Combating Medical Image Tampering using Deep Transfer Learning

Ghazanfar Latif [1, a)], Ghassen Bin Brahim [1, b)], Nazeeruddin Mohammad [2, c)], and Jaafar Alghazo [3, d)]

[1] *Computer Science Department, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia.*
[2] *Cybersecurity Center, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia.*
[3] *Artificial Intelligence Research Initiative, University of North Dakota Grand Forks, ND, USA.*

[a)] *Corresponding author: glatif@pmu.edu.sa*
[b)] gbrahim@pmu.edu.sa
[c)] nmohammad@pmu.edu.sa
[d)] jaafar.alghazo@und.edu

Abstract. Tampered Medical images and scans are a very serious issue in the medical field. The results of tampered scans can range from mild to serious results. For example, a tampered scan or medical image can lead to misdiagnosing a patient with a serious condition or a patient who suffers from a serious condition can be misdiagnosed as healthy leading to a delay in receiving proper treatment. In this paper, we propose a deep learning-based methodology to detect tampered/fake cancers in 3D CT scans of human lungs. We use a publicly available dataset of deepfakes that include both tampered and genuine cancers and propose the use of a modified Convolutional Neural Network (CNN), in particular the AlexNet with Transfer Learning. The model was pre-trained on a large dataset and fine-tuned on the medical images dataset to detect tampering. Experimental results achieve a high level of accuracy in detecting tampering. Using the AlexNet with Transfer learning, we achieved an accuracy of 89.47%, Recall of 89.47%, Precision of 89.47%, and F1 measure of 89.47%. This is higher accuracy than similar methods using the same dataset reported in the extant literature. Achieving this high accuracy for a multiclass complex problem is considered an excellent achievement. Expert radiologists are mostly unable to distinguish between real and tampered cancer images. The results demonstrate the potential for deep learning-based models in improving accuracy and efficiency in tampered medical scan detection.

Keywords. Medical Image Forgery, Medical Image Tampering, Deep Learning, Malicious Tampering, Convolutional Neural Networks, CNN.

INTRODUCTION

Medical image tampering and forgery can have serious consequences, since they can lead to inaccurate diagnoses and treatments. With the increasing use of digital medical imaging, it became easier for individuals to alter and modify medical images, either intentionally or unintentionally. Tampering with medical images can involve various techniques such as adding or removing image content, modifying image properties, or manipulating image metadata [1]. One of the main reasons why people would commit medical image forgery is to conceal or alter the presence of diseases, tumors, or other abnormalities in the image. This is usually done to avoid treatment or to obtain insurance coverage, for instance. Additionally, medical image forgery could be used for research misconduct, where researchers manipulate medical images to falsify their results [2]. For these reasons, medical institutions and regulatory bodies need to establish policies and guidelines to ensure the authenticity and integrity of medical images. Additionally,

awareness of the possibility of image tampering among healthcare professionals is becoming very important, and aimed to take the necessary precautions to ensure the accuracy and reliability of the images being used during diagnosis and treatments.

Various techniques have been proposed to address the issue of medical image tampering and forgery. These include digital watermarking, encryption, digital signatures, and artificial intelligence-based techniques [3]. Machine learning and deep learning can also be used for the detection of medical image tampering and forgery by training models to recognize and identify patterns, anomalies, and inconsistencies in the metadata of medical images, which can be an indication of data tampering [4-5]. For example, if the metadata indicates that the image was captured using a certain device, but the image characteristics are inconsistent with that device, the algorithm can flag it for further investigation.

Machine learning and deep learning-based models can also be trained on images to identify unusual shapes, structures, and regions of the image [6-8]. These models have the ability to detect and analyze patterns in images, such as changes in texture, color, or sharpness, to identify areas that are inconsistent with the rest of the image. They can also be used to compare multiple images of the same subject and identify differences between them. Machine learning algorithms can also be trained to identify the source device that captured the image [9]. This can help detect forgeries that were created using non-medical imaging devices. Though these AI-based techniques can be considered valuable tools serving in the detection of medical image tampering and forgery, however, still classification errors may occur, therefore, these techniques need to be used in conjunction with other security measures to ensure the authenticity and integrity of medical images and results.

In this research work, deep learning-based Convolutional Neural Network (CNN) models are used to classify medical images of lungs into four classes: (1) the "True-Benign (TB)" class –which includes images that show no cancerous growths in the lungs, (2) the "True-Malignant (TM)" class –which includes images that show real cancerous growths in the lungs, (3) the "False-Benign (FB)" –which includes images that show cancerous growths that have been removed from the lungs, and (4) the "False-Malignant" class –which includes images that show the presence of fake cancerous growths being added to the lungs. The proposed CNN-based classification model will attempt to accurately analyze and classify images into either of these 4 categories, which will help medical professionals make more informed and faster diagnoses leading to more effective treatment decisions.

The rest of the paper is organized as follows. Section 2 provides an overview of the literature about existing machine learning-based techniques for the detection of tampered images. Section 3 describes the proposed methodology along with a description of the used dataset. Section 4, details the model performance in terms of prediction accuracy, and then a comparative study is presented in relation to the existing work. Finally, in Section 5, conclusions are drawn, and future research directions are suggested.

RECENT STUDIES

In [10], Qadir et al. proposes a Deep learning-based model to detect medical image forgery. The proposed model was designed to uncover image forgery using the “Copy Move Forgery” technique; a simple method of tampering with images in which an object is copied, removed, and replaced in the same image. The model is based on deep learning and is tested on a dataset of 1000 medical images. Results show that the proposed algorithm outperforms existing methods in terms of accuracy and robustness.

Mirsky et al. in [11] describe a framework for automatically injecting and removing medical evidence from 3D medical scans such as those produced from CT and MRI. The proposed framework consists of two conditional GANs (cGAN) which perform “in-painting” (image completion) on 3D imagery. The authors demonstrate using a use case how such an attack can go unnoticed by expert radiologists and state-of-the-art deep learning/AI-based models. In [12], Solaiyappan and Wen investigate the capabilities of machine learning algorithms for image tampering detection. Specifically, the authors experiment and evaluate with eight different machine learning algorithms. These include three conventional machine learning methods (Support Vector Machine, Random Forest, and Decision Tree) and five deep learning models (DenseNet121, DenseNet201, ResNet50, ResNet101, VGG19). These algorithms were used to classify tampered and untampered images. Results show that deep learning with interest region localization achieved a near-accurate performance in detecting the tumor-injected scans.

In [13], Al_Azrak et al. propose a new method for detecting copy-move forgery in images based on the combination of the Discrete Cosine Transform (DCT) and the Discrete Fourier Transform (DFT) with deep learning. The proposed method is tested on a dataset of 1000 images and achieves a high detection rate of 98.5%. Ghai et al. in [14] provide an automated deep learning-based fusion model for detecting and localizing copy-move forgeries (DLFM-CMDFC), which combines models of generative adversarial networks (GANs) and densely connected

networks (DenseNets). Results showed that a high accuracy level in terms of detecting image forgery was achieved with the proposed convolutional neural network-based model.

In [15], Reichman et al. introduce a new large-scale dataset of tampered Computed Tomography (CT) scans generated by different methods: LuNoTim-CT dataset, which is considered as the most comprehensive testbed for comparative studies of data security in health care. The authors further propose a deep learning-based framework, namely ConnectionNet, to detect the tampering of medical images. Experimental results show that the proposed ConnectionNet is effective during the detection of tampered images created by various methods.

In [16], Alheeti et al. propose an intelligent deep learning-based detection method for malicious tamper cancer using deep neural networks (DNN). The experimental results show that the proposed method based on DNN can enhance performance detection of malicious tampering of cancer imagery. Manjunatha and Patil aim to study existing methodologies for detecting passive image tampering using deep learning techniques [17]. The focus of their research is on tamper detection using deep learning techniques. Different tampered image datasets such as MICC, CASIA, and UCID have been used by existing tamper detection methodologies for validating tampering detection accuracies. The proposed method uses deep learning to build an efficient tampering detection mechanism and has demonstrated good performance in terms of TPR, FPR, and F1-Score.

In [18], Khallaf and Alasadi discuss the various strategies developed for image forgery detection. The strategies are divided into two categories: conventional techniques and deep neural networks-based techniques. Performance results show that the proposed deep learning-based method during the detection of image alteration achieves an accuracy of 98.5% which is considered higher than that achieved by prior existing methods.

Thakur and Jindal propose a hybrid deep learning (DL) and machine learning-based approach for passive image forgery detection [19]. The DL algorithm classifies images into 2 classes: forged versus non-forged categories, whereas color illumination localizes forgery. The proposed method was tested on public datasets, and the results were compared to other algorithms. Experimental results show that the proposed method outperforms other algorithms in terms of accuracy.

In [20], Suganya et al. propose a method for detecting copy-move forgery in medical images. The method uses a technique called Golden Ball Optimization (GBO) to identify the key points (KPs) in the image. The features at the identified KPs are evaluated using SURF, and dimensionality reduction is performed using principal component analysis. The features are then clustered using the most valuable player-based optimization to detect the forgery. Experimental results were conducted on a dataset of size 300 medical images. Performance results show that the proposed model exhibits higher precision, specificity, sensitivity, and accuracy compared to existing methods.

PROPOSED METHODOLOGY

The proposed methodology is shown in Figure 1. The tampered and actual scans are input into the model which consists of a convolutional layer with a ReLU activation function. This is followed by a pooling layer and another convolutional layer with ReLU activation and again followed by a pooling layer and dropout layer. This portion of the model is used for feature learning. The output of this phase is input to a flatten layer followed by a dense layer and SoftMax activation function. This portion of the model results in feature reduction. The output of the model classifies the scans into one of four classes; False-Benign (FB), False-Malicious (FM), True-Benign (TB), and True-Malicious (TM).

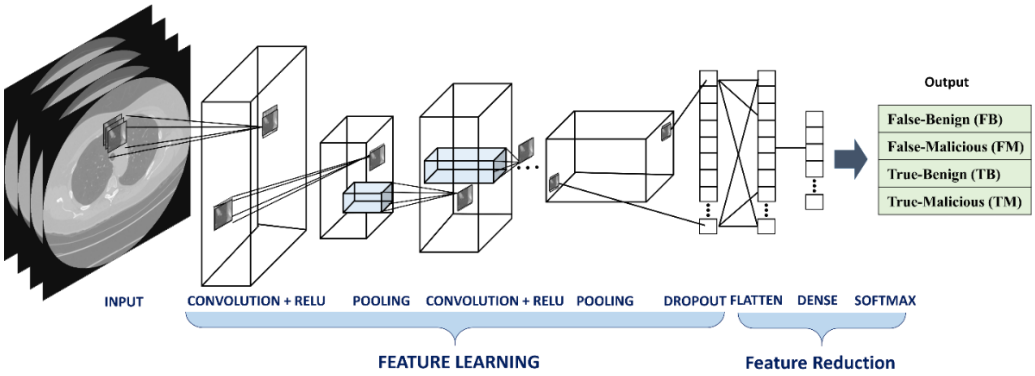
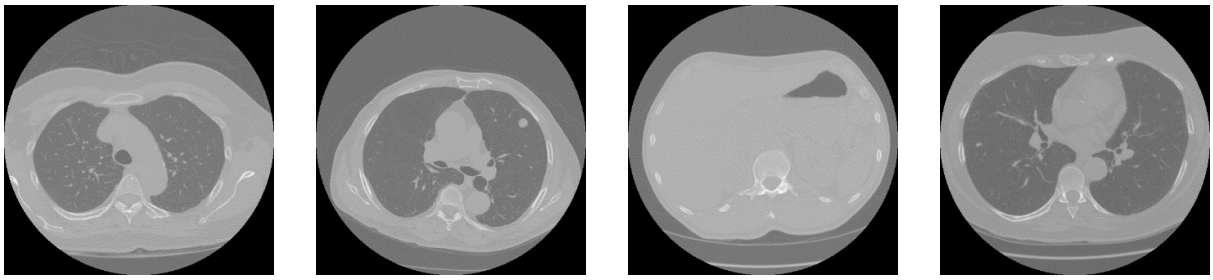


FIGURE 1. Proposed Methodology for Fake Scan Detection.

Experimental Dataset

In the experiments, deepfakes 3D CT scans of human lungs were used, some of which were tampered with by removing real cancer and injecting fake cancer. These scans are publicly available online [21]. The main objective of this dataset is to distinguish between genuine and fake cancers and to detect tampered medical scans. The dataset consists of two sets stored in DICOM format, one with 80 scans and the other with 20 scans. A single scan is a series of 512x512 images, typically consisting of 100-300 slices along the z-axis, where cancers can occupy multiple slices. To assess the level of difficulty in detecting fake images, three expert radiologists evaluated the dataset and were unable to reliably distinguish between real and fake cancers. During the blind trial, the first 80 scans were presented to the radiologists without informing them of the tampering, while in the open trial, the 20 scans were presented after informing them of the truth and asking them to identify the tampered images. The dataset is labeled according to the following four classes: (1) True-Benign (TB) for locations without cancer, (2) True-Malicious (TM) for locations with real cancer, (3) False-Benign (FB) for locations where real cancer was removed, and (4) False-Malicious (FM) for locations where fake cancer was injected. The slice number is also provided along with the labels.



False-Benign (FB) 72 images | False-Malicious (FM) 41 images | True-Benign (TB) 15 images | True-Malicious (TM) 35 images

FIGURE 2. Samples of the four classes from the deepfakes 3D CT scans dataset used for the experiments.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of artificial neural network that uses a mathematical operation called convolution instead of general matrix multiplication in at least one of its layers. They are designed to process pixel data and are used during image recognition and processing. CNNs are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN), based on the shared-weight architecture of the convolution kernels or filters that slide along input features and provide translation invariance [22].

CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers are the first layer of a convolutional network and are responsible for detecting features in the input image. Pooling layers are used to reduce the spatial dimensions of the output from the convolutional layers, which helps to reduce the computational complexity of the network. Fully connected layers are used to classify the input image based on the features detected by the convolutional layers. CNNs have been used in a wide range of applications, including image classification, object detection, face recognition, and natural language processing. They have become a popular choice for many computer vision tasks due to their ability in achieving a high-accuracy performance.

AlexNet Model based Transfer Learning

AlexNet is a convolutional neural network (CNN) architecture that was introduced in 2012 by Krizhevsky et al. [23]. It is a deep neural network that consists of 5 convolutional layers, 3 fully connected layers, and a softmax layer. It has a total of 60 million parameters and 650,000 neurons. On the other hand, Transfer learning is a machine learning technique that involves using a pre-trained model as a starting point for a new task. The idea is to use the knowledge that the pre-trained model has learned from one task to help it learn a new task more quickly and accurately. Transfer learning is particularly useful when the new task has a small amount of data available, as it can help to prevent overfitting.

AlexNet transfer learning based-model is a technique that involves using the AlexNet model as a pre-trained model for a new task. The pre-trained AlexNet model is used as a starting point, and then the model is fine-tuned for the new task. Fine-tuning involves training the model on the new task using a small amount of data, and then gradually increasing the amount of data until the model can perform well on the new task.

There are several advantages of using AlexNet model based on transfer learning. First, it can help to reduce the amount of data needed to train a model. This is particularly useful when the new task has a small amount of data available. Second, it can help to prevent overfitting, which is a common problem in machine learning. Third, it can help improve the accuracy of the model, as the pre-trained model has already learned many features that are useful for the new task.

However, there are also some disadvantages to using the AlexNet Transfer Learning based model. First, the pre-trained model may not be well-suited to the new task, which can lead to poor performance. Second, the pre-trained model may have learned features that are not relevant to the new task, which can also lead to poor performance. Finally, fine-tuning the model can be time-consuming and computationally expensive.

EXPERIMENTAL RESULTS AND DISCUSSIONS

The dataset was divided into 80% training and 20% testing. The 80% was further divided into 20% for validation and the rest for training. We first experimented with various transfer learning models, namely LeNet, GoogleNet, VGG16, and AlexNet as shown in Table 1. The training accuracies were very high, with the lowest training accuracy of 95% for both LeNet and GoogleNet, and the highest training accuracy of 100% was achieved using VGG16 and AlexNet. The validation accuracy was also high, ranging between 84.85% for LeNet and 93.94% for AlexNet. The training loss was lowest at 0.0046 for the AlexNet.

TABLE 1. Experimental Results for different CNN Models based on the Training and Validation data

Transfer Learning Model	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
LeNet	0.2244	95.00%	0.4706	84.85%
GoogleNet	0.2592	95.00%	0.4611	87.88%
VGG16	0.1732	100%	0.9154	75.00%
AlexNet	0.0046	100.00%	0.5424	93.94%

Testing accuracies ranged between 75% (for LeNet) and 89.47% (for AlexNet), as shown in Table 2. It is observed in Table 2 that for all metrics used, namely Accuracy, Precision, Recall, and F1 measure, the highest metrics were achieved when using the AlexNet transfer Learning Model. Using the AlexNet, the F1 score is 85.71%, Recall 85.00% and Precision 88.26%. The high level of accuracy using the AlexNet is very promising and is a step forward toward achieving an optimal solution for fake scan detection. The problem with fake scans is that it is extremely hard to detect even for professional radiologists and thus obtaining an optimal solution using machine learning is a difficult task.

TABLE 2. Experimental Results for different CNN Models based on Test Accuracy, Precision, Recall and F1 Measure.

Transfer Learning Model	Accuracy	Precision	Recall	F1 measure
LeNet	75.00%	80.00%	76.67%	72.73%
GoogleNet	82.46%	83.33%	82.54%	82.46%
VGG16	80.17%	80.31%	79.84%	80.13%
AlexNet	89.47%	85.71%	85.00%	88.26%

Figure 2 shows the plot of training accuracy vs the validation accuracy using the AlexNet transfer learning model. We can observe that as the number of iterations increases the training and validation accuracies increase striving to reach 100%. This suggests that the model is learning the underlying patterns in the training data and is generalizing well to new, unseen data. The model was monitored to ensure that overfitting does not occur. In the actual testing phase, we showed that there was no overfitting. The assessment of the model showed that the training and validation are in line with hyperparameter tuning to achieve the best accuracy.

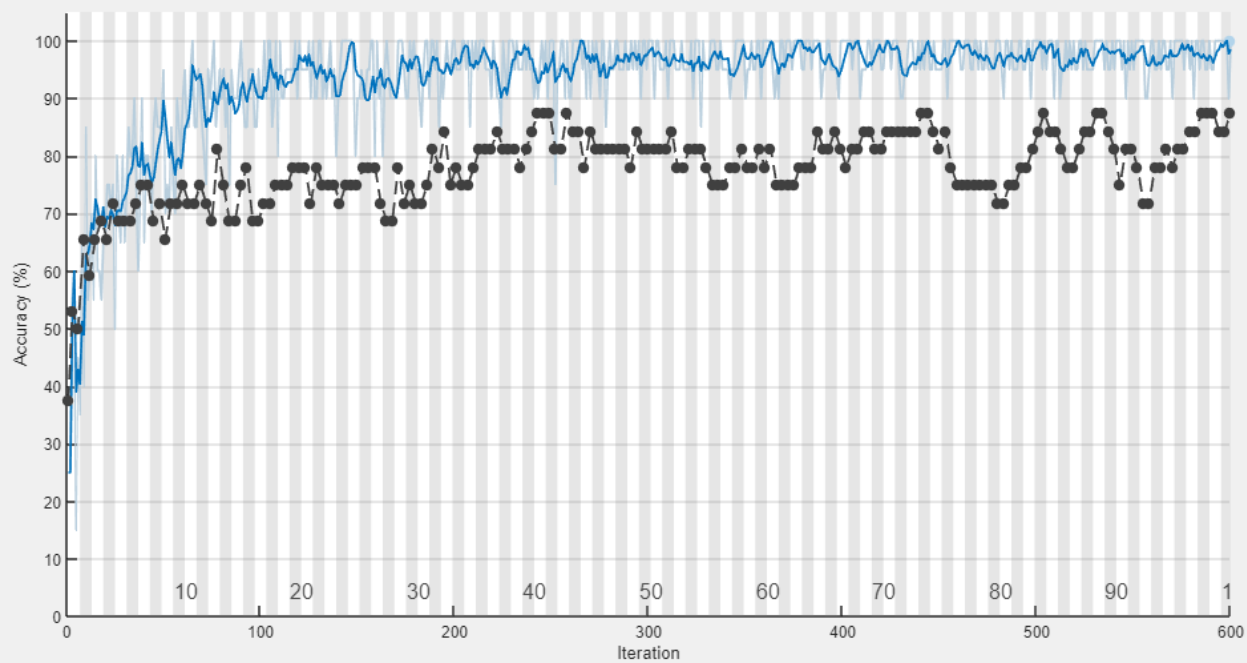


FIGURE 3. Learning Curve-based Comparison of Training Accuracy vs Validation Accuracy

Figure 3 shows the learning curve-based comparison of the training loss vs validation loss. We can observe that as the number of iterations increases the loss approaches zero which confirms that the model is learning to fit the training data very well and is generalizing well to new, unseen data. This is also a good indication that the model did not overfit or underfit.

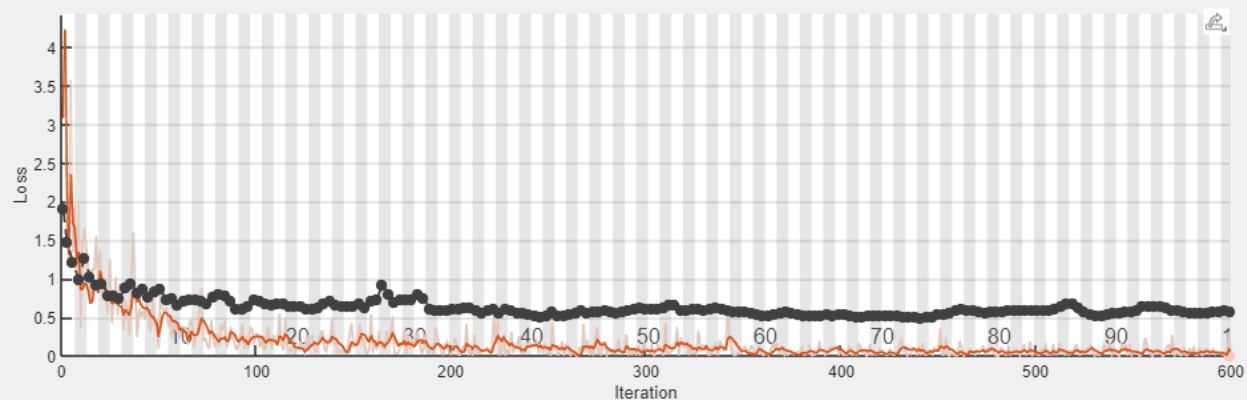


FIGURE 4. Learning Curve-based Comparison of Training Loss vs Validation Loss

CONCLUSION AND FUTURE DIRECTIONS

The use of deepfake technology to create false medical records and manipulate medical images can have serious consequences for patient health and safety, as well as the integrity of medical research. It is important for healthcare professionals, researchers, and policymakers to be aware of these risks and take steps to mitigate them. Additionally, the ethical implications of using deepfake technology in medicine must be carefully considered. The potential for misuse and abuse of this technology highlights the need for ethical guidelines and regulations to be put in place to ensure that it is used for the benefit of patients and society as a whole. In this paper, we propose a deep learning-based methodology to detect fake cancers in 3D CT scans of human lungs. We use a publicly available dataset of deepfakes

that include both fake and genuine cancers and propose the use of a modified Convolutional Neural Network (CNN), in particular the AlexNet with Transfer Learning. The model was pre-trained on a large dataset and fine-tuned on the medical images dataset to detect fake cancers. Experimental results achieve a high level of accuracy in detecting fake cancers. Using the AlexNet with Transfer learning, we achieved an accuracy of 89.47%, Recall of 89.47%, Precision of 89.47%, and F1 measure of 89.47%. This is higher accuracy than similar methods using the same dataset reported in the extant literature. Expert radiologists are mostly unable to distinguish between real and fake cancers. The results demonstrate the potential for deep learning-based models in improving accuracy and efficiency in tampered medical scan detection. Achieving this high accuracy for a multiclass complex problem is considered an excellent achievement.

Future work will include exploring other deep learning algorithms and modifying them to achieve an optimal solution for fake scan detection. Reaching an optimal solution is very important since radiologists find it extremely difficult to detect tampered images. The researchers are also planning to search for other publicly available datasets and combine them to have one complex dataset that can serve as a benchmark for researchers working in the detection of fake scans domain.

ACKNOWLEDGMENTS

This work was supported by the Cybersecurity Research Grant 2022 by Prince Mohammad Bin Fahd University, Saudi Arabia. The authors also would like to acknowledge the support of Prince Mohammad bin Fahd University, KSA, for providing the facilities in the College of Computer Engineering and Science to perform this research work.

REFERENCES

1. Chaitra, B., & Reddy, P. V. (2022). Digital image forgery: taxonomy, techniques, and tools–a comprehensive study. *International Journal of System Assurance Engineering and Management*, 1-16.
2. Kaushik, B., & Kaushik, K. (2023). Forensics in Medical Imaging: Techniques and Tools. In *Unleashing the Art of Digital Forensics* (pp. 165-180). Chapman and Hall/CRC.
3. Sharma, P., Kumar, M., & Sharma, H. (2022). Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation. *Multimedia Tools and Applications*, 1-34.
4. Solaiyappan, S., & Wen, Y. (2022). Machine learning based medical image deepfake detection: A comparative study. *Machine Learning with Applications*, *8*, 100298.
5. Al_Azrak, F. M., Sedik, A., Dessowky, M. I., El Banby, G. M., Khalaf, A. A., Elkorany, A. S., & Abd. El-Samie, F. E. (2020). An efficient method for image forgery detection based on trigonometric transforms and deep learning. *Multimedia Tools and Applications*, *79*, 18221-18243.
6. Latif, G. (2022). DeepTumor: Framework for Brain MR Image Classification, Segmentation and Tumor Detection. *Diagnostics*, *12*(11), 2888.
7. Alghazo, J. M., Latif, G., Alzubaidi, L., & Elhassan, A. (2019). Multi-language handwritten digits recognition based on novel structural features. *Journal of Imaging Science and Technology*, *63*(2), 20502-1.
8. Latif, G., Abdelhamid, S. E., Mallouhy, R. E., Alghazo, J., & Kazimi, Z. A. (2022). Deep Learning Utilization in Agriculture: Detection of Rice Plant Diseases Using an Improved CNN Model. *Plants*, *11*(17), 2230.
9. Freire-Obregon, D., Narducci, F., Barra, S., & Castrillon-Santana, M. (2019). Deep learning for source camera identification on mobile devices. *Pattern Recognition Letters*, *126*, 86-91.
10. Qadir, M., Tehsin, S., & Kausar, S. (2021, April). Detection of Copy Move Forgery in Medical Images Using Deep Learning. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE.
11. Mirsky, Y., Mahler, T., Shelef, I., & Elovici, Y. (2019, August). CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. In *USENIX Security Symposium* (Vol. 2019).
12. Solaiyappan, S., & Wen, Y. (2022). Machine learning based medical image deepfake detection: A comparative study. *Machine Learning with Applications*, *8*, 100298.
13. Al_Azrak, F. M., Sedik, A., Dessowky, M. I., El Banby, G. M., Khalaf, A. A., Elkorany, A. S., & Abd. El-Samie, F. E. (2020). An efficient method for image forgery detection based on trigonometric transforms and deep learning. *Multimedia Tools and Applications*, *79*, 18221-18243.
14. Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People*, (ahead-of-print).

15. Reichman, B., Jing, L., Akin, O., & Tian, Y. (2021). Medical Image Tampering Detection: A New Dataset and Baseline. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I* (pp. 266-277). Springer International Publishing.
16. Alheeti, K. M. A., Alzahrani, A., Khoshnaw, N., & Al-Dosary, D. (2022, March). Intelligent deep detection method for malicious tampering of cancer imagery. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* (pp. 25-28). IEEE.
17. Manjunatha, S., & Patil, M. M. (2021, February). Deep learning-based Technique for Image Tamper Detection. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1278-1285). IEEE.
18. Khallaf, H. D., & Alasadi, A. H. (2022). Detection of Image Tempering: Conventional and Deep Learning-based Techniques. *Journal of Education for Pure Science-University of Thi-Qar*, *12*(2), 162-171.
19. Thakur, A., & Jindal, N. (2020). Hybrid deep learning and machine learning approach for passive image forensic. *IET Image Processing*, *14*(10), 1952-1959.
20. Suganya, D., Thirunadana Sikamani, K., & Sasikala, J. (2022). Copy-move forgery detection of medical images using golden ball optimization. *International Journal of Computers and Applications*, *44*(8), 729-737.
21. UCI Machine Learning Repository: Deepfakes: Medical Image Tamper Detection Data Set (2020) Uci.edu, 2020. Available from: https://archive.ics.uci.edu/ml/datasets/Deepfakes%3A+Medical+Image+Tamper+Detection.
22. Widiastuti, N. I. (2019, November). Convolution neural network for text mining and natural language processing. In *IOP Conference Series: Materials Science and Engineering* (Vol. 662, No. 5, p. 052010). IOP Publishing.
23. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2019). ImageNet classification with deep convolutional neural networks. 2012: 1097–1105. *Last accessed Oct*, *1*.