСПЕНСЬКИЙ,
ЮРІЙ БОНДАРЧУК

СТЕГОАНАЛІЗ ЗОБРАЖЕНЬ НА ОСНОВІ ШІ ЗА УМОВ ОБМЕЖЕНИХ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ

У даному дослідженні розглядається проблематика сучасного стегоаналізу, що полягає у дихотомії між високоефективними, але обчислювально затратними State-of-the-Art (SOTA) моделями штучного інтелекту [1], [2], та легковаговими архітектурами, які є швидкими, але нездатними самостійно виявляти слабкі стеганографічні сигнали [3], [4]. Було висунуто гіпотезу, що поєднання класичної інженерії ознак, зокрема використання фільтрів Spatial Rich Model (SRM) для підсилення залишкових шумів [5], [6], із сучасним методом самонавчання (Self-Supervised Learning, SSL) для регуляризації та покращення здатності до узагальнення [7], [8], може наділити легковагову згорткову нейронну мережу необхідними властивостями для ефективної роботи. Для перевірки було проведено комплексний порівняльний експеримент за участі чотирьох моделей: базової легковагової архітектури [3], моделі з SRM-фільтрами [6], важкої SOTA-архітектури SRNet (Residual Network) моделі [1], та запропонованої гібридної моделі [10], [13]. Експеримент проводився на складному гетерогенному наборі даних, що включав зображення, оброблені трьома різними стеганографічними алгоритмами з двома рівнями навантаження [11]. Оцінка ефективності здійснювалася на двох наборах даних: на тестовій вибірці з того ж домену даних (in-distribution) та на абсолютно новому, сторонньому датасеті для перевірки здатності до узагальнення (out-of-distribution) [11], [12]. Результати експерименту повністю підтвердили запропоновану гіпотезу.

Ключові слова: комп'ютерний зір, стеганографічний алгоритм, модель нейронної мережі, оцінка ефективності, комплексний експеримент.

Oleksandr Uspenskyi, candidate of technical sciences, associate professor, associate professor at the computer science and artificial intelligence technologies in the field of cybersecurity academic department, Institute of special communication and information protection at the National technical university of Ukraine “Igor Sikorsky Kyiv polytechnic institute”, Kyiv, Ukraine, ORCID 0000-0001-6953-421X, uspensky@ukr.net.

Bondarchuk Yurii, cadet, master's degree student, Institute of special communication and information protection at the National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, ORCID 0009-0002-3198-5087, ri01.bondarchuck.yuri@gmail.com.

Успенський Олександр Анатолійович, кандидат технічних наук, доцент, доцент кафедри комп'ютерних наук та технологій штучного інтелекту у сфері кібербезпеки, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

Бондарчук Юрій Михайлович, курсант, магістрант, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.