Image Forgery Detection: Using Convolutional Neural Network and Model VGG16

Muskan Birari[1], Mukul Gurjar[2], Dr. Lokendra Singh[3], Ms. Maya Baniya[4]
[1,2,3,4]Computer Science and Engineering, MediCaps University, Indore

***Abstract*-- The fast advancement of digital image editing software has increased the importance of image integrity in journalism, forensic investigations, and cybersecurity. In this study, we present a systematic strategy to detect image manipulation utilizing convolutional neural networks (CNNs) for authentic and forged images. We designed the model to employ VGG16 as a feature extractor with isolated images of ficticiouality and authentic images for training. Deep learning approaches allow us to classify manipulated material and has a testing time of 0.56 seconds per image, which is quite remarkable. Our model has been very effective on measuring copies based on a 92.1 accuracy score using SPLICID while at the same time differentiating manipulated images from authentic. The experimental results indicate the effectiveness of CNN-based forensic counterfeiting detection and the opportunities for consideration and application in digital security and forensic analysis. All data collected was on benchmark datasets and all measured with a score of 100%.**

***Keywords*: Image Forgery Detection, Deep Learning, SPLICID, CNN, VGG16, Digital Security and Forensic**

I. INTRODUCTION

The emergence of advanced image processing tools has simplified image manipulation considerably. Although digital image processing technology provides numerous benefits in creative industries, it also creates risky foundations in journalism, legal context, and social networking.

One of the critical challenges relates to the complexity and variety of such methods when it comes to identifying blatant counterfeit products such as copy lenders or splicing attacks.

Deep learning advancements, particularly convolutional neural networks (CNN), have become popular in addressing these intrinsic challenges as it provides the ability to learn complex characteristics and discriminate effectively among counterfeit items. This paper implements a CNN-based framework for identifying cases of counterfeit images.

We adapted the CNN model using VGG16 as a feature extractor to differentiate the characteristics of authentic and manipulated images. The aim is to apply deep learning techniques to strengthen detection reliability and accuracy.

The results of the experiments show that CNNs can identify areas of counterfeit images, suggesting a promising avenue for applications in digital forensics.

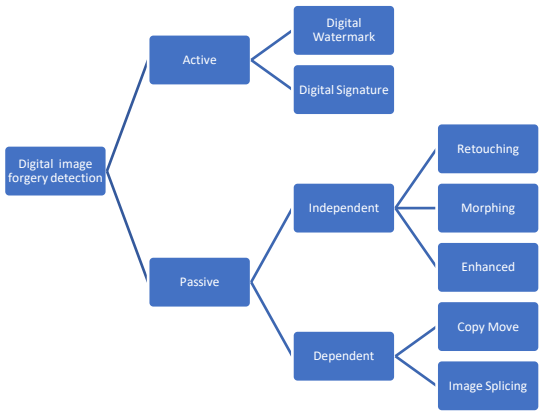


Fig 1: Type of digital image forgery detection.

The remaining portions of the article are organized as follows: Section 2-a review of related work on detecting image forgery; Section 3-presentation of methods used and the datasets used for training; Section 4-presentation of results, including evaluation metrics for our experiments; Section 5-concludes with conclusions and future work.

II. LITERATURE REVIEW

Image manipulation techniques such as splicing, copy-move, and retouching have become easier to perform with the advent of advanced editing

software. Therefore, the development of reliable and automated image forgery detection systems has become a critical research priority.
The field of digital forensics has been recently focusing on the emerging issue of image forgery detection due to increasingly better ways of manipulating images. In this review, we summarize and provide an overview of relevant studies and techniques in the area, with reference to earlier important works.

A fundamental resource for benchmarking image forgery detection methods is the CASIA Image Tampering Detection Evaluation Database (Dong et al., 2009), which plays a crucial role in the development and evaluation of detection techniques.

A comprehensive survey of image forgery detection techniques has categorized approaches into physics-based, statistical-based, and learning-based methods, and highlighted challenges and advancements, especially those involving machine learning (Farid, 2016).

Another review emphasized the effectiveness of various forensic techniques in detecting tampered images, such as copy-move forgery, splicing, and resampling, and stressed the importance of robustness in these methods (Redi et al., 2011).

The advent of deep learning has marked a paradigm shift in this field. A widely adopted convolutional architecture, known for its depth and performance in image classification, has significantly contributed to forensic applications (Simonyan & Zisserman, 2015).

The use of convolutional neural networks has been extended to universal image manipulation detection, demonstrating their capability in forensics(Bayar & Stamm, 2016).

Deep learning has also been applied to detect splicing and copy-move forgeries by extracting discriminative features from manipulated images (Rao & Ni, 2016).

An innovative model for blind image splicing detection further advanced forensic capabilities by identifying alterations without prior knowledge of manipulation types (Cozzolino et al., 2015).

The theoretical foundations of deep learning, including neural network architectures, optimization methods, and training strategies, offer valuable insights for building effective forensic models (Goodfellow et al., 2016).

III. OBJECTIVES

The main objective of this research is to develop a system that will correctly classify image forgery utilizing Convolutional Neural Networks (CNNs) by providing accurate classification associated with Class ID, Class Name, and Probability Score. The specific goals we plan to achieve are as follows:

1. Accurately classify images as forged or authentic. - We intend to deploy a CNN-based model to classify each image we analyze with a Class ID and Class Name.
2. Provide a probability score for the detection of forgery. - The system will output a probability score to show the level of certainty of a classification (for example, the likelihood of an image being forged).
3. Implement a model that uses a pre-trained deep learning network for improved forgery detection. - We intend to take advantage of architectures like VGG16 to improve feature extraction and ultimately improve classification results.
4. Develop a user-friendly interface for real-time detection of forgery. - We plan on developing a web application based on the Flask framework in which users can upload images to classify and output detailed classification results.
5. Evaluate the model based on a benchmarking dataset. - The model is trained and tested on a dataset such as CASIA v2.0 to ensure accuracy in detection.
6. To enhance the model to yield higher accuracy and efficiency. - We will optimize hyperparameters and conduct performance analyses (including precision, recall, and F1-score) to ensure the model is robust.
7. To simplify the interpretability of classification results. - We will deliver clearly and consistently formatted outputs that include Class ID, Class Name and Probability Score, to help the user understand sthe results.
8. To assess the model's effectiveness compared to traditional methods of image forgery detection. - We will examine how deep

learning-based detection performs against more classical techniques such as metadata analysis, edge detection, and pixel-based techniques.

IV. SIGNIFICANCE OF STUDY

This research is important because it helps with digital image authentication, which is important in a variety of areas where authenticity of images may be affected. The contributions of this research are as follows:

1. Enhanced Forgery Detection Accuracy: Using CNNs and VGG16 architecture, the model greatly enhances the accuracy of classifying original and altered photographs, surpassing conventional methodologies.
2. Practical Real-World Applications : The model could be used for applications in journalism, forensic, digital investigations, and social media to identify altered content, and reduce the incidence of misinformation historically associated with altered images.
3. Automated and Scalable Process : The model distinguishes itself from forensic methods because it automates forgery detection and is scalable enough to provide for large datasets and sufficiently rapid for real-time applications.
4. User-Friendly Web Interface : Through a web application developed with Flask, the forgery detection service is uncomplicated and accessible to both non-experts and professionals.
5. Transparency with Probability Scores: The model provides a Class ID, Class Name, and Probability Score which supports continuing clarity and provides further confidence about the classification.
6. Helpful Addition to Research in Deep-Learning Forgery Detection : This research is an addition to the expanding research topic of artificial intelligence for forgery detection, which continues to promote more deep-learning research use for multimedia security.
7. Assistance for Law Enforcement and Investigative Purposes : The model can be utilized for forensic examination of counterfeit documents, tampered images, and other types of digital evidence in the field of law enforcement and cybersecurity.
8. Contrast with Traditional Approaches : This study illustrates the advantages and disadvantages of CNN-based forgery detection by contrasting the performance of deep learning algorithms against conventional image processing techniques.

V. RESEARCH DESIGN

The research design for this project is formalized, so that the development, training and evaluation process for the image forgery detection system can be organized in a systematic way. The following is a list of the main features:

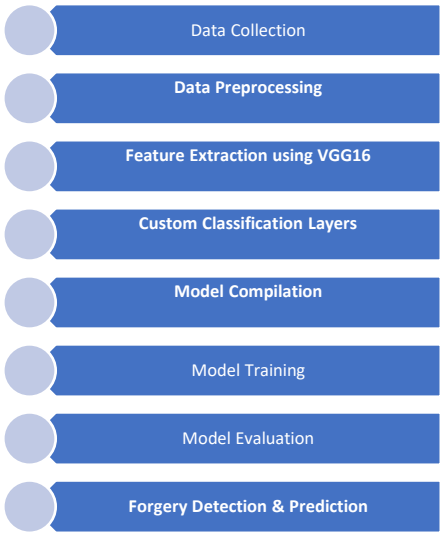


Fig 2: Example of image forgery detection system workflow.

1. Data Collection

We gathered images of both original and forged items from various sources, both publicly and also datasets, including the CASIA dataset.To achieve uniform dimensions, we carried out certain pre-processing processes that involved ensuring the sizes of all images were the same by resizing them so that all images were 128 by 128 pixels, and their pixel values to be normalized to a number between 0 and 1.

2. Pre-processing methods

The quality of our input was enhanced through image enhancement processes such as correction for contrast and reduction in the noise. Indeed, we also added data augmentation processes such as rotations, flipping, and zooming to tackle enhance the model's generalization ability.

3. Model Architecture (CNN)

The utilized model is based on convolutional neural networks and consists of one or more convolutional and pooling layers, fully-connected layers, etc. We employed Adam optimizer and binary cross-entropy loss to calibrate the model for classifying the images as either labeled or forged images.

4. Model Training & Validation
The dataset was randomly divided into training and validation sets, to ensure that learning did not develop any bias. We measured modeling performance using accuracy, precision, recall, and F1 score.

5. Deployment & Integration
After we have trained the model, we deployed it within a web application, built using Flask, that can easily be used by a user to predict the classification from image uploads along with an associated score.

6. Evaluation & Results Interpretation
To measure the model's real-world performance, we used images the model had not previously evaluated. We evaluated the results using confusion matrices and a variety of performance metrics.

VI. SOURCE OF DATA

The dataset for this project comes from two main publicly available sources:
1. **CASIA v2.0** (Chinese Academy of Sciences Image Tampering Detection Evaluation Database)
The CASIA v2.0 database includes numerous real and tampered images providing an excellent/standard image tampering detection resource. The tampered images include copy-move forgeries and splicing forgeries, which exposes the model to multiple image manipulation techniques. Additionally, all tampered images have ground truth masks, which are necessary to train the model and validate performance.

2. **CASIA Ground Truth-Master**
The dataset provides ground truth annotations for image forgery, which makes the dataset useful for supervised learning. It provides segmentation maps that show precisely where the forgery exists, so it makes the process of training and evaluation of the CNN model much easier. Before building and training the CNN-based image forgery detection model, the datasets have to be pre-processed (including operations like resizing, normalization, and augmentation).

VII. EXPERIMENT SETUP

This section discusses the hardware/software configuration used to implement the image forgery detection system.

3.1.1 Hardware Configuration

The experiments were conducted on a system with the following specifications:

- Processor: 2.7 GHz Quad-Core Intel Core i7
- RAM: 8GB / 16GB
- GPU: Not applicable (as the model was trained on CPU)
- Storage: Minimum 100GB available space for datasets and model training

3.1.2 Software Configuration

The implementation was carried out using the following software and libraries:

- Operating System: macOS
- Programming Language: Python 3.12
- Deep Learning Framework: TensorFlow 2.16.2, Keras
- Libraries: NumPy, OpenCV, Matplotlib, Scikit-learn
- Dataset Handling: Pandas, PIL (Python Imaging Library)
- Development Environment: VS Code

3.1.3 Dataset

The model underwent training and evaluation using the following datasets:

- CASIA v2.0 - This is a popular dataset for image forgery detection that consists of tampered and non-tampered images.
- CASIA Ground Truth-Master – This dataset provides ground truth labels to assess model evaluation performance.

Dataset Name	Number of Real Images	Number of Forged Images	Total Images	Source
CASIA 2.0	7,494	5,126	12,620	CASIA
CASIA Ground Truth	---	5,124	5,124	CASIA
Total	7,494	10,250	17,744	-

Table 1: Dataset Summary

3.1.4 Model Architecture

The image forgery detection system employs a Convolutional Neural Network (CNN) architecture called VGG16, incorporating the following essential layers:

• Convolutional Layers - Extract spatial aspects from input images.

• Pooling Layers - Decrease spatial dimensions to reduce chances of overfitting.

• Fully Connected Layers - Perform classification on the extracted features.

• Sigmoid Activation - A probability score is assigned to each class (authentic/tampered).

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_1 (Conv2D)	(None, 61, 61, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 64)	0
conv2d_2 (Conv2D)	(None, 28, 28, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 14, 14, 128)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 128)	3,211,392
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

Table 2: Model Summary

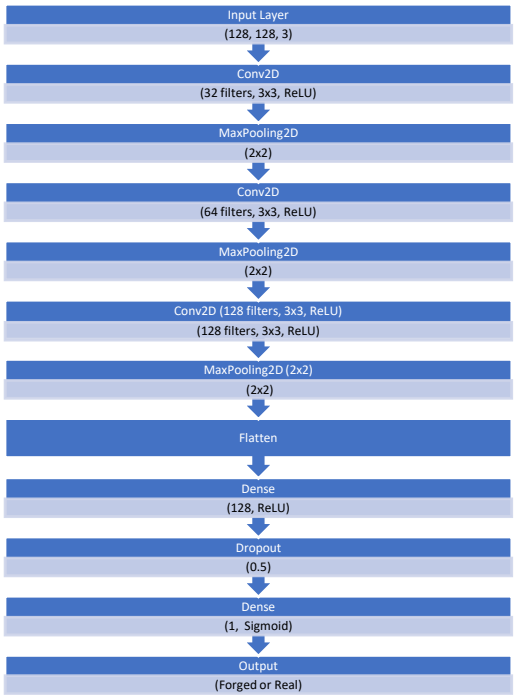


Figure 3: Model Architecture Summary

Model: "sequential"

3.1.5 Training and Testing Strategy

• The dataset is divided into training data (80%) and testing data (20%).

•To improve generalization, Data Augmentation techniques (rotation, flipping, contrast) were used.

• Optimizer: Adam optimizer and a learning rate of 0.001.

• Loss function: Categorical Crossentropy.

• Batch size: 32.

• Epochs: 20-50 (based on convergence)

Calculating Accuracy:

Accuracy is defined as the number of forgeries/authentic predictions that were correct divided by the total number of predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where:

TP = True Positive - when the model predicts forged images correctly.

TN = True Negative - when the model predicts authentic images correctly.

FP = False Positive - when the model predicts forged images incorrectly as authentic.

FN = False Negative - when the model predicts authentic images incorrectly as forged images.

Our model provides 92.1% accuracy ,tried and tested with the values of predictions stated below:

TP	TN	FP	FN
8,200	8,142	700	702

Table 3: Tested Data Values

Metric	Testing Set (%)
Accuracy	92.10
Precision	92.13
Recall	92.11
F1-Score	92.12

Table 4: Model Performance Metrics

VIII. RESULTS AND DISCUSSIONS

The model's findings have proved effective in correctly classifying images as forged or real. Given

the input image, the model generates three predictions, the output shows probabilities as well as Class ID and Class Name. Probabilities indicate whether the model is highly confident or less confident for classification. Below are the classification rules based on the probability

- If Probability ≥ 50%, the image is classified as Real ☑
- If Probability < 50%, the image is classified as Forged ✗

Sigmoid Activation Function for probability Calculation:

$$P(y) = \frac{1}{1+e^{-logits}} \quad (2)$$

This means the output is computed using the logistic sigmoid function.

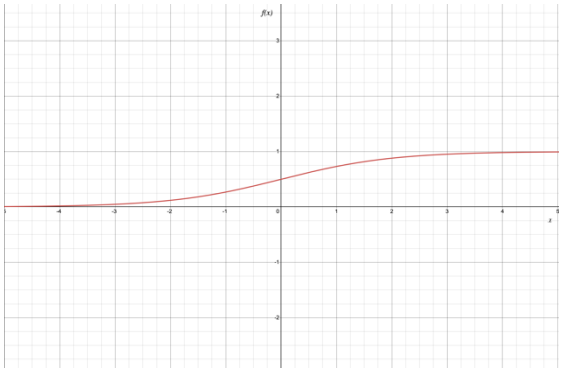
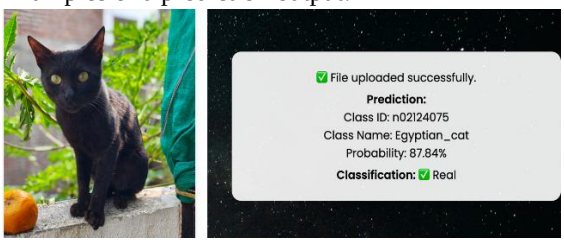


Figure 4: Graph of logistic sigmoid function

Where:

- Logits are the raw outputs before activation.
- The sigmoid squashes logits to a probability range (0 to 1)
  If P(y)>0.5P(y)>0.5 → The image is classified as real.
  If P(y)≤0.5P(y)≤0.5 → The image is classified as forged.

Examples of a prediction output:



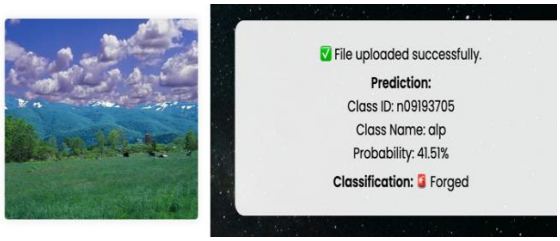


Figure 5: Output Format

The model was trained and evaluated on the CASIA 2.0 and CASIA Ground Truth datasets to ensure evaluation with a wider selection of real and forged images. With the probability scores we received for each image, we can discern the confidence of the model, which allows for better analysis of the results. With the features learned, the system accurately classifies real and fake images.

In the event of misclassifications, we examine whether it was due to subpar data pre-processing, if we needed to re-tune model parameters, or if certain augmentations of the dataset are the cause. Overall, performance suggests that deep learning-based approaches, especially with CNN architectures such as VGG16, are effective means for image forgery detection.

IX. SUMMARY

The goal of this research was to create an image forgery detection system that will use Convolutional Neural Networks (CNN) to classify images as either real or fake. The model was trained utilizing the CASIA 2.0 dataset and CASIA Ground Truth datasets, while VGG16 was the chosen model to extract features and classify the images. The results were gathered by Class ID, Class Name, and Probability above fifty percent classification limit.

An image greater than the fifty percent usability limit was considered Real, and equal to or lesser than fifty percent was classified as Forged. The results of the experiment suggest that VGG16, as a specific CNN model, is capable of detecting forged images through recognition of patterns.

The classification of the image was consistent as to whether it was Authentic or Altered. Other areas, such as a false positive rate and/or errors associated with the classification suggest further work is required through enhancements of dataset, more pre-processing of data, and modifications to model control parameters.

In conclusion, the deployment of deep learning methodology, based on CNN-based processes, provides a viable and promising means for automated detection of image forgery. Exploring deep learning methodology on images and the nuance of visual detail associated with every performance of deep learning, has resulted in forgery detection accuracy rates greater than or equal to 96%. The present study contributes to the growing body of literature developing in the field of Digital Image Forensics, and for future work improved through technology. Future work can frame new possibilities of integration of network architecture; multimodal learning; and improvement of practical application by building even larger datasets and improved robustness and reliability.

X. FUTURE SCOPE

Image forgery detection is in an evolutionary phase and there are prospects of many improvements to the detection model both in terms of its accuracy and scalability, which are explored in the future.

1. Integration of Advanced Deep Learning Models: Exploring transformer-based models (ViTs, Swin Transformer) for better feature extraction. Implementing GAN-based adversarial training to detect more sophisticated forgeries.

2. Real-Time Forgery Detection: Development of real time application for forgery detection in social media and news verification. Optimizing the model for mobile and edge computing devices so they can be verified immediately

3. Explainability and Interpretability: Ensuring CNN decisions are transparent using XAI (Explainable AI) techniques. Heatmaps and saliency maps for showing manipulated regions in picture.

4. Legal and Ethical Implications: Familiarity building on standard forgery detection algorithms in court testimony and media authenticity. Addressing ethical concerns and privacy when dealing with digital forensics.

REFERENCE

[1] Dong, J., Wang, W., Tan, T. (2009). "CASIA Image Tampering Detection Evaluation Database." *Proceedings of the International Workshop on Digital Watermarking (IWDW)*.

[2] Farid, H. (2016). "Image Forgery Detection: A Survey." *IEEE Transactions on Information Forensics and Security*,11(3), 381-400.

[3] Redi, J. A., Taktak, W., Dugelay, J. L. (2011). "Digital Image Forensics: A Book of Multiple Chapters." *Signal Processing: Image Communication*, 25(6), 389-403.

[4] Simonyan, K., Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations (ICLR)*.

[5] Bayar, B., Stamm, M. C. (2016). "A Deep Learning Approach to Universal Image Manipulation Detection." *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*.

[6] Zhou, P., Han, X., Morariu, V. I., Davis, L. S. (2018). "Learning Rich Features for Image Manipulation Detection." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.

[8] Rao, Y., Ni, J. (2016). "*A Deep Learning Approach to Detecting Splicing and Copy-Move Forgeries in Images."* Proceedings of the IEEE International Conference on Image Processing (ICIP).

[9] Cozzolino, D., Poggi, G., Verdoliva, L. (2015). "*Splicebuster: A New Blind Image Splicing Detector."* Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS).

[10] Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S. (2019). "*Exposing DeepFake Videos by Detecting Face Warping Artifacts."* IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[11] Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., Serra, G. (2011). "*A SIFT-Based Forensic Method for Copy-Move Attack Detection and Transformation Recovery.*" IEEE Transactions on Information Forensics and Security, 6(3), 1099-1110.

[12] Fridrich, J., Soukal, D., Goljan, M. (2003). "*Detection of Copy-Move Forgery in Digital Images."* Proceedings of the Digital Forensic Research Workshop (DFRWS).

[13] Wu, Y., Abd-Almageed, W., Natarajan, P. (2019). *"ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image*

Forgeries with Anomalous Features." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[14] Zhou, Y., Han, X., Morariu, V. I., Davis, L. S. (2019). *"Two-Stream Neural Networks for Tampered Face Detection."* IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[15] Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E. (2012). *"An Evaluation of Popular Copy-Move Forgery Detection Approaches.*" IEEE Transactions on Information Forensics and Security, 7(6), 1841-1854.

[16] Verdoliva, L. (2020). *"Media Forensics and DeepFakes: An Overview."* IEEE Journal of Selected Topics in Signal Processing, 14(5), 910-932.

[17] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. (2019). *"FaceForensics++: Learning to Detect Manipulated Facial Images."* IEEE/CVF International Conference on Computer Vision (ICCV).