Advanced Image Forgery Classification with ResNet, Vision Transformers, and Edge-Aware Techniques

Kedhar Phanindra Sai Gurram*, Nikit K C

Abstract

Image forgery detection is an important task in digital forensics, as it can help identify and prevent the misuse of manipulated images. However, existing methods for image forgery detection are often limited by the type and quality of the forgeries, as well as the availability of large and diverse datasets. To address this issue, a novel edge enhancement based ResNet50-ViT architecture has been proposed. Edge enhancement on the input images has been performed using Canny edge detection algorithm. These images are then fed to ResNet50 for feature extraction and Vision Transformer has been used to classify them into copy-move, inpainted, spliced or authentic images. The proposed model has been trained on 12000 images which include images from publically available forgery datasets such as CASIAv2, NIST and COMOFOD. The experimental results show that the proposed model outperforms established image classification models on the multiclass forgery dataset.

1. Introduction

The ubiquity of digital media in today's society is evident in the diverse forms and platforms it takes. It has undergone a rapid and massive expansion in the past decade, reaching billions of people and transforming various aspects of society. This expansion has spurred the development of highly advanced and sophisticated media manipulation tools. Along with these tools the advent of generative AI models such as Midjourney and Dall-E has increased the possibilities of media manipulation ten folds. As a result of this, edited media is imperceptible to the human eye in the majority of cases. Even though audio and video manipulation have seen significant development, image manipulation has had the most amount of research and development over the years. This raises the question of credibility on any form of digital media.

Image forgery, also known as digital image tampering or image manipulation, refers to the deliberate alteration, modification, or creation of visual content with the intention to deceive or mislead viewers. Image forgery can be broadly classified into 3 classes, copy-move forgery, image splicing and image inpainting. Copy-move forgery means duplicating or cloning one or more regions within an image and pasting them onto other parts of the

*Corresponding author.
Email address: kedhar.gurram2002@gmail.com (Kedhar Phanindra Sai Gurram)

April 25, 2025

Figure 1: Original vs Forged images for various Image Forgery Types

same image. Image splicing involves combining objects from multiple images to create a composite image that appears as a single, unaltered photograph. Image inpainting refers to the process of filling in missing or damaged regions of an image with plausible content. Figure 1 shows the above described kinds of image forgeries alongside their untampered images.

The field of image forgery detection has advanced considerably in recent years, with numerous studies exploring and applying different strategies and algorithms to detect and counter image manipulation. The research can be broadly divided into two subdomains: classical feature extraction based and deep-learning based. Classical feature extraction methods involve identifying some common patterns in forged images, they often involve extracting handcrafted features from the images. Some of the commonly used techniques are Local Binary Patterns (LBP), Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) etc. Classical feature extraction based approaches suffer from high false alarm rate because [1] of their sensitivity to image variations & their limited ability of generalization. Deep learning based approaches on the other hand identify the complex features underlying in the images. Therefore, they are more robust and precise for forgery detection. Furthermore, the research has been done on various classes of image forgery using both classical and deep learning based methods.

Previous research in this field has only addressed specific types of image forgery, such as copy-move, splicing, or inpainting. However, there is a lack of comprehensive studies that

cover the wide range and diversity of image forgery techniques. The proposed research tries to solve this problem by proposing a hybrid ResNet50 - Vision Transformer model to identify various kinds of image forgeries that are performed on an image. The work focuses on the edge discontinuity that is common in forged images, it uses edge enhancement and feature extraction along with the state of the art classification technology: the Vision Transformer for better identification of the image forgery patterns and their classification. ResNet50 model has been used to perform feature extraction on edge enhanced images as it offers a good middle ground between performance and accuracy. The results have been thoroughly compared with various established architectures for image classification on the dataset.

2. Literature Survey

Several methods have been proposed to detect image forgery using different features and models. The work can be classified as classical feature extraction based and deep learning based. Regarding feature extraction based approaches. Zhang et al. [2] proposed a method for image splicing detection based on moment features from a multi-size block discrete cosine transform (MBDCT) and image quality metrics (IQMs). They achieved an accuracy of 89.16% on the Columbia Dataset . de Carvalho et. al [3] proposed a forgery detection method for image splicing that extracts texture- and edge-based features from illuminant estimates on image regions of similar material and achieved detection rates of 83% on images collected from the internet. Amerini et. al [4] proposed a SIFT-Based Forensic Method for copy–move forgery detection. Pun et. al. [5] proposed a novel copy-move forgery detection scheme using adaptive oversegmentation and feature point matching. The problem with classical feature extraction based methods is that they can only identify a specific type of tampering by identifying certain features in the image [6]. Deep learning models can be employed to alleviate these problems, as they can identify and learn the underlying patterns in a wide variety of forgeries.

Image inpainting has become a major research hot spot in computer vision applications [7]. The use of deep learning technologies such as Generative Adversarial networks (GAN) in image inpainting has greatly overcome the deficiency of traditional inpainting algorithms and produced more realistic and coherent results. Image inpainting detection has become a challenging task and various researchers have employed deep learning based approaches for it. Wu et. al. [8] proposed an end-to-end Image Inpainting Detection Network (IID-Net) that utilizes hierarchical enhancement, Neural Architecture Search (NAS)-based feature extraction, and global/local attention modules to detect and localize inpainted regions with pixel accuracy. Xiao et. al. [9] proposed a high-pass filter attention full convolutional network (HPACN) which integrates squeezed excitation blocks (SE) and concurrent spatial and channel attention (scSE) to enhance feature extraction and improve detection and localization of image inpainting operations. Hu et al. [10] proposed an edge-aware transformer framework for accurate detection of image inpainting forgeries, encompassing both deep learning-based and traditional methods. Their approach leverages a two-stream Transformer architecture to learn global body features and fake edge features separately, while a multi-modality cross attention module effectively combines the extracted information.

For copy-move forgery detection, Kumar et al. [11] proposed a hybrid deep convolutional neural network using VGG and inceptionV3 models. Abdalla et al. [12] proposed a novel approach for image forgery detection and localization using scale variant convolutional neural networks (SVCNNs). The study focuses on copy-move forgery detection and localization by incorporating sliding windows of various scales into customized CNNs to generate possibility maps indicating image tampering. Pomari et. al. [13] proposed deep learning based method for detecting photographic splicing by utilizing Illuminant Maps and Convolutional Neural Networks. Jaiswal et. al. [14]proposed a deep learning based splicing detection network using pretrained ResNet50 on CASIA v2 dataset. There has been significant research on Image Forensics and forgery detection. However, existing research in the field of image forgery detection predominantly focuses on specific types of image forgeries, limiting the scope of analysis to individual forgery techniques. However, there is a lack of comprehensive studies that incorporate multiple kinds of forgeries, thereby presenting a research gap in the field. This limitation highlights the need for a holistic approach that addresses the detection and analysis of various types of image forgeries in a unified framework. By addressing this research gap, our study aims to provide a comprehensive solution that encompasses multiple image forgery techniques, contributing to a more robust and versatile forgery detection methodology.

3. Methodology

The purpose of this section is to provide the motivation for this research and to introduce the proposed architecture of the system. The section is divided into two sub sections: Motivation and proposed architecture.

3.1. Proposed Architecture

The problem of forgery classification is not a trivial image classification task. Modern forgery techniques are highly sophisticated and challenging to detect. To accurately identify the type of forgery, the model needs to learn complex features and patterns in the images such edge irregularities, camera noise, lighting inconsistencies and other cues. Deep learning models are suitable for this task, as they can learn high-level representations from raw data. Among the various deep learning architectures for image classification, such as VGGNet, ResNet, InceptionV3, EfficientNet etc., Vision Transformer (ViT) is one of the most recent and state-of-the-art models in the field of computer vision. It can outperform very deep convolutional neural networks on complex visual recognition tasks, such as image classification, object detection, and semantic segmentation, while requiring significantly fewer computational resources. However, the basic ViT does not perform well on the forgery classification problem, as it lacks the ability to extract local features from images without the availability of an extremely large data set. Therefore, this paper proposes a novel model that combines a CNN ResNet50 for feature extraction with a ViT for the problem of Multiclass Image Forgery detection. The model also incorporates edge features extracted by Canny Edge Detector, which are useful for detecting edge inconsistencies that are common in forged images. The proposed model achieves accuracy of over 80% on the multi-class forgery classification problem.

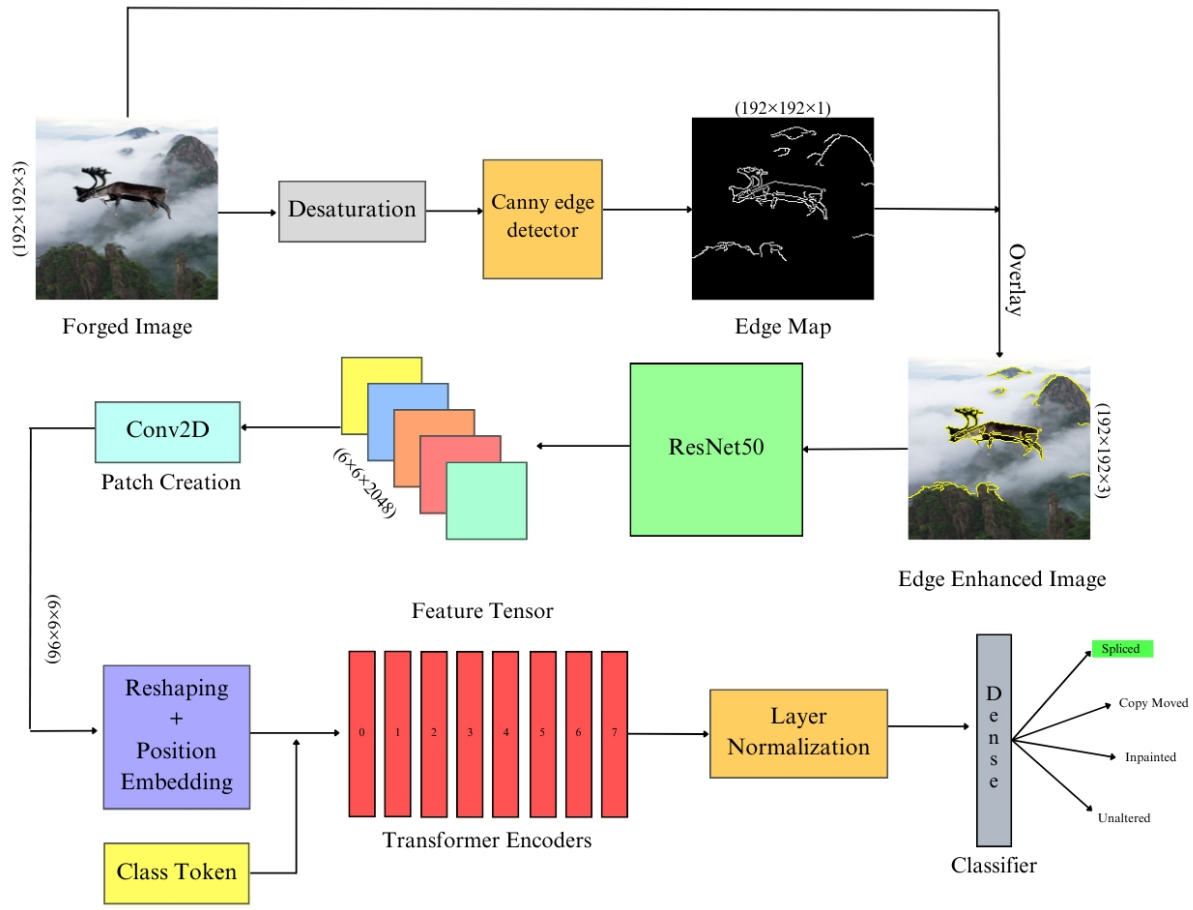

Figure 2: Proposed Edge Focused ResNet50-ViT Architecture

3.1.1. Preprocessing

The images have been resized to 192x192 pixels. The resizing has been performed because of the variations in the image sizes of the data-set. The "lanczos" interpolation is used for resizing the images as it is one of the best algorithms that preserves the fine details of the image during resizing which are critical for the proposed work. Next, edge enhancement is performed using the Canny edge detector [15], which is a multi-stage algorithm that effectively detects a diverse range of edges within an image. This detector has a low error rate and accurate localization of edge points on the center of the edge. It is also resilient to image noise and does not create any false edges because of it. The detected edges are then overlaid on top of the images in high contrast yellow color, which is not found in natural images, to enhance the edge features. The choice of threshold for edge detection is challenging, as the model must balance between over-focusing and under-focusing on edge features, especially the forged ones.

3.1.2. Feature Extraction

A CNN block is used to extract features from the edge-enhanced images and feed them to the Vision transformer, rather than using raw inputs. The chosen CNN model is ResNet50 which has 50 layers and offers good accuracy and minimal performance degradation the choice of this particular CNN is a result of Ablation survey performed on various CNN models, the results are shown in Table 3. The proposed architecture uses the pretrained version of the model with weights on the ImageNet dataset, which enables faster learning and convergence on a limited dataset by means of transfer learning. The last convolutional layer of ResNet produces feature maps, which are reshaped using a custom patch creation layer and relayed to the transformer for generating patch embeddings. Table 2 shows the kernel size and the number of filters used for the patch creation convolution layer. The convolutional blocks have been frozen from learning and are only used for feature extraction.

3.1.3. Vision Transformer

This section describes the backbone of the proposed architecture: Vision Transformer (ViT). The challenging problem of Forgery Classification requires the model to learn high-level representations from the data and achieve high accuracy during prediction. ViT is the perfect fit for such a task. ViT is a neural network architecture that applies the transformer model, originally developed for natural language processing to computer vision tasks. ViT consists of an encoder that transforms a sequence of image patches into a sequence of feature vectors and a classifier that predicts the output label based on these vectors. The encoder is composed of multiple layers of self-attention and feed-forward neural networks, which enable the model to capture the global and local dependencies among the patches. The classifier used in the model is a simple linear layer that takes the feature vector corresponding to the class token as input and outputs a probability distribution over the classes. ViT can localize the tampered regions in the image by using the attention maps of the encoder layers, which can provide interpretability and evidence for the classification decision. However there are a few challenges that a ViT model suffers from, to achieve higher accuracies, the vision transformers require training on very large datasets like the JFT-300M this is the result of the low locality inductive bias [16]. The lack of availability of such a large dataset is one of the big challenges that this work tries to solve. The proposed work improves the accuracy of ViT by over 19.2 % on the image forgery classification problem.

4. Experiments and Results

This section comprises the details of the datasets used, the experimental setup and the evaluation metrics that show the performance of the proposed architecture on the datasets. Finally the results have been compared with other work in this field. The model has been implemented in Tensorflow 2.10 using Keras API running on a computer with Windows 10 and NVIDIA Tesla A4000 GPU along with 128 GB of RAM.

4.1. Dataset Description

There is publically available dataset that contains all 3 types of forgeries used in the work. Therefore, a combination of multiple publicly available datasets have been used for

training and testing. CASIA v2 dataset [17] comprises 7481 authentic images and 5123 forged images, out of which 3294 are spliced (different sources) and 1828 copy move forged images (same source). A portion of the CASIA V2 has been used for the training. Along with this, COMOFOD dataset [18] has been used for copy and move forged images; it consists of 3750 forged images. NIST dataset [19] has 13470 spliced images out of which 2500 images have been taken for training and 400 for testing. For inpainting the dataset proposed in [20] has been used. Table 1 shows the number of images taken for training and testing. A total of 12000 images have been taken for training the model & 1960 images have been used for testing.

Dataset	Format	Original		Copy Moved		Spliced		Inpainted	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing
CASIA v2	TIFF/JPEG	3000	500	1000	250	500	100	-	-
NIST	PNG	-	-	-	-	2500	400	-	-
COMOFOD	PNG	-	-	2000	250	-	-	-	-
Inpainting Dataset	JPEG	-	-	-	-	-	-	3000	500

Table 1: Forgery Dataset Description

4.2. Hyperparameters

The performance of Deep Learning based models depends heavily on the hyperparameters especially in the case of vision transformers, the performance metrics can vary greatly according to the parameters. The hyperparameters used in the proposed model are listed in Table 2 . These parameters have been determined by extensive testing and experimentation.

Hyperparameter	Value
Input Shape	(192, 192, 3)
Encoder Layers	8
ViT Heads	12
Number of Filters for Patch Creation (Conv2D)	96
Multi Layer Perceptron Dimensions	384
Patch Size	32
Learning Rate	0.001
Dropout Rate	0.4
Batch Size	16
Validation Split	0.2
Training Epochs	30

Table 2: Hyperparameters of the Proposed Architecture

4.3. Evaluation Metrics

The performance of the proposed architecture has been evaluated on various metrics such as, accuracy, precision, recall and score. These metrics are described below.

$$\text{Accuracy} = \frac{\eta_{TP} + \eta_{TN}}{\eta_{TP} + \eta_{TN} + \eta_{FP} + \eta_{FN}} \tag{1}$$

$$\text{Precision} = \frac{\eta_{TP}}{\eta_{TP} + \eta_{FP}} \tag{2}$$

$$\text{Recall} = \frac{\eta_{TP}}{\eta_{TP} + \eta_{FN}} \tag{3}$$

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Here, η_{TP} is the refers to the number of true positive instances correctly classified by the model. For each class i.e Authentic, Spliced, Copy Moved, Inpainted there will be 1 true positive instance. η_{TN} refers to the number of true negative instances that are correctly classified by the model. For each class there will be 3 instances of true negative classes. Similarly, η_{FP} and η_{FN} refer to the number of instances classified incorrectly by the model as positive and negative for a class respectively.

4.4. Ablation Survey

This section contains the ablation survey done on various CNN architectures for feature extraction, the choice of the CNN was a difficult task because of the small size of the dataset very dense models can very easily over fit the data. Therefore we conducted a survey with the CNN block being the ablation parameter on various CNN architectures available. Table summarizes the findings.

CNN Model	Depth	Parameters (M)	Average time per epoch (s)	Test Accuracy (%)
VGG16	16	138.4	80	85.09
MobileNetV2	55	4.3	75	78.41
InceptionV3	183	23.9	888	85.10
ResNet50	107	25.6	128	91.96
ResNet101	103	25.6	120	70.51

Table 3: Ablation Survey of Various CNNs

From the table it is evident that ResNet50 offers a good balance between the network depth, training-time, and number of parameters. All the models were tested on best validation weights after 30 epochs. VGG16 offered the lowest training times because of the simplicity of the model, however there were large fluctuations in validation accuracy on the dataset. MobileNetV2 Performed surprisingly well given the small number of parameters in the model. InceptionV3 performed slightly worse than ResNet50 but the training

4.5. Performance on Multi-class Forgery Dataset

This section contains the performance analysis of the proposed model on the dataset described in Table 1. The graph of Training vs Validation accuracy been plotted in Figure 3. It is evident from the curves that the model achieves very fast convergence as well as accuracy. The Confusion Matrix of the model tested on 1997 images from the dataset is

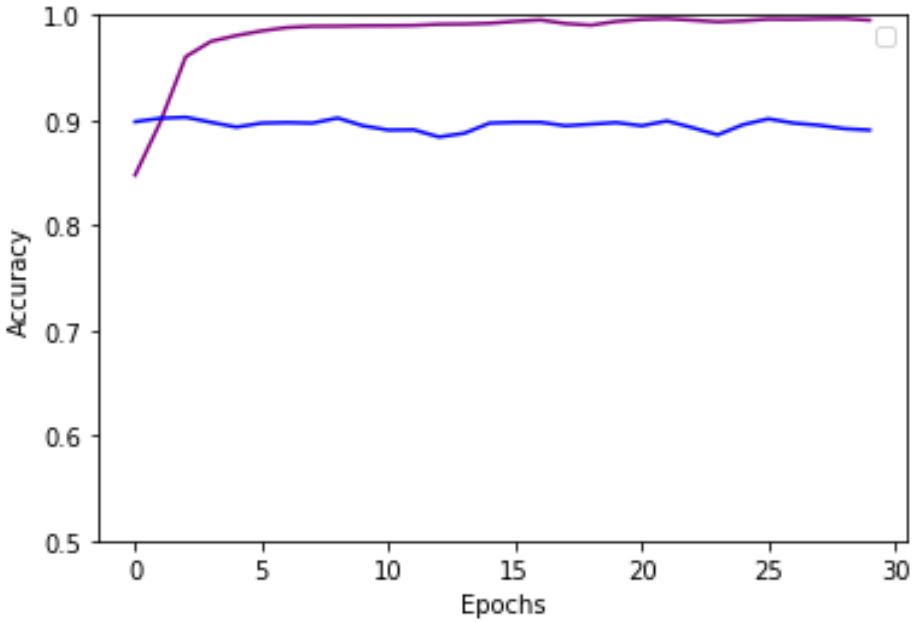

Figure 3: Training vs Validation Accuracy

plotted in Figure 4.

Table 4 shows the performance metrics of the proposed model on the Multiclass Forgery Dataset.

Accuracy	Precision	Recall	F1 score
91.96	92.03	92.02	91.99

Table 4: Performance on Multiclass Forgery Dataset

4.6. Comparison with state of the art classification models

Since there has been no previous work that accounts for all kinds of forgeries that the proposed model identifies, we compared the performance of the proposed architecture with some established deep learning models on the same dataset. Table 5 summarizes the findings. The standard implementation of the Vision Transformer model yielded comparatively lower accuracy in our experiments. One plausible explanation for this outcome could be the relatively smaller size of the dataset employed in our proposed work. The VGG16 model,

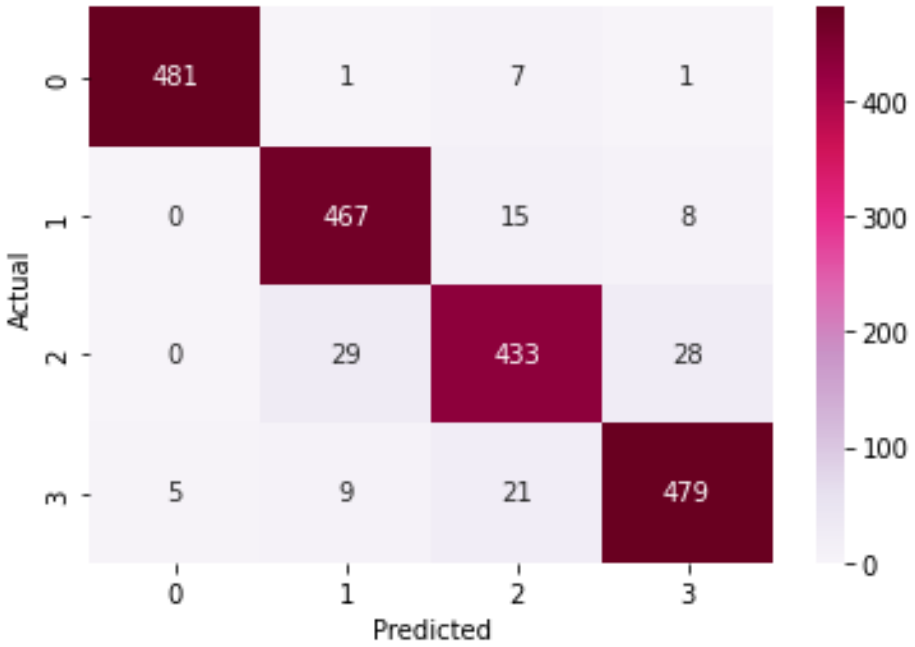

Figure 4: Confusion Matrix of The Proposed Model on classes: copy-moved(0), inpainted(1), authentic(2), spliced(3)

due to its simpler architecture, exhibited severe underfitting, resulting in an inability to effectively identify complex relationships and patterns within the forged images. Similarly MobileNetV2 suffered from high fluctuations during validation because of the simplicity of the model making it unfit for the proposed problem. On the other hand, the InceptionV3 model performed relatively well, achieving an accuracy of over 83%. However, we observed significant fluctuations in the validation curve during training, indicating a potential issue of overfitting. Similarly, the deeper ResNet101 model displayed similar tendencies, where the model tended to memorize the training data rather than generalize and learn underlying patterns. The ResNet50 model exhibited a similar trend, albeit with fewer fluctuations

Classification Model	Test Accuracy(%)
ViT	68.64
VGG16	25.50
MobileNetV2	84.28
ResNet50	83.14
ResNet101	84.68
InceptionV3	83.2
Proposed	**87.88**

Table 5: Performance of Established Deep Learning Models on Multiclass Forgery Problem

compared to its deeper counterpart. To address the challenge of overfitting, we proposed a model that mitigates complexity by freezing the layers of ResNet and utilizing them solely for feature extraction. This approach ensures that the model remains suitably complex while assisting the Vision Transformer in accurate classification.

5. Conclusion & Future Scope

A hybrid deep learning approach that combines CNNs and ViT for multi-class image forgery classification is proposed in this paper. Edge enhancement is performed on the images using the Canny edge detection algorithm and then they are passed to the pretrained ResNet50 model for feature extraction. The ViT architecture is modified to accept these feature vectors and perform classification based on them. An accuracy of 87.88% is achieved by the proposed approach on a dataset of 12,000 images that are either original or forged using techniques such as copy-move, splicing, or inpainting. An ablation study is conducted to justify the choice of CNN for feature extraction and the results are compared with established classification models. However, some limitations are faced by the proposed approach. One of them is the lack of a large and consistent dataset that incorporates different types of forgeries. Inconsistencies are created by a mixture of different datasets that can affect the learning of deep learning models. The low locality inductive bias of ViT can be addressed and its performance can be improved by a larger dataset. Moreover, more classes of forgeries and AI-generated images can be included in the future research.

References

[1] W. Ye, Q. Zeng, Y. Peng, Y. Zhang, X. Li, A two-stage detection method of copy-move forgery based on parallel feature fusion, J Wireless Com Network 30 (1) (2022) 30.

[2] Z. Zhang, J. Kang, Y. Ren, An effective algorithm of image splicing detection, in: 2008 International Conference on Computer Science and Software Engineering, Vol. 1, 2008, pp. 1035–1039. doi:10.1109/CSSE.2008.1621.

[3] T. J. de Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, A. de Rezende Rocha, Exposing digital image forgeries by illumination color classification, IEEE Transactions on Information Forensics and Security 8 (7) (2013) 1182–1194. doi:10.1109/TIFS.2013.2265677.

[4] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, G. Serra, A sift-based forensic method for copy–move attack detection and transformation recovery, IEEE Transactions on Information Forensics and Security 6 (3) (2011) 1099–1110. doi:10.1109/TIFS.2011.2129512.

[5] C.-M. Pun, X.-C. Yuan, X.-L. Bi, Image forgery detection using adaptive oversegmentation and feature point matching, IEEE Transactions on Information Forensics and Security 10 (8) (2015) 1705–1716. doi:10.1109/TIFS.2015.2423261.

[6] Z. J. Barad, M. M. Goswami, Image forgery detection using deep learning: A survey, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 571–576. doi:10.1109/ICACCS48705.2020.9074408.

[7] Image inpainting based on deep learning: A review, Displays 69 (2021) 102028. doi:10.1016/j.displa.2021.102028.

[8] H. Wu, J. Zhou, Iid-net: Image inpainting detection network via neural architecture search and attention, IEEE Transactions on Circuits and Systems for Video Technology 32 (2022) 1172–1185. doi:10.1109/TCSVT.2021.3075039.

[9] C. Xiao, F. Li, D. Zhang, P. Huang, X. Ding, V. Sheng, Image inpainting detection based on high-pass filter attention network, Computer Systems Science and Engineering 43 (2022) 1145–1154. doi:10.32604/csse.2022.027249.

[10] L. Hu, L. Yuanman, J. You, L. Rongqin, X. Li, An Edge-Aware Transformer Framework for Image Inpainting Detection, 2022, pp. 648–660. doi:10.1007/978-3-031-06788-4_53.

[11] S. Kumar, S. K. Gupta, M. Kaur, U. Gupta, 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. doi:10.1016/j.jvcir.2022.103644.

[12] Y. Abdalla, M. T. Iqbal, M. Shehata, Convolutional neural network for copy-move forgery detection, Symmetry 11 (10) (2019). doi:10.3390/sym11101280.

[13] T. Pomari, G. Ruppert, E. Rezende, A. Rocha, T. Carvalho, Image splicing detection through illumination inconsistencies and deep learning, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 3788–3792. doi:10.1109/ICIP.2018.8451227.

[14] A. K. Jaiswal, R. Srivastava, Image splicing detection using deep residual network, in: Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, 2019. doi:10.2139/ssrn.3351072.
URL https://ssrn.com/abstract=3351072

[15] J. Canny, A computational approach to edge detection, Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-8 (1986) 679 – 698. doi:10.1109/TPAMI.1986.4767851.

[16] S. Lee, S. Lee, B. C. Song, Improving vision transformers to learn small-size dataset from scratch, IEEE Access 10 (2022) 123212–123224. doi:10.1109/ACCESS.2022.3224044.

[17] J. Dong, W. Wang, T. Tan, Casia image tampering detection evaluation database, in: 2013 IEEE China Summit and International Conference on Signal and Information Processing, 2013, pp. 422–426. doi:10.1109/ChinaSIP.2013.6625374.

[18] D. Tralic, I. Zupancic, S. Grgic, M. Grgic, Comofod — new database for copy-move forgery detection, in: Proceedings ELMAR-2013, 2013, pp. 49–54.

[19] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, A. K. Roy-Chowdhury, Hybrid lstm and encoder–decoder architecture for detection of image forgeries, IEEE Transactions on Image Processing 28 (7) (2019) 3286–3300. doi:10.1109/TIP.2019.2895466.

[20] N. Kumar, T. Meenpal, Semantic segmentation-based image inpainting detection, in: M. N. Favorskaya, S. Mekhilef, R. K. Pandey, N. Singh (Eds.), Innovations in Electrical and Electronic Engineering, Springer Singapore, Singapore, 2021, pp. 665–677.