DeepFake Detection for Chest-CT Images

1st Shatakshi Shree
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
shatakshi.shree2020@vitstudent.ac.in

2nd Agrim Sharma
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
agrim.sharma2020@vitstudent.ac.in

3rd Sriganesh Raj
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
sriganesh.raj2020@vitstudent.ac.in

***Abstract*—Medical imaging is the non-invasive process of producing internal visuals of a body for the purpose of medical examination, analysis, and treatment. Medical imaging can be used to detect and diagnose a range of diseases, injuries, and abnormalities. Medical imaging plays a crucial role in modern medicine, allowing medical professionals to accurately diagnose and treat a wide range of conditions. It also allows for minimally invasive procedures and surgical interventions. Numerous attacks on clinics and hospitals occurred in 2018, resulting in serious data breaches and delays in medical services. When an attacker has access to medical records, they are able to do far more than just demand a ransom or sell the information. One method of creating deepfakes is to inject and remove tumors from medical imaging. Failure to recognise medical deepfakes might result in significant losses of hospital resources or even death. Deepfakes can be used to create fake medical images or videos that compromise the privacy and dignity of patients. For example, deepfakes can be used to create fake images or videos of patients in compromising positions or situations. Deepfake detection is necessary in medical image analysis to prevent the creation and spread of manipulated medical images that can have serious consequences for patient diagnosis and treatment. Hence, we will attempt to build a machine learning model and train it to carry out detection between original images and deepfakes created and analyze them.**

***Index Terms*—DeepFake, ResNet18, ResNet50, DCGAN, CNN**

I. Introduction

DeepFake technology has become increasingly prevalent in recent years, with the ability to generate realistic and convincing images and videos using machine learning techniques. While DeepFake technology has many potential applications, it also has the potential to be used for malicious purposes such as creating fake medical images.

In particular, the use of DeepFake technology to generate fake Chest-CT images has become a concern for the medical community. The ability to generate realistic Chest-CT images can have serious consequences for patient diagnosis and treatment, as fake images can lead to misdiagnosis and inappropriate treatment.

To address this issue, researchers have been developing methods for DeepFake detection in Chest-CT images. These methods involve the use of machine learning algorithms that can identify features in Chest-CT images that are indicative of fakery.

Detecting DeepFake Chest-CT images is a critical step towards ensuring the accuracy and reliability of medical diagnosis and treatment. By developing effective detection methods, medical professionals can better identify and address any potential issues with DeepFake Chest-CT images, ensuring that patients receive the appropriate care they need.

II. Literature Survey

The work proposed in [10] attempts to address the detection of such attacks with a structured case study. Specifically, eight different machine learning algorithms are evaluated, which include three conventional machine learning methods (Support Vector Machine, Random Forest, Decision Tree) and five deep learning models (DenseNet121, DenseNet201, ResNet50, ResNet101, VGG19) in distinguishing between tampered and untampered images. The findings of this work show near perfect accuracy in detecting instances of tumor injections and removals. In this study [7] author presents generative adversarial neural networks capable of generating realistic images of knee joint X-rays with varying osteoarthritis severity. It offers 320,000 synthetic (DeepFake) X-ray images from training with 5,556 real images. The models are validated regarding medical accuracy with 15 medical experts and for augmentation effects with an osteoarthritis severity classification task. Survey of 30 real and 30 DeepFake images for medical experts was devised. The result showed that on average, more DeepFakes were mistaken for real than the reverse. The result signified sufficient DeepFake realism for deceiving the medical experts. In work [8], at present, data driven approaches to classifying medical images are prevalent. However, most medical data is inaccessible to general researchers due to standard consent forms that restrict research to medical journals or education. Our study focuses on GANs, which can create artificial fundus images that can be indistinguishable from actual fundus images. Before using these fake images, it is essential to investigate privacy concerns and hallucinations thoroughly. As well as, reviewing the current applications and limitations of GANs is very important. A talked in the work [2] Medical Deepfake pertains to application of AI-triggered deepfake technology on to medical modalities like Computed Tomography (CT) scan, X-Ray, Ultrasound etc. The tampering attacks involve either insertion or removal of certain disease conditions, tumors in/from the modality under analysis. This paper implements and demonstrates a practical, lightweight technique which aims to accelerate deepfake detection for biomedical imagery by detecting malignant tumors injected in modalities of healthy patients. The developed technique makes use of convolutional reservoir networks (CoRN), which enable ensemble feature extraction and results in improved classification metrics. [3] The paper comprehensively evaluates the applicability of the recent top ten state-of-the-art Deep Convolutional Neural Networks (CNNs) for automatically detecting COVID-19 infection using chest X-ray images. Moreover, it provides a comparative analysis of these models in terms of accuracy. This study identifies the effective methodologies to control and prevent infectious respiratory diseases. Our trained models have demonstrated outstanding results in classifying the COVID-19 infected chest x-rays. [4] In this study, we propose a chest X-ray image classification method based on feature fusion of a dense convolutional network (DenseNet) and a visual geometry group network (VGG16). This paper adds an attention mechanism (global attention machine block and category attention block) to the model to extract deep features. A residual network (ResNet) is used to segment effective image information to quickly achieve accurate classification. The paper [5] explains that Kidney–ureter–bladder (KUB) imaging is a radiological examination with a low cost, low radiation, and convenience. Although emergency room clinicians can arrange KUB images easily as a first-line examination for patients with suspicious urolithiasis, interpreting the KUB images correctly is difficult for inexperienced clinicians. Recently, artificial-intelligence-based computer-aided diagnosis (CAD) systems have been developed to help clinicians who are not experts make correct diagnoses for further treatment more effectively. [1] In this study, we proposed a CAD system for KUB imaging based on a deep learning model designed to help first-line emergency room clinicians diagnose urolithiasis accurately. This study identifies real and deepfake images using different Convolutional Neural Network (CNN) models to get the best accuracy. It also explains which part of the image caused the model to make a specific classification using the LIME algorithm. To apply the CNN model, the dataset is taken from Kaggle, which includes 70 k real images from the Flickr dataset collected by Nvidia and 70 k fake faces generated by StyleGAN of 256 px in size. For experimental results, Jupyter notebook, TensorFlow, NumPy, and Pandas were used as software, InceptionResnetV2, DenseNet201, InceptionV3, and ResNet152V2 were used as CNN models. [9] This paper has analyzed the features related to the computer vision of digital content to determine its integrity. This method has checked the computer vision features of the image frames using the fuzzy clustering feature extraction method. By the proposed deep belief network with loss handling, the manipulation of video/image is found by means of a pairwise learning approach. This proposed approach has improved the accuracy of the detection rate by 98 percent on various datasets. The author in [6] presents a survey of algorithms used to create deepfakes and, more importantly, methods proposed to detect deepfakes in the literature to date. We present extensive discussions on challenges, research trends and directions related to deepfake technologies. By reviewing the background of deepfakes and state-of-the-art deepfake detection methods, this study provides a comprehensive overview of deepfake techniques and facilitates the development of new and more robust methods to deal with the increasingly challenging deepfakes.

III. Design

The project's design is split into 3 parts - creation of deepfakes, training of the models on the original dataset and then choosing the best model and training it on the deepfakes dataset. The model used for deepfake creation was DCGAN, and there were four models used for training on the original dataset (custom-made CNN, ResNet18, ResNet50 and MobileNetV2). The custom-made CNN was then trained on the dataset of created deepfakes.

IV. Methodology

The first step in our project was to create deepfake images from our original dataset of chest-CT images. For the creation of deepfakes, we plan to use a Generative Adversarial Network (GAN). GANs are used for teaching a deep learning model to generate new data from that same distribution of training data. They are made up of two different models - a generator and a discriminator.

Then, we train 4 models on our original dataset, namely CNN, ResNet18, ResNet50 and MobileNetV2. In ResNets, the technique of skip connections is used. The skip connection connects activation functions of a layer to further layers by skipping some layers in between, and this structure forms a residual block. ResNets are made by stacking these blocks together.

To illustrate the difference between deepfake images and original ones we select the model which achieved the highest accuracy on the original dataset and run that model on the newly created deepfake images. Our proposed methodology should achieve a comparatively lower accuracy on the deepfake images, hence allowing us to detect them.

A. DCGAN

DCGAN stands for "Deep Convolutional Generative Adversarial Network". It is a type of generative adversarial network (GAN) that uses deep convolutional neural networks (CNN) to generate new images. The DCGAN architecture is made up of two networks - a generator network and a discriminator network - that are trained in parallel. The generator network takes random noise as input and generates fake images, while the discriminator network takes real images and fake images and tries to distinguish between them. During training, the generator network tries to create images that are realistic enough to fool the discriminator network, while the discriminator network tries to correctly identify the fake images generated by the generator network. The goal is to optimize the adversarial loss function. This competition between the generator and discriminator networks results in the generation of increasingly realistic images. DCGAN is particularly effective in generating high-quality images of complex objects and scenes. In the medical field, DCGAN has been used to generate synthetic medical images that can be used for training and testing machine learning algorithms.

B. CNN

CNN stands for Convolutional Neural Network. It is a type of artificial neural network that is commonly used in deep learning applications for image and video analysis, natural language processing, and other pattern recognition tasks. The key difference between a CNN and a traditional neural network is the use of convolutional layers. In a CNN, each convolutional layer applies a set of filters to the input data. These filters are small matrices of numbers that are multiplied with a corresponding portion of the input data, generating a set of feature maps that represent the presence of certain patterns or features in the input. CNNs are highly effective in image and video analysis tasks because they are able to automatically learn and extract features from raw data, without the need for manual feature engineering. They have been used in a wide range of applications, including object recognition, facial recognition, image segmentation, and natural language processing.

C. ResNet18

ResNet18 is a convolutional neural network architecture that was introduced by researchers at Microsoft in 2015. It is part of the ResNet family of models, which are known for their deep architectures and high performance on a variety of image recognition tasks.

ResNet18 is a 72-layer architecture with 18 deep layers. The architecture of this network aimed at enabling large amounts of convolutional layers to function efficiently. However, the addition of multiple deep layers to a network often results in a degradation of the output. This is known as the problem of vanishing gradient where neural networks, while getting trained through back propagation, rely on the gradient descent, descending the loss function to find the minimizing weights. Due to the presence of multiple layers, the repeated multiplication results in the gradient becoming smaller and smaller thereby "vanishing" leading to a saturation in the network performance or even degrading the performance.

The primary idea of ResNet is the use of jumping connections that are mostly referred to as shortcut connections or identity connections. These connections primarily function by hopping over one or multiple layers forming shortcuts between these layers. The aim of introducing these shortcut connections was to resolve the predominant issue of vanishing gradient faced by deep networks. These shortcut connections remove the vanishing gradient issue by again using the activations of the previous layer. These identity mappings initially do not do anything much except skip the connections, resulting in the use of previous layer activations. This process of skipping the connection compresses the network; hence, the network learns faster. This compression of the connections is followed by expansion of the layers so that the residual part of the network could also train and explore more feature space. The network is considered to be a DAG network due to its complex layered architecture and because the layers have input from multiple layers and give output to multiple layers.

The introduction of residual blocks overcomes the problem of vanishing gradient by implementation of skip connections and identity mapping. Identity mapping has no parameters and maps the input to the output, thereby allowing the compression of the network, at first, and then exploring multiple features of the input.

One of the benefits of ResNet18 is its relatively small size compared to other deep neural network architectures. This makes it easier to train and deploy in resource-constrained environments such as mobile devices or embedded systems.

Overall, ResNet18 has proven to be a powerful tool for image recognition tasks, thanks to its innovative architecture and the use of residual connections to improve training and performance. Its success has helped to inspire further research into deep neural network architectures, and it continues to be an important tool for image recognition and computer vision applications.

D. ResNet50

ResNet-50 is a convolutional neural network that is 50 layers deep. You can load a pretrained version of the neural network trained on more than a million images from the ImageNet database. The ResNet architecture is considered to be among the most popular Convolutional Neural Network architectures around. The requirement for a model like ResNet arose due to a number of pitfalls in modern networks at the time. Difficulty in training deep neural networks: As the number of layers in a model increase, the number of parameters in the model increases exponentially. More expressive, less different: A neural network is often considered to be a function approximator. It has the ability to model functions given input, target and a comparison between the function output and target. Adding multiple layers into a network makes it more capable to model complex functions. Vanishing/Exploding Gradient: This is one of the most common problems plaguing the training of larger/deep neural networks and is a result of oversight in terms of numerical stability of the network's parameters. ResNet, due to its architecture, does not allow these problems to occur at all, because of the skip connections which do not allow it as they act as gradient super-highways, allowing it to flow without being altered by a large magnitude.

Skip Connections: The ResNet paper popularized the approach of using Skip Connections. In mathematical terms, it would mean y=x+F(x) where y is the final output of the layer.

In terms of architecture, if any layer ends up damaging the performance of the model in a plain network, it gets skipped due to the presence of the skip-connections.

E. MobileNetV2

MobileNetV2 is a convolutional neural network architecture that seeks to perform well on mobile devices. It is based on an inverted residual structure where the residual connections are between the bottleneck layers. Residual blocks connect the beginning and end of a convolutional block with a skip connection. By adding these two states the network has the opportunity of accessing earlier activations that weren't modified in the convolutional block. This approach turned out to be essential in order to build networks of great depth. A residual block connects wide layers with a skip connection while layers in between are narrow.

V. Results

Among the 4 models we trained on the original dataset, our custom-made CNN model gave the highest accuracy of around 99.6%. The same model, when trained on a dataset of 500 deepfakes generated by our DCGAN model from the original dataset, gives an accuracy of 50%. Thus, our hypothesis that the accuracy of the model drops steeply when the learning from a set of original models is transferred to a deepfake image was found to be true for the given use case. Future work in the project can be the extension and verification of this methodology for other images in the medical field and the generalization of the idea for various other sectors where deepfakes are an issue.

VI. Conclusion

Building a Convolutional Neural Network (CNN) model from scratch for detecting chest-CT deepfake images is a complex task that requires a deep understanding of both computer vision and deep learning techniques. However, it is possible to create a robust and accurate model by following a series of steps. First, we collected a dataset of chest-CT images. The dataset was properly curated and labeled to ensure that the model is trained on high-quality data. Next, the CNN model architecture was designed, which involves selecting appropriate layers, filter sizes, activation functions, and pooling methods. This was done based on the specific requirements of the deepfake detection task and the size and complexity of the dataset. Once the model architecture was finalized, the model was trained using the dataset. The model was validated and tested during the training process to ensure that it is learning the features of the images and is not overfitting or underfitting. Finally, the trained model was evaluated using our test dataset of deepfake images. The model's performance metrics was analyzed to determine its effectiveness in detecting chest-CT deepfake images. In conclusion, building a CNN model from scratch for detecting chest-CT deepfake images is a challenging task, but we were able to build a model which helped achieve our goal.

References

[1] Abir, W.H., Khanam, F.R., Alam, K.N., Hadjouni, M., Elmannai, H., Bourouis, S., Dey, R., Khan, M.M.: Detecting deepfake images using deep learning techniques and explainable ai methods. Intelligent Automation & Soft Computing. pp. 2151–2169 (2023)

[2] Budhiraja, R., Kumar, M., Das, M., Bafila, A.S., Singh, S.: Medifaked: Medical deepfake detection using convolutional reservoir networks. In: 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT). pp. 1–6. IEEE (2022)

[3] Ghaffar, Z., Shah, P.M., Khan, H., Zaidi, S.F.A., Gani, A., Khan, I.A., Shah, M.A., et al.: Comparative analysis of state-of-the-art deep learning models for detecting covid-19 lung infection from chest x-ray images. arXiv preprint arXiv:2208.01637 (2022)

[4] Lingzhi, K., Jinyong, C.: Classification and detection of covid-19 x-ray images based on densenet and vgg16 feature fusion [j]. Biomedical Signal Processing and Control **77** (2022)

[5] Liu, Y.Y., Huang, Z.H., Huang, K.W.: Deep learning model for computer-aided diagnosis of urolithiasis detection from kidney–ureter–bladder images. Bioengineering **9**(12), 811 (2022)

[6] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., Huynh-The, T., Nahavandi, S., Nguyen, T.T., Pham, Q.V., Nguyen, C.M.: Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding **223**, 103525 (2022)

[7] Prezja, F., Paloneva, J., Pölönen, I., Niinimäki, E., Äyrämö, S.: Deepfake knee osteoarthritis x-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. Scientific Reports **12**(1), 18573 (2022)

[8] Salini, Y., HariKiran, J.: Deepfakes on retinal images using gan. International Journal of Advanced Computer Science and Applications **13**(8) (2022)

[9] Saravana Ram, R., Vinoth Kumar, M., Al-shami, T.M., Masud, M., Aljuaid, H., Abouhawwash, M.: Deep fake detection using computer vision-based deep neural network with pairwise learning. Intelligent Automation & Soft Computing **35**(2) (2023)

[10] Solaiyappan, S., Wen, Y.: Machine learning based medical image deepfake detection: A comparative study. Machine Learning with Applications **8**, 100298 (2022)