EfficientNetB0 Ensemble Model for Unified Deepfakes Detection

Samiya Afzal Minhas
Software Engineering Department
University of Engineering and Technology
Taxila, Pakistan
19-SE-47@students.uettaxila.edu.pk

Saad Mushtaq*
Software Engineering Department
University of Engineering and Technology
Taxila, Pakistan
19-SE-68@students.uettaxila.edu.pk

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

2023 17th International Conference on Open Source Systems and Technologies (ICOSST) | 979-8-3503-8132-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOSST60641.2023.10414228

***Abstract*—In recent years, we have witnessed the generation of exceptional authentic deepfake images and videos due to the availability of cutting-edge Artificial Intelligence and deep learning techniques. Deepfakes represent synthetic multimedia content used to propagate disinformation for defamation, political unrest, manipulating elections, committing crimes, etc. In this paper, we present a novel ReLU-Swish EfficientNet (RSE-Net) for deepfakes detection. Our proposed RSE-Net is capable of reliably detecting deepfakes videos that are generated using different techniques. Our model leverages an ensemble of EfficientNet architectures, which are combined using a fusion technique to enhance the model's performance in detecting deepfakes. We suggested the ReLU activation in conv2D layers in place of regular Swish activation in EfficientNetB0 first variant as ReLU is computationally more efficient and reduces the risk of overfitting. We evaluated our model on two large-scale challenging deepfake datasets: FaceForensics++ and CelebDF. Our RSE-Net attained an average accuracy of 99.7% on the FaceForensics++ dataset, and 96.09% on the CelebDF dataset. Furthermore, our model generalizes well and effectively detects deepfake videos in real-world scenarios. Thus, it is a valuable tool for analyzing and detecting potentially manipulated content.**

***Index Terms*—Deepfake detection, Celeb-DF, FaceForensics++, Ensemble model, Fused-EfficientNet**

I. Introduction

Deepfake technology has developed quickly in recent years, making it possible to produce manipulated media that is incredibly realistic and convincing. Deepfakes pose serious risks to various fields, including politics, entertainment, and personal privacy. Deepfakes are artificially produced images, videos, or audio recordings that show people saying or doing things they never actually did [1]. In the digital age, it has become necessary to identify and mitigate the negative impacts of deepfakes. Deepfake technology is becoming more prevalent, which raises questions about the integrity and validity of multimedia content. The possibility of false information, reputational harm, and social engineering attacks grow more worrisome with the ability to produce fake media that is practically indistinguishable from real footage. The rapid spread of deep fakes through social media platforms and open-source tools [2], [3] increase their impact. Hence it is essential to advance effective procedures for spotting and thwarting these deceptions.

Deepfake detection has been a topic of significant interest in recent years. Agarwal et al. (2019) were among the pioneers and proposed a method using Recurrent Neural Networks (RNN) to detect deepfakes by focusing on temporal video inconsistencies. This work laid the foundation for Rössler et al. (2019), who developed a large-scale video dataset, FaceForensics++, to benchmark facial manipulation detection methods, emphasizing the importance of diverse and high-quality datasets for training detection models [4]. Building on these baseline works, Li et al. (2018) suggested a strategy to detect face-warping artifacts by focusing on spatial rather than temporal inconsistencies [5]. Nguyen et al. [6] further explored this spatial-temporal dichotomy (2019) and provided a comprehensive review of deep learning techniques for creating and detecting deepfakes. This study also highlights the ongoing arms race between deepfake generation and detection. Tolosana et al. (2020) surveyed a broader range of face manipulative methods and their detection and combined spatial, temporal, and other cues [7]. The study emphasized the necessity of employing multi-modal detection methods. Nguyen et al. (2020) extended this multi-model approach by introducing the use of Capsule Networks for deepfake detection. This approach could potentially capture complex spatial and temporal patterns more effectively than traditional CNNs or RNNs [8]. Recent works have also explored innovative approaches to deepfake detection. Kuang et al. (2021) proposed a dual-branch neural network that can detect inconsistencies both spatially and temporally for Deepfake video detection [9]. The Bi-LSTM network was employed [10] to capture the temporal inconsistencies of optical flow and the EfficientNet [11] is used to extract optical flow properties.

Deepfake detection is still a difficult problem due to several challenges that arise from the diverse aspects involved in creating and detecting deepfakes. Some of the reasons are rapidly evolving deepfake technology, generative models, diversity in deepfake creation, occlusions, tilted faces, facial pose, race, and gaze. Therefore, there exists a need for robust detection techniques that can detect altered media produced by different deepfakes generation techniques. We introduced a novel ReLU-Swish EfficientNet (RSE-Net) that can reliably counter deepfakes generated from multiple techniques. Our work's significant contributions are:

- We introduce a novel ReLU-Swish EfficientNet that utilizes an ensemble approach with EfficientNetB0 as the base architecture to accurately recognize various deepfakes.
- Our proposed model demonstrates the capability to identify various forms of deep fakes under different manipulation techniques, diverse lighting conditions, side profile faces, and individuals with various skin tones, genders, and ages.
- We conducted numerous tests on the FF++ and CelebDF datasets, to show the effectiveness of the proposed model against the current deep fakes detection techniques.

II. Proposed Method

The proposed deepfakes detection method is discussed in this section.

A. EfficientNetB0 Base Architecture

EfficientNetB0 is a CNN architecture [13] based on a mobile inverted bottleneck convolution (MBConv). It is an improved version of the inverted residuals and linear bottlenecks structure used in MobileNetV2. The EfficientNetB0 architecture starts with a stem, which is a superficial convolutional layer. This is followed by several MBConv layers with varying filter sizes and numbers of filters. Each MBConv layer consists of a squeeze-and-excitation (SE) module, which progressively adjusts features according to the channel. The MBConv layers are grouped into seven blocks, each with multiple filters and repeats. At the end of the architecture, there is a global average pooling, a dropout layer, and a fully connected layer for finetuning the model.

B. Proposed Method Architecture

This research work proposes a novel deep-learning model that utilizes an ensemble approach with EfficientNetB0 as the base architecture. Figure 1 shows the architecture of our proposed model. In Model A, we introduce a modification to the activation function in the conv2D layers of the EfficientNetB0 architecture. We replaced the Swish activation function with the ReLU activation function as it is computationally more efficient and reduces the risk of overfitting by suppressing negative values. All layers in model A are set to be trainable. The mathematical representation of ReLU is:

$$f(z) = max(0, z). \qquad (1)$$

In Model B, we froze the trainable layers to prevent them from being updated during the training process. By freezing these layers, we focus on utilizing the learned representations from Model A while reducing the overall number of trainable parameters. This approach helped us to avoid overfitting and enhanced the model's generalization capability. Then the outputs of Model A and Model B were concatenated into a single pipeline. This fusion leveraged the complementary information extracted by both models and exploited their collective knowledge for improved deepfake detection. By merging the two models into a single ensemble, we effectively combine their strengths and enhance the performance of the system .

The outputs of these two models are then combined followed by a GlobalAveragePooling layer. This is then connected to a Dense layer with 512 units and a 'ReLU' activation function. A Dropout layer is introduced for regularization, and then a Dense layer with a 'softmax' activation function is used for multi-class classification. Adam optimizer and a categorical cross-entropy loss function that is appropriate for multi-class classification problems are used in the model's training. The learning rate is reduced when the validation accuracy stops improving, which is a common technique to achieve better convergence.

1) Pooling Layer: This layer decreases the dimensionality of the key points vector by removing unnecessary information. The pooling layer helps the suggested model prevent by aggregating pixel information and enhances its ability to handle spatial translations in the input. In the proposed method, the computed features are obtained from the average pooling layer and fed into the additional dense layers.

2) Dense Layer: Following the layer of pooling, we included a single dense layer at the end of the model architecture. It has a ReLU activation function and 512 neurons. It is responsible for highlighting the altered areas in the input samples by eliminating undesirable information and improving the accuracy of detecting visual deepfakes. The addition of a dense layer enhances the ability of the proposed model to compute a more effective collection of aspects of the video frame. Finally, the extracted information is passed to the softmax layer in the form of deep key points.

3) Softmax Layer: It is the final layer of the proposed model which is used for deep fakes classification. Our proposed model employs the softmax activation function in the last fully connected layer to convert the outputs of the previous layer into a probability distribution.

III. Experiments and Results

A. Datasets

In this study, we used two benchmark datasets for evaluating our deepfake detection model: FaceForensics++ and Celeb-DF. FaceForensics++ is a widely-used dataset [12] in deepfake detection research. It consists of manipulated and original videos, providing a diverse set of samples for training and evaluating deepfake detection algorithms. The dataset is generated using various manipulation methods, making it an ideal resource for benchmarking deepfake detection techniques.

The CelebDF dataset [13] consists of 590 original YouTube videos with people of various ages, ethnicities, and genders. The dataset includes 5,639 corresponding deepfake videos produced using an improved synthesis procedure. The high visual quality of the deepfake videos in the Celeb-DF dataset presents a more challenging task for detection algorithms,

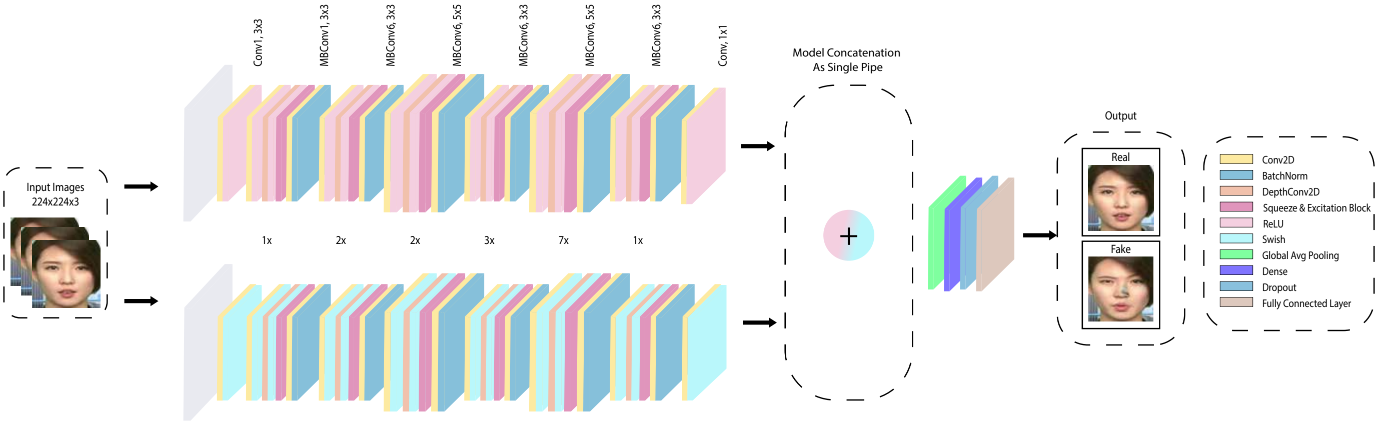

Fig. 1. Proposed Model Architecture.

making it a valuable resource for evaluating deepfake detection models.

B. Evaluation of Proposed Model

To evaluate the effectiveness of the proposed RSE-Net model, it was trained on five different subsets of the Face-Forensics++ dataset so that it can detect deepfake videos that are generated through different techniques like face swap, deepfakes, face2face, face shifter, and neural texture. The training dataset was divided into train and validation subsets with ratios of 80 and 20 respectively. . We used balanced training and validation subsets that had an equal number of videos in each class. The hyper-parameters that were used for the training of our model are batch size of 16, epochs set to 25, and learning rate of 0.0001. We used accuracy as an evaluation metric and categorical cross entropy for calculating the loss. Table I shows the precision, recall, and accuracy value of the model. Our proposed RSE-Net achieved the highest accuracy of 98.8% on the deepfakes subset and overall accuracy, precision, and recall greater than 95% on every subset. Only minor semantic changes and mouth regions are modified in fake videos that are generated through the neural texture technique. The Faceshifter algorithm uses a fusion of two generative adversaries to produce fake videos. The accuracy of 97.8% and 97.1% on the face shifter and neural texture subset indicate the effectiveness of the model against the challenges of deepfake detection of forgeries like neural texture and face shifter.

TABLE I
MODEL EVALUATION USING FACEFORENSICS++ DATASET

Subset	Accuracy %	Precision %	Recall %
Deepfakes	98.8	98.23	99.2
FaceSwap	98.7	98.9	98.4
Face2Face	97.4	97.4	97.4
FaceShifter	97.8	97.6	98.1
NeuralTextures	97.1	97.1	97.1

We also trained our model on the CelebDF dataset to check its effectiveness on challenging and benchmark datasets. We divided the training dataset into train and validation datasets with ratios of 80 and 20 respectively. Both sets had two classes namely real and fake. The same hyper-parameters were used for training as mentioned in the above paragraph. Our proposed RSE-Net achieved 99.65%, 97.3%, and 96.09% training, validation, and testing accuracy respectively on the CelebDF dataset. These positive results show that our model can reliably identify deep fakes even in the presence of tough situations including a very side-protruded face, bad lighting, a wide range of skin tones, and a variety of gender and age.

C. Comparative Analysis

We contrasted the accuracy of the proposed RSE-Net with existing techniques to show its effectiveness against existing works [14]–[20] that have used different subsets of FF++. This comparative analysis is shown in Table II. From Table II we conclude that our RSE-Net obtained an average gain of 6.03%, 12.25%, 24.99%, 65.37%, and 22.08% for Deep-Fakes, FaceSwap, Face2Face, FaceShifter, and NeuralTextures subsets, respectively. This indicates the effectiveness of our model against different deepfake generation methods and challenging conditions. And proposed model also surpasses all other existing methods in terms of accuracy on the FF++ dataset.

D. Comparison Of Different Activation f(x) in Model-A and Dense Layer

To compare the accuracy of the model with different activation functions, one at a time, in Conv2D Layer, we replaced the Swish activation function in the Conv2D layers of Model-A with other activation functions and compared model accuracy. Table III shows that the proposed model attained the highest accuracy with the ReLU activation function in the Conv2D layers of Model-A. As ReLU is simple, introduces non-linearity in the network, its gradient is unsaturated and provides a zero gradient for positive inputs which promotes better gradient flow during backpropagation.

TABLE II
COMPARISON WITH EXISTING APPROACHES ACCURACY ON FF++ DATASET SUBSETS.

Models	FF++ Subsets				
	DeepFakes	FaceSwap	Face2Face	FaceShifter	NeuralTextures
Zaher et al. [14]	93.30	91.96	93.75	-/-	77.10
CviT [15]	93	69	-/-	46	60
Xie et al. [16]	93.08	74.67	91.61	-/-	65
Khalid et al. [17]	86.20	86.10	71.20	-/-	95.30
Demir et al. [18]	93.28	91.62	59.69	-/-	57.02
Hafsa et al. [19]	93.39	93.01	92.11	84.91	78.6
Hafsa et al. [20]	97.14	98.8	98.5	96.07	92.07
Proposed Model	98.8	98.7	97.4	97.8	97.1

TABLE III
ACCURACY COMPARISON OF DIFFERENT ACTIVATION F(X) IN CONV2D LAYERS OF MODEL-A

Activation f(x)	Training	Validation	Test
Swish	97	88.75	86.8
ReLU	99.65	95.95	95.47
GeLU	98.7	89.1	84.95
SELU	98	94	93.75
ELU	99	92.6	91.5

To analyze the effect of changing ReLU in the dense layer with other activation functions on model accuracy, we replaced it with other activation functions and compared model accuracy. The model achieved an accuracy greater than 90% with different activation functions in its dense layer. Table IV shows that we achieved the highest accuracy with the ReLU activation function in the dense layer of the proposed model.

TABLE IV
ACCURACY COMPARISON OF DIFFERENT ACTIVATION F(X) IN DENSE LAYER

Activation f(x)	Training	Validation	Test
ReLU	99.65	95.95	95.47
GeLU	97.1	93.45	93.9
SELU	99	92.15	92.19
ELU	97	94.2	92.75
Tanh	98	93.65	92.125
Exponential	99	93.35	87.6
LeakyReLU	98.1	91.9	91.675
PReLU	99.05	96.45	95

E. Generalizability Evaluation

To analyze how well our proposed model generalizes against different subsets of faceforensics++ and datasets we conducted an experiment in which we tested every model on the subset that it was trained and every other subset of faceforensics++ dataset and CelebDF dataset itself. Table V shows that every model achieved an accuracy greater than 96% when it was tested against the subset on which it was trained. On the other hand, accuracy dropped to 55% in the case of cross-subset testing. The model trained on the subset NeuralTexture achieved an accuracy of 73.5% on the deepfake subset and 65% on the face-shifter subset. Overall, we got an average accuracy of 60% in this cross-subset experiment. This is because different subsets of faceforensics++ are generated through different deepfakes techniques and are very dissimilar from each other.

IV. CONCLUSION

This research work has presented a novel fused ReLU-Swish EfficientNet model architecture for enhancing the performance of deepfake detection. Extensive experimentation on CelebDF and FaceForensics++ datasets demonstrated that the proposed model achieved superior performance in distinguishing between genuine and manipulated videos over contemporary works. Our proposed model demonstrates the capability to identify various forms of deepfakes manipulated using different techniques, and under diverse lighting conditions, side profile faces, and individuals with different skin tones, genders, and ages. In addition to correctly classifying the deepfakes, our proposed model outperforms the current techniques in terms of detection performance. The proposed model holds significant potential for real-world applications, such as ensuring the integrity of media content and safeguarding against the malicious use of deepfake technology. Further research should focus on presenting more transparent deep learning models for deepfakes detection and improving cross-corpora evaluation.

ACKNOWLEDGMENT

The authors would like to thank the Multimedia Signal Processing (MSP) research lab at UET-Taxila Pakistan and the SMILES lab at Oakland University Rochester MI, USA for providing the lab resources to conduct this research.

REFERENCES

[1] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 53(4):3974–4026, 2023.

[2] FaceSwap. FaceSwap.dev, URL https://faceswap.dev/. Open Source multi- platform Deepfakes software.

[3] iperov. DeepFaceLab. Github Repo, URL https://github.com/iperov/DeepFaceLab/. the leading software for creating deep fakes.

[4] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1–11, 2019.

[5] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face-warping artifacts. arXiv preprint arXiv:1811.00656, 2018.

[6] Thanh Thi Nguyen, Cuong M Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. arXiv preprint arXiv:1909.11573, 1(2):2, 2019.

[7] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and Beyond: A survey of face manipulation and fake detection. Information Fusion, 64:131–148, 2020.

TABLE V
ACCURACY COMPARISON WITH CROSS-DATASET EVALUATION.

Dataset	Models					
	DeepFakes	FaceSwap	Face2Face	FaceShifter	NeuralTextures	CelebDF
Deepfake	98.8	55.25	57.35	57.05	73.5	59.4
FaceSwap	54.9	98.7	55.5	55.2	47.75	55.2
Face2Face .	55.1	55.9	97.4	55.3	60.85	55.15
FaceShifter-	57.8	55.35	54.9	97.8	65	54.75
NeuralTextures	55.75	54.75	55.4	56.15	97.1	50.2
CelebDF	57.05	55.15	56	52.425	57.57	96.095

[8] Huy H Nguyen, Junichi Yamagishi, and Isao 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), pages 2307–2311. IEEE, 2019.

[9] Liang Kuang, Yiting Wang, Tian Hang, Beijing Chen, and Guoying Zhao. A dual-branch neural network for deepfake video detection by detecting spatial and temporal inconsistencies. Multimedia Tools and Applications, 81(29):42591–42606, 2022.

[10] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11):2673–2681, 1997.

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.

[12] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1–11, 2019.

[13] Pu Sun Honggang Qi Yuezun Li, Xin Yang and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In IEEE Conference on Computer Vision and Patten Recognition (CVPR), 2020.

[14] Ahmed H. Khalifa, Nawal A. Zaher, Abdallah S. Abdallah, and Mohamed Waleed Fakhr. Convolutional neural network based on diverse gabor filters for deepfake recognition. IEEE Access, 10:22678–22686, 2022. doi: 10.1109/ACCESS.2022. 3152029.

[15] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126, 2021.

[16] Daniel Xie, Prosenjit Chatterjee, Zhipeng Liu, Kaushik Roy, and Edoh Kossi. Deep- fake detection on publicly available datasets using modified alexnet. In 2020 IEEE symposium series on computational intelligence (SSCI), pages 1866–1871. IEEE, 2020.

[17] Hasam Khalid and Simon S Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 656–657, 2020.

[18] Ilke Demir and Umur Aybars Ciftci. Where do deep fakes look? synthetic face detection via gaze tracking. In ACM Symposium on Eye Tracking Research and Applications, pages 1–11, 2021.

[19] Hafsa Ilyas, Aun Irtaza, Ali Javed, and Khalid Mahmood Malik. Deepfakes examiner: An end-to-end deep learning model for deepfakes videos detection. In 2022 16th International Conference on Open Source Systems and Technologies (ICOSST), pages 1–6, 2022. doi: 10.1109/ICOSST57195.2022.10016871

[20] Hafsa Ilyas, Ali Javed, Muteb Mohammad Aljasem, and Mustafa Alhababi. Fused swish-relu efficient-net model for deepfakes detection. In 2023 9th International Conference on Automation, Robotics and Applications (ICARA), pages 368–372. IEEE, 2023.