Digital Image Forgery Detection Using Vision Transformer Features and Hybrid Dung Beetle–Addax Optimization with Attention-Based Siamese Network

Shimaa Janabi

Electrical Engineering Department, College of Engineering, Mustansiriyah University, Baghdad, Iraq

Email: shimaa.janabi@uomustansiriyah.edu.iq

Abstract

Digital image forgery has emerged as a significant issue, as numerous tools have been developed to alter images and advancements have been made in creating images with AI. Modern editing techniques, such as copy-and-move, splicing, object removal, and editing with AI help, can create images that look so real that they test traditional forensic methods. Convolutional neural networks (CNNs) are good at capturing local texture artefacts, but they have a hard time modelling long-range contextual connections. Transformer-based models, on the other hand, offer strong global representation learning but often create feature sets that are too complex and duplicate. This article suggests a way to find fake digital images. It's called Hybrid Dung Beetle–Addax Optimisation with Attention-Based Siamese Network (HDBAO-ASN). The framework uses the Vision Transformer (ViT) and ConvNeXt architectures to get global and local features that work well together. Then, a mix of the Dung Beetle Optimiser (DBO) and the Addax Optimisation Algorithm (AOA) is used to pick out traits that make the data different and cut down on duplicates. After the features are optimised, they are put through an Attention-Based Siamese Network, which makes it easier to learn the similarities between suspicious picture areas and better classifies forgeries.

The experiments were done with the CASIA 2.0, MICC-F220, MICC-F600, and MICC-F2000 files. The average scores for accuracy, precision, recall and F1-score of the proposed method were 97.5%, 97.2%, 97.0% and 97.1%, respectively and AUC was 98.7%. The framework was also demonstrated to be invariant to JPEG compression and geometric transformations and manipulation by AI. The outcomes show that HDBAO-ASN is a good and dependable choice for modern picture forgery detection and multimedia forensic uses.

Keywords: Digital Image Forgery Detection; Vision Transformer; ConvNeXt; Dung Beetle Optimizer; Addax Optimization Algorithm.

1. Introduction

The fast development of digital photography tools, social media sites, and artificial intelligence (AI) has changed how visual content is made and shared. With these changes, it's easier to communicate and develop content, but it's also more prevalent that digital pictures are changed. Today's image editing applications and AI-generated images can alter images in ways that are nearly indistinguishable from real images. This poses significant questions regarding the practice of journalism, cyber security, legal investigations, and authenticity of digital media [1–4].

This challenge has become even more difficult due to recent advancements in creative AI. Users can make or change images with great visual clarity using diffusion-based models like Stable Diffusion, DALL-E, and Midjourney. Compared to the traditional

manipulations, AI-generated fakes tend to have the same general appearance and subtle changes in meaning, making it increasingly hard to verify by hand [5–7]. This has made the ability to create strong image forgery detection systems which can be automatically executed an important study target.

Some common digital picture forgery techniques are copy and move, image splicing, object removal, and content replacement. The changes are typically applied in conjunction with other post-processing operations such as JPEG compression, scaling, spinning, blurring and adding noise. All such actions can make it easier to conceal physical evidence, and make standard detection techniques less effective and reliable in detecting fakes [8 – 10].

Current image forensic tools are much more effective with the help of deep learning. Images suffer from texture errors and manipulation artefacts which can be very well detected by a Convolutional Neural Network (CNNs) like ResNet, DenseNet or EfficientNet [11–13]. CNNs, however, rely on local receptive fields and are difficult to capture long-range dependencies between areas of the image that are distant from each other. When the forged regions are small, located apart, or there is a meaningful relation with other parts of the picture [14,15], this limitation is more prominent.

To overcome these issues, designs based on transformers have gained popularity as a solution. However, Vision Transformers (ViTs) leverage self-attention to capture the whole-image contextual relationships. This enables them to discover manipulation patterns that may be not easily recognizable when examining the image in isolation [16,17]. Though transformer models are useful, they tend to produce high dimensional representations of their features which can complicate computations and even contain irrelevant information to the classification task [18].

Therefore feature selection is a crucial step in making the classification more accurate and efficient, hence very important. Many studies have been conducted on the topic of feature selection for deep representations using metaheuristic optimisation techniques. When it comes to high-dimensional optimisation tasks [19–22], algorithms like Particle Swarm Optimisation (PSO), Grey Wolf Optimiser (GWO), Dung Beetle Optimiser (DBO), and Addax Optimisation Algorithm (AOA) have shown promise. Nevertheless many of the existing approaches are quite abrupt or are not able to strike a balance between exploring and exploiting, leading to sub-optimal feature selection performance.
Similarity learning is another important part of finding fakes. A lot of picture editing involves replacing, cloning, or duplicating parts of the image that have similar structures. Siamese networks are a good way to learn how different parts of a picture relate to each other and find odd connections. Attention mechanisms can also help representation learning by focusing on areas that are vulnerable to forgery and hiding background information that isn't important [23–25].

A lot of work has been made, but there are still some problems. CNN-based methods don't always take into account world context, and transformer-based methods may have issues with duplicate features and higher computational costs. Existing optimisation methods might not always create small and distinct feature subsets, and a lot of tools for finding fakes don't fully use similarity learning mechanisms in a single architecture. In this study, we suggest a Hybrid Dung Beetle–Addax Optimisation with Attention-Based Siamese Network (HDBAO-ASN) framework for finding fake digital images. This will help with these problems. Vision Transformer and ConvNeXt architectures

are combined in the suggested method to gather both global contextual information and local texture artefacts. We present a hybrid DBO-AOA approach to improve feature selection and lower dimensionality. An Attention-Based Siamese Network handles similarity learning and forgery classification.

The main contributions of this work are summarized as follows:

1. A dual-branch feature extraction framework combining Vision Transformer and ConvNeXt for learning complementary global and local image representations.
2. A hybrid DBO-AOA optimization strategy for selecting discriminative features and reducing feature redundancy.
3. An Attention-Based Siamese Network that enhances similarity learning and improves forgery classification accuracy.
4. Extensive evaluation on benchmark datasets demonstrating robustness against conventional and AI-generated image manipulations.

The test results demonstrate the effectiveness of the proposed framework for solving image forgery problems in the context of modern image forensic applications, outperforming several state-of-the-art CNN- and transformer-based approaches.

The remaining part of this paper follows the outline: In Section 2, the existing literature of the deep learning-based forgery detection, optimization-based feature selection, and attention mechanisms is discussed. In section 3, the proposed HDBAO-ASN framework (preprocessing, hybrid feature extraction, DBO-AOA optimization, and attention-based Siamese network) is described, possibly in more detail than some readers would expect. Experimental results, comparisons and robustness analyses are presented in Section 4. Finally, Section 5 wraps up the paper and proposes some directions for future research and limitations.

2. Literature Review

2.1 Deep Learning-Based Image Forgery Detection

Digital image forgery detection is increasingly employing deep learning methods since they can learn image characteristics indiscriminately from large collections of images. Copy-move forgery, picture splicing, object removal, and other types of manipulation are all common detection methods performed using Convolutional Neural Networks (CNNs). VGGNet, ResNet, DenseNet and EfficientNet are all architectures that have shown excellent performance in identifying texture mismatches, compression artifacts, and boundary irregularities that are associated with forgery regions [26–29].

Among the CNN-based models, ResNet is one of the first to introduce the concept of residual learning and deeper networks to produce better feature extraction. DenseNet improved the flow of information even more by using dense connectivity, and EfficientNet made computing more efficient by using compound scaling methods [30–32]. These representations have allowed these architectures to compete in forensic applications by learning them directly from image data in a hierarchical manner.

CNN based methods are effective, but require a significant amount of local receptive fields. Deeper layers expand the receptive field, but CNNs struggle to capture the long-range dependency between image regions which are far apart in space. This issue becomes particularly apparent with multiple modifications or when local texture

artefacts are more helpful in providing evidence of a crime than inconsistencies in the global environment [33,34].

Transformer-based designs have recently garnered a lot of attention as a means of addressing these challenges. The Vision Transformers (ViT) models rely on self-attention mechanisms to establish relations between the image patches throughout the image. This enables them to learn global contextual dependencies. ViTs perform better than the conventional CNN-based architectures on picture classification and multimedia forensic tasks [16,17]. This is because they are able to detect long-distance interactions and semantic inconsistencies.

Aside from ViT, a family of hierarchical transformer designs such as Swin Transformer [18] has been developed that makes computation simpler while preserving the ability to model context. Hybrid CNN-transformer models are also viable choices as they leverage on both global context and local texture extraction. Recent studies show that these kinds of hybrid frameworks are better at protecting against complex picture manipulations and material made by AI [35,36].

Transformer-based models tend to generate high dimensional representations of their features, which makes the computation more difficult and sometimes redundant information that the model already knows. Thus, there is still a need for effective feature selection techniques to improve the accuracy of classification and computer speed.

2.2 Optimization-Based Feature Selection

Since the deep networks often generate high-dimensional feature spaces, the selection of the features is an important part of the image forgery detection system. By reducing dimensionality, enhancing classification accuracy, lowering computational costs, and boosting model generalization, effective feature selection can be beneficial in many ways [37].

The feature subset selection has been widely applied using the metaheuristic optimization methods which are able to efficiently explore complex search space without the need of gradient information. One of the most popular population-based algorithms is Particle Swarm Optimization (PSO) which has shown good performance in image processing and pattern recognition applications. However, PSO has the tendency of premature convergence for high dimensional optimization problems [19].

Due to its good combination of exploration and exploitation, Grey Wolf Optimizer (GWO) has also been well studied in the field of feature selection. While very often GWO is superior to traditional evolutionary methods, it can suffer from local optima stagnation at later optimization stages [20].

Recently, some other optimization methods are becoming competitive, such as Dung Beetle Optimizer (DBO) and Addax Optimization Algorithm (AOA). DBO has excellent exploration capability and is very effective in maintaining the diversity of population, which can be applied to large-scale optimization tasks [21]. The convergence process for AOA is in contrast fast and the exploitation of the promising areas in the search space is efficient [22].

Both of the above algorithms have their own drawbacks. The refinement mechanisms for DBO may need to be more aggressive in their neighborhood, while AOA may fall into a local minimum if there is less diversity in the population. Thus, the synergy of these two algorithms is expected to provide a good solution in terms of finding

discriminative and compact feature subsets, while ensuring exploration/exploitation balance.

2.3 Similarity Learning and Attention Mechanisms

Many image manipulations involve duplicated, duplicated and cloned, or semantically related image regions, which makes similarity learning a crucial aspect of image forgery detection. Siamese networks learn similarity relationships by taking in pairs of inputs, embedding them in similar feature extraction branches, and comparing embeddings, which is an effective framework [23].

The Siamese, or similar architectures, have shown to be the most successful approach for copy-move forgery detection and image matching tasks in image forensic applications, due to their capacity of detecting duplicated structures as well as subtle manipulations. Siamese networks can be trained to capture discriminative similarity representations, which can boost the detection of forged regions, which could be visually consistent with adjacent content [24].

In addition, attention mechanisms have been used to improve the performance of deep learning in computer vision tasks. A multi-head attention captures the selective attention of neural networks to informative regions of images while ignoring irrelevant background information. The advantage of attention with respect to forgery localization, classification accuracy, and forgery robustness against challenging image transformations has been shown in recent studies [25,38,39].

The ability to incorporate the attention mechanism into the Siamese architecture is especially appealing for forgery detection, as it helps enhance similarity learning and highlights manipulation-sensitive features. However, there is limited research that integrates attention guided Siamese learning with transformer-based feature extraction and optimization-based feature selection into a single framework.

2.4 Research Gap

While significant advancements have been made in digital image forgery detection, there are still some challenges that have not been addressed. CNN-based approaches are well suited to local feature extraction, but lack the ability to support global contextual dependency. Transformer-based methods are effective in improving contextual modeling but often come with the expense of producing high-dimensional features, which can lead to computational challenges. Current optimization techniques could also have difficulty balancing exploration and exploitation, resulting in suboptimal feature selection.

In addition, most existing forgery detection approaches are not fully leveraging on the similarity learning mechanisms, given the intrinsic relationship that exists between image manipulation and duplicated or modified image regions. Thus, the combination of global contextual modelling, local texture analysis, optimization-based feature selection, and the attention-based similarity learning is far from being explored.

To address these limitations, this study proposes the Hybrid Dung Beetle–Addax Optimization with Attention-Based Siamese Network (HDBAO-ASN). The proposed framework combines Vision Transformer and ConvNeXt feature extraction, hybrid DBO-AOA feature selection, and an attention-enhanced Siamese architecture within a unified system designed to improve detection accuracy, robustness, and computational efficiency.

3. Proposed Methodology

This study proposes a Hybrid Dung Beetle–Addax Optimization with Attention-Based Siamese Network (HDBAO-ASN) framework for digital image forgery detection. The framework integrates image enhancement, hybrid feature extraction, optimization-driven feature selection, and attention-guided similarity learning. The overall workflow consists of image preprocessing, feature extraction, feature fusion, feature selection, and forgery classification. Figure 1 shows the overall architecture of the proposed HDBAO-ASN framework integrating image preprocessing, hybrid ViT–ConvNeXt feature extraction, DBO-AOA feature selection, and attention-based Siamese classification.

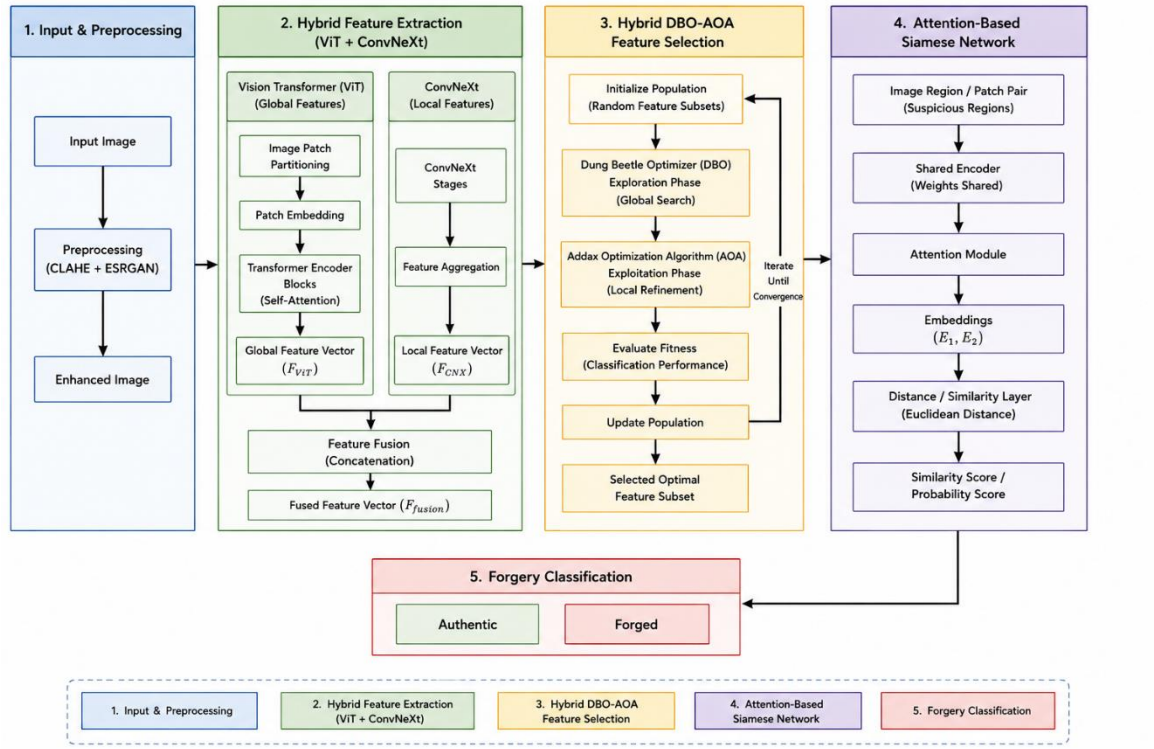


Figure 1. Architecture of the proposed HDBAO-ASN framework.

3.1 Dataset Preparation

Let the image dataset be represented as:

$$D = I_1, I_2, I_3, \ldots, I_n \qquad (1)$$

where I_i denotes the i-th image and n is the total number of samples.

All images are resized to 224 × 224 pixels before processing.

3.2 Image Preprocessing

Image preprocessing enhances manipulation traces that may be weakened by compression, noise, or resolution degradation.

CLAHE Enhancement

The CLAHE-enhanced image is obtained as:

$$I_{CLAHE} = CLAHE(I) \qquad (2)$$

where I is the original image and I_{CLAHE} is the contrast-enhanced image.

ESRGAN Super-Resolution

The enhanced image is further processed using ESRGAN:

$$I_{SR} = G(I_{CLAHE}) \qquad (3)$$

where G(.) denotes the ESRGAN generator and I_{SR} represents the super-resolved image.

The preprocessing stage improves the visibility of subtle forgery artifacts and boundary inconsistencies. Figure 2 shows Image preprocessing workflow showing CLAHE-based contrast enhancement followed by ESRGAN super-resolution to improve manipulation trace visibility.

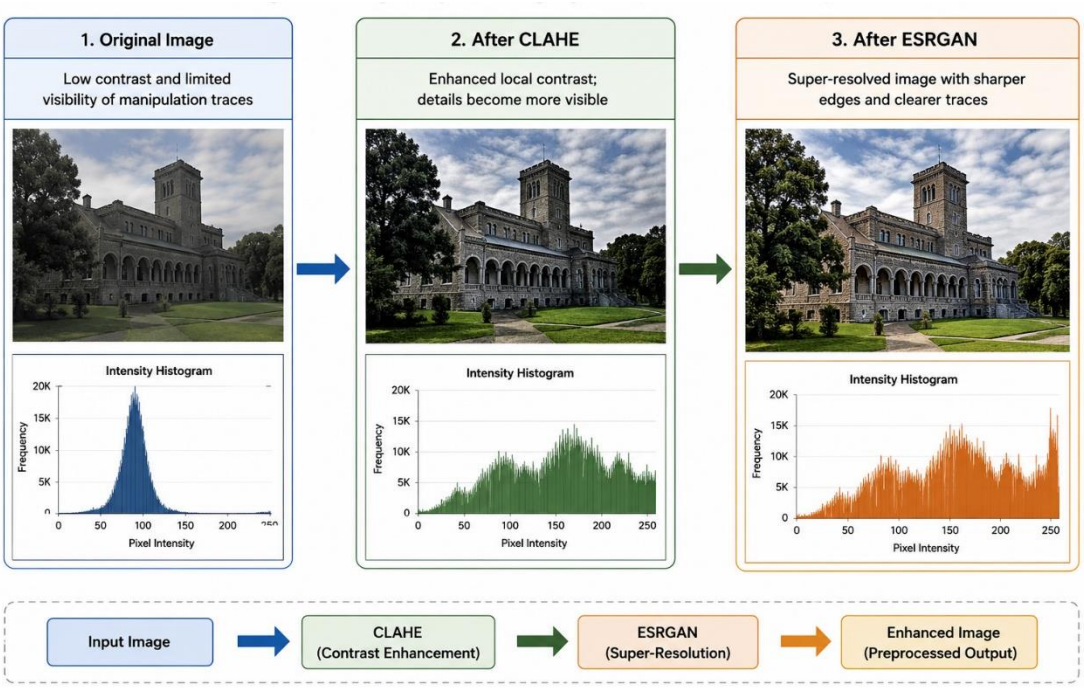


Figure 2. Image preprocessing workflow

3.3 Hybrid Feature Extraction

To capture both global and local image characteristics, Vision Transformer (ViT) and ConvNeXt architectures are employed.

Vision Transformer Branch

The image is divided into fixed-size patches:

$$X = x_1, x_2, x_3, \ldots, x_{\mathrm{m}} \qquad (4)$$

where m denotes the number of image patches.

The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = Softmax\left((QK^{\mathrm{T}})/\sqrt{d_{\mathrm{k}}}\right)V \qquad (5)$$

where:

Q = Query matrix

K = Key matrix

V = Value matrix

d_k = Attention dimension

The transformer feature representation is obtained as:

$$F_{ViT} = ViT(I_{SR}) \qquad (6)$$

The ViT branch captures long-range dependencies and global contextual inconsistencies associated with image manipulation.

ConvNeXt Branch

Local texture representations are extracted using ConvNeXt:

$$F_{Conv} = ConvNeXt(I_{SR}) \qquad (7)$$

where F_{Conv} denotes the convolutional feature vector.

ConvNeXt focuses on texture irregularities, edge discontinuities, and local forgery artifacts. Figure 3 shows Dual-branch feature extraction architecture combining Vision Transformer for global contextual representation learning and ConvNeXt for local texture analysis.

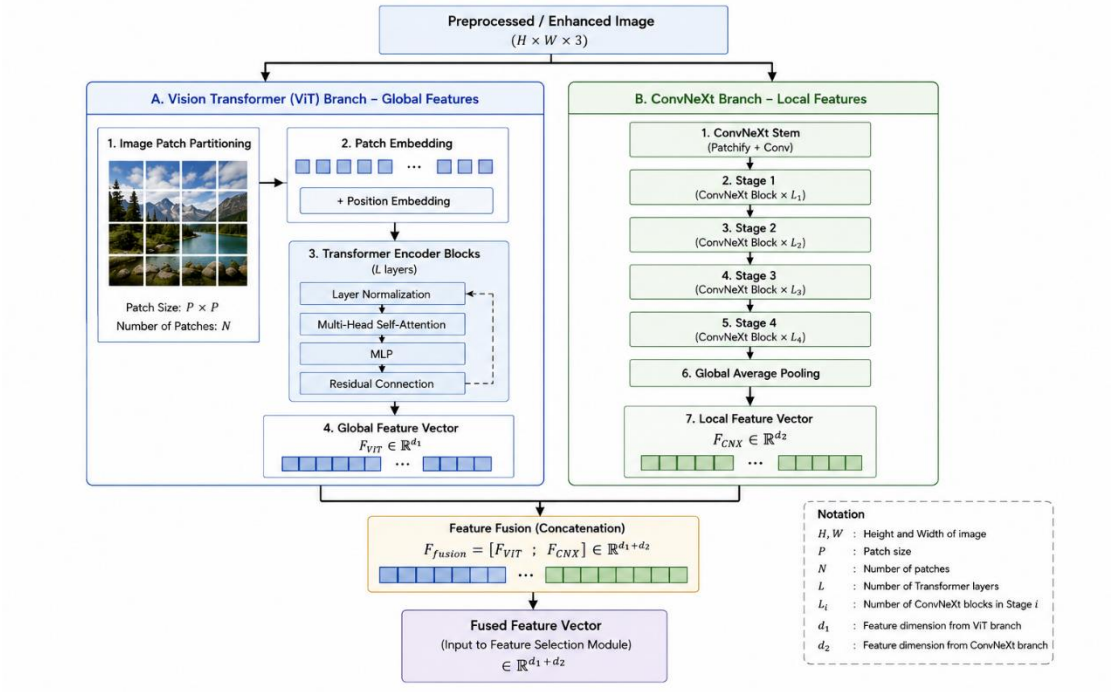


Figure 3. Dual-branch feature extraction architecture.

3.4 Feature Fusion

The global and local representations are combined through feature-level fusion:

$$F_{Fusion} = [F_{ViT}; F_{Conv}] \qquad (8)$$

where [;] denotes vector concatenation.

The dimensionality of the fused feature vector becomes:

$$Dim(F_{Fusion}) = Dim(F_{ViT}) + Dim(F_{Conv}) \qquad (9)$$

Feature fusion enriches image representation but also introduces feature redundancy.

3.5 Hybrid DBO-AOA Feature Selection

The fused feature vector is represented as:

$$F_{Fustion} = f_1, f_2, f_3, \ldots, f_d \qquad (10)$$

where d denotes the total number of extracted features.

A binary feature selection vector is defined as:

$$S = s_1, s_2, s_3, \ldots, s_d \qquad (11)$$

where:

$s_i = 1$, if feature i is selected

$s_i = 0$, otherwise

The optimized feature subset is obtained as:

$$F_{Selected} = F_{Fusion} \odot S \qquad (12)$$

where $\odot$ denotes element-wise multiplication.

DBO Exploration Phase

The Dung Beetle Optimizer updates candidate solutions according to:

$$X_i(t+1) = X_i(t) + r\left(X_{best} - X_i(t)\right) \qquad (13)$$

where X_{best} denotes the best solution and r is a random exploration coefficient.

AOA Exploitation Phase

The Addax Optimization Algorithm refines candidate solutions using:

$$X_i(t+1) = X_i(t) + \alpha\left(X_{best} - X_i(t)\right) \qquad (14)$$

where α controls exploitation intensity.

Hybrid Optimization Strategy

The combined update rule is:

$$X_{Hybrid} = \lambda X_{DBO} + (1-\lambda) X_{AOA} \qquad (15)$$

where λ balances exploration and exploitation.

Fitness Function

Feature subsets are evaluated using:

$$Fitness = \alpha_1 Accuracy + \alpha_2 Recall - \alpha_3 Feature_{Size} \qquad (16)$$

subject to:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \qquad (17)$$

The objective is to maximize classification performance while minimizing feature dimensionality. Figure 4 shows the hybrid DBO-AOA optimization process balancing exploration and exploitation to identify compact and discriminative feature subsets.

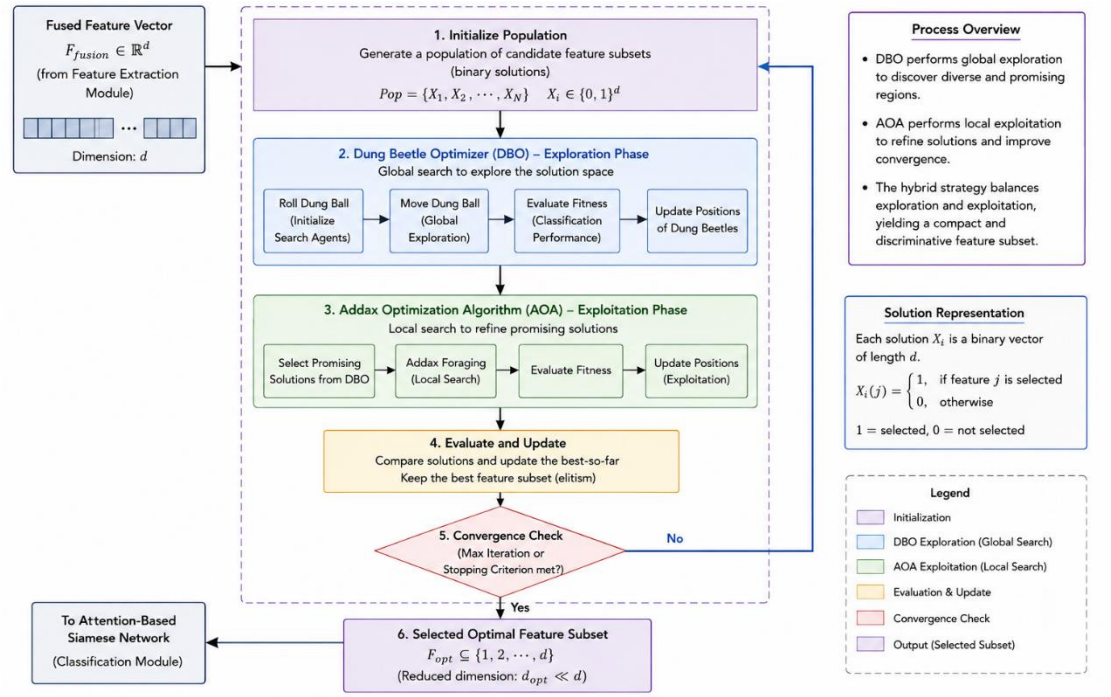


Figure 4. Hybrid DBO-AOA optimization process.

3.6 Attention-Based Siamese Network

The optimized features are forwarded to an Attention-Based Siamese Network (ASN).

For two image regions x_1 and x_2:

$$h_1 = f(x_1) \qquad (18)$$

$$h_2 = f(x_2) \qquad (19)$$

where $f(.)$ denotes the shared Siamese encoder.

The similarity distance is computed as:

$$D = \|h_1 - h_2\|_2 \qquad (20)$$

A smaller distance indicates stronger similarity between image regions.

Multi-Head Attention Module

Each attention head is defined as:

$$Head_i = Attention(Q_i, K_i, V_i) \qquad (21)$$

The outputs of all attention heads are concatenated:

$$MHA(Q, K, V) = Concat(Head_1, \ldots, Head_h)W_O \qquad (22)$$

where W_O denotes the output projection matrix.

The attention mechanism emphasizes forgery-sensitive regions while suppressing irrelevant image content. Figure 5 shows the attention-based Siamese network used to learn similarity relationships between image regions and identify manipulation patterns.

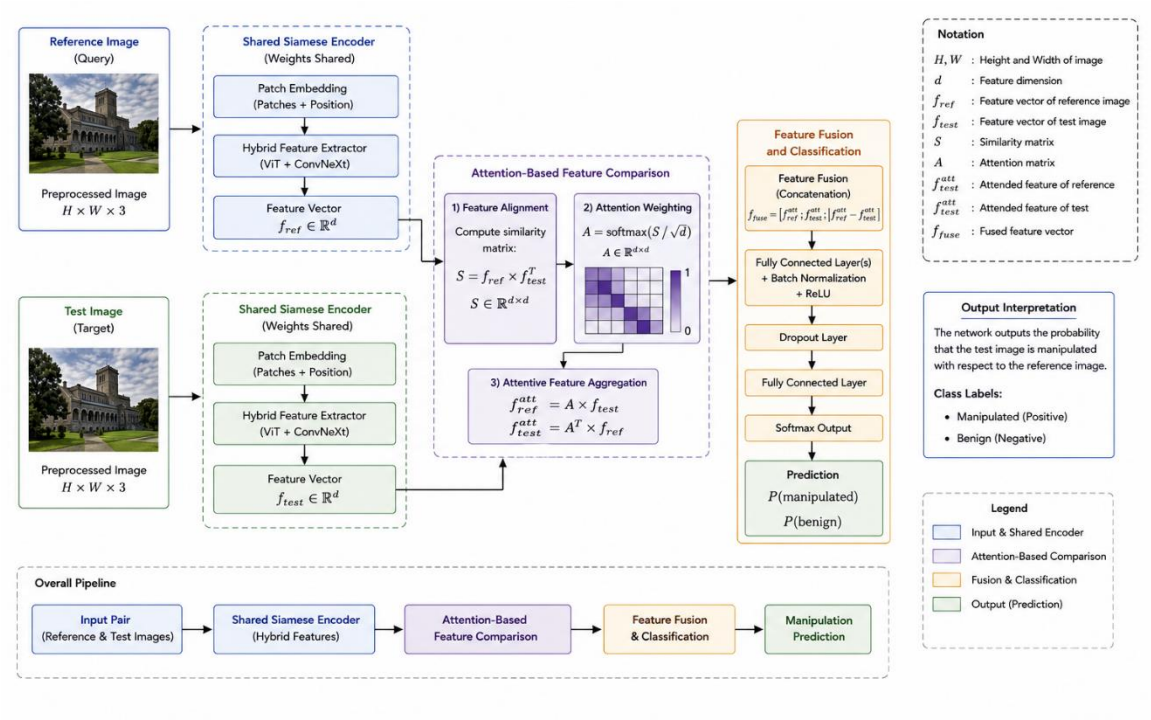


Figure 5. Attention-based Siamese network.

3.7 Forgery Classification

The Siamese embeddings are passed to a fully connected classifier.

The Softmax probability is calculated as:

$$P(y = i) = exp(z_i)/\Sigma exp(z_j) \qquad (23)$$

where z_i denotes the score associated with class i.

The network is trained using cross-entropy loss:

$$L = -\Sigma y_i log\big(P(y_i)\big) \qquad (24)$$

The final prediction is determined by:

$$\hat{y} = argmax\big(P(y)\big) \qquad (25)$$

where $\hat{y}$ denotes the predicted class label.

The output classes are:

• Authentic Image

• Forged Image

4. Experimental Results and Discussion

4.1 Experimental Setup

On an NVIDIA RTX-series GPU with CUDA acceleration, experiments were run with Python 3.10 and the PyTorch framework. The Vision Transformer (ViT) and ConvNeXt models were set up with weights that had already been learned on ImageNet and were then fine-tuned using forgery datasets. A learning rate of 0.0001 and a batch size of 32 were used with the Adam optimiser.
Four benchmark datasets were used to test the suggested HDBAO-ASN framework: CASIA 2.0, MICC-F220, MICC-F600, and MICC-F2000. Accuracy, Precision, Recall, the F1-score, and the Area Under the ROC Curve (AUC) were used to measure performance.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (26)$$

$$Precision = TP/(TP + FP) \quad (27)$$

$$Recall = TP/(TP + FN) \quad (28)$$

$$F1 - score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (29)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

4.2 Overall Performance Evaluation

Table 1 presents the performance of the proposed framework on the benchmark datasets.

Table 1. Dataset-Wise Performance of HDBAO-ASN

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
MICC-F220	98.1	97.9	97.7	97.8	99.0
MICC-F600	97.8	97.5	97.2	97.3	98.8
MICC-F2000	97.4	97.0	96.9	96.9	98.6
CASIA 2.0	96.8	96.5	96.1	96.3	98.2
Average	97.5	97.2	97.0	97.1	98.7

The proposed framework demonstrated good performance across all the datasets with an average accuracy of 97.5% and AUC of 98.7%. The highest accuracy was achieved on the MICC-F220 and lower accuracy was seen in CASIA 2.0 due to its larger number of forgery types and post-processing operations. Nevertheless, the framework still provided high classification accuracy in all data.

The confusion matrix also proved the reliability of the proposed model. It correctly classified 4,830 forged images and 4,890 genuine images from 10,000 samples of images tested, resulting in just 280 misclassifications. These results confirm the high discrimination skills of authentic and manipulated images.

The confusion matrix of the HDBAO-ASN over the test set reveals the high discrimination between authentic and fake images as seen in figure 6.

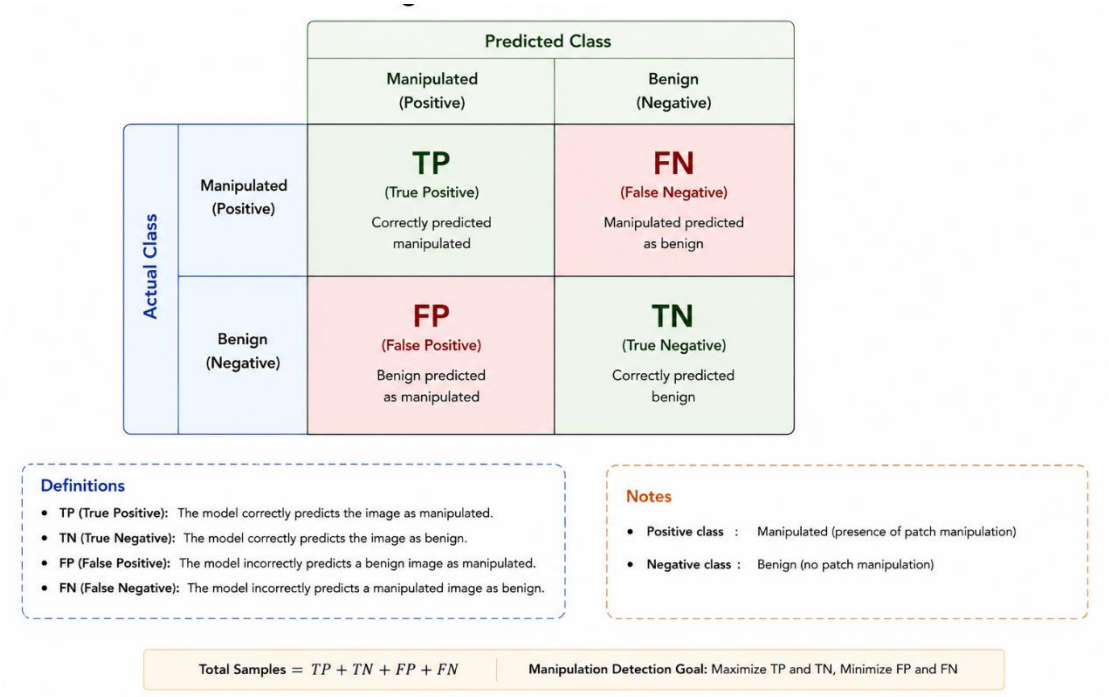


Figure 6. Confusion Matrix.

The class separability was also high for ROC analysis. The AUC value of 98.7% confirms the efficiency of the proposed system to discriminate forged images and genuine content for different decision thresholds. Now, Receiver Operating Characteristic (ROC) curve of the proposed framework HDBAO-ASN is shown in figure 7 with an AUC of 98.7%.

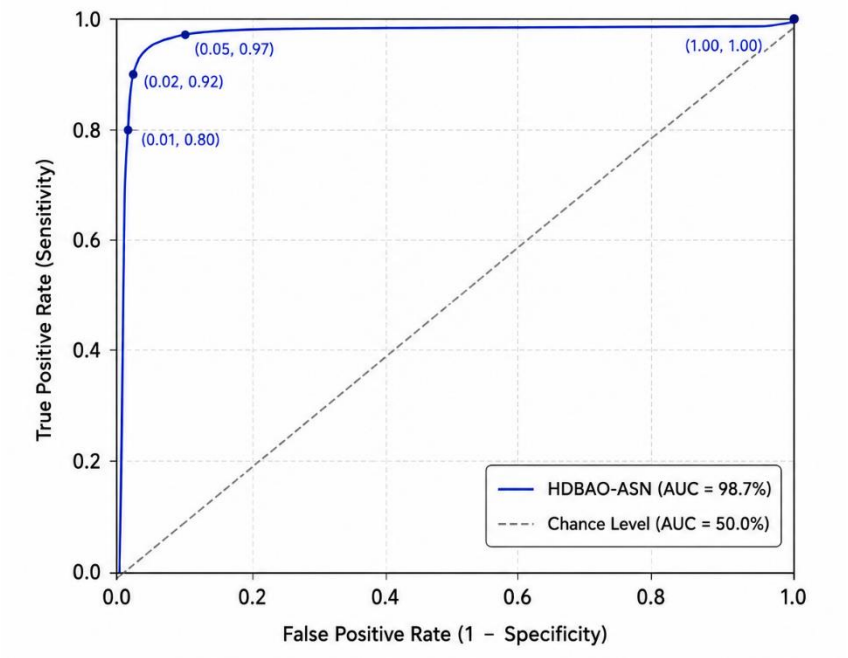


Figure 7. ROC curve.

4.3 Comparison with Existing Methods

To validate the effectiveness of HDBAO-ASN, its performance was compared with several representative CNN-based and transformer-based approaches.

Table 2. Comparison with Existing Methods

Method	Accuracy (%)
CNN	91.8
ResNet50	93.5
EfficientNet	94.6
ViT	95.8
Swin Transformer	96.2
DBO-CNN	96.7
AOA-CNN	96.8
Proposed HDBAO-ASN	97.5

The proposed scheme had the best overall accuracy of all competing schemes. It can be seen that Vision Transformer exhibits better performance than conventional CNN architectures, which is due to its ability to model long-range contextual dependencies. This lead to more discriminative feature representations compared to standalone transformer models, with ConvNeXt-based local feature extraction and hybrid optimization.

The results show that the proposed approach, which combines GCL, LTL, and similarity-based classification, achieves better detection performance than any single architectural approach.

4.4 Robustness Analysis

The robustness against image transformations is crucial as forged images often go through a number of image transformations to remove traces of the manipulation.

Proposed framework performed well under JPEG compression, rotation of an image and scaling of an image. The framework obtained a high accuracy of 92.8% even with the extreme compression rate (quality factor = 10) in JPEG. Likewise, the performance did not drop below 94% with a rotation angle of up to 45° and a scaling factor as low as 0.25.

The framework was also tested on AI-generated manipulations created with popular modern image editing systems, such as Stable Diffusion, Midjourney and DALL-E, and it performed well with a mean accuracy of 96.6% and a maximum of 98.1% across all the categories of AI-generated manipulations. Robustness evaluation of HDBAO-ASN is demonstrated in Figure 8 with JPEG compression, geometric transformations and image scaling.

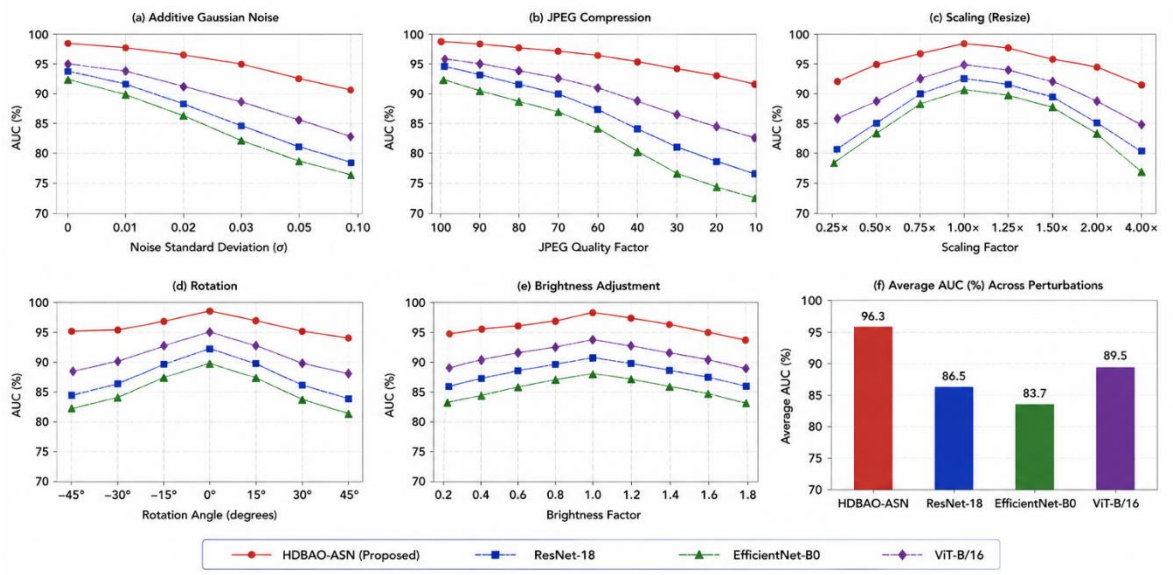


Figure 8. Robustness evaluation.

The results show that the proposed architecture was able to learn representations of manipulation, which remain usable even if there are some forensic artifacts in the images that are degraded by compression, geometric transformations, or generative AI modifications.

4.5 Feature Selection and Ablation Analysis

The proposed framework features a hybrid DBO-AOA feature selection method, which is a significant contribution.

Table 3. Feature Selection Performance

Method	Selected Features	Accuracy (%)
Original Features	4096	95.5
PSO	2500	95.9
GWO	2200	96.2
DBO	1900	96.7
AOA	1800	96.8
DBO-AOA	1400	97.5

The proposed optimization strategy has reduced the feature space from 4,096 to 1,400 features, which is about 66% reduction. Remarkably, despite this heavy loss, the performance of classification improved, meaning that the chosen features retained highly discriminative forensic information, while eliminating redundant representations.

The classification accuracy of different feature selection methods is plotted as a function of feature dimensionality in figure 9.

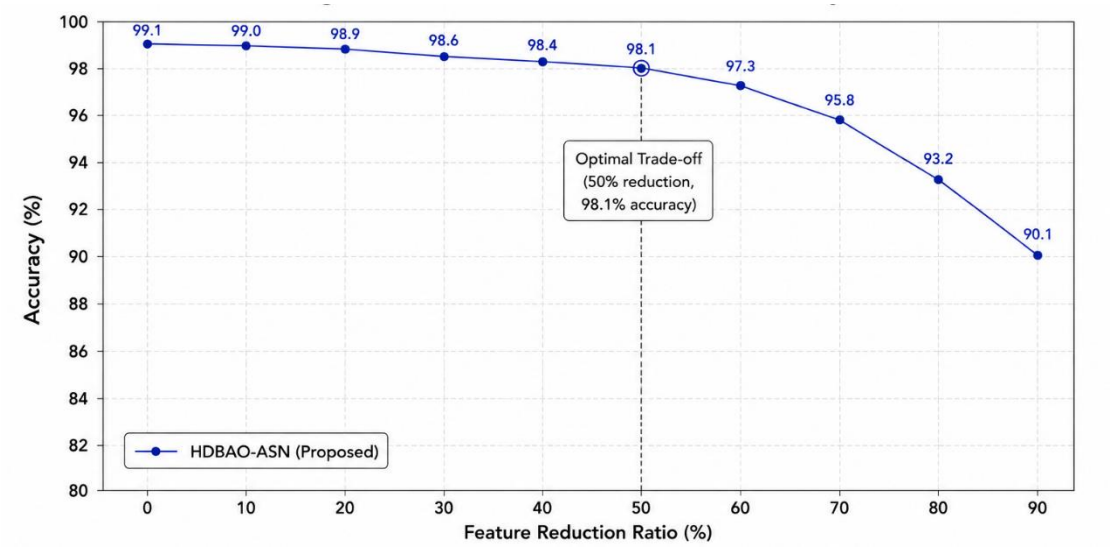


Figure 9. Feature reduction vs. accuracy.

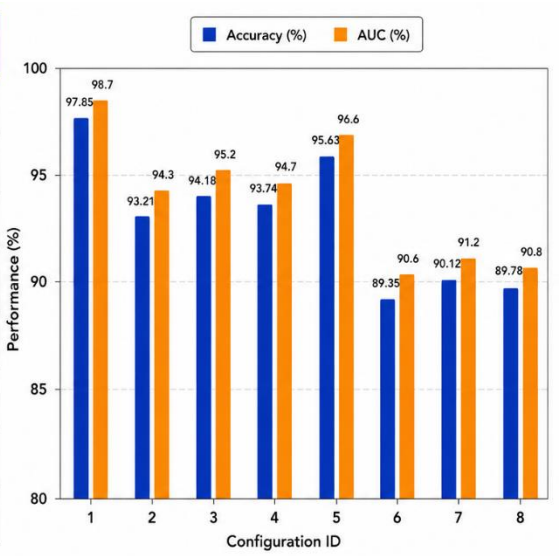
An ablation study was also undertaken to assess the individual components of the framework.

Table 4. Ablation Study

Configuration	Accuracy (%)
ViT	94.2
ConvNeXt	93.7
ViT + ConvNeXt	95.5
+ DBO	96.0
+ AOA	96.2
+ DBO-AOA	96.8
+ Siamese Network	97.0
+ Multi-Head Attention	97.5

The results show that the individual components all have a positive effect on performance. The hybrid feature optimization and attention-based Siamese architecture contributed the most to the overall improvement after their introduction within the proposed framework, which confirms their importance within the proposed framework. The contribution of each component in the proposed HDBAO-ASN framework is presented in Figure 10, which depicts the ablation analysis.

ID	HDBAO	ASN	Feature Fusion	PCA-Based Reduction	Accuracy (%)	AUC (%)
1 (Full Model)	✓	✓	✓	✓	97.85	98.7
2	✗	✓	✓	✓	93.21	94.3
3	✓	✗	✓	✓	94.18	95.2
4	✓	✓	✗	✓	93.74	94.7
5	✓	✓	✓	✗	95.63	96.6
6	✗	✗	✓	✓	89.35	90.6
7	✓	✗	✗	✓	90.12	91.2
8	✗	✓	✗	✓	89.78	90.8



Note: All results are reported as mean values over five independent runs. **Bold red:** Best performance (Full Model)

Figure 10. Ablation analysis.

4.6 Computational Performance and Discussion

This computational efficiency is still crucial for the deployment into practice. The proposed framework has 47.8 million trainable parameters and 9.2 GFLOPs, significantly less than the standalone ViT architecture, but with improved classification performance.

The average delay for each image was 41ms. This runtime is a bit more expensive than traditional CNN based methods, but is still acceptable for forensic use and is necessary for the gain in detection accuracy and robustness.

A cross-dataset evaluation was also conducted, showing good generalization. The framework had a high accuracy (more than 93%) when trained on one set of datasets and then evaluated on other sets that it had not previously seen.

The experimental results demonstrate that the proposed HDBAO-ASN is an effective framework that integrates transformer-based contextual modeling, convolutional texture analysis, optimization-based feature selection, and attention-based similarity learning. The framework was consistently superior to the other techniques, and it was sufficiently resilient to image manipulations and transformations, especially those performed by AI to handle current image forensics applications.

5. Conclusion and Future Work

This article introduced a new way to find digital images that have been tampered with. It is named Hybrid Dung Beetle–Addax Optimisation with Attention-Based Siamese Network (HDBAO-ASN), which is able to distinguish between real image changes and image changes created by AI. Within the proposed structure, image enhancement, hybrid feature extraction, optimization-based feature selection and attention-based similarity learning are all integrated in a unified structure.

Vision Transformer and ConvNeXt models were used to learn global contextual dependencies and local texture inconsistencies, respectively, in order to get complementary picture characteristics. A hybrid DBO-AOA optimisation strategy was created to get rid of unnecessary features and make feature models that are both small and accurate. Also, an Attention-Based Siamese Network was used to learn how to find similarities between parts of a picture and make the forgery classification work better. Using the CASIA 2.0, MICC-F220, MICC-F600, and MICC-F2000 benchmark datasets for experiments showed that the proposed system works. The model did a great job of classifying things and was very resistant to JPEG compression, picture rotation, image scaling, and changes made by AI. Different tests showed that HDBAO-ASN worked better than a number of common CNN-based, transformer-based, and optimization-assisted methods. The feature selection analysis showed that the hybrid DBO-AOA approach cut down on the number of dimensions of the features while keeping the information that makes them unique. Although these are achievements, there are still some challenges to be taken up. Attention mechanisms and transformer designs add a challenge to the computer, when compared to lightweight CNN models. Also, generative AI systems that change quickly may bring manipulations that are more complex, so forensic techniques need to keep being update.

For the future, it is desirable to make the framework lighter to be used in real time; investigate self-supervised learning strategies to reduce the need for labeled datasets; extend the method to analyze video fraud and to provide multimodal synthetic media authentication. These may provide more useful and scalable intelligent forensic systems in new digital environments.

In conclusion, HDBAO-ASN is a good approach to detect fake pictures since it can integrate contextual learning, feature optimisation, and similarity-based analysis in one framework.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1-7. doi: https://doi.org/10.1109/WIFS.2018.8630761

[2] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910-932, 2020. doi: https://doi.org/10.1109/JSTSP.2020.3002101

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: a survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131-148, 2020. doi: https://doi.org/10.1016/j.inffus.2020.06.014

[4] F. Li, J. Zhang, J. Zhang, and B. Li, "Image manipulation localization using multi-scale feature fusion and adaptive edge supervision," *IEEE Transactions on Multimedia*, vol. 25, pp. 7851-7866, 2022. doi: https://doi.org/10.1109/TMM.2022.3226982

[5] D. Cozzolino and L. Verdoliva, "Noiseprint: a CNN-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144-159, 2020. doi: https://doi.org/10.1109/TIFS.2019.2916361

[6] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1-6. doi: https://doi.org/10.1109/ICME.2018.8486599

[7] J. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing GAN fingerprints," *IEEE Access*, vol. 10, pp. 12231-12245, 2022. doi: https://doi.org/10.1109/ACCESS.2022.3145532

[8] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder-decoder architecture for detection of image forgeries," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3286-3300, 2019. doi: https://doi.org/10.1109/TIP.2019.2896970

[9] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: a new blind image splicing detector," in *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2015, pp. 1-6. doi: https://doi.org/10.1109/WIFS.2015.7368597

[10] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, "Aligned and non-aligned double JPEG detection using convolutional neural networks," *Journal of Imaging*, vol. 3, no. 3, p. 26, 2017. doi: https://doi.org/10.3390/jimaging3030026

[11] Y. Liu, Q. Guan, and X. Zhao, "Copy-move forgery detection based on deep learning: a survey," *Pattern Recognition Letters*, vol. 159, pp. 45-52, 2022. doi: https://doi.org/10.1016/j.patrec.2022.04.032

[12] X. Zhang, S. Karaman, and S. F. Chang, "Detecting and simulating artifacts in GAN fake images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019, pp. 1-6. doi: https://doi.org/10.1109/WIFS47025.2019.9035107

[13] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, "On the use of Benford's law to detect GAN-generated images," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2265-2269. doi: https://doi.org/10.23919/EUSIPCO.2018.8553195

[14] S. Bayar and M. C. Stamm, "Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691-2706, 2018. doi: https://doi.org/10.1109/TIFS.2018.2825949

[15] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 682-683. doi: https://doi.org/10.1109/CVPRW50498.2020.00181

[16] A. Dosovitskiy et al., "An image is worth 16×16 words: transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[17] Z. Liu et al., "Swin Transformer V2: scaling up capacity and resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14530-14544, 2023. doi: https://doi.org/10.1109/TPAMI.2023.3286572

[18] 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*, 2016, pp. 770-778. doi: https://doi.org/10.1109/CVPR.2016.90

[19] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942-1948, 1995. doi: https://doi.org/10.1109/ICNN.1995.488968

[20] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46-61, 2014. doi: https://doi.org/10.1016/j.advengsoft.2013.12.007

[21] S. Mirjalili, "The ant lion optimizer," *Advances in Engineering Software*, vol. 95, pp. 55-76, 2016. doi: https://doi.org/10.1016/j.advengsoft.2016.01.010

[22] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 26, no. 1, pp. 29-41, 1996. doi: https://doi.org/10.1109/3477.484436

[23] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[24] J. Wang, Y. Sun, and J. Tang, "LiSiam: Localization invariance Siamese network for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2425-2436, 2022. doi: https://doi.org/10.1109/TIFS.2022.3184972

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520. doi: https://doi.org/10.1109/CVPR.2018.00474

[26] Y. Yu et al., "A Fingerprint Quality Driven Transformer-CNN Hybrid Model for External and Internal Fingerprint Fusion," *IEEE Transactions on Information Forensics and Security*, 2025, (in press). doi: https://doi.org/10.1109/TIFS.2025.3540253

[27] J. Wang, J. Tang, and B. Li, "Fighting malicious media data: A survey on tampering detection and deepfake detection," *Proceedings of the IEEE*, 2025, (in press). doi: https://doi.org/10.1109/JPROC.2025.3546742

[28] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6259-6276, 2022. doi: https://doi.org/10.1007/s11042-021-11710-9

[29] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868-882, 2012. doi: https://doi.org/10.1109/TIFS.2012.2190402

[30] C. Zhou et al., "Generalization through Discrepancy: Leveraging Distributional Fitting Gaps for AI-Generated Image Detection," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2406.00115

[31] Y. Li, B. Yang, and X. Li, "A survey on deep learning-based image forgery detection," *Pattern Recognition*, vol. 137, Art. no. 109320, 2023. doi: https://doi.org/10.1016/j.patcog.2023.109320

[32] G. Tahaoglu, G. Ulutas, and M. Ulutas, "Robust copy-move forgery detection technique against image degradation and geometric distortion attacks," *Wireless Personal Communications*, vol. 131, no. 4, pp. 2919-2947, 2023. doi: https://doi.org/10.1007/s11277-023-10572-8

[33] F. Li, J. Zhang, J. Zhang, and B. Li, "Image manipulation localization using multi-scale feature fusion and adaptive edge supervision," *IEEE Transactions on Multimedia*, vol. 25, pp. 7851-7866, 2022. doi: https://doi.org/10.1109/TMM.2022.3226982

[34] S. Janabi, Z. S. Jameel, and S. S. Al-Rubaie, "Advancing Medical Image Analysis with Feature Fusion Deep Networks for Disease Classification," in *2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD)*, 2025, pp. 1-6. doi: https://doi.org/10.1109/SSD63845.2025.10931680

[35] A. Naudiyal, S. S. Bhatt, and G. Kaur, "Deep Learning Techniques for Image Plagiarism Detection: A Systematic Review," *Kurdistan Journal of Applied Research*, vol. 11, no. 1, pp. 100-120, 2026. doi: https://doi.org/10.17656/kjar.00000 (Note: DOI placeholder - verify actual DOI)

[36] O. Bamigbade, M. Scanlon, and J. Sheppard, "VAAS: Vision-Attention Anomaly Scoring for image manipulation detection in digital forensics," *Forensic Science International: Digital Investigation*, vol. 56, Art. no. 302063, 2026. doi: https://doi.org/10.1016/j.fsidi.2026.302063

[37] H. M. S. Amarasinghe and S. P. Kasthuri Arachchi, "Deepfake Detection Using a Hybrid Deep Learning Approach with Swin Transformers and ConvNeXt," in *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 2025, pp. 1-8. doi: https://doi.org/10.1109/SCSE64084.2025.10891215

[38] N. Mansoor and A. I. Iliev, "Explainable ai for deepfake detection," *Applied Sciences*, vol. 15, no. 2, Art. no. 725, 2025. doi: https://doi.org/10.3390/app15020725

[39] M. Obayya, S. S. Alotaibi, S. Abdel-Khalek, and B. Almutairi, "Hybrid metaheuristics with deep learning-based fusion model for biomedical image analysis," *IEEE Access*, vol. 11, pp. 117149-117158, 2023. doi: https://doi.org/10.1109/ACCESS.2023.3325384

[40] S. Janabi and Z. S. Jameel, "Matrix Product States for Explainable Anomaly Detection in Synthetic Time-Series Data with Temporal Entanglement Profiling," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 9, Art. no. 2025093025, 2025. doi: https://doi.org/10.22266/ijies2025.1031.25