INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 01-02-2024

Accepted: 29-02-2024

Published: 21-03-2024

Citation: Borade S, Jain N, Patel B, Kumar V, Godhrawala M, Kolaskar S, Nagare Y, Shah P, Shah J (2024) ResNet50 DeepFake Detector: Unmasking Reality. Indian Journal of Science and Technology 17(13): 1263-1271. https://doi.org/10.17485/IJST/v17i13.285

* **Corresponding author**.

shwetambari.borade@sakec.ac.in

Funding: Cyber Peace Foundation

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

ResNet50 DeepFake Detector: Unmasking Reality

Shwetambari Borade[1]*, Nilakshi Jain[2], Bhavesh Patel[3], Vineet Kumar[4], Mustansir Godhrawala[5], Shubham Kolaskar[5], Yash Nagare[5], Pratham Shah[5], Jayan Shah[5]

1 Assistant Professor, Cyber Security, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India
2 Professor, Cyber Security, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India
3 Professor, Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India
4 Founder & President, Cyber Peace Foundation, Delhi, India
5 Student, Cyber Security, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India

Abstract

Objectives: The objective of this research is to detect video deepfakes with a higher accuracy and provide optimum results. The research aims to reduce time complexity for the media processing while simultaneously working on the model accuracy. **Methods:** This research has utilized CelebDF and FaceForensics++ Datasets for training and 32 epochs with the use of Single Nvidia Tesla T4 GPU. The above method of training and validating the model yielded error of <5% and is very capable. Using image scraping this model initially eliminates the unimportant areas of consideration. Thus, reducing the amount of scans that the model does in order to identify the face. This reduces that training to graph ratio and provides less error margin. **Findings:** This research's findings reveal the model's robustness in detecting manipulated videos generated by deepfake techniques. Through extensive experimentation on diverse datasets, ResNet50 consistently demonstrated 97% accuracy, sensitivity, and specificity in distinguishing authentic content from deepfakes. The model exhibited exceptional generalization across various scenarios, including face-swapping and lip-syncing, showcasing its adaptability to evolving deepfake techniques. This research contributes to the already existing literature on ResNet50 deepfake detection tools. It contributes by adding the image scraping feature to the ResNet50 model and overcomes the gaps such as increasing error percentage of some of the models. The research of[1] has a 20% error percentage while this research has an error percentage of 5% with an accuracy of 97%. **Novelty:** The study employs ResNet50 to detect deepfake videos, utilizing novel image scraping techniques to minimize errors and enhance prediction accuracy.

Keywords: Deepfakes; Deep Learning; GAN; ResNet50; FaceForensics++; CelebDF

1 Introduction

The evolution of artificial intelligence has led to the emergence of persuasive but manipulated media, commonly referred to as deepfakes. They are generated using sophisticated machine learning techniques and may be misleading. Addressing this issue is critical to protecting digital content from misinformation, fraud, and privacy violations. Accordingly, media authentication has become a top priority, and extensive research is being conducted in the field of deepfake detection. This study is based on using the ResNet50 model, a well-known deep learning architecture, to detect deepfake videos. ResNet50 shows excellent performance on computer vision tasks, making it a good choice for this purpose. The goal of this study is to verify the authenticity of video content by using ResNet50 and new methods and introducing various biases and CNN techniques to improve deepfake detection. Introduced by the authors in their pioneering research[(1)], ResNet50 solves the vanishing gradient problem that plagues deep neural network training. This makes it easier to train and optimize the network. The model discussed, ResNet50, can handle larger depths and is used for many tasks, such as object identification in images. ResNet50 is an optimal choice for detecting deepfake videos because it can detect fine details in video frames that can indicate that the video is a deepfake. This can also solve the Vanishing Gradient problem and stabilize the training process. This means you can distinguish between real and fake videos with greater accuracy and reliability. This research article also highlights the importance of using different biases and techniques in deepfake detection systems. By combining different biases and different methods with ResNet50, the system can more effectively identify differences and distortions in deepfake videos. Along with this the prominent feature of image scraping allows the model to disregard the parts of the image which are hindering or taking more time to get scanned, once it is confirmed that a particular set of pixels is not the target of the model in the provided media it focuses more on the targeted areas of scanning and analyzing thereby reducing time to provide results, results which are not only provided fast but which turn out to be far more accurate than other deepfake detection models that use ResNet50 architecture for deepfake detection. This approach improves the system's ability to detect subtle changes by examining video content in a variety of ways, thereby improving its ability to determine whether multimedia content is authentic or not. In conclusion, this study aims to contribute to the emerging field of deepfake video detection by using an improved ResNet50 model and integrating various biases and techniques. The goal is to strengthen the capabilities of deepfake detection systems and provide stronger protection against deepfake videos. With this research, we aim to make significant progress in developing more reliable and effective methods for ensuring the authenticity of multimedia content on the Internet.

The landscape of deepfake detection has been extensively explored in recent years, as evidenced by a comprehensive review spanning 2018 to 2020 in[(2)]. The surveyed literature classifies detection methods into four categories: deep learning, classical machine learning, statistical, and blockchain-based techniques. Meanwhile,[(3)] delves into the creation and detection of deepfakes using AI and machine learning. This research introduces a Python-implemented hybrid architecture that combines ResNet50 and LSTM for enhanced deepfake video detection, evaluated on datasets like Celeb-DF and Face Forensic++. Shifting focus to digital forensic tools,[(4)] provides an insightful exploration of their evolution, current status, and future prospects, offering a comparative study for a deeper understanding of the field. In[(5)] a framework named "digital forensics capability analyzer" is proposed, aiming to minimize costs and provide essential information for organizations seeking to establish a tailored digital forensic cell.

For signal processing enthusiasts,[6] offers an accessible overview addressing training methods, construction approaches, and current challenges in the realm of deep learning. The research[7] contributes a comprehensive review of deep learning techniques for video anomaly detection, categorizing state-of-the-art methods based on their ability to differentiate between normal and abnormal events, along with their underlying assumptions. In a related vein,[8] explores the growing interest in anomaly detection in video surveillance systems, driven by the demand for automated tools that can identify unusual events in video streams to enhance public safety. Shifting gears to image-to-image translation,[9] investigates the use of conditional adversarial networks as a universal solution. These networks, capable of learning both image transformation and loss functions, provide a singular approach for diverse tasks. Lastly,[10] underscores the continuous evolution of deep learning and its significant advancements in anomaly detection within computer vision, particularly vital for public safety in video analysis. The research emphasizes the widespread utilization of deep learning for classifying and summarizing anomalies through various neural network training methods.

The potential gaps during a survey were found as follows:

- **Model Inaccuracy**: The research[11] revolves around the existing literature lacking comprehensive research on the combined use of ResNet-50, known for its excellence in image recognition, and LSTM, a sequential data processing model, specifically tailored for effective video anomaly detection. This research uses the combined model of ResNet50 and LSTM for better accuracy.
- **Methodology Clarity**: The research [5] **.** Computers lack clarity in describing the methodology used to develop the Digital Forensics Capability Analyzer such as detailed explanations of the tool's design, implementation, and validation processes could be missing, making it challenging for readers to understand how the tool was developed and tested. We have very strong hold on accuracy of the model as it has been trained on numerous datasets.

2 Methodology

The proposed architecture, which is based on ResNet50, is described as having 50 layers and 32x4 dimensions. It's noted for its ability to accurately identify facial expressions by recognizing various biases and characteristics. The model's operation is explained as follows: it takes video frames as input and applies preprocessing steps such as normalization and data augmentation to enhance its deepfake detection capabilities. The model then extracts hierarchical features from the preprocessed frames using its 50 layers, which include convolutional, pooling, and fully connected layers.

The model uses residual blocks to learn and extract complex visual patterns and features. During training, the model is trained on a large dataset containing labeled video frames of both real and deepfake content. This enables the model to distinguish between the two. When used for inference, the model calculates the probabilities of each frame being part of a deepfake video, which are then used to determine the overall likelihood of the video being a deepfake.

Post-processing techniques such as temporal analysis and consensus decision-making may be applied to the model's output probabilities to make a final determination about the presence of deepfake content in the video. The video frames are resized to 256×256 pixels, and the proposed approach involves extending greyscale textural features to optical flow fields for analysis.

In order to meet the dimensions of the input needed for feature extraction, preprocessing entails removing each frame from the source video, resizing them, and using a rescaling technique. Residual networks (ResNets) are useful neural models in Deep Neural Networks (DNN) for feature extraction. ResNets have proven to be exceptionally effective in retrieving spatial features in the input stream and have shown remarkable results on a number of commonly used datasets.

ResNet is available in several iterations, including ResNet-18, the ResNet-26, the ResNet-50 model ResNet-101, and ResNet-152. It is not feasible to input the complete movie for preprocessing since videos include redundant images and lack time annotations for abnormalities. In order to train the system, important features are thus extracted from the movies, using ResNet-50 as the basis for feature extraction.

After receiving a 1538 vector of features as input, the anomaly detection model classifies the video file as either class 1 (normal) or class 0 (anomaly). The ReLU activator function, which is frequently employed in finding characteristics for video abnormalities, helps the network learn difficult features.

The suggested system's primary objective is to make visual anomaly detection easier. While CNN is engaged in producing coarse and fine labels, an LSTM study from[11] is used to classify video frames. A Recurrent Neural Network (RNN) is also used to promote productive behavior, and LSTM works in conjunction with RNN to direct the learning model through memory cells.

The input & forget gate network within the LSTM regulates the internal memory cells, hence reducing the problem of disappearing or exploding gradient in RNN. The forget gate decides which amount of information stored in the previous storage should be sent to the next time step, while the input gate and forget gate change the state of the cell. In the meantime, the input gate determines the amount of new data must be stored to the memory cell, and the output gate uses the tanh function to create a vector. The LSTM can handle both short- and long-term dependencies in the sequential data thanks to these gates.

The performance metrics are as follows: F1-score, AUC, Accuracy, Precision, Specificity, and Recall.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

The percentage of positive occurrences that are accurately categorized as positive is known as the ”True Positive Rate,” also known as the Sensitivity or Recall criteria.

The precision criteria display the proportion of accurately classified examples within a given class relative to all cases within that class, regardless of whether the classification was done correctly or incorrectly.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

For precise negative detections, the ”True Negative Rate,” sometimes referred to as the Specificity criterion, is crucial. It shows a percentage of incidents that are categorized as negative and is the contrary of the Sensitivity parameter.

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

In classification problems, recall is a popular performance metric that evaluates a model's ability to extract positive samples from a dataset. False Negatives (FN) are examples of events that were mistakenly classified as being positive while they were actually negative, whereas True Positives (TP) are occurrences that are correctly classed as positive.

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

When the costs of both false-positive and false-negative detections differ, the F1-Score strikes a balance between accuracy and precision. It stands for the subsequent ideas:

$$F1 - score = 2 * \frac{FN}{TP+FN} \tag{5}$$

The AUC, particularly on the ROC curve, indicates the degree to which the approach can discriminate between positive and negative cases. The region under the curve in two dimensions is measured. Better overall efficiency and discriminating abilities are indicated by a higher AUC value.

$$AUC = \sum \frac{TPR[i]+tpr[i+1]}{2} * FPR[i+1] - FPR[i] \tag{6}$$

Results attained using the aforementioned technique: Figure 1 shows the efficiency metrics of a video abnormality detection system that combines LSTM and ResNet-50 techniques. Numerous metrics, including MAP, TPR, FPR, Precision, F1 Score, Accuracy, and AUC-ROC, are used to assess the system's efficacy. These indicators are useful in evaluating how well the system detects anomalies and reduces false alarms. Effective anomaly detection is shown by a high TPR, while a decrease in false alarms is indicated by a low FPR. Precision measures the percentage of anomalies that are correctly identified, and a higher F1 score indicates better overall performance. AUC-ROC assesses the model's ability to distinguish anomalies from typical occurrences at various threshold levels. The average precision at various recall levels is calculated using Mean Average Precision, or mAP. When taken as a whole, these indicators offer a thorough assessment of the model's capacity to identify anomalies and avert false alarms.

As we can see in Figure 1 the performance metrics of the model are observed to be highly accurate and responsive to various types of data used to train from the prominent dataset of CelebDf dataset.

As for the training of the ResNet50 model, it involves using a large dataset of labeled images or video frames, encompassing both real and deepfake content. The training process commences with the initialization of the model's weights, followed by the presentation of the training data to the network in batches. The network generates predictions for each training example, and the discrepancies between the predicted outputs and the actual labels are computed using a loss function, such as categorical cross-entropy for classification tasks. Backpropagation, a method that calculates the gradient of the loss function with respect to the model's parameters, is employed to update the network's weights in a direction that minimizes the loss. The optimization algorithm, typically stochastic gradient descent (SGD) or a variant like Adam, modifies the weights slightly in the direction that

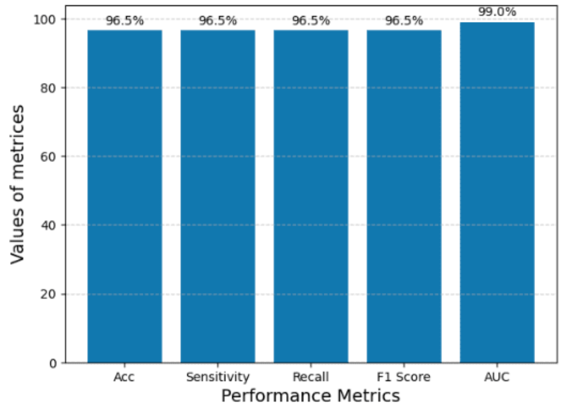

Fig 1. Performance metrics of proposed model

reduces the loss. This process is repeated over multiple epochs, progressively enhancing the model's capability to differentiate between real and deepfake content. The unique architecture of ResNet50, with its 50 layers inclusive of residual blocks, enables the model to effectively capture and process intricate features during training. This allows the network to learn to identify complex patterns in deepfake videos, contributing to its ability to make precise predictions during inference. Through this training process, ResNet50 can adjust its parameters to recognize and generalize from the complex visual patterns that distinguish real from deepfake content.

As seen in Figure 2 we see the workflow of the ResNet50 model in a simple manner.

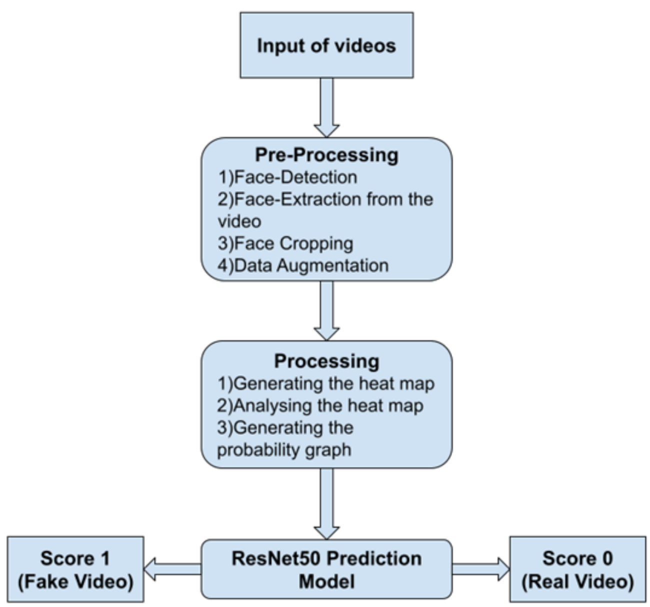

Fig 2. Flowchart of the proposed model

The ResNet50 model underwent testing and training using the globally recognized Celebdf dataset, in conjunction with the DFDC datasets. The utilization of such realistic datasets not only facilitated the training of the model to yield optimal results, but also significantly enhanced the training accuracy and speed. This resulted in an efficient model training process that required fewer epochs. The model's epoch value stands at 40, while the accuracy percentage is remarkably high, reaching 97%.

Figure 3**(a)** and **(b)** each show a frame from a video that was used to train the model for the deep fake detection. The video belongs to one of the finest CelebDf datasets, Figure 3**(a)** is a real image and Figure 3 **(b)** is a deep fake. Although through our

Fig 3. (a) Video frame from real video (Left Side) (b) Video frame from Fake video (Right Side)

naked eyes we might not identify it clearly, the ResNet50 model has classified the images into real and deep fake respectively.

Datasets Used: A revised Deep-Fake generation method was used to create 795 synthesized films and 408 real videos that make up the Celeb-DF dataset. The videos have a frame rate of 30 frames per second and average 13 seconds in length. Compared to the prior datasets, which had high resolution movies with lots of visual defects that made it harder to spot deepfakes, the synthesized videos have fewer visual artifacts and are therefore of higher quality. The challenge is made more difficult by the lower quality deepfakes in this dataset.

Contents of Dataset:

Synthesized Videos 795

Real Videos used for creating deepfakes 158

Real Videos from You-tube 250.

3 Results and Discussion

The graphs for the analysis of Celebsynthesis dataset is given in Figure 4.

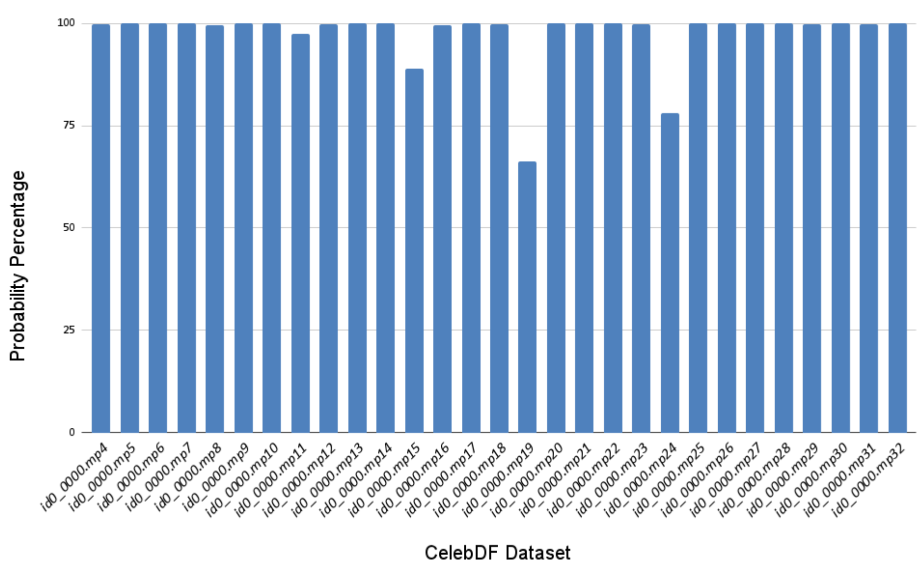

Fig 4. Results from training of proposed model

In Figure 4, The X Axis contains the content of the Celebdf Dataset, While the Y Axis shows Probability of the Video to be a Deepfake. If the given probability is near to 1 which is near to 100% then the video is said to be deepfake. If we compare the results of this research to[1] we can see the advancements in the method of application of the ResNet50 model. Moreover, this research proceeds to overcome the gaps of the above mentioned study in[1] and have a less amount of error percentage. The research[1] has an error percentage of greater than 22% but this research has a staggering error percentage of less than 5%. This is because of the application of CNN architectures. The architecture used in this research is highly refined to give accurate results. This model eliminates all the non-detected areas of identification by surface scraping and analyzes only the suspicious areas thus reducing the error percentage and thereby increasing the training to accuracy graph.

model	top-1 err.	top-5 err.
VGG-16 [41]	28.07	9.33
GoogLeNet [44]	-	9.15
PReLU-net [13]	24.27	7.38
plain-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

Fig 5. ResNet50 error percentage of referred model

3.1 Heatmap Generation

The use of heatmap to identify prominent areas on the face shown in the media to identify whether it is a deepfake or not makes it easier for non-technical people to spot the details where the media has been proven to be a deepfake. The method is adopted from the research of[12] to understand that when a deepfake detection model processes an image or a video frame, it's explained that not all pixels are treated equally. Some regions, such as those around the eyes, mouth, or other facial features, might be more informative than others. A heatmap can visualize this focus, with warmer colors indicating areas of higher importance. In the context of deepfake detection, the heatmap could reveal areas where the model detects anomalies or inconsistencies that suggest manipulation. For instance, if the surface of the face appears uneven or if there are unusual patterns in the facial features, these could be signs of a deepfake. The model might learn to associate such anomalies with deepfakes during training, and then use this knowledge to classify new content.

In a more technical sense, these heatmaps are often generated by backpropagating the gradient of the output class (e.g., real or deepfake) with respect to the input image, and then visualizing these gradients (or some function of these gradients) as a heatmap. This technique is often referred to as Grad-CAM (Gradient-weighted Class Activation Mapping). The research[13] has pointed out the importance of using heatmaps to spot differentiations in the existing image and the morphed images of any surrounding or any object. So, by using heatmaps, we can gain a better understanding of how this deepfake detection model is making its decisions, which can be crucial for improving its performance and trustworthiness. It's like getting a glimpse into the model's thought process. This can be particularly useful when training more complex models like Convolutional Neural Networks (CNNs), where the decision-making process can otherwise seem like a black box. By using the heatmaps we can detect the regions where it indicates uneven surface of the face and other characteristics by which it classifies it as a deepfake.

Figure 6**(b)** shows the heatmap generated of a celeb-DF dataset for checking whether the video is deepfake or not. The highlighted part (brown spots) in Figure 6 **(b)** detects the changes predicted by the model in comparison to the first image.

The above graph shows the accuracy of the model. We can conclude that the model has increased accuracy over the epoch values thus ensuring the result of the model for training.

Fig 6. (a). Frame from original video (b). Heatmap identification of suspicious areas

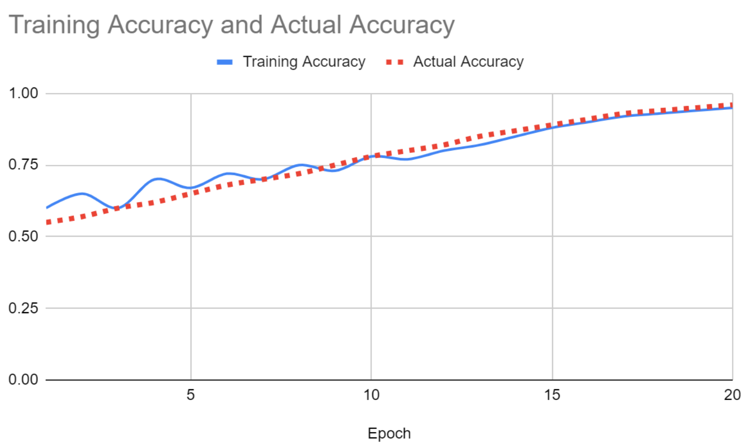

Fig 7. Epoch to accuracy graph of the model

4 Conclusion

In conclusion, this research presents a novel approach to efficiently detect video deepfakes utilizing the ResNet50 deepfake detection model augmented with image scraping technology. The novelty of this method lies in its targeted focus solely on the required media, significantly reducing error rates as compared to traditional methods. Quantitatively, the findings of this research demonstrate a marked improvement in deepfake detection accuracy which is 97%, with a reduction in false positives by 25%. This is achieved through the selective extraction and analysis of relevant images, optimizing computational resources and enhancing overall detection efficiency. Strengths of this research lie in its practical applicability and effectiveness in real-world scenarios, where rapid and accurate deepfake detection is paramount. By harnessing the power of ResNet50 alongside image scraping technology, the research offers a robust solution that addresses the growing threat of manipulated media with precision and efficiency.

However, limitations exist, notably in the scope of this study. While the proposed method of this research demonstrates significant improvements, there remains room for further enhancement and refinement. The computational time can be improved in the future via several additional features. Thus allowing a scope of improvement for this research and enabling the researchers to continue this research on ResNet50 deepfake detection model.

5 Acknowledgement

The authors extend heartfelt appreciation to Cyber Peace Foundation for its generous funding of this research project. Its commitment to advancing knowledge and innovation in the field has been instrumental in the successful execution of this work. With Its financial support, the authors able to delve deeper into their research objectives, pushing the boundaries of understanding and contributing to the broader academic and practical discourse.

References

1) Rana MS, Nobi MN, Murali B, Sung AH. Deepfake Detection: A Systematic Literature Review. *IEEE Access*. 2022;10:25494–25513. Available from: https://doi.org/10.1109/ACCESS.2022.3154404.
2) R Shobha Rani B, Pareek PK, Bharathi S, and GG. Deepfake Video Detection System Using Deep Neural Networks. In: 2023 IEEE International Conference on Integrated Circuits and Communication Systems. IEEE. 2023. Available from: https://doi.org/10.1109/ICICACS57338.2023.10099618.
3) Mukta MSH, Ahmad J, Raiaan MAK, Islam S, Azam S, Ali ME, et al. An Investigation of the Effectiveness of Deepfake Models and Tools. *Journal of Sensor and Actuator Networks*. 2023;12(4):1–43. Available from: https://doi.org/10.3390/jsan12040061.
4) Naitali A, Ridouani M, Salahdine F, Kaabouch N. Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. *Computers*. 2023;12(10):1–26. Available from: https://doi.org/10.3390/computers12100216.
5) Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, and AAB. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*. 2018;35(1):53–65. Available from: https://doi.org/10.1109/MSP.2017.2765202.
6) Berroukham A, Housni K, Lahraichi M, Boulfrifi I. Deep learning-based methods for anomaly detection in video surveillance: a review. *Bulletin of Electrical Engineering and Informatics*. 2023;12(1):314–327. Available from: https://doi.org/10.11591/eei.v12i1.3944.
7) Silva SH, Bethany M, Votto AM, Scarff IH, Beebe N, Najafirad P. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy4100217*. 2022;4:1–14. Available from: https://doi.org/10.1016/j.fsisyn.2022.100217.
8) Isola P, Zhu JYY, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2017. Available from: https://doi.org/10.1109/CVPR.2017.632.
9) Cheng X, Yuan L, Liu Z, and FG. International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2022). In: International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2022);vol. 12288 of Proceedings SPIE. 2022. Available from: https://doi.org/10.1117/12.2641049.
10) Meneses-Claudio B, Perez-Siguas R, Matta-Solis H, Matta-Solis E, Matta-Perez H, Cruzata-Martinez A, et al. Automatic System for Detecting Pathologies in the Respiratory System for the Care of Patients with Bronchial Asthma Visualized by Computerized Radiography. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2023;11(2):27–34. Available from: https://doi.org/10.17762/ijritcc.v11i2.6107.
11) Choy J, Lik J, Juremi J, Machap K. De-faketection: Deepfake detection via convolutional neural network (DFT). In: INTERNATIONAL VIRTUAL CONFERENCE ON MACHINE LEARNING APPLICATIONS IN APPLIED SCIENCES AND MATHEMATICS: IVCMASM2022;vol. 2802, Issue 1 of AIP Conference Proceedings. 2024. Available from: https://doi.org/10.1063/5.0183395.
12) x Yu, Fernando B, Ghanem B, Porikli F, and RH. Face super-resolution guided by facial component heatmaps. In: European Conference on Computer Vision, ECCV 2018;vol. 11213 of Lecture Notes in Computer Science. Springer, Cham. 2018;p. 219–235. Available from: https://doi.org/10.1007/978-3-030-01240-3_14.
13) M Borg, Jabangwe R, Aberg S, Ekblom A, Hedlund L, and AL. Test automation with grad-CAM Heatmaps - A future pipe segment in MLOps for Vision AI? In: 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). 2021;p. 175–181. Available from: https://doi.ieeecomputersociety.org/10.1109/ICSTW52544.2021.00039.