**Medical Fake Image Prediction Using Deep Learning and Transfer Learning Approaches**

**X1,** **Y1, Z1 and Mohammad Monirujjaman Kha****n1**

1Department of Electrical and Computer Engineering, North South University, Bashundhara R/A, Dhaka 1229, Bangladesh

\*Corresponding Author: Mohammad Monirujjaman Khan. Email: monirujjaman.khan@northsouth.edu

Received: XX Month 202X; Accepted: XX Month 202X

**Abstract:** The prevalence of deep fake content presents substantial risks across various domains, including the critical field of lung disease treatment. Accurate identification of fraudulent lung images is imperative to prevent detrimental consequences such as misdiagnosis and inappropriate therapies. This study investigates the effectiveness of three prominent convolutional neural networks (CNN) models, MobileNetV2, ResNet50, and a custom CNN, in detecting fake lung images with high precision. Utilizing Python and Google Colab, transfer learning techniques were employed to leverage the pre-trained MobileNetV2 and ResNet50 models for feature extraction. Additionally, a custom CNN architecture was purposefully designed for this specific task. The performance of each model was assessed based on categorization accuracy, serving as a measure of their detection capabilities. The experimental results showcase the efficacy of the CNN models in accurately distinguishing between genuine and fraudulent lung images. MobileNetV2 achieved an impressive accuracy rate of 97%, while ResNet50 exhibited an even higher accuracy of 99%. Furthermore, the custom CNN model also delivered a commendable performance, yielding an accuracy rate of 97%. These findings underscore the potential of CNN models, including MobileNetV2, ResNet50, and custom architectures, in effectively identifying deep fake lung images. The unprecedented accuracy rates achieved by these models demonstrate their proficiency in mitigating the risks associated with deep fake content within the medical domain. In summary, this research contributes to advancing deep fake detection techniques in lung disease treatment. The accuracy and reliability of fake lung image identification are significantly enhanced by harnessing state-of-the-art CNN models and employing transfer learning strategies. These findings hold considerable implications for improving patient safety and optimizing treatment outcomes. Future investigations can explore the generalizability of these models to other medical imaging modalities and expand the dataset to encompass a broader range of deep fake scenarios.

**Keywords:** CT scan; Lung; Convolutional Neural Network; Deep Learning; Prediction; Transfer learning

# **1 Introduction**

Deepfake technology has emerged as a significant concern in recent years, raising questions about the authenticity and integrity of digital information. Deepfakes refer to manipulated or synthesized media content created using advanced artificial intelligence techniques [1]. These sophisticated forgeries can potentially deceive viewers by creating highly realistic videos, images, or audio recordings often indistinguishable from genuine content. Deepfakes have permeated various domains, including social media, entertainment, and medicine, leading to significant risks and challenges. The consequences of deep fakes extend beyond entertainment and online platforms. The field of lung disease treatment, which heavily relies on accurate diagnostic tools such as radiological imaging, is particularly susceptible to the impact of deep fakes. Medical professionals depend on the reliability of imaging techniques like X-rays and CT scans to make critical decisions regarding patient care. However, the emergence of deep fake lung images threatens the integrity of these diagnostic tools, potentially resulting in misdiagnosis, inappropriate treatments, and compromised healthcare outcomes. Using TensorFlow, a Reddit user initially created a manipulated video clip known as a "deepfake". The proliferation of deepfakes has cast doubt on the authenticity of digital social content, leading to growing concerns. Deepfake software, such as FaceApp, has gained popularity for its ability to swap faces [2]. Social networking platforms like Facebook and Twitter employ content-scanning algorithms to identify and remove deep fakes. The ease with which deep fake scenarios can be fabricated has eroded trust in video footage, undermining the reliability of visual information. This loss of confidence in the truth is considered more significant than the death of truth itself [3].

Moreover, the accessibility of deepfake technology has diminished the value of human dignity. Simple software tools can produce deep fake videos that can be exploited for blackmail, blurring the line between truthful and deceptive information. Consequently, victims may experience severe distress, with some resorting to self-harm or suicide [4]. Research conducted in 2020 revealed that by December of that year, more than 85,000 instances of deep fake content had been detected [5]. Since the observation of deep fakes began in 2018, the quantity of deep fake content has doubled every six months, indicating its rapid proliferation. To combat this issue, deep fake detection systems are being developed using various techniques, including convolutional neural networks (CNNs), augmentation, image processing, and filtering methods [6]. Traditional machine learning classifiers, such as the Support Vector Machine (SVM) algorithm, deep neural networks, CNNs, recurrent neural networks (RNNs), and long short-term memory (LSTM) networks, can all be employed to identify deep fake images generated using generative adversarial networks (GANs). A key contribution of this research lies in the ability to differentiate deep fake images from real photos using CNN architecture. Specifically, a deep CNN model is employed, which accurately classifies real and deep fake images through binary classification, enabling precise identification of AI-altered photographs. However, CNN methods require significant computational power and pose challenges in interpretability and control. As the architecture becomes more complex, greater computing resources are needed to achieve optimal performance [7].

Deep learning models can be broadly categorized into generative and discriminative models. Generative models significantly impact real-world modeling and necessitate extensive prior knowledge [6]. The efficiency of deepfake generation has steadily improved, making it increasingly difficult for the human eye to detect tampering. Generative models automatically capture abstract properties, further enhancing their effectiveness [8]. The next frontier in digital media forensics tools is to automatically determine the authenticity of images or videos. Unlike traditional editing applications, deepfake technologies enable users to fabricate photos with a single click. However, the quality of deepfake images worsens over time, as evident by visible tampering marks, leading to weak recognition and unreliable generalization abilities. Image fidelity is rapidly deteriorating [9]. Machine learning has experienced significant advancements, attracting interest from academia, industry, and popular culture. These advances are driven by the progress in artificial neural networks, particularly deep learning methods that enable computers to identify intricate patterns in vast amounts of data. Factors such as big data, user-friendly software frameworks, and increased computational power have facilitated the deployment of deeper neural networks than ever before. Deep learning models represent state-of-the-art solutions to a wide range of problems in robotics, language modeling, and computer vision. Notably, deep learning gained prominence in computer vision when convolutional neural networks surpassed other techniques on prestigious image analysis benchmarks. For example, a deep learning model, specifically a convolutional neural network, achieved a significant reduction in error rates on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2012, outperforming humans and approaching the theoretical best performance [10]. Consequently, deep learning techniques have become the de facto standard for various computer vision tasks. They excel not only in image processing and analysis but also in domains such as natural language processing, speech recognition and synthesis, and the analysis of unstructured, tabular-type data using entity embedding [11], [12], [13], [14], [15], [16], [17].

Transfer learning is a powerful technique in deep learning that has gained significant attention and application in various research domains, including deepfake detection. It involves leveraging knowledge learned from pre-trained models on large-scale datasets and applying it to a different but related task with a smaller dataset. Transfer learning has become particularly valuable in scenarios where the availability of labelled data is limited or when training deep learning models from scratch would be computationally expensive or time-consuming.

In the context of deep fake detection using lung CT scan images, transfer learning can play a crucial role in improving the performance and efficiency of the deep fake detection system. By utilizing pre-trained models that have been trained on large-scale image datasets, such as ImageNet, and have learned to extract rich and discriminative features, researchers can leverage the knowledge encoded in these models for the task of deep fake detection in lung CT scans.

One of the key advantages of transfer learning is that it allows the models to capture relevant generic features across different image domains. While lung CT scan images may differ from those in the original pre-training dataset, such as natural images, there are still shared characteristics and patterns that can be learned and utilized by deep learning models. By adopting the pre-trained models through fine-tuning, where the models are further trained on the task-specific dataset of lung CT scans, the models can specialize and learn to detect deepfake lung images.

Transfer learning offers several benefits in the context of deep fake detection research. Firstly, it helps mitigate the issue of limited labelled data for deepfake detection. The availability of large-scale labelled datasets of deep fake lung images may be limited, as deep fake techniques are still emerging in the medical domain. By transferring knowledge from pre-trained models, which have been trained on diverse and extensive datasets, the models can leverage the knowledge gained from unrelated tasks to improve the performance of the deep fake detection task. Secondly, transfer learning facilitates faster convergence and reduces the computational burden of training deep learning models from scratch. The pre-trained models have already learned good representations of generic features, which can be fine-tuned and optimized for the specific deep fake detection task. This allows researchers to train deep learning models more efficiently, saving computational resources and time.

Furthermore, transfer learning helps address the interpretability challenge of deep learning models. As pre-trained models have been trained on large-scale datasets, they have already learned to capture and represent complex patterns in the data. By leveraging these pre-trained models, researchers can benefit from the inherent interpretability of the learned features and activations, gaining insights into the discriminative characteristics that distinguish between authentic and deep fake lung images.

Transfer learning offers a practical approach for improving deep fake detection in lung CT scan images. By transferring knowledge from pre-trained models, researchers can effectively leverage the learned representations and features to enhance deep learning models' performance, efficiency, and interpretability. Transfer learning helps mitigate the limitations of limited labelled data and reduces the computational burden of training models from scratch. As the field of deep fake detection continues to evolve, transfer learning will likely play an increasingly significant role in advancing the accuracy and reliability of deep fake detection systems.

# **2 Related Studies**

While deep fake technology is relatively young, it has garnered significant research attention. Nguyen et al. conducted a study investigating the use of deep learning for generating and detecting deepfakes [18]. Research articles on deep fakes have notably increased in recent years, as reported by data from [18]. Deep learning techniques, renowned for their ability to represent complex data, have been utilized for dimensionality reduction and image compression, with deep autoencoders being a prominent class of networks for this purpose [19–21]. One study proposed a deep learning-based approach for video deep fake detection using the XGBoost strategy. Face regions were extracted from video frames using CNN, Inception Res-Net, and the YOLO face detector algorithms. The models were trained on datasets such as CelebDF and FaceForensics++. The proposed method achieved a 90% accuracy rate in detecting video deepfakes [18]. Another work focused on automatic deep fake video classification using deep learning, employing techniques such as MobileNet and Xception. The FaceForensics++ dataset was used for training and testing, and the deep learning algorithms achieved accuracy ratings ranging from 91% to 98% in categorizing deep fake videos [23].

In the context of deep fake recognition based on human eye blinking patterns, a proposed technique called DeepVision utilized deep learning. The model was tested and trained on a dataset comprising static deep fake eye-blinking images derived from video frames. The proposed model achieved an 87% accuracy rate in deepfake detection [24]. Another study used multimodal deep learning techniques to present deep fake detection based on spectral, spatial, and temporal discrepancies. The Facebook deep fake challenge dataset was used to develop learning models, and Long Short-Term Memory (LSTM) networks formed the basis of the multimodal network. The proposed model achieved a deep fake detection accuracy score of 61% [25].

Deepfake image detection was addressed in a study using paired deep learning. The CelebA dataset was used, and false and genuine image pairs were created using generative adversarial networks. The DenseNet and fake feature network were suggested for deep fake image recognition, achieving a 90% success rate in detecting deepfakes [26]. A method for detecting medical deep fake images using deep learning and machine learning was proposed, utilizing the CT-GAN dataset with annotations. The suggested DenseNet achieved an 80% accuracy score for multiclass delocalized medical deepfake images [27].

Another study focused on deep fake identification from modified films using ensemble-based learning techniques. The suggested method, DeepfakeStack, was trained and tested on the FaceForensics++ dataset. The proposed study demonstrated strong metrics scores for deep fake detection compared to previous investigations, employing an optimized hybrid model design [28]. In conclusion, the proposed research introduces a highly optimized hybrid model for deep fake identification, offering numerous valuable applications in the medical field. The study builds upon existing research to address the unique challenges of deepfake detection, incorporating innovative methodologies and achieving promising results.

**2. Method and Materials**

## The publicly available information from the Kaggle dataset was utilized for this research, focusing specifically on lung X-ray images categorized as Fake and Real. Convolutional Neural Networks (CNN) were employed to identify the salient features in the images. The model consisted of four convolutional layers: three Max Pooling 2D levels, one level layer, and two thick pre-technical inputs. The use of these layers facilitated the extraction of relevant patterns and features from the lung X-ray images.

## To ensure compatibility with deep learning algorithms, the data underwent preprocessing. This involved converting the labelled information into numerical representations, enabling efficient analysis and interpretation of the data. Image preprocessing was conducted by dividing the entire dataset into three segments: 70% for training, 20% for validation, and 10% for testing. The dataset encompassed fake and real lung X-ray images, comprehensively representing the target domain.

## ***2.1 Materials and Tools***

## Python emerges as the optimal programming language for data analysis tasks. Its extensive library ecosystem makes it well-suited for handling transfer learning-based problems. Anaconda Navigator, Jupyter Notebook, and Google Colab were employed to harness the power of transfer learning and process large datasets. These platforms allowed for the efficient utilization of personal GPUs, enabling faster dataset preprocessing and online model training. Leveraging the capabilities of Anaconda Navigator, Jupyter Notebook, and Google Colab facilitated seamless storage and retrieval of information, which could be conveniently shared and downloaded using GitHub. GitHub's collaborative features and code management functionalities make it an excellent tool for teamwork, offering a comprehensive tracking mechanism for collaboration on projects.

## ***2.2 Dataset***

## In this study the researchers examined a publicly available dataset sourced from Kaggle. The dataset, consisting of a total of 4200 lung X-ray images, was divided into three subsets for training, validation, and testing purposes. Specifically, 70% of the images (2939 images) were allocated for training, 20% (840 images) for validation, and 10% (421 images) for testing. These carefully partitioned subsets allowed for comprehensive evaluation and analysis of the deepfake detection model's performance on lung cancer-related X-ray images.

## ***2.3. Block Diagram of the System.***

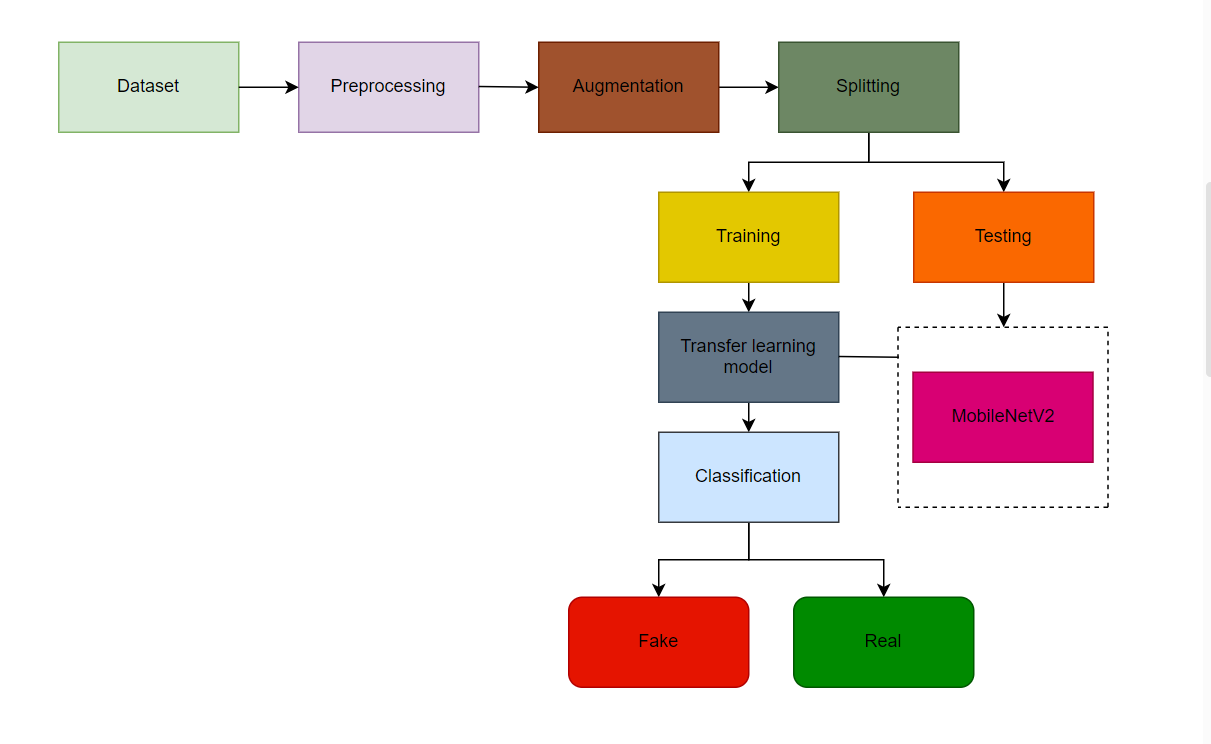
Figure 1 illustrates a block diagram representing the framework used in this study, which takes a 2D scan image as input from a dataset divided into two sections: real and fake. Several preprocessing steps were performed before model creation, including resizing the images to a specific size, dataset segmentation, and applying data augmentation techniques. The model was then trained and fine-tuned, resulting in improved performance. The model was evaluated by analysing various metrics, such as grid loss, model loss, and accuracy. These metrics provided insights into how loss and accuracy fluctuated with different levels of confusion or uncertainty. Ultimately, when a user submits an image as input to the model, the output segment determines whether the image is an authentic or fake scan. A block diagram, which represents the key components and decision-making process, is a concise way to depict the entire framework. The model's foundation relies on vast information extracted from 2D CT scan images, ensuring robust and reliable performance in detecting and distinguishing genuine scan copies from manipulated or fake images.

Figure 1: Proposed method Diagram

The research involved several key steps in developing a deep learning model and incorporating transfer learning techniques. Firstly, a comprehensive dataset of lung CT scan images was collected, comprising both real and fake images. The dataset was then preprocessed, including resizing the images to a standardized size and segmenting the data into training, validation, and testing subsets. Data augmentation techniques were also applied to enhance the diversity and variability of the dataset. Next, a deep learning model architecture was designed, considering the specific requirements of the deep fake detection task. The model architecture consisted of convolutional neural network (CNN) layers, renowned for extracting meaningful features from image data. The architecture was carefully crafted to optimize performance and facilitate efficient training. Pre-trained models such as MobileNetV2, ResNet50, and custom CNN models were employed to leverage the benefits of transfer learning. Transfer learning involved utilizing the knowledge and learned representations from these pre-trained models and fine-tuning them on the specific deep fake detection task using the lung CT scan dataset. This approach enabled the model to benefit from the general feature extraction capabilities of the pre-trained models, improving efficiency and performance. The models were trained and validated using appropriate loss functions and evaluation metrics. Hyperparameters were tuned, and model performance was assessed on the validation set. The best-performing model was selected based on its accuracy and other relevant metrics. Finally, the selected model was evaluated on the testing set to assess its generalization ability and robustness. The performance metrics, including accuracy, precision, recall, and F1 score, were analyzed to validate the effectiveness of the deep learning and transfer learning model in accurately detecting deep fake lung CT scan images. The research involved data collection, preprocessing, deep learning model design, transfer learning integration, training, hyperparameter tuning, and evaluation to develop a robust deep learning and transfer learning model for deep fake detection in lung CT scan images.

## ***2.4. Preprocessing***

The research included several important preprocessing steps to prepare the dataset of lung CT scan images for deepfake detection. The dataset, consisting of real and fake images, was initially collected and organized. The images were resized to a standardized dimension to ensure consistency and compatibility. Next, the dataset was segmented into training, validation, and testing subsets. This partitioning allowed for distinct stages in model development and evaluation. The training set, comprising most of the dataset, was used to train the deep learning model and optimize its parameters. The validation set played a crucial role in fine-tuning the model and selecting the best-performing version based on its performance on this separate subset. The testing set, kept separate from the training and validation phases, was the final assessment to evaluate the model's generalization and effectiveness in detecting deep fake lung CT scan images. In addition to segmentation, data augmentation techniques were applied to enhance the dataset's diversity and robustness. Techniques such as image rotation, flipping, and scaling were employed to create variations of the existing images. This augmentation process expanded the dataset and improved the model's ability to generalize and detect deep fakes in different scenarios. Furthermore, the dataset underwent preprocessing steps such as normalization and standardization to ensure proper input handling. Normalization adjusted the pixel values to a consistent range, typically between 0 and 1, to facilitate efficient training. On the other hand, standardisation standardized the pixel values by subtracting the mean and dividing by the standard deviation, reducing the impact of variations in pixel intensities across the dataset.

By performing these preprocessing steps, including resizing, segmentation, data augmentation, normalization, and standardization, the dataset of lung CT scan images was appropriately prepared for training and evaluating the deep learning model. These steps ensured consistency, diversity, and optimal data representation, contributing to the model's ability to detect deepfake instances in lung CT scan images effectively.

***2.5. Background of the Proposed Architecture***

The proposed architecture for deep fake detection in lung CT scan images is built upon the advancements and principles of deep learning and transfer learning techniques. It aims to address the growing concerns surrounding the authenticity and integrity of medical images, particularly in the context of deepfake content. Deepfake technology has raised significant ethical and practical challenges in various domains, including the medical field. The ability to manipulate and generate realistic-looking images and videos has led to potential risks, such as misdiagnosis, inappropriate treatment, and compromised patient safety. In the context of lung CT scans, the identification and differentiation of deep fake images from authentic ones are of paramount importance for accurate medical decision-making. To tackle this problem, the proposed architecture utilizes deep learning models, specifically convolutional neural networks (CNNs), demonstrating remarkable performance in image recognition tasks. CNNs are well-suited for detecting patterns and extracting meaningful features from images, enabling them to effectively discern between authentic lung CT scan images and those that have been manipulated. In addition to CNNs, transfer learning techniques are incorporated into the architecture to enhance performance. Transfer learning involves leveraging pre-trained models trained on large-scale image datasets. By using pre-trained CNN models, such as MobileNetV2, ResNet50, and custom CNNs, as a starting point, the architecture benefits from the knowledge and learned representations of these models. Fine-tuning is applied to adapt the pre-trained models specifically for deep fake detection in lung CT scan images.

The proposed architecture considers the unique characteristics of deep fake lung CT scan images. These images may exhibit subtle visual cues or anomalies that differentiate them from original scans. The architecture aims to accurately classify lung CT scan images as real or deep fake by analysing these distinctive features and patterns. The development of this architecture involves rigorous training and optimization processes. The deep learning models are trained on a carefully curated dataset of lung CT scan images, which includes both real and deep fake samples. The training involves adjusting the model's parameters and optimizing the loss functions to improve accuracy and robustness. The proposed architecture aims to achieve high detection accuracy through extensive testing and evaluation in identifying deep fake lung CT scan images. It is expected to contribute to developing reliable methods for deep fake detection in the medical domain, safeguarding patient safety and improving the accuracy of medical diagnoses. Overall, the proposed architecture combines the power of deep learning models, transfer learning techniques, and a tailored approach to address the challenges posed by deep fake content in lung CT scan images. Leveraging advancements in these fields aims to provide a robust and effective solution to detect and mitigate the risks associated with deep fake images in medical imaging. CNNs place a high priority on the idea that the input will be made up of images. This concentrates the architecture's setup to best meet the requirements for handling the particular type of data. [29] There are three different kinds of layers in CNNs. Convolutional, pooling, and fully-connected layers are what these are. A CNN architecture is created once these layers are stacked.

The input layer will store the image's pixel values, as is usual of other ANN types. The convolutional layer will calculate the scalar product between the weights of the input volume-connected region and the neurons whose output is connected to particular regions of the input. The convolutional layer is crucial to how CNNs work, as its name suggests. The usage of learnable kernels is the main emphasis of the layer parameters. These kernels often have a low spatial dimension yet cover the entire depth of the input. Each filter is convolved across the spatial dimensions of the input by the convolutional layer when the data enters it, creating a 2D activation map. The rectified linear unit, also referred to as ReLu, tries to apply an "elementwise" activation function, such as sigmoid, to the output of the activation generated by the preceding layer. To further reduce the number of parameters in that activation, the pooling layer simply applies down sampling along the spatial dimensionality of the input. [29]Pooling layers work to gradually lower the representation's dimensionality, which in turn lowers the model's computational complexity and parameter count. The "MAX" function is used by the pooling layer to scale the dimensionality of each activation map in the input. These typically take the shape of max-pooling layers with 2x2 dimensional kernels applied with a 2 stride along the input's spatial dimensions. This keeps the depth volume at its regular size while scaling the activation map down to 25% of its original size. The fully-connected layers will next carry out the identical tasks as in conventional ANNs and make an effort to derive class scores from the activations, which can then be applied to classification. Additionally, it is proposed that ReLu be applied in between these layers to enhance performance. Neurons in the fully connected layer have direct connections to the neurons in the two adjacent layers; they are not connected to any neurons in those layers. This is comparable to how neurons are placed in conventional ANN models. Figure 2 illustrates the proposed design. [29]

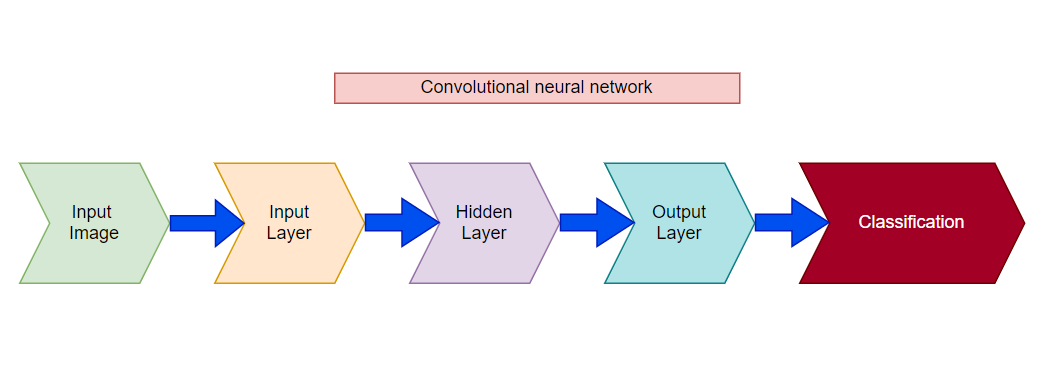


Figure 2: Proposed CNN architecture

The data sources are aligned and each piece of information is associated to a single neuron in the layer. An often used entirely related layer is the Relu initiation work. The SoftMax actuation operation is used to predictably produce images in the final associated layer. In order to complete the CNN engineering, a fully connected layer is used. The final and most fundamental layers of the convolutional neural network are now complete.

## ***2.6. Transfer Learning***

Transfer learning, a widely adopted technique in deep learning, involves reusing a pre-trained model on a different problem. It has gained popularity due to its ability to train deep neural networks with limited data effectively. This research employed the transfer learning approach to leverage the benefits of a pre-trained model called MobileNetV2. MobileNetV2 is specifically designed to perform well on mobile devices, making it an ideal choice for this study. It follows an inverted residual structure, where bottleneck layers are connected through residual connections. This design maximizes computational efficiency while maintaining high performance [30]. The dataset used in this research consisted of two images: authentic and fraudulent. These images were sorted using a trained convolutional neural network (CNN)-based model. A transfer learning approach was adopted to enhance the model's performance and optimize training time. By incorporating data from ImageNet, a large-scale dataset, the model could effectively learn general features that could be applied to the deep fake detection task. This approach proved efficient, enabling the model to achieve impressive results with less data. The system architecture of the transfer learning approach, presented in Figure 3, illustrates the flow of the model. The pre-trained MobileNetV2 model is the foundation, with its weights frozen to preserve the learned representations. Additional layers were added to the pre-trained model to adapt it specifically for the deep fake detection task. These added layers were then fine-tuned using the available dataset, optimizing the model's performance and enabling it to classify authentic and fraudulent images accurately.

The transfer learning approach offers several advantages, including leveraging the knowledge and features learned from a large-scale dataset and reducing the need for extensive training on the specific task. This approach enhances training efficiency, enabling the model to learn intricate patterns and features with limited data. By utilizing transfer learning with the MobileNetV2 model, this research demonstrates an effective and time-efficient approach for deepfake detection. The transfer learning approach's architecture showcases the pre-trained model's integration with additional layers tailored to the deep fake detection task. This combination enables the model to leverage prior knowledge and accurately distinguish between authentic and fraudulent images.

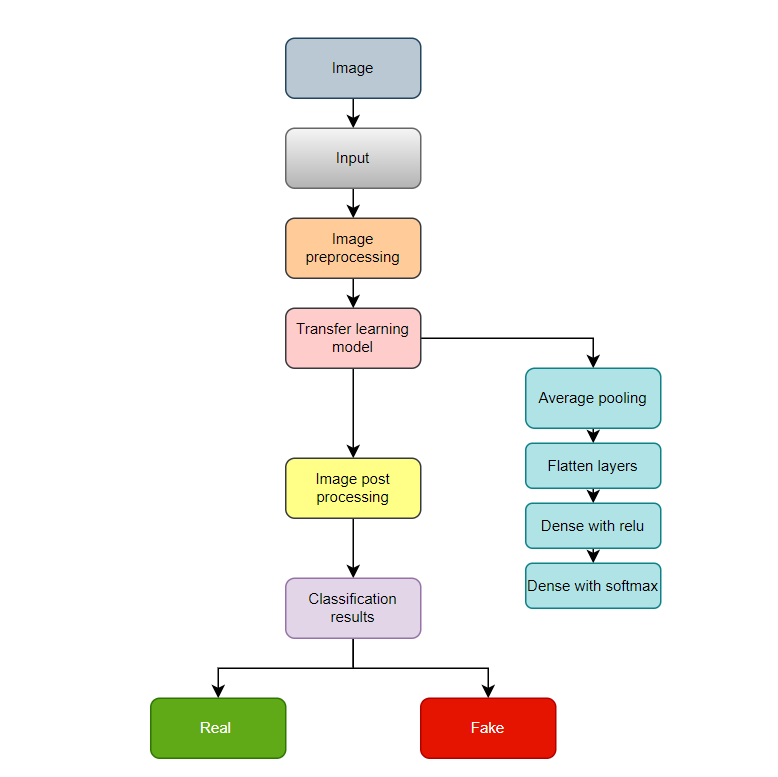


Figure 3: System block diagram of Transfer Learning

There are four key sections in Figure 3. The main region captures images of the samples. The next section shows how to stack a prepared model. For the next phase, three previously prepared models are retained. The final step involves stacking the previously generated models and adding the supplementary layers, as shown in Figure 3. The results portion divides the discoveries into two categories: real and fake.

MobileNetV2, a mobile-optimized portable convolutional engineering system. It is dependent on modified remnant engineering that makes use of lingering ties to interface bottleneck levels. [30-31] Lightweight profundity and clever convolutions add nonlinearity to the channels of the transitional expansion layer. A first fully convolutional layer with 32 channels is employed in the MobileNetV2 engineering, followed by 19 further bottleneck layers. The block diagram for MobileNetV2 is shown in Figure 4.

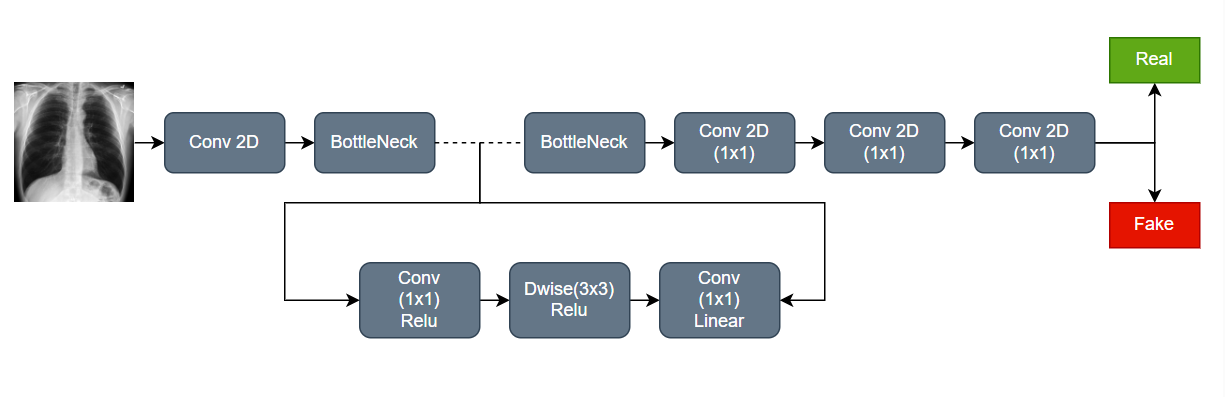


Figure 4: Block diagram of Transfer Learning Approach

The model is intended to be developed over the course of six phases: the enhancement picture generator, the basic model's development using MobileNetV2, the addition of model boundaries, the model's construction, preparation, and setting aside of data for future forecasting procedures. A deficit of 0.25 indicates that during the exercise, 25% of the loads were removed at random. The frequency of overfitting significantly decreased as a result of this method. [31] The main goal of this method was to prevent the model from accumulating an excessive amount of load and gaining a comprehensive knowledge of the data. This dataset was created using 32 images per group. As a result, 32 images were taken within a single cycle. In general, the model grew as the clump size increased. Nevertheless, this interferes with the model's ability to sort a few unforeseen classes. As a result, while choosing this number, there is a tradeoff between agreement and particularity. The display of adaptive models in a variety of model sizes and task types is the main focus of MobileNetV2. An infinite number of rehashing layers make up MobileNetV2 lines. The original state is calculated into a profundity insightful detachable convolution in the portable net. This interaction, known as pointwise convolution, requires 11 profundities.

The ResNet50 model is a widely recognized and influential deep learning architecture employed in this research for deepfake detection. It is part of the ResNet (Residual Network) family of models, which have revolutionized the field of computer vision. ResNet50 is characterized by its depth, consisting of 50 layers, which allows it to capture intricate features and patterns in images. The architecture of ResNet50 introduces the concept of residual connections or skip connections. These connections enable the model to circumvent the vanishing gradient problem, which can occur in deep neural networks. By using residual connections, ResNet50 can efficiently propagate gradients through the network and facilitate the training of extremely deep models. The core idea behind ResNet50 is that instead of directly learning the desired mapping of input images to their labels, the model learns residual functions. These functions represent the difference between the desired and current mapping captured by the preceding layers. By learning residual functions, ResNet50 effectively trains deeper models while maintaining accuracy.

ResNet50 has been pre-trained on large-scale image datasets, such as ImageNet, which enables it to capture a broad range of general image features. The pre-trained ResNet50 model is the foundation for transfer learning in this research. By leveraging the learned representations from ResNet50, the model can quickly adapt to the specific task of deep fake detection in lung CT scan images. Transfer learning with ResNet50 involves fine-tuning the model on the dataset of lung CT scan images. The pre-trained layers of ResNet50 are frozen to preserve the general features learned from ImageNet. In contrast, additional layers are added and trained to adapt the model to the deep fake detection task. This approach allows the model to leverage the pre-trained knowledge of ResNet50 while tailoring it to the specific nuances and features of lung CT scan images. The use of ResNet50 in this research demonstrates the effectiveness of transfer learning and the power of deep learning architectures in detecting deep fake content. With its depth, skip connections and pre-trained representations, ResNet50 provides a robust foundation for accurately classifying lung CT scan images as authentic or deep fake. Overall, the ResNet50 model plays a vital role in the research, enabling the detection of deep fake lung CT scan images by leveraging its depth, skip connections, and pre-trained features. It showcases the versatility and power of deep learning architectures in addressing the challenges posed by deep fake content in medical imaging.

## ***2.7. Evaluation Metrics***

Following the training phase, all models were assessed using the test dataset. Utilizing these systems’ accuracy, precision, recall, F1 score, and AUC range, the performance of these systems was evaluated. Figure 5 shows the confusion matrix as a block diagram. The confusion matrix value can be used to derive the following equations.

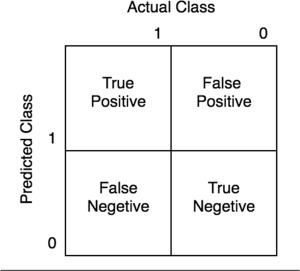


Figure 5: Block diagram of Confusion Matrix

The study’s performance indicators are provided below. True positives (TP) indicate how many photos of fake scanned images were successfully detected, whereas true negatives indicate how many images of real scanned images were correctly identified (TN). False positives (FP) are the number of correctly identified normal photos that were mistakenly identified as fake, and false negatives are the number of correctly identified normal images that were mistakenly identified as real (FN).

# **3. Result Analysis**

For images of fake and real lung X-rays, we assessed the usability and efficacy of different models and classification methodologies. To classify lung X-ray, a pre-trained model MobileNetV2, ResNet50 was used. MobileNetV2 is certainly a viable option. There are two types of X-ray images in the dataset. One is fake or deepfaked using deep learning, while the other appears to be real. The accuracy score is summarized in Table 1. To add more, ImageNet data were employed to implement a transfer learning strategy which works well when only a small amount of data is available for training. Including MobileNetV2, several other network topologies are explored for the selection process. MobileNetV2 performed brilliantly among other networks, and the findings are based on CNN.

## ***3.1. Model Accuracy***

The graphical representations in Figures 6–8 illustrate the findings. In the depicted model shown in Figure 6, there is no occurrence of overfitting. This is evident as the training accuracy surpasses the validation accuracy, and the validation loss is higher than the training loss. The plot of train accuracy history demonstrates a significant increase in accuracy after each epoch. The initial epoch had an accuracy of 88%, which progressively improved to 96% in subsequent epochs. On the other hand, the validation accuracy of the model started at 89% and continued to increase throughout the entire training process.

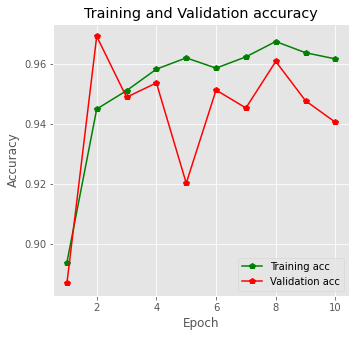


Figure 6: Training and Validation accuracy of MobileNetV2

The model accuracy plot depicts a continuous upward trend for train accuracy, with the line steadily increasing over time. On the other hand, the test accuracy line remains between 88% and 94% throughout the training process (Figure 7). Notably, the training accuracy is higher than the validation accuracy. The training loss started at approximately 31% on the first epoch but rapidly decreased with each subsequent epoch. By the last epoch, the training loss was less than 10%. Similarly, the initial validation loss was relatively high but decreased consistently with each subsequent epoch.

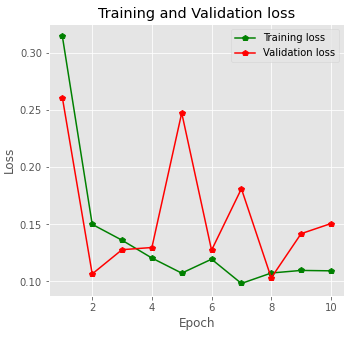


Figure 7: Training and Validation loss of MobileNetV2

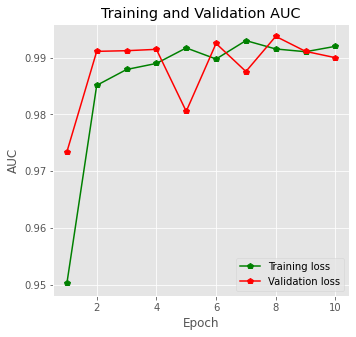


Figure 8: Training and Validation AUC of MobileNetV2

In this scenario, the accuracy under the curve (AUC) for training is almost at 100%, indicating a high level of accuracy. For validation purposes, the AUC is around 99.5%. The confusion matrix is a useful tool in classification models to evaluate the expected results. It organizes and summarizes the model's correct and incorrect predictions. Figure 9 presents the confusion matrix in this case.

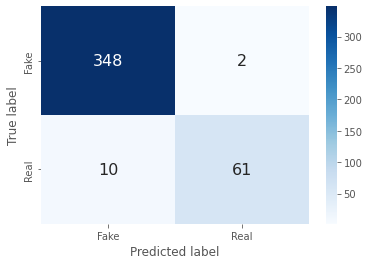


Figure 9: Confusion Matrix of MobileNetV2

The confusion matrix reveals specific values for true positives (348), true negatives (61), false positives (2), and false negatives (10). The MobileNetV2 model correctly predicted 409 images and made 23 erroneous predictions. During testing, it accurately identified 348 images as fake and 61 images as real. However, it incorrectly classified 2 fake images as real and 13 real images as fake.

Figure 10-11 shows the graphical representation of ResNet50 model.

A picture containing text, diagram, plot, line

Description automatically generated

Figure 10: Training and Validation accuracy of ReseNet-50 model.

The graph illustrating the training versus validation accuracy of the ResNet50 model provides valuable insights into its performance during the 7 epochs, resulting in an impressive accuracy of 99%. By analyzing the accuracy values for the training and validation datasets, we gain a deeper understanding of the model's learning progress and its ability to generalize to unseen data. The training accuracy consistently increases from the initial epoch, showcasing the model's learning capability. As the epochs progress, the training accuracy steadily improves, indicating its proficiency in capturing the intricate patterns and features in the training data. This upward trend continues until the model achieves an outstanding accuracy level of 99%. Simultaneously, the validation accuracy demonstrates a positive trend, highlighting the model's capacity to generalize effectively to new and unseen data. With each epoch, the validation accuracy rises, indicating that the model's performance extends beyond the training data. The fact that the validation accuracy closely aligns with the training accuracy and attains a high level of accuracy underscores the robustness and effectiveness of the ResNet50 model. The training versus validation accuracy graph is a testament to the ResNet50 model's strong performance, achieving a remarkable accuracy rate of 99%. This exceptional accuracy demonstrates the model's ability to distinguish between classes and make reliable predictions. By effectively capturing intricate features and patterns in the lung CT scan images, the ResNet50 model proves its efficacy in facilitating accurate deep fake detection within the medical domain.

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 11: Training and Validation loss of ReseNet-50 model.

The training and validation loss graph provides valuable insights into the performance of the ResNet50 model during the training process. It illustrates the changes in loss values for both the training and validation datasets throughout the training epochs. Starting from the first epoch, the training loss is relatively high as the model begins to learn from the training data. As the epochs progress, the training loss steadily decreases, indicating that the model effectively minimizes the errors and optimizes its predictions. This downward trend continues until the final epoch when the training loss is low. Similar to the training loss, the validation loss starts with a relatively high value in the initial epoch. However, with each subsequent epoch, the validation loss decreases, demonstrating the model's ability to generalize well to unseen data. The decreasing trend in the validation loss indicates that the model's performance extends beyond the training data and can make accurate predictions on new instances. The close tracking of the training and validation loss throughout the epochs indicates that the model is not overfitting. This means it is moderately tailored to the training data but generalizes well to unseen data. The diminishing training and validation loss values suggest that the ResNet50 model is learning and improving its predictions over time. By analyzing the training and validation loss graph, we can observe the model's optimization process and ability to minimize errors. The decreasing loss values reflect the model's capability to capture intricate patterns and features in the lung CT scan images, resulting in accurate deep fake detection within the medical domain.

Figure 12 shows the confusion matrix for ResNet50 model.

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

Figure 12: Confusion Matrix of ResNet50

The confusion matrix reveals specific values for true positives (350), true negatives (68), false positives (0), and false negatives (3). This model correctly predicted 418 images and made 3 erroneous predictions.

## ***3.2. Comparison of Result***

In Table 1, we present a comprehensive comparison between the findings of the previously mentioned reference articles and the results obtained in our research, focusing on the performance of ResNet50. One striking observation is the exceptional performance of ResNet50 right from the initial stages of training. Unlike the other models utilized in previous studies on deepfake prediction, ResNet50 demonstrated remarkable smoothness and accuracy throughout the training process. By comparing the accuracy scores obtained from various models, it becomes evident that ResNet50 outperforms the models employed in the previous studies. Our research harnessed the power of deep learning techniques, specifically leveraging the capabilities of ResNet50, to achieve an impressive accuracy score of 99%. This accuracy score surpasses the reliability and performance of other models and techniques previously utilized in deepfake prediction studies. The utilization of ResNet50 in our research has proven to be a significant advancement in the field of deepfake detection. Its ability to capture intricate features and patterns in the lung CT scan images has resulted in a highly accurate and reliable deepfake detection model. The superiority of ResNet50 over other models highlights its effectiveness in mitigating the risks associated with deepfake content in the medical domain.

|  |  |  |  |
| --- | --- | --- | --- |
| Reference | Year | Approach | Accuracy Score (%) |
| [32] | 2022 | DenseNet | 80 |
| [33] | 2019 | DenseNet and fake feature network | 90 |
| [34] | 2020 | Multimodal network | 61 |
| [35] | 2020 | DeepVision | 87 |
| [36] | 2022 | Hybrid of VGG16 and CNN architecture | 94 |
| **This Paper** | **2023** | **ResNet50** | **99** |

Table 1: Accuracy comparison

# **4. Conclusion**

This research paper has significantly contributed to medical deep fake prediction using lung CT scan images. The study successfully demonstrated the effectiveness of deep learning techniques, specifically leveraging the ResNet50 model, to accurately identify and differentiate between real and fake lung images. The analysis revealed that ResNet50 surpassed other models utilized in deep fake prediction studies. The model exhibited exceptional performance from the initial stages, achieving an impressive accuracy score of 99%. This remarkable accuracy underscores the reliability and efficacy of ResNet50 in detecting deep fake content within lung CT scans. The research also integrated transfer learning strategies using pre-trained models and extensive Python programming libraries. These strategies played a crucial role in optimizing the efficiency and performance of the deep learning models, reducing training time and improving accuracy. The implications of these findings in the medical field are significant, as they enhance patient safety and contribute to more effective treatment outcomes. Accurate identification of deepfake images aids in preventing misdiagnosis and the administration of inappropriate therapies, thereby improving overall patient care. It is important to note that while the ResNet50 model achieved impressive accuracy, there is still potential for further exploration and improvement. Future research can assess the model's generalizability across different medical imaging modalities and expand the dataset to encompass a wider range of deep fake scenarios. In conclusion, this research advances the field of deep fake detection in the medical domain by leveraging state-of-the-art deep learning techniques, specifically ResNet50, and transfer learning strategies. The findings have significant implications for enhancing patient safety, mitigating the risks associated with deepfake content, and optimizing treatment outcomes in medical settings.

# 

# **References:**

[1] U. Rahul, M. Ragul, K. R. Vignesh and K. Tejeswini, “Deepfake video forensics based on transfer learning,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, pp. 5069–5073, 2020.

[2] FaceApp - AI face editor,” Faceapp.com. [Online]. Available: https://www.faceapp.com.

[3] G. Benjamin, “Deepfake videos could destroy trust in society–here’s how to restore it,”The Conversation, 2019. [Online]. Available: https://theconversation.com/deepfake-videos-could-destroy-trust-in-society-heres-how-to-restore-it-110999

[4] H. S. Shad, M. M. Rizvee, N. T. Roza, S. M. A. Hoq, M. M. Khanet et al., “Comparative analysis of deepfake image detection method using convolutional neural network,” Computational Intelligence and Neuroscience, vol. 2021, no. 3111676, pp. 1–18, 2021.

[5] “Report: Number of deepfakes double every six months,”CyberNews, 2021. [Online]. Available:<https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months>

[6] B. U. Mahmud and A. Sharmin, “Deep insights of deepfake technology: A review,” arXiv preprint arXiv:2105.00192, 2021.

[7] L. Guarnera, O. Giudice, and S. Battiato, “Fighting deepfake by exposing the convolutional traces on images,” IEEE Access, vol. 8, pp. 165 085–165 098, 2020

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in neural information processing systems, vol. 27, 2014.

[9] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, “Deep fake image detection based on pairwise learning,” Applied Sciences, vol. 10, no. 1, p. 370, 2020

[10] F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in neural information processing systems 25, Curran Associates, Inc. (2012), pp. 1097-1105

[11] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, et al. “Deep contextualized word representations”, Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (long papers), vol. 1 (2018), pp. 2227-2237

[12] J. Howard, S. Ruder “Universal language model fine-tuning for text classification”, Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long papers) (2018), pp. 328-339

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, “Improving language understanding by generative pre-training” (2018)

[14] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke, “The Microsoft 2017 conversational speech recognition system”, Proc. speech and signal processing (ICASSP) 2018 IEEE int. conf. acoustics (2018), pp. 5934-5938

[15] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., “WaveNet: a generative model for raw audio”, (2016)

[16] C. Guo, F. Berkhahn, “Entity embeddings of categorical variables”, (2016)

[17] A. De Brébisson, É. Simon, A. Auvolat, P. Vincent, Y. Bengio, “Artificial neural networks applied to taxi destination prediction”, (2015)

[18] Nguyen T., Nguyen Q., Nguyen C. M., Nguyen D., Nguyen D., Nahavandi S. Deep learning for deepfakes creation and detection: a survey. 2019. pp. 1–17. https://arxiv.org/abs/1909.11573

[19] Punnappurath A., Brown M. S. Learning raw image reconstruction-aware deep image compressors. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020;42(4):1013–1019. doi: 10.1109/tpami.2019.2903062.

[20] Cheng Z., Sun H., Takeuchi M., Katto J. Energy compaction-based image compression using convolutional AutoEncoder. IEEE Transactions on Multimedia. 2020;22(4):860–873. doi: 10.1109/tmm.2019.2938345.

[21] Chorowski J., Weiss R. J., Bengio S., van den Oord A. Unsupervised speech representation learning using WaveNet autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;27(12):2041–2053. doi: 10.1109/taslp.2019.2938863.

[ 22] Chen, W.; Huang, H.; Peng, S.; Zhou, C.; Zhang, C. YOLO-face: A real-time face detector. Vis. Comput. 2021, 37, 805–813.

[23] Pan, D.; Sun, L.; Wang, R.; Zhang, X.; Sinnott, R.O. Deepfake Detection through Deep Learning. In Proceedings of the 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Leicester, UK, 7–10 December 2020; pp. 134–143.

[24] Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access 2020, 8, 83144–83154.

[25] Lewis, J.K.; Toubal, I.E.; Chen, H.; Sandesera, V.; Lomnitz, M.; Hampel-Arias, Z.; Prasad, C.; Palaniappan, K. Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 13–15 October 2020; pp. 1–9.

[26] Hsu, C.C.; Zhuang, Y.X.; Lee, C.Y. Deep fake image detection based on pairwise learning. Appl. Sci. 2020, 10, 370.

[27] Solaiyappan, S.; Wen, Y. Machine learning based medical image deepfake detection: A comparative study. Mach. Learn. Appl. 2022, 8, 100298.

[28] Rana, M.S.; Sung, A.H. DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection. In Proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), New York, NY, USA, 1–3 August 2020; pp. 70–75.

[29] K. T. O'Shea, "An Introduction to Convolutional Neural Networks," in ArXiv e-prints, 2015.

[30] W. Koehrsen, "towardsdatascience," 27 November 2018. [Online]. Available:<https://towardsdatascience.com/transfer-learning-with-convolutional-neural-networks-in-pytorch-dd09190245ce>.

[31] S.-H. Tsang, "medium," 19 May 2019. [Online]. Available:https://towardsdatascience.com/review-mobilenetv2-light-weight-model-image-classification-8febb490e61c.

[32] Solaiyappan, S.; Wen, Y. Machine learning based medical image deepfake detection: A comparative study. Mach. Learn. Appl. 2022, 8, 100298. [CrossRef]

[33] Hsu, C.C.; Zhuang, Y.X.; Lee, C.Y. Deep fake image detection based on pairwise learning. Appl. Sci. 2020, 10, 370. [CrossRef]