A Comparative Study of EfficientNetB4 and VGG19 Models for Deepfake Detection

Sandesh Shrestha
Computer Information Science
Minnesota State University, Mankato
Mankato, MN, United States
sandesh.shrestha@mnsu.edu

Saumya Gautam
Computer Information Science
Minnesota State University, Mankato
Mankato, MN, United States
saumya.gautam@mnsu.edu

Eun Soo Park
Computer Information Science
Minnesota State University, Mankato
Mankato, MN, United States
eunsoo.park@mnsu.edu

Naseef Mansoor
Computer Information Science
Minnesota State University, Mankato
Mankato, MN, United States
naseef.mansoor@mnsu.edu

***Abstract*—With the increasing production of hyper-realistic altered images, the need for effective deepfake detection technology has become critical. These altered images pose significant threats to security, privacy, and the spread of misinformation, complicating the distinction between authentic and manipulated content. This challenge has far-reaching implications, from social media and politics to personal relationships. This study focuses on detecting deepfake human face images using a balanced dataset of 140,000 images, comprising 70,000 real faces sourced from Nvidia's Flickr dataset and 70,000 fake faces generated by StyleGAN. In this research, we compare the performance of EfficientNetB4 and VGG19 models, to identify subtle manipulations in high-quality deepfake images. Our findings demonstrate that the EfficientNetB4 model achieves an accuracy of 98.54%, while the VGG19 model reaches 99.11% in the base model, highlighting the effectiveness of these models in advancing deepfake detection technology.**

Keywords — Deepfake Detection, GANs, Convolutional Neural Network, EfficientNetB4, VGG19

I. INTRODUCTION

With the growing advancement in artificial intelligence (AI) and deep learning technology, new challenges have emerged, such as the creation of hyper-realistic manipulation of images known as deepfakes. The term "deepfake" is a combination of the words "deep learning" and "fake" [1]. These synthetic images closely mimic real individuals' appearances and expressions making it difficult for the human eye to distinguish between authentic and fake images. Deepfakes have been used in several cases such as creating misleading images of public figures or impersonating individuals for fraud, which in turn, present serious risks to their privacy and security.

Deepfake technology relies heavily on Generative

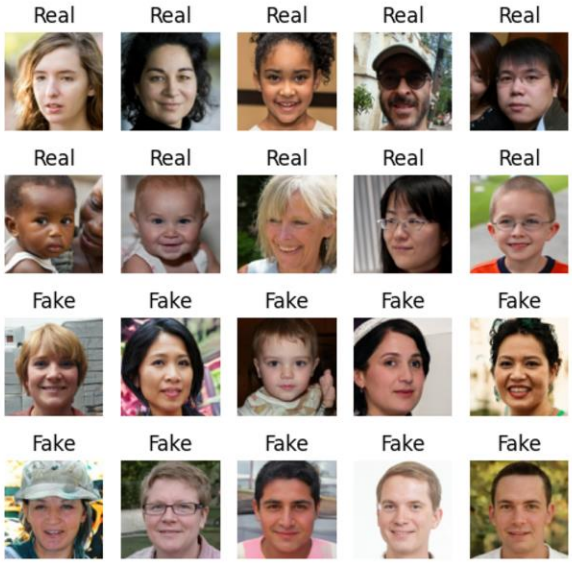


Figure 1. Sample images from the Flickr dataset demonstrating real and fake images generated using Style-GANs [13].

Adversarial Networks (GANs). In this process, two neural networks, the generator, and the discriminator, engage in a competitive learning process to produce highly realistic media [2]. The generator produces images, while the discriminator attempts to differentiate these from real images. Several recent cases show the real-world impact of deepfakes and reinforce the need for advanced detection tools. For example, in the realm of pornography, deepfake technology has been widely exploited for non-consensual purposes including the creation of revenge porn. In South Korea, a major controversy emerged with the "Nth Room" scandal [4] in 2020. This case involved a network of chat rooms on Telegram where users shared sexually exploitative content, including videos of women and minors that were altered using deepfake technology. The perpetrators used these manipulated contents to blackmail victims and profit from non-consensual pornography from 260,000 users. The incident exposed the dangers of deepfake technology when used maliciously, leading to public outrage against deepfake

pornography and revenge porn. It highlighted the urgent need for technological and legal measures to combat deepfake abuse, especially in cases where victims suffer from identity misuse and privacy violations. Similarly, a striking example occurred in 2023 [5], when a report revealed that AI-generated deepfake images were being utilized to create sexually explicit content involving children. These images are often indistinguishable from real photographs, making them especially dangerous as they can easily circulate online, complicating efforts to identify and apprehend perpetrators. The exploitation of deepfake technology not only deepens the trauma for victims but also underscores the urgent need for advanced detection tools. These examples reinforce the urgent need for advanced detection tools that can accurately identify manipulated images, ensuring the integrity of information and protecting individuals from misuse.

Among the different types of GANs, StyleGAN is highly renowned for its ability to produce hyper-realistic human faces. Its unique architecture allows for precise control over style features such as lighting, facial details, and background, producing images that are challenging to differentiate from genuine photographs [3]. In this study, we use a dataset where the fake images are generated specifically by StyleGAN. This model demonstrates the sophisticated manipulations possible with GANs, creating deepfakes that generally avoid the typical visual artifacts often found in images produced by simpler GAN models. This comparative analysis allows us to assess the effectiveness of each model architecture in detecting StyleGAN-based manipulations in high-quality deepfakes.

The significance of deepfake detection lies in its potential to protect information integrity, maintain public trust, and uphold privacy rights in digital spaces. As tools like StyleGAN become widely accessible, anyone with minimal technical knowledge can create realistic fake images, increasing the potential for misuse across various domains. Several existing models have been proposed for deepfake detection, including CNN architectures like Xception, Inception, and ResNet, each demonstrating varying degrees of effectiveness. However, there remains a critical need for comparative studies to evaluate these models comprehensively. Effective detection methods are important not only to prevent individual and societal harm but also to support ethical standards in media and communications. In this study, we focus on comparing EfficientNetB4 and VGG19 to evaluate their performance in detecting complex deepfake images. EfficientNetB4 is known for its computational efficiency and accuracy, while VGG19's depth allows it to capture intricate features within images, making these models promising for identifying subtle manipulations characteristic of StyleGAN-generated faces. Through our research, we contribute to the field by exploring the capabilities of EfficientNetB4 and VGG19, comparing the base and the fine-tuned models, and ultimately contributing to the enhancement of deepfake detection methodologies.

II. Related Work

Recent studies in deepfake detection have explored a variety of methodologies, with a significant focus on convolutional neural networks (CNNs) and their architecture. While prior studies like [6] authors highlighted VGG19's limited generalization ability, its widespread use and effectiveness in image classification tasks make it a valuable baseline for comparison. Including VGG19 allows for a deeper exploration of its specific strengths and weaknesses relative to EfficientNetB4, especially in the unique context of human face based deepfake detection. In parallel, authors in [7] compared the performance of VGG16, VGG19, and ResNet50, revealing that ResNet50 outperformed the others in terms of precision and recall. However, the authors also mentioned that its increased complexity resulted in longer training times, which could hinder its deployment in time-sensitive applications. Further contributing to the discussion, authors in [8] emphasized the efficacy of deep learning techniques, asserting that more sophisticated models tend to yield better performance in deepfake detection. Yet, these advanced models often come with substantial computational resource requirements, making them less accessible for broader use. Authors in [9] explored various CNN architectures, illustrating improvements in detection rates but also highlighting vulnerabilities to adversarial attacks, which remain a significant concern in maintaining robustness against evolving deepfake strategies. Additionally, authors in [10] presented a framework that leverages transfer learning in CNNs, further underscoring the role of advanced techniques in exposing deepfakes. Their work suggests that while transfer learning can enhance detection capabilities, it also raises questions about the trade-off between model complexity and operational efficiency. In [11] authors highlighted the challenges in creating models that balance robustness and efficiency, an ongoing issue as deepfake technologies evolve. This perspective emphasizes the importance of evaluating both EfficientNetB4 and VGG19, as these models represent contrasting approaches to addressing these challenges. EfficientNetB4, with its emphasis on resource efficiency, offers a promising pathway toward meeting these demands.

Our study focuses on EfficientNetB4 and VGG19 due to their distinct characteristics and historical significance in the field of image classification. VGG19, introduced in [12], is renowned for its simplicity and depth, making it a staple in many image-processing tasks. Its layered architecture has proven effective in capturing intricate features, which is crucial for deepfake detection in human face images. Conversely, EfficientNetB4, introduced in [11], represents a paradigm shift in model efficiency and performance by optimizing depth, width, and resolution through a compound scaling method. This allows EfficientNetB4 to achieve high accuracy while requiring fewer parameters and computational resources compared to traditional models. The combination of these architectures offers a unique opportunity to explore their comparative effectiveness in the context of deepfake detection.

III. Methodology

In this section, we discuss the dataset used in this work along with the methodology for designing the models for deepfake detection.

A. Dataset

The dataset used in this research is taken from the Flickr dataset collected by Nvidia [13]. This dataset is specifically collected for the task of deepfake detection and consists of two primary classes:

- Real images: Authentic, real, and unaltered human face images.

- Fake images: StyleGAN-generated or manipulated deepfake images.

It consists of 140,000 images, systematically divided into training, validation, and test sets to facilitate comprehensive model evaluation. The training set consists of 50,000 images of real faces and 50,000 images of deepfakes. In the validation set, there are 10,000 real images and 10,000 deepfake images. Similarly, the test set comprises 10,000 real images and 10,000 deepfake images, which will be used to evaluate our final model. Each image has a resolution of 256 x 256. Some sample images from the dataset are shown in Figure 1.

B. Data Preprocessing

To enhance the performance of the deep learning models and improve their generalization capabilities, data preprocessing plays an important role. Since the input layer for the VGG19 and EfficientNet has a dimension of 224 × 224 pixels, the images from the dataset were resized to the acceptable input dimensions for the models. To preserve the color information and maintain compatibility with the model's architecture, each image was processed in RGB color mode.

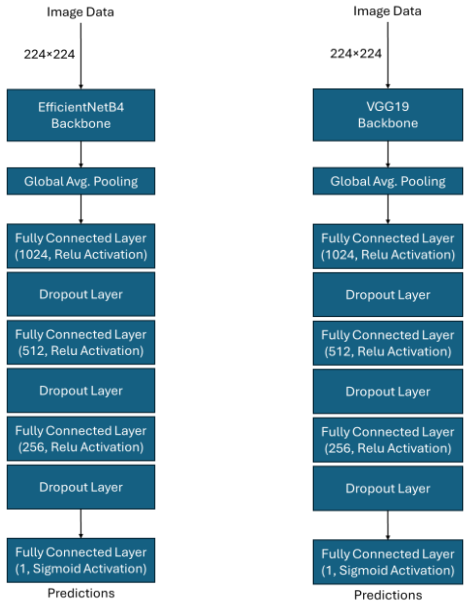
Data augmentation can improve model robustness by introducing variability in the dataset. Artificial augmentation like horizontal flip, vertical flip, crop, brightness, and contrast provide generalization. However, in our study, we decided not to apply artificial data augmentation because the dataset already exhibits significant variability. The images in the dataset feature a wide range of lighting conditions, facial orientations, and expressions, as well as natural variations in contrast. This inherent diversity reduces the necessity for additional augmentation to achieve generalization.

C. Model Training

Figure 2 shows the accuracy of different models on the ImageNet dataset. From the figure, we can see that EfficientNets are medium-sized models while their performance is comparable to large models (high number or parameters). In this paper, we conducted a comparative analysis of two CNN-based models: EfficientNetB4 and VGG19. These models were selected based on their complementary strengths. EfficientNetB4 was chosen for its better performance with medium parameters in image classification tasks. On the other hand, VGG19, with its deep architecture and simple design, provides a robust baseline for comparison, offering insights into how a classic architecture performs against a modern, highly optimized model.

a. EfficientNetB4 Model

EfficientNet models emerged as a response to the need for more efficient neural networks that could achieve state-of-the-art performance while using fewer parameters and computational resources [11]. The baseline model, EfficientNetB0 was scaled to create larger models like EfficeintNetB4. EfficientNetB4 consists of 19 million



a) b)

Figure 3. Architecture of a) EfficientNetB4 fine-tuned and b) VGG19 fine-tuned model used for deepfake detection.

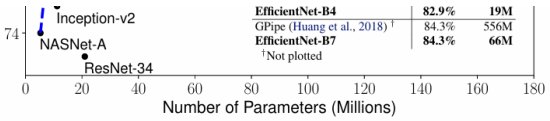


Figure 2. ImageNet Accuracy for different models with varying number of training parameters [11].

parameters which is fewer compared to other CNN models. The fine-tuned architecture begins with the pre-trained EfficeintNetB4 backbone for feature extraction, followed by a global average pooling layer to condense spatial features. It includes fully connected layers with 1024, 512, and 256 neurons, each using ReLU activation for non-linearity, and dropout layers to reduce overfitting. The final layer is a fully connected layer with a single neuron and sigmoid activation for binary classification, designed to classify the deepfake images.

b. VGG19 Model

VGG19 is a deep convolutional neural network model developed by the Visual Geometry Group (VGG) at the University of Oxford. VGG19 consists of 19 layers: 16 convolutional layers and 3 fully connected layers. It is mostly used for the image classification [14]. VGG19 has consistent performance across various image classification benchmarks, including the ImageNet dataset, achieving high accuracy and low error rates [12]. In the fine-tuned architecture, the last 20 layers of the pre-trained VGG19 backbone are frozen to retain learned spatial features, followed by fully connected layers with 1024, 512, and 256 neurons and dropout layers to

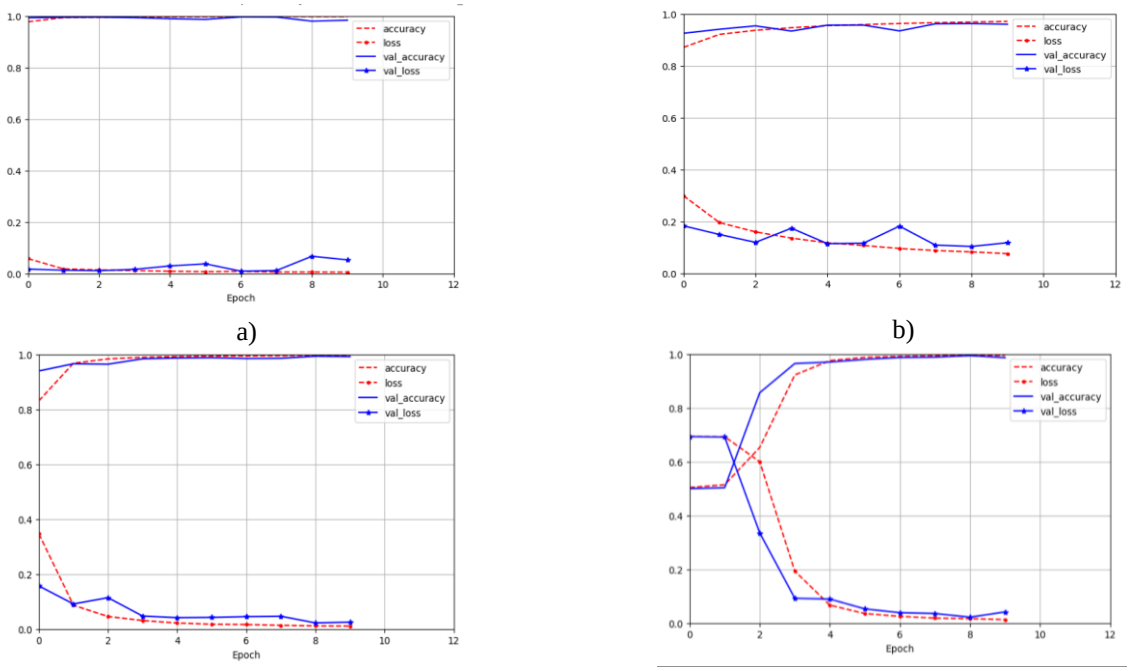


Figure 4. Training and Validation accuracy and Loss for all the four models
a) EfficientNet Base, b) EfficientNet Fine-tuned, c) VGG 19 Base, and d) VGG 19 Fine-tuned.

reduce overfitting. Then a fully connected layer with a single neuron and sigmoid activation is added to classify deepfake and real images.

IV. Experimental Results

We trained our models for 25 epochs using the training data set, with validation accuracy monitored at every epoch to assess performance and prevent overfitting. We have used a batch size of 32 while training the model to ensure computational efficiency with hardware constraints. We have used the Adam optimizer with a learning rate of 0.001 for stable and efficient model training. All models used a sigmoid activation function in the final layer for binary classification. We used the binary cross entropy loss function. Inference was performed on the test dataset to evaluate the models' fitness.

Initially, we train the EfficientNetB4 and VGG19 models on the training dataset. These are what we call the base models: EfficientNetB4 Base, and VGG19 Base. The base models are pre-trained on the ImageNet dataset, meaning they have a backbone network initialized with weights learned from the ImageNet dataset, followed by a single neuron classification layer with sigmoid activation. After training the base models, we fine-tune them by adjusting different hyperparameters. The hyperparameters for all models were carefully selected to ensure consistent and fair comparative analysis.

We experimented with different hyperparameters for the models to determine the best set. We experimented with different numbers of layers after the backbone network and selected three dense layers as they provided the best performance. The three dense layers have neurons of 1024, 512, and 256, respectively. Each dense layer used ReLU activation to introduce non-linearity, while the classification layer applied sigmoid activation for binary classification. We also experimented with different dropout rates and found that 0.2, 0.3, and 0.5 provided the best results.

A. Comparative Performance Analysis

In this section, we compare the performance of all four models: EfficientNetB4 Base, EfficientNetB4 fine-tuned, VGG19 Base, and VGG19 fine-tuned. Table 1 lists the training, validation, and test accuracy of these models. The results show that the EfficientNetB4 Base model outperformed the EfficientNetB4 Fine-tuned model in validation and test accuracies. The EfficientNetB4 Base model achieved a test accuracy of 98.54%, compared to 95.54% test accuracy for the EfficientNetB4 Fine-tuned model. This is attributed to the binary cross-entropy loss for the EfficientNetB4 Base models compared to the EfficientNetB4 Fine-tuned model as shown in Figure 4 a) and b).

A similar pattern can be also observed for VGG19 Base and VGG19 Fine-tuned models. However, unlike the EfficientNetB4 models, the difference in accuracy between the VGG19 Base and VGG19 Fine-tuned models is minimal. The VGG19 Base model achieved a higher 0.11% accuracy compared to the VGG19 Fine-tuned model. Furthermore, as can be seen from Figure 4, the loss for VGG19 is much lower than that of EfficientNetB4 for both the Base and Fine-tuned models.

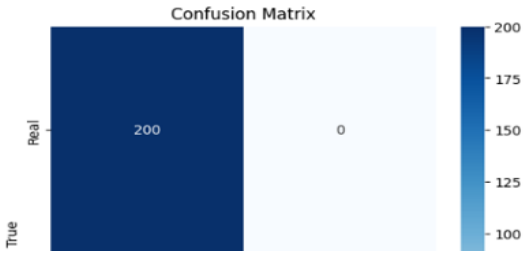


TABLE 1. EVALUATION METRICS FOR COMPARATIVE STUDY

	Accuracy		
Model	**Train**	**Validation**	**Test**
EfficientNetB4 Fine-tuned	97.83%	96.30%	95.54%
EfficientNetB4 Base	99.80%	98.30%	98.54%
VGG19 Fine-tuned	99.1%	99.25%	99.00%
VGG19 Base	99.64%	99.20%	99.11%

Although the fine-tuned models are expected to have better accuracy than base models, in our experiment the base models performed better. This can be due to the introduction of additional training parameters in the fine-tuned models, which could have introduced unnecessary complexity, reducing the ability of the model to generalize.

B. Experimentation with Data Augmentation

One of the primary challenges we faced during this study was the computational cost associated with training deep learning models on large datasets. Training the models on datasets with 100,000 images per epoch, incorporating data augmentation techniques such as rotations, width shifts, height shifts, zooms, and horizontal flips as shown in Table 2, resulted in significantly longer training times. The use of data augmentation is expected to improve model robustness and generalization, allowing the models to better handle variations in real-world data and avoid overfitting to the training set.

Despite utilizing GPU (P100), the training time for each epoch with data augmentation as shown in Table 2 was approximately 24 hours. This highlights the immense computational resources and time needed to train these models effectively. To mitigate this issue, we experimented with smaller subsets of the dataset to evaluate the performance of the models more efficiently. This approach allowed us to assess model performance while reducing the overall training time and computational burden. In our reduced dataset experiment, we experimented with a dataset of 2000 images for training and 400 images each for validation and testing. We applied the same data augmentation techniques as shown in Table 2 on the training and validation dataset. However, the models trained in a reduced dataset could not generalize effectively, leading to poor performance on fake image detection and bias toward real images.

The confusion matrix for the EfficientNetB4 fine-tuned model and the VGG19 fine-tuned model is shown in Figure 5. From the confusion matrix, it can be seen that the accuracy is only 50%, and the precision for the "Real" image is also 50%, showing a high rate of false positives. However, the recall is 100% because the model correctly identifies all "Real" images without missing any. The confusion matrix shows strong biases toward the real images, with no fake images correctly identified.

V. CONCLUSION

In this study, we conducted a comparative analysis of EfficientNetB4 and VGG19 models to evaluate their performance in detecting deepfake human face images generated by StyleGAN. Both models achieved high accuracy across training, validation, and test datasets. VGG19 consistently outperformed EfficientNetB4, achieving a slightly higher test accuracy of 99.11% compared to the base model of EfficientNetB4 at 98.54%, indicating its ability to capture intricate image features. On the other hand, as EfficientNetB4 is more computationally efficient than VGG19, it is suitable for resource-constrained applications. Interestingly, the fine-tuned versions of both models provided no improvements. By leveraging the strengths of these CNN architectures, our study contributes to the growing body of knowledge in deepfake detection and supports the development of reliable and efficient tools to mitigate the risks posed by deepfake technology.

TABLE 2. DATA AUGMENTATION PARAMETERS

Parameter	Value
Rotation	30
Width shift	0.2
Height shift	0.2
Zoom	0.2
Horizontal flip	True

REFERENCES

[1] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes," ACM Computing Surveys, vol. 54, no. 1. Association for Computing Machinery, Apr. 01, 2021. doi: 10.1145/3425780

[2] I. J. Goodfellow et al., "Generative Adversarial Nets," Advances in neural information processing systems 27, 2014.

[3] T. Karras NVIDIA and S. Laine NVIDIA, "A Style-Based Generator Architecture for Generative Adversarial Networks," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.

[4] R. Kim, "Everything You Need to Know About the Nth Room Case in 'Cyber Hell'", netflix.com/tudum/articles. Accessed: Nov. 1. 2024. [Online] Available: https://www.netflix.com/tudum/articles/everything-to-know-about-the-nth-room-case-in-cyber-hell

[5] K. Chan, "British man sentenced to 18 years for using AI to make child sexual abuse imagery", apnews.com. Accessed: Nov. 1. 2024. [Online] Available: https://apnews.com/article/ai-deepfakes-child-sexual-abuse-6b73afb86e95068a2a444f299fd7ca30

[6] O. Jannu et al, "Comparative Analysis of Deepfake Detection Models," in 2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024, Institute of Electrical and Electronics Engineers Inc., 2024, doi: 10.1109/I2CT61223.2024.10543823.

[7] Z. N. Ashani, et al, "Comparative Analysis of Deepfake Image Detection Method Using VGG16, VGG19 and ResNet50," Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 47, no. 1, May 2025, pp. 16–28, doi: 10.37934/araset.47.1.1628.

[8] M. Taeb and H. Chi, "Comparison of Deepfake Detection Techniques through Deep Learning," Journal of Cybersecurity and Privacy, vol. 2, no. 1, Mar. 2022, pp. 89–106, doi: 10.3390/jcp2010007.

[9] Raveena et al, "Exploring Deepfake Detection: A Comparative Study of CNN Models," in 2024 International Conference on Intelligent Systems for Cybersecurity, ISCS 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ISCS61804.2024.10581012.

[10] S. Suratkar et al, “Exposing DeepFakes Using Convolutional Neural Networks and Transfer Learning Approaches,” in 2020 IEEE 17th India Council International Conference, INDICON 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/INDICON49873.2020.9342252.

[11] M. Tan and Q. v Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” arXiv preprint arXiv, 2019. pp. 6105-6114

[12] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in ICLR, Sep. 2014.

[13] Xhlulu. "140k Real and Fake Faces", Kaggle, 2021, https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces/data.

[14] S. Mascarenhas and M. Agarwal, “A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification,” in Proceedings of IEEE International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications, CENTCON 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 96–99. doi: 10.1109/CENTCON52345.2021.9687944.