Fused Swish-ReLU Efficient-Net Model for Deepfakes Detection

Hafsa Ilyas
Software Engineering Department
University of Engineering and Technology
Taxila, Pakistan
hafsailyas97@gmail.com

Ali Javed
Software Engineering Department
University of Engineering and Technology
Taxila, Pakistan
ali.javed@uettaxila.edu.pk

Muteb Mohammad Aljasem
Electrical Engineering Department
Wayne State University
Detroit, MI, USA
muteb.aljasem@wayne.edu

Mustafa Alhababi
Electrical Engineering Department
Wayne State University
Detroit, MI, USA
FT7950@wayne.edu

2023 9th International Conference on Automation, Robotics and Applications (ICARA) | 978-1-6654-8921-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICARA56516.2023.10125801

***Abstract*—With the rapid development of sophisticated deepfakes generation methods, the realism of fake content has reached the level where it becomes difficult for human eyes to identify such high-quality fake images/videos, thus increasing the demand for developing deepfakes detection methods. The diversity in deepfakes images/videos in terms of ethnicity, illumination condition, skin tone, age, background setting, and generation algorithms makes the detection task quite difficult. To better address the aforementioned challenges, we present a novel Swish-ReLU Efficient-Net (SRE-Net) that is robust to the identification of deepfakes generated using different face-swap and face-reenactment techniques. More precisely, we fused two EfficienNet-b0 models, one with the ReLU and the other with the Swish activation function along with layer freezing to achieve better detection results. Our SRE-Net attained the average accuracy and precision of 96.5% and 97.07% on the FaceForensics++ dataset, and 88.41% and 91.28% on the DFDC-preview dataset. The high detection results demonstrate the effectiveness of SRE-Net while detecting the deepfakes generated using different manipulation algorithms.**

Keywords*—Deepfakes detection, fused Swish-ReLU Efficient-Net, FaceForensics++, DFDC-preview

I. Introduction

Over the past few years, the generation of fake multimedia content including fake images and videos commonly known as “Deepfakes” is increasing tremendously. Deepfakes can be categorized as face-swap, face-reenactment, lip-syncing, and entire face synthesis. Generative adversarial networks and autoencoders based deepfakes generation approaches have significantly improved the realism and quality of fake content. Such realistic deepfakes images/videos are now becoming indiscriminate for the human eyes. Additionally, there are many open-source software and applications (i.e., FaceSwap Live, DeepFace Lab [1], Reface, etc.) that enable the non-tech person to create deepfakes in the real world. In this age of information technology, where social media platforms like Facebook, Twitter, Instagram, Youtube, etc., are the greatest source of information, such fake multimedia content can be spread in no time and may have a serious social and political impact. Deepfakes can be used to fool the face and speech recognition systems, tamper the digital evidence, spread false information, impersonate and defame renowned personalities and disrupt the political processes [2].

To counter the threats associated with visual deepfakes generation, various deep learning-based methods are developed in recent years that have attained good performance for detecting manipulated images or videos. For instance, Sabir et al. [3] introduced the approach that combines face alignment with recurrent neural network (RNN) and evaluated it on the FaceForensics++ (FF++) dataset. This approach [3] provides good detection results for static images but is unable to perform better on videos [2]. Similarly, Montserrat et al. [4] introduced a pipeline having 3 components: (i) face detection with multi-task cascaded convolution neural network (MTCNN), (ii) feature extraction using convolution neural network (CNN), and (iii) prediction using Automatic Face Weighting with gated recurrent unit (GRU). The approach [4] was evaluated on Deepfake Detection Challenge dataset (DFDC) and the reported accuracy was 91.88%. In [5], a five-layered 3D convolutional neural network (3DCNN) was presented that learned the spatio-temporal face features for the detection of facial manipulation. 3DCNN was only evaluated on FaceSwap, DeepFakes, and Face2Face subsets of the FF++ dataset. Two deep learning models to detect face swap manipulation from a single image based on differences between the face and its context were presented in [6]. One of the models was trained on the face and the other on the context (e.g., ear, hair, etc.). This approach achieved an accuracy of 75% on Celeb-DF and 66% on the FF++ dataset. In [7], a convolutional vision transformer (CviT) was presented that learned the local and global features of the fake and real images using the attention mechanism. CviT was evaluated on the FF++ dataset and attained the highest accuracy of 93% on the DeepFakes subset. However, CviT is unable to provide good detection results for other subsets of the FF++ dataset.

Deepfakes detection is still a challenging task since the deepfakes generation techniques are getting improved day by day. So, the need for detection methods exists that are able to detect the fake multimedia content generated using various deepfakes generation algorithms such as FaceSwap, NeuralTextures, FaceShifter, Face2Face, etc. Therefore, in this research work, we proposed a novel fused Swish-ReLU Efficient-Net (SRE-Net) for the accurate detection of different types of deepfakes. The fusion of two models will help to better

This work was supported by the grant of Punjab Higher Education Commission of Pakistan via Award No. (PHEC/ARA/PIRCA/20527/21).

978-1-6654-8921-8/23/$31.00 ©2023 IEEE

Authorized licensed use limited to: OAKLAND UNIVERSITY. Downloaded on May 29,2023 at 15:22:12 UTC from IEEE Xplore. Restrictions apply.

capture the distinct features in the fake and real images and ultimately helps to improve the detection accuracy.

The key contributions of our research work are:

- We propose a novel and robust Swish-ReLU Efficient-Net that employs a fusion of EfficientNet-b0 with ReLU activation and EfficientNet-b0 with Swish activation along with layer freezing to reliably detect the deepfakes.
- Our proposed model is able to detect different types of deepfakes including face-swap and face-reenactment with challenging conditions of varied manipulation techniques, diverse illumination conditions, non-face frames, side-posed faces, different skin tones, gender, and age.
- We performed extensive experiments on FF++ and DFDC-preview datasets to demonstrate the effectiveness of our model against the existing deepfakes detection methods.

II. Proposed Methodology

A. *Pre-processing*

In the pre-processing, we extracted the frames from videos and then used MTCNN [8] for detecting the faces from each frame. According to our model's requirement, the detected face images are then resized to 224×224 resolution having three channels. These resized facial frame images are then fed to our SRE-Net for the detection of deepfakes.

B. *Architecture of Proposed SRE-Net*

In comparison to the traditional CNN models, EfficientNet models proved to be more accurate and have shown good performance in transfer learning tasks [9]. The compound scaling coefficients used in EfficientNet models allow the uniform scaling of depth, width, and resolution. From the EfficientNet models, the EfficientNet-b0 version has the lowest parameters and floating-point operations [9]. So, based on the fastest speed and lowest computational complexity, we utilized EfficientNet-b0 in our novel fused SRE-Net for deepfakes detection. Our fused model has a total of 8.75 million parameters which is less than most of the existing CNN models including Xception, Inception-v3, InceptionResent-v2, DenseNet, and EfficientNet (b2-b7). EfficientNet-b0 is a robust and computationally efficient model based on the linear and inverted bottleneck residuals of MobileNet-v2 along with squeeze and excitation (SE) blocks. The proposed fused Swish-ReLU Efficient-Net model is shown in Fig. 1.

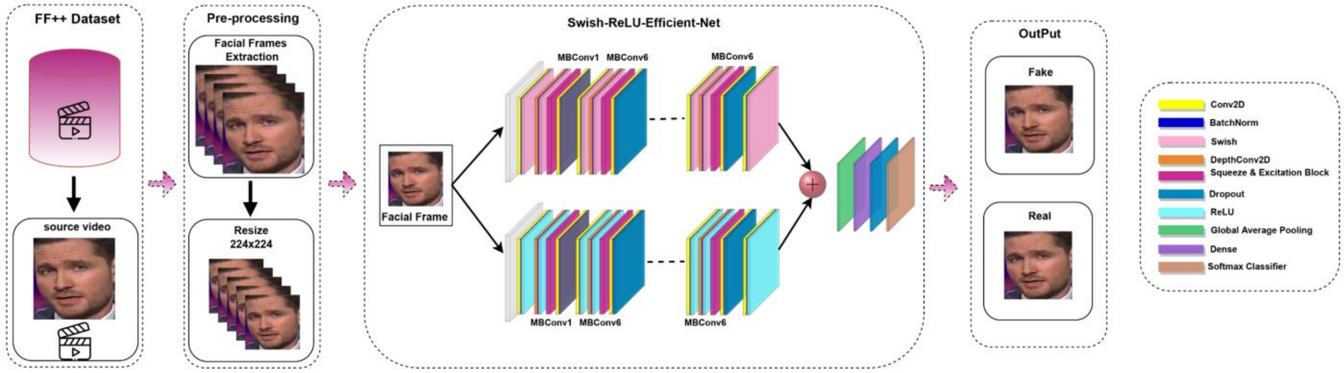

Fig. 1. Architecture Diagram of Swish-ReLU Efficient-Net.

Our proposed fused model includes two pre-trained EfficientNet-b0 models named, Model S and Model R. These two models are different in terms of the activation function used in them. The Model S is composed of EfficientNet-b0 with Swish activation function which helps in the strong regularization, saturation avoidance, improves optimization, robustness, and gradient flow. The Swish activation function can be computed as:

$$f(x) = x \times Sigmoid\ (\beta x) \tag{1}$$

The initial layers of Model S are frozen, which means that the model preserves the features of ImageNet and is partially trained on our deepfakes datasets. The layer freezing also helps in reducing the computation time of the model. Whereas in Model R, we introduced the ReLU activation function in EfficientNet-b0. ReLU is the fast and simple, widely used activation function in CNN models. The gradients of the ReLU activation function are not saturated and help to accelerate the convergence compared to other activation functions (i.e., Tanh, Sigmoid). The ReLU activation function is computed as:

$$f(x) = max\ (0, x) \tag{2}$$

The Model R is fully trained on deepfakes datasets used for experiments as the layers of the model are not frozen. These two models (Model S and Model R) having different activation functions are fused using the concatenation method. The models fusion help in the extraction of the features that are rich and preserve more information thus help in the accurate detection of deepfakes. After that, new layers including global average pooling, dense, and dropout are integrated into the model. The global average pooling layer can better interpret the features obtained from the fused model by imposing a connection between feature maps and categorization. While the dropout layer helps to overcome the overfitting problem. Finally, a

Authorized licensed use limited to: OAKLAND UNIVERSITY. Downloaded on May 29,2023 at 15:22:12 UTC from IEEE Xplore. Restrictions apply.

classification layer with a softmax classifier is employed, which computes the cross-entropy loss for the final prediction.

III. Experiments and Results

The performance of the proposed model is evaluated on standard, diverse, and challenging FaceForensics++ [10] and DFDC-preview [11] datasets. For the performance evaluation, standard metrics such as accuracy, precision, and recall are computed.

A. Datasets

FF++ dataset contains 1000 real videos and five subsets of fake videos each comprising 1000 videos. The fake videos are generated using the five different deepfakes generation algorithms named as DeepFakes, FaceSwap, Face2Face, NeuralTextures, and FaceShifter. For the evaluation of our proposed model, we split the videos in each folder into training, validation, and test sets (720/140/140).

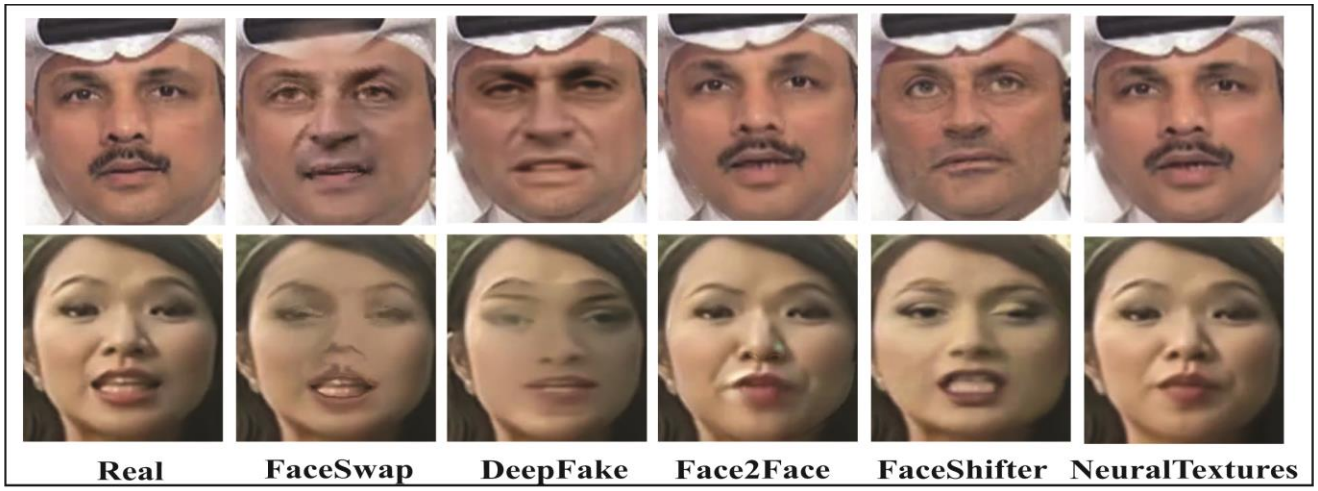

Fig. 2. FaceForensics++ Dataset.

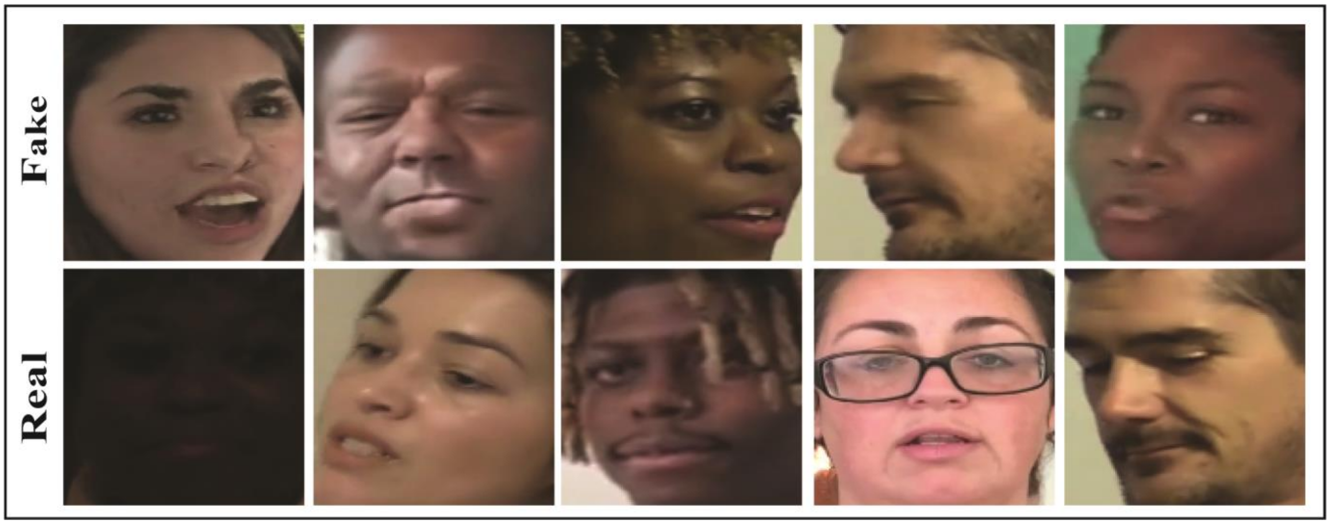

Fig. 3. DFDC-preview Dataset.

DFDC-preview is the preview of the challenging DFDC dataset. This dataset is composed of real and fake videos of actors having different gender, skin tones, ages, varied illumination conditions, and background settings. For our experimentation, we divided the dataset into two non-overlapping subsets i.e., training and testing, with a split ratio of 80:20. After splitting the datasets, we extracted the frames from each video of the datasets for the evaluation of our fused SRE-Net. A few samples of FF++ and DFDC-preview datasets are shown in Fig. 2 and Fig. 3, respectively.

B. Performance Evaluation

To evaluate the performance of SRE-Net on deepfakes generated via different algorithms, we designed a five-stage binary classification experiment on the FF++ dataset. For this purpose, at each stage, we used the 1000 real videos and 1000 fake videos from one of the fake subsets of the FF++ dataset. The videos are splitted into training, validation, and testing sets. In the first stage, we evaluated our model on FaceSwap and Real subsets of the FF++ dataset. Likewise, in the next four stages, we consider the fake videos from the remaining fake subsets (DeepFakes, Face2Face, NeuralTextures, FaceShifter)

Authorized licensed use limited to: OAKLAND UNIVERSITY. Downloaded on May 29,2023 at 15:22:12 UTC from IEEE Xplore. Restrictions apply.

separately. The performance of the SRE-Net on the FF++ dataset is shown in Table I.

TABLE I. PERFORMANCE EVALUATION

Subsets of FF++ Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
FaceSwap	98.89	99.42	98.36
DeepFakes	97.14	94.62	99.93
Face2Face	98.5	99.85	97.14
NeuralTextures	92.11	91.84	92.43
FaceShifter	96.07	99.62	92.50

It is important to note that our proposed model attained the highest accuracy reaching 99% on the FaceSwap subset of the FF++ dataset. Since the NeuralTextures algorithm introduced very few semantic changes and only the mouth region is altered while generating fake videos/images whereas the FaceShifter algorithm generates challenging fake content via fusing two complex GAN's models. Thus, the detection of fake multimedia content generated using NeuralTextures and FaceShifter algorithms are quite difficult. Our fused model achieved the lowest accuracy of 92.11% on NeuralTextures generated deepfakes which is relatively good. The accuracy, precision, and recall above 90% on all the subsets of the FF++ dataset indicate that our model has the capability to detect deepfakes generated using a variety of face-swap and face-reenactment manipulation techniques.

To show the robustness of our model on diverse and challenging deepfakes dataset, we evaluated our method on the DFDC-preview dataset. After the model is trained on the facial frames in the training subset, we tested it on the test set to evaluate the performance. Our proposed model attained an accuracy of 88.41%, precision of 91.28%, and recall of 94.27% on the DFDC-preview dataset. Such good results demonstrate that our method is capable of detecting deepfakes in the presence of challenging conditions such as the extremely side-posed face, poor illumination condition, varied skin tones, gender, and age.

C. Comparison with Existing Methods

To demonstrate the effectiveness of the proposed SRE-Net against existing state-of-the-art methods, we compared the results in terms of accuracy on the FF++ and DFDC-preview datasets. The comparative analysis for DFDC-preview and FF++ is shown in Tables II and III, respectively. The missing numbers in Table III indicate that the results are not provided on these subsets of the FF++ dataset by the corresponding methods.

TABLE II. COMPARATIVE ANALYSIS FOR DFDC- PREVIEW DATASET

Models	**Accuracy (%)**
FInfer [15]	80.39
CNN-LSTM [16]	81.91
CWSA net [16]	85.28
HCiT [14]	89.73
SRE-Net (Proposed)	88.41

From Table II, it can be observed that our model attained the highest detection accuracy on DFDC-preview dataset except from the one model [14]. Detection of deepfakes videos included in the DFDC-preview dataset is quite challenging as the videos include extremely side-posed faces, poor illumination conditions, and realistic fake faces. The detection accuracy of SRE-Net is very close to the best-performing model [14], which reveals that our proposed model can detect the deepfakes of the challenging dataset quite accurately.

It can be inferred from Table III that our model attained an average accuracy gain of 3.14% for DeepFakes subset, 9.95% for FaceSwap subset, 12% for Face2Face, 50% for FaceShifter, and 22% for NeuralTextures subset. Achieving such good results on all subsets of the FF++ dataset is challenging in the presence of non-face frames, varying illumination conditions, and different deepfakes generation methods. However, our method outperforms all the existing methods while detecting different kinds of deepfakes included in the FF++ dataset.

D. Comparison with Deep Learning Models

To show that our fused SRE-Net performs better than existing deep learning models, we conducted an experiment where we compared the performance of our model on FF++ dataset against the deep learning models including XceptionNet, DenseNet, VGG, ResNet50, and Inception-v3. In Table IV, comparison results in terms of accuracy are presented.

TABLE IV. COMPARISON WITH EXISTING DEEP LEARNING MODELS

Deep Learning Models	**Accuracy (%) on FF++ dataset**	**No. of Parameters**
DenseNet	93	14.3 M
Xception	94	22.9 M
VGG	95	138.4 M
Inception-v3	95.5	23.9 M
ResNet	92	25.6 M
SRE-Net (Proposed)	96.5	8.75 M

TABLE III. COMPARATIVE ANALYSIS FOR FF++ DATASET IN TERM OF ACCURACY

Models	**Subsets of FF++ Dataset**				
	DeepFakes (%)	***FaceSwap (%)***	***Face2Face (%)***	***FaceShifter (%)***	***NeuralTextures (%)***
Sabir et al. [3]	96.9	96.3	94.35	--	--
3DCNN [5]	91.63	86.95	87.85	--	--
Khalifa et al.[12]	93.30	91.96	93.75	--	77.10
CviT [7]	93	69	--	46	60
Demir et al. [13]	93.28	91.62	59.69	--	57.02
HCiT [14]	96	97.82	95.85	--	86.29
SRE-Net (Proposed)	97.14	98.89	98.5	96.07	92.11

Authorized licensed use limited to: OAKLAND UNIVERSITY. Downloaded on May 29,2023 at 15:22:12 UTC from IEEE Xplore. Restrictions apply.

Table IV shows that SRE-Net attained an average detection accuracy of 96.5% which is higher than the comparative deep learning models. So comparatively, our fused model has achieved the highest deepfakes detection accuracy with a smaller number of parameters. The Inception-v3 attained the accuracy of 95.5% closer to our SRE-Net but at the expense of a greater number of parameters. Thus, from the comparison in Table IV, we can conclude that our fused SRE-Net is computationally efficient and provides higher accuracy for the detection of deepfakes as compared to the existing deep learning models.

IV. Conclusion

In this paper, we have proposed a novel fused Swish-ReLU Efficient-Net for the detection of deepfakes that is reliable and robust to the variation of deepfakes generation algorithms. More precisely, our method detects the face-swap and face-reenactment deepfakes generated via different approaches. Our novel fused SRE-Net comprises the fusion of two variants of EfficientNet-b0 models having different activation functions. The proposed model is evaluated on FaceForensics++ and DFDC-preview datasets and shows outstanding detection performance irrespective of challenging conditions such as deepfakes generated via diverse algorithms having varied illumination conditions, skin tones, and non-face frames. Our proposed fused SRE-Net not only accurately classifies the deepfakes but also shows remarkable detection performance against the existing methods. We also compared the detection accuracy and number of parameters of our model with existing deep learning models. In the future, we will extend our research for detecting the deepfakes generated through unseen manipulation algorithms and also develop a method that is robust towards adversarial attacks while detecting deepfakes.

Acknowledgment

This work was supported by the grant of Punjab Higher Education Commission (PHEC) of Pakistan via Award No. (PHEC/ARA/PIRCA/20527/21).

References

[1] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C.S. Facenheim, L. RP, J. Jiang, and S. Zhang, “DeepFaceLab: Integrated, flexible and extensible face-swapping framework.” arXiv preprint arXiv:2005.05535, 2020

[2] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. J. A. I. Malik, "Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward," pp. 1-53, 2022.

[3] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. J. I. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," vol. 3, no. 1, pp. 80-87, 2019.

[4] D. M. Montserrat et al., "Deepfakes detection with automatic face weighting," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 668-669.

[5] A. Kohli, A. J. M. T. Gupta, and Applications, "Light-weight 3DCNN for DeepFakes, FaceSwap and Face2Face facial forgery detection," pp. 1-13, 2022.

[6] Y. Nirkin, L. Wolf, Y. Keller, T. J. I. T. o. P. A. Hassner, and M. Intelligence, "DeepFake detection based on discrepancies between faces and their context," 2021.

[7] D. Wodajo and S. J. a. p. a. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021.

[8] K. Zhang, Z. Zhang, Z. Li, and Y. J. I. s. p. l. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," vol. 23, no. 10, pp. 1499-1503, 2016.

[9] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning, 2019, pp. 6105-6114: PMLR.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1-11.

[11] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. J. a. p. a. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," 2019.

[12] A. H. Khalifa, N. A. Zaher, A. S. Abdallah, and M. W. J. I. A. Fakhr, "Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition," vol. 10, pp. 22678-22686, 2022.

[13] I. Demir and U. A. Ciftci, "Where do deep fakes look? synthetic face detection via gaze tracking," in ACM Symposium on Eye Tracking Research and Applications, 2021, pp. 1-11.

[14] B. Kaddar, S. A. Fezza, W. Hamidouche, Z. Akhtar, and A. Hadid, "HCiT: Deepfake video detection using a hybrid model of CNN features and vision transformer," in 2021 International Conference on Visual Communications and Image Processing (VCIP), 2021, pp. 1-5: IEEE.

[15] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. J. I. T. P. A. M. I. Qin, "FInfer: Frame Inference-based Deepfake Detection for High-Visual-Quality Videos," pp. 1-9, 2022.

[16] Y. Lu, Y. Liu, J. Fei, Z. J. S. Xia, and C. Networks, "Channel-Wise Spatiotemporal Aggregation Technology for Face Video Forensics," vol. 2021, 2021.

Authorized licensed use limited to: OAKLAND UNIVERSITY. Downloaded on May 29,2023 at 15:22:12 UTC from IEEE Xplore. Restrictions apply.