IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 11, No. 5, May, 2024, pp. 1227-1235
Journal homepage: http://iieta.org/journals/mmep

Detecting Hidden Data in Images Using Convolutional Neural Networks

Erick Delenia[1], Yoggy Harisusilo Putra[1], Bayu Aditya Triwibowo[1], Ntivuguruzwa Jean De La Croix[1,2], Tohari Ahmad[1*]

[1] Department of Informatics, Institut Teknologi Sepuluh Nopember, Kampus ITS, Surabaya 60111, Indonesia
[2] African Center of Excellence in Internet of Things, College of Science and Technology, University of Rwanda, Kigali 3900, Rwanda

Corresponding Author Email: tohari@its.ac.id

https://doi.org/10.18280/mmep.110511

Received: 19 December 2023
Revised: 27 February 2024
Accepted: 8 March 2024
Available online: 30 May 2024

Keywords:
CNN, Deep Learning, information security, national security, network infrastructure, steganalysis

ABSTRACT

Recent advancements in Deep Learning (DL) have driven the development of innovative methodologies, particularly within the domain of steganalysis for spatial domain images. Steganalysis, as the counterpart to steganography, is dedicated to uncovering concealed data within the content, making a digital image. Convolutional Neural Networks (CNNs), grounded in DL principles, have been influential in pushing the boundaries of this field. Despite the development of various CNN architectures that have raised the precision in detecting images with steganographic payload, current models contend with challenges related to the detectability of low payload capacities and suboptimal processes for feature learning. In response, this study introduces a novel CNN architecture to enhance steganalysis and improve the accuracy of detecting covert data in spatial domain images. The proposed model introduces a strategic integration of maximum and average pooling, a tandem approach meticulously designed to amplify the network's proficiency in capturing intricate details and multiple layers of information. Moreover, the proposed CNN architecture is structured into three principal stages: preprocessing, feature extraction, and classification. The preprocessing stage comprises Input, regular convolution layer, and Batch Normalization. The feature extraction stage employs the ReLU as a non-linear activation function based on its capacity to expedite computation by bypassing the need for exponentials and divisions. The classification stage introduces the multi-scale inception module to enhance the probabilistic feature classification. The proposed model's correctness in probabilistic classification through the receiver operating characteristic curve (ROC AUC) yields an AUC of 0.95, reflecting a prediction correctness of 95%. Furthermore, the results show that the proposed model outperforms the results of previous research studies in terms of accuracy and improves the existing works with a percentage ranging from 2.3 to 2.9%.

1. INTRODUCTION

In the contemporary digital landscape, ensuring secure data transmission through public networks for confidential communications is of utmost importance. Secure data transmission in public networks is crucial for protecting sensitive information from unauthorized access and manipulation. Data hiding protocols ensure the confidentiality of data, preventing eavesdropping, while mechanisms for authentication and integrity verification safeguard against impersonation and tampering. The reliability of communication channels is enhanced, contributing to business continuity and fostering customer trust [1]. Implementing robust security measures is essential to navigate the shared and potentially vulnerable space of public networks. Data hiding (DH) is a strategy employed to obscure sensitive information within non-sensitive data, enhancing the difficulty for unauthorized individuals to identify or access concealed content. This technique encompasses the integration or encryption of confidential data into seemingly innocuous files or structures. The significance of data hiding lies in its ability to bolster the security of sensitive information, providing an additional layer of protection against unauthorized access and ensuring the confidentiality of the hidden data [2, 3]. Common methodologies involve steganography, where information is discreetly embedded within various media [4], each serving as a potential carrier for covertly transmitting sensitive data. Applying data hiding techniques is crucial for organizations aiming to safeguard critical information, uphold privacy, and mitigate the risks associated with data breaches or unauthorized disclosures. The overarching goal of DH remains consistent: the conservation of the quality of the modified digital object (known as stego) in a way that allows it to traverse insecure channels without compromising the integrity of the concealed information [5].

The widespread use of DH techniques, while having beneficial applications, has inadvertently created a potential avenue for the covert transmission of malicious data, leading

to tangible threats to societal security through executing illicit plans [6]. Steganalysis has been developed as a critical countermeasure in response to this growing challenge. Steganalysis aims to preserve the inherent integrity of digital media by actively addressing and preventing the nefarious misuse of data-hiding methodologies [7]. To address this complex issue effectively, contemporary steganalysis methods emphasize both blind steganalysis [7, 8] and locative steganalysis [9, 10], strategically dealing with different aspects of detecting and locating potential steganographic payloads hidden within digital images. The increasing risks associated with data hiding highlight the crucial need for effective steganalysis practices, as they play a pivotal role in identifying and mitigating potential threats arising from secret transmissions of malicious content, thereby safeguarding the integrity of digital communication channels and societal well-being.

Within the expansive field of steganalysis, the crucial role of machine learning (ML) models is evident, employing classical ML techniques such as support vector machines, linear regression, principal component analysis, nearest neighbor, and K-means clustering [11]. However, conventional ML methods face challenges in effectively handling the growing volume of data, mainly due to the separation of feature extraction and classification phases. Deep learning (DL) has emerged as a powerful and adaptable solution to address these challenges, seamlessly integrating feature extraction and classification into an end-to-end learning process [12]. Convolutional Neural Networks (CNNs) within the Deep Learning paradigm are widely recognized for their ability to discern essential features without human supervision [13]. Despite the extensive use of CNNs in steganalysis, there is a discernible need for improvement. The role of ML and DL, particularly the advantages offered by CNNs, is pivotal in advancing steganalysis techniques to effectively address evolving challenges and cope with the increasing complexity of data.

This paper introduces a new CNN architecture, drawing inspiration from cutting-edge research, with the primary goal of significantly advancing the accuracy of detecting hidden data within spatial domain images. The proposed multifaceted enhancements incorporate innovative strategies to elevate the model's capabilities. These include the implementation of mix pooling to reduce spatial dimensions judiciously, the integration of multiple separable convolution layers to enhance feature extraction, the adoption of dropouts to effectively mitigate overfitting, and the incorporation of multi-scale layers facilitated by inception modules, as visually represented in Figure 1. The proposed method is meticulously evaluated through a comprehensive experimental examination to substantiate and underscore its notable outperformance over existing approaches. This research marks a substantial contribution to the field, offering a refined and advanced approach to hidden data detection within spatial domain images, with implications for diverse applications and furthering the state-of-the-art in CNN-based steganalysis techniques.

The rest of the paper consists of related work reported in Section 2, methodology detailed in Section 3, experimental results presented in Section 4, and conclusions drawn in Section 5. The related section presents the general framework of image steganalysis and related work in spatial domain image steganalysis. The methodology section presents the step-by-step process of the proposed method; the experimental results section explains the results of this research and comparisons with previous related research. The overall work is then summarized in the conclusion section.

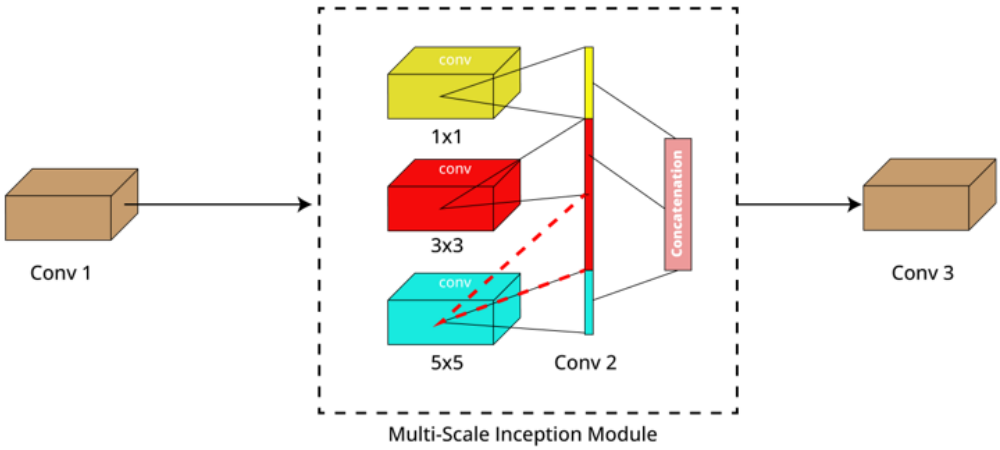


Figure 1. Schematic of the multi-scale inception module

2. RELATED WORKS

Within this section, an in-depth exploration unfolds, delving into the rational framework that underlies the research. A meticulous analysis of related studies, methodologies, and findings is undertaken to provide a nuanced understanding of the research context. The following plot provides an extensive review of antecedent studies suitable to the research objectives.

In 2016, the steganalysis model Xu-Net [14] introduced a novel CNN architecture tailored for steganalysis, demonstrating a profound grasp of steganalysis dynamics. Notable characteristics of this architecture include the computation of absolute values for elements within feature maps generated from the initial convolutional layer, the imposition of constraints on data value ranges through saturation regions defined by a hyperbolic tangent (tanh) in the early stages of the network, and a reduction in modeling strength through the incorporation of 1×1 convolutions in deeper layers. Furthermore, the model employs a hybrid of TanH and Rectified Linear Unit (ReLU) non-linear activation functions in the designed CNN for steganalysis. The discernment from their experimentation suggests that the

utilization of TanH in specific groups, as opposed to ReLU, results in enhanced performance, ostensibly attributed to the effective confinement of data value ranges within the saturation regions of TanH. However, substituting more ReLUs in deeper layers with TanH may lead to suboptimal outcomes, potentially linked to challenges in gradient backpropagation associated with TanH. The model implementation revealed that Xu-Net achieved an accuracy of 72.7% for S-UNIWARD at 0.4 bits per pixel (bpp). These findings demonstrate that Xu-Net presents a sophisticated CNN architecture for steganalysis, showcasing improved performance with nuanced activation function selections. However, it is essential to critically evaluate certain drawbacks, such as the potential suboptimal outcomes associated with the TanH activation function in deeper layers, which warrant further exploration for comprehensive model enhancement.

In the subsequent year, 2017, the research outlined by Ye et al. [15] introduced an innovative CNN architecture, departing from conventional practices by employing a set of high pass filters derived from Spatial Rich Models (SRM) residual maps. This difference from the standard use of high pass filters [7] is grounded in the non-randomized initialization of these filters with SRM values. Particularly noteworthy is the integration a novel Truncated Linear Unit (TLU) activation function designed to enhance the Signal-to-Noise Ratio (SNR), particularly in scenarios of low SNR during the steganography insertion process. Moreover, the improved performance is attributed to incorporating knowledge regarding channel selection or potential pixel alterations [16]. Significantly, a notable enhancement in performance is observed by implementing transfer learning from a network trained on datasets with high payload to a network trained on images with low payload. This innovative architecture achieved an accuracy of 68.7% for S-UNIWARD at 0.4 bpp. Based on these findings, it is remarkable that Ye et al. [15] introduced an unconventional yet effective CNN architecture, emphasizing non-randomized filter initialization from SRM residual maps and introducing a novel TLU activation function to improve SNR. Applying transfer learning from high to low payload datasets further enhances steganalysis performance.

In the following year, 2018, the emergence of the innovative CNN architecture, Yedroudj-Net, drew inspiration from prior research, offering a comprehensive structure across three distinct stages: preprocessing, feature extraction, and classification [17]. Notably, 30 fundamental high pass filters derived from SRM, known for their non-trainable nature and fixed 5×5 size, were employed in the preprocessing phase. Transitioning to the feature extraction stage, Yedroudj-Net comprised five blocks, each incorporating a convolution layer, batch normalization, and selective use of average pooling. Activation functions applied in these blocks included Truncation and ReLU in blocks 1, 2, and 3-5, respectively, with a significant shift in kernel size from 5×5 in blocks 1 and 2 to 3×3 in blocks 3 to 5. The subsequent classification stage utilized three fully connected layers housing 256, 1024, and 2 neurons in the final layer corresponding to the output classes. The SoftMax activation function was applied to generate a distribution of two class labels. This architectural configuration resulted in an accuracy of 72.7% for S-UNIWARD at 0.4 bpp, establishing Yedroudj-Net as a noteworthy contribution to the field.

Additionally, in 2018, the study presented by Boroumand et al. [18] introduced a CNN architecture operating within spatial and frequency domains, specifically targeting the JPEG domain. The study emphasized the efficacy of the frequency domain in eliminating the need for manual and device heuristics required by other networks to detect steganographic noise. The network utilized filter banks inspired by SRM to initialize weights in the preprocessing layer. Subsequent adjustments to these weights during training were made to amplify noise introduced by the steganography algorithm while minimizing its impact on image content. The resultant SR-Net achieved an accuracy of 81.3% for S-UNIWARD at 0.4 bpp, underscoring its effectiveness.

In 2020, the steganalysis research detailed by Zhang et al. [19] unveiled Zhu-Net, drawing inspiration from the architecture of Yedroudj-Net. Distinguishing itself, Zhu-Net incorporates the ReLU activation function after each convolution layer, strategically employed to expedite network convergence. This architecture stands out by leveraging shortcuts and the Xception module to augment overall performance, deviating from Yedroudj-Net's choice to abstain from using shortcuts or the inception module. Notably, Zhu-Net's feature extraction stage adopts a pair of separable convolutions, signaling a departure from conventional approaches. In this context, Zhu-Net showcases substantial advancements compared to existing CNN-based networks. Despite the experimental results demonstrating a promising accuracy improvement, reaching 89%, the model reveals a notable drawback related to the complexity of the CNN and a considerable demand for an extensive dataset to enhance classification correctness. This underscores the need for critical evaluation and areas for improvement in addressing these challenges for the continued advancement of steganalysis using CNN architectures.

In 2021, GBRAS-Net was introduced as a steganalysis model by Reinel et al. [20], taking inspiration from Zhu-Net. The architecture of GBRAS-Net is organized into three distinct stages: preprocessing, feature extraction, and classification. In the classification stage, GBRAS-Net adopts an HPF bank filter for kernel initiation. Diverging from Zhu-Net in the feature extraction stage, GBRAS-Net combines separable and depth-wise convolution with skip connections. The subsequent classification stage incorporates global average pooling as input before transitioning to the SoftMax function for prediction. Empirical findings highlight GBRAS-Net's superior performance compared to previously proposed CNNs. Despite achieving a promising accuracy improvement of around 90%, the model reveals a notable drawback related to the complexity of the CNN and a substantial need for an extensive dataset to enhance classification correctness. This underscores the necessity for thorough critical evaluation and identifies key areas for improvement, contributing to the ongoing progress in steganalysis through CNN architectures.

In the futuristic realm of 2023, a recently introduced CNN architecture, as described by Ntivuguruzwa et al. [7], has put forth an innovative design, building upon preceding research to enhance the proficiency of CNNs in uncovering concealed data within spatial domain images. During preprocessing, the authors incorporated 30 SRM filters, deriving 25 from 3×3 kernels and the remaining five from 5×5 kernels. Furthermore, in this stage, the proposed method employed the tangent hyperbolic as a non-linear operator, intensifying non-linearity to optimize the efficiency of the deep network and foster improved network convergence. To expedite training time, the kernel values were intentionally set as non-trainable. Transitioning to the feature extraction stage, a unique combination of depth-wise separable convolution with regular

convolution layers was implemented, capitalizing on the advantages of increased model expressiveness, reduced storage size, and the segregation of channel correlations. Each regular convolution layer utilized the LeakyReLu activation function to enhance non-linearity, complemented by the batch normalization layer for normalizing the distribution of feature maps. The strategic use of LeakyReLu prevented vanishing gradients by converting negative values into small positive ones, facilitating backward communication and ensuring positive model weights. The rapid convergence achieved with LeakyReLu contributed to swift network convergence and heightened training stability. In the subsequent classification stage, multi-scale average pooling was integrated to retain spatial information from the preceding layer, thereby augmenting feature expression. Three sequentially arranged dense layers followed this. The output from the last fully connected layer was then directed to the SoftMax layer to transform generated features into probabilistic classes, ultimately yielding class labels. The multi-scale average pooling featured a 3-scale pyramid pool with sizes (4,4), (2,2), and (1,1). Despite achieving promising accuracy improvements of around 90%, surpassing previously proposed models, the model exhibits a noteworthy drawback that may lead to overfitting, given the imperative need for enhancing the model's feature learning ability. This emphasizes the crucial need for comprehensive assessment and pinpointing key areas requiring enhancement, ultimately playing a role in the continual advancement of steganalysis through CNN architectures.

Building upon existing research, this study introduces a novel CNN designed to enhance the detection hidden data within spatial domain images. The core of these enhancements involves the integration of mix pooling, dropout, and multi-scale convolution, strategically applied to bolster the stability of the classification stage. However, a comprehensive examination of the current literature highlights notable shortcomings, such as low detection rates, suboptimal outcomes, and increased complexity. Leveraging insights derived from the recognized limitations in previous studies, this paper addresses these challenges by proposing an innovative model. The objective is to mitigate complexity and classification accuracy concerns by introducing a cutting-edge steganalysis approach. This includes integrating proficient feature selection and optimization techniques, marking a significant advancement in response to identified deficiencies in the existing body of literature.

3. METHODOLOGY

This section explains the proposed methodology starting from the dataset, the proposed CNN architecture, and compares the proposed architecture with related research discussed in the related work session. In essence, this section serves as a comprehensive exposition of the methodology, commencing with the foundational dataset, traversing through the details of the CNN architecture, and culminating in a meticulous comparative analysis with relevant research.

3.1 Dataset

The dataset employed for training and testing the proposed CNN model involves resizing the original cover images, generating stego images using specific steganographic algorithms, and subsequently partitioning the dataset into training, validation, and testing sets while ensuring a balanced distribution between stego and cover images. The dataset utilized for training and testing the proposed CNN model is sourced from the Break Our Steganographic System (BOSSBase 1.01) database [21], which is publicly accessible and comprises Portable Gray Map (PGM) images with 8-bit grayscale. The BOSSBase database encompasses 10,000 cover images, each with a 512×512 pixels resolution. The principal characteristics and preprocessing steps applied to the dataset are described as follows:

1. Resizing cover image size:
   - The initial cover images, sized 512×512 pixels, undergo resizing to 256×256 pixels. This resizing step reduces computational complexity in subsequent operations while preserving reasonable image quality.
2. Steganographic images generation:
   - Post-resizing, stego images are generated using two distinct steganographic algorithms: Wavelet Obtained Weights (WOW) and Spatial Universal Wavelet Relative Distortion (S-UNIWARD).
   - The steganographic process involves embedding information with a payload of 0.4 bpp.
   - This results in two separate stego image datasets, one for WOW and another for S-UNIWARD, containing 10,000 images.
3. Data distribution:
   - Following stego image generation, three datasets are formed: 10,000 cover images, 10,000 stego images from the WOW algorithm, and 10,000 stego images from the S-UNIWARD algorithm.
   - Two distinct models are constructed, each utilizing a different dataset: Model 1 incorporates Cover Images + Stego Images from WOW. In contrast, Model 2 employs the dataset containing Cover Images + Stego Images from S-UNIWARD.
   - The total dataset of 20,000 images is partitioned into three subsets:
     - 8,000 images for training
     - 2,000 images for validation
     - 10,000 images for testing
   - The data split maintains a 50:50 proportion between stego and cover images across all subsets.

3.2 Hyper-parameters selection

The rationale behind this training methodology's chosen hyperparameters and optimization techniques is rooted in a strategic approach to bolster the model's learning capacity, expedite convergence, and enhance overall performance. The decision to employ Conv2D and SeparableConv2D convolutional layers with the Glorot kernel initializer is driven by the acknowledged efficacy of Glorot initialization in mitigating the vanishing/exploding gradient problem, thereby fostering stable and efficient learning. Within the Batch Normalization layer, the deliberately selected parameters, including a modest momentum of 0.2, an epsilon value of 0.001, and a renorm momentum set at 0.4, collectively aim to expedite adaptation to data distributions, mitigate overfitting risks, and ensure consistent and reliable updates during training. Moving to the optimization strategy, the adoption of the Adam optimizer is motivated by its adaptive learning rate capabilities and proven effectiveness in handling sparse

gradients. A learning rate of 0.001 strikes a balance between convergence speed and precision, while an epsilon value of 1e-08 safeguards against division by zero issues, ensuring numerical stability throughout the optimization process.

3.3 CNN architecture

The architecture delineated in this research is visually represented in Figure 2. It is structured into three principal stages: preprocessing, feature extraction, and classification. The ensuing discourse provides a detailed elucidation of each of these three stages.

3.3.1 Preprocessing

The preprocessing stage comprises Input, regular convolution layer, and Batch Normalization. The model's input is configured with dimensions (256,256,1). Subsequently, the input undergoes processing in the regular convolution layer, where a 30 SRM filter bank with a 5×5 kernel is applied. The utilization of this filter bank is aimed at enabling the model to capture a more diverse range of texture information and image features. The convolutional operation employs a stride of (1,1), and padding is set to 'same' to ensure that the output from preprocessing maintains the same size as the input. In the preprocessing layer, the chosen activation function is 3TanH, an extension of the TanH function. TanH itself produces output within the range of -1 to 1. The output of the TanH activation function is defined as illustrated in Eq. (1).

$$f(x) = \left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}\right) \tag{1}$$

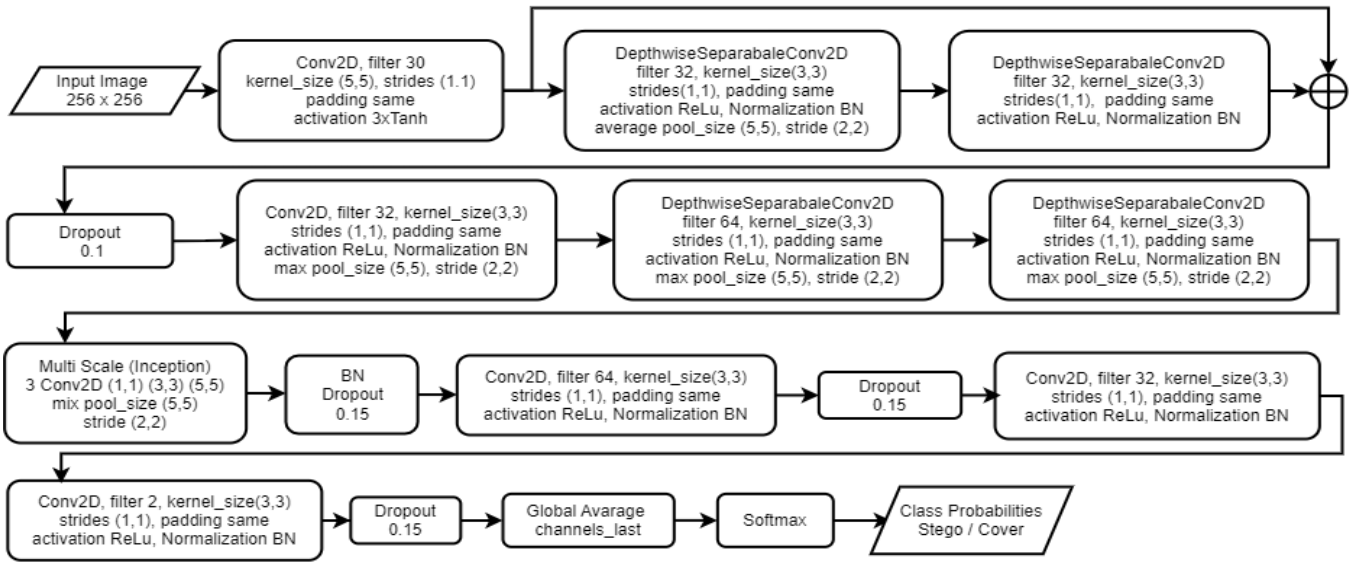


Figure 2. Architecture of the proposed CNN

The primary advantage conferred by this function lies in its ability to generate a zero-centered output, thereby facilitating the backpropagation process [22, 23]. In this preprocessing layer, the weights are designated as non-trainable, signifying that throughout the training phase, these weights remain static and are not subject to updates or evaluations. After the non-trainable weights, batch normalization is employed. Batch Normalization exerts a positive influence on gradient flow throughout the network. It accomplishes this by mitigating the gradient's dependence on the scale of the parameter or its initial value. This reduction in dependence enables the utilization of higher learning rates without the associated risk of divergence.

3.3.2 Feature extraction

Diverging from the preprocessing stage, the feature extraction stage employs the non-linear activation function ReLU. ReLU is preferred for its capacity to expedite computation by bypassing the need for exponentials and divisions, thereby enhancing overall computational speed [22]. The mathematical formulation of the ReLU activation function is expressed in Eq. (2).

$$f(x) = \max(0, x) = \begin{Bmatrix} x_i, if\ x_i \geq 0 \\ 0, if\ x_i < 0 \end{Bmatrix} \tag{2}$$

The ReLU activation function serves to rectify input values below zero by setting them to zero, thus mitigating the issue of missing gradients. The feature extraction stage is structured with a configuration of 9 blocks. Blocks 1 and 2 incorporate a depth-wise convolution layer with 1×1 filters for dimension reduction, followed by a Separable convolution layer featuring 32 3×3 filters and applying the ReLU activation function, succeeded by Batch Normalization. Block 1 is connected to Block 2 through a skip connection, followed by a dropout mechanism. Moving Block 3, it encompasses a convolutional layer with 32 filters of size 3×3, employing 'same' padding. Subsequently, Batch Normalization is applied, followed by Mix Pooling with a kernel size of 2×2, stride of 2×2, and 'same' padding. Mix Pooling is introduced in this study to address the limitations of both Max pooling and Average pooling. Max pooling, focusing solely on the largest element, can lead to losing salient features when most elements exhibit high magnitudes, potentially yielding undesirable results. Conversely, average pooling encounters challenges when a substantial portion of activations in the pooling zone is zero, leading to a significant reduction in the characteristics of convolution features [24]. Eq. (3) outlines the formula for pooling.

$$S_j = \lambda\, max\, a_i + c \tag{3}$$

The pooling output depends on the value of λ where λ is binary 0 or 1. If $\lambda = 0$ it means the pooling output or S_i is expressed in Eq. (4); otherwise, if $\lambda = 1$, then S_j is obtained using the relation in Eq. (5) for average pooling.

$$S_i = (1-\lambda)\frac{1}{|R_j|}\sum_{i \in R_j} a_i \tag{4}$$

$$S_i = \max a_i \tag{5}$$

In Blocks 4 and 5, a Separable convolution layer is employed with 64 filters of size 3×3, maintaining the same structure as Block 3, followed by batch normalization and mix pooling. Block 6 introduces a variation by incorporating a multi-scale approach from the inception module. This entails three convolution layers with filters of 64, 64, and 32, each featuring different kernel sizes (1×1, 3×3, and 5×5). These layers are arranged in parallel, enabling the network to capture features at varying scales simultaneously. Batch normalization is applied, followed by dropout with a probability value of 0.15. Moving to Block 7, it comprises a 64-filter convolution layer with a 1×1 kernel, followed by batch normalization and dropout with a probability value of 0.25. Blocks 8 and 9 share a similar composition, featuring a convolution layer succeeded by batch normalization. The distinction lies in the convolution layer's number of filters and kernel size; Block 8 incorporates 32 filters with a 3×3 kernel, while Block 9 applies two filters with a 1×1 kernel. Both blocks conclude with dropout, utilizing a probability value of 0.25. Blocks 8 and 9 have the same composition, namely a convolution layer followed by batch normalization, differentiated by the number of filters and kernel size in the convolution layer where block 8 has 32 filters with a kernel size of 3×3 and block 9 applies two filters with a kernel size of 1×1, followed using dropout with a probability value of 0.25.

3.3.3 Classification

In the classification stage, the process involves employing global average pooling as the final channel, which is subsequently connected to the SoftMax function to transform the generated features into probabilistic classes. The modeling phase incorporates an optimizer designed to dynamically adjust the learning rate for each weight, thereby facilitating faster convergence.

3.4 Model evaluation

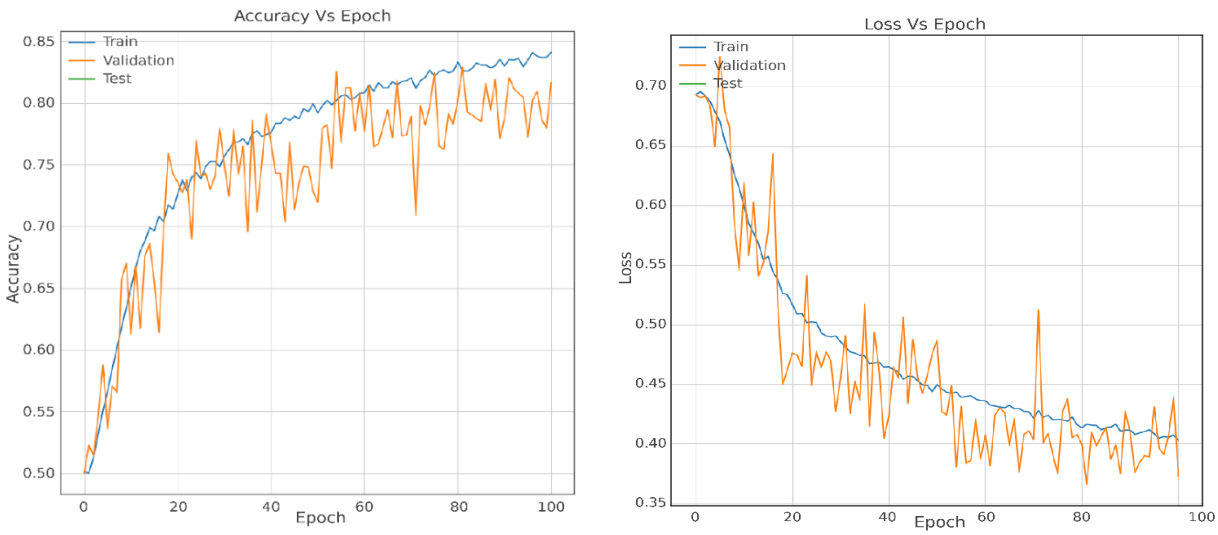
The evaluation metrics employed to assess the performance of the proposed model predominantly revolve around accuracy, as of Eq. (6). Accuracy, which measures the proportion of correctly classified instances among total predictions, serves as a straightforward and illuminating indicator of the model's overall correctness. This metric hinges on the outcomes of the four classes in the classification results: True Positive (TP), representing concealed images accurately predicted as such; True Negative (TN), denoting visible images accurately predicted as visible; False Positive (FP), indicating visible images inaccurately predicted as concealed; and False Negative (FN), representing concealed images inaccurately predicted as visible. Additionally, the model's overall performance is further gauged using the AUC-ROC (Area Under the Curve) metric. AUC-ROC values nearing 1 signify commendable performance, while those approaching 0.5 suggest a performance akin to chance.

Together, these evaluation metrics offer a comprehensive and nuanced assessment of the proposed model's classification accuracy and its overall efficacy in distinguishing between concealed and visible images.

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN} \times 100\right)\% \tag{6}$$

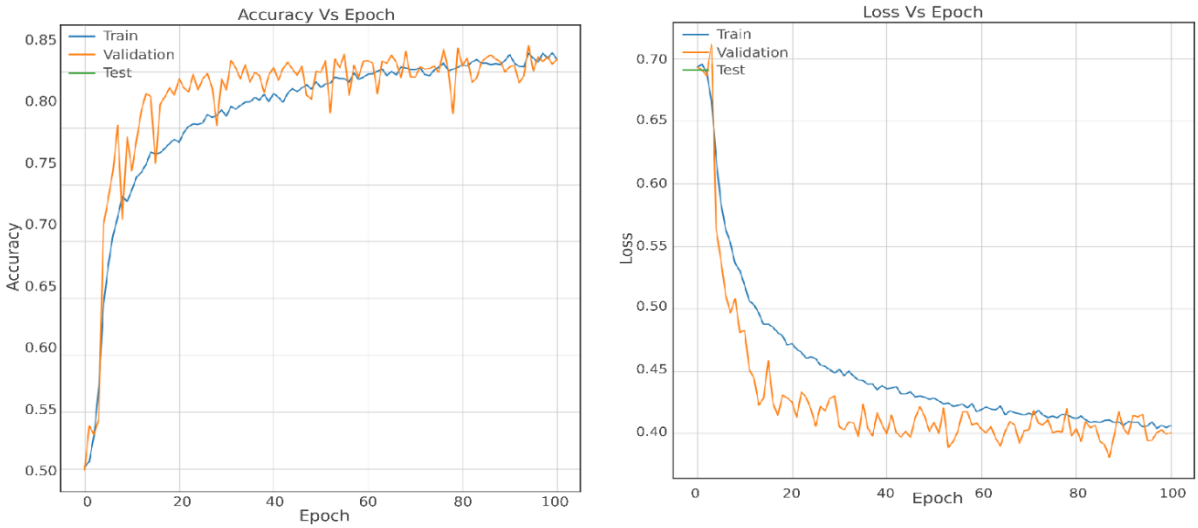
4. RESULTS AND DISCUSSION

The results analysis in this section, grounded in the experimental setups outlined in Section 3, provides a comprehensive evaluation of the proposed model's performance concerning the S-UNIWARD and WOW steganographic algorithms at a 0.4 bpp payload. The comparative benchmarking against state-of-the-art models, with accuracy as the key metric, reveals insightful strengths and weaknesses. Figures 3-5 visually depict the model's accuracy progression during training, notably showcasing stable performance across epochs for both S-UNIWARD and WOW stego datasets.



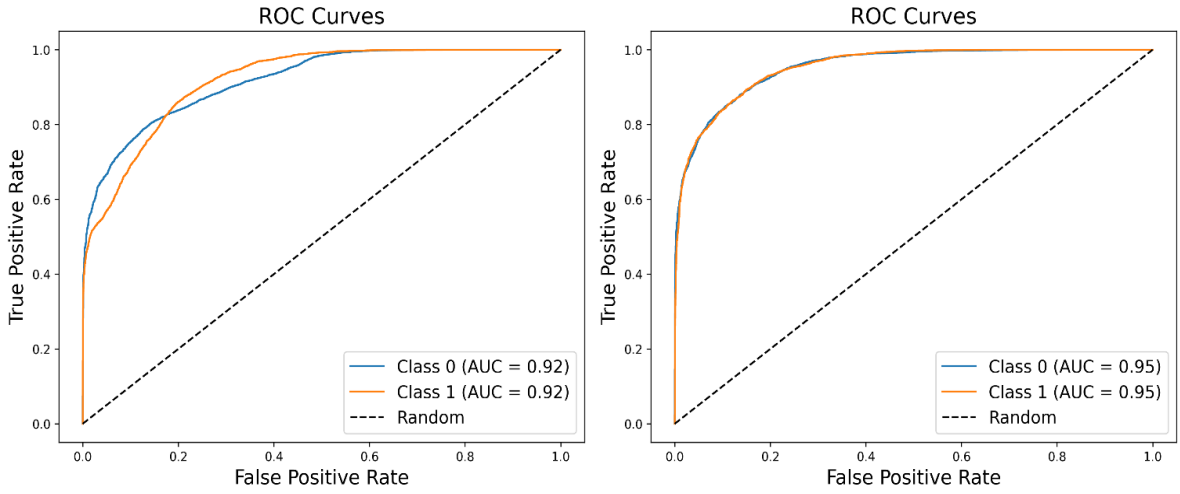
(a) Accuracy (b) Loss

Figure 3. Training and validation curves of the proposed CNN with S-UNI 0.4 bpp



(a) Accuracy (b) Loss

Figure 4. Training and validation curves of the proposed CNN with WOW 0.2bpp



(a) S-UNIWARD (b) WOW

Figure 5. ROC curves

Note: Legend: "Class 0" represents the stego images, "Class 1" represents the cover image. Random represents the unclassified inputs.

Table 1. Comparison of the accuracy between the proposed method and the state-of-the-art works

Steganographic Algorithm	S-UNIWARD	WOW
	0.4 bpp	0.4 bpp
Yedroudj-Net	77.2	84.1
Zhu-Net	80.1	84.4
GBRAS-Net	81.4	85.9
Proposed Method	84.5	87.6

Table 1 provides a comprehensive overview of the accuracy results from model comparisons, showcasing the proposed CNN model against Yedroudj-Net, Zhu-Net, and GBRAS-Net. The testing accuracy for S-UNIWARD and WOW stego datasets with a 0.4 bpp payload is systematically presented for each model. Starting with the baseline Yedroudj-Net model, subsequent enhancements with the Zhu-Net model yielded improved accuracies, further elevated by the GBRAS-Net model. Notably, the proposed CNN model demonstrated superior performance, achieving testing accuracies of 84.5% and 87.6% for the S-UNIWARD and WOW datasets, respectively. This represents a substantial improvement, surpassing GBRAS-Net by 2.9% for the S-UNIWARD dataset and 2.3% for the WOW dataset. The table succinctly captures the comparative accuracy gains achieved by the proposed model, showcasing its efficacy in steganalysis tasks.

The impact of the proposed enhancements, namely mix pooling, multi-scale inception module, and dropouts, on the model's performance is significant. These enhancements contribute to stability without compromising accuracy or succumbing to overfitting. Notably, the proposed CNN model outshines Yedroudj-Net, Zhu-Net, and GBRAS-Net, achieving testing accuracies of 84.5% and 87.6% for the S-UNIWARD and WOW datasets, respectively. This surpasses GBRAS-Net by 2.9% and 2.3%, underscoring the efficacy of the model's design choices. The incorporation of mix pooling, the multi-scale inception module, and dropouts enhances the model's ability to discern concealed and visible images with improved accuracy. While the results demonstrate notable success, future research avenues may explore additional optimizations in model architecture and investigate features or preprocessing methods to bolster robustness across diverse datasets further. The current study lays a strong foundation for advancements in steganalysis methodologies, offering promising avenues for future exploration and refinement.

5. CONCLUSIONS

This study makes significant contributions by introducing a meticulously designed steganalysis model for the precise detection of concealed data within spatial-domain images. The proposed CNN architecture integrates innovative elements, including mix pooling, dropouts, and a multi-scale initialization module. These enhancements collectively result in substantial improvements over previous works, as particularly demonstrated in the application to WOW and S-UNIWARD stego datasets at a 0.4 bpp payload. The observed enhancements in accuracy, ranging from 2.3% for S-UNIWARD to 2.9% for WOW, underscore the model's effectiveness in advancing steganalysis methodologies. The broader implications of this research extend to its potential applications in real-world scenarios, where the accurate detection of concealed data is of paramount importance. The proposed model's superior performance, outperforming previous works, positions it as a significant milestone in steganalysis. The refined architectural choices, such as mix pooling, dropouts, and the multi-scale initialization module, enhance the model's adaptability and generalizability. These qualities make the model applicable in diverse settings, including cybersecurity, digital forensics, and information security, where identifying hidden information within images is crucial. This paper specifically presents a steganalysis model that not only enhances accuracy in detecting concealed data but also contributes to the broader field of image-based information security. The strategic use of mix pooling, dropouts, and a multi-scale initialization module showcases significant improvements over existing methodologies. The observed enhancements in accuracy underscore the model's efficacy. This study serves as a stepping stone for future advancements in steganalysis, offering a robust and adaptable model with implications beyond the academic realm.

Future work in this domain could explore more advanced optimization techniques, further refinement of the model architecture, and the exploration of novel features or preprocessing methods to enhance the steganalysis model's robustness and applicability across diverse datasets. Additionally, investigating the model's performance under varying payload conditions and different steganographic algorithms would contribute to a more comprehensive understanding of its capabilities and limitations.

FUNDING

The research was supported by the Institut Teknologi Sepuluh Nopember, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2024.

REFERENCES

[1] Ferreira, W.D., Ferreira, C.B., Da Cruz Júnior, G., Soares, F. (2020). A review of digital image forensics. Computers & Electrical Engineering, 85: 106685. https://doi.org/10.1016/j.compeleceng.2020.106685

[2] Aminy, M.R.H., De La Croix, N.J., Ahmad, T. (2023). A reversible data hiding approach in medical images using difference expansion. In 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), Bangkok, Thailand, pp. 358-362. https://doi.org/10.1109/CICN59264.2023.10402139

[3] Anandha, R.D.A., De La Croix, N J., Ahmad, T. (2023). A steganographic scheme to protect medical data using radiological images. In 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), Bangkok, Thailand, pp. 369-374. https://doi.org/10.1109/CICN59264.2023.10402248

[4] Arsyad, H., De La Croix, N.J., Ahmad, T. (2023). A steganographic approach to secure data using pairs-based difference expansion. In 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), Bangkok, Thailand, pp. 363-368. https://doi.org/10.1109/CICN59264.2023.10402286

[5] Mayer, O., Stamm, M.C. (2019). Forensic similarity for digital images. IEEE Transactions on Information Forensics and Security, 15: 1331-1346. https://doi.org/10.1109/TIFS.2019.2924552

[6] Mayer, O., Stamm, M. C. (2019). Forensic similarity for digital images. IEEE Transactions on Information Forensics and Security, 15: 1331-1346. https://doi.org/10.1109/TIFS.2019.2924552

[7] Ntivuguruzwa, J.D.L.C., Ahmad, T. (2023). A convolutional neural network to detect possible hidden data in spatial domain images. Cybersecurity, 6(1): 23. https://doi.org/10.1186/s42400-023-00156-x

[8] De La Croix, N.J., Ahmad, T. (2023). Toward hidden data detection via local features optimization in spatial domain images. In 2023 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, pp. 1-6. https://doi.org/10.1109/ICTAS56421.2023.10082736

[9] De La Croix, N.J., Ahmad, T. (2023). Toward secret data location via fuzzy logic and convolutional neural network. Egyptian Informatics Journal, 24(3): 100385. https://doi.org/10.1016/j.eij.2023.05.010

[10] De La Croix, N.J., Ahmad, T., Ijtihadie, R.M. (2023). Pixel-block-based steganalysis method for hidden data location in digital images. International Journal of Intelligent Engineering & Systems, 16(6): 375-385. https://doi.org/10.22266/ijies2023.1231.31

[11] Eid, W.M., Alotaibi, S.S., Alqahtani, H.M., Saleh, S.Q. (2022). Digital image steganalysis: Current methodologies and future challenges. IEEE Access, 10: 92321-92336. https://doi.org/10.1109/ACCESS.2022.3202905

[12] Alzubaidi, L., Zhang, J., Humaidi, A.J., et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8: 53. https://doi.org/10.1186/s40537-021-00444-8

[13] Qian, Y., Dong, J., Wang, W., Tan, T. (2015). Deep learning for steganalysis via convolutional neural networks. Media Watermarking, Security, and Forensics, 9409: 171-180.

[14] Xu, G., Wu, H.Z., Shi, Y.Q. (2016). Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 23(5): 708-712. https://doi.org/10.1109/LSP.2016.2548421

[15] Ye, J., Ni, J., Yi, Y. (2017). Deep learning hierarchical representations for image steganalysis. IEEE Transactions on Information Forensics and Security, 12(11): 2545-2557.

https://doi.org/10.1109/TIFS.2017.2710946
[16] Zhang, W., Li, C., Peng, G., Chen, Y., Zhang, Z. (2018). A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mechanical Systems and Signal Processing, 100: 439-453. https://doi.org/10.1016/j.ymssp.2017.06.022
[17] Yedroudj, M., Comby, F., Chaumont, M. (2018). Yedroudj-net: An efficient CNN for spatial steganalysis. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 2092-2096. https://doi.org/10.1109/ICASSP.2018.8461438
[18] Boroumand, M., Chen, M., Fridrich, J. (2018). Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 14(5): 1181-1193. https://doi.org/10.1109/TIFS.2018.2871749
[19] Zhang, R., Zhu, F., Liu, J., Liu, G. (2019). Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. IEEE Transactions on Information Forensics and Security, 15: 1138-1150. https://doi.org/10.1109/TIFS.2019.2936913
[20] Reinel, T.S., Brayan, A.A.H., Alejandro, B.O.M., et al. (2021). GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis. IEEE Access, 9: 14340-14350. https://doi.org/10.1109/ACCESS.2021.3052494
[21] Bas, P., Filler, T., Pevný, T. (2011). Break our steganographic system: The ins and outs of organizing BOSS. In 13th International Conference, IH 2011, Prague, Czech Republic, pp. 59-70. https://doi.org/10.1007/978-3-642-24178-9_5
[22] Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378. https://doi.org/10.48550/arXiv.1811.03378
[23] De La Croix, N.J., Ahmad, T., Han, F. (2023). Enhancing secret data detection using convolutional neural networks with fuzzy edge detection. IEEE Access, 11: 131001-131016. https://doi.org/10.1109/ACCESS.2023.3334650
[24] Li, Y., Bao, J., Chen, T., Yu, A., Yang, R. (2022). Prediction of ball milling performance by a convolutional neural network model and transfer learning. Powder Technology, 403: 117409. https://doi.org/10.1016/j.powtec.2022.117409