Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.040257
Article

Deep Fakes in Healthcare: How Deep Learning Can Help to Detect Forgeries

Alaa Alsaheel, Reem Alhassoun, Reema Alrashed, Noura Almatrafi, Noura Almallouhi and Saleh Albahli*

Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
*Corresponding Author: Saleh Albahli. Email: salbahli@qu.edu.sa
Received: 11 March 2023; Accepted: 19 June 2023; Published: 30 August 2023

Abstract: With the increasing use of deep learning technology, there is a growing concern over creating deep fake images and videos that can potentially be used for fraud. In healthcare, manipulating medical images could lead to misdiagnosis and potentially life-threatening consequences. Therefore, the primary purpose of this study is to explore the use of deep learning algorithms to detect deep fake images by solving the problem of recognizing the handling of samples of cancer and other diseases. Therefore, this research proposes a framework that leverages state-of-the-art deep convolutional neural networks (CNN) and a large dataset of authentic and deep fake medical images to train a model capable of distinguishing between authentic and fake medical images. Specifically, the paper trained six CNN models, namely, ResNet101, ResNet50, DensNet121, DenseNet201, MobileNetV2, and MobileNet. These models had trained using 2000 samples over three classes: Untampered, False-Benign, and False-Malicious, and compared against several state-of-the-art deep fake detection models. The proposed model enhanced ResNet101 by adding more layers, achieving a training accuracy of 99%. The findings of this study show near-perfect accuracy in detecting instances of tumor injections and removals.

Keywords: Deep learning; image processing; medical imaging; artificial intelligence

1 Introduction

The medical field faces a severe threat with the advent of deep fakes because deep learning creates fakes that look so real that specialists mistake them for original images. Medical imaging is a technique of visually representing the body's internal functions and tissues to reveal them, serving as a guide to diagnosis and treatment [1]. There are several medical imaging techniques available for better treatment, such as X-rays, CT (Computed Tomography) scans, MRI (Magnetic Resonance Imaging), and Positron-Emission Tomography (PET) scans [2]. PET scans provide unique information about the scanned area applied to different phases of medicine and treatment.

The advent of medical imaging and the recent digitalization of the healthcare industry brought many positive changes. Moreover, it opened the doors to solving problems related to privacy and authenticity. On the other hand, a hacker could gain access to sensitive and confidential data or perform worst actions, even tampering with the data, which could mislead the doctors and, in extreme cases, could lead to death. Intruders try to upload deep fakes into databases; deep fakes are synthetic media in which existing images or videos are replaced with other people's images [3]. Deep fakes threaten the credibility of information because they leverage the technologies of machine learning and deep learning to create media based on knowledge gathered from authentic media, which might be an advantage in other fields but a huge disadvantage in the medical field.

Every new technology was initially intended to bring good to the world, but eventually, its disadvantages surface, and deep fakes are no exception. The weaponization of deep fakes can have a massive impact on every aspect of life, including the economy and national security. It can inflict harm on individuals and democracy all over the world. Deep fakes will further erode declining trust in the media [4]. Deep fakes pose an especially significant threat in the medical field because it deals with human lives, and any mistake or error could lead to a chain of terrible events.

The authenticity issue must be addressed, as record tampering can harm hospitals and patients. Hence, a system that automatically identifies and detects these deep fakes is required now more than ever, and that is where the paper leverages machine learning and deep learning. Since deep fakes are a product of deep learning, it seems only reasonable to use deep learning and machine learning to combat them. However, as with all deep learning models, performance is constantly evolving as more training data is collected to approach perfection in the model; this applies both to the algorithms that produce deep fakes and to those that detect them, the models that try to detect these deep fakes have to be on high alert and undergo constant retraining to keep up with the latest developments.

The contributions of related research works are discussed as follows:

- Provide a comprehensive study of the latest state-of-the-art techniques in detecting deep fake samples.
- Experiment with analyzing several machine learning and deep learning algorithms to compare their performance and optimize the results.
- Study a CNN approach capable of computing a reliable collection of deep fake samples to improve the effectiveness of medical deep fake detection.
- Choose an appropriate dataset that s quality data and helps optimize the performance and efficiency of the trained models.

The motivation of the proposed research model is to get the optimum solution for detecting deep fakes in healthcare. It also aims to identify how deep fakes prevent the best healthcare solutions. It will stress the need for time to identify and propose the best solution.

The rest of the research manuscript is organized as follows. Section 2 discusses the state-of-the-art approaches to deep fakes and their detection in healthcare. Section 3 outlines the proposed experimental setup and the results obtained from the proposed research model. Finally, Section 4 presents the conclusion and suggests future work based on the proposed research model.

2 Related Work

In this section, works are discussed that have been conducted on detecting and handling deep fakes in medical imaging using artificial intelligence, specifically machine learning and deep learning approaches. State-of-the-art approaches are discussed for handling deep fakes in healthcare. However,

before going into these works, it would be useful to introduce the world of artificial intelligence as it relates to deep fakes. A few significant points are discussed that have a great role in identifying deep fakes in healthcare.

2.1 Machine Learning

Machine learning (ML) is a component of artificial intelligence that allows computers to acquire knowledge and use this knowledge to make decisions without being explicitly programmed. ML does this by using algorithms to train systems using datasets relevant to the task. As with every field, ML has sub-divisions, such as supervised learning, where the model is trained with samples where the correct answers are already known. An example of supervised learning can be seen in Fig. 1, where various classes have been separated into categories.

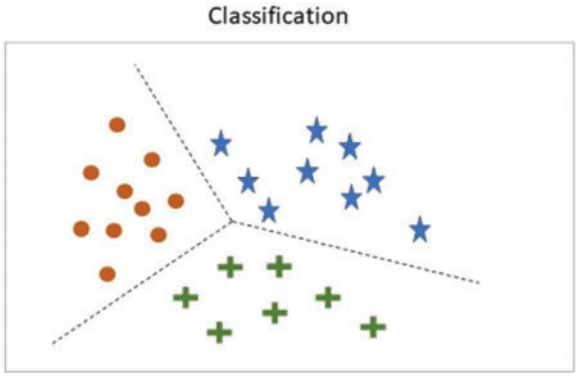


Figure 1: Classification using supervised learning

2.2 Deep Learning

Deep learning (DL) is a class of machine learning algorithms that employ layers to extract progressively higher-level features from raw input [5]. It uses mechanisms that are similar to those operating in the human brain. Deep learning works by passing input media (usually pictures and videos) through layers that extract features in the media. In this way, the system makes decisions after training.

2.3 Classification

Classification is a supervised machine learning method where the model tries to predict and classify input data and give it a correct label. It does this by classifying the output into two categories: in this case, real or fake. This research aimed to enhance classification performance by using deep fakes. Thus, the paper looks at several classification models and sees which produces the most accurate results.

2.4 Deep Fakes

The development of generative deep learning algorithms has reached a stage where it is challenging to distinguish between real and fake content. Deep fakes are digital forgeries that use advanced deep learning methods to create or modify audio or visual content to mislead the audience. It involves training generative neural network architectures or generative adversarial networks. Deep Fakes' fundamental component is machine learning, which enables them to be generated more quickly and

cost-effectively. The Deep Fakes pose the most significant problem because machine learning models constantly evolve and upgrade, so keeping up with the trend is necessary. The invention of generative adversarial networks (GANs) and autoencoders was a giant step toward creating undetected deep fakes. A GAN is unique because it combines two neural networks that have each been thoroughly trained in deep learning recognition. The first, the generator, is tasked with producing fake images. The second, the discriminator, is tasked with determining if this media is fake or real. The components of an autoencoder are an encoder, which reduces a picture to a latent space with fewer dimensions, and a decoder, which reconstructs the original image from the latent representation. Deep fakes use this architecture by encoding images with a universal encoder.

2.4.1 How Far Deep Fakes Go

In the last few years, creating fake content like images and videos has become more common, utilizing artificial intelligence (AI) digital manipulation techniques. Deep fake technology came to light in November 2017 when an anonymous user on the social media platform Reddit posted an algorithm that took advantage of existing artificial intelligence algorithms to create realistic fake videos [6]. One method of faking involves swapping someone's face with that of a target person in a photo or video and creating content to mislead people into believing that the target person has said words someone else has said.

More constructively, deep fake technology is also used in healthcare. One potential route for generating deep fakes in medical scans is associated with injecting and removing tumors, as shown in Fig. 2. Most medical images are grayscale and low resolution, making it more difficult for people and algorithms to detect manipulation. This is in contrast to images of nature scenes, which often have high-resolution color and texture details. If a medical deep fake goes undetected, it might cost the hospitals a lot of money or someone's life.

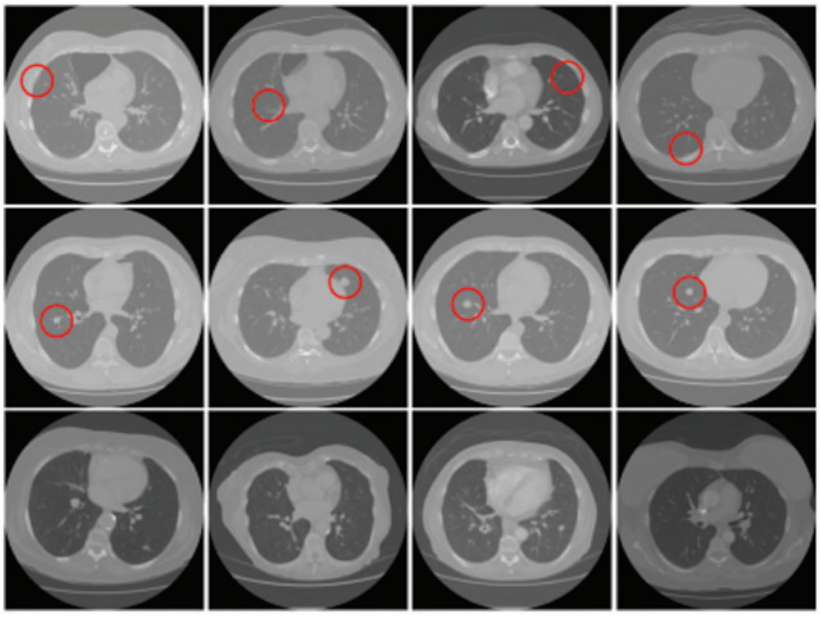


Figure 2: Tampered raw CT-scan images. Row 1: Removed samples, Row 2: Injected samples, Row 3: Untampered samples

2.4.2 Deep Fake Detection

Several techniques have been presented to detect manipulation in medical images. The two categories of these methods are active detection and passive detection, as shown in Fig. 3. Active detection methods, like digital watermarking and signatures, require an authentication code to be embedded using specialized hardware or software before an image is distributed. Watermarking techniques are active detection methods that demand embedded information to authenticate the image. But the two biggest problems with active detection are (i) adding more information after the image was taken and (ii) the effect of the watermark on the quality of the image [7].

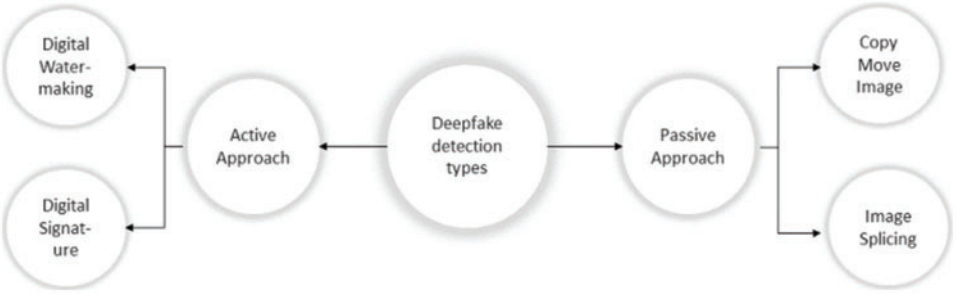


Figure 3: Deep fake detection types

On the other hand, passive detection methods, like copy-move and image splicing, are done by comparing the frequency domain properties or statistical data of the image to identify changes in local features and the entire image. The copy-move and image-splicing techniques are frequently used. To hide the area of interest from the observer, copy-move involves duplicating an uninteresting area over the target area. The target area can also be duplicated using this method, and the frequency of interest regions can be increased. Image splicing differs from copy-move in that the duplicated region of interest for image splicing comes from an external image. The following are the key benefits of passive detection techniques: (i) No earlier data is needed to validate the image; (ii) it prevents visual damage to the image resulting from the watermark information being embedded in the picture.

2.5 Review of Related Work

Having understood deep fakes and how they affect medicine, the research looks into recent works which were targeted at studying these deep fakes and using artificial intelligence techniques to detect and handle them.

The work done by Alsirhani et al. [8] explains the challenges of Generative Adversarial Networks (GAN) and their two main medical imaging applications. The first application was centered around generative elements images. In contrast, the second application was centered around the discriminant component; the discriminator D can be used as a detector when abnormal or fake images are presented. This study was applied to a wide range of applications using GAN to work on reconstruction, image synthesis, segmentation, classification, detection, and registration, to detect abnormal images.

Ragab et al. [9] discovered that applying CT-GAN to creating 3D medical images offered more convincing results than 2D scans, proving how realistic the transformations are and how deceiving they can be. This means it is quite easy to insert malicious data points and pollute an authentic dataset; in this experiment, malignant samples were added, and benign samples were removed from a dataset of 3D CT scans of lung cancer. Then, this was implemented with D, using a conditional GAN (cGAN) to perform in-painting (image completion); cGAN is a GAN with a generator and discriminator but

is conditioned on additional input with further information to generate and discriminate images more effectively for injection. The generator will always complete the images. Their work confirms how vulnerable databases and hospitals are to such attacks, with an accuracy of 61% for detecting an injection and 39% for detecting a removal, and the accuracy for cancer removal was 90% from 95.8%. The detection of cancer injection was 70% from 99.2%.

Ragab et al. [10] studied previous work on detecting GAN-generated images and discovered there was no effective solution, so they experimented with two methods, the principal component analysis (PCA) and support vector machine (SVM), to classify GAN-generated images using a two-stage cascade framework which works on as little as 1% of the original image to detect forgeries. The classification baffled doctors, as they could not differentiate between real and fake images, proving how important this research is. The final proposed model was able to classify Cycle Generative Adversarial Network (CycleGAN) tampered medical images and real images with an accuracy of 99.8%.

Ghadi et al. [11] employed various Convolutional Neural Networks (CNN) to compare their ability to detect GAN-generated deep fake images. Since deep fakes employ machine learning to generate excellent fakes, differentiating between synthesized and real images becomes an even more tedious task. This study explored eight CNN-based architectures, including DenseNet169, DenseNet121, DenseNet201, VGG16, VGG19, VGGFace, ResNet50, and a customized model, to classify deep fake images and evaluated them using five metrics: F1-score, an area under the ROC (Receiver Operating Characteristic) curve, recall, accuracy, and precision.

The paper by Siddharth et al. [12] proposed a medical image deep fake detection system based on three machine learning methods and five deep learning models to identify and differentiate tampered and untampered images. The CT-GAN dataset was used to build learning techniques used in the experiment. The results of the study showed that deep learning with region-of-interest localization would classify tumor injection scans more effectively. The DenseNet model got the best accuracy score of 80% for multiclass delocalized medical deep fake images.

Suk et al. [13] addressed the issues that could result from data manipulation and image regeneration in the medical field. The study dataset used 4 images of lesions that were included for fundus data manipulation (normal, diabetic retinopathy, glaucoma, and macular degeneration). All this was based on Sparse CNN to fuel the manipulation detection system using U-Net and Cycle GAN and ended up with a detection ability of 91%.

The paper by Reichman et al. [14] proposed a deep-learning-based framework, ConnectionNet, which automatically detects tampered images. The proposed ConnectionNet works on small tampered regions in the images, yields promising results, and can serve as a reference for future research into medical imaging. LuNoTim CT is a fresh dataset that was used. It contains a sizable number of Computed Tomography (CT) scans tampered with by three methods: copy-move forgery, classical inpainting, and deep inpainting. The suggested framework had a deep fake detection accuracy score of 85%.

Gite et al. [15] focused their study on Tuberculosis (TB) caused by the Mycobacterium, which affects the lungs. The major method doctors use to diagnose TB is from images produced by X-rays. The work used deep learning algorithms by comparing U-Net, fully convolution network (FCN), semantic segmentation model (SegNet), and U-Net++ to see which performed best. After analysis and comparing techniques for lung segmentation, the models were evaluated, and U-Net++ got better accuracy than U-Net. At the end of the study, U-Net++ achieved more than 98% accuracy, U-Net, SegNet, and FCN, 95%, 84%, and 78%.

Riza et al. [16] used a novel deep fake predictor (DFP) based on a hybrid of VGG16 and a convolutional neural network. A dataset comprised of both deep fake image samples and real image samples was used for the training. The proposed DFP approach achieved 95% precision and 94% accuracy for deep fake detection, and after comparing it with other state-of-the-art models, the DFP proved to be the better model.

Dustin et al. [17] proposed photographic and video deep fakes. This research study discussed plastic surgery, where imaging technology plays an important role. Emerging technologies were highlighted in the research study, finding that the importance of medical imaging is such that deep fakes have a major impact.

From all the works discussed in this section, the importance of eradicating deep fakes in medicine is evident; Table 1 summarizes all the works discussed.

Table 1: A comparison of the related work diagnosis

Reference	Methodology	Findings	Gaps identified
Alsirhani et al. [8]	Research other works using GAN	–	No gap since no work was done, just a review of previous works using machine learning to detect deep fakes
Ragab et al. [9]	CT-GAN	Cancer removal accuracy went from 90% to 95.8%. Cancer injection accuracy went from 70% to 99.2%	Mistakes by the radiologist would completely sabotage the work done by the CT-GAN since identifying deep fakes is already difficult as it is.
Ragab et al. [10]	PCA, SVM	93.5% accuracy in detecting CT slices and a better result for CT scans compared to other models	There is no correlation with sub-images
Ghadi et al. [11]	DenseNet169 DenseNet121 DenseNet201 VGG16 VGG19 VGGFace ResNet50 A customized model,	ResNet50 had the best accuracy of 97%	The customized model could be improved by fine-tuning its hyperparameters and training on more real data
Siddharth et al. [12]	Support Vector Machine Random Forest Decision Tree DenseNet121 DenseNet201 ResNet50 ResNet101 VGG19	DenseNet121had the best accuracy of 80.4%	Performed better when there was the localization of the areas of interest, not the entire real-time images

(Continued)

Table 1 (continued)

Reference	Methodology	Findings	Gaps identified
Suk et al. [13]	U-Net Cycle GAN	91% Accuracy	The model does not account for doctors' expertise
Reichman et al. [14]	ConnectionNet	Accuracy of 85%	The experiment is only limited to CT scan images.
Gite et al. [15]	U-Net FCN SegNet, U-Net++	U-Net++ had the best accuracy of 98%	The experiment focused only on tuberculosis, so it cannot be applied to any other illness
Riza et al. [16]	Deep Fake Predictor (DFP)	95% precision 94% accuracy	The system was used on a broad selection of image samples and might not be perfect for medical images

3 Method

The methodology of the framework follows all traditional deep learning, where data is collected and then passed through a preprocessing machine. In this case, the layers extract the features from the image samples. The logic takes place, the classification is done, and the output is the classified sample. Fig. 4 shows the methodology framework, designed in a simple pattern.

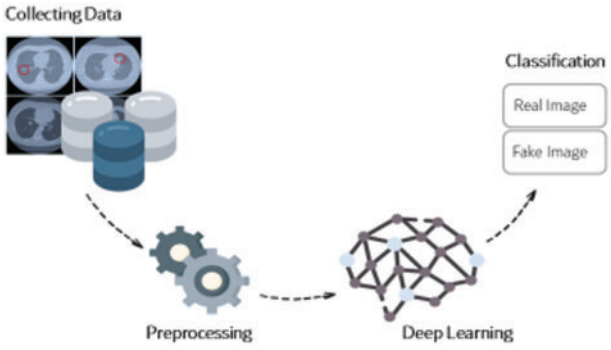


Figure 4: The methodology framework

This work was a state-of-the-art framework for detecting deep fakes in medical images using real image samples to optimize the classification process. After the dataset was collected, the preprocessing phase was applied: the dataset was cleaned, augmented, and balanced to ensure it was properly tailored for the task ahead. Using tampered lung CT scans, high-level deep learning models were applied to detect deep fake samples: ResNet, DenseNet, and other Convolutional Neural Networks (CNN) models. A general framework of the system is shown in Fig. 5.

The deep learning models used in this research study are discussed below, with all parameters used in the experiments. All the deep learning models presented below were used in the research experiments.

3.1 ResNet101

ResNet is short for Residual Networks, a classic neural network used as a backbone for many computer vision tasks. The ResNet family includes the ResNet101 model, which is 101 layers deep [18]. According to previous results, Resnet101 was one of the most promising candidate models for this type of research. This model has 44.6 million parameters, a size of 167 MB, and an image input size of 224 by 224. Fig. 6 shows the ResNet101 architecture visually.

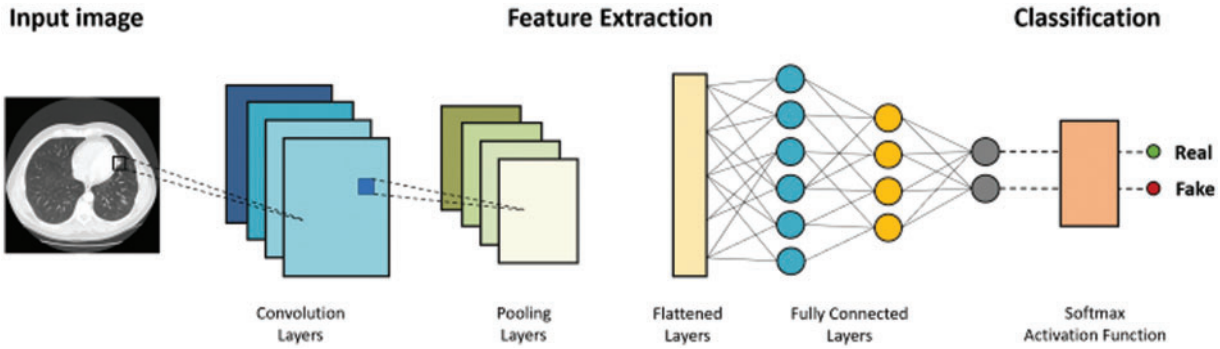


Figure 5: Proposed framework for proposed research model

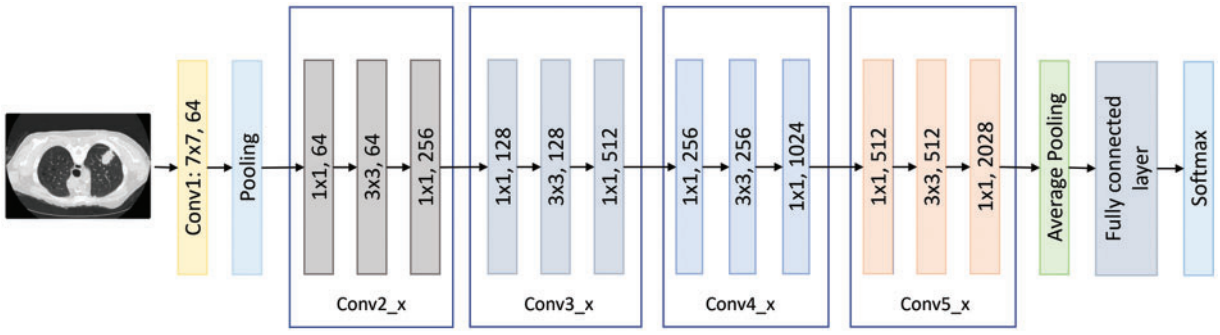


Figure 6: The ResNet101 flow chart

3.2 ResNet50

This is a convolutional neural network of 50 layers that form networks by stacking residual blocks. This model has learned rich feature representations for a wide range of images and performed well for various purposes, particularly classification problems [18]. The 50-layer ResNet employs a bottleneck design as its main building block. A residual bottleneck block uses 1×1 convolutions, known as a "bottleneck," which reduces the number of parameters and matrix multiplications. This enables much faster training of each layer. It uses a stack of three layers rather than two layers [19]. ResNet50 has 25.6 million parameters, a size of 96 MB, and an image input size of 224 by 224. Fig. 7 shows the ResNet50 architecture visually.

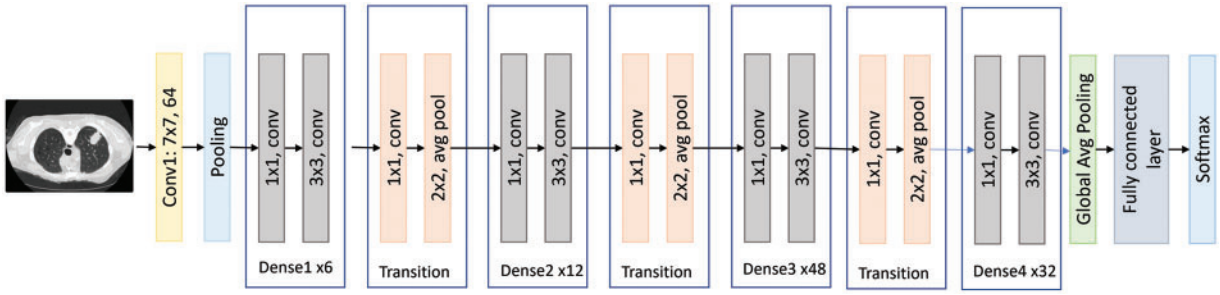


Figure 7: The DenseNet201 flow

3.3 MobileNet

MobileNet is a class of Convolutional Neural Networks. It is the first mobile computer vision model for TensorFlow, which was open-sourced by Google and used separable depth-wise convolutions. It dramatically reduces the number of parameters to create lightweight deep neural networks. MobileNets are small, low-latency, low-power models parameterized to meet the resource constraints of various use cases. They can be built upon for classification, detection, embedding, and segmentation [20]. The paper now has an excellent starting point for fast training for classification, detection, and other common tasks. Fig. 8 shows the MobileNet architecture visually.

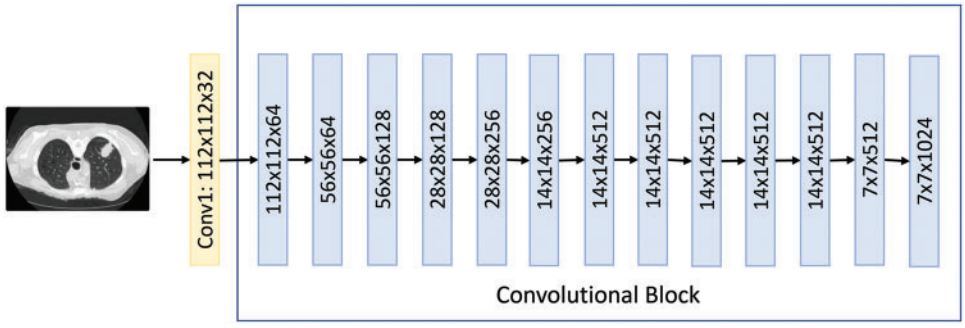


Figure 8: The MobileNet flow

3.4 MobileNetV2

MobileNetV2 is the same as the original MobileNet, except that MobileNetV2 employs inverted residual blocks with bottlenecking features. The MobileNetV2 architecture was based on an inverted residual structure where the input and output of the residual block were thin bottleneck layers. Unlike traditional residual models, which used expanded representations in the input, MobileNetV2 uses lightweight depth-wise convolutions to filter features in the intermediate expansion layer [21]. Also, it has a lower number of parameters compared to MobileNet. Fig. 9 shows the MobileNetV2 architecture visually.

The models discussed above were comprised of layers that carry out the training tasks. These layers are generally discussed below:

- Convolutional layers are the main building block of the neural network. They transform the input media by applying a filter for features to be extracted by forming feature maps. This is usually the first layer of every neural network because that's when the work begins.

- Pooling layer: Is used to reduce the dimension of the feature maps, thereby reducing the number of parameters needed for training and the computational time and power needed. It summarizes the features from the convolutional layer to make them effective and usable.
- Convolution block: This is a combination of the convolutional layer and the pooling layer and is used by some models as an essential feature extraction component. It usually consists of one or more convolutional layers, followed by one or more pooling layers, which are used to reduce the dimensions of the feature map while retaining its integrity.
- Dense layer: This is a deeply connected neural network layer whose neurons are connected to the neurons of the preceding layer. It receives its input from the pooling block and begins the classification process based on the extracted features by feeding the inputs from the pooling layer into its neurons.

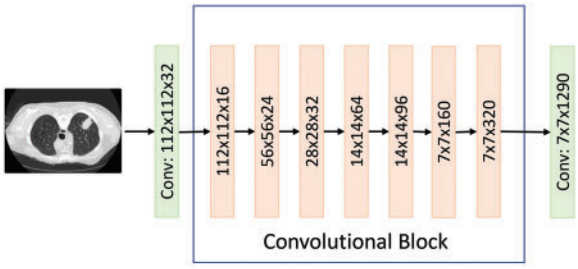


Figure 9: The MobileNet2 flow

4 Experimental Setup and Results

This section discusses the dataset used for the experiments, then explains the procedure for using the designed models to automatically detect deep fake medical images and the outcome of the evaluation of the designed models.

4.1 Evaluation Metrics

To evaluate detection performance on lung CT-Scan-based deep fakes of the research model, the accuracies of the models had to be tested. Accuracy was an evaluation metric used for machine learning and deep learning models. Accuracy defines the measure of the relationship between the total number of correct predictions or classifications made and the total number of predictions or classifications made. The formula for accuracy is shown in Eq. (1).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

4.2 Dataset

The research study attempted to detect real cancer samples from deep fake images to improve the classification performance. The type of data required for the project was a dataset consisting of tampered and untampered medical deep fake images. To obtain sufficient data, the proposed research study chose this dataset [22], which includes deep fakes in CT scans of human lungs that were tampered with to remove real cancer and inject fake cancer. A preview of the dataset can be seen in Fig. 10.

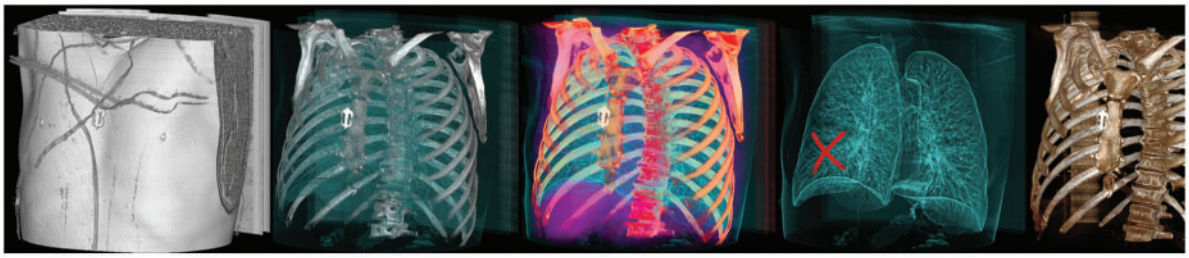


Figure 10: Sample images from the dataset

4.2.1 Dataset Description

Dataset preparation requires a dataset to be transferred initially. In the current research study, the dataset was uploaded into Google Drive, and then the drive was mounted to access the model. After gaining access to the dataset, the researchers created a new class called Untampered (UT) that combined two types of scans: True-Benign (TB) and True-Malicious (TM). The tampered scans were in different classes: False-Benign (FB) and False-Malicious (FM), as shown in Table 2:

Table 2: Classes in the research dataset

Class	Acronym	Description	Train data sample	Test data sample
Untampered	UT	Real scans, no cancer injected or removed	23	20
False-Benign	FB	Scans include a region where real cancer has been removed	49	17
False-Malicious	FM	Scans include a region where fake cancer has been injected	24	18

The distribution of the dataset classes can be seen graphically in Fig. 11 with the False-Benign class having the largest representation in the dataset.

4.2.2 Dataset Preprocessing

The proposed research study maximized the dataset size to 2000, with 80% for training data and 20% for testing data, with some techniques used to avoid over-fitting that occurred in the initial experiment. Fig. 11 shows the complete distribution of the dataset with their classes.

Data Augmentation

Due to unbalanced training data, the count of FM and UT is less than half the FB class. FM and UT images may need to be augmented to balance the amount of data with the FB class. In FM and UT classes, the following four data augmentation strategies were used: [23]

- Flip over the x-axis, y-axis, and both axes.
- Combinations of x, y shifts; 4 units in a specified direction.

- Rotation of 360 degrees in 30 or 45-degree increments.
- Shear images.

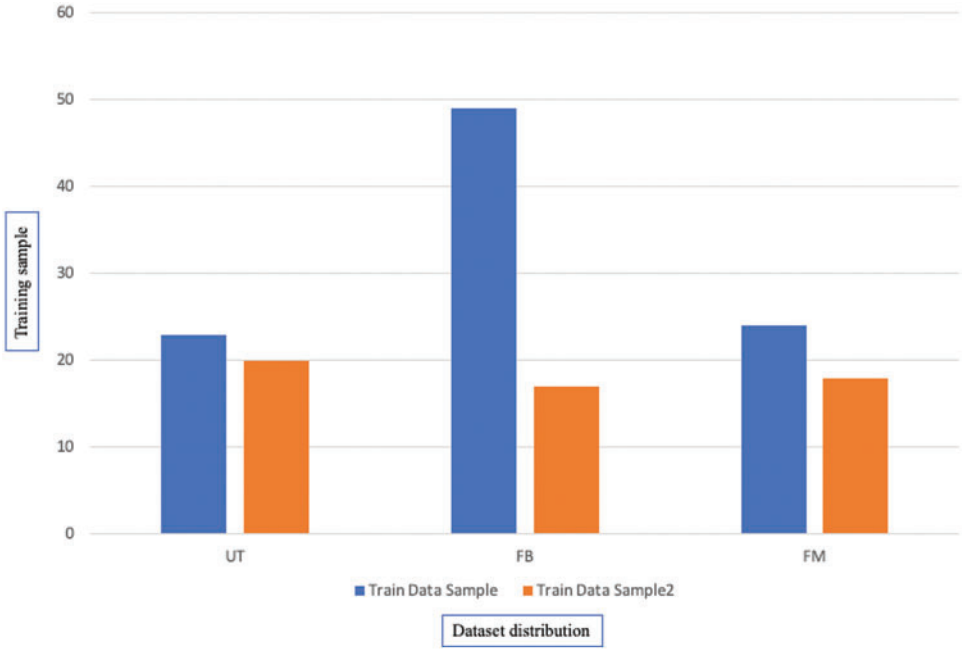


Figure 11: The distribution of the dataset classes

The generalizability of a model may be improved, and over-fitting can be minimized by creating new, artificially-augmented datasets on which it can be trained. Data augmentation is a common technique used to solve this issue, such as for the FM and UT classes. Fig. 12 shows the result of this augmentation step.

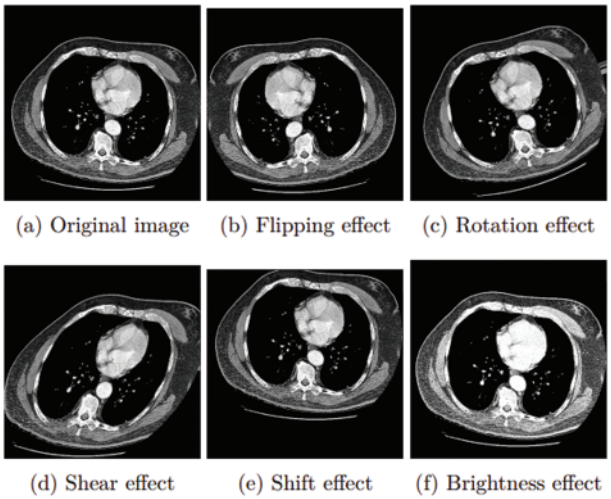
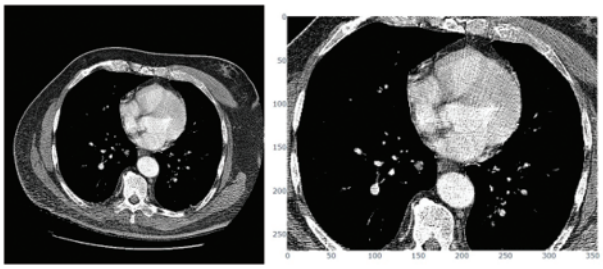


Figure 12: Augmentation approaches

Localization of a Region Of Interest (ROI)

The research study attempted to improve the model's accuracy. The researchers worked on the localization of a region of interest (ROI) [24]. In medical images, localization of a region of interest (ROI) is the process of locating a particular part. Finding the tumor's position within the scanned images is the most important part of the dataset. Deep learning with localization of the region of interest has been found in several studies to be the most effective method for classifying tumor injection and removal. Fig. 13a represents a sample of the dataset [22] with negative space. In this case, the research focuses on the lung, as it is the region of interest. The proposed research model can extract it by using the OpenCV library to read images from a height of 90–120 pixels and a width of 446–389 pixels, as shown in Fig. 13.



(a) Original Image From The Dataset (b) Localization Of A Region Of Interest

Figure 13: ROI approach

However, because the ROI in this section was the entire lung, the proposed research study could not achieve optimized performance from this approach. Tumor site localization was necessary to obtain more satisfying results, but the challenge was to differentiate between naturally occurring and artificially generated tumors.

Data Distribution and Balancing

The proposed research study encountered the issue that the number of FB samples is nearly twice that of FM and UT samples, indicating that the training and test data were extremely unbalanced, causing an overfit and reducing model performance, for which the three classes must be balanced. Balancing was an important technique because a model trained on unbalanced data may be unable to forecast the results for the minority class accurately. To balance all three classes, some balancing techniques may be used to solve this issue. One of these is an oversampling technique used to develop extra synthetic samples for the minority class. The research study used this technique to balance the dataset. After augmentation and balancing of the data, 2000 samples were generated, 1600 for the training and 400 for the testing set, and distributed [25], as seen in Table 3.

Table 3: The final distribution of the dataset

	Train data sample	Test data sample
FB	533	133
FM	532	134
UT	535	133

The distribution can be seen graphically in Fig. 14 showing a fair distribution that was perfect for training models and getting effective results, where the x-axis is the sample points.

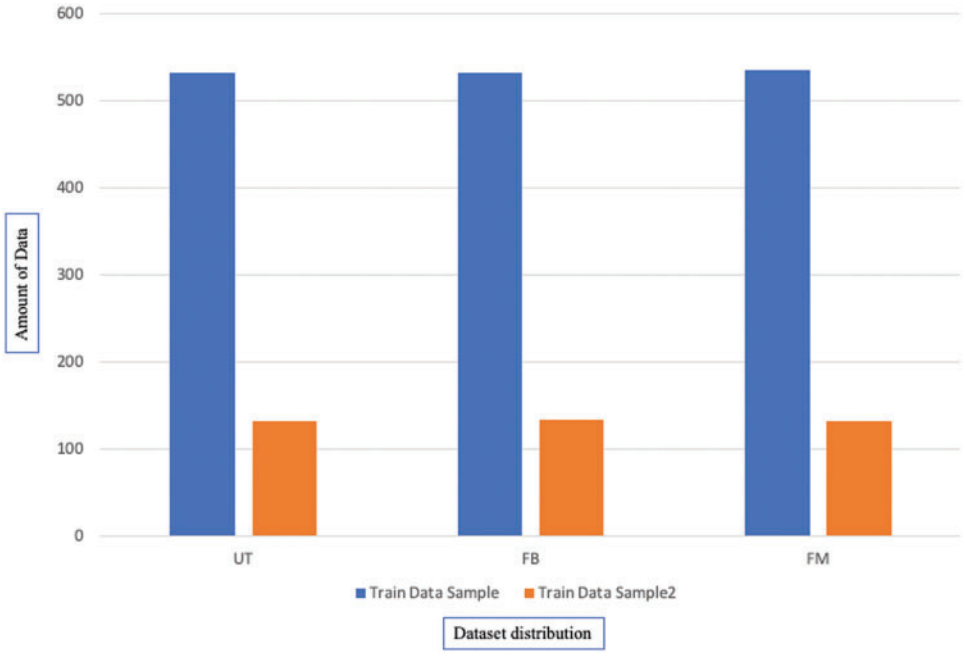


Figure 14: Distribution of the dataset classes after processing

4.3 Implementation

To achieve the paper's aim of improving classification performance, the research study used the model architecture as a base and froze all pre-trained convolution layers. This GAP led to adding Normalization and Regularization layers, a Fully-Connected Dense layer with rectified linear activation (ReLU), and a Softmax activated 3-neuron layer. It was, therefore, necessary to build ResNet101 with a 128-neuron Fully-Connected Dense Layer, ResNet50 with 32-neuron layers, DenseNet121 with 80-neuron layers, DenseNet201 with 1024 neurons, MobileNet with 256 neurons, and MobileNetV2 with 128 neurons [26].

4.4 Experimental Results

Experimental results were computed after the successful completion of the experiments. Experimental results were used to decide whether the proposed model was better than those which had been used before.

4.4.1 Results Before Augmentation

The research study initially trained the six models with 151 samples but with no data preprocessing and an imbalanced dataset in three classes. This resulted in an over-fitting problem, as shown in Table 4, which shows the training and testing performance. Over-fitting happens when a model performs well on training data but not well on test data. This problem arose with the proposed models because the training data size was insufficient, with the models trained on limited training samples

for several epochs. However, additional testing and experimentation were needed to find the optimal performance due to the limited number of available testing and training samples.

Table 4: Model performances before data augmentation

Model	Train accuracy	Train loss	Test accuracy	Test loss
ResNet50	1	0.02	0.88	0.39
ResNet101	1	0.009	0.84	0.49
MobileNet	0.97	0.59	0.80	0.71
MobileNetV2	1	1.08	0.88	1.40
DenseNet201	1	0.01	0.94	0.35
DenseNet121	1	0.002	0.86	0.40

4.4.2 Results After Augmentation

To reduce over-fitting, the research re-run the training with data augmentation for each of the three classes. As can be seen in Table 5, the performance of the models improved as a result of the addition of more data. However, due to the unbalanced classes, there was still a gap between training and testing. Now the research study had the additional task of improving performance and drastically simplifying the over-fitting problem.

Table 5: Model performance after data augmentation

Model	Train accuracy	Train loss	Test accuracy	Test loss
ResNet50	0.98	0.03	0.79	1.51
ResNet101	0.94	0.27	0.71	0.73
MobileNet	0.99	0.04	0.92	0.29
MobileNetV2	0.99	0.05	0.84	0.46
DenseNet201	0.99	0.01	0.78	0.74
DenseNet121	0.99	0.76	0.86	0.87

4.4.3 Result after Data Balancing

Now the models have been trained again after balancing the classes. As seen in Table 6, the best performance was achieved on the balanced dataset. The performance of test accuracy improved, avoided over-fitting, and bridged the gap between training and testing accuracy in all six models. The performance of ResNet101 was the best.

After attempts to avoid and reduce over-fitting, it can be seen that data augmentation and balancing lead to the best performance in all six models, as shown in Fig. 15. Now, the research study attempts to choose which of these six models has the highest accuracy, where the X-axis is plotted with the algorithm's name given in Table 6 representing the accuracy of training and testing.

Table 6: Model performances after data balancing

Model	Train accuracy	Train loss	Test accuracy	Test loss
ResNet50	0.95	0.16	0.93	0.32
ResNet101	**0.99**	**0.09**	**0.94**	**0.40**
MobileNet	0.97	0.21	0.92	0.47
MobileNetV2	0.93	0.22	0.88	0.40
DenseNet201	0.86	0.50	0.82	0.67
DenseNet121	0.88	0.31	0.85	0.49

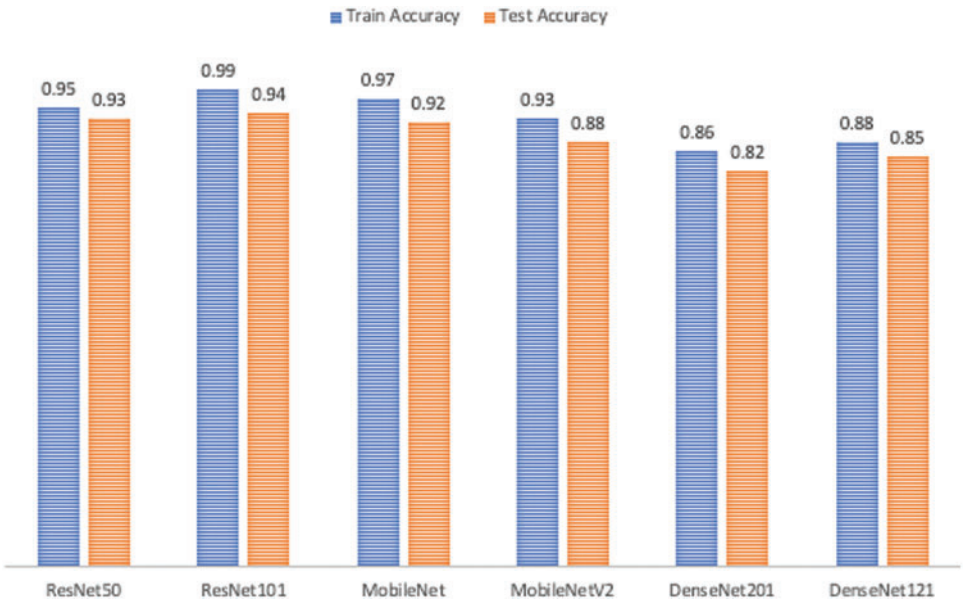


Figure 15: Performance comparison

4.4.4 Chosen Model-ResNet101

The chosen model must distinguish between the three classes and detect whether or not the input samples are fake. After comparing the results of the proposed models, as shown in Fig. 15, ResNet101 proved to be the highest-performing model when trained with the dataset chosen. The model originally had a 99% training accuracy and 94% validation accuracy. The Below-the-Plot Figures in Figs. 16 and 17 show an accuracy of 5.2 and a loss of 5.3 for ResNet101, respectively.

The training parameters for customized ResNet101 are shown in Table 7.

4.4.5 Evaluation of Selected Model

To improve the classification performance of the proposed model, this research imported Regularize 11, l2 and added GlobalAveragePooling2D, Flatten, Batch Normalization, Dropout, and Dense Layers, as shown in Fig. 18. Also, the proposed model used Model Checkpoint to monitor validation accuracy and loss to store higher weight in the file path. This research used the Early Stopping form of regularization; if no improvement takes place, training is stopped to avoid over-fitting problems.

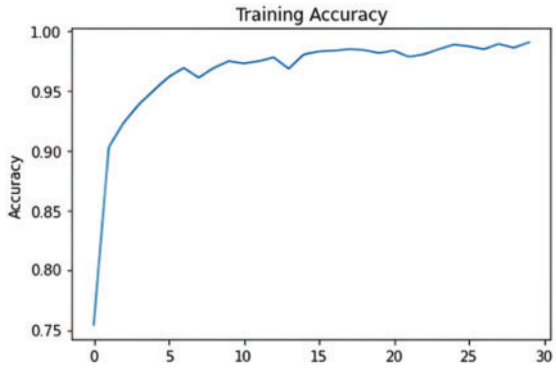


Figure 16: ResNet101 accuracy

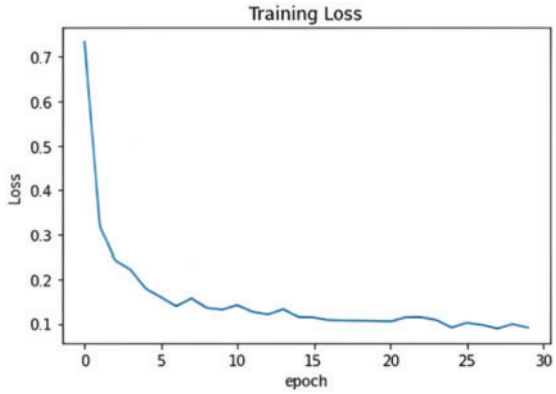


Figure 17: ResNet101 loss

Table 7: Description of hyperparameters

Network parameters	Value
Epochs	30
Batch size	32
Optimizer	Adam

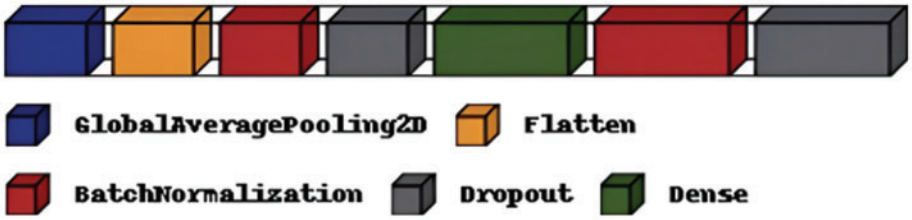


Figure 18: Layers added to ResNet101

Moreover, the researchers trained the model with several optimizers and got the best results with the Adam optimizer; the comparison with different optimizers is shown in Table 8.

Table 8: Comparison with other optimizers

Optimizers	Training accuracy	Testing accuracy
SDG	95%	93%
RMSprop	98%	92%
Adam	**99%**	**94%**

4.4.6 Comparison with Other Models

Current deep fake detection studies use a range of datasets and techniques. However, not all of these studies used the same evaluation metrics as the proposed model, and many used non-medical image datasets, which differ from those used in the proposed model. In this research study, related studies are compared, using accuracy as an evaluation metric and datasets comparable to ours. Table 9 summarizes the studies used for comparison.

Table 9: Comparison with related works

Reference	Model	Accuracy
Ragab et al. [10]	SVM	93.5%
Ghadi et al. [11]	ResNet50	97%
Siddharth et al. [12]	DenseNet121	80.4%
Suk et al. [13]	U-Net	91%
Reichman et al. [14]	ConnectionNet	85%
Gite et al. [15]	U-Net++	98%
Riza et al. [16]	Deep fake predictor (DFP)	94%
Proposed model	**ResNet101**	**99.0%**

4.4.7 Discussion

The results shown in the study by Ragab et al. [10] used a custom model which can classify CycleGAN-tampered and real medical images using a Support Vector Machine (SVG) model and achieved an accuracy of 93.5%. In the study by Ghadi et al. [11], seven CNN-based architectures were trained and tested, which yielded the following results: DenseNet169 with 95.0% accuracy, 97.0% for Dense-Net121; 96.0% for DenseNet201; 92.0% for VGG16; 94% for VGG19; 97.0% for ResNet50; and a customized model with 90.0% accuracy, in detecting GAN-produced deep fakes on face images. Siddharth et al. [12] also trained several deep learning models in the hope of finding the most efficient model; their DenseNet121 model achieved the best result with an accuracy of only 80.4%. Then, Suk et al. [13] addressed data manipulation in fundus lesions using a model based on sparse CNN to detect data manipulated by the U-Net and Cycle General Adversarial Network (Cycle GAN). The model achieved 91.0% accuracy. Reichman et al. [14] produced a deep-learning-based framework, ConnectionNet, which automatically determines whether a medical image has been tampered with. The suggested framework had a deep fake detection accuracy score of 85.0%. Gize et al. [15] proposed the U-Net++, which had improvements over the generic U-Net for identifying deep fake samples, with an accuracy of 98%, which was a massive improvement over the results previously discussed for

the U-Net model. Finally, Riza et al. [16], who proposed a new model, the Deep Fake Predictor (DFP), were able to train the model and achieve an accuracy of 94%.

All the models discussed in this section had excellent results and are fit for production. However, this research was focused on optimizing the processes of detecting deep fakes and brought errors to a bare minimum. The proposed model did this with an accuracy of 99% using the ResNet101 model. Also, the light architecture of this paper's method gives advantages in terms of avoiding the problems of over-fitting, which improves the recognition power of the paper's model. As a result, the research can summarize that this paper's model using ResNet101 is proficient at distinguishing between tampered and untampered scans.

5 Conclusion and Future Work

The reason for improving the classification performance has been argued in detail throughout this study. The proposed research model attempted to enhance the efficiency of detection to prevent the consequences of medical deep fakes. For this purpose, the proposed model spent considerable time and effort finding and preparing a high-quality dataset for the experiment and selecting the appropriate models. The study indicates that data augmentation and balancing were essential pre-processing techniques for identifying real and fake CT scans. As previously noted, ROI localization would also be effective. The proposed model introduced ResNet101, a state-of-the-art algorithm, which proved to be the highest-performing model to train with the dataset chosen. The model originally had a 99% training accuracy and 94% validation accuracy. In the future, we plan to evaluate the proposed method for other types of visual deep fakes, such as Neural Texture, lip-synching, etc.

Acknowledgement: The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Saleh, Alaa; data collection: Reem, Noura Almatrafi; analysis and interpretation of results: Reema, Noura Almallouhi. Alaa; draft manuscript preparation: Alaa, Reem, Reema, Noura Almatrafi, Noura Almallouhi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data is contained within the article.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Albahli, "Efficient GAN-based Chest Radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia," *International Journal of Medical Sciences*, vol. 17, no. 10, pp. 71–84, 2020.

[2] W. L. Maria, S. Blogg, S. Jackson and S. K. Hosking, "NPS MedicineWise: 20 years of change," *Journal of Pharmaceutical Policy and Practice*, vol. 11, no. 3, pp. 1–7, 2022.

[3] Y. Y. Ghadi, I. Akhter, S. A. Alsuhibany, T. A. Shloul, A. Jalal *et al.,* "Classification of deep fake videos using pre-trained convolutional neural networks," in *Proc. Int. Conf. on Digital Futures and Transformative Technologies (ICoDT2)*, New York, NY, USA, pp. 1–6, 2021.

[4] J. Ashish, "Debating the ethics of deep fakes," *Tackling Insurgent Ideologies in a Pandemic World, ORF and Global Policy Journal*, vol. 14, no. 4, pp. 75–79, 2020.
[5] S. Albahli, "Twitter sentiment analysis: An Arabic text mining approach based on COVID-19," *Frontiers in Public Health*, vol. 10, no. 10, pp. 71–84, 2022. https://doi.org/10.3389/fpubh.2022.966779
[6] M. Westerlund, "The emergence of deep fake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 81–98, 2019.
[7] M. Prakash, C. Shyam, O. Hariand and S. Maheshkar, "Authentication of medical images using a passive approach," *IET Image Processing*, vol. 13, no. 13, pp. 2420–2427, 2022.
[8] A. Alsirhani, M. Ezz and A. M. Mostafa, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, no. 5, pp. 967–984, 2020.
[9] M. Ragab, H. A. Abdushkour, A. F. Nahhas and W. H. Aljedaibi, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in *USENIX Security Symp.*, Santa Clara, CA, USA, vol. 2019, 2019.
[10] M. Ragab, H. A. Abdushkour, A. F. Nahhas and W. H. Aljedaibi, "GAN-based medical image small region forgery detection via a two-stage cascade framework," *arXiv preprint arXiv:2205.15170*, vol. 19, no. 21, pp. 512–523, 2022.
[11] Y. Y. Ghadi, I. Akhter, S. A. Alsuhibany, T. A. Shloul, A. Jalal *et al.,* "Comparative analysis of deep fake image detection method using convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 88, no. 6, pp. 910–931, 2021.
[12] S. Siddharth and Y. Wen, "Machine learning based medical image deep fake detection: A comparative study," *Machine Learning with Applications*, vol. 8, no. 4, pp. 813–825, 2022.
[13] K. Y. Suk, H. J. Song and J. H. Han, "A study on the development of deep fake-based deep learning algorithm for the detection of medical data manipulation," *Webology*, vol. 19, no. 1, pp. 4396–4409, 2022.
[14] B. Reichman, L. Jing, O. Akin and T. Yingli, "Medical image tampering detection: A new dataset and baseline," *Pattern Recognition*, vol. 68, no. 4, pp. 266–277, 2021.
[15] S. Gite, A. Mishra and K. Kotecha, "Enhanced lung image segmentation using deep learning," *Neural Computing and Applications*, vol. 8, no. 3, pp. 1–15, 2022.
[16] A. Riza, M. Munir and M. Almutairi, "A novel deep learning approach for deep fake image detection," *Applied Sciences*, vol. 12, no. 9, pp. 920–934, 2022.
[17] C. Dustin, T. Nicholas, G. Cuccolo, A. Ibrahim, A. Furnas *et al.,* "Photographic and video deep fakes have arrived: How machine learning may influence plastic surgery," *Plastic and Reconstructive Surgery*, vol. 145, no. 4, pp. 145–1456, 2020.
[18] D. Jia, W. Dong, R. Socher, L. J. Li, L. Kai *et al.,* "Imagenet: A large-scale hierarchical image database," *Computer Vision and Pattern Recognition*, vol. 129, no. 3, pp. 248–255, 2021.
[19] K. Brett and B. Koonce, "ResNet 50." convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization," *Computer Vision*, vol. 304, no. 2, pp. 63–72, 2021.
[20] W. Wei, I. Yutao, T. Zou, X. Wang, Y. Jieyu *et al.,* "A novel image classification approach via dense-MobileNet models," *Mobile Information Systems*, vol. 27, no. 2, pp. 139–149, 2020.
[21] E. Sandler and A. Howard, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, Salt Lake City, Utah, USA, pp. 4510–4520, 2018.
[22] Medical Deepfakes Lung Cancer Dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/ymirsky/medical-deepfakes-lung-cancer
[23] C. J. Richard, Y. F. Ming, W. Tiffany, A. M. Chen and G. K. Drew, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, 2021.
[24] E. S. Pashentsev, "Malicious use of deep fakes and political stability," *Artificial Intelligence and Robotics*, vol. 82, no. 2, pp. 19–32, 2020.

[25] O. J. Andrei and G. M. Sharon, "Deep fake: A social construction of technology perspective," *Current Issues in Tourism*, vol. 24, no. 13, pp. 1798–1802, 2021.
[26] Q. Huma, A. Farooq, M. Nawaz and T. Nazir, "FRD-LSTM: A novel technique for fake reviews detection using DCWR with the Bi-LSTM method," *Multimedia Tools and Applications*, vol. 11, no. 4, pp. 1–15, 2023.