ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20220403080215

Original scientific paper
Received: 2022-03-30, Accepted: 2022-04-12

AFMB-Net: DeepFake Detection Network Using Heart Rate Analysis

A. Vinay, Nipun Bhat, Paras S. Khurana*, Vishruth Lakshminarayanan, Vivek Nagesh, S. Natarajan, T. B. Sudarshan

Abstract: With advances in deepfake generating technology, it is getting increasingly difficult to detect deepfakes. Deepfakes can be used for many malpractices such as blackmail, politics, social media, etc. These can lead to widespread misinformation and can be harmful to an individual or an institution's reputation. It has become important to be able to identify deepfakes effectively, while there exist many machine learning techniques to identify them, these methods are not able to cope up with the rapidly improving GAN technology which is used to generate deepfakes. Our project aims to identify deepfakes successfully using machine learning along with Heart Rate Analysis. The heart rate identified by our model is unique to each individual and cannot be spoofed or imitated by a GAN and is thus susceptible to improving GAN technology. To solve the deepfake detection problem we employ various machine learning models along with heart rate analysis to detect deepfakes.

Keywords: Convolutional Neural Network (CNN); Celeb-DF; Deepfake; Frame Normalisation; Generative Adversarial Networks (GAN); Motion Magnification Spatio-Temporal Map (MMST Map); MBConv Blocks; PhotoPlethysmoGraphy (PPG)

1 INTRODUCTION

Deepfakes are fake media in which a person's likeness is replaced with that of some-one else in the media. Though content faking isn't recent, deepfakes use advanced ML and AI generation techniques to influence or produce social media content that can easily deceive the naked eye. Using generative neural network architectures like autoencoders or GANs, is the key machine learning approach used to generate deepfakes (GANs). The neural network used to synthesize deepfakes are autoencoders. Autoencoders comprise of an encoder and a decoder. Encoders reduce dimensions of the images and the decoder uses these reduced dimensions to reconstruct the images. Deepfakes use this autoencoder architecture by encoding a human into the latent space with a universal encoder. Attaching a GAN to the decoder is a standard up-grade to this architecture. In an adversarial relationship, a GAN trains a generator and a discriminator. As a result, the generator produces images that closely resemble reality, as any flaws are detected by the discriminator. Deepfake generation algorithms are constantly evolving making deepfakes difficult to detect. This is because once a flaw is detected in the generation algorithm by the detection algorithms, it is easily fixed. Deepfakes are being used for both positive and negative purposes.

It has become really important to be able to classify if a video is a deepfake or not. While there are many neural network-based methods to identify GAN generated deepfakes, none of these methods can keep up with the improving deepfake generation technology as well as generalize to an unknown domain.

2 RELATED WORK

Ipek Ganiyusufoglu et al. [1] proposes the usage of spatio-temporal features, modelled by 3D CNNs to maximize generalization. These spatio-temporal models can transfer some of their understanding of already learned deepfake methods onto new unseen fake kinds. Xuan HauNguyen et al. [2] used 3D convolution kernels to con-struct deep 3D CNN models that extracts spatio-temporal features from frames sequence for detecting deepfakes. David Güera et al. [3] have used RNN in their approach to detect deepfakes. Ekraam Sabir et al. [4] have used recurrent-convolutional network combinations. ResNet and DenseNet are employed for the CNN component in different experiments, along with unidirectional, bi- directional and multi RNNs on FF++ dataset. Lingzhi Li et al. [5] used HRNet in their paper for classification of deepfakes. It works under the assumption that there is a boundary when deepfake image is formed by blending two images. This leads to the existence of intrinsic image discrepancies across the blending boundary. The face X-ray of the input produces a greyscale image, which reveals if it can be decomposed. Pavel Korshunov et al. [6] have used the VGG network to determine deepfakes from generating their own dataset by using GAN based face-swapping algorithm. Forrest N. Ian-dola et al. [7] have published the SqueezeNet paper, which aims to reduce the number of parameters while retaining the accuracy. Xiangyu Zhang et al. [8] have proposed ShuffleNet, which combines modern convolutional neural networks with ShuffleNet Units stacked in stages, which makes use of sparse connections and grouped convolutions. The ShuffleNet unit introduces channel shuffle operations for information cross- flow between the channel groups in grouped convolutions. The paper aims for efficient computations while maintaining a sturdy performance in detection of deepfakes. Mark Sandler et al. [9] proposed the MB-Conv Blocks in the MobileNet. The MBConv Block performs feature space exploration through the use of its inverted residual blocks which help extract more information from the frames through expansion and squeezing. Peng Zhou et al. [10] propose a method of using a two-stream neural network for deepfake detection where in one stream extracts the facial features from the input and the other stream extracts the steganalysis features. Darius Afchar et al. [11] propose a neural network architecture which is widely used in the domain of deepfakes - MesoNet, which is powerful and compact. Huy H. Nguyen et al. [12] proposes the use of capsule networks to detect deepfakes which is proven to perform better at the video level than at the frame level. Chih-Chung Hsu et al. [13] a discriminator is used in this compared to learning from a binary classifier to effectively and easily detect deepfakes. Lakshmanan

Nataraj et al. [14] published their work where co-occurrence matrices are extracted on three color channels and a CNN Model is trained. Huy H. Nguyen et al. [15] proposed the concept of multitask learning which is used to detect areas of manipulation in the fake video. Run Wang et al. [16] published FakeSpotter where they monitored the behaviour of the neuron, analyzed the pattern in which the neurons are activated to capture features. Andreas Rossler et al. [17] introduces a new deepfake dataset - FaceForensics++ along with Patrick Kwon et al. [18] who introduce the KoDF, a large-scale Korean deepfake dataset. Bojia Zi et al. [19] have also introduced another challenging real-world dataset for deepfake detection - WildDeepfake. Thanh Thi Nguyen at al [20] summarized of techniques of deepfake creation and detection at that time. Recently many new approaches have been introduced in the deepfake detection domain like Domain adaptation [21], GANs [22, 23], Generalization [24], Dynamic face augmentation [25], N-Shot learning [26, 27], Reverse engineering [28-30], Self-supervised learning [31, 32] and Vision transformers [33, 34].

2.1 Heart Rate Analysis

Lijun Yin et al. [35] performed heart rate analysis to detect deepfakes, their method-ology PPG cells are generated by extracting facial features using face detectors, identifying faces with highest PPG stability and extracting regions of interest, aligning non-linear ROI into a rectangular image so that they form a face video, each image is now divided into 32 squares and the chromaticity of the PPG signal in each square is found, using the 32 PPG chromaticity values, the chromaticity of the entire image is found. Roberto Caldelli at al [36] work aims to exploit inter-frame correlations which are used as features for classification. J. Hernandez-Ortega et al. [41] proposed the Motion model and Appearance model with the normalized frames as input as published in their work. The paper published by Steven Fernandes et al. [37] proposes a simpler method for extraction of heart rates from deepfakes compared to the existing techniques using a neural ODE. The heart rate of the original videos is learnt by the Neural ODE using which it can predict the heart rate of all manipulated versions of the original video. Creating a deepfake dataset using a commercial website, 320 videos from DeepFake TIMI are also used. Heart rates of the facial videos were extract-ed. This could be performed using any of the methods listed below:

- Measuring changes in skin colour caused by blood flow patterns, this is performed by finding the average RGB values in the regions of interest.
- The heart rate was calculated using the Fourier Fast Transform The average optical intensity in the forehead was found due to its stability across all frames and the heart rate was found.

Colours in the frames are amplified and the intensity of these colours are then used to identify the flow of blood. The colour red is given special importance as it implies more flow of blood in the given region.

2.2 Attention Feature Map

Hyeonseong Jeon et al. [38] proposed FDFTNet which is a mechanism that is employed to increase the accuracy of popular CNN Models. The core of this approach is to introduce an image-based self-attention module called attention feature map that uses only the attention module and the downsampling layer. This module is added to the pre-trained model and fine-tuned on data to search for new features in the feature space to detect fake images.

3 METHODOLOGY

3.1 Pre-processing Stage

Our study uses the heart rate analysis method in order to detect deepfakes as proposed by Hua Qi et al. [42]. The videos used are from the Celeb-DF dataset [39].

3.1.1 Extracting Faces and Alignment of Frames

- Preparing the dlib library for face detection.
- Preparing MTCNN [40] for face extraction.
- After detecting and extracting the faces from each frame, we process the frame by using 81 face landmarks predictor and remove the eyes and the inner mouth regions from the face.
- We save the newly extracted and processed face as a jpg file, as shown in Fig. 1, representative of each frame along with the facial landmarks as a NumPy file. (Each jpg image from a given video are stored in sequential order in each folder representative of that video)

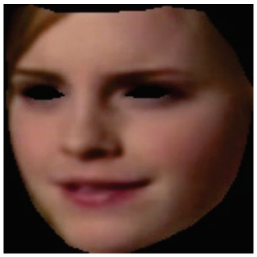

Figure 1 Extracted face after aligning and removing eyes and mouth

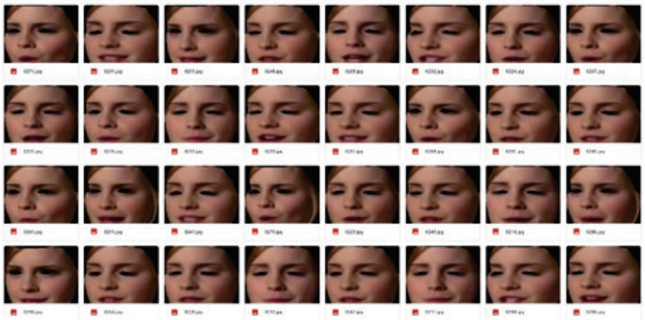

Figure 2 Resized faces

3.1.2 Resizing the Frames

- The output from previous step is the input in this step. These files are all of different dimensions. The aim of this step is to make all the frames of equal dimensions.

- The original dimensions of each jpg file representing a frame are resized to the dimensions 300 × 300.
- The resized frames, as shown in Fig. 2, are now saved as jpg files.

3.1.3 Generating Face Video

- We iterate through the folders containing frames of the aligned faces.
- Using the ffmpeg tool, we now combine these resized and aligned frames to generate a face video.

3.1.4 Generating Motion Magnified Video

- For this step, we use an open-source tool called PyEVM. A frame from a motion magnified video is shown in Fig. 3.
- Spatial decomposition is performed on the video using a Laplacian Pyramid.
- The spatially decomposed videos are now passed through a temporal bandpass filter.
- The resulting signal is amplified.
- This amplified signal is now added back so as to generate the motion magnified video.

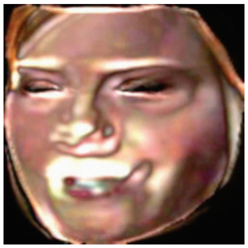

Figure 3 A frame in motion magnified video

3.1.5 Creating Motion Magnified Spatio-Temporal Map (MMST Map)

- Each frame in the motion magnified video is divided in 5 × 5 grid of Region of Interest (ROI) blocks.
- Pooling operation is performed for each block and each color channel of the frame.
- The MMST map is generated for each video in which each row represents a particular ROI block and the corresponding columns of the row represents the temporal variation of the block across all the frames present in the video. A visual representation of the MMST Map is shown in Fig. 4.

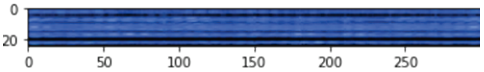

Figure 4 MMST Map

3.1.6 Assigning Spatial and Temporal Weights to the MMST Map

In this step, spatial and temporal weights are assigned to the MMST Map. These weights are used to help with detection of deepfakes.

- The MMST map is passed through a pretrained MesoNet for the spatial weights and the weight matrix W is generated.
- The MMST map is passed through a pretrained LSTM for the temporal weights and the temporal weight matrix $\boldsymbol{T}$ is generated.
- The spatial weight matrix $\boldsymbol{S}$ is multiplied with the temporal weight matrix $\boldsymbol{T}$, resulting in $\boldsymbol{W}$, as shown in Eq. (1).

$$\boldsymbol{W} = \boldsymbol{S} * \boldsymbol{T}^{\mathrm{T}} \tag{1}$$

- The MMST Map, represented by X is now multiplied by the product of Step 4 which is used as the input for the detection models, as shown in Eq. (2).

$$\boldsymbol{A} = X * \boldsymbol{W}^{\mathrm{T}} \tag{2}$$

3.2 Detection Stage

3.2.1 AFMB-Net - Attention Feature Map + ResNet-18 + MB-Conv Blocks

We are using an attention feature map along with MB-Conv blocks and ResNet-18 as represented in Fig. 5. The attention feature map is a method to help ameliorate the accuracy of the baseline model along with the MB-Conv Blocks that perform feature space extraction thus further helping us to extract the most relevant features in identification of deepfakes. By using the attention feature map, we can develop long term dependencies through the use of softmax outputs thus helping reduce computations. If the CNN were to be used to develop the same dependencies, it would take a large number of computations as it performs computation only one time and only sees the kernel size.

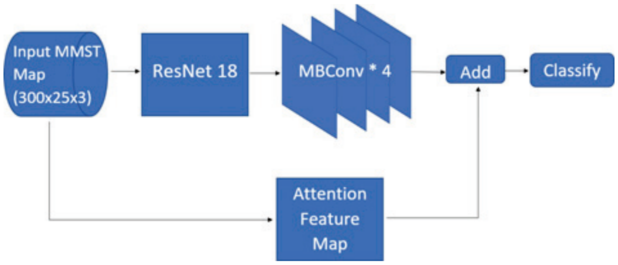

Figure 5 AFMB-Net Architecture

3.2.2 FrameNorm - Frame Normalization Method

The Motion model and Appearance model, shown in Fig. 6, as proposed in the works of J. Hernandez-Ortega et al. [41] are used. This model is slightly different as we have taken the input of this model as the motion magnified videos instead of the MMST Map.

- The difference of frame at time t and at time t–1 of motion magnified video is taken and normalization is performed on the difference.
- The frame at time t is normalized and saved simultaneously with the previous step.

- The output of first step is then passed to the motion model and the output of second step is passed to the spatial model.
- The outputs of both are multiplied and passed through a fully connected dense layer.

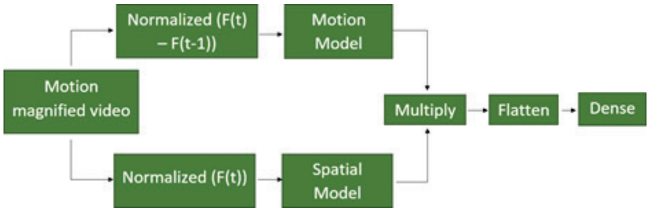

Figure 6 FrameNorm Architecture

4 RESULTS AND DISCUSSION

Our results obtained are depicted through Tab. 1. The AFMB-Net architecture uses the ResNet-18 model as a backbone. The ResNet is a powerful CNN architecture widely used in image classification. This is mainly due to use of deep residual networks in the architecture which reuse the activation functions from the previous layers. The ResNet-18 is a lightweight version of the ResNet that provides high accuracy. The Celeb-DF dataset is a challenging dataset for deepfake detection where most models fall short. In order to enhance the performance of the ResNet for deepfake detection, we have used MBConv blocks and an Attention Feature Map. The Attention Feature Map works by giving special attention to the RGB channels. The MBConv blocks work by performing the squeezing and expansion of the input. This performs feature space extraction on the input and thus further helps extract important features. The AFMB-Net architecture gave us a training accuracy and testing accuracy of 97.91% and 95.19% respectively. Using the above-mentioned intuition, the ResNet-18, MBConv blocks and the Attention Feature Map have helped achieve this accuracy.

Table 1 Models Performance

Model Architecture	Metric	
	Training Accuracy	Testing Accuracy
AFMB-Net	0.9791	0.9519
FrameNorm	0.8538	0.7538

The FrameNorm architecture on the other hand gave us a testing accuracy of 75.38% and training accuracy of 85.38%. By the use of the Motion model, we aimed to extract maximal information between two consecutive frames in order to help us detect any inconsistencies over time (similar to the Heart Rate Analysis). The use of motion magnified videos as input allowed us to further enhance any inconsistencies between consecutive frames along the temporal and spatial domains. Similarly, the Appearance model was used to detect any inconsistencies within the spatial domain of a single frame. The use of motion magnified videos further helped detect any inconsistencies. This combination of Motion and Appearance model helped us classify deepfakes.

5 CONCLUSION

Through the course of our work, we have aimed to show the use of the heart rate analysis method in the detection of deepfakes. We believe that this method is fool-proof as the heart rates of individuals cannot be generated by GANs and are unique to each person. The proposed methodology is to implement a biological signal extractor in the form of PPG cells or spatio-temporal cells which is passed through different Machine Learning Models. The dataset used for our experimentation is the Celeb-DF dataset. This dataset provides high quality deepfake videos that are difficult to tell apart from real videos using the naked eye alone. Using the attention feature map, a machine learning method that is used to increase the accuracy of Convolutional Neural Networks, the MBConv Blocks for feature space extraction and developing long term dependencies, we see an increase in the accuracy achieved on the ResNet-18, 95%. The FrameNorm architecture however gave us a testing accuracy of 75.38%.

Some limitations encountered were:

- Objects in front of the Face: Any objects in front of the face will cause the model to not be able to identify/detect the face thus hindering the creation of MMST Map.
- MMST Map Dimensions: The size of the MMST Map is of the dimensions 300 × 25 × 3, this may have led to lower levels of accuracy than if the model had height and width of same or similar dimensions.
- As illustrated by the work carried out, it is apparent that our intuition on the use of heart rates for the detection of deepfakes is rightly justified and yields good results in the detection of deepfakes.

Notice

The paper was presented at the International Congress of Electrical and Computer Engineering (ICECENG'22), which took place in Bandırma (Turkey), on February 9-12, 2022. The paper will not be published anywhere else.

6 REFERENCES

[1] Ganiyusufoglu, I., Ngo, L., M., Savov, N., Karaoglu, S., & Gevers, T. (2020). Spatio-temporal features for generalized detection of deepfake videos. *arXiv*: 2010.11844. https://doi.org/10.48550/arXiv.2010.11844

[2] Nguyen, X. H., Tran, T. S., Nguyen, K. D., Truong, D. T., et al. (2021). Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. *Forensic Science International: Digital Investigation, 36,* 301108. https://doi.org/10.1016/j.fsidi.2021.301108

[3] Güera, D. & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. *The 15th IEEE international conference on advanced video and signal based surveillance (AVSS),* 1-6. https://doi.org/10.1109/AVSS.2018.8639163

[4] Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *arXiv*: 1905.00582. https://doi.org/10.48550/arXiv.1905.00582

[5] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face x-ray for more general face forgery detection.

CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5000-5009. https://doi.org/10.1109/CVPR42600.2020.00505

[6] Korshunov, P. & Marcel, S. (2018). Deepfakes: a new threat to face recognition? Assessment and detection. *CoRR.*

[7] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50× fewer parameters and 0.5 mb model size. *arXiv*: 1602.07360. https://doi.org/10.48550/arXiv.1602.07360

[8] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848-6856. https://doi.org/10.1109/CVPR.2018.00716

[9] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[10] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two stream neural networks for tampered face detection. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1831-1839. https://doi.org/10.1109/CVPRW.2017.229

[11] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-7. https://doi.org/10.1109/WIFS.2018.8630761

[12] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Using capsule networks to detect forged images and videos. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2307-2311. https://doi.org/10.1109/ICASSP.2019.8682602

[13] Hsu, C.-C., Lee, C.-Y., & Zhuang, Y.-X. (2018). Learning to detect fake face images in the wild. *International Symposium on Computer, Consumer and Control (IS3C),* 388-391. https://doi.org/10.1109/IS3C.2018.00104

[14] Nataraj, L., Mohammed, T. M., Manjunath, B., Chandrasekaran, S., Flenner, A., Bap-py, J. H., & Roy-Chowdhury, A. K. (2019). Detecting GAN generated fake images using co-occurrence matrices. *Electronic Imaging, 2019*(5), 532-1-532-7. https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532

[15] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2019, 1-8. https://doi.org/10.1109/BTAS46853.2019.9185974

[16] Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., & Liu, Y. (2019). FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI-20)*, 3444-3451. https://doi.org/10.24963/ijcai.2020/476

[17] Rossler, D., Cozzolino, L., Verdoliva, C., Riess, J., Thies, M., & Niessner, M. (2019). *FaceForensics++:* Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1-11. https://doi.org/10.1109/ICCV.2019.00009

[18] Kwon, P., You, J., Nam, G., Park, S., & Chae, G. (2021). KoDF: A large-scale Korean deepfake detection dataset. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10724-10733. https://doi.org/10.1109/ICCV48922.2021.01057

[19] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y.-G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2382-2390. https://doi.org/10.1145/3394171.3413769

[20] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nahavandi, S. et al. (2019). Deep learning for deepfakes creation and detection: A survey. *arXiv*: 1909.11573. https://doi.org/10.48550/arXiv.1909.11573

[21] Chen, B. & Tan, S. (2021). Featuretransfer: Unsupervised domain adaptation for cross-domain deep-fake detection. *Security and Communication Networks, 2021.* https://doi.org/10.1155/2021/9942754

[22] Borji, A. (2021). Pros and cons of GAN evaluation measures: New developments. *arXiv*: 2103.09396. https://doi.org/10.48550/arXiv.2103.09396

[23] Xuan, X., Peng, B., Wang, W., & Dong, J. (2019). On the generalization of GAN image forensics. *Chinese conference on biometric recognition*, 134-141. https://doi.org/10.1007/978-3-030-31456-9_15

[24] Kim, M., Tariq, S., & Woo, S. S. (2021). Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1001-1012. https://doi.org/10.1109/CVPRW53098.2021.00111

[25] Das, S., Seferbekov, S., Datta, A., Islam, M., Amin, M. et al. (2021). Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3776-3785. https://doi.org/10.1109/ICCVW54120.2021.00421

[26] Aneja, S. & Niessner, M. (2020). Generalized zero and fewshot transfer for facial forgery detection. *arXiv*: 2006.11863. https://doi.org/10.48550/arXiv.2006.11863

[27] Liu, Y., Stehouwer, J., Jourabloo, A., & Liu, X. (2019). Deep tree learning for zero-shot face anti-spoofing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4680-4689. https://doi.org/10.1109/CVPR.2019.00481

[28] Skibba, R. (2020). Accuracy eludes competitors in facebook deepfake detection challenge. *Engineering, 6*(12), 1339-1340. https://doi.org/10.1016/j.eng.2020.10.008

[29] Tariq, S., Jeon, S., & Woo, S. S. (2021). Am I a real or fake celebrity? Measuring commercial face recognition web APIs under deepfake impersonation attack. *arXiv*: 2103.00847. https://doi.org/10.48550/arXiv.2103.00847

[30] Asnani, V., Yin, X., Hassner, T., & Liu, X. (2021). Reverse engineering of generative models: Inferring model hyperparameters from generated images. *arXiv*: 2106.07873. https://doi.org/10.48550/arXiv.2106.07873

[31] Aneja, S., Bregler, C., & Niessner, M. (2021). COSMOS: Catching out-of-context misinformation with self-supervised learning. *arXiv*: 2101.06278. https://doi.org/10.48550/arXiv.2101.06278

[32] Röthlingshöfer, V., Sharma, V., & Stiefelhagen, R. (2019). Self-supervised face-grouping on graphs. *Proceedings of the 27th ACM International Conference on Multimedia*, 247-256. https://doi.org/10.1145/3343031.3351071

[33] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv*: 2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[34] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[35] Ciftci, U. A., Demir, I., & Yin, L. (2020). How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. *IEEE International Joint Conference on Biometrics (IJCB),* 1-10. https://doi.org/10.1109/IJCB48548.2020.9304909

[36] Caldelli, R., Galteri, L., Amerini, I., & del Bimbo, A. (2021). Optical flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters, 146*, 31-37. https://doi.org/10.1016/j.patrec.2021.03.005

[37] Fernandes, S., Raj, S., Ortiz, E., Vintila, I., Salter, M., Urosevic, G., & Jha, S. (2019). Predicting heart rate variations of deepfake videos using neural ode. *Proceedings of the IEEE/CVF Inter-national Conference on Computer Vision Workshops*. https://doi.org/10.1109/ICCVW.2019.00213

[38] Jeon, H., Bang, Y., & Woo, S. S. (2020). FDFTnet: Facing off fake images using fake detection fine-tuning network. *IFIP International Conference on ICT Systems Security and Privacy Protection. Springer*, 416-430. https://doi.org/10.1007/978-3-030-58201-2_28

[39] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 3207-3216. https://doi.org/10.1109/CVPR42600.2020.00327

[40] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[41] Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2020). DeepFakesON-Phys: Deepfakes detection based on heart rate estimation. *arXiv*: 2010.00400. https://doi.org/10.48550/arXiv.2010.00400

[42] Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., & Zhao, J. (2020). DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms. *Proceedings of the 28th ACM International Conference on Multimedia*, 4318-4327. https://doi.org/10.1145/3394171.3413707

Authors' contacts:

A. Vinay, Professor
PES University,
Ring Road Campus, Bangalore, Karnataka, India
a.vinay@pes.edu

Nipun Bhat
PES University,
Ring Road Campus, Bangalore, Karnataka, India
bhatnipun29@gmail.com

Paras S. Khurana
(Corresponding author)
PES University,
100 Feet Ring Road, BSK III Stage, Ring Road Campus,
Bangalore-560085, Karnataka, India
paras.s.khurana@gmail.com

Vishruth Lakshminarayanan
PES University,
Ring Road Campus, Bangalore, Karnataka, India
vishruth2270@gmail.com

Vivek Nagesh
PES University,
Ring Road Campus, Bangalore, Karnataka, India
vivek.nagesh1@gmail.com

S. Natarajan, Doctor
PES University,
Ring Road Campus, Bangalore, Karnataka, India
natarajan@pes.edu

T. B. Sudarshan, Doctor
PES University,
Ring Road Campus, Bangalore, Karnataka, India
sudarshan@pes.edu