International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 279-296
https://doi.org/10.70917/ijcisim-2026-3583
Article

Machine learning Model for Digital Image Steganalysis

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

[1*]NES's Gangamai College of Engineering Nagaon, Dhule, 424005, India 1jaanhavi1@gmail.com
[2]NES's Gangamai College of Engineering Nagaon, Dhule, 424005, India
nileshsuryawanshipatil088@gmail.com

Abstract: Digital image steganalysis plays a vital role in detecting hidden information embedded within images, yet identifying advanced steganographic techniques remains a significant challenge. This paper presents a hybrid feature extraction framework that combines deep features from Residual Network (ResNet)-50 with handcrafted statistical, texture, and transform-domain features to enhance detection performance. Using the ALASKA2 dataset, images are preprocessed and augmented before extracting Subtractive Pixel Adjacency Matrix (SPAM), Local Binary Pattern (LBP), co-occurrence, and wavelet residual features, which are fused with deep representations. Principal Component Analysis-SHapley Additive exPlanations (PCA)-(SHAP) -based feature selection approach reduces dimensionality while improving interpretability by retaining the most informative features. The selected features are classified using an ensemble of eXtreme Gradient Boosting (XGBoost) and Deep Convolutional Neural Network (DCNN) models. Experimental results demonstrate an accurate score of 98.57%, outperforming recent state-of-the-art methods and highlighting the effectiveness of the proposed model for robust digital image steganalysis.
Keywords: Digital image steganalysis, ResNet-50, XGBoost, Deep Convolutional Neural Network, Ensemble Model.

1. Introduction

In the current digital era, where the flow of confidential information over public networks has become a norm, the requirement for secure and stealthy communication methods has become more pressing than ever before [1]. One such method that has received considerable attention in recent years is digital image steganography, a method of hiding confidential information within digital images in a manner that conceals the presence of the hidden message itself [2, 3]. This stealthy communication method has found numerous applications in diverse fields, ranging from secure communication to copyright protection, intelligence, and even digital watermarking [4, 5].

Steganography does not encrypt message contents; it hides the existence of a message in a carrier such as an image, audio, or video [6,7]. Images are gaining popularity as a cover for hiding information because they are used more than any other type of file and contain lots of redundant information; thus, excess data can help hide more information [8]. Modern developments in steganographic techniques have led to significant advancements that make them harder to detect [9]. Steganographic techniques alter the image statistics and structures in a slight manner. This change may be so subtle that it can escape detection by human visual systems or the statistical detector [10,11].

Steganalysis is a growing area of study dealing with finding out if an image has hidden data inside it. However, before looking for hidden data in an image, there is a preliminary and incredibly important step associated with steganalysis that is critical to its success; this is the extraction of usable features from the image [12,13]. The extraction process must identify features that would provide the finest details associated with the presence of a hidden payload, such as residuals from noise during the embedding operation; pixels that contain disruptive values or patterns; and textures that are not consistent with the surrounding textures [14-16]. The viability of identifying the presence of a hidden payload increases with the strength and relevance of features obtained through analysis of an image [17-19].

Steganalysis techniques based on engineered features [20] have not proven to be robust against adaptive and content-aware embedding approaches [21]. DL approaches are efficient but are less interpretable and require large computational resources as well [22,23]. Machine Learning (ML) and DL based approaches, on the other hand, provide a balanced solution that offers performance and interpretability. ML algorithms and ensemble techniques effectively

Copyright: © 2026 by the authors

discern stego images from cover images. The performance of such models relies heavily on the quality and robustness of the extracted input features [24-26]. As shown in Figure 1, the processes of how ML is utilized and performed for image steganalysis exist.

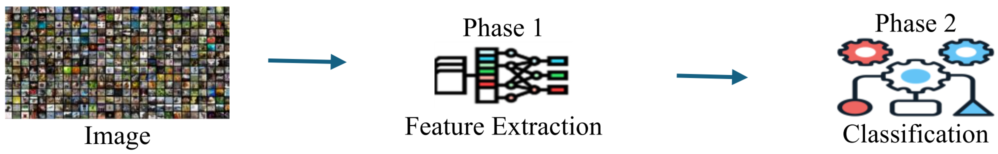


Figure 1: ML-based Steganalysis [27]

This study concentrates on the design of ML and DL algorithms solely for feature extraction in digital image steganalysis. Instead of developing a complete end-to-end detection system, the main objective of this study is to design a reliable ML and DL algorithm solely for the extraction of significant features from digital images for steganalysis purposes. The proposed algorithm combines statistical features, transform domain features, and image texture features to create a comprehensive set of features that can emphasize steganographic marks. The extracted features can be further used by ML algorithms or forensic software in the next phase. The contributions of this study are as follows:

Developed a unified feature-extraction scheme combining ResNet-50 deep features with statistical and residual descriptors for robust steganalysis representation.

Designed an enhanced preprocessing and augmentation pipeline to strengthen model generalization under diverse image variations.

Implemented a hybrid PCA–SHAP feature-selection strategy to retain only the most informative and discriminative features.

Employed a dual-model classifier using Ensemble XGBoost and DCNN to achieve high-confidence separation of stego and cover images.

By leveraging statistical descriptors, transform-domain residuals, and dimensionality reduction techniques, the proposed methodology offers a standalone feature extraction module capable of supporting a wide range of steganalysis systems, thereby enhancing their interpretability, adaptability, and effectiveness.

This paper is organized as follows: Section 1 presents the in-depth knowledge of the background and introduction of image steganalysis. Section 2 provides a summary of the previous work, and Section 3 presents the research methodology of the current research, along with the architecture of the methodology. Section 4 discussed the implemented results, and finally, the conclusion and possible future direction are stated in Section 5.

2. Literature Review

In this section, various previous studies based on steganalysis using ML and DL models are studied and summarized. Yang et al. (2025) [28] proposed a deep reinforcement learning-based Discrete Cosine Transform (DCT) image steganography method using Proximal Policy Optimization (PPO) for adaptive block selection and for dynamic reward weight adjustment, achieving a Peak Signal-to-Noise Ratio (PSNR) of 42.73 dB, Structural Similarity Index Matrix (SSIM) of 0.985, and reduced detection error rates to 11.2% on BOSSBase at 0.2 bits per pixel, demonstrating improved imperceptibility and security under low-payload conditions. Liu et al. (2025) [29] developed CLPSTNet, a curriculum learning-based multi-scale CNN incorporating Inception and dilated convolutions, and achieved PSNR of 43.62 dB, SSIM of 0.992, and decoding accuracy of 97.8% on ALASKA2, along with consistently low steganalysis scores across VOC2012 and ImageNet datasets 2504.16364v1. Shehab and Alhaddad (2025) [30] designed a Long Short-Term Memory (LSTM)-fused CNN for secure telemedicine steganalysis, replacing fully connected layers with LSTM to enhance temporal feature extraction, achieving 93.47% accuracy on BOSSBase and 91.82% on ALASKA2, with 32% fewer trainable parameters than conventional CNNs, thus offering a lightweight and accurate framework.

In 2025, Alrusaini [31] evaluated the robustness of five DL models for image steganalysis that were EfficientNet, SRNet, ResNet, Xu-Net, and Yedroudj-Net against common image transformations using the BOSSBase dataset. EfficientNet exhibited the highest resilience across all transformations, while Xu-Net and Yedroudj-Net showed significant degradation, particularly under noise, highlighting a need for more transformation-tolerant architectures. Li et al. (2025) [32] presented a steganalysis strategy based on the integration of active learning with off-policy deep reinforcement learning (DRL) and Differential Evolution for hyperparameter optimization, which enabled the learning model to selectively label only informative samples. Their strategy demonstrated high accuracy in steganalysis, with F-measures of 93.152% on BossBase 1.01 and 91.834% on BOWS-2, proving its high efficiency with only a few labeled samples. Ma et al. (2025) [33] designed a network enhancement framework based on the application of evolutionary algorithms for optimizing CNN training in steganalysis, focusing on improving Xu-Net and Yedroudj-Net. Their technique improved detection accuracy by 1.1% and 1.3%, respectively, and accelerated convergence by adjusting parameters in the feature enhancement modules dynamically. Wardhani et al. (2024) [34] presented a hybrid classical-quantum framework that integrated EfficientNet with quantum Convolutional Neural Networks (CNN), tested on the ALASKA2 dataset, where the hybrid framework showed 97.03% classification accuracy with EfficientNet-B3 and proved a 2.5% improvement over the classical-only CNN model, confirming its superiority in detecting hidden data. In 2024, Kuznetsov et al. [35] focused on improving the detection of Spread Spectrum Image Steganography (SSIS) by fine-tuning SRNet specifically on SSIS data, addressing its initial underperformance on this method. While the SSIS detection improved significantly after fine-tuning by achieving an accuracy of 0.98, the model's performance on other methods slightly decreased, suggesting a trade-off between specialization and generalizability. Martyniak and Czaplewski (2024) [36] examined architectural modifications of two leading CNNs, Yedroudj-Net and GBRAS-Net, designed for the steganalysis of digital images by testing 17 and 20 variants, respectively, to identify improvements in accuracy and parameter efficiency. Experiments found that CNN's Yedroudj-Net and GBRAS-Net both achieved an accuracy of 95% on the BOSSBase 1.01 dataset.

Kuznetsov et al. (2024) [37] conducted a detailed comparison of three SRNet implementations (custom, TensorFlow-based with an additional dense layer, and PyTorch-based), evaluating their steganalysis performance across different payloads on the BOSSBase and BOWS2 datasets, and found that the TensorFlow variant consistently yielded the highest F1-score of 85.95% at 0.5 bpp, while the custom model balanced performance and computational efficiency with fewer training epochs. Zhu et al. (2023) [38] introduced a modular image steganalysis approach, comprising three modules of pixel-level anomaly detection using the Subspace Approximation with Adjusted Bias (SAAB) transform. Results showed that the proposed model achieved comparable detection error rates against S-UNIWARD of 27.86 and WOW of 24.05. Mikhail et al. (2023) [39] proposed an ensemble transfer learning framework incorporating CNN, Inception, AlexNet, and ResNet50 models for detecting stego images. The ensemble method selected the best-performing classifier based on voting, achieving 86% accuracy on the BOSSBase dataset for payloads of 0.2–0.4 bpp, and demonstrating that the ensemble approach surpassed individual models in detection robustness. Liu et al. (2023) [40] developed JPEG-GraphNet, a graph attention learning-based deep model enhanced with a steganographic feature enhancement module (SFE) and Graph Attention Layers (GAL) to better preserve weak stego signals and capture global features in JPEG images. Findings showed that it achieved an accuracy of 91.7% at 0.4 bpp, outperforming SRNet and J-UNIWARD-based methods across multiple datasets, including ALASKA2, BOSSBase, and BOWS2 2023-4. Hammad et al. (2022) [41] proposed a low-complexity classification algorithm based on Segmentation-based Fractal Texture Analysis with Generalized Discriminant Analysis (SFTA-GDA). Experiments reported that the SFTA-GDA configuration achieved the highest accuracy of 90% with significantly reduced feature dimensionality compared to conventional high-dimensional steganalysis classifiers on the IStego100K dataset 2022-1.0.

Despite recent advances in the research, several research gaps have been identified. First, while Yang et al. [28] and Li et al. [32] used reinforcement learning for embedding and detection, an integrated adaptive framework remains unexplored. Liu et al. [29] and Wardhani et al. [34] proposed advanced architectures, but their real-time feasibility on low-resource systems is unclear. Third, Alrusaini [31] and Martyniak [36] revealed model weaknesses under transformations, indicating a need for more robust, generalized designs. Lastly, Zhu et al. [38] and Hammad et al. [41] offered lightweight models, but balancing simplicity with accuracy, especially for video or multimodal steganalysis, remains a challenge. In the next section, by considering these research gaps, a hybrid methodology for digital image steganalysis is proposed by focusing on the feature extraction algorithm.

3. Research Methodology

In this section, the proposed methodology for image-based forensic classification using the ALASKA2 dataset is described in detail. It begins with preprocessing steps like normalization, resizing, and noise reduction, followed by data augmentation through rotation, flipping, scaling, and noise injection. Feature extraction is then carried out using ResNet-50 for deep features, along with statistical and residual descriptors such as SPAM, co-occurrence matrices, wavelet-based, and LBP features. The extracted features are optimized using PCA and SHAP for dimensionality reduction and ranking. Finally, the selected feature vectors are classified using an ensemble of

XGBoost and DCNN models, and the system's performance is evaluated across both normal and stego classes. Figure 2 shows the architecture of the proposed methodology.

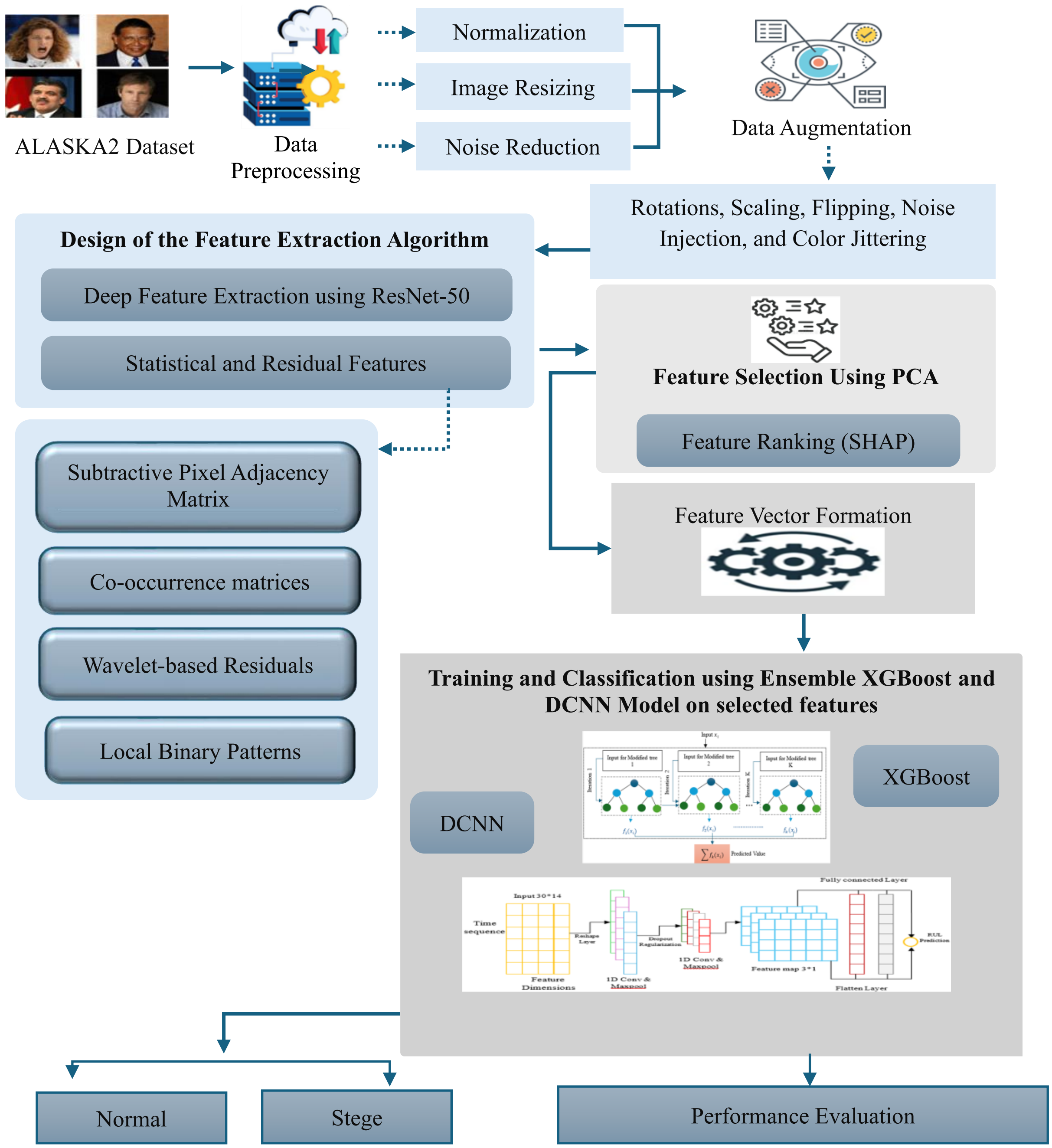


Figure 2: Proposed Methodology

3.1 Dataset: ALASKA2

The ALASKA2 dataset [42] is a large-scale and realistic image dataset designed to support research in digital image steganalysis. It is freely available at [42], and it contains a total of 80,000 color JPEG images, comprising 5000 images in the training set and 1,000 images in the testing set. All images are in JPEG format, compressed using three different quality factors, 95, 90, and 75, to reflect typical real-world image variations encountered in consumer photography. Each image is either a clean (cover) image or has been subtly modified to contain hidden information (stego), with an even distribution to support binary classification. Payloads are not fixed across the dataset: each image has a unique embedding density, with more complex images assigned higher payloads to maintain imperceptibility.

3.2 Data Preprocessing

The raw image data from the dataset undergoes a systematic data preprocessing [43] pipeline designed to enhance the quality and uniformity of inputs before feature extraction. Initially, all input images are subjected to image resizing to ensure a consistent spatial dimension.

$$I_{resized} = \text{Resize}(I_{raw}, H, W) \quad (1)$$

where H and W represent the target height and width, respectively. This is followed by normalization, wherein pixel intensities are scaled to a common range (typically [0,1] or mean-subtracted form), computed as

$$I_{norm}(x, y) = \frac{I_{resized}(x,y) - \mu}{\sigma} \quad (2)$$

where μ and σ denote the mean and standard deviation of the image. Noise reduction is applied using smoothing or denoising filters to suppress unwanted artifacts that could mask steganographic traces. Additionally, data augmentation techniques such as rotation (θ), horizontal and vertical flipping, noise injection, scaling, and color jittering are employed to improve model generalizability and robustness. These transformations produce a diverse set of augmented images I_{aug}, which enhances the learning model's ability to detect subtle stego patterns across varied spatial configurations and lighting conditions. The preprocessed data serves as the input for subsequent feature extraction modules.

3.3 Data Augmentation

In order to enrich the training sample diversity and boost the generalization capability of the model, data augmentation techniques were systematically applied to the preprocessed image of the dataset [44]. This step tries to produce changes that would happen in real life without changing the steganographic traces. The augmenting pipeline, together with geometric transformations like rotation R_θ (I), i.e., rotates the image I by angle θ and flips F_{axis} (I), i.e., the axis refers to horizontal or vertical flip. Scaling transformations $S_\alpha(I)$, where α is the scaling factor, adjust image size to simulate zoom effects. Additionally, color jittering alters brightness, contrast, and saturation to generate variants $J(I, \Delta_b, \Delta_c, \Delta_s)$ where Δ denotes the change in respective attributes. To make the model more robust against adversarial noise, random noise injection $I_{noisy} = I + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$, is performed. These augmented images retain the stego signatures while ensuring variability in spatial and spectral characteristics, thereby helping the DL and ensemble classifiers learn invariant features and improve performance across diverse stego manipulations.

3.4 Design of Feature Extraction using ResNet-50

ResNet-50 [45] is a deep CNN with 50 layers, built using residual blocks that facilitate the training of very deep architectures by employing shortcut connections to mitigate the vanishing gradient problem [46]. In the proposed steganalysis framework, ResNet-50 is used as the core DL model for extracting high-level semantic and discriminative features from both cover and stego images. Unlike traditional CNNs that learn a direct mapping $H(x)$, ResNet learns residual mapping:

$$F(x) = H(x) - x \Rightarrow H(x) = F(x) + x \quad (3)$$

Where x is the input and $F(x)$ is the learned residual. This allows the network to learn only the difference from identity mapping, improving convergence and generalization.

In this methodology, the preprocessed and augmented images are passed through the ResNet-50 network pre-trained on ImageNet, but the final classification layer (softmax) is removed, and intermediate convolutional features are extracted as deep feature representations. This strategy ensures that the model captures fine-grained texture and residual artifacts introduced during steganographic embedding. Figure 3 shows the architectural representation of the feature extraction process of ResNet50.

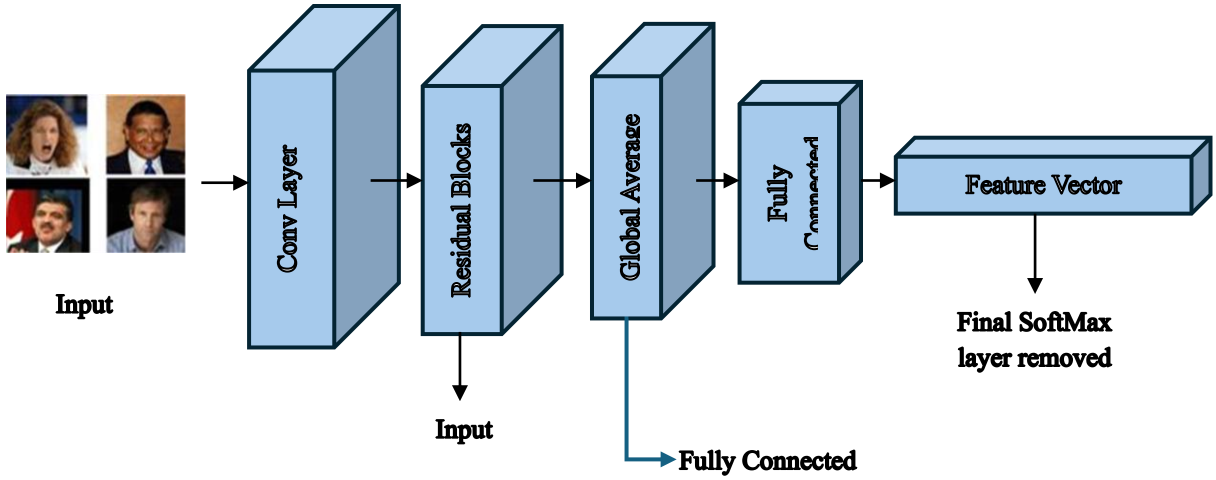


Figure 3: Feature Extraction using ResNet50.

The extracted feature vector $f_i \in \mathbb{R}^n$ for each image is then concatenated with handcrafted features (such as SPAM, LBP, wavelet residuals, etc.) to form a comprehensive feature set.

Pseudocode for Deep Feature Extraction Using ResNet-50

```
Input:
  - Preprocessed image dataset: D = {I₁, I₂, ..., I_N}
  - Pre-trained ResNet-50 model
Output:
  - Deep feature matrix: F = [f₁, f₂, ..., f_N]

1. Initialize ResNet-50:
  Load ResNet-50 with ImageNet weights
  Remove the final classification (fully connected) layer
2. Define Feature Extractor Function:
  def extract_features(image):
    Resize image to (224 x 224 x 3)
    Normalize pixel values to [0, 1] or mean-subtracted
    Pass the image through ResNet-50 up to the last pooling layer
    Flatten output feature map to vector f ∈ ℝⁿ
    return f
3. Initialize empty matrix F = []
4. For each image I in dataset D:
  f_i = extract_features(I)
  Append f_i to F
5. Return feature matrix F
```

3.5 Feature Selection Using PCA

PCA [47] operates by converting high-dimensional feature spaces into their lower-dimensional counterparts through its process of projecting original features onto orthogonal components that show maximum variance. The process of reducing feature dimensions through feature extraction from deep statistical and residual elements requires a vital step because it helps remove duplicate features while improving classifier accuracy. Mathematically, if $X \in \mathbb{R}^{n\times d}$represents the feature matrix with $n$samples and $d$features, PCA computes a transformation matrix $W \in \mathbb{R}^{d\times k}$such that:

$$Z = XW \tag{4}$$

Where $Z \in \mathbb{R}^{n\times k}$is the reduced feature representation and $k < d$. The matrix $W$is constructed from the top $k$eigenvectors of the covariance matrix $\Sigma = \frac{1}{n}X^{\top}X$, corresponding to the largest eigenvalues, captures maximum information with minimal features.

The selected components undergo additional refinement through the calculation of SHAP values, which serve for ranking the features [48]. The SHAP framework determines feature significance by measuring how each feature affects model predictions, thus enabling researchers to find the most effective features. The SHAP value ϕ_i for a feature i is computed as:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x) - f_S(x)] \tag{5}$$

where F is the set of all features, $S$is a subset, and $f_S(x)$is the model output when trained on features in S. This combination of PCA for unsupervised dimensionality reduction and SHAP for supervised feature importance ensures that the final feature vector retains maximum information content while remaining compact and interpretable, ready for downstream classification using ensemble XGBoost and DCNN models.

3.6 Feature Vector Formation

Following the extraction of deep and handcrafted features from each image, the Feature Vector Formation [49] step compiles and refines the combined information into a compact, discriminative profile for classification purposes. The deep features $f_{\text{deep}} \in \mathbb{R}^{2048}$, obtained from the final pooling layer of the ResNet-50 model, are concatenated with statistical and residual features $f_{\text{stat}} \in \mathbb{R}^{m}$ including SPAM, LBP, co-occurrence matrices, and wavelet-based residuals to form a unified high-dimensional vector $f_{\text{raw}} \in \mathbb{R}^{d}$, where $f_{\text{raw}} = [f_{\text{deep}} \parallel f_{\text{stat}}]$ and $d = 2048 + m$. PCA is applied to project the feature vector onto a lower-dimensional subspace, yielding:

$$f_{\text{pca}} = W^{\top}(f_{\text{raw}} - \mu) \tag{6}$$

Where $W$is the matrix of the top $k$eigenvectors of the covariance matrix and $\mu$is the mean vector. Finally, SHAP values ϕ_i are computed for each principal component to evaluate feature importance, and the top-ranked components are selected to form the final optimized vector $f_{\text{final}} \in \mathbb{R}^{r}$, which is then used for training an ensemble and DL classifiers for steganalysis.

3.7 Training and classification using Ensemble XGBoost and DCNN Model

Following the formation of a compact and discriminative feature vector, the classification phase employs a hybrid strategy integrating both traditional ML and DL paradigms to enhance the robustness and generalizability of steganalysis. Specifically, two parallel classifiers are trained: an ensemble-based Extreme Gradient Boosting (XGBoost) model and a Deep Convolutional Neural Network (DCNN). The XGBoost classifier [50] combines scalable features with regularization functions and high interpretability to operate on the optimized feature vectors $f_{\text{final}} \in \mathbb{R}^{r}$. The system builds an ensemble of decision trees through sequential construction, which creates new trees that reduce the residual errors produced by earlier trees based on the established objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \text{with} \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \tag{7}$$

Where l is the loss function (e.g., logistic loss), f_k are individual regression trees, T is the number of leaves, and $\lambda$and $\gamma$are regularization terms to penalize model complexity. Simultaneously, utilizing the deep feature maps (obtained straight from ResNet-50) instead of concatenation, the DCNN classifier [51,52] uses the convolutional layers to explicitly capture local spatial dependencies and hierarchical abstractions. The architecture of DCNN mainly contains convolutional layers, ReLU activations, batch normalization, and dropout layers for regularization. Finally, a SoftMax layer is applied that outputs probabilities for the class labels (cover vs stego). The performance of both classifiers is inferred using evaluation metrics including accuracy, precision, recall, and F1-score. The outputs are also compared to examine consistency, sensitivity to embedding schemes, and resilience to image distortions. This dual-model architecture makes use of the strengths of both paradigms. XGBoost's structured learning on tabular features and DCNN's end-to-end learning from image tensors lead to overall improvement in detection accuracy. Also, it ensures adaptability to varying steganographic techniques.

3.8 Detection and Performance Analysis

The last step of the suggested framework for steganalysis is to test the classification performance of trained models (XGBoost & DCNN) based on rigorous detection metrics. The objective is to evaluate the accuracy and reliability of the models in distinguishing between cover images (unaltered) and stego images (with inserted payload).

Let TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively. These values are derived from the confusion matrix. The performance metrics are computed as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{12}$$

$$\text{AUC} = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \tag{13}$$

Where $TPR = \text{Recall}$ and $FPR = \frac{FP}{FP+TN}$

4 Results and discussion

In this section, a complete evaluation of the proposed steganalysis system is provided using the ALASKA2 dataset based on the effectiveness of using a hybrid feature extraction method with an ensemble classification method. The hyperparameter setting used in the approach is provided in Table 1.

Table 1: Hyperparameter Settings

Hyperparameter	**Value**
No. of trees	300
Max. Depth	6
Learning Rate	0.05
Conv layers	4
Kernel Size	3*3
Activation	ReLU
Dropout	0.5
Batch Size	32

Figure 4 shows the preprocessing step where the images in the ALASKA2 dataset were resized to have a uniform resolution and normalized to have an intensity range of [0,1]. This step ensures that the inputs to the ResNet-50 architecture are standardized and that the training process is not biased due to the differences in the magnitudes of the pixels. It can be observed from the figure that the cover images and stego images maintained their features after the normalization process.

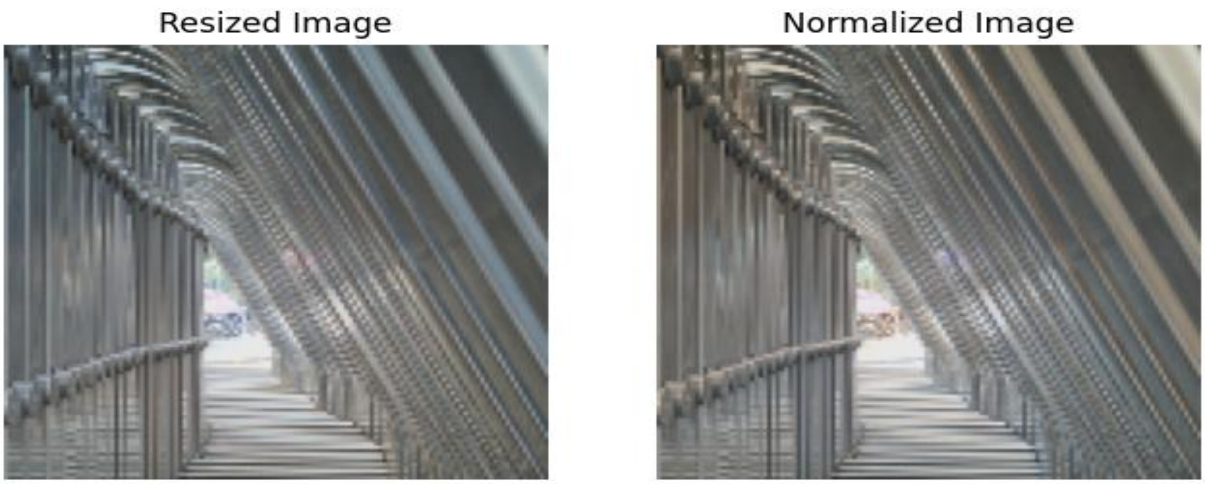


Figure 4: Preprocessing stage showing resized and normalized image

The result of data augmentation for both, as shown in Figure 5 (a) cover, and 5 (b) stego samples, includes rotation, scaling, flipping, noise addition, and color jitter. These are done to increase the size of the training data and prevent overfitting. The data augmentation process for cover images is done in such a way that the natural image is retained. For stego images, the patterns of the embedded signal are also maintained.

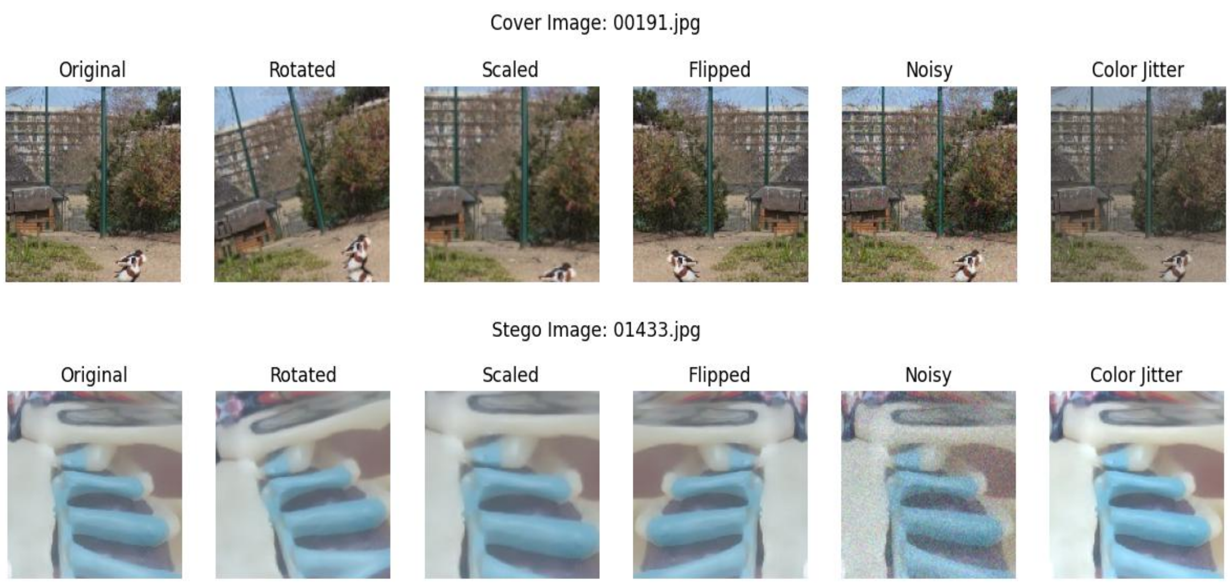


Figure 5: Data augmentation (a) Cover image (b) Stego image

The cumulative variance plot in Figure 6 reveals that the first 50 principal components capture approximately 60% of the dataset's variance. This balance optimizes computational efficiency while preserving most discriminative information for classification.

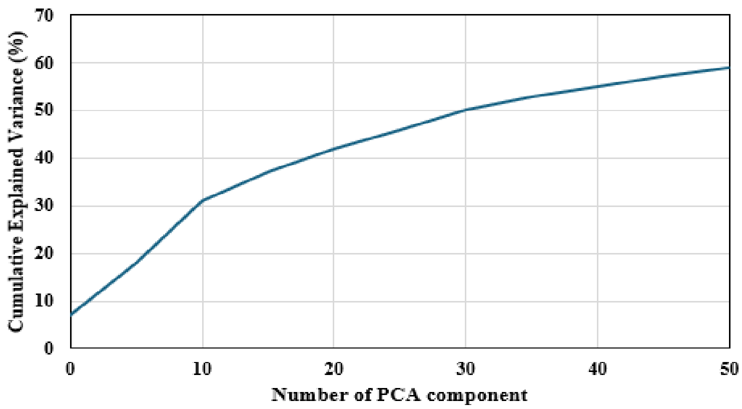


Figure 6: PCA cumulative variance plot

The PCA scatter plot in Figure 7 demonstrates how the first two components of the data spread out across different regions, which shows that the classification challenge arises from the overlapping areas between cover and stego clusters that need more complex solutions than basic linear boundaries.

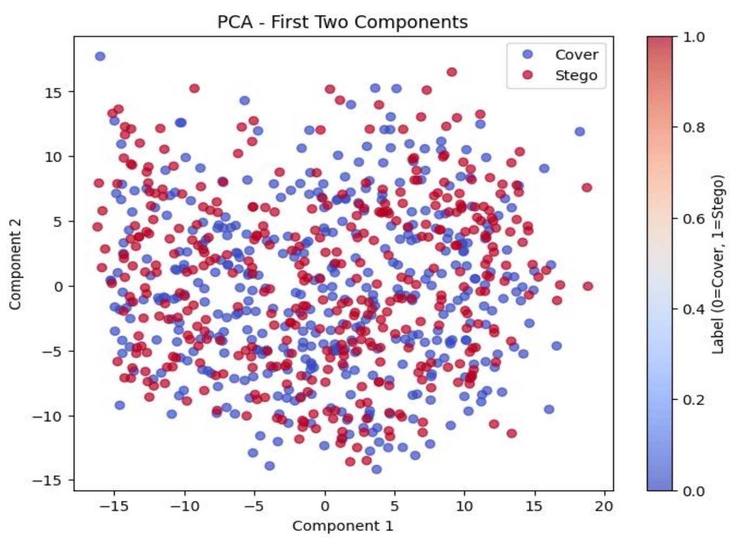
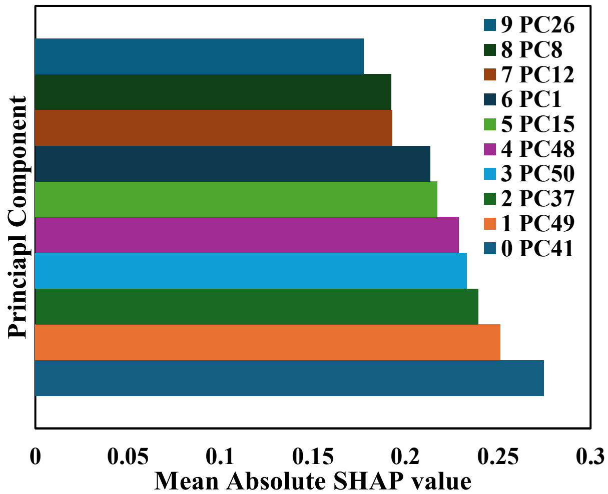
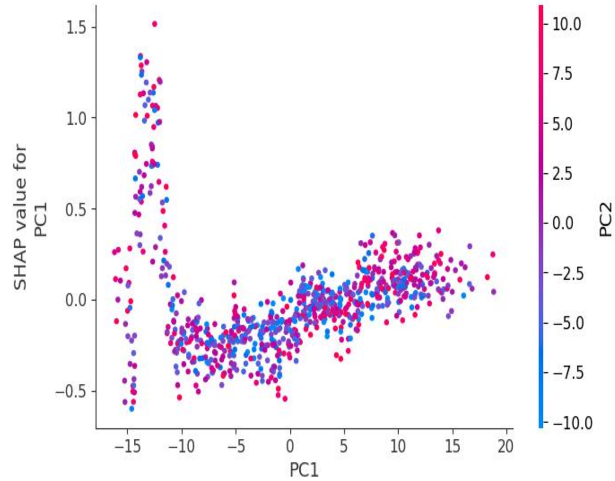


Figure 7: PCA scatter plot

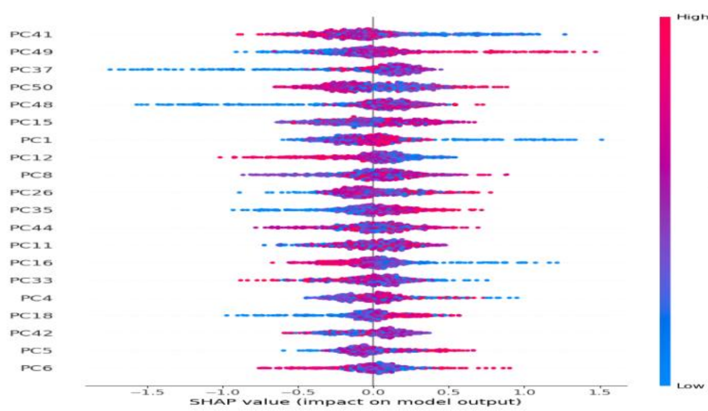
In Figure 8 (a), a bar chart of the mean SHAP values for all principal components is presented, again highlighting the dominance of the top-ranked PCs from before. This verification of the ranking and representation further supports the accuracy of the feature selection process. In Figure 8 (b), this is further expanded to include a SHAP beeswarm plot, which illustrates the role of positive SHAP values in marking predictions as stego images and negative values in marking as covers. The color bar represents the ranges of high-impact feature values. On the other hand, Figure 8 (c) presents a SHAP dependency plot, which highlights the connection between two high-impact principal components. The gradient bar represents non-linear relationships, which suggest that the boundary of the classification is a combination of multiple features, not individual ones.



8(a)



8(b)



8(c)

Figure 8: (a) Mean SHAP value bar chart, figure 8(b) SHAP dependency plot and figure 8 (c) SHAP beeswarm plot of feature impacts.

The ResNet-50 feature vectors for cover (blue) and stego (red) images are visualized in the two principal components in Figure 9. The classes are somewhat mixed, which means that there are discriminative features, but the classes are not perfectly separated in this feature space. The steganalysis task is difficult; that is, the difference is too little, so a very sophisticated classifier is needed to distinguish them.

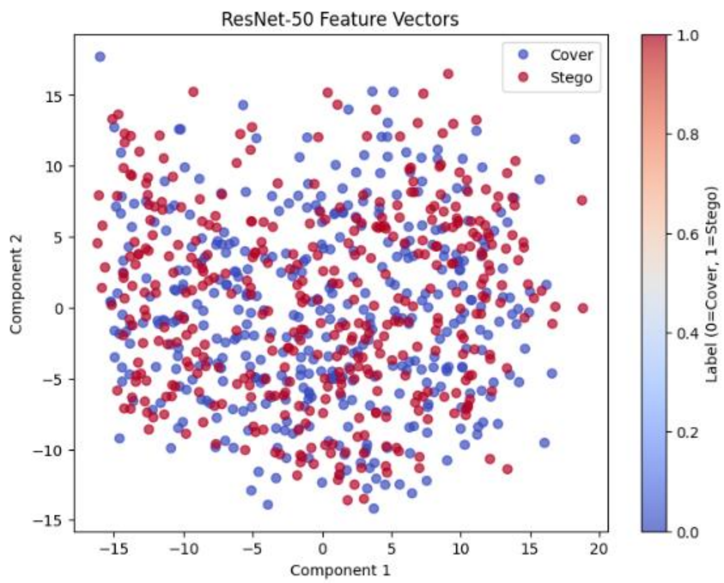


Figure 9: ResNet-50 feature vectors

The heatmaps presented in Figure 10 correspond to the Cover Wavelet Obtained Weights (WOW) and Stego WOW images for four samples each. The cover images displayed on the top row contain less high-intensity areas, meaning lower embedding artifacts. In contrast to this, the bright regions in the stego images (bottom row) are more scattered and concentrated, signifying the pixel-based alteration effects of the WOW stego. The distinct difference in intensity patterns exhibits traces of embedding, which can be used by a steganalysis algorithm for classification.

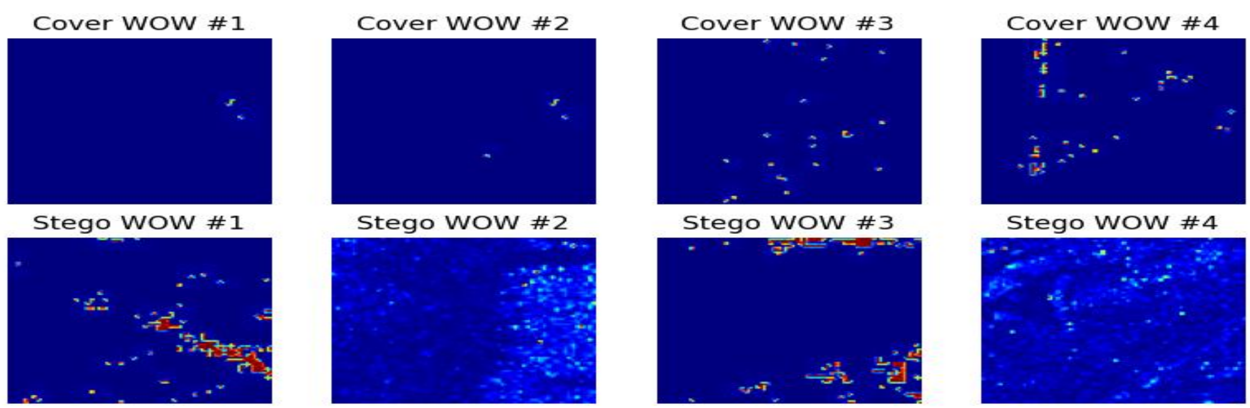


Figure 10: Heatmap visualizations comparing Cover WOW and Stego WOW

The blockchain metadata records for image files are displayed in Table 2. The first record is the "Genesis Block" in index 0, followed by two successive records. The metadata records comprise index, date, image path, image hash, previous block hash, and current block hash. Data integrity is preserved by cryptographically linking through previous_hash, which means that any alteration of the previous image would make the next blocks invalid, thus the images are secured on the blockchain.

Table 2: Blockchain-based metadata records

No.	Timestamp	Image path	Image hash	Previous hash	Hash
0	1748939758.662975	Genesis Block	0	0	65a050f352800b4c9c60cdb38f5dcc24bbd15035927ad5e3c54456fae2a1769

1	1748939758.6645162	/content/output_wow_cost/cover/00001 - Copy.jpg	5777a411d777af7675663c791342fe4a332bd485dffe50a453bde0f585782b75	65a050f352800b4c9c60cdb38f5dcc24bbd15035927ad5e3c54456fae2a1769	261f376219ee3c909f710a54e9afc8e50aa10da74d02f796b9850834460f70d8
2	1748939758.6652157	/content/output_wow_cost/cover/00001.jpg	49631c65a3989bef22d98469163a0ee0d0f4699ac086ff648e76460f35681af2	261f376219ee3c909f710a54e9afc8e50aa10da74d02f796b9850834460f70d8	be6a67cc9f9d3769716241b0438351e73f8fb4bd3eef98e2ba96f328700d5a0

Figure 11 below shows a graphical representation of the comparison between the original and decrypted feature vectors for the first 100 values. The two graphs that show the original and the decrypted features are almost identical, and this shows that the decryption process works well in restoring the original information. The almost identical graph shows that the encryption and decryption process used in the system is reliable. It is important to ensure that the feature information is not affected during the decryption process, especially when the features are used for classification in ML. This is a clear indication that the security processes incorporated into the system do not affect the quality of information used for analysis.

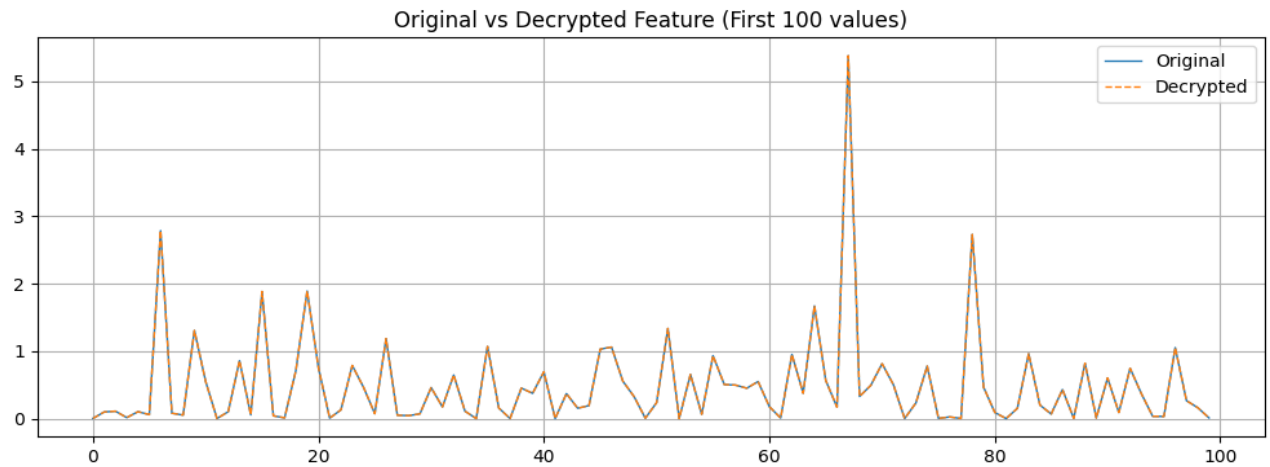
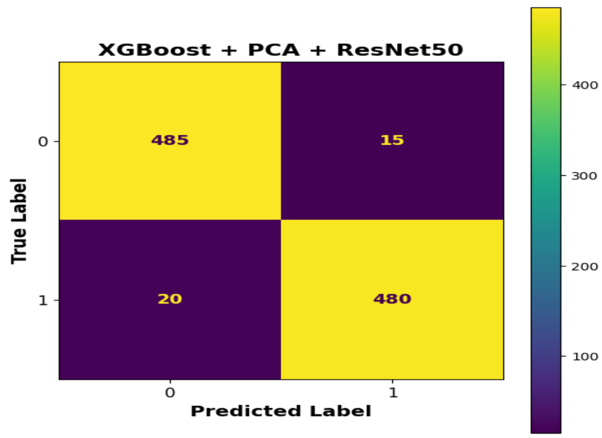
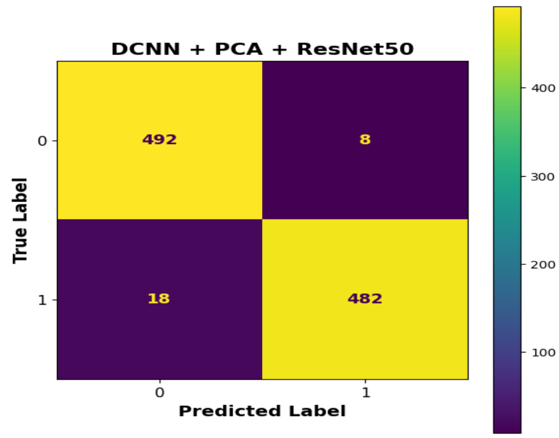


Figure 11: Visual comparison between the original and decrypted feature vectors

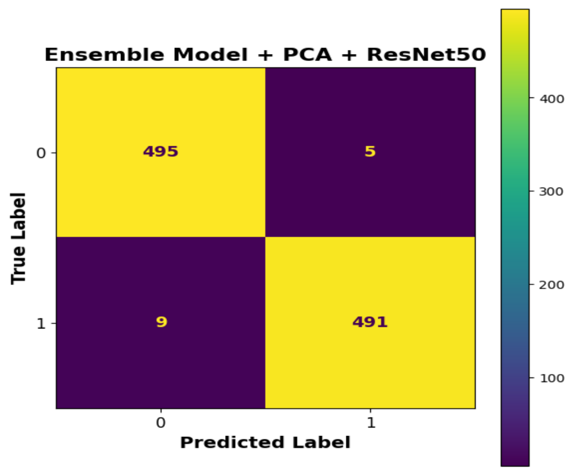
The confusion matrices for the three best-performing models, XGBoost + PCA + ResNet50, DCNN + PCA + ResNet50, and the proposed Ensemble Model + PCA + ResNet50, are shown in Fig. 12 (a), 12 (b), and 12 (c). The XGBoost-based model has a moderate number of false positives and false negatives, but the DCNN-based model shows better class separation with fewer errors. The proposed ensemble model has the best diagonal dominance, correctly classifying 495 samples of class 0 and 491 samples of class 1 with only 14 errors, thus having the best discriminative ability. The progressive decrease in the off-diagonal values of the confusion matrices of the three models confirms that the combination of deep feature extraction, PCA, and ensemble learning significantly improves binary classification accuracy.



12 (a) XGBoost



12 (b) DCNN



12 (c) Ensemble Model

Figure 12: Confusion Matrix of XGBoost, DCNN and Ensemble Model

The performance of various learning approaches for classification is compared in Table 3, considering various features of engineering and fusion settings. The accuracy of the baseline models (XGBoost and DCNN) is moderate, but it improves with ensemble learning and dimensionality reduction techniques via PCA. The addition of deep feature learning through ResNet50 further improves the classification accuracy of all models. Among all the models, the Ensemble Model + PCA + ResNet50 performs the best, with an accuracy of 98.57%, precision of 99.48%, recall of 98.46%, and F1-score of 98.97%, thus establishing its superiority in terms of generalization and robustness.

Table 3: Ablation study

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+ResNet50	96.02	95.54	95.88	95.71

DCNN+PCA+ResNet50	97.38	97.01	97.22	97.11
Ensemble Model+PCA+ResNet50	98.57	99.48	98.46	98.97

5 Comparative analysis

Table 4 below shows a comparative study of the latest state-of-the-art techniques reported in the literature for the same or similar classification problems. The previous studies conducted using EfficientNet-B0, CNN, DCI, and CLPSTNet reported accuracies between 82.4% and 98%, showing consistent improvement in DL-based techniques over the past few years. The current study (2025), which combines an Ensemble Model with PCA and ResNet50, outperforms all previous methods by reporting an accuracy of 98.57%, thus setting a new standard. This shows the strength of combining dimensionality reduction techniques, deep feature learning, and ensemble techniques for improving classification accuracy and performance.

Table 4: Comparative Analysis

Author	Year	Technique	Accuracy
Hong et al. [53]	2023	EfficientNet-B0	82.4%
Wardani et al. [34]	2024	CNN	97%
Lerch-Hostalot et al. [54]	2024	DCI	92.2%
Liu et al. [29]	2025	CLPSTNet	98%
Proposed Study	2025	Ensemble Model+PCA+ResNet50	98.57%

6 Conclusion and future scope

As digital image steganography methods become increasingly complex, the existing steganalysis systems are finding it increasingly difficult to accurately identify the hidden messages in images. A powerful hybrid technique for digital image steganalysis is proposed in this paper based on discriminative feature learning. The proposal incorporates features from ResNet-50 to learn the embedded characteristics of the images. The included features contain SPAM, LBP, co-occurrence, and wavelet features. The enhanced augmentation and preprocessing module helped to improve the generalization ability of the model. Meanwhile, the PCA–SHAP approach ensured high interpretability when carrying out the dimensionality reduction, as it selects the most salient components only. Experimental assessment on the ALASKA2 dataset demonstrated that the Ensemble Model utilizing PCA and ResNet-50 surpassed every other model with a high accuracy of 98.57%. Thus, the proposed approach of combining handcrafted features, DL, and ensemble classification is effective for high- confidence steganalysis. Future work aims to enhance the proposed methods for cross-dataset testing, implement them in real-time on resource-constrained devices, and cover video and multimodal steganalysis. Moreover, self-supervised learning along with lightweight models would be examined to increase the scalability and real-world forensic applicability of the proposed approach.

Acknowledgement

We would like to express our sincere gratitude to all those who contributed to the successful completion of this research. We are particularly thankful for the guidance, support, and resources that were made available throughout the course of this work. The insights and encouragement we received played a vital role in shaping this study, and we truly appreciate the assistance provided at every stage.

References

1. Peng, J., 2020. Secure covert communications over streaming media using dynamic steganography. Ph.D. dissertation, University of West London, London, U.K. https://repository.uwl.ac.uk/id/eprint/6943/ (accessed 30 June 2026).
2. Evsutin, O., Melman, A., Meshcheryakov, R., 2020. Digital steganography and watermarking for digital images: A review of current research directions. IEEE Access, 8, 166589–166611. https://doi.org/10.1109/ACCESS.2020.3022779

3. Subramanian, N., Elharrouss, O., Al-Maadeed, S., Bouridane, A., 2021. Image steganography: A review of the recent advances. IEEE Access, 9, 23409–23423. https://doi.org/10.1109/ACCESS.2021.3053998
4. Aberna, P., Agilandeeswari, L., 2024. Digital image and video watermarking: Methodologies, attacks, applications, and future directions. Multimedia Tools and Applications, 83(2), 5531–5591. https://doi.org/10.1007/s11042-023-15806-y
5. Wang, Z., Byrnes, O., Wang, H., Sun, R., Ma, C., Chen, H., Wu, Q., Xue, M., 2023. Data hiding with deep learning: A survey unifying digital watermarking and steganography. IEEE Transactions on Computational Social Systems, 10(6), 2985–2999. https://doi.org/10.1109/TCSS.2023.3268950
6. De La Croix, N. J., Ahmad, T., & Han, F. (2024). Comprehensive survey on image steganalysis using deep learning. array, 22, 100353.https://doi.org/10.1016/j.array.2024.100353
7. Choudhary, S., Husain, S., 2023. Analysis of cryptography encryption for network security and image steganography technique. Algorithms, 7(10). https://doi.org/10.55041/IJSREM26359
8. Ao, L., Lin, Y., Yang, Y., 2025. A phishing software detection approach based on R-tree and the analysis of the edge of stability phenomenon. Electronics, 14(14), 2862. https://doi.org/10.3390/electronics14142862
9. Kumar, A., Rani, R., Singh, S., 2023. A survey of recent advances in image steganography. Security and Privacy, 6(3), e281. https://doi.org/10.1002/spy2.281
10. Apau, R., Asante, M., Twum, F., Hayfron-Acquah, J.B., Peasah, K.O., 2024. Image steganography techniques for resisting statistical steganalysis attacks: A systematic literature review. PLoS ONE, 19(9), e0308807. https://doi.org/10.1371/journal.pone.0308807
11. AbdelRaouf, A., 2021. A new data hiding approach for image steganography based on visual color sensitivity. Multimedia Tools and Applications, 80(15), 23393–23417. https://doi.org/10.1007/s11042-020-10224
12. Bashir, B., Mir, R.N., Qureshi, S. Unveiling the evolving landscape of image steganalysis: A comprehensive survey, challenges, and emerging trends. https://doi.org/10.2139/ssrn.5579870
13. Luo, W., Wei, K., Li, Q., Ye, M., Tan, S., Tang, W., Huang, J., 2024. A comprehensive survey of digital image steganography and steganalysis. APSIPA Transactions on Signal and Information Processing, 13(1). https://doi.org/10.1561/116.20240038
14. Zhou, K., Li, J., Xiao, Y., Yang, J., Cheng, J., Liu, W., Luo, W., Liu, J., Gao, S., 2021. Memorizing structure-texture correspondence for image anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 33(6), 2335–2349. https://doi.org/10.1109/TNNLS.2021.3101403
15. 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
16. Rahman, S., Uddin, J., Zakarya, M., Hussain, H., Khan, A.A., Ahmed, A., Haleem, M., 2023. A comprehensive study of digital image steganographic techniques. IEEE Access, 11, 6770–6791. https://doi.org/10.1109/ACCESS.2023.3237393
17. Hardan, H., Khashan, O.A., Alshinwan, M., 2025. Edge-based data hiding and extraction algorithm to increase payload capacity and data security. Computers, Materials & Continua, 84(1). https://doi.org/10.32604/cmc.2025.061659
18. Chaganti, R., Ravi, V., Alazab, M., Pham, T.D., 2021. Stegomalware: A systematic survey of malware hiding and detection in images, machine learning models, and research challenges. arXiv preprint, arXiv:2110.02504. https://doi.org/10.48550/arXiv.2110.02504
19. Wu, Z., Guo, J., Zhang, C., Li, C., 2021. Steganography and steganalysis in voice over IP: A review. Sensors, 21(4), 1032. https://doi.org/10.3390/s21041032
20. Zhong, X., Das, A., Alrasheedi, F., Tanvir, A., 2023. A brief, in-depth survey of deep learning-based image watermarking. Applied Sciences, 13(21), 11852. https://doi.org/10.3390/app132111852
21. Ahmed, S.T., Hammood, D.A., Chisab, R.F., Al-Naji, A., Chahl, J., 2023. Medical image encryption: A comprehensive review. Computers, 12(8), 160. https://doi.org/10.3390/computers12080160
22. Shaikh, T.A., Rasool, T., Verma, P., Mir, W.A., 2024. A fundamental overview of ensemble deep learning models and applications: Systematic literature and state of the art. Annals of Operations Research, 1–77. https://doi.org/10.1007/s10479-024-06444-0
23. Tadesse, D.M., Kebede, S.R., Debele, T.G., Waldamichae, F.G., 2025. Application of deep reinforcement learning in various image processing tasks: A survey. Evolving Systems, 16(1), 13. https://doi.org/10.1007/s12530-024-09632-2
24. Abimannan, S., El-Alfy, E.-S.M., Chang, Y.-S., Hussain, S., Shukla, S., Satheesh, D., 2023. Ensemble multi-featured deep learning models and applications: A survey. IEEE Access, 11, 107194–107217. https://doi.org/10.1109/ACCESS.2023.3320042
25. Balasubramanian, A.A., Al-Heejawi, S.M.A., Singh, A., Breggia, A., Ahmad, B., Christman, R., Ryan, S.T., Amal, S., 2024. Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology. Cancers, 16(12), 2222. https://doi.org/10.3390/cancers16122222
26. Swetha, M., Godi, A.R., 2025. SLCCC-Net: Hybrid steganography and AI system for secure cancer classification from histopathological images in Internet of Medical Things applications. MethodsX, 103398. https://doi.org/10.1016/j.mex.2025.103398

27. De La Croix, N.J., Ahmad, T., Han, F., 2024. Comprehensive survey on image steganalysis using deep learning. Array, 22, 100353. https://doi.org/10.1016/j.array.2024.100353
28. Yang, R., Liu, L., Han, B., Hu, F., 2025. Deep reinforcement learning-based DCT image steganography. Mathematics, 13(19), 3150. https://doi.org/10.3390/math13193150
29. Liu, F., Zhang, T., Zhang, C., 2025. CLPSTNet: A progressive multi-scale convolutional steganography model integrating curriculum learning. arXiv preprint, arXiv:2504.16364. https://doi.org/10.48550/arXiv.2504.16364
30. Shehab, D., Alhaddad, M., 2025. Image steganalysis using LSTM fused convolutional neural networks for secure telemedicine. Frontiers in Medicine, 12, 1619706. 6 https://doi.org/10.3389/fmed.2025.161970
31. Alrusaini, O.A., 2025. Deep learning for steganalysis: Evaluating model robustness against image transformations. Frontiers in Artificial Intelligence, 8, 1532895. https://doi.org/10.3389/frai.2025.1532895
32. Bohang, L., Li, N., Yang, J., Alfarraj, O., Albelhai, F., Tolba, A., Shaikh, Z.A., Alizadehsani, R., Pławiak, P., Yee, P.L., 2025. Image steganalysis using active learning and hyperparameter optimization. Scientific Reports, 15(1), 7340. https://doi.org/10.1038/s41598-025-92082-w
33. Ma, Y., Zhang, X., Wang, J., Jin, R., Nasimov, R., Zhang, H., 2025. Digital image steganalysis network strengthening framework based on evolutionary algorithm. Scientific Reports, 15(1), 7472. https://doi.org/10.1038/s41598-025-91390-5
34. Wardhani, R.W., Putranto, D.S.C., Ji, J., Kim, H., 2024. Toward hybrid classical deep learning-quantum methods for steganalysis. IEEE Access, 12, 45238–45252. https://doi.org/10.1109/ACCESS.2024.3381615
35. Kuznetsov, O., Frontoni, E., Chernov, K., Kuznetsova, K., Shevchuk, R., Karpinski, M., 2024. Enhancing steganography detection with AI: Fine-tuning a deep residual network for spread spectrum image steganography. Sensors, 24(23), 7815. https://doi.org/10.3390/s24237815
36. Martyniak, R., Czaplewski, B., 2024. Architectural modifications to enhance steganalysis with convolutional neural networks. In: Proceedings of the International Conference on Computational Science. Springer Nature Switzerland, Cham, Switzerland, pp. 50–67. https://doi.org/10.1007/978-3-031-63751-3_4
37. Kuznetsov, A., Luhanko, N., Frontoni, E., Romeo, L., Rosati, R., 2024. Image steganalysis using deep learning models. Multimedia Tools and Applications, 83(16), 48607–48630. https://doi.org/10.1007/s11042-023-17591-0
38. Zhu, Y., Wang, X., Chen, H.-S., Salloum, R., Kuo, C.-C.J., 2023. Green steganalyzer: A green learning approach to image steganalysis. APSIPA Transactions on Signal and Information Processing, 12(1). https://doi.org/10.48550/arXiv.2306.04008
39. Mikhail, D.Y., Hawezi, R.S.H., Kareem, S.W., 2023. An ensemble transfer learning model for detecting stego images. Applied Sciences, 13(12), 7021. https://doi.org/10.3390/app13127021
40. Liu, Q., Yang, Z., Wu, H., 2023. JPEG steganalysis based on steganographic feature enhancement and graph attention learning. Journal of Electronic Imaging, 32(3), 033032–033032. https://doi.org/10.1117/1.JEI.32.3.033032
41. Hammad, B.T., Ahmed, I.T., Jamil, N., 2022. A steganalysis classification algorithm based on distinctive texture features. Symmetry, 14(2), 236. https://doi.org/10.3390/sym14020236
42. Kaggle, 2026. ALASKA2 Image Steganalysis. https://www.kaggle.com/ (accessed 22 May 2026).
43. Dehbozorgi, P., Ryabchykov, O., Bocklitz, T., 2024. A systematic investigation of image preprocessing on image classification. IEEE Access, 12, 64913–64926. https://doi.org/10.1109/ACCESS.2024.3395063
44. [44] Malik, K.R., Sajid, M., Almogren, A., Malik, T.S., Khan, A.H., Altameem, A., Rehman, A.U., Hussen, S., 2025. A hybrid steganography framework using DCT and GAN for secure data communication in the big data era. Scientific Reports, 15(1), 19630. https://doi.org/10.1038/s41598-025-01054-7
45. Khoria, S., Singh, R., 2025. Predicting lung cancer using ResNet-50 deep residual learning model using ReLU-memristor activation function. Engineering Journal, 29(3), 45–57. https://doi.org/10.4186/ej.2025.29.3.45
46. Kheddar, H., Hemis, M., Himeur, Y., Megias, D., & Amira, A. (2024). Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions. Neurocomputing, 581, 127528. https://doi.org/10.1016/j.neucom.2024.127528
47. Mehrabinezhad, A., Teshnehlab, M., Sharifi, A., 2024. A comparative study to examine principal component analysis and kernel principal component analysis-based weighting layer for convolutional neural networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 12(1), 2379526. https://doi.org/10.1080/21681163.2024.2379526
48. Wang, H., Liang, Q., Hancock, J.T., Khoshgoftaar, T.M., 2024. Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. Journal of Big Data, 11(1), 44. https://doi.org/10.1186/s40537-024-00905-w
49. Shanmugam, L., Shaji, A., Abinaya, S., Sourish, J.S., 2025. Enhanced gated MLP-based lung disease classification using grey wolf optimized (GWO) multi-feature extraction with cumulative learning. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3611029
50. Jamil, M., Trpcheska, H.M., Popovska-Mitrovikj, A., Dimitrova, V., Creutzburg, R., 2025. Advancing image spam detection: Evaluating machine learning models through comparative analysis. Applied Sciences, 15(11), 6158. https://doi.org/10.3390/app15116158
51. Al-Rawashdeh, R., Rahman, M.M., Niazi, M., 2025. Robust image steganography approach based on edge detection combined with CNN algorithm. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3582159

52. Farooq, N., Mir, R.N., 2024. Image steganalysis using deep convolution neural networks: A literature survey. International Journal of Sensors, Wireless Communications and Control, 14(4), 247–264. https://doi.org/10.2174/0122103279296370240529075507
53. Hong, E., Lim, K., Oh, T.-W., Jang, H., 2023. Lightweight image steganalysis with block-wise pruning. Scientific Reports, 13(1), 16148. https://doi.org/10.1038/s41598-023-43386-2
54. Lerch-Hostalot, D., Jimenez, D.M., 2024. Single-image steganalysis in real-world scenarios based on classifier inconsistency detection. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, pp. 1–6. https://doi.org/10.1145/3664476.3670911