Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Transfer Learning-Based MobileNetV2 Framework for Medical Deepfake Detection Using Chest X-Ray Images

Roshani Anandrao Parate[1], Dr.Kirti Jain[2]

[1]PhD Scholar Computer Scienece and Engineering, Sanjeev Agrawal Global Educational University(SAGE) Bhopal, Madhyapradesh, email: parate.roshani121@gmail.com
[2]Professor Sanjeev Agrawal Global Educational University(SAGE) Bhopal,Madhyapradesh,email: Kirtijain9@gmail.com

ABSTRACT

Medical deepfakes are an emerging cybersecurity threat to healthcare systems because image-manipulated radiology can influence the interpretation of the diagnostic results and clinical decision making. The artificial intelligence-generated or modified medical images may conceal diseases, introduce novel abnormalities, or endanger patient safety with false graphical data. Consequently, more believable medical deepfake detection systems are required to provide trustful healthcare imaging environments. The paper shows a transfer learning based lightweight convolutional neural network-based framework to detect manipulated chest X ray images. The experimental data was built on the basis of publicly available pneumonia chest X ray images in FAKE and REAL categories. Artificial manipulations of a binary classification task to produce false medical images include blur, noise addition, brightness addition, and compression changes. The proposed framework used the MobileNetV2 framework and the pretrained ImageNet weights to learn effective feature extraction and classification. The testing accuracy was determined to be approximately 98 percent by using 720 chest X ray images in experimental testing. The findings show that lightweight transfer learning models are applicable in successfully verifying medical images and improving clinical confidence in medical imaging systems.

KEYWORDS: Medical Deepfake Detection, Chest X-Ray, Transfer Learning, MobileNetV2.

1. INTRODUCTION

Technologies based on artificial intelligence are quickly changing healthcare systems by enhancing the accuracy of diagnoses, cutting back on manual labor, and aiding quicker clinical decision making within hospitals and research institutions. Medical imaging has been a major area where AI has developed because of the vast amounts of visual data generated by radiological processes such as chest X rays, magnetic resonance imaging, and computed tomography scans which must be processed automatically. Dee learning algorithms are increasingly finding their way into health care environments to predict diseases, classify medical images, segment medical images, and monitor patients.

Despite these advantages, the use of medical deepfakes has posed serious safety and ethical concerns in the medical field. Manipulated medical images may deliberate falsify diagnostic results, and thus confuse the doctor and influence treatment. False radiology images may cause imaginary abnormalities or conceal actual diseases, as such, it is more probable to diagnose and consequently provide late medical care [1]. Hence, healthcare facilities should have powerful artificial intelligence tools that are capable of identifying the tampered medical content.

The existing medical deepfake detection algorithms tend to be based on computationally expensive architectures, which are inapplicable to lightweight models. The present paper therefore provides a convolutional neural network-based Medical deepfake detector model using MobileNetV2 transfer learning [2]. The simulated fake images of the chest x rays were done using blur, noise, brightness and compression manipulations. Around 98 percent classification accuracy was reached on 720 testing images with experimental testing.

2. RELATED WORK

Convolutional neural networks have contributed significantly to the studies in contemporary medical imaging because they are able to identify intricate patterns within radiological images without human involvement. More medical applications of artificial intelligence algorithms to identify pneumonia, diagnose tumors, analyze diabetic retinopathy, and segment organs are emerging [3]. The best medical imaging technologies improve the efficiency of the diagnostic process by reducing the human factor and maximizing prediction accuracy in any clinical application. The study of chest X ray analysis is one of the most researched fields as respiratory diseases demand quick diagnosis and constant image analysis.

Artificial intelligence techniques to identify manipulated digital media and deepfake pictures have also been examined by researchers. Deepfake Detection algorithms usually analyze discrepancies in texture, lighting, spatial artifacts, and pixel distributions created in the process of manipulation of images. A number of studies use convolutional neural networks, recurrent neural networks, and transformer architectures to differentiate authentic and manipulated content. The healthcare environment has placed more relevance on the detection of deepfakes as the technology has the capacity to compromise the accuracy of the diagnostic process and patient safety in the event of falsifying images [4].

The transfer learning techniques have performed well in the image classification tasks given the pretrained models that reduce the computational complexity and training time. Medical image analysis systems typically employ systems based on MobileNetV2, ResNet50, EfficientNet, and DenseNet models [5]. The dense layers plus the sigmoid activation functions are often used in research studies of binary classification of authentic and manipulated images.

However, a majority of the recent studies rely on large-scale computational systems and huge datasets that cannot be implemented in lightweight clinical practice. There are also frameworks, which are more preoccupied with facial deepfakes, as opposed to manipulation of radiological images. Also, there are no studies studying effective medical deepfake detectors that may be implemented into a real healthcare environment [6].

Deep Learning in Chest X-ray Analysis

The recent research demonstrates deep learning as effective in the analysis of chest X-ray. CNNs are trained to identify hierarchical features and contribute to the identification of pneumonia, tuberculosis, COVID-19, and other thoracic diseases. This is however performance dependent on the size of the dataset, image quality and the variation in the population and requires very robust approaches [7].

CNN-Based Medical Image Classification

The CNN models, such as ResNet, DenseNet, VGG, and EfficientNet, have shown a good performance in terms of medical image classification. They can identify complicated patterns, which facilitates an analysis. More sophisticated architectures, however, require a lot of computational power and big data, which are challenging to resource-constrained healthcare [8].

Transfer Learning and Lightweight Models

Transfer learning trains the CNN representations on a pretrained model and transfers the learnt features to medical classification, reducing training and computation time. MobileNetV2 is interesting because depthwise separable convolution and inverted residual networks reduce the number of calculations and still have the feature extraction capabilities. Slim models have the capability to strike a balance between efficiency and performance [9].

Medical Image Manipulation and Clinical Security

Limited efforts have been made to detect altered medical images deliberately. Brightness, noise, texture, and changes in compression can potentially affect radiological interpretation. The existing literature of deepfakes depicts CNN-based detection of visual artefacts, but most of the literature is of the face and not of medical radiology. Special medical image integrity methods are required. Explainable AI has the ability to promote transparency and trust [10].

Research Gap and Proposed Framework

The existing literature shows progress of medical image classification, transfer learning, and deepfake detection but it lacks studies that integrate these areas into lightweight medical image verification. The proposed framework of MobileNetV2 aims to fill this gap by separating authentic and manipulated chest X-rays that are processed by preprocessing, transfer learning, binary classification, and evaluated by accuracy, precision, recall, F1-score, and confusion analysis based on the confusion matrixes [11].

3. PORPOSED MODELLING

3.1. System Architecture

The proposed system architecture is a sequential medical deepfake detection pipeline that will strive to attain the optimal classification performance. The ready dataset is first scanned into chest X ray images which are then transmitted to the preprocessing module. The pictures are subjected to resizing and normalization and then inputted into the MobileNetV2 transfer learning model. The features extracted are fed to the dense classification layers in which the sigmoid activation is used [12].

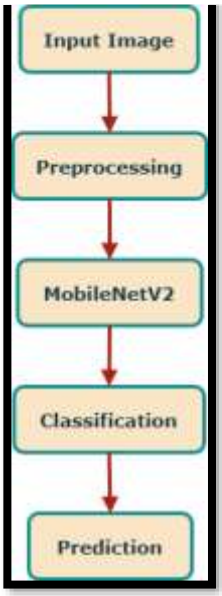

Figure 1 Proposed medical deepfake detection framework architecture

3.2. Dataset Preparation

The experimental data were built on the pictures of the chest x ray found in the publicly available pneumonia imaging repository. The data were divided into training, validation and testing datasets of REAL and FAKE image folders. Original chest X ray scans were termed as REAL images [13]. FAKE samples were made by artificial manipulation of images as FAKE with Gaussian blur, addition of brightness noise, and compression changes. The last dataset consisted of 1200 training images, 32 validation images and 720 testing images used in the experimental evaluation.

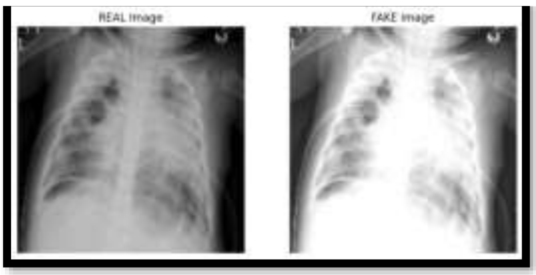

Figure 2 Sample REAL and FAKE chest X ray images used for experimental analysis

3.3. Image Preprocessing

Preprocessing on the images was done to enhance consistency of training and maximize the performance of the models when using them to classify the images. The model was trained using all the chest X ray images which were resized to standard resolution of 224 x 224. The pixels were normalized by the rescaling values of zero to one making the process of feature extraction stable [14]. There was limited use of data augmentation techniques as manipulated versions of the data have already been generated using the blur, brightness, compression, and random noise addition techniques.

3.4. Deep Learning Model

The proposed medical deepfake detector employed the MobileNetV2 convolutional neural network architecture because it has reduced computational architecture and usefulness in features extraction. Transfer learning was done with pretrained ImageNet weights but omitting the original classification layers

[15]. The pretrained feature extraction component was fixed during the training, in order to reduce the computational complexity and overfitting. Global average pooling and dropout layers were introduced before the last dense classification layer. Sigmoid activation was also used to classify REAL and FAKE with binary classification.

3.5. Conceptual Clinical Assurance Module

The conceptual clinical assurance module was to help healthcare professionals in medical image verification processes. The framework can theoretically generate scores of confidences that will reflect the reliability of the prediction of every classified chest X ray picture [16]. Verification with the help of AI has a potential to help radiologists identify suspicious medical images before diagnostic interpretation. The implementation in the future in distributed hospital settings would facilitate safe clinical decision making and mitigate risks related to manipulated medical imaging content and unapproved diagnostic changes.

3.6. Experimental Setup and Dataset

The experimental structure was coded in Python language in the Jupyter Notebook platform. TensorFlow and Keras libraries were used to create deep learning models and transfer learning. Image preprocessing, manipulation, and visualization was applied using open CV in the whole process of experimentation. NumPy, Matplotlib and Scikit learn other libraries helped with the numerical analysis, graphical representation and performance analysis capabilities [17].

Table 1 Dataset distribution

Dataset Split	REAL	FAKE	Total
Train	600	600	1200
Validation	16	16	32
Test	360	360	720

The medical deepfake dataset of trained models consisted of two classification levels namely REAL and FAKE. The training data were 1200 images (600 REAL and 600 FAKE). The validation dataset had 32 images (16 REAL and 16 FAKE) and the testing dataset had 720 images (360 REAL and 360 FAKE). Training was done by first resizing all the images to 224 x 224 [18]. The images were trained and tested with a 224 x 224 resolution.

Ten epochs of experimental training were performed with a batch size of 32. To converge effectively, Adam optimization algorithm was selected using a learning rate of 0.001. Binary cross entropy loss function was used to learn classification with cross entropy loss function.

Table 2 Model training parameters

Parameter	Value
Image Resolution	224 × 224
Batch Size	32
Epochs	10
Optimizer	Adam
Learning Rate	0.001
Loss Function	Binary Crossentropy

Accuracy, precision, F1 score and confusion matrix analysis were used to evaluate a model performance. The evaluation through experimentation was around 98 percent classified on testing.

4. RESULTS AND DISCUSSIONS

In this section all the results and the discussions should be made.

4.1. Training Performance

Stable learning performance was shown by the training process on ten experimental epochs. The accuracy of the training began at about 78 percent, and steadily improved over the course of learning until it peaked at about 97 percent in later epochs. The validation accuracy was constantly around 100 percent due to the lack of samples of images in the validation dataset.

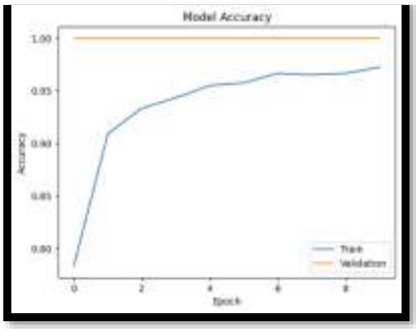

Figure 3 Training and validation accuracy during model training

Simultaneously, both the loss of training values decreased to a smaller range of 0.45 to under 0.10 indicating improved classification and reduced errors during prediction [19]. The experimentation also recorded relatively low values of validation losses. The resulting accuracy and loss curves provided the same convergence and transfer learning in training.

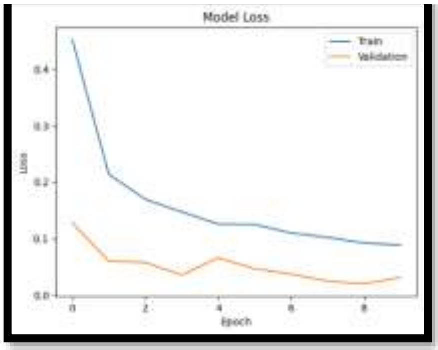

Figure 4 Training and validation loss during experimental learning

4.2. Classification Results

The trained MobileNetV2 structure was shown to have good classification accuracy in the test of 720 X ray images of the chest. The analysis of the confusion matrices revealed that both REAL and FAKE categories had an equal predictive power. The system was able to correctly classify 352 REAL images and 352 FAKE images with only eight misclassified images per category. The values of precision, recall, and F1 score of REAL and FAKE classes were about 0.98, which suggests that predictive reliability is similar in terms of manipulated and authentic medical images. The examples of visual prediction also demonstrated the capacity of the framework to distinguish between brightness manipulated, noisy and blurred chest X ray images, and original medical scans.

Table 3 Classification Performance Metrics

Metric	REAL	FAKE
Precision	0.98	0.98
Recall	0.98	0.98
F1-Score	0.98	0.98

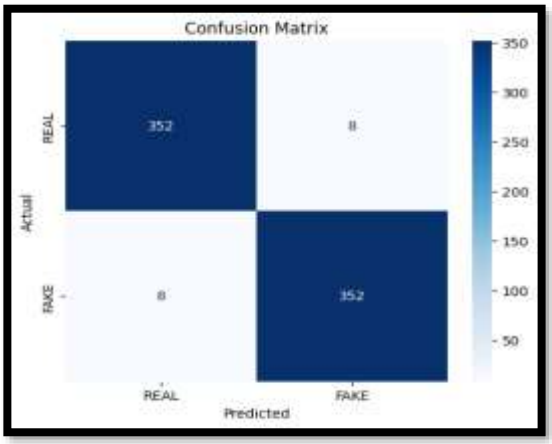

Figure 5 Confusion matrix

4.3. Discussion

The trained MobileNetV2 model was found to achieve good results in respect of classification when it was tested on 720 chest X ray images. Confusion matrix analysis demonstrated that there was no differing predictive power between REAL and FAKE. The system had 352 REAL images and 352 FAKE images that were correctly classified and only eight misclassification was made in each category [20]. The precision, recall, and F1 of REAL and FAKE classes were about 0.98, indicating that manipulated and authentic medical images are equally predictable. The visual forecasts also indicated that the framework was able to distinguish between brightness manipulated, noisy, and blurred images of chest X rays and original medical images.

5. CONCLUSION

Conclusion: The paper presented a light medical deepfake detector that was trained over the MobileNetV2 transfer learning model to detect manipulated chest X-ray images. The proposed procedure included the image preprocessing and binary classification techniques to distinguish between REAL and FAKE medical images generated due to simulated manipulations, including blur, brightness adjustment, noise addition, and compression. Experimental test offered an accuracy of testing of about 98-percent that suggested the proposed framework may be utilized well in the detection of manipulated medical images with low degree of computational complexity. The implications of these findings include that the lightweight transfer learning models could be a possible solution to improving the reliability and security of medical imaging systems.

Future Work: The given framework demonstrated good results on simulated image manipulations, but additional studies can enhance its application in real healthcare settings. Future research opportunities may be to test the model on more heterogeneous medical images, use more complex deepfake generation algorithms, and look at explainable artificial intelligence (XAI) methods to increase transparency of the model and clinical trust. Additionally, it can be further extended through support to real-time deployment and validation to multiple healthcare organizations, thereby enhancing its usability and applicability to medical image security.

REFERENCES

[1] Dhar, T., Dey, N., Borra, S. and Sherratt, R.S., 2023. Challenges of deep learning in medical image analysis—improving explainability and trust. IEEE Transactions on Technology and Society, 4(1), pp.68-75.
[2] Conze, P.H., Andrade-Miranda, G., Singh, V.K., Jaouen, V. and Visvikis, D., 2023. Current and emerging trends in medical image segmentation with deep learning. IEEE Transactions on Radiation and Plasma Medical Sciences, 7(6), pp.545-569.
[3] Mahmood, T., Rehman, A., Saba, T., Nadeem, L. and Bahaj, S.A.O., 2023. Recent advancements and future prospects in active deep learning for medical image segmentation and classification. IEEE Access, 11, pp.113623-113652.
[4] Prabha, R., Razmah, M., Sridevi, S., Roopa, D. and Asha, R.M., 2022, May. A big wave of deep learning in medical imaging-analysis of theory and applications. In 2022 6th International conference on intelligent computing and control systems (ICICCS) (pp. 1321-1327). IEEE.
[5] Malik, A., Kuribayashi, M., Abdullahi, S.M. and Khan, A.N., 2022. DeepFake detection for human face images and videos: A survey. Ieee Access, 10, pp.18757-18775.
[6] Rana, M.S., Nobi, M.N., Murali, B. and Sung, A.H., 2022. Deepfake detection: A systematic literature review.

IEEE access, 10, pp.25494-25513.

[7] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I.E. and Mazibuko, T.F., 2023. An improved dense CNN architecture for deepfake image detection. IEEE Access, 11, pp.22081-22095.

[8] Elhanashi, A., Saponara, S. and Zheng, Q., 2023. Classification and localization of multi-type abnormalities on chest X-rays images. IEEE Access, 11, pp.83264-83277.

[9] Hussain, A., Amin, S.U., Lee, H., Khan, A., Khan, N.F. and Seo, S., 2023. An automated chest X-ray image analysis for covid-19 and pneumonia diagnosis using deep ensemble strategy. IEEE Access, 11, pp.97207-97220.

[10] Malik, H., Anees, T., Chaudhry, M.U., Gono, R., Jasiński, M., Leonowicz, Z. and Bernat, P., 2023. A novel fusion model of hand-crafted features with deep convolutional neural networks for classification of several chest diseases using X-ray images. IEEE Access, 11, pp.39243-39268.

[11] Muhammad Asad Arshed, S. Mumtaz, Ștefan Cristian Gherghina, Neelam Urooj, S. Ahmed, and C. Dewi, "A Deep Learning Model for Detecting Fake Medical Images to Mitigate Financial Insurance Fraud," Computation, vol. 12, no. 9, pp. 173–173, Aug. 2024, doi: 10.3390/computation12090173.

[12] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review," Wiley interdisciplinary reviews. Data mining and knowledge discovery/Wiley interdisciplinary reviews. Data mining and knowledge discovery, vol. 14, no. 2, Nov. 2023, doi: 10.1002/widm.1520.

[13] A. Raza, K. Munir, and M. Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection," Applied Sciences, vol. 12, no. 19, p. 9820, Sep. 2022, doi: 10.3390/app12199820.

[14] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, "Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning," Information, vol. 16, no. 4, pp. 270–270, Mar. 2025, doi: 10.3390/info16040270.

[15] R. Siddiqi and S. Javaid, "Deep Learning for Pneumonia Detection in Chest X-Ray Images: A Comprehensive Survey," Journal of Imaging, vol. 10, no. 8, pp. 176–176, Jul. 2024, doi: 10.3390/jimaging10080176.

[16] E. Khan, M. Z. U. Rehman, F. Ahmed, F. A. Alfouzan, N. M. Alzahrani, and J. Ahmad, "Chest X-Ray Classification for the Detection of COVID-19 Using Deep Learning Techniques," Sensors, vol. 22, no. 3, p. 1211, Feb. 2022, doi: 10.3390/s22031211.

[17] Z. Mustafa and Heba Nsour, "Using Computer Vision Techniques to Automatically Detect Abnormalities in Chest X-Rays," Diagnostics, vol. 13, no. 18, pp. 2979–2979, Sep. 2023, doi: 10.3390/diagnostics13182979.

[18] Shubham Mathesul et al., "COVID-19 Detection From Chest X-Ray Images Based on Deep Learning Techniques," Algorithms, vol. 16, no. 10, pp. 494–494, Oct. 2023, doi: 10.3390/a16100494.

[19] C. Gu and M. Lee, "Deep Transfer Learning Using Real-World Image Features for Medical Image Classification, With a Case Study on Pneumonia X-Ray Images," Bioengineering, vol. 11, no. 4, p. 406, Apr. 2024, doi: 10.3390/bioengineering11040406.

[20] M. Aleem et al., "Seeing Through the Fake: Explainable AI With Multiple CNNs for Deepfake Detection," IEEE Access, vol. 14, pp. 131–162, 2026, doi: 10.1109/access.2025.3649