Grenze International Journal of Engineering and Technology, June Issue

Comprehensive Detection of Image Manipulations using Deep Learning

Divya Garikapati[1], Kethepalli Niharika Sri Sandhya[2], Mulpuri Sai Varun[3] and Karthik Mullapati[4]
[1-4]Lakireddy Bali Reddy College of Engineering Mylavaram, Andhra Pradesh, India
Email: garikapati.divya5@gmail.com, knss1710@gmail.com, mulpurisaivarun@gmail.com, karthikmullapati951@gmail.com

Abstract*—Digital forensics, security, and journalism have all faced difficulties as a result of the proliferation of deepfake technology and sophisticated image manipulation techniques like splicing, tampering, and faceswapping, which have seriously damaged the legitimacy and dependability of digital media. In order to overcome this, we present a novel multi-model deep learning framework that combines Error Level Analysis (ELA), Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Vision Transformers (ViT) to detect and classify image forgeries with remarkable robustness and precision. ViT outperforms conventional GAN- and CNN-based models without overfitting by using global attention mechanisms to assess image-wide associations and detect tiny anomalies in texturing, lighting, and structural details, resulting in over 95% accuracy in detecting deepfakes. CNNs detect pixellevel anomalies and structural inconsistencies, ELA highlights compression artifacts to identify tampered regions, and GANs improve training by producing a variety of synthetic forgeries and exposing the model to a range of manipulation techniques. In order to guarantee resilience under a variety of manipulation scenarios, the system was trained on an extensive dataset of actual and altered photos, utilizing preprocessing approaches such as data augmentation and class-weighted loss functions. This scalable solution is an essential tool for media verification, misinformation detection, and digital forensics, setting a new standard for protecting the integrity of digital content. Experimental results show state-of-the-art performance, with high precision and recall across all forgery types, even under difficult conditions like compression and post-processing.

***Index Terms*— Deepfake detection, Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Image tampering, Image splicing, Multi-Model Approach.**

I. INTRODUCTION

In the digital epoch, images have lost no little integrity to be able to be operated easily with super powerful tools and AI-driven manipulation techniques. Basic Sense of Image Forgery Definition: Changing or modifying an image with the intention of deceiving viewers to accept a distorted or even totally fabricated reality. Such manipulations of images have a wide spectrum of purposes, from entirely harmless artistic needs to the most pernicious aims: propaganda lies, forged documents, or defamation of reputation. Improved technology for image retouching accelerates humanity's incapacity to distinguish between a picture and reality and thus causes heavy damage to the credibility of the media of images.

Image forgery has ramifications that are quite multi-layered, especially in today's networked world. Images

Grenze ID: 01.GIJET.11.2.540

spread rapidly through social media and other online platforms.

A. Types of Image Forgery

Splicing
Splicing refers to the process of implanting several cut sections of an image from different sources so as to formulate one single, altered image. Manipulative technique: This is one of the very common media manipulation techniques in which scenes or events are composed from elements taken from several photographs or videos. Splicing is done to ensure the manipulated content looks natural; therefore, it is accepted as reality.

Tampering
Image tampering is the manipulation of certain areas of an image-adding, removing, or replacing objects in the scene. For example, an object can be digitally removed from the image or something new inserted in the image to alter its context. Tampering is usually carried out for a change of meaning or an aspect of the image. Therefore, it is considered a powerful tool for the dissemination of fake information and promotion of false narratives. Because such manipulated images may be interpolated in a way that does not easily reveal manipulation to the naked eye, detection is possible only by closer inspection.

Deepfake Generation
Deepfakes represent the most advanced and sophisticated form of image forgery. Such manipulations are developed using Generative Adversarial Networks (GANs), a class of deep learning algorithms capable of generating highly realistic images and videos based on the idea of learning from large datasets. Deepfakes are especially effective in facial manipulation wherein the faces of the people can be swapped or changed to create fake videos or images. Deepfakes is a new generation of technology with which close-to-understandable appearances and behaviors of the real world could be synthesized: from voice replications, facial expressions, to movements. This has raised great concerns over the areas of identity theft, privacy invasion, and the spread of misinformation.

II. LITERATURE SURVEY

The technology of deep fake is under rapid development mainly through the contribution of Generative Adversarial Networks and neural networks. This led to the skyrocketing of research in order to devise viable techniques of manipulated image and video detection. Various studies addressed this problem from very different perspectives, designing algorithms that detect the presence or absence of manipulation traces by the local and global characteristics of images.

A. AMTEN-net for Fake Face Detection

It presented an advanced technique in the form of the model called AMTEN-net, that contains a specific preprocessing layer, named AMTEN: designed precisely to emphasize manipulation traces in facial images, amplifying the area where potential manipulation artifacts are located and downplaying the original, unaltered features of the image to assist the CNN in concentrating on such areas. AMTEN-net achieved a very high detection accuracy of 98.52% on the manipulated face image databases with robustness against post-processing manipulation- like compression and blurring effects. However, a challenge for the model lies in adapting rapidly evolving deep fake techniques, such as those produced by new GAN models (for example, DeepFake, Face2Face, StyleGAN). Such advanced techniques create highly realistic faces, so the test set models need to generalise well to new kinds of synthetic images.

B. CNN Error Level Analysis for The Detection of Image Forgery

The other used a CNN in combination with Error Level Analysis to detect forgery in the images based on the different compressing levels. ELA works on the basis that it presents uneven compression levels across the image, which means displaying the tampered region. Such a method was pretty efficient using a CNN when it came to detecting ordinary types of forgeries; these include splicing or copying and moving. Although highly effective, the approach of CNN-ELA is highly sensitive to the quality and diversity of its training dataset. The architecture of CNN mostly fails to generalize across types of manipulations. It postulates a need for more extensive as well as diverse datasets to make it adaptive toward newly emerging manipulation techniques.

C. Hierarchical Fine-Grained Image Forgery Detection And Localization

Some have indeed suggested a hierarchical strategy for the detection and localization of subtle manipulations, where images are analyzed at multiple levels of detail. This approach allows CNNs with an attention mechanism

to look around areas within the image that suspect tampering and to the detection of subtle edits that otherwise easily pass through a simpler detection system. This hierarchical structure would allow a model to take on different resolutions while processing image details, which would be essential in getting subtle traces of tampering. This method, though highly precise in localizing forgeries, suffers from difficult challenges of scalability and efficiency; at least for real-time applications where the computational cost may even prohibit use of such techniques. Improvement in generalizing to a large number of forgery types is still another challenge.

TABLE I:

S.No.	Title	Methodology	Techniques/Algorithms	Results	Identified Gaps
1	Image Forgery Detection using Deep Learning	Data Preparation, Preprocessing	CNN, VGG- 19	High accuracy, low loss	Limited generalization for unseen data
2	Image Forgery Detection using CNN	Preprocessing, Feature Extraction	CNN, Error Level Analysis (ELA)	High accuracy for common forgeries	Heavy reliance on dataset quality and diversity
3	Enhancing Digital Image Forgery Detection	Data Collection, Transfer Learning	CNN, Mobile NetV 2, Transfer Learning	High accuracy (95%)	Real-time implementation challenges
4	Fake Face Detection using AMTEN-net	Amplification of Manipulation Traces	AMTEN, CNN	98.52% accuracy, robust to edits	Limited adaptability to newer deep fake techniques like StyleGAN
5	Image Forgery Detection: A Survey	Review of Deep Learning Techniques	CNNs, MFCNs, Mask R-CNN	High accuracy, improved detection	High computational needs, limited scalability
6	Hierarchical Fine-Grained Detection and Localization	Multi-level detail analysis	CNN, Attention Mechanisms	High precision in subtle forgeries	High computational cost, challenging for real-time application
7	Full-Resolution End-to-End Trainable CNN Framework	Full-resolution image analysis	Fullresolution CNN	Effective at splicing, copymove	High memory demands, impractical for real-time on large images
8	Comprint: Image Forgery Detection with Compression Fingerprints	Compression artifact analysis via Siamese network	Siamese NN, Compression Artifact Analysis	High accuracy with JPEG images	Limited to compressed formats, less effective on lossless images
9	Robust Image Forgery Detection over OSNs	Robust to OSN-specific operations	SNN, Residual Learning	High accuracy in OSN scenarios	Limited effectiveness beyond OSNs, dependency on noise modeling
10	DF-Net for Image Forgery Detection	Preprocessing, Multi-scale feature extraction	U-Net, scSE Blocks	High accuracy across datasets	Limited generalization to non-OSN images
11	Boundary-based Forgery Detection by Shallow CNN	Focus on low- resolution images, boundary identification	Shallow CNN	Effective for low-res, simple scenes	Limited handling of complex backgrounds and generalization limitations

III. METHODLOGY

A. Preparation and Preprocessing of Data

Data Collection

In this work, we collect data using datasets available on public platforms like GitHub, Kaggle, and CASIA. These would then be a rich source of real and manipulated images for training and evaluation of the techniques. We selected datasets such that as extensive of the deep fake generation methods and tampering techniques are

covered. Images were accessed from GitHub that had already been generated with the state-of-the-art GAN models and exhibited a range of deep fake techniques, both DeepFakeLab and StyleGAN. This meant very realistic deep fakes of public figures. Kaggle provided such a data set because it included a large number of labeled real and fake images with many kinds of manipulations on the images, from quite high-quality deep fakes to those involving splicing and face-swapping, among others. These images were valuable for this kind of task because their diversity in image quality, lighting, and background would greatly increase the chances of generalizability across many real-world cases. We also used datasets from CASIA-the Chinese Academy of Sciences Institute of Automation; specialized datasets included this, like CASIA1-Splicing and CASIA-Tampering. Images were manipulated in these datasets by different techniques, including splicing, cloning, and pixel-level tampering. CASIA datasets are well-annotated, which contributed to the controlled set of quality tampered images, thus enabling better performance evaluation on traditional image manipulation methods. The combined dataset contains about 50,000 images, evenly distributed between real and fake images spanning a wide range of deep fake generation techniques. Several preprocessing steps of the data were taken-standardized image resizing and normalization, augmentation, etc.-to improve generalization of the model. Although the dataset seems to be a rather solid basis, there were class imbalances with different kinds of manipulations, and variability in image quality of GANs, which have been mitigated with class-weighted loss functions and careful strategies for training models. This is a quite complete dataset with all its diversity and careful preprocessing being critical in training a model to really distinguish real from manipulated images.

TABLE II:

Dataset	Source	Type of Manipulation	Description	Total Images
CASIA v1 (Splicing)	CASIA	Splicing Manipulation	CASIA1 is one of the early datasets, which contains spliced images. Splicing is a very common tampering technique where parts from an image are inserted into another image to create a forged image.	921 spliced, authentic 800
CASIA v2 (Tampering)	CASIA	Various Tampering Methods	This version has spliced and other manipulations of images along with the copy-move forgery over wide range of scenes. The dataset is massively exercised for tampering detection.	5123 tampered, authentic 749
Deep fake and real images	Kaggle	GAN-Based Deep Fakes	This dataset contains manipulated images and real images. The manipulated images are the faces which are created by various means.	100,000 images

Preprocessing

To ensure compatibility with the CNN, VGG16, and Vision Transformer (ViT) architectures, all images were resized to 224x224 pixels and standardized for color channels and resolution. Data augmentation techniques, including random rotation, flipping, and color adjustments, were applied to enhance variability, improve model robustness, and reduce overfitting. Class-weighted loss functions were implemented to address imbalances across different manipulation types, ensuring fair representation during training.

B. CNN-based Forgery Detection Model

Architectures

Then, the CNN-based model for forgery detection will be designed with multiple layers that will capture and learn the distinguishing patterns and features signalling image manipulation. The model begins with four convolutional layers that progressively increase the number of filters (32, 64, 128, and 256). This will enable the model to learn low-level to high-level features. The ReLU activation is applied after each convolutional layer, introducing nonlinearity to the model and thus enabling it to learn more complex patterns in the data. Max-pooling layers are applied after all the convolutional layers with the purpose of reducing spatial dimensions and thereby compactifying feature maps, which help in computational efficiency as well as overfitting reduction.

Finally, here it passes through two fully connected layers where the final feature map is flattened. The model is allowed to classify the input image as "real" or "forged" by giving an output probability value using a softmax activation function. With this capability of learning both global and local patterns, it effectively detects several image manipulations, such as splicing, tampering, and other very minute alterations by the CNN.

Training Now, the proposed CNN model is trained using a cross-entropy loss function-a standard binary classification loss function-and an Adam optimizer with which it updates its model weights with the aim of optimizing convergence during training. The model learns for 50 epochs, while the learning rate is 10^4 and the batch size is of 32 images. Early stopping is performed based on the validation accuracy to prevent overfitting

that would ensure good generalization of the model toward unseen data. Thus, training the CNN model could successfully identify both real and forged images in an efficient manner since it learns high-level semantic features as well as low-level textures specific to forgeries.

C. CNN + VGG16 Model Overview

This CNN + VGG16 model overarches VGG16 architecture; it is a powerful, pre-trained model significantly used in different image classification tasks. VGG16 was trained on a large ImageNet dataset and thus enabled to extract common visual features toward other tasks. In forgery detection, VGG16 works as a feature extractor. The architecture is preserved up to the first layers of VGG16 since those first layers have been well learnt very rich general features on the ImageNet dataset. Further layers of VGG16 are kept fine-tuned for adaptation on this particular forgery task.

Design

The basic VGG16 architecture is kept; afterwards, specific layers are added for forgery classification. This is followed by the application of global average pooling to reduce spatial dimensions of the feature map created for compact representations of the learned features, and further through some dense layers with the application of ReLU activation for fine-tuning the representation features. This model ends in a softmax layer for binary classification, by which its output declares an image as real or forged.

Training Strategy

The first layers of VGG16 are frozen while training, which is to lock in the features learned from the ImageNet dataset. The feature extraction ability of strong VGG16 will be retained, and the model will adapt to the forgery detection task with the last few layers fine-tuned. For fine-tuning, the learning rate has been set to as low a value as 10^-5 so that the pre-trained weights are not terribly disturbed. Using binary cross-entropy loss with the Adam optimizer and having a batch size of 32, the model fine-tunes the VGG16 network, thereby generalizing good features from ImageNet to general images, and it also learns forgery detection-specific features in the process of training it.

D. CNN + VGG16 + GAN Model for Forgery Detection

Overview of the Prediction of Images using GAN

For the purpose of using in this model, GANs are used to generate some of the forgeries synthetically. Two networks constitute GAN. These include the generator and discriminator. Forgeries are generated by the generator and the discriminator has to differentiate between the real and fake images. The presence of GANs in the model ensures that the CNN + VGG16 model was exposed to a wider range of forgeries, assisting in augmenting the capability of the network to detect various manipulation strategies. Once again, the adversarial nature of GANs forces the discriminator to refine its ability to detect slight manipulation.

E. Architecture and Training

Generator Network

The generator within the GAN framework serves as a convolutional neural network whose architecture has been designed to synthesize fake images. This way, it learns to manipulate the latent space representation in an iterative way for producing images that look more real. The generator's goal here is to produce fake-looking forgeries that could pass as real which the discriminator fails to differentiate them as fakes.

Discriminator Network

The discriminator would be a combination of CNN and VGG16. It will be trained to classify an image as being real or forged. At first, the discriminator is trained separately for 30 epochs on the real images and the synthetic forgeries created by the generator; then, it begins adversarial training. The training algorithm of the generator and discriminator is an alternating one. The generator attempts to create more realistic forgeries, while the discriminator is the adversary it is against; it attempts to improve its discriminative capability to distinguish the real from the forged ones. Adversarial training would force the model to learn more robust features for distinguishing between real and forged images.

Adversarial Training Algorithm

The adversarial training algorithm simply switches its training between training the generator and the discriminator. The loss function of the generator is mean squared error, leading to generate realistic images, while it is binary cross-entropy for the discriminator. It is trained so that the generator develops an image that is

difficult for the discriminator to classify as fake. While the discriminator gets more and more ideas of how to distinguish the real pictures from the ones generated, by evaluation of subtle features that differ them.

Training Details

The batch size is 32 and the learning rate is set as 10^-4 like in the CNN and CNN + VGG16 models. Training details for this model are designed so it will face real images as well as GANgenerated forgeries so that the system may identify slight manipulations as well as on forgery in various real-world scenarios.

Based on the additional forgeries generated through GANs, the model is capable of performing better generalization and attains higher accuracy in detecting manipulated images. The generator presents new types of manipulation that the CNN + VGG16 discriminator might not have encountered during the training process, thus further strengthening the overall model of forgery detection.

F. Vision Transformers (VIT) for Forgery Detection

Overview

Vision Transformers (VIT) represent a transformative approach to image forgery detection, excelling particularly in deepfake classification. Unlike traditional convolutional networks that focus on localized features, VIT leverages attention-based mechanisms to examine global relationships within an image. This enables VIT to detect subtle and intricate anomalies that often go unnoticed by other models, making it an indispensable tool for sophisticated forgery detection.

Performance of VIT

Our implementation of VIT achieved an outstanding accuracy of over 95% in deepfake detection, outperforming both GAN-based and CNN-based models. This level of accuracy was consistent across diverse datasets containing a wide range of image qualities and manipulation techniques, highlighting VIT's robustness and generalization capabilities in real-world scenarios.

G. Formulae

1.GAN Loss Formula for Deep fake Detection:

$$L_{Gen} = -E\ [log\ (Disc(Gen(\xi)))]$$

Where:

- ξ represents random noise, which is input to the generator.
- Gen(ξ) denotes the generated image.•Disc(Gen(ξ)) is the discriminator's probability that the generated image is real.

2. Discriminator Loss Function:

$$LDisc = -E[\log(Disc(Ireal))] - E\ [\log\ (1 - Disc(Gen(\xi)))]$$

Where:

- Ireal is an actual image from the real dataset.
- Disc(Ireal) is the discriminator's classification output for the real image.

3. Activation (ReLU):

The ReLU activation function introduces non-linearity, enhancing the model's ability to learn complex patterns:

$$ReLU(x) = max(0, x)$$

4. Pooling Layer:

Pooling reduces the dimensionality of feature maps, retaining essential features while discarding redundant information:

PF(i,j)=max(FM(i,j),FM(i+1,j),FM(i,j+1),FM(i+1,j+1))

Pooling Layer = PF

Feature Map = FM

5. Fully Connected Layer (Output):

Output=W ⋅Flattened Features +b

Where:

- W represents the weight matrix.
- Flattened Features is the vectorized form of the final feature map.
- b is the bias term.

IV. RESULTS

The ViT model shows superior performance in distinguishing "Real" and "Fake" images, with precision, recall, and F1-scores of approximately 0.9926 for both classes. The model performs with an accuracy of 99.26% on a balanced dataset of 15,360 samples, showcasing robust generalization and reliability. This high performance of the model in distinguishing between actual and altered images is based on its capacity to utilize attention well in order to capture fine details in images, making it excellent for this differentiation task. Good performance on all metrics means it can sustain relatively low false-positive and false-negative rates and ultimately achieve exceptional accuracy and reliability of classification.

Apart from the above results, the model has excellent performance in tampering, splicing, and deepfake detection tasks, with high accuracy and low loss values in both training and validation. It obtained an accuracy of 89.62% with a loss of 0.2782 on the training set and 90.16% with a loss of 0.2821 on the validation set. This consistency between training and validation suggests that the model has learned to generalize and detect manipulated images without overfitting. These results are indicative of a strong feature extraction and classification ability, which will make the model adaptable for any tampering detection application.

However, the result for splicing and deepfake detection gives more nuanced results. In splicing detection, a precision of 0.8058 and accuracy of 74.93% were achieved; however, it had a rather lower recall at 0.5570. This means that although this model is excellent at reliably indicating spliced images, it might tend to miss even the slightest examples of spliced images. On the other hand, it performed very well on deepfake detection with an accuracy of 86.61%, precision of 0.8885, and recall of 0.8361. It indicates that even though there is room for improvement in splicing detection, particularly when subtle alterations are involved, the model is very well optimized to detect deepfakes, hence being a robust tool for the detection of any form of image manipulation.

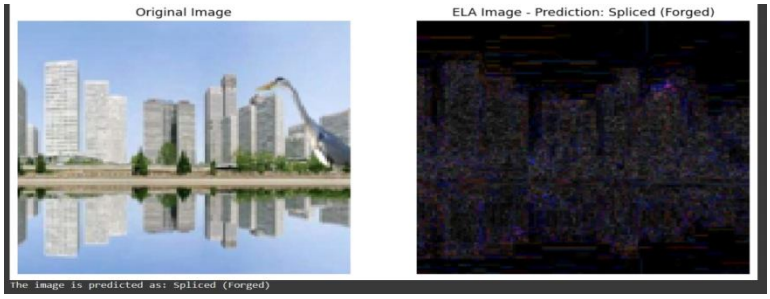

Fig 1: Splicing image prediction

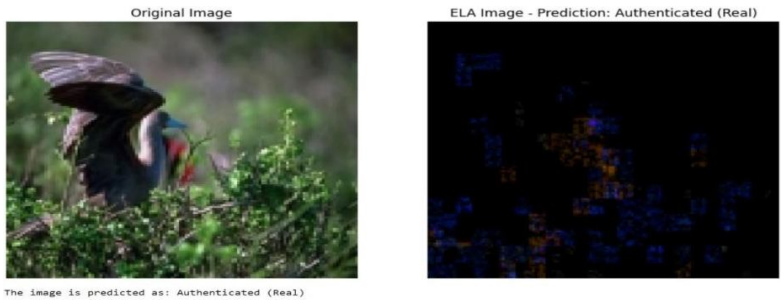

Fig 2: Real image prediction

Fig 3: Deepfake Detection

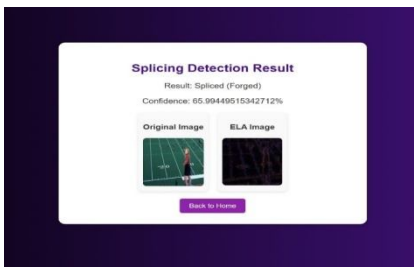

Fig 4: Splicing detecting

V. Conclusion

Improving techniques of image forgery pose a serious threat to the integrity of digital media. The three types of manipulations splishing, tampering, and deep-faking present unique challenges for their respective detection. A more sophisticated forgery technique necessitates a corresponding advanced AI based detection system. Thus, incorporation of deep learning techniques, especially CNNs and the hybrid models, shall provide a promising way forward in their fight against image forgery so that the digital content will continue to remain trustworthy and reliable in the midst of misinformation and manipulative digital means.

References

[1] https://ieeexplore.ieee.org/document/9951952
[2] https://ieeexplore.ieee.org/document/10205377
[3] https://ieeexplore.ieee.org/document/10226188
[4] https://www.researchgate.net/publication/371957250_IMAGE_FORGERY_DETECTION
[5] https://openaccess.thecvf.com/content/CVPR2023/html/Guo_Hierarchical_FineGrained_Image_Forgery_Detection_and_Localiz ation_CVPR_2023_paper.html
[6] https://www.researchgate.net/publication/342996815_A_Full-Image_Full-Resolution_Endto-End-Trainable_CNN_Framework_for_Image_Forgery_Detection
[7] https://ieeexplore.ieee.org/abstract/document/9074408?casa_token=2iQlFJ1TtFoAAAAA:Y-DcayLtbJnD-a204i_teEvAltJNTOG7JXMnliYQg1Lbu4Fu8lXdwIppoyNGKI_3PEj9L9Q9Y
[8] https://www.sciencedirect.com/science/article/pii/S107731422100014X
[9] https://ieeexplore.ieee.org/abstract/document/4806202