DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

Unmasking Deepfakes Based on Deep Learning and Noise Residuals

Wildan J. Hadi[1], Suhad M. Kadhem[2], Ayad R. Abbas[3]
[1]*Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq*
[2,3]*Computer Sciences Department, University of Technology, Baghdad, Iraq*
[1]*wildanjh_comp@csw.uobaghdad.edu.iq* , [2]*110102@uotechnology.edu.iq*, [3]*110010@uotechnology.edu.iq*

Abstract— *The main reason for the emergence of a deepfake (deep learning and fake) term is the evolution in artificial intelligence techniques, especially deep learning. Deep learning algorithms, which auto-solve problems when giving large sets of data, are used to swap faces in digital media to create fake media with a realistic appearance. To increase the accuracy of distinguishing a real video from fake one, a new model has been developed based on deep learning and noise residuals. By using Steganalysis Rich Model (SRM) filters, we can gather a low-level noise map that is used as input to a light Convolution neural network (CNN) to classify a real face from fake one. The results of our work show that the training accuracy of the CNN model can be significantly enhanced by using noise residuals instead of RGB pixels. Compared to alternative methods, the advantages of our method include higher detection accuracy, lowest training time, a fewer number of layers and parameters.*

Index Terms— *Deepfake, Deep Learning, Steganalysis Rich Model, Convolution Neural Network.*

I. INTRODUCTION

Recently, the efficiency of computers has increased significantly in terms of simulating reality. One aspect of this evolution is possible to see in modern cinema, which has been dependent on these computers to generate personalities, landscape and presented in a manner that cannot be distinguished from reality [1]. Another form of computer efficiency is found in modern gaming, where this advancement contributed to establishing an accurate simulation in the gaming environment [2].

The recent development has not been limited to the efficiency of computers but also included artificial intelligence (AI) techniques. AI is not a set of theories, but in fact, is treated as a "paradigm-change" technology that has been entered in many application areas [3],[4].Deep Learning is a subfield of AI that has been referred to by many researchers and developers. Deep learning algorithms such as generative adversarial networks (GANs) [5],[6] or Autoencoders (AEs) [7],[8] can be applied to images, text, video, and voices to create fake media which show it is realistic but is not [9].

The advent of modern computers with powerful Artificial Intelligence techniques (especially Deep Learning) enabled many artificial intelligence applications to appear and spread, "deepfake" is one of these applications. Deepfakes appeared by taking celebrities' faces and putting them in adult videos. This phenomenon affected celebrities and included politicians, presidents, and singers, where their faces were placed in fake videos to say and do things they had never done [10]. An example of deepfake is shown in *Fig. 1*, this image is a simple part of fake videos starring Tom Cruise [11]. In 2018, the common face-swapping program needed large sets of data to output Deepfakes. After that, similar applications become less complex and more flexible. For example, China's ZAO application fakes many TV shows and movies by swapping faces [10].

Received 14/October/2021; Accepted 31/December/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

As we noted, the trends in technology have contributed significantly to the increase in the techniques of generating deepfakes, which have recently become more challenging to detect and recognize. For example, most of the models applied to The DeepFake Detection Challenge (DFDC) dataset [12] suffered much worse performance when applied to invisible data than this found in this dataset [13]. Also, most of these methods are based on RGB image contents as input to their models. Therefore, it is necessary to build a deepfake detection model that achieves a high generalization ability.

The steganalysis rich model (SRM) [14] will be employed in our work because of the positive results it has gained in the field of image forensics for evaluating data steganography. First, noise residuals are extracted using an SRM filter, and the resultant noise content (instead of RGB image content) is given to the light Convolution Neural Network (CNN), which can strongly distinguish deepfake videos from real ones.

The following are the main contributions of the current work:

- Proposing a new model for detecting deepfake based on deep learning and noise residuals. The results achieve good performance when compared with alternative methods.
- The lowest training time is required to use a noise map (gathered by SRM filter) instead of RGB image content.
- Combining multimedia forensics tools with deep learning increases detection accuracy and exploiting this in the future can make a qualitative leap in detecting deepfakes.

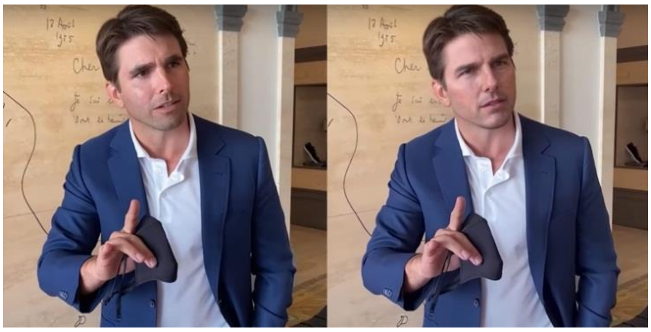

FIG.1. AN EXAMPLE OF DEEPFAKES [11]. LEFT THE REAL IMAGE.RIGHT THE FAKE IMAGE.

II. RELATED WORK

Multimedia Forensics-based methods: Before the advent of deep learning and deepfake technology, multimedia forensics were used to expose manipulation in digital media. One of the essential sections of multimedia forensics is that when the JPEG-compressed image gets tampered with and compressed again, the result is that defects appear throughout the image except for the affected area [15]. On the other hand, exploiting and detect anomalies caused by double compression in the video is usually possible but more difficult due to the complication of the video encoding algorithm [16]. Another aspect of multimedia forensics is analyzing local noise using statistical tools in the spatial domain [17]. This context, called error level analysis(ELA), is widely used in research for its simplicity. Glitches in the manufacture of the camera sensor also cause these sensors to cause minor deviations from their imposed behavior. These deviations create a pattern similar to a noise pattern called photo-response non-uniformity (PRNU) noise, which can be considered a camera fingerprint. So if the image is tampered with, this manipulated area does not contain This pattern [16]. Facial landmarks position were also analyzed by

Received 14/October/2021; Accepted 31/December/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

researchers in reference [18] to exploit them as a sign for deepfakes. Specifically, face synthesis models create fake faces that keep the facial expression of the original faces. This mismatch in the location of facial landmarks of two faces is used for deepfakes detection.

Deep Learning-based methods The methods used to detect deepfakes are usually based mainly on Convolutional Neural Networks (CNN), due to several reasons. First, many online platforms such as YouTube have huge videos and images of many celebrity's actors available free that can be used to train neural networks. Second, many public databases such as Face Forensic ++ [19], Celeb-DF [20], TIMIT [21], UADFV [18], and others exist and provide many data sufficient to train any CNN model [22]. For example, in [23], two different structures of the CNN neural network were used, each one with a low number of layers to concentrate on the mesoscopic characteristics of images. The first consisting of four layers of successive convolutions and pooling. Also, the second structure similar to the first one except the first two convolutional layers replaced by an alternative of the inception module. The two models were tested on several databases and achieved promising results. The authors of [24] used a capsule network for forgery detection. The suggested model includes three essential capsules with two outputs to classify fake from real. Transfer learning also has a large share in the field of deepfake detection. Much research relied on the principle of transfer learning in forgery detection. One of these research relied on assembling different trained CNN models [25]. In reference [26], the authors proposed the Fake spotter approach, which recognizes fake faces from real ones by observing the neuron behaviors of deep face recognition systems using a simple binary classifier. The database used for the evaluation of this model consists of Face Forensics ++, Celeb-DF, and fake faces generated with GANs.The overall performance reaches more than 90%. Temporal inconsistencies at video frame levels are also exploited for detecting deepfakes. The authors of [27] propose a two-step deepfake detection step, which firstly extracts feature from input video frames using CNN followed by capturing temporal inconsistencies using recurrent neural network RNN. Database used for this model collected from different websites. The accuracy of this model is about 94%. Another method based on deep learning is SSTNet used in [28] to detect tampering. The proposed method used CNN for learning steganalysis features to help in detecting hidden anomalies. This step is followed by RNN to exploit temporal inconsistencies between frames. Experimental is applied on Face forensics ++ database.

Deep learning and multimedia forensics-based methods Most recent research has focused on building models based on multimedia forensics tools with deep learning techniques to improve the accuracy of deepfake detection. The authors of [29] built a model based on error level analysis (ELA) in detecting defects, and then the resulting images are entered into a binary classifier. The proposed model achieved good results, reaching the accuracy of deep fake detection of more than 97%. A different example of integrating multimedia forensics tools with deep learning techniques is found in [30]. The authors relied on the RGB image contents with noise contents to learn rich features useful in detecting manipulation. The tampered regions are identified using both RGB and noise contents of the input images. In reference [31] multimedia forensics and deep learning are also used for deepfakes detection. Combining triplet network and GoogleLeNet as two-stream network model for learning local noise maps and tampering artifacts. The proposed method gave promising results.

Received 14/October/2021; Accepted 31/December/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

III. THE PROPOSED METHOD

In this section, the steps required for deepfake detection is described in detail. Firstly, we show how noise residuals' maps are extracted from the RGB input dataset using SRM filters. Then after preparing the noise residuals maps of real and fake faces, the architecture of binary classifier using CNN model is described.

A. Noise Residuals Processing

The RGB image contents are not sufficient to be the basis on which to detect all kinds of manipulation. Therefore, it is necessary to search for another way that is relied upon alone or with the RGB contents to detect media manipulation. One of these things is to pay attention to noise rather than the high level of semantic image content. We are based on noise residuals as input to our model using SRM filters [14]. This noise residuals' calculations are obtained by passing three high-pass filters as shown in *Fig. 2,* which are given good performance among the rest 30 kernels as applied in [30].

$$\frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

FIG. 2. SRM THREE FILTERS TO GATHER NOISE MAP.

More precisely, the process is to convolve the RGB face image with a filter of size 5*5 using SRM filters mentioned above to gather noise distribution. This mean convertingmm all detected faces from the input video frames to noise maps. This process can be treated as preprocessing convolution layer before going to the deep learning step. *Fig. 3* shows the effectiveness of the SRM filter on both real and fake images.

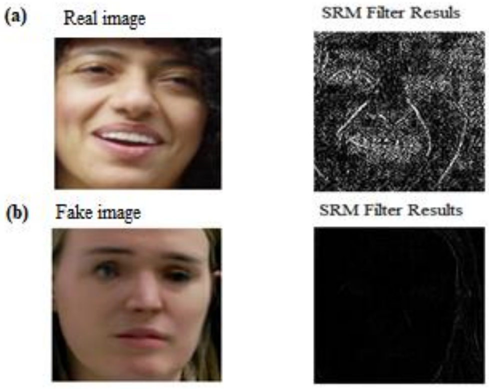

FIG.3. APPLYING SRM FILTER ON FAKE AND REAL IMAGES. (A) AN EXAMPLE OF APPLYING SRM FILTERS TO THE REAL IMAGE. (B) THE RESULTS OF APPLYING SRM FILTERS ON THE FAKE IMAGE.

B. Deepfakes Detection

Fig. 4 shows the general structure of the proposed model. The model relied on the Face Forensics database, where the frames were extracted from each video, then the front face detector in dlib (an open source library) was used to extract the facial area. Next, the resulted images were preprocessed by

Received 14/October/2021; Accepted 31/December/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

convolving them with SRM filters (mentioned above) to gather noise for each image. Finally, the preprocessed images were given to lightweight CNN for binary classification after dividing them to training and testing sets. The network architecture being used consists of only two convolution layers because the network is trained to extract the features of SRM images. So, it is easy to distinguish between real and fake images.

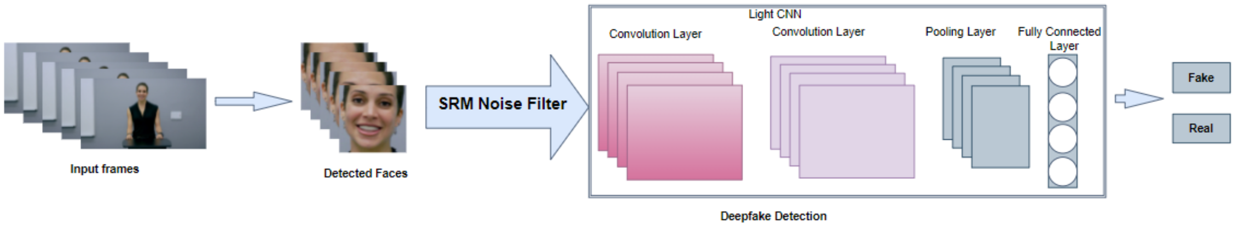

FIG. 4. ILLUSTRATION OF OUR MODEL. THE FACE AREA FROM INPUT FRAMES IS EXTRACTED USING DLIB AND THEN PROCESSED BY SRM FILTERS TO GATHER NOISE MAPS TO FINALLY INPUT THEM TO CNN BINARY CLASSIFIER.

IV. EXPERIMENTS

To train the proposed model, we used the Face Forensics database. This database consists of rich YouTube videos with high quality and realistic forgeries to help in image classification models. So that after gathering noise maps from the database, they were divided into training and testing sets for CNN training.

A. Training Model

After extracting ROI (using dlib) from the input frames, noise maps are gathered. These maps represent the input to our CNN model instead of original semantic image content. The CNN architecture consists of two convolution layers with one max pooling to learn SRM filtered input images. The batch size is adjusted as 64, and the ADAM optimization method is used. The proposed model arrives about 98% in terms of training accuracy and the validation accuracy is near to 97% as shown in *Fig. 5*. Just with 30 epochs, my training and testing accuracy convergence very quickly.

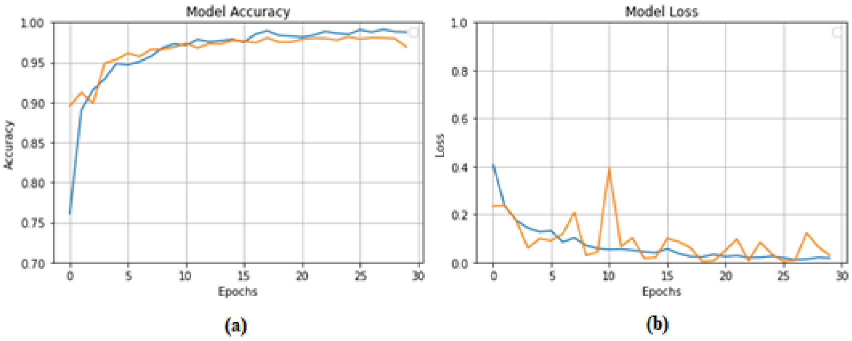

FIG. 5. PERFORMANCE OF CNN MODEL ON SRM FILTERED IMAGES. (A) VALIDATION ACCURACY VS TRAINING ACCURACY. (B) VALIDATION LOSS VS TRAINING LOSS.

Received 14/October/2021; Accepted 31/December/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

B. Comparison with alternative Methods

The performance of the proposed model was compared with alternative methods like two-stream Neural networks [31], head poses [18], and MesoNet [23]. The performance of these methods reached 85.1, 89.0, 84.3, respectively. The advantages of our proposed model include reducing the required numbers of training epochs because of including the noise residuals maps instead of RGB maps in the training set. Also, the accuracy of training and validation steps reaches high values. This means that the features in the input noise maps can be used successfully to recognize fake video frames. However, when comparing our model with these alternative methods, the following problems can be recognized:

1- If there is no GPU environment, it is hard to train such models. On the other hand, our method can be implemented without a GPU environment.

2- The principle on which other models rely to detect forgery is not recognized, whereas our model detects forgery using noise residuals maps.

V. CONCLUSIONS

It's gotten tough to identify whether this video is fake or not since the emergence of modern artificial intelligence techniques. In this research, It was created a model that can distinguish between the real video and the one created using artificial intelligence methods. To increase deepfake detection accuracy, we built a model based on multimedia forensic methods combined with deep learning. The proposed model was evaluated and its performance was compared to that of alternative methods. This indicates the effectiveness of learning noise residuals instead of semantic RGB image contents. The proposed model helps to reduce the required training time, as well as the required resources.

REFERENCES

[1] Jameel, W. J., Kadhem, S. M., and Abbas, A. R., " Detecting Deepfakes with Deep Learning and Gabor Filters", *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 1,pp. 18-22, 2022.

[2] "How AI and simulated reality are changing the way we look at sports and gaming." https://yourstory.com/2021/07/ai-simulated-reality-changing-sports-gaming/amp (accessed Sep. 07, 2021).

[3] Z. A. Mohammed, M. N. Abdullah, and I. H. Al Hussaini, "Predicting Incident Duration Based on Machine Learning Methods," *IRAQI J. Comput. Commun. Control Syst. Eng.*, vol. 21, no. 1, pp. 1–15, 2021.

[4] A. Q. Albayati and S. H. Ameen, "A Method of Deep Learning Tackles Sentiment Analysis Problem in Arabic Texts," *IRAQI J. Comput. Commun. Control Syst. Eng.*, vol. 20, no. 4, pp. 9–20, 2020.

[5] X. Mao and Q. Li, Generative Adversarial Networks for Image Generation. Springer Nature, 2021.

[6] X. Wang, W. Li, G. Mu, D. Huang, and Y. Wang, "Facial expression synthesis by u-net conditional generative adversarial networks," in Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, 2018, pp. 283–290.

[7] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in 2017 seventh international conference on affective computing and intelligent interaction (ACII), 2017, pp. 370–376.

[8] H. Zhu, Q. Zhou, J. Zhang, and J. Z. Wang, "Facial aging and rejuvenation by conditional multi-adversarial autoencoder with ordinal regression," arXiv Prepr. arXiv1804.02740, 2018.

[9] S. Karnouskos, "Artificial Intelligence in Digital Media: The Era of Deepfakes IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY 1 Artificial Intelligence in Digital Media: The Era of Deepfakes," 2020, doi: 10.1109/TTS.2020.3001312.

[10] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?," Bus. Horiz., vol. 63, no. 2, pp. 135–146, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681319301600.

[11] "Creating These Convincing Tom Cruise Deepfakes Wasn't Easy." https://www.makeuseof.com/creating-convincing-tom-cruise-deepfakes-wasnt-easy/ (accessed Sep. 09, 2021).

[12] B. Dolhansky et al., "The deepfake detection challenge dataset," arXiv Prepr. arXiv2006.07397, 2020, [Online]. Available: https://arxiv.org/abs/2006.07397.

[13] S. A. Khan and H. Dai, "Video Transformer for Deepfake Detection with Incremental Learning," arXiv Prepr.

Received 14/October/2021; Accepted 31/December/2021

DOI: https://doi.org/10.33103/uot.ijccce.22.3.10

arXiv2108.05307, 2021.

[14] J. Fridrich and J. Kodovsky, “Rich models for steganalysis of digital images,” IEEE Trans. Inf. Forensics Secur., vol. 7, no. 3, pp. 868–882, 2012.

[15] J. Lukáš and J. Fridrich, “Estimation of primary quantization matrix in double compressed JPEG images,” in Proc. Digital forensic research workshop, 2003, pp. 5–8, [Online]. Available: http://www.ws.binghamton.edu/fridrich/Research/Doublecompression.pdf.

[16] L. Verdoliva, “Media forensics and deepfakes: an overview,” IEEE J. Sel. Top. Signal Process., vol. 14, no. 5, pp. 910–932, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9115874/.

[17] S. Lyu, X. Pan, and X. Zhang, “Exposing region splicing forgeries with blind local noise estimation,” Int. J. Comput. Vis., vol. 110, no. 2, pp. 202–221, 2014.

[18] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8261–8265, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8683164/.

[19] D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and A. Nießner, MatthiasRössler, “Faceforensics: A large-scale video dataset for forgery detection in human faces,” arXiv Prepr. arXiv1803.09179, 2018, [Online]. Available: https://arxiv.org/abs/1803.09179.

[20] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216, [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.html.

[21] P. Korshunov and S. Marcel, “Deepfakes: a new threat to face recognition? assessment and detection,” arXiv Prepr. arXiv1812.08685, 2018, [Online]. Available: https://arxiv.org/abs/1812.08685.

[22] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann, “Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features,” J. Imaging, vol. 7, no. 7, p. 108, 2021.

[23] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[24] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311.

[25] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, “Video face manipulation detection through ensemble of cnns,” in 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5012–5019.

[26] R. Wang et al., “FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces,” arXiv Prepr. arXiv1909.06122, 2019, [Online]. Available: https://arxiv.org/abs/1909.06122.

[27] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8639163/.

[28] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, “Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2952–2956.

[29] W. Zhang and C. Zhao, “Exposing Face-Swap Images Based on Deep Learning and ELA Detection,” in Multidisciplinary Digital Publishing Institute Proceedings, 2019, vol. 46, no. 1, p. 29, [Online]. Available: https://www.mdpi.com/2504-3900/46/1/29.

[30] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Learning rich features for image manipulation detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1053–1061.

[31] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Two-stream neural networks for tampered face detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1831–1839.

Received 14/October/2021; Accepted 31/December/2021