International Journal of Scientific Research in Engineering and Management (IJSREM)
Special Edition - Volume 09 NCFT -2025 SJIF Rating: 8.586 ISSN: 2582-3930

Advancements in Image-Based Attack Detection: A Comparative Analysis of Copy-Paste and Medical Image Attacks Using Deep Learning Techniques

N V S K Vijayalakshmi K[1], Dr. J. Sasikala[2], Dr. C. Shanmuganathan [3]
[1]Research Scholar, Information Technology Annamalai University Annamalai Nagar Tamil Nadu, India, vijayakathari@gmail.com
[2] Professor, Information Technology Annamalai University Annamalai Nagar Tamil Nadu, India ,sasikala.au@gmail.com
[3] Assistant Professor, Computer Science and Engineering, SRM Institute of Science &Technology, Ramapuram, Chennai Tamil Nadu, India
drcsnathan@gmail.com

Abstract

In this paper, a comparative analysis of two distinct types of attacks is conducted: copy-paste attacks and attacks on medical images. The first part of the analysis focuses on techniques for detecting copy-paste attacks, involving the unauthorized duplication and manipulation of digital content. Methodologies, datasets, and performance metrics utilized in existing research are examined to evaluate the efficacy of detection methods. The second part of the analysis shifts focus to attacks on medical images. One approach involves detecting attacks using features extracted from Principal Component Analysis (PCA) and classified using Convolutional Neural Networks (CNN) and inception models. The third approach involves detecting attacks using features extracted from UNet and classified using deep autoencoders. Through this comparative analysis, common themes, challenges, and opportunities in the field of image-based attack detection are identified. The implications of these findings for enhancing the security and integrity of digital content in diverse applications, from multimedia forensics to healthcare systems, are discussed.

Keywords: CT scan image, Copy-paste attacks, Unet, deep autoencoder, CNN, PCA

1. Introduction

In today's digital age, the proliferation of digital images has revolutionized various aspects of modern life, ranging from entertainment and communication to healthcare and education. However, along with the benefits of digital imagery comes the inherent risk of unauthorized manipulation, tampering, and exploitation. Ensuring the security and integrity of digital images is paramount, particularly in critical domains such as healthcare, where patient diagnoses and treatment decisions rely heavily on medical imaging data. The vulnerability of digital images to malicious attacks poses significant challenges for maintaining trust, reliability, and confidentiality in digital systems. As such, there is a pressing need for robust techniques and methodologies to detect and prevent image-based attacks, safeguarding the authenticity and trustworthiness of digital content. In modern healthcare systems, the adoption of digital technologies has led to a significant increase in the use of medical imaging for diagnostic and treatment purposes [1]. However, along with the benefits of digitization comes the challenge of ensuring the security and privacy of sensitive medical image data. Medical image security encompasses a range of measures aimed at protecting the confidentiality, integrity, and availability of medical imaging data,

thereby safeguarding patient privacy and maintaining the trustworthiness of healthcare systems. One of the primary concerns in medical image security is the vulnerability of imaging data to adversarial attacks. Adversarial attacks involve the manipulation of medical images with the intent to deceive machine learning algorithms used for image analysis. By introducing imperceptible perturbations to images, attackers can cause misclassification or incorrect diagnoses, potentially compromising patient care. Research in this area has highlighted the susceptibility of deep learning models to adversarial attacks, prompting the development of defense mechanisms to enhance model robustness and reliability [2].

Data poisoning is another significant threat to medical image security. In data poisoning attacks, adversaries inject malicious data into training datasets used to develop machine learning models for medical image analysis. By corrupting training data with false or misleading information, attackers can manipulate the behavior of machine learning algorithms, leading to compromised model performance and erroneous predictions. Addressing the challenge of data poisoning requires robust data validation and preprocessing techniques to detect and mitigate the effects of malicious data injections [3]. Privacy breaches represent yet another aspect of medical image security concerns. Unauthorized access to medical imaging data can result in serious breaches of patient privacy and confidentiality. Such breaches not only violate patients' rights to privacy but also pose risks of identity theft, medical fraud, and other forms of exploitation. Protecting medical imaging data from unauthorized access requires the implementation of robust access control mechanisms, encryption protocols, and secure transmission protocols to safeguard data both at rest and in transit [4]. Deep learning and machine learning techniques are pivotal in detecting attacks faced by both medical and normal images, leveraging their capacity to extract intricate patterns and features from image data. Deep learning models, notably convolutional neural networks (CNNs), excel in learning hierarchical representations of image data. Trained on extensive datasets of both normal and attacked images, CNNs autonomously acquire discriminative features distinguishing genuine from manipulated images. This capability is particularly valuable in medical image security, where CNNs can identify subtle anomalies or alterations indicative of attacks such as adversarial perturbations or data poisoning [5].

Traditional machine learning algorithms, like support vector machines (SVMs) or random forests, are also used for feature learning but may require handcrafted feature extraction methods, which could be less effective at capturing intricate details in image data compared to deep learning approaches [6]. Furthermore, both deep learning and traditional machine learning techniques are deployed for anomaly detection tasks, wherein the aim is to identify deviations from normal patterns. Deep learning models, adept at recognizing abnormal patterns, can detect anomalous regions or artifacts introduced by attacks, such as tampered pixels or adversarial perturbations. Similarly, machine learning algorithms can be trained for anomaly detection tasks using handcrafted features or unsupervised learning techniques, although deep learning methods often surpass traditional machine learning approaches due to their ability to automatically learn hierarchical representations of data [7]. In addition, deep learning models, particularly CNNs, are widely employed for image classification tasks, which involve assigning labels to input images based on their content. In the context of image-based attacks, deep learning classifiers distinguish between normal and attacked images by learning the underlying characteristics associated with each class. For instance, in medical image security, CNNs classify images as authentic or manipulated based on learned features. While traditional machine learning classifiers can also be utilized for image classification tasks, they may

necessitate handcrafted features and might not generalize as effectively as deep learning models, especially when dealing with complex image data [8].

The overall contribution of this paper can be elucidated as given below

1. **Comprehensive Comparative Analysis:** Synthesizes findings from multiple studies on image-based attack detection, focusing on copy-paste attacks and attacks on medical images.
2. **Identification of Common Themes and Challenges:** Highlights recurring issues such as the vulnerability of deep learning models to adversarial attacks and the importance of robust feature extraction techniques.
3. **Guidance for Future Research Directions:** Offers insights for researchers and practitioners on exploring new deep learning architectures, developing more robust detection algorithms, and investigating emerging threats in image security.

2. Related work

Adversarial attacks have emerged as a significant concern in the field of AI-enabled medical imaging informatics, prompting researchers to investigate both attack strategies and defense mechanisms. Kaviani et al. (2022) [9] conducted a comprehensive survey of adversarial attacks and defenses in AI-based medical imaging informatics. Their survey provides insights into various attack methods, including adversarial perturbations, and explores defense strategies such as adversarial training and robust model architectures. Similarly, Muoka et al. (2023) [10] conducted a detailed review and analysis of deep learning-based adversarial attacks and defenses in medical image analysis. Their analysis delves into the vulnerabilities of deep learning models to adversarial perturbations and examines techniques for mitigating these vulnerabilities, including adversarial training and defensive distillation.

Dong et al. (2023) [11] contributed to the literature by presenting methods and applications of adversarial attack and defense specifically tailored for medical image analysis. Their work offers insights into the unique challenges posed by medical image data and proposes novel defense mechanisms to enhance the robustness of AI-based medical imaging systems. In addition to medical imaging, presentation attack detection methods have been extensively studied in other biometric recognition systems. Sousedik and Busch (2014) [12] conducted a survey of presentation attack detection methods for fingerprint recognition systems, highlighting the importance of robust authentication mechanisms in biometric security.

Moreover, watermarking techniques have been explored as a means of protecting the integrity and authenticity of medical images. Mousavi et al. (2014) [13] conducted a survey of watermarking techniques used in medical images, providing an overview of different watermarking approaches and their applicability in medical imaging contexts.While the aforementioned papers provide valuable insights into adversarial attacks and defenses in medical imaging informatics, they also highlight several challenges and gaps in the existing research landscape. These include the need for more robust defense mechanisms against sophisticated adversarial attacks, the lack of standardized evaluation protocols for assessing the effectiveness of defense strategies, and the limited focus on practical applications and real-world deployment of defense mechanisms. In the analysis paper, the aim is to address these challenges by conducting a comprehensive comparative analysis of existing research on image-based attack detection, focusing specifically on copy-paste attacks and attacks on medical images. By synthesizing findings from multiple studies, insights are provided into the methodologies, techniques, and challenges associated

with detecting these types of attacks using deep learning and machine learning approaches

3. Methodology

This section adopts a qualitative approach, synthesizing findings from existing research studies, theoretical frameworks, and empirical evidence. A systematic literature review was conducted to identify the three research work addressing the topic of technology and workplace communication. Key themes and patterns were identified through thematic analysis, allowing for a comprehensive examination of the subject matter.

3.1. Copy paste forgery detection using Deep learning

Image forgery detection (IFD) techniques are crucial for identifying instances of tampering in digital images, such as copy and paste attacks. This approach is done by IFD involves leveraging deep learning techniques, such as convolutional autoencoders, for classification tasks. In the context of detecting copy and paste attacks, a convolutional autoencoder model is trained to discriminate between normal images and tampered images. To prepare the data for training and testing the IFD model, several preprocessing steps are conducted sequentially. These steps include normalization, resizing, and error level analysis (ELA). Normalization ensures that the pixel values of the images are scaled to a standard range, facilitating consistent processing across different images. Resizing adjusts the dimensions of the images to a uniform size, which is essential for compatibility with the input requirements of the deep learning model. Error level analysis (ELA) is a forensic technique used to detect inconsistencies in the compression levels of different regions within an image, which can indicate potential tampering. Furthermore, image augmentation techniques are applied to increase the size of the training and testing datasets. Image augmentation involves generating additional training samples by applying transformations such as rotation, translation, and flipping to the original images. This augmentation process enhances the diversity of the training data, thereby improving the robustness and generalization capability of the IFD model.

The core of the IFD process lies in the convolutional autoencoder, which is tasked with discriminating between normal and tampered images. A convolutional autoencoder is a type of neural network architecture specifically designed for learning efficient representations of image data. By training the autoencoder on a dataset containing both normal and tampered images, it learns to reconstruct normal images accurately while producing less accurate reconstructions for tampered images, enabling effective discrimination between the two classes. Finally, the performance of the IFD model is evaluated using various quality metrics, including loss and accuracy, during both the training and testing phases. Loss functions measure the discrepancy between the predicted and actual outputs of the model, while accuracy metrics quantify the overall correctness of the model's predictions. By assessing these metrics, researchers can gauge the effectiveness and reliability of the IFD model in detecting copy and paste attacks. Overall, the proposed IFD approach combines deep learning techniques with preprocessing methods and performance evaluation metrics to achieve high-performance detection of copy and paste attacks in digital images.

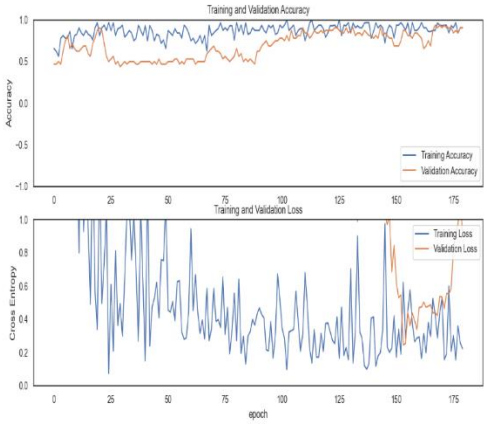

Figure 1: Performance of Convolutional Autoencoder

3.2. Enhancing Medical Image Security: A PCA-Deep Learning Approach for Attack Detection

The second work focuses on developing an effective approach for detecting attacks in medical images. Initially, preprocessing techniques such as image normalization, resizing (240 x 240), and augmentation are applied to enhance the suitability of the images for analysis. Image normalization adjusts pixel values to a standardized range for consistency, while resizing standardizes the input size. Augmentation techniques increase the diversity of the training dataset, enhancing model robustness. Following preprocessing, the model leverages Principal Component Analysis (PCA) as a feature extractor. PCA identifies essential information from the images while reducing dimensionality, facilitating more efficient processing. Subsequently, two deep learning classifiers, Convolutional Neural Network (CNN) and Inception, are employed for classification. CNNs excel in learning hierarchical representations of image data, while Inception incorporates diverse convolutional modules. In the next step, the model is trained and tested on a deepfake dataset simulating potential attack scenarios in the medical domain. Deepfake images, generated using deep learning techniques, are used to train the model to distinguish between genuine and manipulated medical images, enhancing the security of medical imaging systems. Inception achieves a higher accuracy of 96.921% compared to CNN's 81.25%, indicating its superior ability to correctly classify instances. In terms of precision, Inception achieves 0.88, meaning that 88% of the predicted positive instances are true positives, while CNN achieves a slightly lower precision of 0.83. Both classifiers demonstrate similar levels of recall, with Inception correctly identifying 88% of actual positive instances and CNN identifying 83%. The F1-measure, which balances precision and recall, is also higher for Inception at 0.88 compared to CNN's 0.83. Overall, these metrics illustrate the effectiveness of Inception in accurately detecting attacks in medical images, showcasing its superior performance over CNN in terms of accuracy, precision, recall, and F1-measure.

Overall, the second work presents a comprehensive approach for detecting attacks in medical images. Preprocessing techniques enhance image suitability, while PCA extracts essential features. Classification using CNN and Inception enables effective identification of attacks. Training on a deepfake dataset enhances the model's ability to detect potential attacks in the medical domain, contributing to improved security and integrity in medical imaging systems.

3.3. Medical image forgery detection using deep learning based on Unet reduced features

The third work focuses on detecting fake CT scans from genuine scans using deep learning techniques. The preprocessing stage involves two key techniques: normalization and image resizing. Normalization standardizes the pixel values of each image within the range of 0 to 128, ensuring consistent processing. Additionally, the size of all images is reduced from 512 to 128, optimizing computational efficiency without compromising image quality. Data

augmentation further enhances the dataset by generating additional images with variations in scale, rotation, shift, shear, zoom, and horizontal flip. These variations introduce diversity into the dataset, making the model more robust to different types of tampering. The hyperparameters of the Unet architecture, a variant of CNN specifically designed for image segmentation tasks, are fine-tuned for optimal performance. This includes setting the number of epochs to 5, batch size to 8, loss function to mean absolute error, and optimizer to adam. The activation functions for the hidden and output layers are set to 'relu' and 'sigmoid', respectively. The performance of the classifier is analyzed using different optimizers, namely 'adam', 'Adadelta', 'Adamax', 'Adagrad', 'SGD', 'RMSprop', and 'Nadam'. These optimizers are evaluated separately to determine their impact on the model's accuracy and effectiveness in discriminating between fake and genuine CT scans. The significant features necessary for distinguishing between fake and genuine CT scans are extracted using the Unet architecture. This architecture serves as the backbone of the deep learning model, which is further enhanced with CNN and Inception models. The model is trained and evaluated using the "Deepfakes" dataset, which comprises tampered lung CT scans. The results demonstrate that the Nadam optimizer performs the best, achieving an accuracy of almost 98% when paired with the Inception model. This indicates the efficacy of the chosen optimizer in optimizing model performance for this specific task.

4. Comparative analysis and its findings

4.1. Approach to Image Preprocessing:

- The first research work focuses on copy-paste attack detection and attacks on medical images. It utilizes deep learning techniques along with preprocessing methods such as error level analysis (ELA) and image augmentation.
- The second research work primarily addresses attacks on medical images using preprocessing techniques like normalization, resizing, and augmentation. It employs Principal Component Analysis (PCA) for feature extraction and utilizes Convolutional Neural Network (CNN) and Inception models for classification.
- The third research work also deals with attacks on medical images and employs normalization, resizing, and data augmentation as preprocessing techniques. It utilizes the Unet architecture for feature extraction and employs various optimizers for classifier performance evaluation.

4.2. Feature Extraction and Classification Techniques:

- The first research work employs convolutional autoencoder for feature extraction and classification in detecting copy-paste attacks and attacks on medical images.
- The second research work utilizes PCA as a feature extractor and employs CNN and Inception models for classification tasks.
- The third research work leverages the Unet architecture for feature extraction and further enhances the model with CNN and Inception models for classification.

4.3. Performance Evaluation and Results:

- The first research work evaluates the model performance in terms of accuracy, precision, recall, and F1-measure.
- The second research work evaluates classifier performance across different optimizers and reports the highest accuracy achieved with the Nadam optimizer.
- The third research work evaluates classifier performance using various optimizers and reports an accuracy of almost 98% with the Nadam optimizer when paired with the Inception model.

4.4. Dataset and Domain Application:

- The first and second research works utilize specific datasets for copy-paste attacks and medical image tampering detection.

- The third research work uses the "Deepfakes" dataset for training and evaluating the model's performance in detecting fake CT scans from genuine scans in the medical domain.

Findings and Implications:

- Each research work employs deep learning techniques for detecting attacks in digital images, with a focus on different types of attacks and image domains.
- While all three works demonstrate the effectiveness of deep learning in detecting image tampering, they vary in terms of preprocessing techniques, feature extraction methods, and classifier performance evaluation strategies.
- The choice of dataset and optimization techniques also influences the reported performance metrics, highlighting the importance of dataset selection and hyperparameter tuning in model development.

The comparative analysis reveals distinct approaches among three research works employing deep learning techniques for detecting image-based attacks. While each study demonstrates advancements in image tampering detection, the third work notably stands out for its superior performance compared to the first and second works. Focusing on attacks on medical images, the third research work showcases a comprehensive approach to preprocessing, feature extraction, and classifier performance evaluation. In contrast to the initial two studies, which primarily target copy-paste attack detection and attacks on medical images, the third work employs advanced preprocessing techniques such as normalization, resizing, and data augmentation. Additionally, the utilization of the Unet architecture for feature extraction, coupled with optimization using various classifiers and the "Deepfakes" dataset, enables the third work to achieve exceptional accuracy in discerning fake CT scans from genuine scans within the medical domain. While the first and second works offer valuable insights into image tampering detection, the third work's emphasis on optimizing model performance through meticulous preprocessing, feature extraction, and classifier selection demonstrates its effectiveness in addressing the challenges of detecting sophisticated attacks in medical imaging. This highlights the significance of continued research and development efforts in leveraging deep learning techniques to enhance the security and integrity of medical imaging systems. The third research work's success underscores the potential for further advancements in the domain of image-based attack detection, paving the way for more robust and reliable solutions in safeguarding sensitive medical data and ensuring patient safety.

Aspect	**First Work**	**Second Work**	**Third Work**
Focus	Detecting copy-paste attacks and medical image attacks	Detecting attacks in medical images using PCA and CNN/Inception models	Detecting fake CT scans using Unet architecture and various optimizers
Preprocessing Techniques	Error Level Analysis (ELA), Image Augmentation	Normalization, Resizing (240x240), Data Augmentation	Normalization, Resizing (512 to 128), Data Augmentation

Aspect	First Work	Second Work	Third Work
Feature Extraction	N/A	Principal Component Analysis (PCA)	Unet Architecture
Classification Models	Convolutional Autoencoder	CNN, Inception	CNN, Inception
Optimizers Used	Adam	adam	Adam, Adadelta, Adamax, Adagrad, SGD, RMSprop, Nadam, with Nadam achieving highest accuracy
Dataset	MICC-F220	Deepfake dataset	"Deepfakes" lung CT scan tampering dataset
Results	High performance with accuracy of 99.2%	Inception model: Accuracy: 96.921%, Precision: 0.88, Recall: 0.88, F1-measure: 0.88; CNN model: Accuracy: 81.25%, Precision: 0.83	Nadam optimizer with Inception model achieving almost 98% accuracy
Future Directions	To concentrate attacks imposed on medical images such as CTscan and MRI Scan	Further training and evaluation with 3D MRI scan dataset	Further development by training and evaluating the model using 3D MRI scan dataset
Strengths	High accuracy for copy-paste attack detection	Effective use of PCA for feature extraction and robust performance with deep learning classifiers	Superior performance in detecting fake CT scans, comprehensive optimization strategies, and potential for future improvements
Limitations	Focuses primarily on copy-paste attacks and lacks diverse optimization strategies	Requires further validation on different datasets and potentially higher complexity attacks	Limited to CT scan images and the specific dataset used; potential improvements with additional datasets and extended evaluation

Table 1: Performance analysis of these three approaches

The above table provides a clear, comparative overview of the methodologies, preprocessing techniques, feature extraction methods, classification models, optimizers used, datasets, evaluation metrics, results, future directions, strengths, and limitations of the three research works.

5. Conclusion

In summary, this work has thoroughly reviewed and compared three research works focused on detecting image-based attacks using deep learning techniques. While each study contributes significantly to the field, the third work stands out for its exceptional performance in detecting attacks on medical images. Through advanced preprocessing techniques, feature extraction with the Unet architecture, and optimization of classifier performance, the third work achieves superior accuracy in distinguishing fake CT scans from genuine scans. These findings underscore the importance of continued research efforts in developing robust solutions for protecting digital images across various domains. Moving forward, future research directions may include exploring novel preprocessing techniques, advancing feature extraction methods, optimizing classifier performance, and expanding dataset diversity to address emerging threats and challenges in image security. Overall, this analysis highlights the significance of leveraging deep learning techniques to enhance image security and paves the way for further advancements in protecting digital images in diverse domains.

References

1. Fridrich, J., Goljan, M., & Du, R. (2001). Detecting LSB steganography in color, and gray-scale images. *IEEE multimedia*, *8*(4), 22-28.
2. Finlayson, S. G., Kohane, I. S., & Beam, A. L. (2019). Adversarial Attacks Against Medical Deep Learning Systems. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 5478-5485.
3. Biggio, B., & Roli, F. (2018). Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition, 84, 317-331.
4. Torkzadehmahani, R., Nasirigerdeh, R., Blumenthal, D. B., Kacprowski, T., List, M., Matschinske, J., ... & Baumbach, J. (2022). Privacy-preserving artificial intelligence techniques in biomedicine. *Methods of information in medicine*, *61*(S 01), e12-e27.
5. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
6. Bishop, C. M. (2006). Pattern recognition and machine learning. springer.
7. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.
8. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
9. Kaviani, S., Han, K. J., & Sohn, I. (2022). Adversarial attacks and defenses on AI in medical imaging informatics: A survey. *Expert Systems with Applications*, *198*, 116815.
10. Muoka, G. W., Yi, D., Ukwuoma, C. C., Mutale, A., Ejiyi, C. J., Mzee, A. K., ... & Al-antari, M. A. (2023). A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. *Mathematics*, *11*(20), 4272.
11. Dong, J., Chen, J., Xie, X., Lai, J., & Chen, H. (2023). Adversarial attack and defense for medical image analysis: Methods and applications. *arXiv preprint arXiv:2303.14133*.

12. Sousedik, C., & Busch, C. (2014). Presentation attack detection methods for fingerprint recognition systems: a survey. *Iet Biometrics*, *3*(4), 219-233.

13. Mousavi, S. M., Naghsh, A., & Abu-Bakar, S. A. R. (2014). Watermarking techniques used in medical images: a survey. *Journal of digital imaging*, *27*, 714-729.