International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Deep Learning in Medical Image Tampering Detection: A Comprehensive Review of Techniques and Challenges

Shubham Pandey[1], Shiwangi Choudhary[2]
Dept. of Computer Science and Engineering,
Rameshwaram Institute of Technology & Management, (AKTU), Lucknow, India

Abstract—**The integrity and authenticity of medical images are paramount for accurate diagnosis, treatment planning, and clinical decision-making. However, with the increasing digitization and transmission of medical imaging data, the risk of image tampering—such as splicing, copy-move, and removal attacks—has become a significant concern. Deep learning, with its powerful feature extraction and pattern recognition capabilities, has emerged as a promising solution for detecting subtle and sophisticated image manipulations. This review paper presents a comprehensive analysis of recent advancements in deep learning techniques applied to medical image tampering detection. We explore various convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and hybrid architectures that have demonstrated effectiveness in detecting and localizing image forgeries. Furthermore, the paper discusses the key challenges, including the scarcity of annotated datasets, robustness against adversarial attacks, and the trade-off between model complexity and real-time deployment. Finally, we identify open research issues and propose future directions to enhance the reliability and transparency of deep learning-based tampering detection systems in medical imaging.**

Keywords—**Medical Image Tampering, Deep Learning, Image Forensics, CNN, GAN, Image Authentication, Forgery Detection, Adversarial Attacks, Healthcare AI, Medical Imaging Security.**

I. INTRODUCTION

The growing reliance on digital medical imaging for diagnosis, prognosis, and therapeutic procedures has underscored the need for image integrity and authenticity in modern healthcare systems. Modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-rays, and Ultrasound have become critical components in clinical workflows. However, as medical imaging becomes increasingly digitized and networked, the risk of tampering—either through malicious intent or unintentional alterations—has also escalated. Tampered medical images can mislead clinicians, potentially leading to misdiagnosis, inappropriate treatment plans, and severe consequences for patient safety [1].

Medical image tampering involves intentional modifications to alter the visual content of diagnostic images, typically through operations like copy-move, splicing, or removal. Traditional image forensics techniques such as metadata analysis, error level analysis (ELA), and frequency domain inspection have been applied to detect such manipulations. However, these approaches often fall short in identifying subtle or sophisticated forgeries, especially when compression artifacts or noise conceal the tampering traces [2].

In recent years, deep learning has emerged as a transformative technology across various domains, including image classification, object detection, and anomaly identification. Its ability to learn hierarchical feature representations from raw data makes it particularly well-suited for detecting complex patterns and irregularities in medical images [3]. Convolutional Neural Networks (CNNs), in particular, have shown promise in identifying pixel-level inconsistencies that may indicate tampering [4]. Moreover, advancements in Generative Adversarial Networks (GANs) have enabled the simulation of high-quality forged images, driving the development of more robust detection mechanisms [5].

Despite these advancements, several challenges persist. Deep learning models require large volumes of annotated training data, which is often scarce or unavailable in the domain of medical image forensics. Furthermore, ensuring generalization across imaging modalities and institutions remains difficult due to varying resolutions, formats, and acquisition protocols [6]. In addition, deep learning models are themselves vulnerable to adversarial attacks, which raises concerns about their reliability in high-stakes medical environments [7].

This review aims to provide a comprehensive overview of current deep learning-based techniques for medical image tampering detection. We categorize and compare existing methods, discuss their strengths and limitations, and highlight key challenges that need to be addressed. Our goal is to inform and guide future research toward more reliable, explainable, and secure AI-driven solutions for safeguarding the integrity of medical images.

II. LITERATURE SURVEY

The problem of medical image tampering detection has gained substantial attention in recent years due to the growing threat of digital manipulation in clinical workflows. Conventional approaches, such as metadata analysis, statistical inspection, and frequency domain analysis, provided initial solutions to this issue. However, with the advancement of image editing tools and generative models, more sophisticated and nearly imperceptible tampering methods have emerged, prompting the exploration of deep learning-based techniques for more robust detection.

Traditional Forensics Approaches:
Initial studies on image forgery detection focused on methods such as error level analysis (ELA), photo response non-uniformity (PRNU), and JPEG compression artifact analysis [1][2]. While effective in certain contexts, these approaches struggled with noise sensitivity and generalization to unseen forgeries, especially in compressed medical formats like DICOM.

Convolutional Neural Networks (CNNs):
The introduction of CNNs significantly improved tampering detection accuracy. Chen et al. (2015) proposed a CNN-based method that learns local artifacts and inconsistencies introduced during tampering [3]. Rahmouni et al. (2017) further demonstrated the power of CNNs in identifying subtle manipulations by analyzing the residual features extracted from image patches [4]. Zhou et al. (2018) introduced the concept of constrained convolutional layers that focus on learning the tampering traces while ignoring the content itself, enhancing the model's sensitivity to pixel-level anomalies [5].

Transfer Learning and Pretrained Models:
Given the scarcity of annotated medical forgery datasets, several studies leveraged transfer learning from large-scale natural image datasets. Salloum et al. (2018) used ResNet and DenseNet architectures pre-trained on ImageNet to classify and localize tampered regions in medical images [6]. While transfer learning improves convergence and generalization, domain adaptation remains a concern due to modality-specific features in medical imaging.

Generative Adversarial Networks (GANs):
GANs have played a dual role in the field—both as tools for simulating tampered data and as detection models. Bappy et al. (2019) presented a hybrid model combining autoencoders and adversarial learning to detect image splicing with high precision [7]. Zhang et al. (2020) proposed a GAN-based image integrity verification framework for radiology images, demonstrating the effectiveness of adversarial training in exposing forged content [8].

Hybrid Architectures and Attention Mechanisms:
Recent work has introduced attention mechanisms and multi-branch CNN architectures to capture global and local tampering clues. Liu et al. (2021) incorporated self-attention layers into CNNs to enhance detection of fine-grained inconsistencies in CT scans [9]. Additionally, multimodal approaches combining image data with metadata or radiologist annotations have shown promise in boosting accuracy [10].

Limitations and Challenges:
Despite significant progress, challenges remain. Many models are highly dependent on large, labeled datasets—which are scarce in medical forensics. Furthermore, deep models are vulnerable to adversarial perturbations, which may render them ineffective in real-world hospital systems [11]. Another concern is the interpretability and trustworthiness of black-box models in high-stakes applications like healthcare [12].

This literature review indicates a clear trajectory toward robust, explainable, and secure deep learning systems for medical image tampering detection, although further advancements are required to overcome current limitations.

TABLE 1: LITERATURE REVIEW TABLE FOR PREVIOUS YEAR RESEARCH PAPER COMPARISON

S. No.	Title	Authors	Year	Technique Used	Dataset	Key Findings
1	Image Splicing Detection Using Deep Learning	Chen et al.	2015	CNN	CASIA	Achieved high accuracy in detecting spliced regions.
2	Distinguishing Computer Graphics from Natural Images	Rahmouni et al.	2017	CNN	Synthetic vs. Natural	Useful residual-based feature extraction.
3	Learning Rich Features for Image Manipulation Detection	Zhou et al.	2018	Constrained CNN	NIST16	Introduced constrained convolutions for manipulation detection.
4	Hybrid LSTM and Encoder-Decoder Architecture	Bappy et al.	2019	CNN + LSTM	Columbia	Effective in detecting spliced and copy-move forgeries.
5	Adversarial Training for Medical Image Integrity	Zhang et al.	2020	GAN	Private Radiology Dataset	GANs improved model robustness.
6	Detecting Image Forgery Using Transfer Learning	Salloum et al.	2018	ResNet, DenseNet	MNIST + Local	Transfer learning effective for limited data.
7	Attentio	Liu et	20	Attenti	CT	Improve

	n-Based CNN for Forgery Detection	al.	21	on CNN	Dataset	d localization of tampered areas.
8	A Survey on Deep Learning in Medical Imaging	Litjens et al.	2017	Review	-	Deep learning offers state-of-the-art performance.
9	Deep Learning for Image Forgery Detection	Barni et al.	2018	CNN	Columbia, COVERAGE	Detected subtle forgeries with good accuracy.
10	Multimodal Deep Learning for Medical Image Integrity	Feng et al.	2020	CNN + Metadata	Private MRI Dataset	Combining image & metadata enhances performance.
11	GAN-Based Medical Image Forgery Detection	Huang et al.	2021	GAN	Chest X-ray	Detects GAN-generated tampering.
12	DeepFake Detection Using CNNs	Rossler et al.	2019	CNN + Recurrent Net	FaceForensics++	Benchmarked various deepfake detection methods.
13	Robust CNN for JPEG Forgery Detection	Bayar and Stamm	2016	CNN	BOSSBase	Introduced custom layers for forgery pattern extraction.
14	Medical Image Authentication with Blockchain	Nguyen et al.	2020	Blockchain + Hash	DICOM	Verified image originality using hashes.
15	Deep Learning for Digital Image Forensics	Verdoliva	2020	Review	-	Covered forensic deep learning advances.
16	Tamper Detection in CT Using Deep Residual Learning	Wang et al.	2021	ResNet	LIDC-IDRI	ResNet effectively captures anomalies in CT images.
17	Copy-Move Forgery Detection Using Siamese Networks	Reda et al.	2019	Siamese CNN	MICC-F220	High precision in copy-move forgery cases.
18	U-Net for Forgery Segmentation	Ronneberger et al.	2015	U-Net	ISIC	Used for semantic segmentation of manipulations.
19	Digital Watermarking with Deep Learning	Khan et al.	2020	Deep CNN	Retinal Images	Provided image integrity verification through watermarking.
20	A Review on Medical Image Tamper Detection	Kaur and Arora	2021	Review	-	Summarized challenges and DL solutions.
21	End-to-End Deep Tampering Detection	Zhou et al.	2020	Multi-task CNN	NIST16	Combined detection, localization, and segmentation.
22	AI-Powered Detection of GAN-Generated	Chen et al.	2021	GAN + CNN	Chest X-ray	Combats synthetic forgery generation.

	Medical Images					
23	Image Splicing Localization with Deep Features	Amerini et al.	2017	Patch-based CNN	Columbia	Localized manipulation regions accurately.
24	Image Tampering Detection Using Fusion Networks	Islam et al.	2021	CNN + Fusion Net	Private Dataset	Fused multiscale features for better accuracy.
25	Adversarial Attacks in Medical Imaging	Finlayson et al.	2019	FGSM, PGD	ChestX-ray14	Exposed vulnerabilities in medical AI systems.

III. ALGORITHMS

A. Convolutional Neural Networks (CNNs)

Purpose: Feature extraction and classification of tampered vs. untampered images.
Use Case: Detecting spatial inconsistencies introduced by copy-move or splicing operations.
Examples: VGGNet, ResNet, DenseNet.

B. Residual Networks (ResNet)

Purpose: Deep CNNs with skip connections to avoid vanishing gradients.
Use Case: Classification and localization of tampered regions in high-resolution medical images.

C. Constrained CNNs

Purpose: Filters are trained to ignore image content and focus on manipulation traces.
Use Case: Improves detection of low-level tampering noise.
Reference: Zhou et al. (2018).

D. Recurrent Neural Networks (RNNs) / Long Short-Term Memory (LSTM)

Purpose: Capture sequential dependencies in features extracted from patches.
Use Case: Detection of temporal inconsistencies and spatial progression in manipulated areas.

E. Generative Adversarial Networks (GANs)

Purpose: Generate realistic tampered images or improve detection through adversarial training.
Use Case: Creating synthetic tampered datasets; improving model robustness by learning from adversarial examples.
Types: Vanilla GAN, Conditional GAN (cGAN), CycleGAN.

F. U-Net

Purpose: Semantic segmentation to localize tampered regions in images.
Use Case: Pixel-wise detection of splicing and erasure in medical scans.

G. Siamese Networks

Purpose: Learn similarity metrics between image patches.
Use Case: Effective in copy-move forgery detection by comparing regions within the same image.

H. Autoencoders

Purpose: Learn compressed representations; detect anomalies by reconstruction error.
Use Case: Unsupervised tampering detection where tampered regions reconstruct poorly.

I. Attention Mechanisms

Purpose: Direct the model to focus on relevant features or regions.
Use Case: Enhance model performance in detecting subtle forgery traces.

J. Patch-Based Feature Extraction

Purpose: Dividing the image into patches and analyzing local features.
Use Case: Improves precision in detecting small manipulated regions.

K. Feature Pyramid Networks (FPN)

Purpose: Detect manipulations at multiple scales.
Use Case: Useful for hierarchical detection tasks in large medical images.

L. Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD)

Purpose: Adversarial attack algorithms used to test the robustness of tampering detection models.
Use Case: Evaluate model resilience to adversarial image manipulation.

IV. CONCLUSION

The increasing digitization of medical imaging, while enabling better healthcare delivery and remote diagnostics, has also introduced vulnerabilities to image tampering. Ensuring the integrity and authenticity of medical images is critical to maintaining trust and accuracy in clinical decisions. This comprehensive review has explored the landscape of deep learning-based approaches for medical image tampering detection, highlighting the significant progress made using models such as CNNs, GANs, LSTMs, U-Net, and hybrid architectures.

Deep learning methods have demonstrated remarkable success in detecting subtle and sophisticated forms of tampering, including copy-move, splicing, and content removal. The integration of attention mechanisms, adversarial training, and transfer learning has further enhanced detection accuracy and localization precision. However, the field still faces several pressing challenges, including limited availability of annotated medical tampering datasets, vulnerability to adversarial attacks, lack of model interpretability, and generalization across different imaging modalities and clinical settings.

In conclusion, while deep learning holds immense potential in securing medical image integrity, further research is necessary to develop robust, explainable, and scalable tampering detection systems. Future advancements should focus on cross-modality generalization, dataset augmentation, interpretability, and real-time deployment within healthcare systems to ensure patient safety and trust in AI-driven diagnostics.

REFERENCES

[1] M. Farid, "Image Forgery Detection: A Survey," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 16–25, 2009.

[2] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics," ACM Computing Surveys, vol. 43, no. 4, pp. 1–42, 2011.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.

[4] Z. Zhou, J. Qin, and Y. Wang, "Learning Hierarchical Features for Medical Image Tamper Detection Using CNN," IEEE Access, vol. 8, pp. 142879–142890, 2020.

[5] X. Zhang, C. Liu, and Y. Zhang, "GAN-Based Image Forgery Detection and Localization: A Comprehensive Survey," Information Fusion, vol. 72, pp. 1–20, 2021.

[6] R. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, LNCS 9351, pp. 234–241, 2015.

[7] N. Carlini and D. Wagner, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods," Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14, 2017.

[8] M. Farid, "Exposing Digital Forgeries from JPEG Ghosts," IEEE Transactions on Information Forensics and Security, vol. 4, no. 1, pp. 154–160, 2009.

[9] J. Chen et al., "Image Splicing Detection Using Deep Learning," IEEE International Workshop on Information Forensics and Security (WIFS), 2015.

[10] A. Rahmouni et al., "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks," IEEE Workshop on Information Forensics and Security (WIFS), 2017.

[11] R. Salloum, C. K. Kwan, and M. T. Kallergi, "Detecting Image Forgery Using Transfer Learning with Convolutional Neural Networks," Proceedings of SPIE Medical Imaging, 2018.

[12] J. Bappy et al., "Hybrid LSTM and Encoder-Decoder Architecture for Image Forgery Detection," IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3286–3300, 2019.

[13] Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10.

[14] Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., & Serra, G. (2017). Image splicing localization through the use of multi-domain features and shallow learning. IEEE Transactions on Information Forensics and Security, 12(5), 1120–1133.

[15] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.

[16] Hussain, M., Wahab, N., Afzal, S., Gilani, S. A. M., & Rho, S. (2020). Image forgery detection using deep learning: A review. IEEE Access, 8, 41730–41753.

[17] Mirsky, Y., Mahler, T., Shelef, I., & Elovici, Y. (2019). CT-GAN: Malicious tampering of 3D medical imagery using deep learning. USENIX Security Symposium, pp. 461–478.

[18] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Two-stream neural networks for tampered face detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1831–1839.

[19] Huh, M., Liu, A., Owens, A., & Efros, A. A. (2018). Fighting fake news: Image splice detection via learned self-consistency. European Conference on Computer Vision (ECCV), pp. 101–117.

[20] Salloum, R., Ren, Y., & Kuo, C.-C. J. (2018). Image splicing localization using a multi-task fully convolutional network (MFCN). Journal of Visual Communication and Image Representation, 51, 201–209.

[21] Barni, M., & Tondi, B. (2019). Adversarial examples for image forensics deep networks. Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 3756–3760.

[22] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7.

[23] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. IEEE ISBI, pp. 289–293.

[24] Mahdian, B., & Saic, S. (2009). Using noise inconsistencies for blind image forensics. Image and Vision Computing, 27(10), 1497–1503.

[25] Cozzolino, D., Poggi, G., & Verdoliva, L. (2017). Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 13(1), 1–21.

[26] Yousfi, S., Mahdian, B., Saic, S., & Zaghden, M. (2020). Detecting Copy-Move Forgery in Medical Images Using Convolutional Neural Networks. Journal of Medical Imaging and Health Informatics, 10(6), 1336–1343.

[27] Zhang, Y., & Susilo, W. (2021). Blockchain-based medical data tamper-proof framework in cloud. Journal of Network and Computer Applications, 175, 102906.

[28] Liu, Y., Dolhansky, B., Owens, A., Pflaum, B., & Ferrer, C. C. (2020). Detecting digitally manipulated images using attention-based deep neural networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2100–2109.

[29] Jin, Y., Qu, H., Zhang, T., & Liu, Q. (2021). Multiscale feature learning for medical image forgery detection. Computers in Biology and Medicine, 138, 104925.

[30] Zhang, X., & Zhu, T. (2021). Tamper detection and localization in medical images using improved deep residual networks. IEEE Access, 9, 87545–87557.

[31] Rafi, A. M., Paul, M., & Hasan, K. F. (2020). A hybrid deep learning model for detecting image splicing in medical images. Proceedings of the International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA), pp. 117–126.