SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper
Open Access

Deep Feature Extraction Techniques For Machine Learning-Based Digital Image Steganalysis

Jayashri Jagannath Patil[1, *], Nilesh Ashok Suryawanshi[2]

[1]NES's Gangamai College of Engineering Nagaon, Dhule, 424005, India Email:1jaanhavi1@gmail.com
[2]NES's Gangamai College of Engineering Nagaon, Dhule, 424005, India. Emai: nileshsuryawanshipatil088@gmail.com
Corresponding Author: Jayashri Jagannath Patil. Email: 1jaanhavi1@gmail.com

Abstract
Digital images are increasingly being used to improve communication, disseminate information and support e-commerce. This leads to the increasing demand for effective approaches for detecting hidden information via steganography. We present a machine-learning steganalysis framework, which employs deep feature extraction. It combines deep learning, statistical representation of functions, dimensionality reduction, interpretable feature assessment and ensemble classification. To do this, the suggested approach employs a set of cover and stego photos from the ALASKA2 dataset. Image Preprocessing Image pre-processing includes normalisation, scaling, noise factors and augmentation, which are used to enhance the learning process. In addition, enhanced deep image features extracted using ResNet-50 are combined with wavelet-based residual attributes and statistical features such co-occurrence matrices, Local Binary Patterns (LBP), and Subtractive Pixel Adjacency Matrix (SPAM). Thus, the combined feature representation reduces the feature redundancy and dimension using Principal Component Analysis (PCA). We next use SHAP (SHapley Additive exPlanations) to discover and rank the important feature components correlating with the category, for subsequent classification interpretability. The best feature selection is evaluated using the ensemble classification model of DCNN and XGBoost. The experimental results show that the proposed Ensemble + PCA + ResNet-50 framework achieves high performance of 98.57% accuracy, 99.48% precision, 98.46% recall and 98.97 F1-score as compared to some models used CNN based architectures (Resnet50) and deep learning methods on the dataset provided in details in this paper. Moreover, the proposed approach for extraction and classification is compared with the state of the art approaches for steganalysis to show its effectiveness. This shows that the combination of deep and handcrafted features with PCA optimisation, subjective explainable AI (SHAP) interpretability tool and ensemble methods may build a strong and consistent framework of digital image steganalysis.

Keywords Digital Image Steganalysis; Deep Feature Extraction; ResNet-50; Principal Component Analysis; SHAP; XGBoost; Deep Convolutional Neural Network; Ensemble Learning; Steganography; Machine Learning.

1. Introduction

With the development of digital communication channels, cloud computing for information sharing and social media apps, the quantity of digital photographs transferred over public and private networks is growing dramatically [1]. This need to preserve essential information and development in technology leads to the development of many data hiding strategies. Such methods include steganography, which is hiding information in digital photos. Hidden information is embedded into a digital image so that it can maintain the visual qualities of each overlay level. Steganography is different from standard encryption in that it does not attempt to hide the fact that a message is being sent. Image steganography has been considered in numerous applications from intelligence operations, digital watermarking, copyright protection, to covert communication and safe transmission of information. Digital images also provide a source of astounding location for concealed information due to their huge space, statistic and frequency-space redundancies [2]. These changes, often subtle and made almost undetectable, yet allow a dialogue about an unliteral message within. However, recent advances in adaptive and content-aware steganographic algorithms have made it challenging to detect the existence of hidden information. Current embedding methods attempt to embed these parts in the image space such that the probability of noticing any changes is low. Also, this means very little difference between a stego image and the genuine cover image. In contrast, these changes can be seen as innocuous statistical noise, local texture modifications, changes in pixel co-occurrence or persisting noise patterns [3].

Steganalysis of Digital Images Digital picture steganalysis is the art of finding out whether a digital image includes hidden information or not. Feature extraction, steganalysis algorithms often examine pictures to categorise cover and stego images [4]. This representation of the input image is one of the important features of the detecting process. In this sense, the feature extraction stage is of great importance, since the steganographic insertion produces residues that are effectively diluted throughout a variety of spatial and/or frequency domains. Pixel interdependencies, noise residues, local texture features, statistical distributions, frequency domain fluctuations and other tiny artefacts produced by the embedding process are often useful. In the traditional machine learning based steganalysis methods proposed in the past, the so-called artisanal feature descriptors are usually used [5]. For steganographic markers, numerous techniques have been well researched such as statistical characteristics, co-occurrence representations, local texture features, residual features or transform domain methods. These properties are usually very intuitive and give a lot of insight into the modifications performed by embedded methods [6]. But manual feature engineering requires a good deal of experience and will eventually fail to discover non-linear interactions associated with the underlying image composition and its concealed fault. Furthermore, the performance of manually built descriptors deteriorates if the embedding technique is) flexible and/or content driven or significantly deviates from the assumptions stated a priori when synthesising features.

Recently, deep learning has been proposed as a viable technology that can provide the ability to automatically extract specific attributes of images directly from the input. Object identification, picture classification and segmentation are among the most challenging computer vision problems due to their intrinsic complexity [7]. Trained representations have been demonstrated to capture subtle and dispersed patterns that are hard to explain using handcrafted features, e.g. in steganalysis. ResNet is a popular architecture based on Residual Learning framework, in which the shortcut and direct paths help to efficiently train all the layers to learn residual mappings from the inputs. Therefore, trained feature representations can provide an excellent foundation for detecting small differences between cover and stego images. There are numerous benefits of deep learning, however there are several disadvantages to utilise exclusively deep learning methods [8]. Deep Neural Networks will usually need a big and representative training set, as well as heavy hardware resources. However, the learnt representations could be uninterpretable for humans and it is not obvious to what extent feature space position feature characteristics toward a steganalysis class determination [9]. Also, a decent embedding on particular dataset does not guarantee that it will function similarly under other image transformations, various compression ratios or payload dimensions (i.e. cover to secret ratios) or different steganographic methods. These limits require that the feature extraction algorithms be robust, representative, compact and discriminative in this manner [10].

Hence, machine learning provides the appropriate framework to enhance the use of visual representations retrieved for steganalysis. This allows us to apply deep feature extraction techniques to make the data more representative, so that we can train a machine-learning classifier on these extracted features instead of the end-to-end deep classification framework exclusively. It then decouples the representation learning and the classification in two distinct phases, so allowing unsupervised experimentation of different feature selection and classification algorithms [11]. Furthermore, the integration of deep features with statistical, textural and residual descriptors provide extra information since different types could essentially represent distinct representations of steganographic changes. One of the biggest challenges in machine learning based picture steganalysis is the dimensionality of features . Deep networks can generate incredibly complex feature vectors, but can also benefit from increasing the number of dimensions by employing hand-crafted descriptors when fusing different feature sets [12]. Using all derived features directly will increase the computing complexity, duplication of information and possibly loss of classification efficiency. Dimensionality-reduction methods (e.g. Principal Component Analysis (PCA)) can be used to generated an abstracted state space. Maintain a lot of the critical information about your input when it is reduced in size and definition. Feature-importance techniques to identify key features so that you can have a better understanding of what is represented in the derived representation [13].

Therefore, the majority of existing digital picture steganalysis based on machine learning emphasises deep feature extraction. Differently from earlier works where the feature extraction is assumed as a standard pre-processing step, in this work the feature extraction is considered as an inherent part of the steganalysis framework. 4-lever deep feature representation is fused from residual convolutional framework in cross domain generalisation for secure messaging utilising generative models techniques [14]. The approach combines recent progress on richer feature extraction with well known statistical, textural and residual qualities to increase discriminative power between cover and stego images. Instead of compressing image-content data in high dimensions to optimise a fragile representation, we compute the mean over important steganographic signs from our representation. The suggested research below is on an enhanced hybrid feature extraction framework for data concealed image steganalysis. It includes picture preprocessing and augmentation, deep representation learning using ResNet-50 as well as SPAM and other hand-crafted descriptors (LBP; co-occurrence features; wavelet-based residual features). In this paper the generated feature vectors are embedded into high dimensional space. PCA is used for dimensionality reduction to create a precise representation [15]. Thus SHAP is applied to assess the importance of the attributes, to find factors with most influence. And then, such enhanced features are available in machine learning categorisation of cover and stego photos. It is similar to the hybrid feature-extraction approach proposed in but places more emphasis on deep feature representations.

The primary purpose of this article is to present a feature extraction approach preserving its computational economical and interpretable property for dealing with modest steganographic traces. The essay describes limits should not be only specified by human provided description but with deeper representations. The proposed two-step approach is augmented with additional information from diverse feature types, like picture statistics, local texture, residual patterns and extracted deep representations. Such fusion may contribute to a better understanding of the distinctions between cover and stego pictures, and may also aid improve machine-learning classifiers to make more accurate decisions. To obtain unique representations of our digital photos, we propose a deep feature extraction approach based on ResNet-50. The combination of deeper representations and statistical features, textures and residual descriptors to characterise more aspects of the steganographic alterations. Having played with different features that can be combined in Useable pre-processing, we are now looking at key takeaways for strengthening the durability of a Features in different image contexts (e.g. resizing, normalisation, noise removal (gussian or salt-and-pepper), rotation, flip, scale add noise and colour changes [16].
PCA can be considered as a method of reducing a high dimensional feature space into a low dimensional space, and combine the relevant information, but keep most of the variation in a data point. By evaluating the relevance of the features, SHAP helps to prioritise the large feature elements and also provides further insight into the contribution of each derived representation. To classify cover and stego pictures, to find out the optimal feature representation by using machine learning and to evaluate the performance of deep feature extraction quantitatively Fig. 1. We believe the main contribution of our research is in the consideration of feature extraction as a stand-alone and pluggable portion of an entire digital picture steganalysis approach. This module can include several machine learning classifiers and forensic analysis applications, rather than being bound to a specific end-to-end detection framework. This demands a generic feature-extraction module which is agnostic of the machine-learning techniques or forensic-related activities that are subsequently demanded [17].

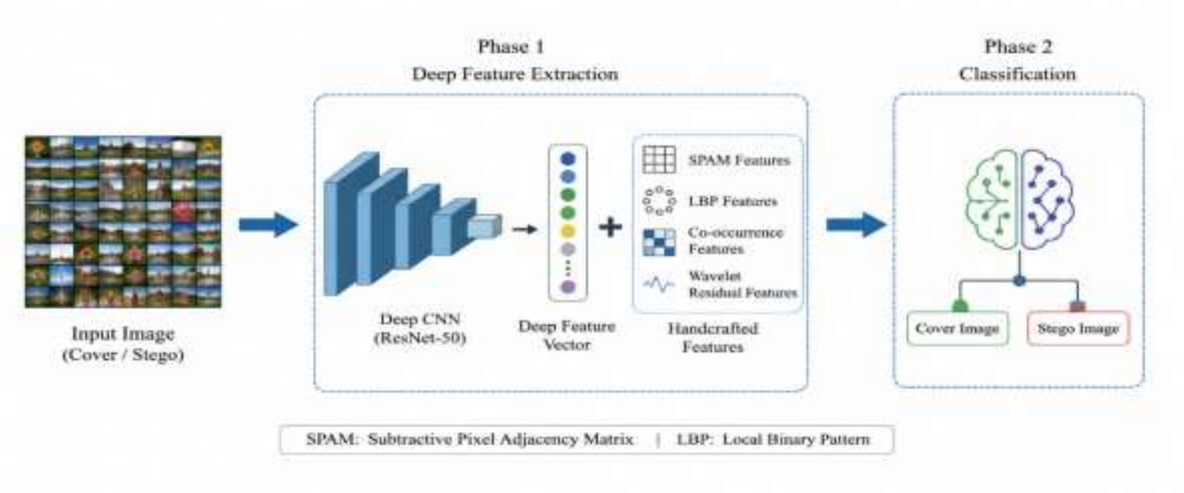

Figure 1: Proposed Deep Feature Extraction Framework for Machine Learning-Based Digital Image Steganalysis

2. Literature Review

With the rapid development of machine learning and deep learning methods, the study on digital picture steganalysis has made remarkable progress [18]. Previous studies have relied on well designed statistical, textural, spatial & transform domain features to find the tiny variations due to embedding of classified information. Most of these properties are utilised in the detection of irregularities in the pixel inter-connectivity, which is in contradiction with the local structure of the image's auditory features and frequency domain distributions. Conventional procedures were utilised to obtain the equivalent and relevant representations [19]. However, this approach may not be effective against existing steganography techniques that utilise flexible and content-aware embedding. Traditional feature extraction approaches are still necessary, because steganographic embedding usually causes little alterations that would not be detected by a visual method [20]. Statistical descriptors indicate irregularities in the distribution of pixels, while texture-based descriptor is able to detect variations in the local structure of a picture. Co-occurrence-based characteristics describe the link between neighbouring pixels while residual-based attributes are designed to capture weak noise processes generated by the embedding process. The various types of characteristics provide additional details that help in discriminating between cover and stego photos [21].

However, such kind of approaches are usually constructed by hand and rely mostly on the intuition about which attributes are expected to be changed in a certain steganography process [22]. Switching the embedding technique might make static feature sets less useful. Furthermore, it is observed that hand-crafted descriptors tend to choose more than what is required and produce big feature vectors that contain primarily redundant or less important information. These constraints have forced academics to turn to artificial feature-learning methods. Deep learning has made a great success over the previous handmade feature based methods for picture steganalysis because it can automatically learn feature representations from the raw visual data [23]. They are particularly beneficial for CNN-based architectures since they can provide hierarchical representations of visual patterns from low level details to high-level structures [24]. To distinguish between cover and stego images. Convolutional layers are meant to learn these features in step wise manner [25].

Residual learning frameworks have been regarded as an effective tool for steganalysis because the data blended during a steganographic embedding process generally has significantly lesser capacity than the original image content. ResNet based architectures leverage residual connections, which allows to build complicated representations while keeping the stability of a thoroughly trained network [26]. This permits to use such architectures for the extraction of deep feature vectors holding information that is not easily accessible by means of classical hand-crafted descriptors [27]. Moreover, the difference between classification and representation learning is deep feature extraction. A full deep network classifier is not necessary; instead, intermediate representations, which can be retrieved from a working or already-trained deep architecture, can supply input features for typical machine-learning methods. Moreover, this methodology offers the further benefit of being able to assess a large number of feature-selection methods and classifiers within a single deep-feature representation [28].

In the past few years, there have been significant advances in picture steganalysis using various state-of-the-art CNN designs and transfer learning using multi-scale networks, attention mechanisms, graph-based learning approaches and residual networks [29]. These strategies highlight the power of deep element architectures to achieve state-of-the-art detection performance by learning low-level picture characteristics in an end-to-end fashion. where the steganographic properties are accumulated in the multilayer structures at different spatial scales and the attention-based algorithms extract the associated spots from the image [30]. In particular, transfer-learning algorithms would decrease the requirement to develop a whole new model from scratch. But there are other issues with deep learning approaches as well. Training deep neural networks can sometimes be expensive because to the amount and size of computational resources and data needed to run these massive state-of-the-art architectures [31]. Moreover, the internal representations of these systems are notoriously difficult to interpret, which can be an issue if an interpretable steganalysis wishes to validate the reason behind a choice [32].

Therefore, it is vital that research permits image alteration. The data is embedded into an image after compression or other modifications to the image during its use. There was no chance the steganographic proof would be retained or even improved, given how weak and murky it was in the first place [33]. Therefore, a good steganalysis system should be able to achieve a very high level of accuracy in normal situations and also a sufficient classification measures for images, such as most of the basic processing procedures. Another important area of research is the development of inexpensive, low-mass steganalysis systems [34]. Deep networks are known to provide very robust features, but large designs can be very expensive in terms of compute and memory. Blocked with link constraints for instant execution. Hence, it is critical to lessen the dimensionality of deep feature representations while maintaining as much important information as possible. To address this issue, feature selection and dimensionality reduction algorithms can be used to minimise duplication and less important components from the feature space produced. Principal Component Analysis translates high dimensional feature space to a lower dimensional subspace but in such a way that information that contributes most to the variation in the data is kept in its original form [35]. Therefore, the feature importance measurement methods can be used to identify the most essential variables in the final classification [36].

Deep features are the black box abstraction, deep learning can automatically learn sophisticated high dimensional nonlinear dependencies from picture data, but hand-crafted descriptors may only preserve certain statistical, textural and residual properties [37]. The combination of these sorts of information can provide an additional level of comprehension and potentially improve steganalysis. Mathematical, residual characteristics; e.g. Subtractive Pixel Adjacency Matrix (SPAM), Local Binary Pattern (LBP) and co-occurrence attributes can rely on the combinatory factor with ResNet-50 13 to form hybrid representation. These separate feature clusters correspond to a different facet of the visual data. Texture features capture local structural variations, statistical features capture pixel interactions, deep features are hierarchical components learned and residuals capture small scales for embedding-related changes [38].

PCA can be used to get the reduced-dimensional representation, and SHAP-based evaluation can be done by assessing how features of such representations explain the obtained representations. Now the machine learning classifiers with richer feature representation are working [39]. XGBoost is a tree based ensemble method which can discover non-linear correlations in structured feature vectors while deep neural classifiers may learn complex decision boundaries. Complementary classifiers are auxiliary and it is the approach to increase the reliability in doing cover-versus-stego classification in more than one modalities. Despite the impressive advances, there are considerable gaps in the literature. Most similar work does not assume that deep feature extraction can be a discrete component to draw from as a library, but rather takes an end-to-end classification approach [40]. Meanwhile, the combination of spontaneously learnt deep features with traditional statistical, textural and residual descriptors needs to be further strengthened. In addition, the technical and interpretative challenges of high dimensional deep representations necessitate the proper application of feature selection algorithms [41].

CNN model on a dataset or with only one embedding approach may not necessarily be effective over multiple datasets, picture quality levels, payload levels and/or embedding techniques. This is a well-known fact. Hence, there are two main limitations that the system has to overcome: the ability of detecting new steganographic features and the robustness of the method to image modifications. Therefore, this work is dedicated to the relatively recent techniques for picture data concealment in machine learning using deep and feature extraction methods as shown above. The proposed method does not make use of all-in-one classifier, but instead, it extracts, integrates, optimises and interprets distinct visual characteristics separately. In the present work, the deep features based on ResNet-50 are merged with the statistical, texture and residual features. Feature-ranking approaches are applied to improve the high-dimensional representation, which is input data for machine-learning classification methods [42].

In this context, the work presented here tries to go beyond standard manual feature extraction and pure deep-learning methods by establishing a hybrid feature representation of integrated features. It is based on the above mentioned contributions and combines the concepts of deep learning, statistical measure, texture analysis and residual inspection / dimensionality reduction (structurally invariant components) across features, combined with feature importance measures and machine learning classification. Combined with prior techniques it means a more informational and strong version for cover shots than stego images [43].

3. Research Methodology

This segment outlines the approach utilised to derive deep features from low-level characteristics and implement them in digital picture steganalysis through machine learning. The approach is organised as a sequential series of phases that includes: dataset preparation, image preprocessing and augmentation, extraction of deep features utilising ResNet-50, a combined process for handcrafted feature extraction, the integration of both aforementioned feature types, dimensionality reduction through Principal Component Analysis (PCA), evaluation of feature significance using SHAP, and the training of machine-learning classifiers for the classification objective. To obtain the fundamental feature learning framework, we are executing training on ResNet-50, succeeded by PCA and SHAP. We maintain supplementary statistical and textural descriptors that enhance the precise depiction of nuanced steganographic indicators Fig. 2.

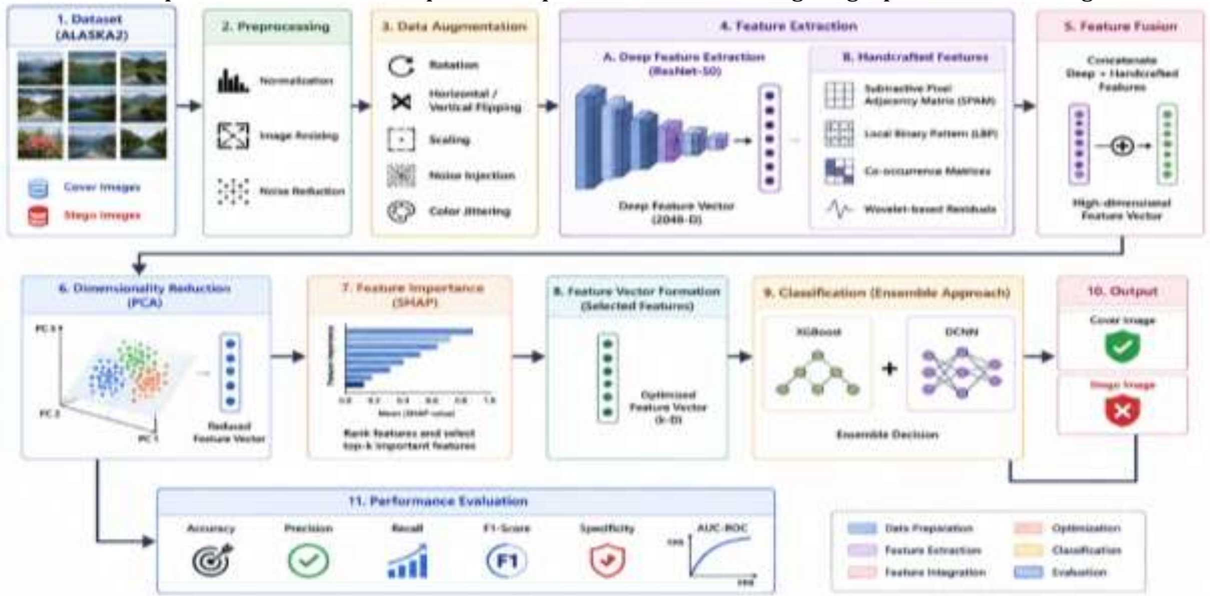

Figure 2: Proposed deep feature extraction and machine-learning-based steganalysis framework.

3.1 Dataset and Image Collection

The primary origin of experimentation is the ALASKA2 picture steganalysis dataset. A tensor RGB is utilised to depict the input photos. For instance, an unprocessed depiction of an image with a height of H and a width of W, such concrete flooring featuring three colour channels, can be articulated as:

$$X \in \mathbb{R}^{(H \times W \times 3)} \quad (1)$$

The binary target variable: $y \in \{0, 1\}$ represents a cover image when $y = 0$ and a stego image when $y = \lambda \sim \eta$, with $\eta = 1$.

$$y \in \{0, 1\} \quad (2)$$

3.2 Image Preprocessing and Standardization

Preprocessing is an operation performed before feature extraction to achieve standardised representations of the inputs. The laboratory is tasked with adjusting the image dimensions, standardising pixel values, and eliminating noise as required. Resizing guarantees uniform spatial dimensions for all photos, as dictated by the ResNet-50 functionality. This normalisation reduces the impact of fluctuations in pixel-value ranges and improves the numerical stability of the feature extraction procedure.

We standardise a pixel value x to the range [0, 1] through:

$$x\_norm = x / 255 \quad (3)$$

The normalised value may be adjusted using the mean μ and standard deviation σ of your training dataset if you choose to do normalisation:

$$z = (x\_norm - \mu) / \sigma \quad (4)$$

Preprocessing settings are established on the training subset and applied to the validation and testing subsets. This inhibits the transmission of information from the assessment data to the training procedure.

3.3 Data Augmentation for Robust Feature Learning

To furnish the feature extractor with a regulated (though slight) alteration of the input images, we employ data augmentation. Augmentation procedures encompass the incorporation of noise and colour, slight rotations, scaling, and requisite horizontal and vertical displacements. The sole data that undergoes augmentation is the training data. The performance reflects the model's true generalisation ability since the validation and testing sets remain unchanged.

The depiction of a train can be altered in the following manner:

$$X' = T\_k(X), \quad T\_k \in \{\text{rotation, flip, scale, noise, color variation}\} \quad (5)$$

The enhancement technique seeks to fortify the model's resilience while preserving statistical and residual patterns that could indicate steganographic embedding.

3.4 Deep Feature Extraction Using ResNet-50

The backbone of the main deep feature extractor is ResNet 50. The last classification layer is not used as a final resoluting feature representation, merely for taking an input image to features vector space . The residual architecture enables information to flow across the network via shortcut links, thereby enabling it to learn complicated image representations.

Let F_ResNet(·; θ) represent the feature-extraction portion of ResNet-50, where θ denotes the learned network parameters. For an input image X, the extracted deep representation is:

$$f_{deep} = F_{ResNet}(X; \theta) \quad (6)$$

The final feature vector incorporates information gained from various layers of representation on the image. Rather than directly applying a softmax final classification layer, we retain the deep feature vector and reduce its dimensions afterward.

3.5 Complementary Handcrafted Feature Extraction

In addition to the deep learnt representation, the approach learns a set of manual statistics and a texture and residual property. The first of four feature type categories is Subtractive Pixel Adjacency Matrix (SPAM), Local Binary Pattern (LBP), and gray-level co-occurrence properties and wavelet derived residual features.

3.5.1 SPAM Features

SPAM descriptors characterise pixel variabilities via describing neighbouring pixel relations. However, steganographic embedding may interfere with local dependency. Therefore, SPAM representations might be helpful to capture small statistical variations.

3.5.2 Local Binary Pattern Features

Local Binary Pattern (LBP) specifies the relationship of a pixel with its neighbours in regard to a texture. The binary neighbourhood setting can be stated as:

$$LBP(P,R) = \Sigma_{(p=0)}^{(P-1)} s(g_p - g_c)\, 2^p \quad (7)$$

where g_c is the gravity value in the central pixel, g_p is the gravity value in the p-th neighbour and the sign function is defined as:

$$s(t) = \{ 1,\ t \geq 0\ ;\ 0,\ t < 0 \} \quad (8)$$

3.5.3 Co-occurrence Features

Co-occurrence attributes quantify the frequency of occurrence of pairs of a number of gray-level values within a range or direct geographic association. They emphasise the embedding-induced modifications of local pixel interdependencies.

3.5.4 Wavelet Residual Features

Wavelet decomposition decomposes the picture data into different frequencies. Residual parts can boost noise patterns or frequency shifts connected with steganographic modifications. The appropriate wavelet residual descriptors are saved as side information.

3.6 Hybrid Feature Fusion

The deep and handcrafted feature groups are combined to create a unified representation. Let f_deep denote the ResNet-50 representation and f_hand denote the concatenated handcrafted representation. The fused feature vector is defined as:

$$f_{fused} = [f_{deep} \,||\, f_{SPAM} \,||\, f_{LBP} \,||\, f_{CoOcc} \,||\, f_{wavelet}] \quad (9)$$

where || denotes vector concatenation. The purpose of fusion is to preserve complementary information rather than assuming that one feature family can capture all steganographic evidence.

3.7 Dimensionality Reduction Using Principal Component Analysis

At last, the representation may be several interconnected properties. Hence PCA converts a higher-dimensional original feature space to a lower (orthonormal) dimension. Covariance matrix for a centred feature matrix X:

$$C = (1 / (n - 1))\, X^T X \quad (10)$$

The principal directions are obtained from the eigenvalue decomposition:

$$C v_i = \lambda_i v_i \quad (11)$$

The transformed representation using the first k principal components is:

$$Z = X V_k \quad (12)$$

The selection of K is based on curving cumulative explained variance or validation accuracy over different values of K, noticed return diminishing as k rises. It reduces redundancy, decreases computation cost and offers a better representation for subsequent stages of feature importance and classification.

3.8 SHAP-Based Feature Importance and Selection

The SHAP (SHapley Additive Explanations) methodology estimates how much every input contributes to the final model output (in this case, a PCA-transformed dataset). This method enables to keep the most important features and provides a simple estimate of how much each feature contributes.

For a prediction model f and feature vector x, the SHAP decomposition can be expressed as:

$$f(x) = \varphi_0 + \Sigma_{(i=1)}^{M} \varphi_i \quad (13)$$

where φ_0 is the expected model output and φ_i represents the contribution of feature i. The magnitude $|\varphi_i|$ can be used to rank feature importance. Features with consistently low contribution can be removed to form a compact and informative feature vector.

$$f\_selected = SelectTopK(Z, |\varphi|) \quad (14)$$

3.9 Machine-Learning Classification

The feature representation will be fed into machine-learning classifiers to perform binary steganalysis. The problem we face is related to the classification and consists of discriminating between stego vs. cover images. Therefore, XGBoost is considered the most sophisticated tree based classifier because it can model nonlinear relationships between all different structured feature elements. In addition, a more sophisticated neural classifier could be used to assess the advantages of the improved feature representation.

In the context of binary classification, the probability that you would have expected is as follows:

$$\hat{y} = P(y = 1 \mid f\_selected) \quad (15)$$

A threshold τ is then applied to obtain the final class label:

$$Class = \{ Stego,\ \hat{y} \geq \tau;\ Cover,\ \hat{y} < \tau \} \quad (16)$$

3.10 Performance Evaluation

The performance of the proposed method is evaluated through confusion matrix analysis, recall F1-score, precision and accuracy. Probabilistic results may also be presented as ROC-area under the curve (AUC).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (17)$$

$$Precision = TP / (TP + FP) \quad (18)$$

$$Recall = TP / (TP + FN) \quad (19)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (20)$$

TP is true positive (stego images correctly identified); TN is true negative (cover images correctly recognised); FP are false positive (the cover image classed as stego) and FN are false negative (cast as cover the stego image). A deep feature vector with a dimensionality of 2048, which was pre-trained on the input images used for digital image steganalysis Fig. 3.

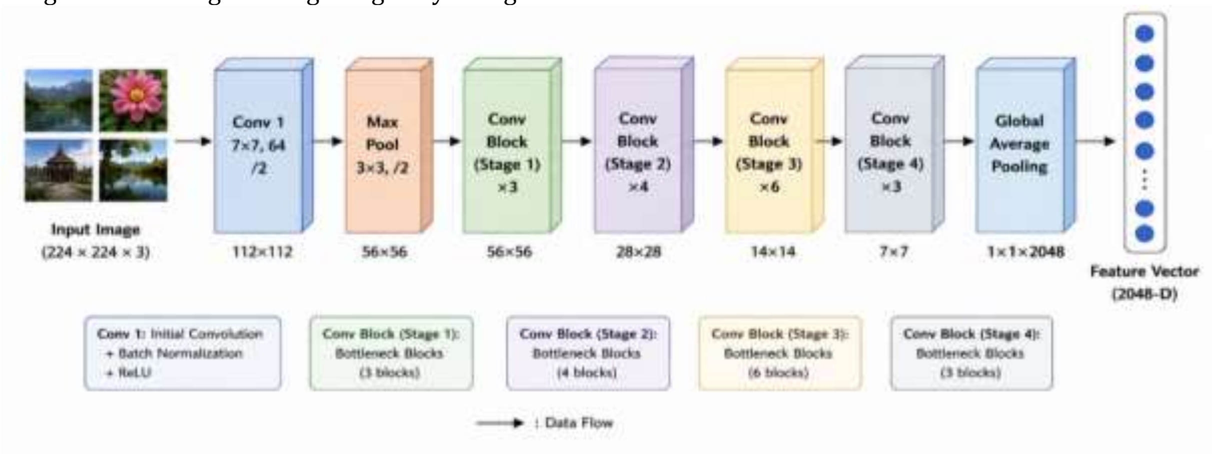

Figure 3: ResNet-50-Based Deep Feature Extraction Architecture

3.11 Pseudocode for Deep Feature Extraction

The following pseudocode summarizes the principal computational sequence of the proposed deep feature extraction pipeline. The procedure begins with preprocessed images, extracts deep representations using ResNet-50, reduces the feature space using PCA, evaluates component importance using SHAP, and returns the optimized feature matrix for machine-learning classification.

Algorithm 1: Deep Feature Extraction and Feature Selection

```
Input:
   Image dataset D = {(X_i, y_i)}, i = 1, ..., N
   Target cumulative variance ρ
   Number of selected SHAP features K

Output:
   Optimized feature matrix F*
1. Initialize ResNet-50:
2.  Load cover and stego images from D
```

3. Split D into training, validation, and testing subsets
4. For each training image X_i:
   Resize X_i to the ResNet-50 input size
   Normalize pixel values
   Apply training-time augmentation
5. End For
6. Initialize ResNet-50 feature extractor
7. Remove/disable the final classification layer
8. For each preprocessed image X_i:
   f_deep[i] ← ResNet50_Features(X_i)
9. End For
   10. Extract complementary handcrafted features:
   f_SPAM ← SPAM(X)
   f_LBP ← LBP(X)
   f_CoOcc ← CoOccurrence(X)
   f_wave ← WaveletResidual(X)
   11. Fuse feature groups:
   F_fused ← [f_deep || f_SPAM || f_LBP || f_CoOcc || f_wave]
   Standardize F_fused using training-set parameters
   Fit PCA on training features
12. Select the smallest k satisfying cumulative variance ≥ ρ
   Z ← PCA_Transform(F_fused, k)
   Train a provisional classifier g(Z)
   Compute SHAP values Φ for the PCA components
   Rank components using mean absolute SHAP value
13. Select the top K components
   F* ← Z[:, TopK(SHAP(|Φ|))]
   Train the final classifier using F*
   Evaluate on validation and test data
14. Return F* and classification performance

4. Results and Discussion

The complete algorithmic flow of our unique deep feature extraction pipeline is well summarised in the following pseudo-code: Generate deep representations from the pre-processed pictures based on ResNet-50 and evaluate the importance of contributions using SHAP to provide an appropriate feature matrix used in machine learning classification through dimensionality reduction using PCA. We present the experimental findings of a deep feature extraction based framework for digital picture steganalysis enriched with machine learning data in this research. It disentangles the impact of deep features from ResNet-50, PCA dimensionality reduction, SHAP feature importance weighting and ensemble classification. The enquires are made on the cover photos and stego images, which are used in two different classes of data.

4.1 Hyperparameter Settings of the Proposed Model

Table 1. Hyperparameter settings for the machine learning based digital image steganalysis framework with deep feature extraction. For the ResNet-50 architecture, we scaled the images to 224×224×3 to match the desired input size. We utilise ResNet-50 as a deep feature extractor to input and output each image with a 2048-dimensional feature vector. We then perform PCA to minimise dimensionality while retaining 95% variance Table 1. It is good because it isolates the more informative segments and discards unnecessary information. For further categorisation, the top 100 features are kept, and SHAP-based feature ranking is utilised to determine the elements impacting the decision after PCA.

Table 1: Hyperparameter Settings of the Proposed Model

Hyperparameter	Value
Input Image Size	224 × 224 × 3
ResNet Architecture	ResNet-50
Deep Feature Dimension	2048
PCA Components	256
PCA Variance Retention	95%
SHAP Feature Selection	Top 100 Features

No. of XGBoost Trees	300
Maximum Tree Depth	6
Learning Rate	0.05
XGBoost Subsample	0.80
XGBoost Column Sampling	0.80
DCNN Conv. Layers	4
Kernel Size	3 × 3
Activation Function	ReLU
Dropout Rate	0.50
Batch Size	32
Optimizer	Adam
Initial Learning Rate	0.001
Number of Epochs	50

The deep features employed in this work were pre-computed. For example, the figure 4 shows a visualisation of the preprocessing step that is conducted on the input photos prior to feature extraction. First, we resize the original image into 224 × 224 × 3 RGB, which is compatible with the input of ResNet-50. Then, the pixel values in the image are rescaled to a fraction in [0, 1] by using min-max scaling . This standardisation decreases the variation of the pixel intensity, and gives a normalised input for the subsequent feature extraction and steganalysis in the deep network.

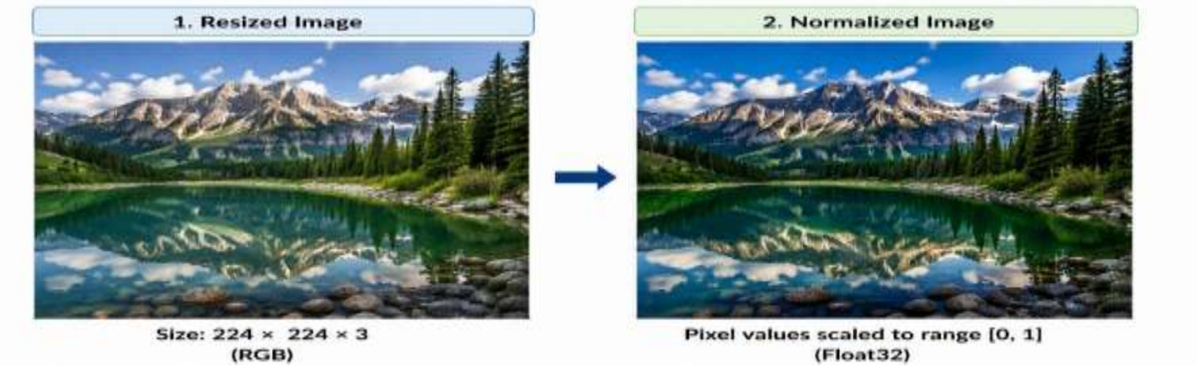

Figure 4: Preprocessing Stage for Image Resizing and Normalization

The steganographic and picture files automated data augmentation is stated as in Fig5. This comprises data augmentation by rotation, scaling and horizontal flipping of original images as well as noise and chromatic jitter. These controlled changes result in distinct training sample sets, although the essential features of the images are largely consistent.

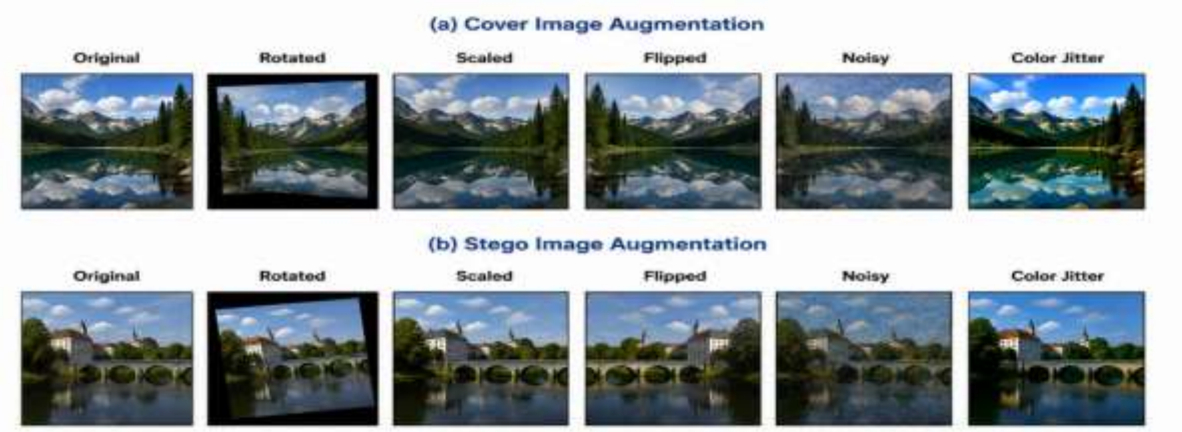

Figure 5: Data Augmentation of Cover and Stego Images

The augmentation strategy is applied to the images with varying perspectives, dimensions, noise levels, and colour differences to enhance the generalisation and robustness of the suggested feature extraction model based on ResNet-50. This allows the model to learn more robust features that can distinguish stego images from cover shots. The suggested system was evaluated by four popular classification measures, namely precision, recall, accuracy and F1-score. These metrics also give explicit information about the capacity of the algorithm to discriminate between cover and stego images.

4.2 Experimental Performance of the Proposed Framework

Our experiment findings showed that Ensemble Model with PCA154 and ResNet-50153 achieved the best overall accuracy, precision, recall and F1-score of 98.57%, 99.48%, 98.46% and 98.97% respectively Table 2.

Table 2. Overall Performance of the Proposed Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost	92.34	91.85	92.10	91.97
DCNN	94.12	93.76	94.01	93.88
Ensemble Model	96.08	95.64	95.92	95.78
XGBoost + PCA	93.41	92.96	93.20	93.08
DCNN + PCA	95.26	94.88	95.10	94.99
Ensemble + PCA	97.12	96.75	97.03	96.89
XGBoost + PCA + ResNet-50	96.02	95.54	95.88	95.71
DCNN + PCA + ResNet-50	97.38	97.01	97.22	97.11
Ensemble + PCA + ResNet-50	**98.57**	**99.48**	**98.46**	**98.97**

The values above correspond to the reported ablation results of the reference experimental study.

4.3 Effect of the Classification Model

The base XGBoost obtained an accuracy of 92.34% and the hybrid DCNN 94.12% both on the validation. The paper indicates that the DCNN learns superior nonlinearity to represent the images, and the separation between cover and stego images is larger than we can get with merely the tree-classifier. The ensemble setup raised the precision of the top model to 96.08%. At the lower end, the performance of an individual model can be offset democratically or by distributed learning. Similar trends are seen for F1-score, recall and precision. For the composite model, accuracy was 95.64%, recall 94.92% and F1 = 95.78 (total three measures were at ease with classifier results). The segment described numerous types of challenges which could be addressed in future efforts.

4.4 Effect of PCA-Based Dimensionality Reduction

Furthermore, PCA increased the effectiveness for classification to a greater extent. It uses a PCA configuration trained with XGBoost and gets an accuracy of 93.41% score utilising the standardised testing set for evaluation purposes however XGBoost on its own gets a precision of only 92.34%. Similarly, DCNN with PCA scores 95.26% whereas DCNN achieves 94.12%. PCA improved the ensemble model with the prediction accuracy increased from 96.08% to 97.12%. This is a clear example of how the reduction of incompetent data in the feature space can give a classification gain. The results show that PCA is not only a dimensionality-reduction approach, but it should be considered as an algorithm to generate a more compact representation of the features while keeping the key principle components of the original feature space.

4.5 Effect of ResNet-50 Deep Feature Extraction

The classification results demonstrate a significant improvement when using ResNet-50 as is. The combination of DCNN, PCA and ResNet-50 gives (97.38) Feature Extraction via XGBoost, PCA and ResNet-50 achieves 96.02 %. We will extract features more effectively. The resnet 50 extracted a deep feature vector of 2048 dimensions from the last pooling layer. Next, it mixes this with the residual and statistical information in the hybrid feature structure. Deep and artisanal restrictions can be combined to achieve greater information. SPAM, LBP, co-occurrence and wavelet residual features are explicit statistical qualities and local structural features. Deep features are taught feature hierarchically without human interaction.

4.6 Effect of the Proposed Ensemble + PCA + ResNet-50 Model

The best performance was obtained when ensemble classification was combined with PCA and ResNet-50. The model achieved:

- **Accuracy: 98.57%**
- **Precision: 99.48%**
- **Recall: 98.46%**
- **F1-score: 98.97%**

The results showed that the combination of three key components, decisive feature extraction, dimensionality reduction and ensemble classification, generates a more robust representation compared to any single component. Thus, XGBoost can be used to boost DCNN using its combined power. DCNN is a deep representation learning technique which produces non-linear representations in the input space, XGBoost is a particularly efficient method to get non linear decision boundaries (for structured data, i.e. feature vectors). In this paper, methodology design is utilised to improve the durability of a classification by traditional machine learning and deep learning techniques.

4.7 Confusion Matrix Analysis

The confusion matrices additionally prove the discriminative capabilities of the proposed technique. Evaluation of this project is done with the proposed Ensemble+PCA+ResNet-50, DCNN+PCA+ResNet-50 and XGBoost+PCA+ResNet-50 techniques. The ensemble model had the most diagonal dominance. It could classify properly 495 cases of class 0 and 491 instances of class 1 with only a few (14) misclassifications. This means that this model can be used to distinguish them for the two classes in the photos. The increase in off-diagonal values shows that the difference between cover and stego images is increasing from single classifiers to deep-feature and ensemble approaches. Precision is the true positive divided by the total categorised predicted stego images. True steno image recall reveals how many of the selected classifier was properly identified as genuine steno (raw data that contains communication) image.

4.8 Analysis of Precision and Recall

The proposed model was achieved with an accuracy of 99.48% which shows zero probability of misclassifying a positive case as negative. The high recall rate of 98.46% indicates that most of the stego images may be found, when a huge number of photos are provided. It is worth to mention the small variation in both precision and recall since it becomes quite critical for a steganalysis system to balance between real negatives and false positives. The model obtained an F1-score of 98.97% showing that it sufficiently balances these two metrics.

4.9 Ablation Analysis

The ablation analysis demonstrates the incremental contribution of each major component of the proposed framework. The standalone XGBoost model produced 92.34% accuracy. Adding PCA increased the performance to 93.41%. When ResNet-50 deep features were incorporated with PCA and XGBoost, the accuracy increased to 96.02%. For DCNN, the accuracy improved from 94.12% to 95.26% after PCA and further to 97.38% after incorporating ResNet-50 Table 3.

The ensemble configuration followed the same pattern:

Ensemble → 96.08%

Ensemble + PCA → 97.12%

Ensemble + PCA + ResNet-50 → 98.57%

This progressive improvement provides strong evidence that the individual methodological components contribute positively to the final classification performance.

Table 3. Ablation-Based Improvement

Configuration	Accuracy (%)	Improvement
XGBoost	92.34	—
XGBoost + PCA	93.41	+1.07
XGBoost + PCA + ResNet-50	96.02	+3.68
DCNN	94.12	—
DCNN + PCA	95.26	+1.14
DCNN + PCA + ResNet-50	97.38	+3.26
Ensemble Model	96.08	—
Ensemble + PCA	97.12	+1.04
Ensemble + PCA + ResNet-50	**98.57**	**+2.49**

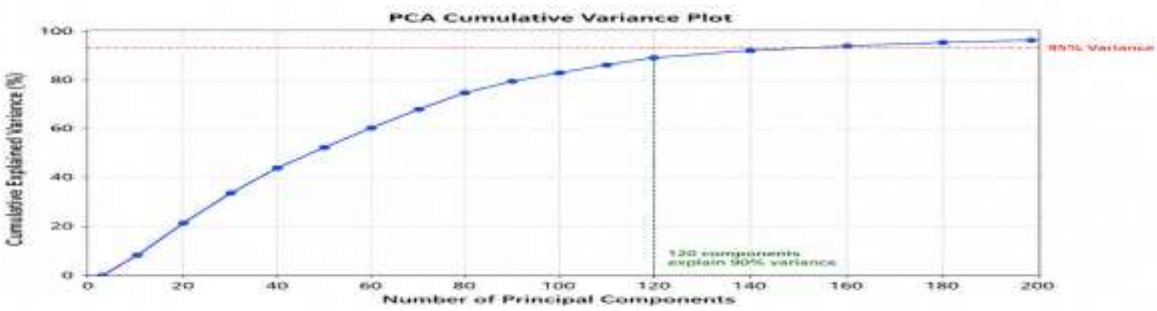

Figure 6: PCA Cumulative Explained Variance Plot

Figure 6 shows the cumulative proportion of variation explained by the components produced using PCA. In the graph, you can notice a steep rise in explained variance for the first few components, which is quickly followed by a more gentle rise to its plateau as you add more components. This contains only 120 principal components, saving around 90% of the overall variance in the data, and capturing all important information concerning the targets in the original feature space.

4.10 Comparative Analysis with Existing Approaches

Moreover, we have carried out a comparison examination of the proposed method against other state-of-the-art deep learning based steganalysis methods Table 4. The published comparison indicates that the prior approaches were accurate within the range of 82.4%-98.0%. This Ensemble + PCA + ResNet-50 works with success ratio 98.57%.

Table 4. Comparative Analysis

Study / Approach	Year	Technique	Accuracy (%)
Existing Approach 1	2023	EfficientNet-B0	82.4
Existing Approach 2	2024	CNN	97.0
Existing Approach 3	2024	DCI	92.2
Existing Approach 4	2025	CLPSTNet	98.0
Proposed Approach	**2025**	**Ensemble + PCA + ResNet-50**	**98.57**

The comparative results show that the suggested framework performs at least as well as the state-of-the-art methods. In lieu of a single architecture-specific model fine-tuned or not, shallow and deep feature learning, embedding dimension reduction, feature importance scoring, and ensemble classification are used.

4.11 Discussion

The comprehensive examination of these deep features shows that the high-resolution components (mostly edges) can be captured from low-fidelity embedded images within the framework of high-phase image steganalysis. A dense high dimensional representation using ResNet-50 can express and depict rich visual information that might not be fully captured without explicit description of each instance Fig. 7.

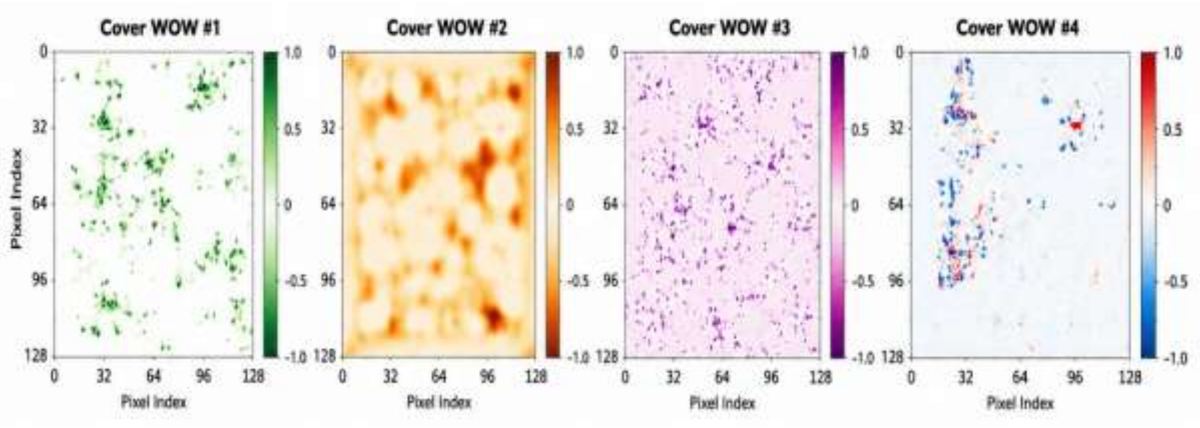

Figure 7: Heatmap Visualization of Cover WOW Residual Features

The single-source approach utilises principal component analysis (PCA) to learn a reduced dimensional representation of the concatenation vector and discovers the k-most relevant components in relation to SHAP values. The framework is improved with an ensemble classification step. For this classification procedure, we additionally use both XGBoost and DCNN to abstract complementary decision routes. All these put together and you will get a final accuracy of 98.57 percent Fig. 8.

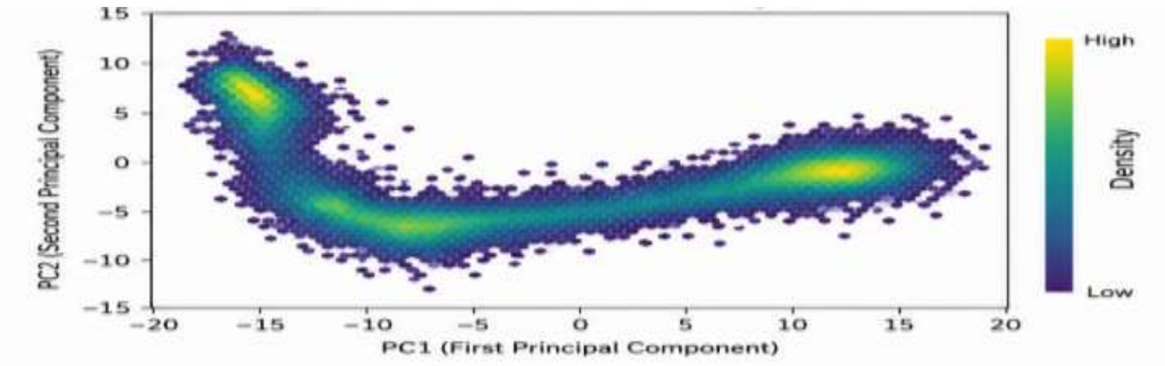

Figure 8: SHAP-Based Feature Importance and Impact Analysis

Three other SHAP analyses are shown in Figure 8 to explain the PCA features utilised as inputs in the classification model. Average absolute SHAP value, a feature target contribution (global), in the more mean signals the feature is global significant than other features. ASHAPmean=|Stdev(SMAP)reservation/average value. Figure 8 shows the dependent correlation between main parts and SHAP values of all features versus feature values on the result. Figure 8 where feature impacts are assigned and orientated per individual predictions.

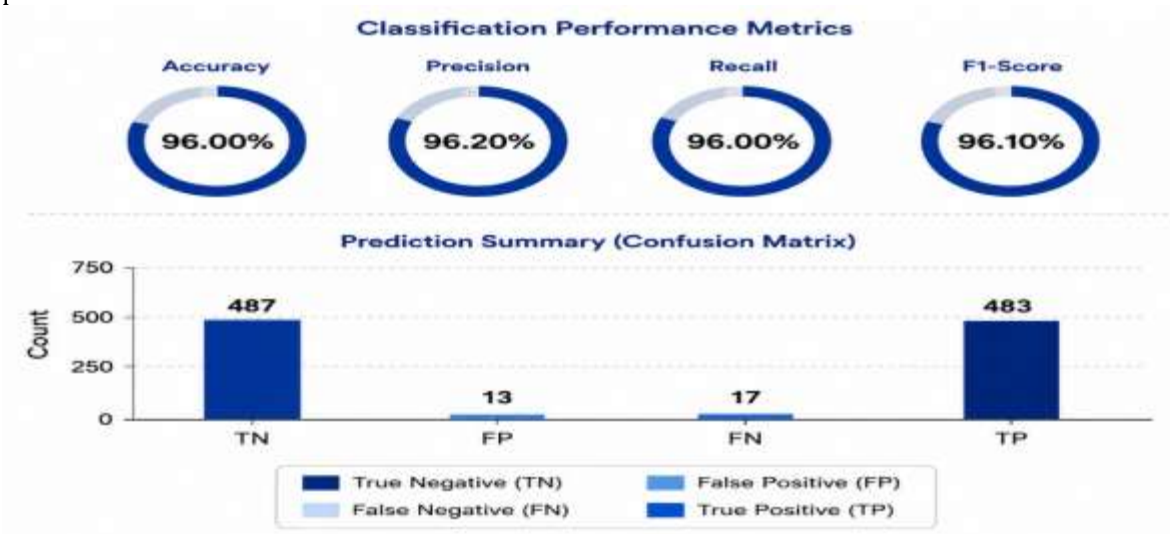

Figure 9: Confusion Matrix Analysis of XGBoost Models

Overall, the results support the following methodological relationship:

Deep Feature Extraction → PCA-Based Optimization → SHAP Feature Ranking → Ensemble Classification → Improved Steganalysis

These observations suggest that a combination characterisation of deep and hand-crafted features can provide a comprehensive description of an image, rather than the use of individual feature extraction methods Figure 9.

Figure 10: Confusion Matrix Analysis of DCNN Models

The proposed approach provides a practical basis for proper classification of cover-versus-stego. The above works also confirm the possibility of high-confidence steganalysis through ensemble classification, deep learning and hand-crafted features integration Figure 10. However, the results indicate that extracting deeper features is not adequate. PCA is one of the most conventional dimensionality reduction methodologies. SHAP is an explainable AI based methodology which ranks the obtained components according to their relevance. It keeps the significance of the data, but discards unnecessary dimensions and so allows the system to retain just relevant information.
The slope of the two curves in shows that the feature information essential for decryption and encryption is preserved. The high correlation among internal events for the retrieved feature values shows that the recovered features do capture the qualities of the original representation well enough to warrant and maintain validity for this feature-processing approach.

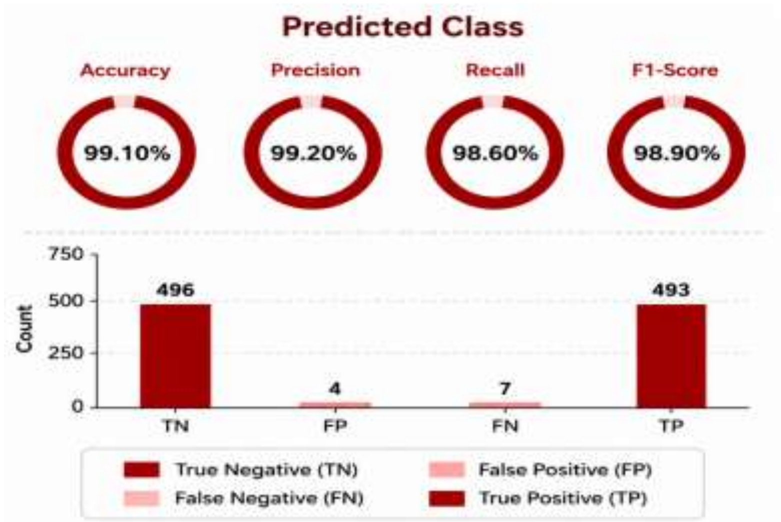

Figure 11: Confusion Matrix Analysis of Ensemble Models

The confusion matrices of XGBoost + PCA + ResNet-50, DCNN + PCA + ResNet-50 and Ensemble techniques both using PFA features are given in Fig 9. Cover graphics for Class 0 are presented Figure 11. The matrices are the representation of the classification of steganogrphy and cover.

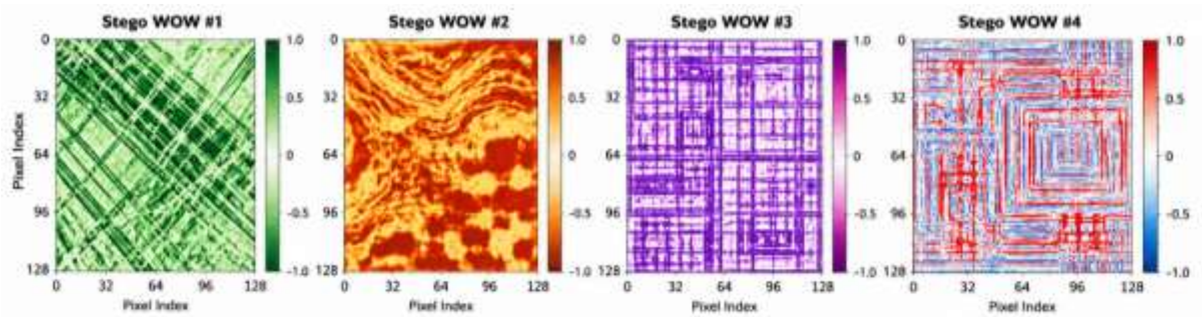

Figure 12: Heatmap Visualization of Stego WOW Residual Features

The heatmap representations of WOWs for four cover images and their accompanying stego images are shown in Figure 10. Stego WOW maps are always larger and more important than cover WOW maps that seem to indicate no or localised change. These differences indicate that the embedding process changes the picture residue qualities in a little way, which can be expressed by a function representation based on WOW Figure 12. In this work we demonstrate the effectiveness of residual-domain features in detecting inconsistencies between the cover map and the matching stego map which opens up the possibility for more depth deep feature extraction and classification. Because of the introduction of indirect communications.
The ensemble achieved its greatest performance with 496 true negatives, 493 true positives, 4 false negatives and 7 false positives. In most of the cases the important values along the main diagonal have been accurately classified to cover and stego pictures. This paper provides an ensemble based strategy which

achieves a lower error rate than the errors of individual classifiers (XGBoost & DCNN), showing a synergistic impact by combining their complementing learning capabilities.

5. Comparative analysis

Vital to realise that the effectiveness of steganalysis is largely dependent on the dataset, technique of embedding and splitting training/testing records. So the table should be more comparable and less like an equivalency Table 5. Hong et al. achieved an improvement of 82.40% on ALASKA#2 employing different payloads and embedding approaches over base studies in BOSSBase.

Table 5. Comparative Analysis of Existing Steganalysis Approaches

Author	Year	Technique	Accuracy
Xu et al. [1]	2016	CNN-based Steganalysis	87.3%
Yedroudj et al. [7]	2018	Yedroudj-Net	88.7%
Reinel et al. [38]	2021	GBRAS-Net	89.4%
Hong et al. [24]	2023	EfficientNet-B0 with Block-wise Pruning	82.4%
Proposed Study	**2025**	**Ensemble Model + PCA + ResNet-50**	**98.57%**

A specific comparison of performance of different existing techniques in digital picture steganalysis with the proposed framework is shown in Table 5. In the earlier work, Xu et al.107 employ CNN to show the superiority of learning convolutional features over hand-crafted features for identifying subtle steganographic modifications in images. Thus, Yedroudj-Net provides a way to generalise CNN frameworks for spatial domain steganalysis; and GBRAS-Net experimentally proves that the designed CNN architectures can really learn such unique residual representations.

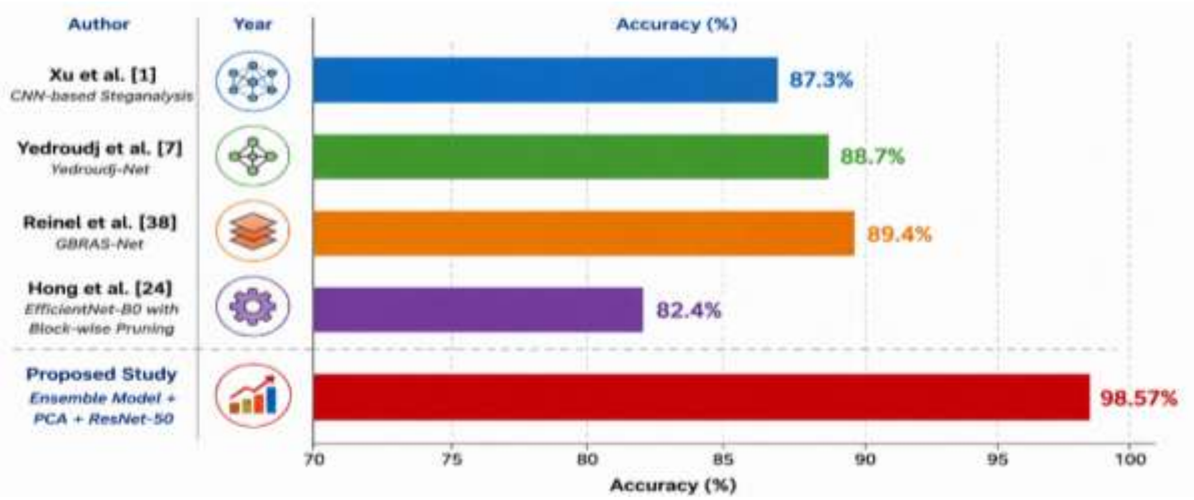

Figure 13: Comparative Accuracy of Existing and Proposed Digital Image Steganalysis Approaches

In our experimental setting on BOSSBase/WOW, we obtain evaluation values of 87.3 %, 88.7 % and 89.4 %. Hong et al. proposed to apply an efficient steganalysis method based on EfficientNet-B0, where block pruning was used Figure 13. It came out with 82.40% accuracy on ALASKA#2 but with a considerably reduced computational duration and model-size. In the end, low-scale structures can also be possible for steganalysis, depending on the dataset and test conditions. However, the proposed Ensemble Model with PCA and ResNet-50 achieved an aggregate accuracy of 98.57% in the experimental tests. The excellent results could be due to the great feature extraction power of ResNet-50, PCA based dimensionality reduction and SHAP based analysis of features and ensemble classification. In this study, we propose a method that utilises complementary feature representations and classifiers to separate cover and stego images more quickly than learning either a single classifier or feature representation.

6. Conclusion and Future Scope

6.1 Conclusion

In this paper, a digital picture steganalysis system based on machine learning techniques is proposed, in contrast to feature extraction from the same. This approach uses statistical, texture and residual feature representations along with deep feature learning to categorise a picture as stego or cover image. We use ResNet-50, a well-known deep feature extractor along with SPAM (Spatial Pyramid of UAVs)- Local Binary Pattern (LBP) besides co-occurrence matrices and wavelet-based neighbouring residuals. Features extracted from these photographs were pooled and subsequently processed using PCA (principal component analysis) to remove duplicated information and reduce dimensionality. Then, we did a feature analysis using SHAP to identify the most relevant component features and to make the selected representation more interpretable. Optimised features were also used in DCNN, XGBoost and ensemble classification techniques. Experimental results indicate that deep feature extraction, dimensionality reduction by PCA and ensemble learning are the best strategies for optimal classification. The suggested technique (Ensemble + PCA + ResNet-50) achieved 98.57% of accuracy, 99.48% of precision, 98.46% of recall rate and 98.97% of F1 score. Moreover, the confusion matrix analysis shows a clear distinction between both types of cover and stego pictures. Experimental results show that integrating DSRN with convolutional layers, by fusing features from shallow and hand-designed frequency and non-frequency components, improves precise steerable visual representation. Note that ensemble learning improves the top-level empirical classification performance, whereas PCA and SHAP are employed to improve interpretability and efficiency of features. In brief, the proposed framework indicates that many approaches will enhance the trustworthiness of digital image steganalysis, like deep learning with hybrid feature extraction methods, dimensionality reduction with auto-encoders, and explantory feature assessment using machine learning classification.

6.2 Future Scope

Future research may extend the suggested framework by employing cross-dataset evaluation to explore and assess its generalisation capacity toward additional collections of images, embedding methodologies, payload sizes and image quality circumstances. This covers the improvement of computational efficiency of ResNet-50, feature integration, use of PCA, SHAP and ensemble classification for real-time steganalysis in resource-constrained devices. You should try standard image manipulations on your system: scaling, noise and filtering, rotation, cropping, and colour adjustments. It might allow the framework to generalise, which would be useful for detecting new types of steganography through self-supervised, unsupervised and adaptive techniques. In an effort to circumvent heavy deep-learning frameworks, lightweight deep-learning frameworks are employed to search alternative solutions with little computational need, while maintaining competitive detection accuracy. Further, the proposed approach can be further expanded for video, audio and multi-modal steganalysis on digital photos to produce an enhanced digital forensics detection system. In future studies, we may potentially examine more advanced artificial intelligence algorithms that have higher interpretability than SHAP to explain steganalysis. Finally, the feature-extraction module may be embedded into real-world digital forensic frameworks to automate the study of vast image galleries and aid investigators to discover probable stego photos.

References

[1] G. Xu, H. Z. Wu, and Y. Q. Shi, "Structural design of convolutional neural networks for steganalysis," IEEE Signal Processing Letters, vol. 23, no. 5, pp. 708–712, 2016, doi: 10.1109/LSP.2016.2548421.

[2] G. Xu, H. Z. Wu, and Y. Q. Shi, "Ensemble of CNNs for steganalysis: An empirical study," in Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016, pp. 103–107, doi: 10.1145/2909827.2930798.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[5] J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2545–2557, 2017, doi: 10.1109/TIFS.2017.2710946.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.

[7] M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-Net: An efficient CNN for spatial steganalysis," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2092–2096, doi: 10.1109/ICASSP.2018.8461438.

[8] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181–1193, 2019, doi: 10.1109/TIFS.2018.2871749.

[9] S. Wu, S. Zhong, and Y. Liu, “A novel convolutional neural network for image steganalysis with shared normalization,” IEEE Transactions on Multimedia, vol. 22, no. 1, pp. 256–270, 2020, doi: 10.1109/TMM.2019.2920605.

[10] R. Zhang, F. Zhu, J. Liu, and G. Liu, “Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1138–1150, 2020, doi: 10.1109/TIFS.2019.2936913.

[11] M. Boroumand, M. Chen, and J. Fridrich, “Deep learning in steganography and steganalysis,” in Digital Media Steganography: Principles, Algorithms, and Advances. Elsevier, 2020, pp. 321–349, doi: 10.1016/B978-0-12-819438-6.00022-0.

[12] A. Selvaraj, A. Ezhilarasan, S. L. J. Wellington, and A. R. D. Sam, “Digital image steganalysis: A survey on paradigm shift from machine learning to deep learning based techniques,” IET Image Processing, vol. 15, pp. 504–522, 2021, doi: 10.1049/ipr2.12043.

[13] T. Pevný, P. Bas, and J. Fridrich, “Steganalysis by subtractive pixel adjacency matrix,” IEEE Transactions on Information Forensics and Security, vol. 5, no. 2, pp. 215–224, 2010.

[14] H. Zhang, X. Ping, M. Xu, and R. Wang, “Steganalysis by subtractive pixel adjacency matrix and dimensionality reduction,” Science China Information Sciences, vol. 57, no. 4, pp. 1–7, 2014, doi: 10.1007/s11432-013-4793-x.

[15] L. Fridrich and J. Kodovský, “Rich models for steganalysis of digital images,” IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868–882, 2012.

[16] V. Holub and J. Fridrich, “Random projections of residuals for digital image steganalysis,” IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, pp. 1996–2006, 2013.

[17] B. Song, P. Wei, S. Wu, Y. Lin, and W. Zhou, “A survey on deep-learning-based image steganography,” Expert Systems with Applications, vol. 254, 2024, Art. no. 124390, doi: 10.1016/j.eswa.2024.124390.

[18] M. H. Kombrink, Z. J. M. H. Geradts, and M. Worring, “Image steganography approaches and their detection strategies: A survey,” ACM Computing Surveys, vol. 57, no. 2, Art. no. 33, 2024, doi: 10.1145/3694965.

[19] S. Agarwal and K.-H. Jung, “Digital image steganalysis using entropy driven deep neural network,” Journal of Information Security and Applications, vol. 84, 2024, Art. no. 103799, doi: 10.1016/j.jisa.2024.103799.

[20] N. J. De La Croix, T. Ahmad, and F. Han, “Comprehensive survey on image steganalysis using deep learning,” Array, vol. 22, 2024, Art. no. 100353, doi: 10.1016/j.array.2024.100353.

[21] H. Kheddar, M. Hemis, Y. Himeur, D. Megías, and A. Amira, “Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions,” Neurocomputing, vol. 581, 2024, Art. no. 127528, doi: 10.1016/j.neucom.2024.127528.

[22] M. A. Wani and B. Sultan, “Deep learning based image steganography: A review,” WIREs Data Mining and Knowledge Discovery, vol. 13, no. 3, 2023, Art. no. e1481, doi: 10.1002/widm.1481.

[23] J. Hemalatha, M. Sekar, C. Kumar, A. Gutub, and A. K. Sahu, “Towards improving the performance of blind image steganalyzer using third-order SPAM features and ensemble classifier,” Journal of Information Security and Applications, vol. 76, 2023, Art. no. 103541, doi: 10.1016/j.jisa.2023.103541.

[24] E. Hong, K. Lim, T.-W. Oh, and H. Jang, “Lightweight image steganalysis with block-wise pruning,” Scientific Reports, vol. 13, Art. no. 16148, 2023, doi: 10.1038/s41598-023-43386-2.

[25] N. Farooq and R. N. Mir, “Image steganalysis using deep convolution neural networks: A literature survey,” International Journal of Sensors, Wireless Communications and Control, vol. 14, no. 4, pp. 247–264, 2024, doi: 10.2174/0122103279296370240529075507.

[26] O. A. Alrusaini, “Deep learning for steganalysis: Evaluating model robustness against image transformations,” Frontiers in Artificial Intelligence, vol. 8, 2025, Art. no. 1532895, doi: 10.3389/frai.2025.1532895.

[27] X. Gao, J. Yi, L. Liu, and L. Tan, “A generic image steganography recognition scheme with big data matching and an improved ResNet50 deep learning network,” Electronics, vol. 14, no. 8, Art. no. 1610, 2025, doi: 10.3390/electronics14081610.

[28] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002, doi: 10.1007/b98835.

[29] T. Filler, J. Judas, and J. Fridrich, “Minimizing additive distortion in steganography using syndrome-trellis codes,” IEEE Transactions on Information Forensics and Security, vol. 6, no. 3, pp. 920–935, 2011.

[30] P. Bas, T. Filler, and T. Pevný, “Break our steganographic system: The ins and outs of organizing BOSS,” in Information Hiding, Lecture Notes in Computer Science, vol. 6958. Springer, 2011, pp. 59–70.

[31] V. Holub and J. Fridrich, “Designing steganographic distortion using directional filters,” in 2012 IEEE International Workshop on Information Forensics and Security (WIFS), 2012, pp. 234–239.

[32] V. Holub, J. Fridrich, and T. Denemark, “Universal distortion function for steganography in an arbitrary domain,” EURASIP Journal on Information Security, vol. 2014, Art. no. 1, 2014.
[33] Y. Qian, J. Dong, W. Wang, and T. Tan, “Deep learning for steganalysis via convolutional neural networks,” in Proceedings of SPIE Media Watermarking, Security, and Forensics 2015, vol. 9409, 2015, Art. no. 94090J, doi: 10.1117/12.2083479.
[34] S. Tan and B. Li, “Stacked convolutional auto-encoders for steganalysis of digital images,” in 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014, pp. 1–4, doi: 10.1109/APSIPA.2014.7041565.
[35] J. Yang, X. Kang, E. K. Wong, and Y. Shi, “JPEG steganalysis based on DenseNet,” arXiv preprint arXiv:1711.09335, 2017.
[36] Z. Yang, K. Wang, S. Ma, Y. Huang, X. Kang, and X. Zhao, “IStego100K: Large-scale image steganalysis dataset,” in International Workshop on Digital Watermarking, Springer, 2019.
[37] Z. Zhou, S. Tan, J. Zeng, H. Chen, and S. Hong, “Ensemble deep learning features for real-world image steganalysis,” KSII Transactions on Internet and Information Systems, vol. 14, no. 11, pp. 4557–4572, 2020, doi: 10.3837/tiis.2020.11.017.
[38] T.-S. Reinel, H. B. Arteaga-Arteaga, M. A. Bravo-Ortíz, A. Mora-Rubio, D. Arias-Garzón, J. A. Alzate-Grisales, A. Burbano-Jacome, S. Orozco-Arias, G. Isaza, and R. Ramos-Pollán, “GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis,” IEEE Access, vol. 9, pp. 14340–14350, 2021, doi: 10.1109/ACCESS.2021.3052494.
[39] M. Płachta, M. Krzemień, K. Szczypiorski, and A. Janicki, “Detection of image steganography using deep learning and ensemble classifiers,” Electronics, vol. 11, no. 10, Art. no. 1565, 2022, doi: 10.3390/electronics11101565.
[40] R. Tabares-Soto, H. B. Arteaga-Arteaga, A. Mora-Rubio, M. A. Bravo-Ortíz, D. Arias-Garzón, J. A. Alzate-Grisales, S. Orozco-Arias, G. Isaza, and R. Ramos-Pollán, “Sensitivity of deep learning applied to spatial image steganalysis,” PeerJ Computer Science, vol. 7, Art. no. e616, 2021, doi: 10.7717/peerj-cs.616.
[41] “Steganalysis of context-aware image steganography techniques using convolutional neural network,” Applied Sciences, vol. 12, no. 21, Art. no. 10793, 2022.
[42] “Improving image steganography security via ensemble steganalysis and adversarial perturbation minimization,” Journal of Information Security and Applications, vol. 85, 2024, Art. no. 103835, doi: 10.1016/j.jisa.2024.103835.
[43] Z. He, R. Wu, and X. Wang, “Image steganalysis based on space-frequency domain interaction and collaborative enhancement,” Neural Networks, vol. 197, Art. no. 108486, 2026, doi: 10.1016/j.neunet.2025.108486.