Kathmandu University Journal of Science, Engineering and Technology, Vol. 18, No. 2, December 2024

Kathmandu University
Journal of Science, Engineering and Technology

Detecting Image Forgeries and Deepfakes: A Comparative Study of CNN and Transformer Models with a Custom-Curated Dataset

Pranish Kafle *[a], Sushan Adhikari †[a], Aayush Man Shakya[a], Nitin Ghimire[a], and Gajendra Sharma[a]

[a]Department of Computer Science and Engineering, Kathmandu University, Nepal.

Abstract

The exponential growth of sophisticated image manipulation technologies, including deepfakes and traditional forgery techniques, has created an urgent demand for robust detection mechanisms in digital forensics and computer vision. This research investigates the comparative effectiveness of two distinct deep learning architectures—InceptionV3 and Vision Transformer (ViT)—for identifying manipulated and synthetic imagery. Our experimental methodology utilized a comprehensive dataset containing over 140,000 images, encompassing conventional manipulation methods such as photoshopping, splicing, copy-move operations, and face-swapping, alongside AI-generated content from StyleGAN, StyleGAN2, and deepfake technologies. Through rigorous evaluation protocols, the InceptionV3 model demonstrated superior performance with 94.0% test accuracy and 93.96% validation accuracy, while the Vision Transformer achieved 88.49% test accuracy. Comprehensive performance analysis across multiple evaluation metrics—including precision, recall, F1-score, and computational efficiency—revealed that InceptionV3 outperformed ViT by 5.51% in accuracy and 10.8% in recall performance. These findings challenge prevailing assumptions about transformer architectures universally surpassing CNNs in computer vision applications. The results indicate that CNN-based architectures, particularly InceptionV3, provide substantial advantages for image forensics applications through enhanced computational efficiency and superior detection capabilities for both manipulated and synthetically generated content. This research contributes valuable insights into architectural selection for deepfake detection systems and establishes benchmarks for future developments in digital media authenticity verification technologies.

Keywords: Image manipulation detection, Deepfake detection, Vision Transformer, InceptionV3, Comparative analysis, Digital forensics, Computer vision

1. Introduction

With the recent rapid advancement of digital technology, images have become one of the primary sources of information. They often contain visual details, such as personal identity, sensitive documents, or evidence relevant to journalism, forensics, and legal proceedings. However, with the digital advancement, the manipulation of images, commonly known as image forgery, is increasingly prevalent and common. Image forgery is the deliberate manipulation of digital images to alter their content with the intention of misleading or falsifying information. Common techniques include splicing (combining parts from different images), copy-move forgery (duplicating regions within the same image), inpainting (removing or filling in objects), and retouching (enhancing or degrading visual elements) [1].

More recently, deepfakes have emerged as a powerful new form of manipulation. Using Generative Adversarial Networks such as StyleGAN, deepfakes generate highly realistic synthetic faces and scenes that traditional forensic techniques struggle to detect.

This study presents a systematic evaluation of pretrained convolutional neural networks (CNNs), such as Inception, for detecting both classical image forgeries and AI-generated deepfakes. To conduct this analysis, we construct a custom dataset by combining authentic images from the CelebA dataset with manipulated counterparts generated through conventional techniques (e.g., splicing, copy-move, inpainting) and synthetic images produced using StyleGAN.

This paper provides a comparative analysis of deep learning-based detection models focusing on their effectiveness against both traditional image manipulation and GAN-generated deepfake images.

2. Related Works

2.1. CNN-Based Image Manipulation Detection

The preliminary studies surrounding CNNs for image manipulation detection were largely focused on developing specialized architectures for forensic purposes. Bayar and Stamm [2] created convolutional layers with restrictions designed only for forensic purposes. They demonstrated that their architecture could exceed the performance of handcrafted features. This paper showed that CNNs could achieve up to 99.97% accuracy in detecting different editing and set the stage for subsequent CNN manipulation attempts. [3]

For deepfake detection, several CNN architectures have shown promising results. MesoNet [4] based approaches were trained to analyze mesoscopic features to determine if the face was AI-generated. XceptionNet has been recognized relative to FaceForensics++ benchmark [5], and InceptionV3 has proven effective across many forensic tasks from detecting manipulated fingerprints with 91.04%-98.07% accuracy [6] to being part of a larger scheme for deepfake endeavors. [7]

Hybrid CNN approaches have thrived as well. VI-NET combines VGG and InceptionV3 for copy-move forgery classification [8]. Many other studies employed ELA and CNN-based creations for practical improvements as well. [9] [10]

[*] All authors contributed equally to this work.

[†] Corresponding author. Email: sushan.adhikari2060@gmail.com

2.2. Vision Transformers in Computer Vision

Vision Transformers (ViTs), introduced by Dosovitskiy et al. [11], adapt the self-attention mechanism from natural language processing to visual data. By dividing images into fixed-size patches and encoding them as tokens, ViTs enable effective modeling of long-range dependencies across the image.

However, ViTs have drawbacks. Zhu et al. [12], for example, compared the performance of ViTs in comparison to CNNs on small datasets. They found that the performance of ViTs on small datasets was drastically different than the performance observed on large datasets. The results of the ViTs were significantly worse. On small datasets, they determined that no inductive biases were present, leading to representational disparities.

Developments in transformer architecture, such as the Data-efficient Image Transformer (DeiT) [13], have proven that a vision transformer could achieve the same performance as CNNs, even trained on less data, simply due to a revised training and distillation strategy. The Swin Transformer [14] created a hierarchical approach and shifted windowing which provides a more computationally efficient vision transformer architecture.

2.3. Vision Transformers for Image Manipulation Detection

The application of transformers for image manipulation detection is an emerging research area. The IML-ViT framework [15] was one of the first to approach the image manipulation localization task systematically with Vision Transformers, demonstrating its performance compared to CNN-based methods.

The appeal of transformers to forensic applications has had mixed results. Some analysis show that transformers surpass CNNs at forensic tasks (i.e. Transformers are better at capturing global inconsistencies of synthetic content), however, other findings reveal that with good tuning, CNNs still perform competitively, especially when trained on smaller datasets. [16]

2.4. Comparisons Between CNN and Transformer Architectures

Many studies have compared CNN and transformer architectures across various computer vision tasks. For example, Bai et al. [17] have analyzed the robustness properties across different perturbation scenarios, finding that transformers and CNNs respond differently across various perturbative efforts. More recently, it's been found that CNNs have an edge over transformers when images are trained with less data. [12]

Furthermore, Murphy et al. [18] compare CNNs versus transformers in a medical application synthesis and provide insights into the architectural trade-offs that could be relevant for forensic applications. Their findings suggest that CNNs often have the advantage in situations requiring more local feature analysis.

3. Methodology

To evaluate the accuracy and effectiveness of various image manipulation detection algorithms, we designed a comprehensive experimental setup comprising custom dataset creation and performance evaluation using standard metrics. The details of the datasets and evaluation criteria are described in the subsections below.

3.1. Datasets Creation

To train and evaluate, we created our custom dataset. As the base dataset of original images, we utilized the CelebFaces Attributes (CelebA) dataset [19], which contains 202,599 cropped and aligned facial images of 10,177 individuals, each annotated with 40 binary attribute labels and 5 facial landmarks. The images cover large pose variations, background clutter, and a diverse range of individuals, supported by a substantial quantity of data and rich annotations.

To simulate manipulated content, we applied a wide range of image manipulation techniques to the original CelebA images. These methods include face swap, face morphing, facial reenactment, image splicing, deepfake simulation, frequency manipulation, digital makeup, compression artifacts, Gaussian blur, median blur, noise addition, brightness and contrast adjustment, color distortion, and resolution manipulation.

Similarly, for the generation of deepfake images, a pre-trained StyleGAN2 model trained on the FFHQ dataset was used[20]. we generated a diverse set of high-resolution photorealistic face images by varying the input latent vectors using different random seed values. These GAN-generated images closely resemble real human faces and are difficult to differentiate with the naked eye.

Table 1 summarizes the composition and sources of the dataset classes used in our experiments.

Class/Source	Number of Images
Fake: AI-Generated	4,630
Fake: StyleGAN2	5,000
Fake: FaceSwap	150
Fake: Photoshopped	1,000
Fake: Photoshopped (CelebA)	20,000
Fake: Photoshopped (Spliced)	6,000
Fake: Spliced	1,995
Fake: StyleGAN	48,446
Real: Flickr	29,102
Real: CelebA-HQ	30,000
Real: Without Photoshop	1,000

Table 1: Composition of Dataset: Types and Quantities of Images

3.2. Model Architecture and Training Setup

We benchmarked two distinct deep learning approaches for manipulation detection: a convolutional model (InceptionV3), a transformer-based architecture (ViT). Each model was trained independently on the same dataset, using model-specific hyperparameters optimized via grid search. Below, we outline each approach:

3.2.1. InceptionV3

For the evaluation, we used InceptionV3, a widely used convolutional neural network that was introduced as an improvement over earlier Inception architectures (also known as GoogLeNet). InceptionV3 is known for its use of inception modules, which simultaneously apply multiple convolutional filters of varying sizes in parallel, facilitating the extraction of both local and global features from input images. This multi-scale processing enables the model to efficiently capture complex data patterns without a significant increase in computational resources [21].

We utilized InceptionV3 pretrained on ImageNet and adapted it using our custom dataset, as detailed in Table 1. The initial layers—up to 991,200 parameters—were frozen to retain low-level feature extraction, while the remaining layers and the classification head were updated during training.

The model was trained using the Adam optimizer with an initial learning rate of 0.001, scheduled via cosine annealing. We used a batch size of 32 and applied standard data augmentations to improve generalization. The dataset was partitioned into training

(60%), validation (20%), and test (20%) sets. To prevent overfitting, a patience of 10 epochs was used and the training was set up for a maximum of 50 epochs.

3.2.2. Vision Transformer Implementation

Our Vision Transformer (ViT) implementation follows the standard ViT-Base architecture, initialized with pre-trained *ImageNet* weights. The input images are split into 16×16 patches, which are linearly embedded and passed through 12 transformer blocks featuring multi-head self-attention mechanisms.

The architecture incorporates strategic freezing of the first 9 transformer blocks (75% of the total) to preserve pre-trained knowledge while allowing fine-tuning of the last 3 blocks and the classification head.

During training, we applied *differential learning rates*: a lower learning rate of 1×10^{-5} for the backbone and a higher learning rate of 1×10^{-4} for the classification head. We also employed focal loss to better handle difficult examples.

3.3. Architecture Selection Rationale

InceptionV3 was selected over other convolutional architectures such as EfficientNet and ResNet-50 due to its ability to capture multi-scale features through inception modules, which combine convolutional filters of varying sizes to model information at different spatial resolutions [21]. Previous work has reported strong results for InceptionV3 in manipulation detection, including 91.04% accuracy in less challenging conditions and 98.07% in more complex scenarios on the SOCOFing dataset [6]. In addition, its parameter count of 22.8 million offers a balanced compromise between computational efficiency and representational power, while ImageNet pre-training facilitates effective transfer learning.

ViT-Base was chosen as the canonical Vision Transformer representative due to its global self-attention mechanism, which enables the modeling of long-range dependencies and the detection of spatially distributed inconsistencies often present in deepfakes [11]. In contrast to hierarchical variants such as Swin Transformer, ViT-Base retains the original architecture proposed in the seminal work, thereby providing a clean baseline for comparative evaluation. With 86.3M parameters, it maintains computational tractability while offering a theoretically strong inductive bias for identifying global anomalies in visual data.

3.4. Training Methodology

Both architectures were trained using stratified dataset splits (60% training, 20% validation, 20% test) to ensure balanced class representation across all sets. WeightedRandomSampler was used during training to maintain equal class representation in each batch. Early Stopping with a patience of 10 epochs was employed to prevent overfitting, and training was conducted for up to 50 epochs.

3.5. Evaluation Metrics

To evaluate the performance of our model, we used common classification metrics such as the confusion matrix, precision, recall, F1 score, and ROC-AUC. The confusion matrix describes the totals of the correct and incorrect classifications and indicates the two types of model errors. Precision refers to how many of the identified cases of deepfakes are real deepfakes, and recall gauges how many real deepfakes the model can detect, which corresponds to missed detections. The F1 score is a single measure that combines precision and recall, and is critical in cases where classes are not fully balanced, such as our dataset. Finally ROC-AUC measures the model's capability to distinguish real and fake across different decision thresholds, which provides a strong measure for all round discriminative performance.

4. Results and Discussion

4.1. InceptionV3 Performance Analysis

After training of the deepfake detection model using InceptionV3 for up to 50 epochs, with early stopping, the model performed optimally at epoch 30, reaching a maximum accuracy score of 93.96%. The training had used the adaptive learning rate approach, attaining optimal convergence at epoch 20, also at maximum accuracy of 93.96%. The regimen of the training indicates successful learning, as the loss curves of the training as well as validation exhibited sustained descent from the initial values of around 0.72 down to the final convergence of 0.30–0.31. Additionally, the accuracy curves of the training as well as validation went in parallel, starting at 50% and reaching terminal values of 93.56% as well as 93.96% respectively, indicating viable generalization without features of overfitting.

A summary of the training and validation metrics is presented below.

Metric	Training Set	Validation Set
Accuracy	93.56%	93.96%
Loss	0.3049	0.3109

Table 2: InceptionV3 training and validation performance metrics

The evaluation metrics utilized in this paper—accuracy, precision, recall, and F1-score—are summarized in Table 3. The model achieved an overall accuracy of 94% on a balanced dataset consisting of both real and fake images. For real content detection, it achieved 92% precision and 96% recall, resulting in an F1-score of 0.94. For fake content detection, the model exhibited 95% precision and 92% recall, also resulting in an F1-score of 0.94. This reflected a model strong ability to detect manipulated images while also maintaining a relatively low false positive rate.

Class	Precision	Recall	F1-Score
Real	0.92	0.96	0.94
Fake	0.95	0.92	0.94
Accuracy	0.94 (on 24,176 samples)		

Table 3: InceptionV3 Performance metrics for deepfake detection per class

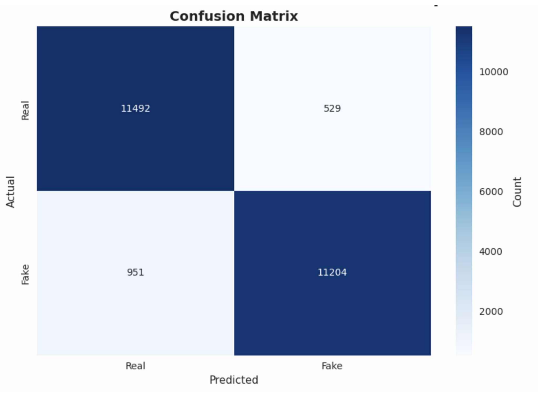

Figure 1: Confusion matrix for classification results

The confusion matrix shown in the figure 1 shows classification results across 24,176 test samples. The model correctly identified 11,492 authentic images and 11,204 manipulated image. The models also misclassified 529 real images as false (False Positive) and 951 fake images as real (False Negative).

The effectiveness of the model was then evaluated via the use of the Receiver Operating Characteristic (ROC) curve which is a plot of a classifier's true positive rate (sensitivity) against its false positive rate as the decision threshold varies [22]. ROC curve provides a comprehensive view of how well the classifier distinguishes between positive and negative cases across all possible thresholds [22]. The ROC curve stayed near the top-left corner, which indicated an excellent classification problem across all decision threshold. The model achieved an AUC of 0.986, demonstrating a high true positive rate while also maintaining a low false positive rate. The results, therefore, support the capability of the model to discriminate between original and manipulated content, even under adverse circumstances.

4.2. Vision Transformer Performance Analysis

The Vision Transformer (ViT)-based deepfake detection model exhibited stable and effective learning dynamics, reaching convergence after 30 training epochs. Both training and validation loss curves steadily decreased, with training loss ending at 0.0583 and validation loss at 0.0798. Training and validation accuracy followed similar upward trends, starting at approximately 50% and culminating in final values of 90.15% (training) and 88.36% (validation). This parallel progression without notable divergence indicates sound generalization, with no distinct signs of overfitting observed.

A summary of the training and validation metrics is presented below.

Metric	Training Set	Validation Set
Accuracy	90.15%	88.36%
Loss	0.0583	0.0798

Table 4: Vision Transformer training and validation performance metrics

The effectiveness of the ViT model was further evaluated on the test set. The model attained an overall accuracy of 88.49%. For fake content detection, the model achieved a precision of 95.00% and a recall (sensitivity) of 81.40%, which collectively yield an F1-score of 87.67%. The specificity, representing the ability to correctly identify real (true negative) samples, was 95.67%. ROC analysis produced an Area Under the Curve (AUC-ROC) of 0.884, signifying robust discriminative capability across a range of decision thresholds.

Metric	Value
Overall Accuracy	88.49%
Precision (Fake)	95.00%
Recall (Sensitivity)	81.40%
F1-Score (Fake)	87.67%
Specificity (Real)	95.67%
AUC-ROC	0.884

Table 5: Test set classification performance metrics for Vision Transformer

Class	Precision	Recall	F1-Score
Real	0.84	0.96	0.89
Fake	0.95	0.81	0.88
Accuracy	0.8849 (on 24,176 samples)		

Table 6: Vision Transformer Performance metrics for deepfake detection per class

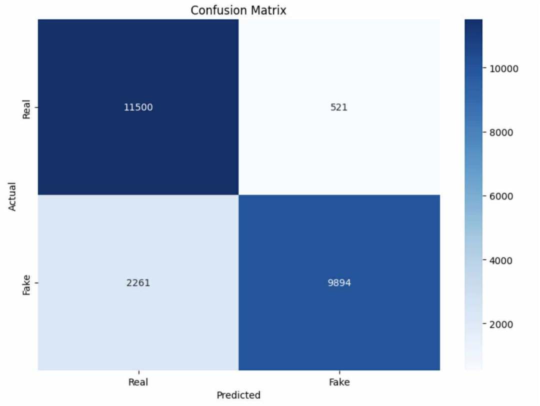

Figure 2: Confusion matrix for Vision Transformer classification results

The confusion matrix in Figure 2 details the classification results, further illustrating the model's performance across the test samples. Additionally, the ROC curve remained close to the top-left corner, confirming that the model maintains a favorable balance between true positive and false positive rates. These results collectively confirm the Vision Transformer's effectiveness for deepfake detection and underline its potential for real-world deployment.

4.3. Comparative Analysis

After conducting independent evaluations of InceptionV3 and Vision Transformer (ViT), a comparative analysis was performed to highlight differences in detection performance and computational efficiency, as summarized in Tables 7 and 8.

In terms of detection performance, Inceptionv3 performs better across the majority of evaluation metrics. It achieves an accuracy of 94.0%, compared to 88.49% for ViT. The recall and F1-score values for InceptionV3 are 92.2% and 93.8%, respectively, whereas ViT achieves 81.40% and 87.67%. Similarly, the area under the ROC curve (AUC-ROC) is also significantly greater for Inception. These values suggest that InceptionV3 attains a more consistent balance between sensitivity and overall classification effectiveness in this context.

Metric	InceptionV3	Vision Transformer	Difference
Accuracy (%)	**94.0**	88.49	+5.51
Precision (%)	**95.5**	95.00	+0.50
Recall (%)	**92.2**	81.40	+10.80
F1-Score (%)	**93.8**	87.67	+6.13
Specificity (%)	95.6	**95.67**	-0.07
AUC-ROC	**0.939**	0.884	+0.055

Table 7: Performance Comparison

The computational comparison reveals substantial differences among InceptionV3 and Vision Transformer. With a model size of 87.3 MB versus ViT's 331.5 MB, InceptionV3 requires nearly four times less storage space. This efficiency extends throughout the pipeline: training time is reduced by 38% (4.2 vs. 6.8 hours), inference speed improves by 31% (12.5 vs. 18.2 ms), and memory usage decreases by 27% (3.8 vs. 5.2 GB). InceptionV3 operates with 22.8 million parameters, while the Vision Transformer parameters—almost a billion parameter- are almost four times more than those of Inception. Despite having fewer parameters, InceptionV3 achieves better or comparable performance across most evaluation metrics, underscoring its remarkable efficiency.

Metric	InceptionV3	ViT	Difference
Model Size (MB)	**87.3**	331.5	+244.2
Training Time (h)	**4.2**	6.8	+2.6
Inference Time (ms)	**12.5**	18.2	+5.7
Memory Usage (GB)	**3.8**	5.2	+1.4
Parameters (M)	**22.8**	86.3	+63.5

Table 8: Computational Requirements Comparison

4.4. Contributions of This Study

This research makes several significant contributions to the field of digital image forensics and deepfake detection.

4.4.1. Comparative Architecture Analysis

We present the a comprehensive comparison between CNN-based InceptionV3 and transformer-based ViT architectures specifically for deepfake and image manipulation detection, providing empirical evidence that challenges common assumptions about transformer superiority in computer vision tasks.

4.4.2. Dataset Integration and Evaluation

Our study introduces a novel evaluation framework utilizing a diverse dataset of over 140,000 images that combines traditional manipulation techniques with modern AI-generated content, enabling robust cross-domain performance assessment. This comprehensive dataset design addresses limitations found in previous studies that focused on single manipulation types or limited synthetic content varieties.

4.4.3. Performance Benchmarking

The research establishes new performance benchmarks for deepfake detection, demonstrating that CNN architectures can achieve superior accuracy (94.0% vs 88.49%) while maintaining significantly better computational efficiency compared to transformer-based approaches. These findings provide practical guidance for selecting architectures in resource-constrained forensic environments.

4.5. Limitations of This Study

Despite the significant contributions, this research acknowledges several important limitations that may influence the interpretation and generalizability of our findings.

4.5.1. Dataset Constraints

While our dataset encompasses over 140,000 images, it may not fully represent the complete spectrum of emerging manipulation techniques or the latest deepfake generation methods. The rapid evolution of generative models like StyleGAN3, DALL-E, and other state-of-the-art systems may introduce novel artifacts that our training data does not capture, potentially affecting detection performance on future synthetic content.

4.5.2. Computational Resource Limitations

Our experimental setup was constrained by available computational resources, which may have limited the exploration of larger model variants or more extensive hyperparameter optimization. Vision Transformers, in particular, are known to benefit from larger datasets and extended training periods, which our resource constraints may not have fully accommodated.

4.5.3. Architectural Scope

This study focuses specifically on InceptionV3 and standard ViT architectures. We did not evaluate hybrid CNN-Transformer models or more recent transformer variants (such as Swin Transformer or ConvNeXt) that might bridge the performance gap observed in our findings. Additionally, the investigation did not explore ensemble methods that could potentially combine the strengths of both architectural approaches.

4.5.4. Generalizability Across Domains

Our evaluation primarily concentrated on face-centric imagery and common manipulation techniques. The findings may not generalize to other image categories, such as natural scenes, medical imagery, or specialized forensic applications, where different visual patterns and manipulation methods are prevalent.

4.5.5. Real-world Performance Validation

The controlled experimental environment may not fully reflect real-world deployment scenarios where images undergo compression, transmission artifacts, or post-processing operations that could affect detection accuracy. The impact of social media platform compression, image quality degradation, and various environmental factors on model performance requires further investigation.

4.5.6. Temporal Generalization

Our study represents a snapshot of current technology capabilities. The rapid advancement in both generation and detection techniques means that performance characteristics may evolve quickly, potentially affecting the long-term validity of our comparative findings.

These limitations highlight opportunities for future research directions and emphasize the need for continued evaluation as both manipulation techniques and detection methodologies continue to advance in this rapidly evolving field.

5. Conclusion

This paper presents an experimental study between two deep learning based manipulation detection architectures, InceptionV3(a CNN-based architecture) and the Vision Transformer, to differentiate between manipulated and real images by training and testing on a custom dataset. Both models underwent a refined transfer learning process and were evaluated on standards of accuracy, precision, recall, and F1 score.

The results of the study indicate that InceptionV3 surpassed the Vision Transformer in almost all tested measurements. In addition, InceptionV3 required less training time, memory, and storage and produced a smaller model size with fewer parameters than ViT. This challenges the common perception that transformer architectures are better than almost every other architecture for vision tasks. Therefore, our findings support the argument that CNN-based models still stand as a highly effective approach to digital image forensics for both detection and resource-intensive purposes.

However, since the technology behind image creation and image manipulation continues to be developed, it should be noted that more studies should be undertaken to assess generalizability since ensembles, hybrid approaches, or more rigorous data augmentation may allow for even more accurate detection. Additionally, applying these findings to less homogenous or more realistically derived datasets and studying the explainability of such models would further solidify the reliability of automatic detectors.

References

[1] Verdoliva L, Media forensics and deepfakes: An overview, *IEEE Journal of Selected Topics in Signal Processing*, 14(5) (2020) 910–932. https://doi.org/10.1109/JSTSP.2020.3002101.

[2] Bayar B & Stamm M C. A deep learning approach to universal image manipulation detection using a new convolutional layer. IHMMSec '16. Association for Computing Machinery, New York, NY, USA (2016). ISBN 9781450342902, p. 5–10. https://doi.org/10.1145/2909827.2930786.

[3] Bayar B & Stamm M C, Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection, *IEEE Transactions on Information Forensics and Security*, 13(11) (2018) 2691–2706. https://doi.org/10.1109/TIFS.2018.2825953.

[4] Afchar D, Nozick V, Yamagishi J & Echizen I. Mesonet: A compact facial video forgery detection network. In: *Proceedings of the IEEE International Workshop on Information Forensics and Security* (2018), pp. 1–7. URL https://arxiv.org/abs/1809.00888.

[5] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J & Nießner M. Faceforensics++: Learning to detect manipulated facial images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1–11. URL https://arxiv.org/abs/1901.08971.

[6] Ratnakar A, Advanced fingerprint alteration detection: A comparative analysis of real and synthetic modifications using inceptionv3 on the socofing dataset, *International Journal of Science and Research Archive*, 13(1) (2024) 564–574. https://doi.org/10.30574/ijsra.2024.13.1.1688.

[7] Nandini B, Kumar S & Gupta S K, Harnessing deep learning for reliable detection of deepfake images using inceptionv3 framework, *International Journal of Computer Applications*, 186(57) (2024) 24–29. https://doi.org/10.5120/ijca2024924298.

[8] Kumar S, Gupta S K, Kaur M & Gupta U, Vi-net: A hybrid deep convolutional neural network using vgg and inception v3 model for copy-move forgery classification, *Journal of Visual Communication and Image Representation*, 89 (2022) 103644. ISSN 1047-3203. https://doi.org/https://doi.org/10.1016/j.jvcir.2022.103644.

[9] Nagm A M, Moussa M M, Shoitan R, Ali A, Mashhour M, Salama A S & AbdulWakel H I, Detecting image manipulation with ela-cnn integration: A powerful framework for authenticity verification, *PeerJ Computer Science*, 10 (2024) e2205. https://doi.org/10.7717/peerj-cs.2205.

[10] Kubal P, Mane V & Pulgam N, Image manipulation detection using error level analysis and deep learning, *International Journal of Intelligent Systems and Applications in Engineering*, 11(4) (2023) 91–99. URL https://www.ijisae.org/index.php/IJISAE/article/view/3457.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J & Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)* (2021). URL https://arxiv.org/abs/2010.11929.

[12] Zhu H, Chen B & Yang C, Understanding why vit trains badly on small datasets: An intuitive perspective, *arXiv preprint arXiv:2302.03751*. URL https://arxiv.org/abs/2302.03751.

[13] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A & Jégou H. Training data-efficient image transformers & distillation through attention. In: *International Conference on Machine Learning* (2021), pp. 10347–10357. URL https://arxiv.org/abs/2012.12877.

[14] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S & Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10012–10022. URL https://arxiv.org/abs/2103.14030.

[15] Sun X, Wang Y, Qin Z, Wu J & Tian Q, Iml-vit: Benchmarking image manipulation localization by vision transformer, *arXiv preprint arXiv:2307.14863*. URL https://arxiv.org/abs/2307.14863.

[16] Maurício J, Domingues I & Bernardino J, Comparing vision transformers and convolutional neural networks for image classification: A literature review, *Applied Sciences*, 13(9). ISSN 2076-3417. https://doi.org/10.3390/app13095521.

[17] Bai Y, Mei J, Yuille A L & Xie C, Are transformers more robust than cnns?, *CoRR*, abs/2111.05464. URL https://arxiv.org/abs/2111.05464.

[18] Murphy Z R, Venkatesh K, Sulam J & Yi P H, Visual transformers and convolutional neural networks for disease classification on radiographs: A comparison of performance, sample efficiency, and hidden stratification, *Radiology: Artificial Intelligence*, 4(6) (2022) e220012. URL https://doi.org/10.1148/ryai.220012.

[19] Liu Z, Luo P, Wang X & Tang X. Deep learning face attributes in the wild. In: *Proceedings of International Conference on Computer Vision (ICCV)* (2015). URL https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

[20] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J & Aila T, Analyzing and improving the image quality of stylegan, *CoRR*, abs/1912.04958. URL http://arxiv.org/abs/1912.04958.

[21] Szegedy C, Vanhoucke V, Ioffe S, Shlens J & Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.

[22] Fawcett T, An introduction to roc analysis, *Pattern Recognition Letters*, 27(8) (2006) 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.