DEEPFAKE DETECTION BASED ON DEEP LEARNING

Lai Minh Tuan*, Pham Tien Manh, Dong Thi Thuy Linh
Academy of Cryptography Techniques

ARTICLE INFO

Received: 14/9/2023
Revised: 24/10/2023
Published: 25/10/2023

KEYWORDS

Deepfakes
Deepfake detection
Deep learning
Transfer learning
Convolutional neural networks

ABSTRACT

The spread of deepfake images and videos in cyberspace is a threat to information security area. Besides, in recent years, deep learning has been increasingly developed and widely applied in many fields, and algorithms have improved performance and accuracy. Hence, the application of deep learning in deepfake detection is a practical research direction. However, applying deep learning models requires a lot of data to solve the problem effectively and takes lots of time to perform training. There is a commonly applied method that helps improve accuracy and takes advantage of pre-trained models with good quality and high accuracy called transfer learning. This study introduces an approach to detect deepfake images using transfer learning methods, including XceptionNet, RestNet101, InceptionResV2, MobileNetv2, VGG19 and DenseNet121, along with comparing it with a traditional CNN model. Through experiments on the Celeb-DF dataset, we demonstrate that DenseNet121 and the softmax classifier perform better than the rest of the methods.

PHÁT HIỆN DEEPFAKE DỰA TRÊN HỌC SÂU

Lại Minh Tuấn*, Phạm Tiến Mạnh, Đồng Thị Thuỳ Linh
Học viện Kỹ thuật mật mã

THÔNG TIN BÀI BÁO

Ngày nhận bài: 14/9/2023
Ngày hoàn thiện: 24/10/2023
Ngày đăng: 25/10/2023

TỪ KHÓA

Deepfakes
Phát hiện deepfake
Học sâu
Học chuyển giao
Mạng tích chập

TÓM TẮT

Việc các hình ảnh, video deepfake tràn lan trên không gian mạng đang trở thành nguy cơ đe dọa an toàn, an ninh thông tin trong kỷ nguyên số. Bên cạnh đó, trong những năm gần đây học sâu ngày càng phát triển và ứng dụng rộng rãi, các thuật toán ngày càng được cải thiện hiệu suất và độ chính xác đáng kể. Do đó, việc áp dụng bài toán học sâu vào việc phát hiện deepfake là một hướng nghiên cứu thiết thực. Tuy nhiên, khi áp dụng các mô hình học sâu đòi hỏi rất nhiều dữ liệu để giải quyết bài toán hiệu quả cũng như cần nhiều thời gian để thực hiện việc huấn luyện. Một phương pháp được áp dụng phổ biến giúp cải thiện độ chính xác cũng như tận dụng được các pre-trained model có chất lượng tốt và độ chính xác cao được gọi là học chuyển giao. Nghiên cứu này giới thiệu một cách tiếp cận để phát hiện deepfake thông qua việc sử dụng phương pháp học chuyển giao, bao gồm XceptionNet, RestNet101, InceptionResV2, MobileNetv2, VGG19 và DenseNet121 và so sánh với mô hình CNN truyền thống. Thông qua các thử nghiệm trên tập dữ liệu Celeb-DF, chúng tôi chứng minh rằng việc sử dụng học chuyển giao đem lại hiệu suất tốt hơn so với mô hình CNN truyền thống và DenseNet121 cùng trình phân loại softmax hoạt động hiệu quả hơn so với các phương pháp khác.

DOI: https://doi.org/10.34238/tnu-jst.8754

* Corresponding author. *Email: lmtuan.1989@gmail.com*

1. Introduction

Ensuring safety and information security are problem that any country has to face. In particular, with the rapid development of social media, the use of technologies to create fake news and misleading information is becoming a big problem. One of the common techniques is deepfake. Deepfakes are digital media - video, audio, and images edited and manipulated using Artificial Intelligence. Deepfakes can be used to swap the faces of celebrities or politicians to bodies in porn images and videos for creating chaos, blackmailing or other purposes [1] - [3].

GAN (Generative Adversarial Networks) [4] is a complex deep learning technology that can be applied to create deepfake images and videos that humans cannot distinguish from authentic ones. Those models are used to train on the dataset, then generate fake images and videos. Deepfake methods require a large training dataset. The larger the data set, the more photo-realistic images and videos. It is threatening to society's security when deepfake methods can be employed to create videos of politicians with forged speeches or celebrities with obscene content.

Deepfake crimes are rising daily. Deepfake media detection is a big challenge and has high demand in digital forensics. To address the threat of deepfakes, the United States Defense Advanced Research Projects Agency (DARPA) initiated a research scheme in media forensics (named Media Forensics or MediFor) to accelerate the development of fake digital visual media detection methods. Recently, Facebook Inc. teaming up with Microsoft Corp and the Partnership on AI coalition have launched the Deepfake Detection Challenge to catalyse more research and development in detecting and preventing deepfakes from being used to mislead viewers [5], [6].

Researchers have introduced several methods to identify fake digital content. The methods used for deepfakes detection are broadly categorized as either machine learning (ML)-based methods or deep-learning (DL) based approaches. Yang et al. [7] proposed a deepfake detection technique that used 2D facial landmarks to compute the 3D head position. To train the SVM classifier, the calculated difference between the head poses was employed as a keypoints vector. This approach shows better performance for detecting deepfakes however, it is unable to compute the landmark alignment from the blurred samples which reduces the robustness of this framework under these cases. The work in [8] utilized the Image Quality Metric (IQM) together with the principal component analysis (PCA) and linear discriminant analysis (LDA) for keypoints computation. The SVM classifier was used to train the obtained and classify the input as fake or real. The results show that the face recognition approaches i.e., Facenet [9] and Visual Geometry Group (VGG) [10] are incompetent to identify visual manipulations. In [11] authors introduced a deepfake detection technique to protect famous personalities. Initially, deepfakes were generated by employing GAN, on which OpenFace2 [12] toolkit was applied to compute the face features and head movements. The binary SVM was trained to distinguish between genuine and forged faces using the measured landmarks. Recently, DL-based approaches have been getting the attention of researchers for deepfakes detection.

Although there has been a lot of research in the field of deepfake detection based on deep learning. However, over time, deepfake methods have been improved with more challenging datasets that make existing detection techniques perform poorly. Therefore, the primary aim of our research is to improve the performance and efficiency of the above problem. The main contributions of this paper include:

- A general procedure for deepfake detection using pre-trained models combined with the concept of transfer learning to solve the problem of overfitting.
- Evaluate the performance of common deep learning models such as XceptionNet, ResNet101, InceptionResV2, MobileNetv2, VGG19, DenseNet121 in detecting deepfake.
- Performance evaluation was performed on the Celeb-DF dataset to show the robustness of the deep learning models.

2. Methodology

The reseach mothodology architectural analysis is examined in Figure 1.

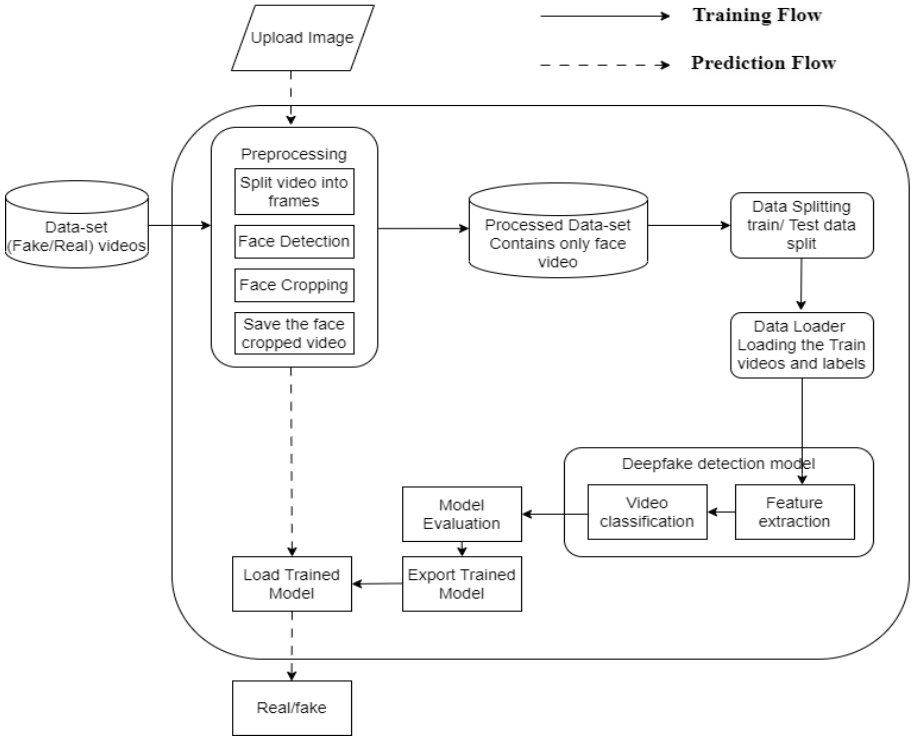


Figure 1. *Workflow diagram of proposed methodology in deepfake prediction*

The deepfake images based on fake and real human faces are utilized for building models. The fake and real faces are structured with the target label into a dataset. The structured deepfake dataset is split into train, validation, and test data portions. We utilized 80% data for training while the remaining 20% data for evalution. The proposed methodology includes the following steps:

- Preprocessing
- Feature computation
- Prediction.

2.1. Preprocessing

In this step, the input video will be obtained from a static data source (such as from a file, database), or a dynamic source (like from a webcam, camera, etc.). Next, extract a certain number of frames on a video. In deepfake videos, the facial region is the main portion of the frame that is manipulated. So, we are concerned with the face area only. A face detector is employed to locate and extract the face region from the rest of the frame. After the face images are extracted, image alignment and data enhancement are performed.

In this study, we use a face detection model to extract the face area out of frames by the DNN Face Extractor method [13].

2.2. Feature computation

Features are extracted using transfer learning techniques with pre-trained models on ImageNet. Transfer learning is an approach where we build pre-trained models for the prediction process. Transfer learning provides the ability to utilize the learned features for high performance in prediction.

Since the pre-trained models have been trained with a fixed-image size, when using them, we need to resize the image to the size required by the model's convolutional neural network. Therefore, the study has resized the image to 224x224 as input for the training model and image prediction.

The CNN model used in our study consists of three Conv2D convolutional blocks (layers.Conv2D) with a MaxPooling layer (layers.MaxPooling2D) in each block. There is a fully connected layer (layers.Dense) with 128 units activated by the ReLU activation function ('relu').

To reduce overfitting, the model uses an additional dropout layer (layers.Dropout(0.2)). The top fully connected layer with 2 units uses softmax activation function to classify the image as real or fake.

The pre-trained CNNs model used in our study are XceptionNet, ResNet101, InceptionResNetV2, MobileNet v2, VGG-19 và DenseNet121.

XceptionNet [14] is an improved form of the Inception-V3 model. In the XceptionNet model, the depthwise separable convolution is introduced instead of using inception modules that are followed by pointwise convolutional layers. A stack of depthwise separable convolution layers with residual connection is applied independently over each input data channel. The pointwise convolutional layer (filter size of 1x1) projects the channel output through a depth-wise convolution into a new channel space.

ResNet-101 (Residual Neural Network) [15] introduces an identity shortcut connection that skips one or more layers. The shortcut connection applies identity mapping, and the computed outcomes are passed to the outputs of the stacked layers. It follows the assumption that only building a stack of identity mappings will achieve similar results as a shallow structure.

InceptionResV2 [16] incorporates the benefits of both the Inception and ResNet models. The inception model better learns the features at different resolutions within the same convolution layer while the ResNet model supports deeper CNN for learning feature without compromising the performance.

MobileNet is a type of convolutional neural network used for image classification developed by Google researchers [12]. They are based on depthwise separable convolutions to build a lightweight deep CNN that reduces computation time and makes a model very small. Depthwise separable convolution is a type of convolution in which only one convolutional filter is applied to each input channel [12].

VGG19 is a deep learning image classification advanced CNN with 19 layers (16 convolutional layers and 3 fully connected layers) [11]. VGG19 is a very deep network that has been trained on millions of images with complex classification tasks. As a result, the network has learned rich feature representations for a wide range of images

A DenseNet121 is a type of convolutional neural network. Each layer in the DenseNet121 architecture is connected to every other layer in a feedforward manner. All previous layers' feature maps will be applied to each layer. As a result, even after passing through many layers, the features are preserved. Due to the reuse of functions, the parameters in the DenseNet121 architecture are reduced [10].

2.3. Prediction

The classifier uses fully connected layers with a softmax activation function with two labels, Fake and Real, for classification and forgery detection.

3. Experiments and results

3.1. Dataset description and evaluation matrices

We tested all models on the Celeb-DF dataset. The Celeb-DF dataset is comprised of 590 real videos and 5,639 DeepFake videos (corresponding to over two million video frames). The average length of all videos is approximately 13 seconds with the standard frame rate of 30

frame-per-second. The real videos are chosen from publicly available YouTube videos, corresponding to interviews of 59 celebrities with diverse distribution in their genders, ages, and ethnic groups. The provided dataset is divided randomly into 80-20 parts. We used 80% of the data for training, while the remaining 20% was for evaluation.

Performance metrics for classification problems include accuracy, F1-measure, precision, and recall. Performance metrics are defined from the baseline as follows [17].

True Positive (TP): It is the case when the model correctly predicted positive outcomes.

True Negative (TN): It is the case when the model correctly predicted negative outcomes.

False Positive (FP): It is the case when the model incorrectly predicted positive outcomes.

False Negative (FN): It is the case when the model incorrectly predicted negative outcomes.

Accuracy: Accuracy is defined as the ratio of the total number of true predicted values to the total number of predictions. Accuracy is used for evaluating classification models. The accuracy of the ML model indicates how many times it was correct overall [17].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: Precision is defined as the ratio of true predicted classification values to the total number of positive classifications [17].

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall: The recall is determined by dividing the total number of positive samples by the number of Positive samples that were correctly identified as Positive [17].

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1 Score: The F1-Score is calculated by taking the harmonic mean of precision and recall and assigning equal weight to each [17]. A higher F1 score indicates better performance from the classifier.

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (4)$$

3.2. Implementation details

The research experimental setup was utilized to build the network. Python with TensorFlow and Keras were utilized to evaluate all experimental studies. The softmax classifier is trained using various feature extraction networks and applied to locate fakes from the Celeb-DF dataset with 20 epochs and a 0.0001 learning rate. The accuracy score, precision score, recall, and F1 score were performance metrics utilized or performance evaluations.

3.3. Results and Discussion

The comparative performance analysis of employed models on unseen test data is analyzed in Table 1. That analysis demonstrates that the VGG-19 model had low-performance scores compared to other transfer learning models.

Table 1. *The comparative performance analysis of employed models*

Model	Accuracy	Precision	Recall	F1 score
CNN	0.8780	0.9142	0.8470	0.8793
InceptionResNetv2	0.9832	0.9876	0.9802	0.9839
MobileNet v2	0.9743	0.9670	0.9847	0.9757
VGG19	0.9642	0.9499	0.9835	0.9664
ResNet101	0.9789	0.9786	0.9812	0.9799
XCeptionNet	0.9768	0.9941	0.9615	0.9775
DenseNet121	0.9842	0.9849	0.9850	0.9850

The bar chart-based performance comparative analysis of employed models is visualized in Figure 2. Among all the employed models, the DenseNet121 obtained the highest accuracy with the value of 98.42% while the InceptionResNetv2 exhibits the second highest accuracy value of 98.32%. VGG-19 model shows the lowest accuracy with a value of 96.42%. Actually, incorrectly detecting a real sample as deepfakes is less costly than missclssifying a deepfake sample as the original. Thereforce, the objective of deepfakes detection techniques is to minimize the false negatives value, hence, optimizing recall is the main priority. Figure 2 demonstrates the DensetNet121 model shows the highest recall value of 98.5%, XceptionNet shows the least value of 96.15%. The F1-score gives an overall analysis of the robustness of the classifier. The DenseNet121 shows the highest F1-score value of 98.50%, while the VGG-19 shows the lowest F1-score value of 96.64%.

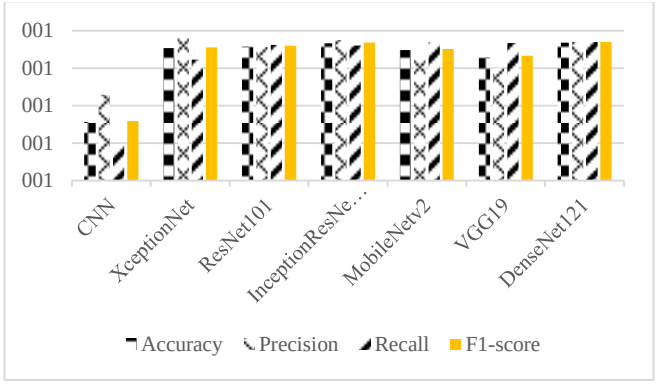
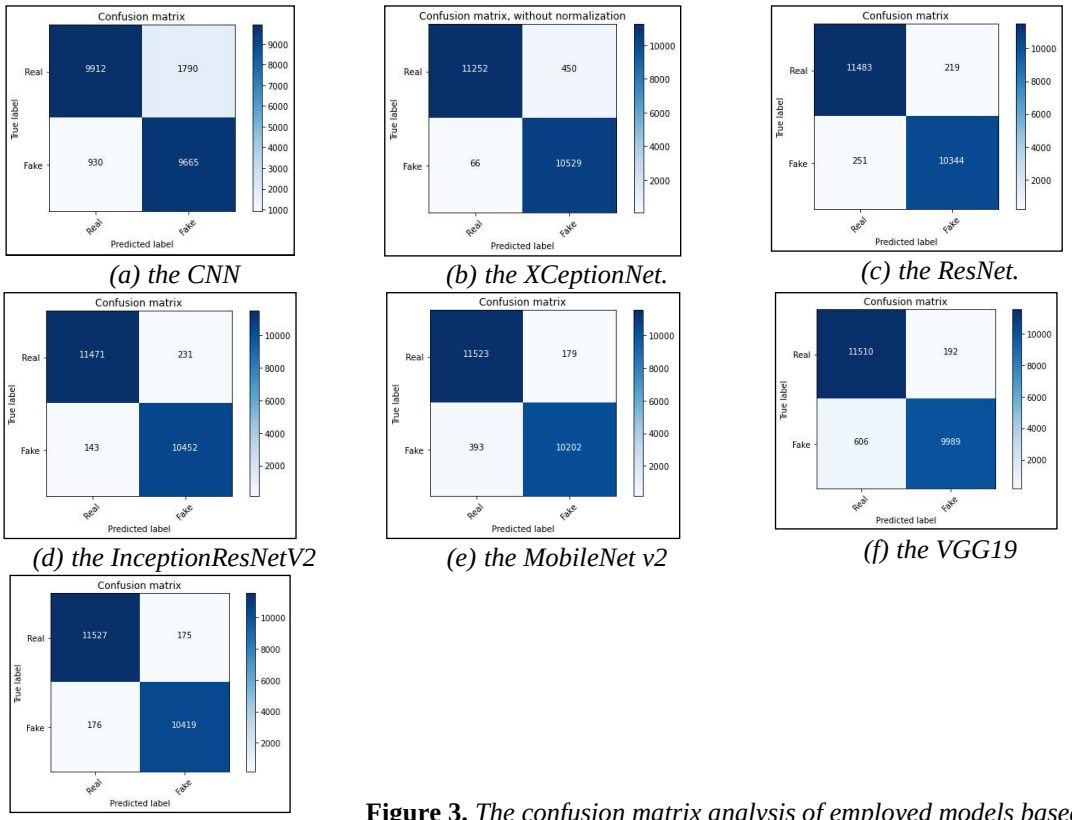


Figure 2. *The bar chart-based performance comparative analysis of employed models*



(a) the CNN *(b) the XCeptionNet.* *(c) the ResNet.*

(d) the InceptionResNetV2 *(e) the MobileNet v2* *(f) the VGG19*

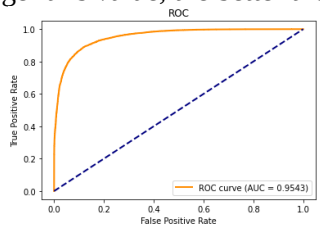
(g) the DenseNet121

Figure 3. *The confusion matrix analysis of employed models based on the validations of performance metrics results*

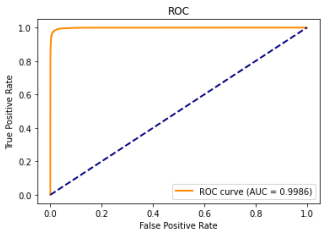
Another important metric that was used to qualify the employed models by this paper is the confusion matrix that is examined in Figure 3. This metric is possible to verify the performance of the algorithm by comparing the predictions with the real values of the labels. The high-class prediction error rate was achived by the VGG19 model, followed by MobileNet and ResNet. XceptionNet has a minumum error rate in class predictions compared to confusion matrix results.

Besides, the experimental results are alose evaluated based on ROC (Receiver Operating Characteristic) curve, which is a graphical chart illustrating the performance of the binary classification system. Each point on the ROC curve is the coordinate correspoding to the TPR on the vertical axis and the FPR on the horizontal axis. An efficient model with low FPR and high TPR, that is, there exists a point on the ROC that is close to the point with coordinates (0, 1) on the graph (upper left corner). The closer the curve is to the upper point, the more efficient the model is.

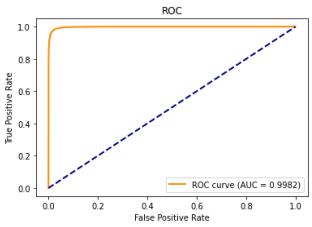
There is another parameter used to evaluate, Area Under the Curve or AUC. It is the area under the orange ROC. The value of this area is a positive number less than or equal to 1. The larger this value, the better the model.



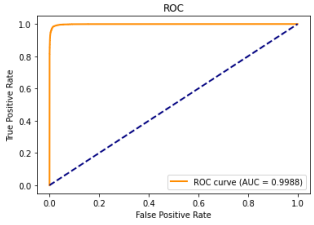
(a) ROC curve of Custom CNN AUC = 0.9543



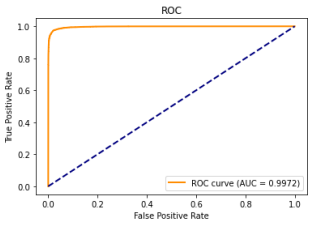
(b) ROC curve of XceptionNet AUC = 0.9986.



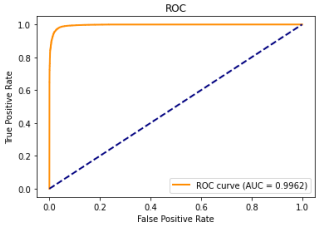
(c) ROC curve of ResNet AUC = 0.9982



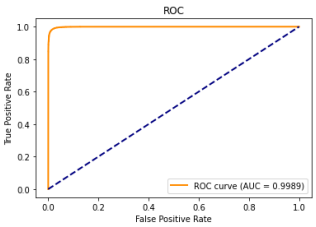
(d) ROC curve of InceptionResNet AUC=0.9988



(e) ROC curve of MobileNet AUC=0.9972



(f) ROC curve of VGG19 AUC=0.9962



(g) ROC curve of DenseNet121 AUC=0.9989

Figure 4. *ROC curve comparison of analyzed models*

Figure 4 represented the results obtained by the ROC curve comparision of analyzed models. DenseNet121 attains the highest AUC of 99,89%, while VGG-19 obtains the lowest AUC value of 99,62%. Therefore, DenseNet121 is highly preferred for deepfakes detection.

4. Conclusion

Deepfake detection using a deep learning-based approach is proposed to help cybersecurity professionals overcome deepfake-related cybercrimes by accurately detecting the deepfake content. In this paper, we apply transfer learning aproaches to compute the representative set of

features from the suspected samples for the task of deepfake detection on dataset Celeb-DF. The work also demonstrates the concept of using transfer learning to compute the representative set of features from the suspected samples.

Several performance evaluation metrics i.e. accuracy, precision, recall, and F1-score, have been utilized in this work to perform an in-depth analysis of how various DL-based feature extractors work. After comparison, it is concluded that DenseNet-121 together with the softmax classifier performs well than the rest of the approaches. For future work, the model presents tests on images from other sources and can be further tweaked with different classifiers to show more accurate classification results.

Acknowledgment

This work has been supported by Academy of Cryptography Techniques.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] L. Nataraj, T. M. Mohammed, B. S. Manjunath, S. Chandrasekaran, A. Flenner, J. H. Bappy, and A. K. Roy-Chowdhury, "Detecting GAN generated Fake Images using Co-occurrence Matrices," *Electronic Imaging*, vol. 31, no. 5, pp. 532-1 - 532-7, 2019.

[2] S. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot… for Now," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 8692-8701.

[3] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to Detect Fake Face Im-ages in the Wild," *IEEE International Symposium on Computer, Consumer and Control* (IS3C), 2018, pp. 388-391.

[4] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," 2018. [Online]. Available: https://arxiv.org/abs/1411.1784. [Accessed Sept. 12, 2022].

[5] A. O. J. Kwok and S. G. M. Koh, "Deepfake: a social construction of technology perspective," *Current Issues in Tourism*, vol. 24, no.13, pp. 1798–1802, 2021.

[6] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

[7] X. Yang, Y. Li, and S. Lyu, "Exposing deepfakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8261-8265.

[8] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," 2018. [Online]. Available: http://export.arxiv.org/pdf/1812.08685. [Accessed Dec. 14, 2022].

[9] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815-823.

[10] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face Recognition Algorithms Surpass Humans Matching Faces Over Changes in Illumination," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1642-1646, 2007.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society* 2015, pp. 1–14.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-10.

[13] OpenCV: DNN-based Face Detection And Recognition, (n.d.). [Online]. Available: https://docs.opencv.org/4.x/d0/dd4/tutorial_dnn_face.html. [Accessed August 31, 2023].

[14] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," In *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[16] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261-2269.

[17] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing Management*, vol. 45, no. 4, pp. 427–437, 2009.