Hasilin, D. L., & Zhuravel, I. M. (2025). Studying the resilience of steganography algorithms to detection by neural networks. *Actual Issues of Modern Science. European Scientific e-Journal, 40*, 87–95. Ostrava.

DOI: 10.47451/tec2025-10-01

The paper is published in Crossref, ICI Copernicus, BASE, Zenodo, OpenAIRE, LORY, Academic Resource Index ResearchBib, J-Gate, ISI International Scientific Indexing, ADL, JournalsPedia, Scilit, EBSCO, Mendeley, and WebArchive databases.

Dmitro L. Hasilin, Ph.D. Student, Department of Information Technology Security, National University "Lviv Polytechnic". Lviv, Ukraine.
E-mail: d.hasilin@ukr.net
ORCID 0009-0000-4636-7666
Ihor M. Zhuravel, Doctor of Engineering Sciences, Senior Researcher, Department Head, Department of Information Technology Security, National University "Lviv Polytechnic". Lviv, Ukraine.
E-mail: izhuravel@ukr.net
ORCID 0000-0003-1114-0124, Scopus 35423018200

Article history:
Received: October 15, 2026
Revised: November 21, 2026
Accepted: December 1, 2026
Published: December 14, 2026

Studying the Resilience of Steganography Algorithms to Detection by Neural Networks

Abstract:
This study analyzes the resilience of steganography algorithms used in the spatial domain against detection using steganalysis deep learning models. We evaluate the impact of geometric transformations on the robustness against detection. The study investigates the effect of image transformations on the embedded message integrity and, more importantly, on the detectability of the algorithms by different models. Proven CNN models, as EfficientNet and SRNet with a Convolutional Block Attention Module, were used to compare detectability after different transformations. As expected, the embedded messages were significantly corrupted, while the models were still able to identify their presence. Introducing noise before embedding reduced detectability and increased robustness of steganographic algorithms. Alongside common LSB and modern S-UNIWARD, evaluating a spatial color space embedding algorithm, previously unknown to the model, significantly decreased accuracy. Due to the complex impact of the Color Space algorithm on uniform areas and smooth color transitions, we observe a higher number of false positives after fine-tuning models. Nevertheless, accuracy and generalization were increased to expected levels consistent with another modern research. The results underscore the need to focus on developing models that can withstand real-world image alterations, as well as improve detection capabilities to keep up with unseen stenographic methods. In practice, this will help practitioners select detection models best suited for operational environments and support future advancements in detection model development and design.

Keywords: steganography, steganalysis, color space, deep learning models.

Abbreviations:

BER is Bit Error Rate,
$ByER$ is Byte Error Rate,
LSB is Least Significant Bit,
$CBAM$ is Convolutional Block Attention Module,
$CNNs$ are Convolutional Neural Networks,
$S-UNIWARD$ is Spatial Universal Wavelet Relative Distortion,
STC is Syndrome Trellis Code.

Introduction

The rising frequency of digital data exchange highlights the demand for reliable steganalysis methods to mitigate potential threats. High data volumes further increase the risk of unsolicited data ingestion. To enhance the security of data pipelines against attacks, there is a necessity to explore the resilience of embedding algorithms not only against attacks but also against detection. This work continues and extends our earlier research (*Hasilin & Zhuravel, 2025*) about the level of resilience to transformations. While trying to add a new dimension to comparison by analyzing robustness to detection. The range of CNN models under consideration was narrowed based on the findings of Alrusaini (*2025*). The two best-performing models identified in that study were used as a starting point, with further exploration aimed at extending support for additional environments—particularly in detecting a broader range of color space–based methods.

The subject of the study is the resilience of spatial-domain steganography algorithms—specifically LSB, S-UNIWARD, and color-space methods—to detection by deep learning models such as EfficientNet and SRNet with Convolutional Block Attention Modules.

The object of the study is the process of information concealment and detection within digital images, particularly the interaction between steganographic algorithms and neural-network-based steganalysis models.

The study aims to evaluate and compare the robustness of contemporary steganographic algorithms against neural network detection, focusing on the effects of image transformations, dataset diversity, and model fine-tuning on detectability and message integrity.

To achieve this aim, the study pursues the following objectives:

- analyze existing approaches to steganography and steganalysis in the context of deep learning development;
- assess the resilience of spatial-domain and colour-space steganographic methods under geometric and noise-based image transformations;
- compare the detection performance of different CNN architectures, including EfficientNet and SRNet-CBAM, in identifying hidden messages;
- examine the influence of dataset composition and fine-tuning on detection robustness and false-positive rates;
- identify optimal conditions for achieving a balance between detection resistance and message integrity in real-world scenarios.

Methods

Steganalysis is the process of detecting hidden messages within digital media. Initially, it relied on statistical analysis, which examined anomalies in pixel distribution. However, these methods were not effective against modern steganographic techniques.

The field evolved significantly with the introduction of deep learning models. Early CNN-based models, such as Xu-Net and Yedroudj-Net, automated the feature extraction process but were vulnerable to common image transformations like compression and resizing.

Later, more advanced architectures emerged to address these vulnerabilities. ResNet introduced residual connections, which help preserve deep feature representations and prevent information loss within the network. This makes it more robust for steganographic detection. EfficientNet uses a compound scaling method to optimize its depth, width, and resolution simultaneously. It also incorporates squeeze-and-excitation SE layers to dynamically adjust feature importance. These features provide exceptional resilience against image transformations, enabling the model to maintain detection accuracy even after an image has been altered (*Alrusaini, 2025*).

In the research, a subset of images from various well-known datasets was used to cover as many variations as possible. The de facto standard is the BOSSBase dataset, which is a grayscale image dataset specifically designed for steganography and steganalysis research. We did not consider it, as one of the methods under research relies on color space (*Margalikas & Ramanauskaitė, 2019*).

COCO, which stands for Common Objects in Context, is a diverse dataset designed for object detection, classification, and context understanding. It includes more than 300,000 images with complex, real-world scenes and multiple labeled objects. The images are in color and vary in resolution. In steganography, COCO is often used because of its variety and realistic content (*Lin et al., 2014*).

ImageNet is an image database organized according to the WordNet hierarchy (*Deng et al., 2009*). It contains over 10 million labeled images covering thousands of object classes. The images are in color and come from diverse sources. In steganography research, subsets of ImageNet are used for training deep neural networks to detect embedded data, taking advantage of the dataset's scale (*Deng et al., 2009*).

DIV2K is a high-quality image dataset primarily used in image SR tasks. It consists of 1,000 images with 2K resolution, meaning they are very detailed and suitable for tasks requiring high visual fidelity. In the context of steganography, DIV2K is used to test how well embedding algorithms perform on high-resolution, high-quality images (*Timofte et al., 2017*).

We've used an equal number of images, 10,000, from the first two datasets mentioned above, to cover them equally. The entire DIV2k dataset was added to make the initial image resolution even more versatile. The initial training dataset, for fine-tuning, was created by scaling all images to the required size, and 10% of the images included embedding using the LSB method.

Literature Review

The growing sophistication of steganography and steganalysis reflects the ongoing competition between concealment and detection technologies. Early research in this field was dominated by statistical methods focused on pixel distribution irregularities (*AbdelRaouf, 2021*).

However, such approaches lost relevance with the emergence of adaptive steganographic algorithms capable of mimicking natural image statistics. The introduction of *Least Significant Bit* (LSB) substitution marked a significant milestone in spatial-domain steganography but simultaneously exposed fundamental vulnerabilities to compression and transformation (*Apau et al., 2024*). Subsequent developments such as S-UNIWARD, which embed information in textured or noisy areas, demonstrated greater resistance to visual and statistical detection by adapting to local image characteristics (*Holub et al., 2014*).

The modern phase of research is increasingly defined by the use of deep learning in steganalysis. CNNs such as Xu-Net and Yedroudj-Net automated the feature extraction process, outperforming classical detectors based on handcrafted features (*Chaumont, 2020*). Nevertheless, these models proved sensitive to geometric transformations, including scaling, rotation, and compression (*Hasilin & Zhuravel, 2025*). To address such weaknesses, architectures such as ResNet and EfficientNet introduced residual connections and compound scaling to improve model stability and accuracy (*Tan & Le, 2019*). Recent studies demonstrated that models enhanced with channel attention mechanisms, such as the CBAM, achieve superior generalisation across diverse embedding methods (*Woo et al., 2018*).

Datasets play a crucial role in training and evaluating steganalysis algorithms. Benchmarks such as ImageNet (*Deng et al., 2009*) and COCO (*Lin et al., 2014*) offer complex, high-resolution colour imagery necessary for learning robust features across diverse contexts. The DIV2K dataset (*Timofte et al., 2017*) provides additional value by introducing high-fidelity images suitable for testing robustness against compression and resizing. However, as noted by Margalikas and Ramanauskaitė (*2019*), traditional grayscale datasets such as BOSSBase may fail to reflect real-world image diversity, which is critical for assessing the detection resistance of colour-space-based algorithms.

The shift towards embedding in colour space marks one of the latest conceptual innovations in steganography. This approach modifies chromatic rather than luminance components, distributing hidden information more uniformly and reducing statistical detectability (*Margalikas & Ramanauskaitė, 2019*). Moreover, using three-dimensional colour cube models enhances resistance to cropping and noise attacks by maintaining global chromatic consistency (*Hasilin & Zhuravel, 2025*). While these methods improve robustness, they often introduce capacity trade-offs and increased computational complexity.

Recent comparative analyses reveal that the detectability of modern steganographic algorithms strongly depends on both embedding strategy and the training diversity of detection models (*Alrusaini, 2025*). Studies emphasise the importance of fine-tuning CNNs on mixed datasets, including unseen embedding methods, to improve detection reliability under transformations. Furthermore, research by Hu, Ni, and Shi (*2018*) confirmed that adaptiveness in embedding entropy domains enhances resilience to JPEG compression, aligning with broader findings on robustness through domain transformation. Complementary experiments show that introducing controlled noise during training can reduce model overfitting and improve generalisation to previously unknown embedding techniques (*Ni et al., 2014*).

In summary, the evolution of steganography from simple spatial substitution to adaptive, colour-space, and deep-learning-resistant methods represents a critical frontier in information security. Simultaneously, the parallel development of detection models—from handcrafted to

CNN-based architectures—illustrates the dynamic co-evolution of attack and defence strategies. The reviewed literature highlights that ensuring message integrity and concealment in real-world conditions requires not only algorithmic innovation but also dataset diversity, model robustness, and cross-domain evaluation frameworks.

Embedding Methods

The most common spatial domain hiding algorithm, the LSB modification, works by altering the least significant bit of a pixel's color component. This change is typically visually imperceptible due to the bit's minimal influence. However, the LSB method is highly vulnerable to various attacks, as even minimal changes in bit sequences can introduce detectable noise, making it easy to expose through statistical analysis (*AbdelRaouf, 2021*).

S-UNIWARD is a content-adaptive steganography algorithm designed to embed secret data into the spatial domain of an image by minimizing detectability. It calculates distortion using relative changes in wavelet coefficients across multiple scales and orientations, allowing it to adapt to local image features. The algorithm embeds data in textured or noisy areas where changes are less noticeable, while avoiding smooth regions. It modifies pixel values directly and often, but not always, uses efficient coding methods like STCs to reduce overall distortion, enhancing its security against detection by reducing the number of modified pixels almost in half compared to LSB (*Holub et al., 2014*).

A contemporary algorithm by Margalikas and Ramanauskaitė (*2019*) shifts data hiding from individual pixels to the color space (the set of colors used in an image). This method gains popularity because current statistical analysis tools often neglect the entire color set. Changes in the color space distribute uniformly across gradients and statistical characteristics, making the container less vulnerable to compression and other image-altering attacks. Furthermore, the color set is highly resistant to pixel cropping, as the loss of individual pixels does not necessarily reduce the overall set of colors.

The text examines an adaptive color space steganography method proposed by Margalikas and Ramanauskaitė (*2019*), which employs an RGB-cube to encode messages. This process involves recursively partitioning the RGB-cube, which represents the image's colors in three-dimensional space, into progressively smaller sub-cubes until each contains at most a single color. Data is then embedded by altering the coordinates (R, G, or B values) of colors within sub-cubes of different sizes that contain exactly one color. To mitigate the known drawbacks of increased color space distortion in shallow areas, we employed a simple yet costly in terms of message capacity modification by limiting the sub-cube size $2 \times 2 \times 2$ (*Hasilin & Zhuravel, 2025*).

Since not all embedding methods support block-based data representation, it is necessary to implement it in the message encoder/decoder. The commonly used BER is effective but sensitive to minor distortions. To address this, ByER is used (*Hasilin & Zhuravel, 2025*). Unlike BER, ByER operates on bytes, making it more resilient to single-bit losses and false-positive matches. It better reflects the amount of information recoverable post-attack and helps assess the robustness of embedding digital keys, images. For instance, QR codes offer strong resistance to modifications.

To reduce the impact of attacks on text data and to conform to established attack resilience evaluation frameworks, a bit repetition coding method was also reused, where each bit is

repeated n times. This allows error correction through a quorum-based approach, ensuring readability even when pixels are lost. However, this method significantly increases data size, making it inefficient for low-resource systems. Moreover, it mainly detects missing bits and may misinterpret certain patterns as valid transitions, thus limiting error correction accuracy (*Hasilin & Zhuravel, 2025*). Another alternative would be to use Reed and Solomon (1960) codes in scenarios where robust error correction is required and the loss of even a few bits would be unacceptable.

Results

For model selection, we relied on the evaluation framework suggested by Alrusaini (*2025*), which assessed model robustness against transformations for methods embedding in the spatial domain. The aim was to keep the detection properties closely compared together with robustness to transformation.

The analysis reveals a practical trade-off between maximizing detection sensitivity in pristine images and ensuring robustness against common image manipulations. EfficientNet emerged as the most practical choice for real-world scenarios, effectively balancing speed, optimized performance, and the ability to adapt to post-embedding transformations. In contrast, the high complexity of SRNet and the inherent fragility of Xu-Net and Yedroudj-Net under significant image distortion pose practical limitations. Transformer models (e.g., ViT, Swin) were deliberately excluded due to high computational overhead and unproven utility in this specific domain (*Alrusaini, 2025*).

Evaluation was implemented using the PyTorch framework (*Paszke et al., 2019*). All models were used with pretrained weights. As part of the preparation of the generic EfficientNet (*Tan & Le, 2019*) model, fine-tuning was done. In the configuration phase, a classification layer was added to configure this generic model with a channel attention module in stego detection mode.

Fine-tuning was done on the train dataset without including color space embedding at first, to test the pretrained model's ability to detect an unseen embedding method. An SRNet model was also used with pretrained weights. As reported in several studies, this model failed to converge when trained on color images. We observed that deviations exceeded expectations. As an alternative, a modification with an added CBAM was used. CBAM was proposed by Woo, Park, Lee, and Kweon (*2018*), and it showed better results.

Because of the different nature of the two embedding methods, we set a goal to adjust the embedding message length to fit the image capacity. This configuration was done to achieve a common embeddment rate equal to 0.4 bpp. In Figure 1 (*Appendix*), we can see that the LSB method, at least not an adaptive one, is not dependent on the pixel values. On the other hand, Color Space method, by design, is highly reliant on the color space distribution. After adding the CBAM module, we observe average performance commonly expected from such models. Levels of false positive cases were back to normal, and specifically low on the known dataset, as could be seen in Figure 2 (*Appendix*).

Introducing noise before embedding reduced detectability by an average of 8%; nevertheless, deviations were quite high, and it is hard to call this reduction consistent and reliable. Further discoveries with a deeper focus on noise sounds like a promising area. After proceeding with a few experiments, replicating attacks in the previous study, which included

compression, cropping, and noise addition, we could conclude that the overall performance of detecting models is falling by up to 10%. That is happening with minimal modification and is consistent with known data (*Alrusaini, 2025*). At the same time, message integrity degradation was high, and neither method was able to withstand those modifications without losing integrity. The decision was made not to proceed with the modification, as there is no sense in verifying the accuracy of detection on an embedded message that is already unextractable.

Alternatively, further experiments show that running detection models on previously unseen embodiment methods reveals accuracy degradation, as could be observed in Figure 3 (*Appendix*).

An additional fine-tuning dataset was made to compare with those results. In the new set, a third of the images with a hidden message were created using the Color space method. The overall number of embedded messages was the same, 20%. The pipeline for this data set configuration is described in Figure 4 (*Appendix*).

Thus, we estimated attack effectiveness. Also, fine-tuning improvements results highlight the importance of training set diversity, even if we are talking about training the last few layers.

Discussion

For a generic method, the detection chance is quite important, together with the effect of the transformation. Nevertheless, we need to keep in consideration the stability of the stegochannel, as there is no value in a method successfully withstanding detection whilst being altered, but losing the ability to deliver the message due to later alteration. The reverse situation doesn't bring much value as well: delivering content to the receiver, acknowledging the fact of being detected, may deliver some value at the moment, but not in the long term. As it is only a matter of time and the competence of the detecting entity, before the transmitted message will be disclosed. The essence of steganographic methods is to guarantee the invisibility of the communication transaction. As soon as this promise is broken, even before some messages are decoded, the entire channel is compromised and cannot be used. At any minute, the technical channel could be destroyed or blocked, and it is required to move to a recovery plan if present in the global system setup.

Conclusion

This study evaluated the robustness and detectability of three different steganography methods: LSB modification, S-UNIWARDS, and color space cube modification. We compared these methods against a deep learning model's ability to detect them under attacks. The goal was to evaluate protection against detection under attacks as a new criterion for selecting steganography algorithms, alongside traditional metrics.

Considering the significant statistical effect, we observe that detectability is very high, and therefore, adaptive methods should be used (*Ni et al., 2014*; *Hu et al., 2018*).

In terms of detection robustness, there is around an 11% advantage compared to LSB. Improvement over S-UNIWARD is less significant at 4-5% which could be explained by the adaptiveness of the algorithm. Overall lower detection accuracy of the SRNet model could be explained by its lightweight design in terms of resource consumption. This was evaluated on a model not trained for this use case, showing the main advantage of using not widely known,

unseen in training sets methods. After the second run and allocating 7% of the training set to the color space transformation cases, we can see that the models extracted a proper feature set, and the advantage reduced to a smaller 6–7% over LSB with high deviations, especially for SRNet. This could be explained by the fact that color space alterations are affecting statistical properties softly, unharmed for gradients. Even so, those changes are detectable by CNN models given training. And deviations could be explained by the nature of the cover image. If it has a lot of unique pixel values, any color space modification begins to look like embedding using adaptive LSB, thus similar to S-UNIWARD, but in case of a high number of unique pixel values, it could degrade to lower than S-UNIWARD values. So, here we have an inverse proportional relationship between the number of unique colors and robustness against detection. To ensure the high robustness of such a stego channel and reduce deviation impact, cover images should be carefully selected and tested. Preference should be given to smooth images without redundancy of high contrast areas. As observed, such covers will provide a smaller capacity, but could leverage the higher detection robustness evaluated in this paper. Such preprocessing verification, using automated tooling, is widely adopted in many areas and should not be a burden to such complex communication channels.

Funding

No external funding was received.

Conflict of Interest

The author declares that there is no conflict of interest.

Acknowledgements:

Not applicable.

References:

AbdelRaouf, A. (2021). A new data hiding approach for image steganography based on visual color sensitivity. *Multimedia Tools and Applications, 80*. https://doi.org/10.1007/s11042-020-10224-w

Alrusaini, O. A. (2025). Deep learning for steganalysis: Evaluating model robustness against image transformations. *Frontiers in Artificial Intelligence*, 8, 1532895. https://doi.org/10.3389/frai.2025.1532895

Apau, R., Asante, M., Twum, F., Ben Hayfron-Acquah, J., & Peasah, K. O. (2024). Image steganography techniques for resisting statistical steganalysis attacks: A systematic literature review. *PLOS ONE, 19*(9), e0308807. https://doi.org/10.1371/journal.pone.0308807

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Chaumont, M. (2020). *Deep learning in steganography and steganalysis*. In M. Hassaballah (Ed.), *Digital media steganography: Principles, algorithms, advances* (pp. 321–349). Elsevier. https://doi.org/10.1016/B978-0-12-819438-6.00022-0

Clark, A. (2015). *Pillow (PIL fork) documentation*. Read the Docs. https://pillow.readthedocs.io/

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Hasilin, D. L., & Zhuravel, I. M. (2025). *Analysis of the robustness of steganographic algorithms against geometric attacks and compression*. *Modern Information Protection*, *(2)*. https://doi.org/10.31673/2409-7292.2025.020965

Holub, V., Fridrich, J., & Denemark, T. (2014). Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, *2014*(1), 1–13. https://doi.org/10.1186/1687-417X-2014-1

Hu, X., Ni, J., & Shi, Y. Q. (2018). Efficient JPEG steganography using domain transformation of embedding entropy. *IEEE Signal Processing Letters, 25(6)*, 859–863. https://doi.org/10.1109/LSP.2018.2818674

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Margalikas, E., & Ramanauskaitė, S. (2019). Image steganography based on color palette transformation in color space. *EURASIP Journal on Image and Video Processing, 2019*(1), 84. https://doi.org/10.1186/s13640-019-0484-x

Ni, J., Hu, X., & Shi, Y. Q. (2014). Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security, 2014(1)*, 1. https://doi.org/10.1186/1687-417X-2014-1

Paszke, A., Gross, S., Massa, F. et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*, 8024–8035.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 6105–6114. PMLR. https://arxiv.org/abs/1905.11946

Timofte, R., Agustsson, E., Van Gool, L., Yang, M., & Zhang, L. (2017). NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. *arXiv preprint*. arXiv:1708.02567. https://arxiv.org/abs/1708.02567

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *arXiv*. https://doi.org/10.48550/arXiv.1807.06521

Appendix

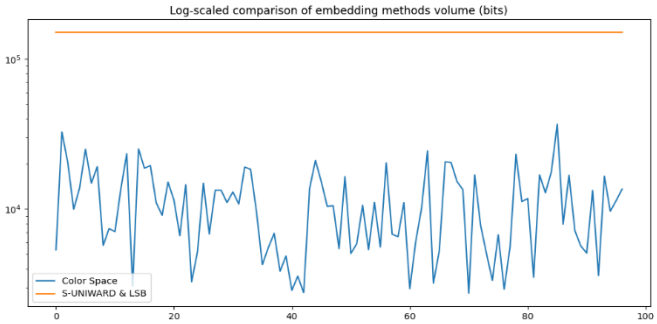

Figure 1. Embedding capacity deviation of all methods on the resulting training set (scaled)

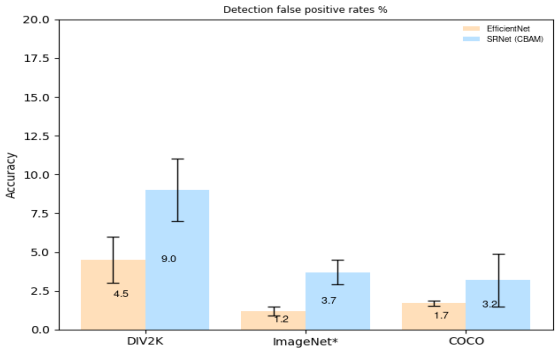

Figure 2. False positive results

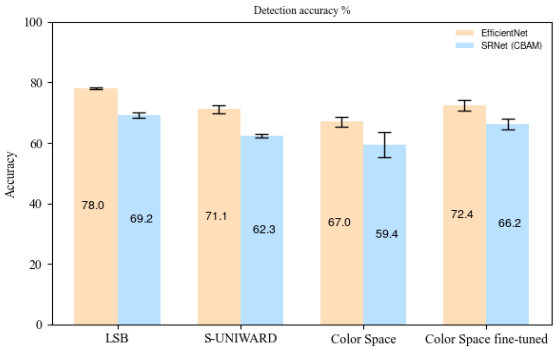

Figure 3. Model accuracy results

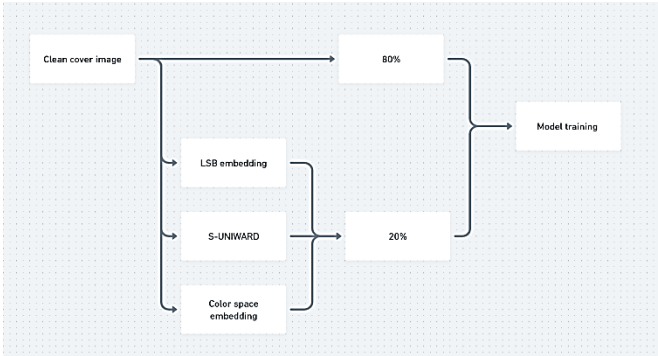

Figure 4. Building fine-tuning dataset flow