Comparative Analysis of Deep Learning Models' Resilience to Adversarial Attacks in Face Recognition

Samuel Salama
Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, USA
ssalama@iu.edu

Chris Haleas
Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, USA
chaleas@iu.edu

***Abstract*—This paper evaluates the resilience of pre-trained convolutional neural networks to adversarial attacks in face recognition tasks. We systematically test against three common adversarial perturbations: Gaussian noise, color shifts, and black patch occlusions. Our experiments measure the classification between real and fake faces, the model contained four classes of facial images: real faces, easy-fake, mid-fake, and hard-fake. The results demonstrate varying levels of vulnerability in modern CNN architectures when faced with these targeted attacks, providing insights for developing more robust facial recognition systems.**

***Index Terms*—adversarial attacks, face recognition, convolutional neural networks, computer vision, model robustness, deepfake detection**

I. Introduction

Facial recognition technology has come about to be one of the most widely deployed applications of computer vision and deep learning. This technology helps transform sectors ranging from security and law enforcement to consumer electronics and social media. The accuracy of modern face recognition systems stems from advancements in Convolutional Neural Networks (CNNs), which have demonstrated an impressive ability to extract and analyze facial features. However, with this large adoption comes questions about the reliability of these systems, specifically in the presence of something called black-box adversarial attacks.

Black-box adversarial attacks are essentially modifications to input images that cause neural networks to make incorrect predictions while maintaining visual similarity to the original images for human observers. These attacks pose significant challenges for security applications that rely on face recognition, as they can allow unauthorized access or impersonations. The vulnerability of CNNs to such attacks stems from their complex non-linear decision boundaries.

We took on a task to see how the resilience of popular CNN architectures perform against three common types of adversarial attacks: Gaussian noise additions, color shifts, and black patch obstructions. We examine these attacks across four categories of facial images: real faces, easy-fake, mid-fake, and hard-fake. By testing several prominent CNN architectures such as InceptionV1 [1], ResNet18 [2], DenseNet121 [3], EfficientNetB0 [4], SqueezeNet1.1 [5], and RegNetX-400MF [6] we aim to identify which models have greater robustness to specific attack types and which present the greatest vulnerability.

Face recognition systems continue to be deployed in increasingly critical applications and understanding their vulnerability to adversarial attacks is essential for developing more robust models. Our findings provide insights into specific architectural characteristics that contribute to adversarial resilience. In this paper, we present our methodology for testing CNN resilience, analyze the comparative performance of different architectures under adversarial conditions, and discuss the implications of our findings for the future development of secure face recognition systems.

II. Background and Related Work

A. Deep Learning for Face Recognition

Convolutional Neural Networks have changed the way we go about facial recognition by allowing systems to learn and extract discriminative facial features automatically without the need for extensive manual feature engineering. Face recognition is a major application of computer vision with many real-world applications across security, authentication, and human-computer interaction. The main advantage of CNNs in face recognition comes from their hierarchical feature extraction abilities, where early layers capture basic features like edges and textures, while deeper layers identify more complex facial structures. Most modern face recognition systems typically employ a pipeline that includes face detection, alignment, feature extraction, and classification or verification. However, further improvements have demonstrated that deeper and more advanced architectures can capture increasingly delicate facial characteristics, leading to better discrimination between individuals. There are many popular CNN architectures that have came about with different strengths and design philosophies for visual recognition tasks. Our study examines persistence across multiple prominent architectures such as InceptionV1, ResNet18, DenseNet121, EfficientNetB0, SqueezeNet1.1, and

RegNetX-400MF against adversarial attacks. These architectures give us a diverse understanding of the fundamental tradeoffs in network design.

B. Adversarial Attacks

Adversarial attacks represent a significant security vulnerability for CNN-based facial recognition systems. These attacks essentially involve modifying input images to cause misclassifications while maintaining visual similarity to the original images. The vulnerability of CNNs to such attacks stems from their highly non-linear decision boundaries in high-dimensional space, which can be taken advantage of. Our research focuses on three common types of adversarial attacks:

- Gaussian Noise Attack: This attack adds random noise sampled from a Gaussian distribution to the input image. Even though it seems simple, Gaussian noise can strongly disrupt the feature extraction process of CNNs, especially when the signal-to-noise ratio is calibrated accordingly.
- Color Shift Attack: This attack exploits the sensitivity of CNNs to color information by altering the color values across an image. This can be through manipulation of hue, saturation, or brightness. Color shifts can be particularly effective against face recognition systems that rely heavily on skin tone and texture features.
- Black Patch Occlusion: This attack places black regions over specific parts of the face, effectively removing information critical for recognition. Strategic placement of black patches can cause the network to miss key discriminative features while maintaining a naturally explainable appearance.

Our approach is complementary to recent work by Sen et al., who demonstrated that even state-of-the-art CNN architectures (ResNet-101, AlexNet, and RegNetY) are highly vulnerable to adversarial attacks [7]. Their study showed how classification accuracy can drop drastically with minimal perturbations, similar to what we have observed in our facial recognition experiments. While they focused on general image classification and proposed defensive distillation as a countermeasure, our research specifically examines how these vulnerabilities make their way into facial recognition systems and which architectural characteristics contribute to adversarial robustness in this domain.

Understanding the resilience of different CNN architectures to these attacks is essential for developing more robust face recognition systems, particularly for applications where security is a major point. By evaluating multiple architectures against these common attack types, we aim to identify architectural characteristics that contribute to adversarial robustness.

III. Methods

A. Data Collection and Preprocessing

The data was obtained from the “Real and Fake Face Detection” Kaggle dataset by the ”CIPLAB” of Yonsei University [8]. It was organized into two directories: one containing real facial images, and the other containing altered facial images. The alterations included replacing facial features such as the left eye, right eye, nose, or mouth. The fake images were further divided into three subjective difficulty levels (easy, medium, and hard), as labeled by the dataset creators in the file names. This labeling allowed us to evaluate model performance across varying levels of manipulation difficulty. We loaded the image file paths and assigned corresponding labels (0 for real and 1 for fake), while also recording each image’s difficulty level. The dataset was split into training (75 percent) and validation (25 percent) sets using stratified sampling with a fixed random state of 42. Standard image transformations specific to each model architecture were applied to prepare the images for input into the neural networks.

B. Model Architecture and Training

We evaluated multiple pre-trained convolutional neural network architectures using transfer learning to identify the most effective model for distinguishing real and fake images. The architectures tested included SqueezeNet1.1, ResNet18, RegNetX-400MF, InceptionV1, EfficientNetB0, and DenseNet121, all of which were pre-trained on the ImageNet dataset. For each model, we replaced the final classification layer with a custom head that outputs two classes: real and fake. We then fine-tuned the network using our dataset. This approach allowed us to leverage the feature extraction capabilities of the pre-trained models while adapting them to our specific classification task. All models were trained for 10 epochs using the Adam optimizer with a learning rate of 0.0001, the cross-entropy loss function, and a batch size of 32. The entire process was implemented using the PyTorch deep learning framework.

C. Adversarial Attack Testing

We assessed model robustness by testing their resistance against three adversarial techniques. Note that our methods involved black-box adversarial attacks, meaning we applied perturbations to the input images rather than altering the models themselves. The way in which we implemented these attacks went as follows:

- **Gaussian Noise Attack**: Adding random noise from a Gaussian distribution with a standard deviation of 0.3 to the validation images.
- **Color Shift Attack**: Applying a color shift attack by adding a uniform value of 0.5 to the RGB channels of the image.
- **Black Patch Attack**: Placing a randomly positioned 190x190 pixel black square on each image.

These attacks were applied to the validation set using a custom dataset class, allowing us to compare model performance under adversarial conditions conditions. We compared performance by calculating the accuracy of each model and visualized the results by plotting the accuracies on a radar chart that you can view in Figure 3.

Examples of an image from our chosen dataset with each of the adversarial attacks can be seen in Figure 1.

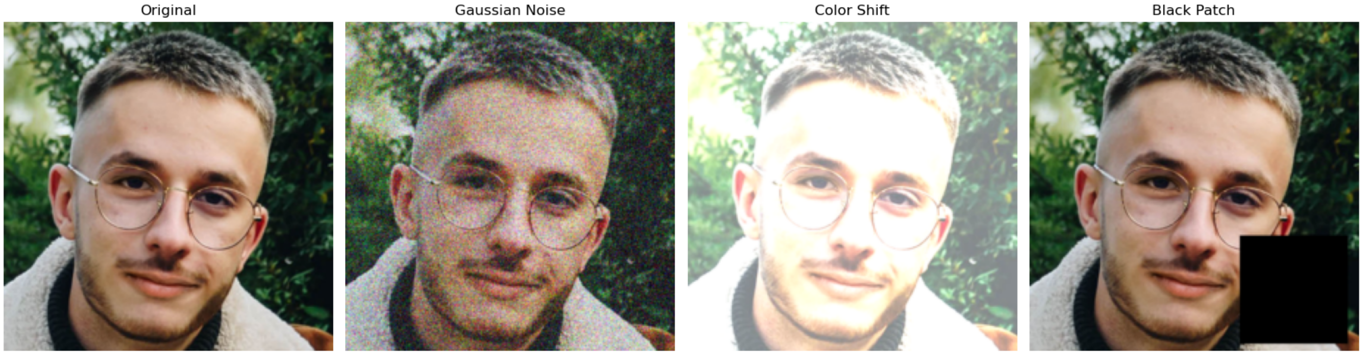


Fig. 1. Original image from our chosen dataset [8] alongside three adversarial attacks on the image

IV. Results

A. Performance on Clean Data

Our evaluation of the six deep learning models showed notable differences in their baseline performance. As shown in Table 1, EfficientNetB0 had the highest accuracy (78.1%) on clean data, outperforming all other models. DenseNet121 (71.0%), RegNetX-400MF (69.7%), and ResNet18 (68.7%) demonstrated moderate performance, while InceptionV1 (68.3%) and SqueezeNet1.1 (65.8%) showed the lowest accuracy rates.

B. Performance Across Difficulty Levels

The models had differing capabilities in detecting fake faces across different difficulty levels (Table 1). However, it is important to note that the difficulty classifications were subjectively labeled by the dataset creators. For 'easy' fake faces, EfficientNetB0 had the highest detection rate (72.5%), while SqueezeNet1.1 struggled the most (31.4%). For 'mid' difficulty fakes, DenseNet121 showed the strongest performance (80.5%), while SqueezeNet1.1 (57.8%) showed the weakest results. When confronted with 'hard' fake images, DenseNet121 (85.2%) had the highest accuracy, whereas SqueezeNet1.1 performed the poorest (52.5%).

C. Real vs. Fake Classification Metrics

Table 1 shows the precision, recall, and F1-scores for both real and fake classifications. We use the F1-score to evaluate the balance of results, as it considers both precision and recall. The model with the highest F1-score for the real class was EfficientNetB0, with a score of 80%, and it also had the highest F1-score for the fake class at 76%, showing that it is the best-performing model overall with a well-balanced performance. DenseNet121 was also quite balanced, with an F1-score of 71% for both real and fake classes. InceptionV1 had similarly balanced results, with an F1-score of 68% for real and 69% for fake. The remaining models were more skewed toward real classifications: RegNetX-400MF had an F1-score of 74% for real and only 64% for fake, ResNet18 had 72% for real and 65% for fake, and SqueezeNet1.1 was the most imbalanced, with an F1-score of 71% for real and just 58% for fake.

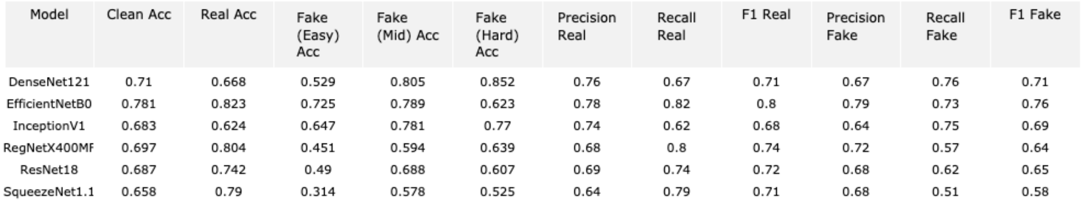
The confusion matrices (Figure 2) offer visualizations of the classification patterns. Notable results include EfficientNetB0, which correctly identified 223 out of 270 real faces (true positives) and 176 out of 240 fake faces (true negatives), showing a relatively balanced performance with low error rates across both classes. In contrast, SqueezeNet1.1 had a strong bias toward classifying images as real, correctly identifying only 122 fake faces (true negatives), yet achieving a high number of correct classifications for real faces (214 true positives).

D. Robustness Against Adversarial Attacks

In Table 2, we compare the clean accuracy of six different models with their accuracy after being subjected to three types of adversarial attacks: Gaussian Noise, Color Shift, and Black Patch. The model most resilient to these attacks was DenseNet121, with the smallest average percentage drop in accuracy at 7.8%, followed by RegNetX-400MF with 8.3%, ResNet18 with 9.67%, SqueezeNet1.1 with 10.23%, and InceptionV1 with 10.67%. Interestingly, although EfficientNetB0 achieved the highest clean accuracy (78.1%), it was the most vulnerable overall, experiencing the largest average accuracy drop of 14.03% across the three attacks.

1) Gaussian Noise Attack: Gaussian Noise was the most effective adversarial strategy, resulting in an average accuracy drop of 15.17% across all models. EfficientNetB0 was particularly affected, suffering a 25.1% decrease in accuracy, which was the largest drop recorded in the entire evaluation. The attack also significantly impacted other models, causing a 15.3% drop in InceptionV1, 14.1% in RegNetX-400MF, 13.8% in SqueezeNet1.1, 13.5% in ResNet18, and 9.2% in DenseNet121. This makes DenseNet121 the most effective against the Gaussian Noise Attack.

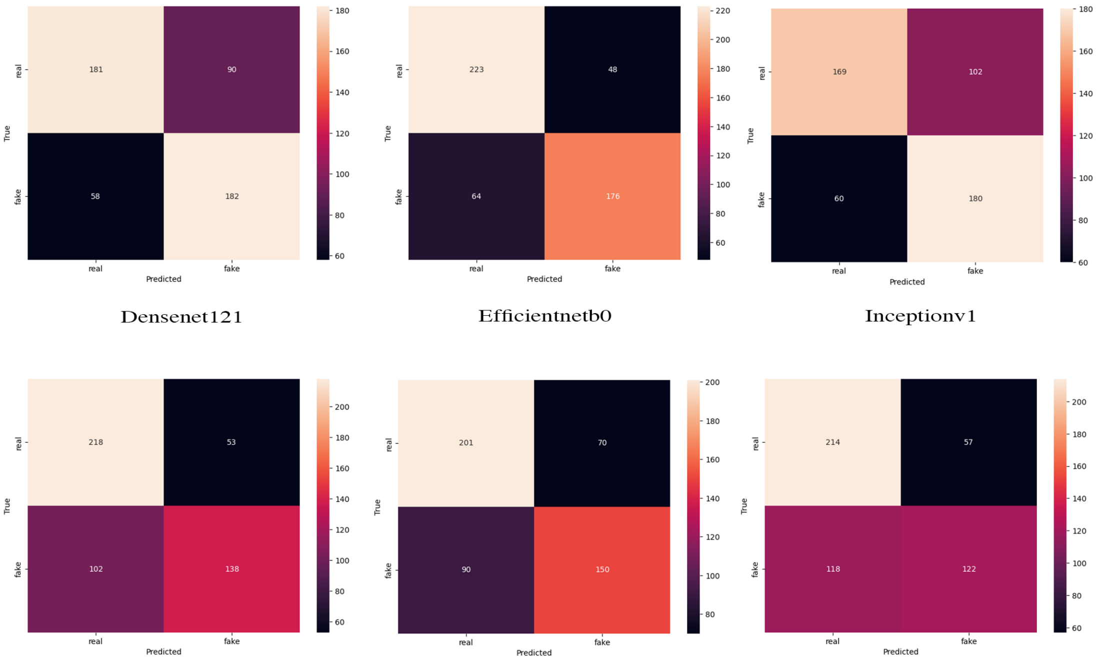
2) Color Shift Attack: The average accuracy drop from the Color Shift attack was 10.27%. Again, EfficientNetB0 was heavily impacted, with a 14.1% drop. Other models experienced more moderate declines, such as 11.6% for SqueezeNet1.1, 11.2% for ResNet18, 9.8% for InceptionV1, 8.4% for DenseNet121, and 6.5% for RegNetX-400MF. This makes RegNetX-400MF the most effective against the Color Shift attack.



Model	Clean Acc	Real Acc	Fake (Easy) Acc	Fake (Mid) Acc	Fake (Hard) Acc	Precision Real	Recall Real	F1 Real	Precision Fake	Recall Fake	F1 Fake
DenseNet121	0.71	0.668	0.529	0.805	0.852	0.76	0.67	0.71	0.67	0.76	0.71
EfficientNetB0	0.781	0.823	0.725	0.789	0.623	0.78	0.82	0.8	0.79	0.73	0.76
InceptionV1	0.683	0.624	0.647	0.781	0.77	0.74	0.62	0.68	0.64	0.75	0.69
RegNetX400MF	0.697	0.804	0.451	0.594	0.639	0.68	0.8	0.74	0.72	0.57	0.64
ResNet18	0.687	0.742	0.49	0.688	0.607	0.69	0.74	0.72	0.68	0.62	0.65
SqueezeNet1.1	0.658	0.79	0.314	0.578	0.525	0.64	0.79	0.71	0.68	0.51	0.58

TABLE I

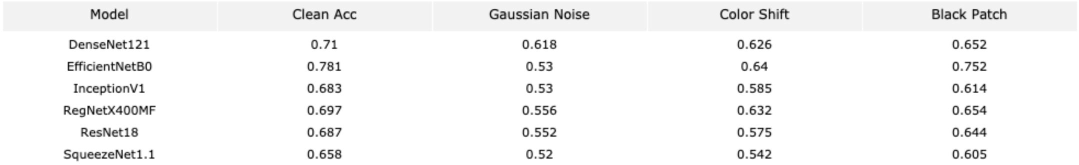
Performance of six models across various metrics, including clean accuracy and real and fake accuracy at different difficulty levels (easy, medium, and hard), along with precision, recall, and F1 scores for both real and fake classifications.



Densenet121 Efficientnetb0 Inceptionv1

Regnetx400mf Resnet18 Squeezenet1.1

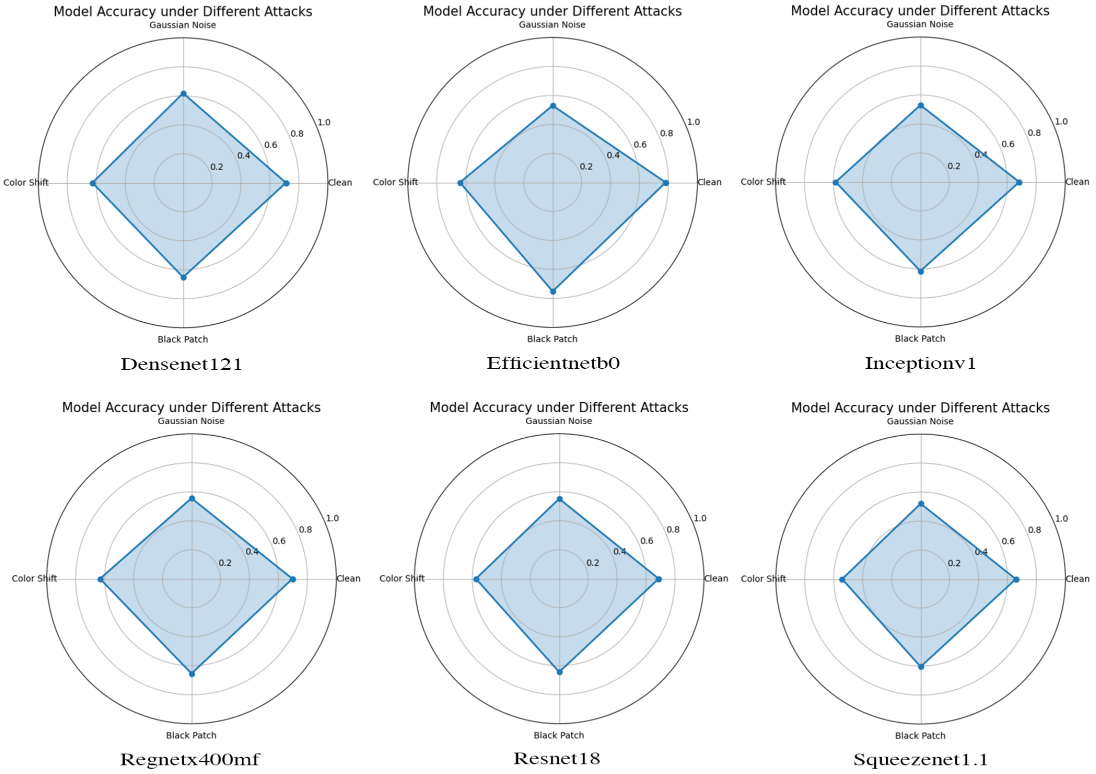
Fig. 2. Confusion matrices showing classification performance of the models on clean data.



Model	Clean Acc	Gaussian Noise	Color Shift	Black Patch
DenseNet121	0.71	0.618	0.626	0.652
EfficientNetB0	0.781	0.53	0.64	0.752
InceptionV1	0.683	0.53	0.585	0.614
RegNetX400MF	0.697	0.556	0.632	0.654
ResNet18	0.687	0.552	0.575	0.644
SqueezeNet1.1	0.658	0.52	0.542	0.605

TABLE II

Performance of six models on a clean validation set and three validation sets with adversarial attacks



Densenet121

Efficientnetb0

Inceptionv1

Regnetx400mf

Resnet18

Squeezenet1.1

Fig. 3. Radar plots of six models on a clean validation set compared to three validation sets with adversarial attacks

3) Black Patch Attack: This was the least effective of the three attacks, causing an average drop of 4.92% in accuracy. The largest drop was 6.9% for InceptionV1, followed by 5.8% for DenseNet121, 5.3% for SqueezeNet1.1, and 4.3% for RegNetX-400MF and ResNet18. EfficientNetB0 demonstrated resilience in this case, with only a 2.9% drop.

4) Summary of Adversarial Attack Results: These results highlight how different convolutional neural network architectures differ in their robustness to adversarial perturbations on validation images. Models like DenseNet121 and RegNetX-400MF show relatively strong defenses, while models with higher clean accuracies, like EfficientNetB0, may still be highly susceptible to attacks. Figure 3 highlights the robustness of the models under different adversarial attacks using a radar plot for visualization.

V. Discussion

A. Interpretation of Results

1) EfficientNetB0 Result Discussion: Despite having the highest clean accuracy, EfficientNetB0 had the largest drops in accuracy under Gaussian Noise and Color Shift attacks, especially with Gaussian Noise, where it experienced a 25.1% decrease. EfficientNetB0 performs exceptionally well on clean data because of its use of compound scaling, which balances depth, width, and resolution to scale the model in a way that optimizes efficiency and performance. This architectural design allows the model to capture specific features and achieve strong classification abilities for data similar to what it is trained on. In addition to its strong performance on clean data, EfficientNet's use of compound scaling also allowed the model to perform well against the Black Patch attack (2.9% accuracy drop). The model's architecture, optimized for specific features, enables it to maintain performance when random obstructions do not cover the critical features necessary for classification. However, this same compound scaling design likely causes the model to overfit to clean data distributions, making it less adaptable when faced with adversarial perturbations that affect the entire image, whether through pixel-level changes or global alterations, rather than just obstructing a specific section. Since the architecture is so tightly tuned to clean features, compound scaling reduces the number of alternative pathways in the network, which puts a limit on the model's ability to recover from disturbances in data. This makes the model less robust when exposed to perturbations like Gaussian Noise and Color Shifts (14.1% accuracy drop).

2) RegNetX-400MF Result Discussion: RegNetX-400MF experienced a high drop in accuracy when faced with Gaussian Noise (14.1% drop), but lesser drops in accuracy for Color Shift and Black Patch attacks. The reason it was more robust against Color Shift (6.5% accuracy drop) and Black Patch attacks (4.3% accuracy drop) may be due to the RegNet family's modular architecture, which processes different parts of the image independently. This flexibility allows the model to handle changes in color distribution, as seen in the Color Shift attack, more effectively because the overall structure and spatial relationships within the image remain intact. While Color Shift affects the global color distribution, it does not disrupt the features on a pixel-level that are key for recognition. Similarly, in Black Patch attacks, where parts of the image are occluded, the model can still rely on non-occluded parts of the image for classification, maintaining the overall integrity of the information. However, the Gaussian Noise attack adds pixel-level changes across the image. This randomness at the pixel level makes it harder for the model to capture meaningful patterns and affects the entire image. The noise interferes with the model's ability to recognize key features, leading to a much larger accuracy drop.

3) DenseNet121 Result Discussion: DenseNet121 had a high clean accuracy of 71% and remained moderately robust against all attacks, with an average of a 7.8% accuracy drop across the attacks. This, in our opinion, makes DenseNet121 the most robust model from our experimentation. The reason it performed well against all attacks is because it has a denser communication flow between layers. DenseNet architecture enables direct connections from any layer to all subsequent layers, rather than allowing it to skip to deeper layers. This density allows the model to have multiple pathways for information flow, ensuring that if one part of the network is disturbed, other paths can make up for it. The increased interconnectedness in the network allows for preservation of important features that can be used at later stages of the model, improving its ability to recover from disturbances. This model architecture allows DenseNet121 to be more resilient to adversarial perturbations like Gaussian Noise (9.2% accuracy drop), Color Shifts (8.4% accuracy drop), and obstructions (5.8% accuracy drop).

4) InceptionV1 Result Discussion: For InceptionV1, the Inception modules, which process input at multiple scales using parallel convolution filters (1×1, 3×3, 5×5) and pooling operations, gave the model differential robustness to various attacks. InceptionV1 showed the largest vulnerability to Gaussian noise (15.3% accuracy drop) because this attack disrupts features at all scales simultaneously, affecting every parallel pathway in the Inception modules and disturbing the important low-level features extracted in the early layers, as it alters even the simple patterns that these layers rely on. The model performed better against Color Shift attacks (9.8% drop) thanks to its batch normalization, which normalizes activations and provides invariance to shifts in the input distribution. Additionally, deeper layers of the model focus more on shapes and patterns rather than color, further aiding robustness. InceptionV1 handled Black Patch attacks best (6.9% drop) because its parallel processing paths allowed the model to extract useful features from non-occluded portions of the image, and the global average pooling along with redundant feature learning across spatial dimensions maintained performance even when parts of the image were obstructed.

5) SqueezeNet1.1 Result Discussion: SqueezeNet1.1's architecture contributed significantly to its observed behavior. SqueezeNet's Fire modules, which compress the network through squeeze layers (1×1 convolutions) followed by expand layers (mix of 1×1 and 3×3 convolutions), create a lightweight architecture that shows varying vulnerability to different attack types. The model showed high accuracy drops under Gaussian noise (13.8%) and Color Shift (11.6%), likely because of the reduced redundancy in feature representations caused by extreme parameter efficiency. These global distortions tend may have overwhelmed the model's compressed pathways, making it harder to retain important features. On the other hand, Black Patch attacks led to a much lower drop (5.3%), suggesting that localized occlusions had a less severe impact on the model's performance. This could be due to SqueezeNet's ability to rely on unperturbed portions of the image, allowing it to maintain performance despite partial occlusions. The Fire modules' sequential design and aggressive dimensionality reduction leave fewer alternative pathways for information flow, but the smaller accuracy drop under the Black Patch attack indicates that the model is more resilient to localized perturbations than to global ones like Gaussian noise and Color Shift.

6) ResNet18 Result Discussion: ResNet18's residual connections, which allow information to flow from earlier to later layers through skip connections, were expected to provide robustness against adversarial attacks. However, the model still showed significant vulnerabilities, especially under Gaussian noise (13.5% accuracy drop) and Color Shift (11.2% drop), showing that the architecture's relatively shallow depth of 18 layers is insufficient to handle global perturbations effectively. These attacks overwhelmed the model's ability to maintain clean feature extraction across layers.

However, ResNet18 showed less of an accuracy drop under Black Patch attacks (4.3% drop), which suggests that the residual connections allowed the model to rely on unperturbed portions of the image, helping it maintain performance by bypassing affected layers and passing information through shortcut paths. Despite this, ResNet18's performance under more global attacks highlights the limitations of its shallow architecture in handling certain adversarial perturbations. While the batch normalization layers help reduce internal covariate shift, they were insufficient to prevent large performance drops when facing global pixel-level attacks, such as Gaussian noise and Color Shift. The model's relatively shallow depth may be a key factor contributing to this lack of robustness.

B. Implications of Research

Our research provides an analysis of the limitations of popular CNN architectures when facing black-box adversarial input examples. The results show that certain architectures, specifically those with denser connectivity, perform better in terms of robustness against adversarial attacks, as they are less tightly tuned to clean data distributions. These models, such as DenseNet121, succeed due to multiple pathways for information flow, allowing them to recover more effectively from data disturbances like noise, color shifts, and obstructions. On the other hand, architectures that do not have as much leeway in their information flows and those that rely too much on tightly tuned parameters for clean data, such as EfficientNetB0, are more vulnerable to adversarial examples as they do not account for data obstructions.

C. Limitations of Research

The limitations of our research include the fact that we only tested six relatively compact models, which were fine-tuned to work with a specific facial dataset. This limits the generalizability of our findings to an application focused on facial recognition, rather than a broader range of data types. In addition, our study only considered three black-box adversarial attacks. There are many other types of adversarial attacks that could be explored, including white-box attacks, where the model itself is targeted rather than just perturbing the input data.

VI. Conclusion

A. Summary of Contributions

This study helps the field of facial recognition trustworthiness by providing an evaluation on how different CNN architectures respond to common adversarial attacks. Our testing of six popular CNN architectures against three types of adversarial perturbations revealed some interesting variations in model resilience. We found that models with the highest baseline accuracy on clean data such as EfficientNetB0 are not necessarily the most robust against adversarial attacks. In fact, EfficientNetB0 demonstrated the largest average accuracy drop 14.03% across all attacks even thought it had the highest clean accuracy of 78.1%. On the other hand, DenseNet121 came about as the most robust architecture, with only a 7.8% average drop in accuracy when subjected to adversarial conditions. The reason that DenseNet121 was able to perform the best is due to its unique architectural characteristics, specifically its dense connectivity pattern that enables direct connections from any layer to all subsequent layers. This design allows for multiple pathways of information to flow, ensuring that if one part of the network is disrupted by adversarial perturbations, than other paths can compensate for it. The increased in interconnectedness preserves important features that remain accessible at later stages of the model, improving its ability to recover from disturbances caused by noise, color shifts, and occlusions.

B. Future Work

Future research should focus on five key areas: (1) testing against more sophisticated attack vectors like FGSM and physical-world attacks; (2) developing countermeasures through adversarial training and input preprocessing; (3) creating hybrid architectures that combine EfficientNetB0's accuracy with DenseNet121's robustness; (4) evaluating performance under real-world conditions with varying lighting and angles; and (5) using visualization techniques to identify which facial features are most vulnerable. Beyond our work on real vs. fake classification, future studies should explore face verification and identification tasks, and investigate how adversarial examples transfer between different architectures. In conclusion, our analysis shows that CNN resilience to adversarial attacks varies significantly based on both architecture and attack type. These findings emphasize that adversarial robustness should be considered alongside accuracy when selecting models for security critical face recognition applications.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ”Going Deeper with Convolutions,” arXiv preprint arXiv:1409.4842, 2014. Available at: https://arxiv.org/abs/1409.4842

[2] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep Residual Learning for Image Recognition,” arXiv preprint arXiv:1512.03385, 2015. Available at: https://arxiv.org/abs/1512.03385

[3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, ”Densely Connected Convolutional Networks,” arXiv preprint arXiv:1608.06993, 2016. Available at: https://arxiv.org/abs/1608.06993

[4] M. Tan and Q. V. Le, ”EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” arXiv preprint arXiv:1905.11946, 2019. Available at: https://arxiv.org/abs/1905.11946

[5] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, ”SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size,” arXiv preprint arXiv:1602.07360, 2016. Available at: https://openreview.net/forum?id=S1xh5sYgx

[6] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, ”Designing Network Design Spaces,” arXiv preprint arXiv:2003.13678, 2020. Available at: https://arxiv.org/abs/2003.13678

[7] J. Sen, A. Sen, and A. Chatterjee, ”Adversarial Attacks on Image Classification Models: Analysis and Defense,” arXiv preprint arXiv:2312.16880, 2023. Available at: https://arxiv.org/abs/2312.16880

[8] CIPLAB @ Yonsei University, ”Real and Fake Face Detection,” Kaggle, 2018. Available at: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection

[9] G. Xu, X. Wang, X. Wu, X. Leng, and Y. Xu, ”Development of Skip Connection in Deep Neural Networks for Computer Vision and Medical Image Analysis: A Survey,” School of Computer Science and Engineering, Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, China, 2023.

[10] S. Chen et al., ”Real-time detection of UAV detection image of power line insulator bursting based on YOLOV3,” J. Phys.: Conf. Ser., vol. 1544, p. 012117, 2020, doi: 10.1088/1742-6596/1544/1/012117.

[11] F. Juraev, M. Abuhamad, S. S. Woo, G. K. Thiruvathukal, and T. Abuhmed, ”Impact of Architectural Modifications on Deep Learning Adversarial Robustness,” Department of Computer Science and Engineering, Sungkyunkwan University and Department of Computer Science, Loyola University Chicago, 2023.