International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Combining Deep Learning and Forensic Analysis for Robust Fake Image Detection

Sonia Yadav[1], Amresh Kumar[2]

[1]Research Scholar, Maharashi Dayanand University, Rohtak-124021, India
[2]AssistantProfessor, CBS Group of Institutions, Jhajjar-124103,India

1. Abstract

The increasing sophistication of image editing tools and AI-generated content has led to a surge in the dissemination of fake images across digital platforms.This study proposes a hybrid method for detecting fake images by integrating Convolutional Neural Networks (CNNs) with conventional image forensic strategies, leveraging the strengths of both approaches.CNNs are employed to detect pixel-level inconsistencies and complex patterns introduced during image manipulation, making them effective against high-quality forgeries such as deepfakes. In contrast, image forensics provides a complementary layer of analysis through metadata inspection, noise inconsistencies, and compression artifact detection using methods like Error Level Analysis (ELA) and Photo-Response Non-Uniformity (PRNU). By integrating these two paradigms, the proposed framework aims to enhance detection accuracy, interpretability, and robustness, particularly in real-world scenarios. Experimental evaluations demonstrate that the hybrid model outperforms standalone methods in detecting various types of fake images.The research also tackles critical issues including resistance to adversarial attacks, the ability to generalize across diverse datasets, and adapting to the rapid advancements in synthetic image generation techniques.This research aims to enhance the reliability and interpretability of fake image detection tools, focusing on their readiness for use in sensitive and high-impact contexts.

Keywords: Fake Image Detection, Deep Learning, Convolutional Neural Networks, Image Forensics, Hybrid Model, PRNU, Error Level Analysis, Deepfakes, Digital Image Manipulation, Adversarial Robustness

2. Introduction

In recent years, the authenticity of digital images has drawn increasing concern due to the rapid evolution of advanced editing software and AI-driven content generation tools. From highly realistic deepfakes to minor image edits, manipulated visuals have reached a level of realism that makes it increasingly difficult for both laypeople and experts to identify altered content. This surge in visual misinformation threatens the integrity of critical domains such as journalism, law enforcement, politics, and social media, where the reliability of visual evidence is paramount.Traditional image forensic techniques, including Photo-Response Non-Uniformity (PRNU), Error Level Analysis (ELA), and chromatic aberration detection, analyze properties like metadata, sensor noise, and compression artifacts to identify tampering [Ref]. These methods are often lightweight, interpretable, and suitable for resource-constrained settings. However, the rise of AI-generated content, especially through techniques

like Generative Adversarial Networks (GANs), has revealed the limitations of traditional forensics when confronted with highly realistic synthetic imagery [Ref].At the same time, deep learning methods—especially Convolutional Neural Networks (CNNs)—have shown significant promise in fake image detection by learning intricate, high-dimensional patterns that traditional techniques and human observers may overlook [Ref]. When trained on varied datasets, CNNs are able to identify a wide range of manipulations by detecting subtle irregularities created during image synthesis or editing.However, their lack of transparency and susceptibility to adversarial attacks present significant drawbacks, especially in high-stakes applications [Ref].This study introduces a hybrid framework that combines the interpretability of classical forensic methods with the powerful pattern-recognition capabilities of CNNs. By combining the distinct advantages of these methods, the proposed system aims to provide a reliable, understandable, and accurate solution for identifying fake images in varied practical settings.

3. Literature Review

The development of fake image detection techniques has closely followed the increasing complexity of image manipulation methods. Early investigations primarily utilized conventional forensic methods to determine if an image was authentic or altered.For instance, Fridrich developed tools such as Error Level Analysis, detection of repeated JPEG compression, and analysis of lighting and shadow mismatches, which were useful for spotting simple image alterations. Similarly, Farid highlighted pixel-level irregularities and geometric distortions as important signs of image forgery. Among sensor-based techniques, Lukas, Fridrich, and Goljan developed Photo-Response Non-Uniformity (PRNU), a method that uses unique sensor noise patterns to trace the origin of an image.To tackle these issues, the research community shifted towards machine learning, especially deep learning, because it excels at automatically learning key patterns and improving detection accuracy.Bayar and Stamm [4] introduced one of the first CNN-based models designed specifically for manipulation detection, which used a constrained convolutional layer to filter out irrelevant image content and highlight tampered areas. Following this, Rao and Ni [5] developed a CNN architecture that learned features directly from both manipulated and untouched image patches, eliminating the need for handcrafted features. Although these models performed well on specific datasets, they still faced challenges when dealing with unseen data and adversarial attacks.Over time, the field saw the emergence of advanced models such as XceptionNet[6], which researchers later adopted for detecting deepfake content with notable success.[7] and MesoNet [8], both of which demonstrated strong performance in detecting synthetic content, including deepfakes generated through GANs. These models take advantage of both spatial and frequency-domain signals to identify subtle irregularities introduced during image synthesis. However, as Goodfellow et al. As highlighted in prior research, CNN-based models typically lack interpretability and can be susceptible to adversarial manipulation, which raises concerns about their dependability in practical applications.With the advancement of research, more advanced models such as XceptionNet were developed and later applied by Rossler et al. and MesoNetfor detecting synthetic media. It shows strong results in detecting synthetic content, including deepfakes generated using GANs. These models took advantage of both spatial and frequency-domain features to catch subtle inconsistencies introduced during image synthesis.However, as Goodfellow et al. pointed out, CNN-based models can be difficult to interpret and are often susceptible to adversarial attacks, which raises concerns about how dependable they are in real-world situations.Advancements in this field have been driven significantly by the growing access to large and varied datasets.Like, the FaceForensics++ dataset [7] got many kinds of

manipulations like DeepFakes, NeuralTextures, and FaceSwap, and it helps a lot in training and checking models. Also, Facebook AI made the Deepfake Detection Challenge (DFDC) dataset [12], it has more than 100,000 videos and is also very important. Lately, Zhou et al. [13] made the Realistic Face Manipulation Dataset (RFMD), which tries to fill the gap between fake and real image situations.Even with all these new ideas, there's still some big problems. These systems still struggle when it comes to handling various kinds of edits, especially the ones they haven't seen before or that are designed to fool them.Plus, they usually don't explain their decisions in a clear way, which makes it tough for users to fully trust the results.Yeh and others said it's super important to not just make the models accurate, but also easy to understand and use in real life. So in the future, researchers gotta build stuff that's not only smart but also tough, clear, and useful.

4. Methodology

In this study, we propose a hybrid approach that combines the strengths of Convolutional Neural Networks (CNNs) with traditional forensic analysis techniques to enhance the detection of fake images.Our objective focuses on enhancing the precision of detection, strengthening resilience to complex tampering, and making the outcomes easier to understand.The approach involves three key steps: preparing the images, extracting features using convolutional neural networks, and analyzing forensic characteristics, which are then integrated through a decision fusion process.

4.1Image Preprocessing

Initially, images are resized to a uniform dimension (such as 256×256 pixels) and adjusted to minimize differences caused by lighting conditions.We apply image enhancement techniques such as histogram equalization to improve feature visibility.A high-pass filtering technique is utilized to isolate noise patterns, making it easier to detect minor anomalies that often result from image alterations.

4.2 CNN-Based Feature Extraction

We employ a customized CNN architecture inspired by Bayar and Stamm (2016) and XceptionNet (Chollet, 2017), consisting of multiple convolutional layers with batch normalization, ReLU activations, and max-pooling. This network is trained on a labeled dataset of real and fake images, enabling it to learn discriminative features at pixel and texture levels.The convolutional neural network produces a probability value that reflects how likely the image is to be counterfeit.

4.3 Forensic Feature Analysis

Simultaneously, we extract forensic features grounded in image processing principles. These include:

- **Error Level Analysis (ELA):** Measures compression discrepancies to reveal altered regions.
- **Photo-Response Non-Uniformity (PRNU):**Analyzes sensor noise inconsistencies unique to authentic cameras.
- **Metadata Consistency Checks:** Examines EXIF data for irregularities.
- **Frequency Domain Analysis:** Uses Discrete Cosine Transform (DCT) coefficients to detect artifacts introduced by GANs or editing.

These features provide interpretable signals that complement the CNN's learnedrepresentations.

4.4 Decision Fusion

The predictions generated by the CNN classifier and the forensic analysis component are combined by assigning different importance weights to each before making the final decision.We experimented with logistic regression and support vector machines (SVM) to fuse these features effectively. The final decision score benefits from both deep feature abstraction and explicit forensic evidence, improving robustness to diverse types of manipulations.

4.5 Implementation and Training Details

We developed our combined model using Python, leveraging the TensorFlow framework alongside the OpenCV library.The CNN is trained for 50 epochs using Adam optimizer with a learning rate of 0.0001 and binary cross-entropy loss. We augmented the dataset with rotations, flips, and noise addition to improve generalization. For forensic features, handcrafted algorithms were implemented and optimized for batch processing.

5. Result & Discussion

The experimental evaluation of the proposed hybrid model was carried out in a well-defined setup utilizing Python 3.9, TensorFlow 2.10, Keras, OpenCV, and various image processing libraries such as NumPy, Scikit-learn, and ExifRead.Model training and testing were conducted on a robust computing platform configured with Windows 11, featuring an Intel Core i7 CPU, 32 GB system memory, and an NVIDIA RTX 3080 graphics card to ensure efficient processing of deep learning operations.To assess the performance of the developed detection model, experiments were carried out using three established and publicly available datasets: FaceForensics++, CASIA TIDE v2.0, and the DeepFake Detection Challenge dataset.These datasets collectively encompass diverse manipulation techniques, including synthetic face generation, region splicing, and duplication-based tampering, providing comprehensive coverage of real-world forgery scenarios.For systematic model training and unbiased assessment, the data from each dataset was divided into three subsets: 80% for training purposes, 10% for validation, and the remaining 10% for final testing.To enhance the robustness of the model and prevent it from overfitting to specific training examples, a range of augmentation techniques were utilized. These included horizontal mirroring, angular rotation, and other transformations to introduce diversity into the training data.Performance was measured using standard metrics such as Accuracy, Precision, Recall, F1-Score, and AUC. The results clearly indicate the superiority of the hybrid model over the standalone CNN and forensic-only approaches. The CNN-only model achieved 91.3% accuracy and an AUC of 0.925, while the forensic-only model lagged behind with 83.6% accuracy and 0.842 AUC. In contrast, the hybrid model delivered a robust performance with 95.8% accuracy, 94.9% F1-score, and a 0.971 AUC, demonstrating its ability to integrate semantic feature learning and low-level forensic signal detection effectively.Manual examination of test images—such as those containing deepfake faces or inserted elements—demonstrated the enhanced capability of the hybrid model. It effectively highlighted manipulated regions using visual cues derived from Error Level Analysis (ELA) and Photo-Response Non-Uniformity (PRNU) maps.The ROC curve also highlighted a significant margin in favor of the hybrid architecture.The discussion of these findings emphasizes the complementary strengths of both components in the hybrid model. CNNs are excellent at learning complex high-level patterns but may

lack interpretability, while forensic techniques offer transparent detection of anomalies such as compression artifacts and sensor inconsistencies.Integrating both analysis streams enabled the model to leverage the unique strengths of each for improved detection performance.Challenges included difficulties in harmonizing the outputs from different components, increased resource consumption during processing, and decreased accuracy when dealing with images that had undergone substantial compression.Furthermore, enhancing the interpretability of predictions driven primarily by CNN features remains a challenge that future work should address.Future enhancements could include explainable AI methods such as Grad-CAM and optimization for real-time applications, making this model suitable for digital forensics, law enforcement, and media integrity verification.

6. Conclusion

This study introduced a hybrid framework for detecting fake images by blending machine learning techniques with traditional image forensic methods. By merging the capabilities of Convolutional Neural Networks (CNNs) with pixel-level forensic tools such as Error Level Analysis (ELA) and Photo-Response Non-Uniformity (PRNU), the system achieved enhanced accuracy in identifying and localizing manipulated visual content. Publicly available datasets—FaceForensics++, CASIA TIDE v2.0, and DFDC—were used to ensure a diverse and realistic training environment, reflecting various tampering strategies including deepfakes, splicing, and copy-move edits.Experimental results demonstrated that this hybrid setup outperformed standalone deep learning models, particularly in detecting subtle manipulations and highlighting tampered regions using interpretive heatmaps. The model's effectiveness was validated through detailed statistical assessment, employing metrics such as overall prediction accuracy, the rate of correctly identified fake instances (precision), the ability to detect actual fakeimages (recall), and the model’s stability in classification performance measured through the area under the ROC curve.Nonetheless, the approach encountered certain challenges, such as increased processing demands, complex result integration, and reduced efficiency on low-quality or compressed images. These areas present opportunities for future enhancement through model optimization, better fusion strategies, and improved interpretability.Overall, the proposed approach proves to be a reliable and effective solution for fake image detection, with strong potential for deployment in digital forensics, content verification, and countering visual misinformation.

References

[1] R. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1–11.

[2] J. Dong, W. Wang, and T. Tan, “CASIA image tampering detection evaluation database,” in Proc. IEEE China Summit & Int. Conf. Signal and Information Processing, Beijing, China, 2013, pp. 422–426.

[3] B. Dolhansky et al., “The DeepFake detection challenge (DFDC) dataset,” arXiv preprint arXiv:2006.07397, 2020.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[5] X. Zhang, S. Wang, and Y. Su, “An overview of image forensics techniques for detecting fake images,” Multimedia Tools and Applications, vol. 80, no. 1, pp. 1–24, 2021.

[6] M. Barni, A. Costanzo, and L. Tondi, “A universal image integrity model for content authentication,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2933–2948, Nov. 2018.

[7] J. Fridrich, “Digital image forensics,” IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 26–37, Mar. 2009.

[8] S. Bayar and M. C. Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer,” in Proc. ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), 2016, pp. 5–10.

[9] Z. Lin, R. Wang, X. Tang, and H. Shum, “Detecting doctored images using JPEG ghosts,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 1–7.

[10] T. Chen, K. Chen, Q. Ma, and J. Shi, “Image tampering localization via PRNU and noise level inconsistencies,” Forensic Science International: Digital Investigation, vol. 39, 2021.

[11] M. Hussain, M. Hussain, and S. A. Malik, “Image forgery detection: A survey,” in Multimedia Tools and Applications, vol. 75, no. 15, pp. 9253–9284, 2016.