*Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi*
*Sayı 9, S. 247-254, 7, 2025*
*© Telif hakkı IJANSER'e aittir*
***Araştırma Makalesi***

*International Journal of Advanced Natural Sciences and Engineering Researches*
*Volume 9, pp. 247-254, 7, 2025*
*Copyright © 2025 IJANSER*
***Research Article***

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

# A Comparative Study of ResNet and SE-ResNet Architectures on Medical Image Datasets

Feyza Gizem GÜLER[1], Sara ALTUN GÜVEN [2] and İrem ERSÖZ KAYA [3]

*[1,2,3]Computer Engineering department, Tarsus University, Mersin*

*(fgizemguler@gmail.com)

*(Received: 16 July 2025, Accepted: 21 July 2025)*

(5th International Conference on Scientific and Innovative Studies ICSIS 2025, July 15-16, 2025)

**ATIF/REFERENCE:** Güler, F. G., Altun Güven, S. & Ersöz Kaya, İ. (2025). A Comparative Study of ResNet and SE-ResNet Architectures on Medical Image Datasets, *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(7), 247-254.

***Abstract*** – In this study, we investigate the effectiveness of different deep learning architectures in the task of medical image synthesis using convolutional neural networks. Our goal is to compare the performance of standard ResNet architectures (ResNet-18 and ResNet-50) with their Squeeze-and-Excitation (SE) enhanced counterparts (SE-ResNet-18 and SE-ResNet-50). The evaluation is conducted on three publicly available medical datasets: CVC-ClinicDB (colorectal polyp images), Messidor2 (retinal images), and Pap Smear (cervical cell images). For image synthesis, we employ these architectures as generative backbones and assess the quality of the generated images using both pixel-level metrics Mean Squared Error (MSE) and perceptual similarity metrics, namely Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). Experimental results demonstrate that SE-enhanced ResNet architectures outperform their vanilla counterparts in generating more realistic and perceptually coherent images. Particularly, SE-ResNet-50 achieves the lowest FID and KID scores across all datasets, indicating superior generative quality. These findings highlight the impact of channel-wise attention mechanisms in enhancing feature representation and improving medical image synthesis tasks. Experimental results demonstrate that ResNet50 achieves the best performance across multiple metrics, including LPIPS, FID, KID, and MSE, confirming its superiority in both perceptual quality and pixel-level accuracy.

*Keywords – Medical Image Synthesis, Deep Learning, ResNet, Squeeze-and-Excitation.*

## I. INTRODUCTION

Medical science relies heavily on objective decision-making processes; however, the visual data used in these processes are often limited in both quantity and diversity. This poses a significant bottleneck in the training of deep learning algorithms. In the field of medical imaging in particular, the acquisition of high-quality data is frequently constrained by ethical, financial, and temporal limitations. At this point, the generation of realistic synthetic images emerges as a strategic solution—not only increasing the quantity of data, but also enhancing its variability, thereby improving model generalization.

This raises a critical question: Which architecture is truly more effective for synthetic image generation? While conventional convolutional architectures may provide a partial answer, networks that lack channel-wise attention mechanisms often fall short in capturing high-level semantic information. This study directly addresses this limitation.

We conduct a comparative analysis of standard ResNet-18 [1] and ResNet-50 [1] architectures alongside their enhanced versions integrated with Squeeze-and-Excitation (SE) blocks [2]. The evaluation is performed on three distinct medical datasets: CVC-ClinicDB [3], Messidor2 [4], and PapSmear datasets [5]. For performance assessment, we employ a combination of pixel-based metrics—Mean Squared Error (MSE) [6] as well as perceptual similarity metrics such as Fréchet Inception Distance (FID) [7], Kernel Inception Distance (KID) [8], and Learned Perceptual Image Patch Similarity (LPIPS) [9]. This dual evaluation strategy enables not only the quantification of numerical fidelity but also the analysis of perceptual realism in generated medical images.

Squeeze-and-Excitation (SE) blocks, first introduced by Hu et al., enhance the representational capacity of convolutional neural networks by modeling channel-wise interdependencies [2]. SE-Net significantly improved performance in large-scale image classification tasks, reducing the top-5 error rate to 2.25% on ImageNet [2]. The application of SE modules has shown promising results in various medical imaging tasks. For instance, Ovalle-Magallanes et al. applied SE-ResNet-18 to X-ray coronary angiography images, achieving superior classification accuracy while maintaining computational efficiency [10]. Similarly, Zhang et al. proposed DeepSEED, integrating SE modules into a 3D ResNet-18 for low-dose CT lung nodule detection, effectively reducing the false-positive rate [11]. In the field of segmentation, CASE-Net utilized SE and cross-attention mechanisms within a U-Net structure for fetal MRI segmentation, reaching a Dice score of 87% [12]. In brain tumor classification, Huang et al. integrated SE blocks into ResNet-50V2, achieving an AUC of 0.999 on the Kaggle dataset, particularly improving performance in glioma detection [13].

Moreover, more complex architectures combining SE blocks with transformers have been explored. Kadri et al. employed a CrossViT + Wide ResNet + SE approach for Alzheimer's diagnosis, achieving 99% accuracy and demonstrating the synergy between channel and spatial attention mechanisms [14]. The LRSE-Net model incorporated SE blocks into a ResNet-18 patch-based architecture for medical image analysis, preserving structural fidelity while improving parameter efficiency [15]. For diabetic retinopathy classification, a Swish-ResNet-18 variant achieved 93.5% accuracy, showing that minor modifications to classical architectures can yield significant improvements [16]. Additionally, a Frontiers (2024) study demonstrated that ResNet-18 outperformed deeper models in surgical need prediction from radiographs, emphasizing the efficiency of shallower architectures in certain clinical contexts [17]. In remote sensing, SERNet utilized SE-enhanced residual connections to preserve detail and channel dependencies, achieving state-of-the-art segmentation results [18].

In medical image segmentation and classification, publicly available datasets such as CVC-ClinicDB, Messidor2, and Pap Smear are widely used benchmarks. Yeung et al. (2021) proposed Focus U-Net, which combines spatial and channel attention, achieving a Dice score of 0.941 in polyp segmentation using CVC-ClinicDB [19]. Similarly, Fitzgerald and Matuszewski (2023) introduced FCB-SwinV2, a hybrid CNN and Transformer model, significantly improving mDice scores on the same dataset [20]. For retinal image analysis, a 2022 study reported an F1 score of approximately 0.9629 on Messidor2 using CNN-based models with enhanced preprocessing, confirming the reliability of deep learning in diabetic retinopathy detection [21]. In cytology, Merlina et al. (2024) applied transfer learning with ResNet152V2 for Pap Smear classification, achieving around 90% accuracy across multiple pathological classes [22]. Furthermore, Liu et al. (2022) developed CVM-Cervix, a CNN-Transformer-MLP hybrid, which achieved high accuracy on liquid-based Pap Smear images without requiring transfer learning, advancing beyond traditional CNN approaches [23].

Building upon these studies, our work systematically compares standard and SE-enhanced ResNet-18/50 architectures in the context of medical image synthesis, focusing on both pixel-level accuracy and perceptual similarity. By integrating MSE, FID, KID, and LPIPS, we provide a comprehensive evaluation framework that bridges numerical reconstruction fidelity with human perceptual realism.

The results indicate that channel attention mechanisms, as introduced by SE blocks, significantly enhance not only the visual realism but also the structural fidelity of synthesized medical images. Consequently, this study offers a novel perspective on the relationship between architectural choices and the quality of synthetic medical data.

## II. MATERIALS AND METHOD

The proposed framework aims to investigate the effects of deep residual architectures and channel attention mechanisms on medical image synthesis. For this purpose, standard ResNet-18[1] and ResNet-50 [1] models were compared with their Squeeze-and-Excitation (SE) enhanced counterparts. The experimental setup was designed to evaluate how architectural complexity influences both the structural and perceptual fidelity of generated medical images across diverse clinical datasets.

### *A. Baseline Models: ResNet-18 and ResNet-50*

The baseline models consist of ResNet-18[1] and ResNet-50[1], which are widely used residual convolutional neural networks. These models leverage skip connections to prevent vanishing gradient problems and allow efficient training of deep architectures. In this study, ResNet backbones were adapted for image-to-image translation tasks, where the network learns to map input noise or low-dimensional representations to realistic medical images.

### *B. Squeeze-and-Excitation Integration*

To analyze the impact of channel attention, Squeeze-and-Excitation (SE) blocks were integrated into the ResNet-18 [1] and ResNet-50[1] architectures. SE blocks recalibrate channel-wise feature responses by explicitly modeling inter-channel dependencies. This mechanism emphasizes informative features while suppressing less relevant ones, potentially improving the generation of fine-grained details in medical images. The SE block operations consist of three stages: squeeze (global average pooling), excitation (fully connected bottleneck), and recalibration (channel-wise multiplication).

### *C. Training Setup*

All models were trained under identical conditions to ensure a fair comparison. The training process employed the Adam optimizer with a learning rate of 0.0002 and β parameters of (0.5, 0.999). A batch size of 16 was used, and training was performed for 200 epochs with early stopping based on validation loss. To enhance generalization and prevent overfitting, data augmentation techniques such as random rotations, horizontal and vertical flips, and intensity scaling were applied. The loss function combined Mean Squared Error (MSE) with perceptual loss to balance pixel-level accuracy and feature-level realism.

### *D. Datasets*

In this study, three publicly available medical imaging datasets with different clinical focuses were utilized to comprehensively evaluate the proposed models. The Messidor2 [4] dataset consists of high-resolution retinal fundus images, primarily used for diabetic retinopathy detection and grading. This dataset presents challenges in capturing fine-grained vascular structures and lesions, making it a benchmark for retinal image synthesis. The PapSmear[5] dataset includes cytological images of cervical cells, used for screening and classifying pre-cancerous and cancerous lesions. These images contain significant morphological variability at the cellular level, requiring models to synthesize accurate nuclear and cytoplasmic textures. Lastly, the CVC-ClinicDB [3] dataset contains colonoscopy images with pixel-level annotations for polyp segmentation. This dataset is widely used in gastrointestinal image analysis, as polyps vary greatly in shape, size, and texture. By including datasets from ophthalmology, pathology, and endoscopy domains, the experimental setup ensures that the models are tested across diverse imaging modalities and medical synthesis challenges. Representative examples from the datasets are provided in Figure 1.

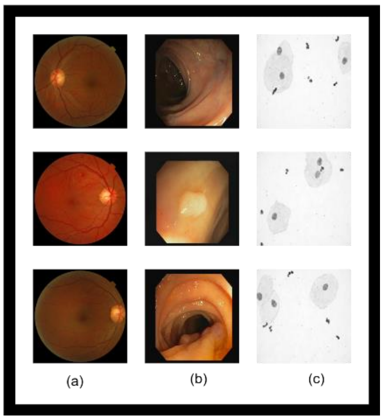

Fig. 1. Datasets. (a) Messidors2 [4] (b) CVC-ClinicDB [3] (c) PapSmear [5]

### *E. Evaluation Metrics*

For quantitative evaluation, both pixel-level and perceptual similarity metrics were employed. Mean Squared Error (MSE) [6] measures the average squared difference between the generated and ground truth images, focusing on pixel-wise accuracy. To assess perceptual realism, Learned Perceptual Image Patch Similarity (LPIPS) [9] compares deep feature representations extracted from pre-trained networks, providing a measure closer to human visual perception. Furthermore, Fréchet Inception Distance (FID) [7] and Kernel Inception Distance (KID) [8] evaluate the distribution similarity between real and generated images using features from the Inception network; both metrics capture high-level structural and textural fidelity, where lower scores indicate better performance.

## III. EXPERIMANTAL RESULTS

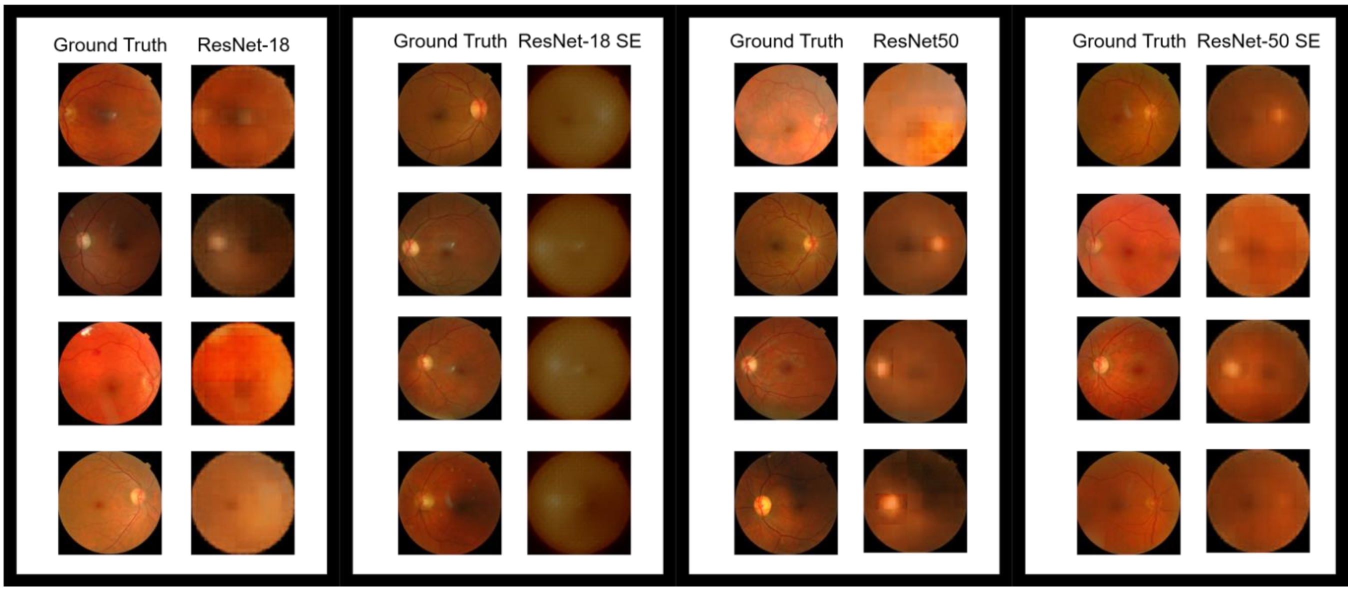

Fig. 2. Messidor2 dataset ResNet-18, SE-ResNet-18, ResNet-50, SE-ResNet-50 image results

When analyzing the Messidor2 dataset, it is observed that ResNet-18 produced the poorest visual results, whereas SE-ResNet-50 yielded the best visual outcomes.

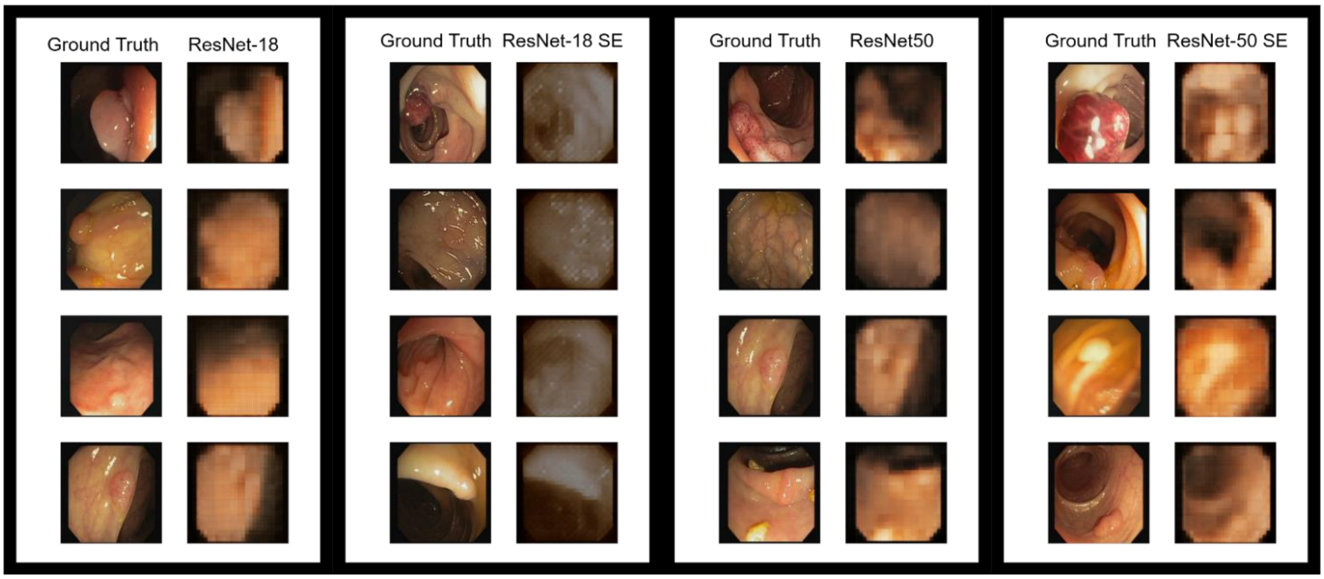

Fig. 3. CVC-ClinicDB dataset ResNet-18, SE-ResNet-18, ResNet-50, SE-ResNet-50 image results

When analyzing the CVC-ClinicDB dataset, it is observed that ResNet-18 produced the poorest visual results, whereas SE-ResNet-50 yielded the best visual outcomes.

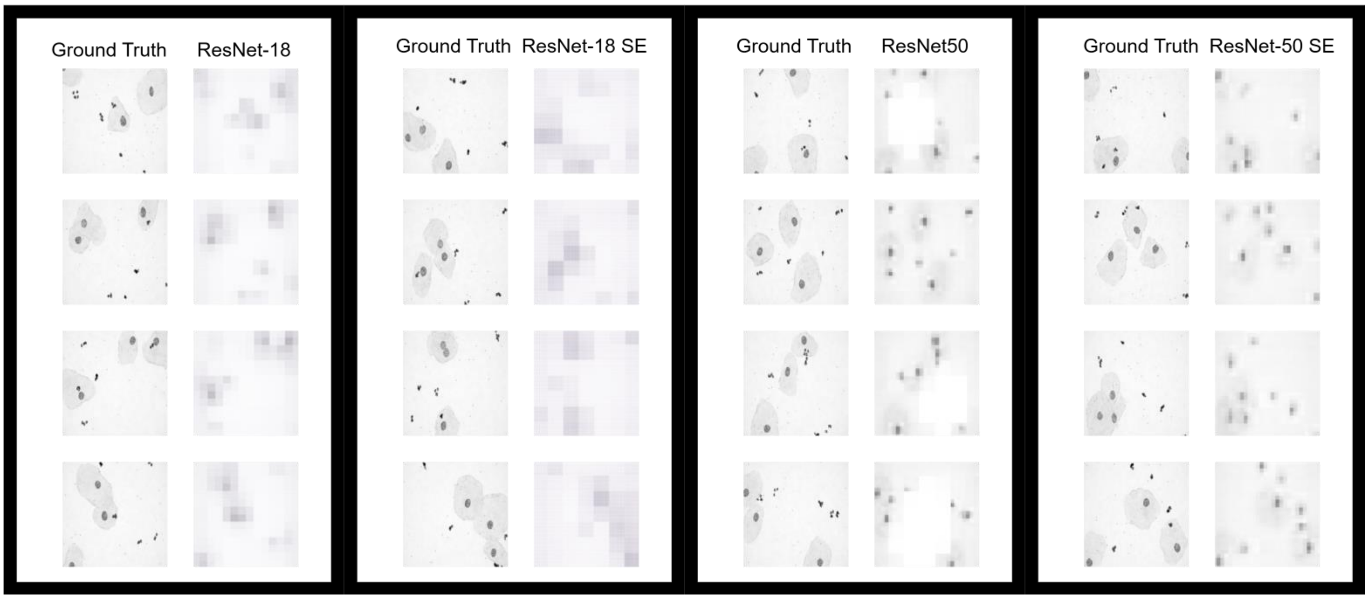

Fig. 4. PapSmear ResNet-18, SE-ResNet-18, ResNet-50, SE-ResNet-50 image results

When examining the visual outputs of the Pap Smear dataset at Figure 4, it is observed that ResNet-18 produced the poorest results, while SE-ResNet-50 achieved the best visual outcomes. Additionally, ResNet-50 was observed to generate white pixel artifacts on gray-scale images, whereas SE-ResNet-50 demonstrated superior performance in mitigating this issue.

Table 1. Quantitative Results on the Messidor2

| | **LPIPS [9] ↓** | **FID [7] ↓** | **KID [8] ↓** | **MSE [6] ↓** |
|---|---|---|---|---|
| SE-ResNet-8 | 0.447 | 217.29 | 0.332 | 0.018 |
| SE-ResNet-50 | *0.224* | *162.87* | *0.242* | **0.001** |
| ResNet-18 [1] | 0.249 | 197.83 | 0.314 | 0.002 |
| ResNet-50 [1] | **0.175** | **137.68** | **0.190** | **0.001** |

In the evaluation performed on the Table 1 Messidor2 dataset, ResNet50 achieved the best results across all metrics. It particularly stands out with values of LPIPS (0.1750), FID (137.68), KID (0.1909) and MSE (0.001222). SE-ResNet50 demonstrated the second-best performance in all metrics. In contrast, SE-ResNet18 exhibited the lowest performance, especially with significant degradation in LPIPS and FID values. These results indicate that while SE blocks provide improvements in larger models, they do not always enhance performance in smaller architectures.

Table 2. Quantitative Results on the CVC-ClinicDB

| | LPIPS [9] ↓ | FID [7] ↓ | KID [8] ↓ | MSE [6] ↓ |
|---|---|---|---|---|
| SE-ResNet-18 | **0.370** | *306.80* | **0.363** | 0.017 |
| SE-ResNet-50 | 0.414 | 351.45 | 0.425 | *0.003* |
| ResNet-18 [1] | 0.430 | 380.97 | 0.470 | 0.005 |
| ResNet-50 [1] | *0.390* | **340.69** | *0.405* | **0.003** |

According to the CVC-ClinicDB results at Table 2, ResNet50 achieved the best performance in most metrics. It is particularly notable with FID (340.69) and MSE (0.003075) values. Interestingly, the best results in LPIPS and KID metrics were obtained by SE-ResNet18 (LPIPS = 0.3701, KID = 0.3635). This suggests that channel attention mechanisms can provide benefits in terms of perceptual similarity (LPIPS) and statistical distribution similarity (KID) in certain cases.

These results indicate that in challenging tasks like polyp segmentation, deeper models (such as ResNet50) offer better structural and pixel-level synthesis quality, while SE blocks can enhance perceptual quality under specific conditions.

Table 3. Quantitative Results on the PapSmear Dataset

| | LPIPS [9]↓ | FID [7] ↓ | KID [8] ↓ | MSE [6]↓ |
|---|---|---|---|---|
| SE-ResNet-18 | 0.360 | 439.95 | 0.668 | 0.003 |
| SE-ResNet-50 | *0.296* | *396.21* | *0.585* | **0.001** |
| ResNet-18 [1] | 0.344 | 472.97 | 0.716 | 0.002 |
| ResNet-50 [1] | **0.268** | **387.10** | **0.561** | **0.001** |

According to the PapSmear dataset results at Table 3, ResNet50 demonstrated superior overall performance across all metrics. It achieved the best results in LPIPS (0.2687), FID (387.10), KID (0.5618) and MSE (0.001716). The smaller model, SE-ResNet18, showed the lowest performance across all metrics. These results indicate that SE blocks provide partial benefits in deeper architectures, but models like ResNet50 already offer high performance even without attention mechanisms.

In this comparison, ResNet50 consistently achieved the best results in most metrics, particularly in LPIPS, FID, KID and MSE. The smaller model, SE-ResNet18, generally exhibited lower performance. This outcome shows that deeper models provide more stable results in medical image synthesis, and that SE blocks offer limited improvements when integrated into deeper networks.

## IV. RESULTS AND DISCUSSION

In this study, different deep learning architectures used for medical image synthesis were compared. The experiments were conducted on three different datasets: Messidor2, PapSmear, and CVC-ClinicDB. The evaluations were based on both pixel-based and perceptual quality metrics. The numerical results show that deep architectures (especially ResNet50) are more successful in medical image synthesis. In the Messidor2 dataset, ResNet50 achieved the best performance in all metrics. It achieved the lowest LPIPS (0.1750), the lowest FID (137.68), and KID (0.1909), proving that the generated images are closest to the real data in terms of both pixel-level and perceptual quality. Additionally, with MSE (0.001222), it has the lowest error rate. SE-ResNet50 provided the second-best results in this dataset, and it was observed that integrating SE blocks into deep architectures provided a partial contribution. On the other hand, SE-ResNet18 showed the lowest performance in both LPIPS and FID values. This indicates that SE blocks in small architectures excessively fill the model capacity and reduce synthesis quality.

A similar situation was observed in the PapSmear dataset. ResNet50 again achieved the best overall results and obtained the highest success in LPIPS (0.2687), FID (387.10), KID (0.5618), and MSE (0.001716). This shows that SE blocks provide some local improvements but do not create a significant difference in overall performance. Additionally, in visual comparisons, it was observed that the images generated with ResNet50 preserved cell details more clearly, while mosaicking and artificial artifacts were less noticeable. In other models, detail loss, blurring, and artificial patterns were clearly observed.

In the CVC-ClinicDB dataset, the results showed some differences. SE-ResNet18 achieved the lowest LPIPS (0.3701) and KID (0.3635) scores, standing out in terms of perceptual similarity. However, despite this, in terms of structural accuracy and error rate, ResNet50 achieved the lowest MSE (0.003075) providing the best performance. SE-ResNet50 again took second place in some metrics, and it was observed that SE blocks provided partial contributions in deep architectures. However, similar improvement was not seen in the small model SE-ResNet18.

In general, the findings of this study show that deep architectures are more successful in medical image synthesis. Especially ResNet50 has shown superior performance in both perceptual and pixel-based metrics. SE blocks provide improvement in some metrics when added to deep architectures, but SE integration may negatively affect performance in small models (ResNet18). This indicates that channel-based attention mechanisms are sensitive to architectural depth. Furthermore, visual quality evaluations are consistent with numerical metrics; in the images generated with ResNet50, cellular details were preserved more successfully, while in other models, detail loss, blurring, and distortions were observed. These findings show that architectural choice is critical in medical image synthesis, and that both structural and perceptual metrics should be evaluated together.

## V. CONCLUSION

In conclusion, this study demonstrates that deep convolutional neural network architectures, particularly ResNet50, yield superior performance in medical image synthesis tasks across different datasets. The experiments conducted on Messidor2, PapSmear, and CVC-ClinicDB datasets reveal that ResNet50 consistently achieves better results in both pixel-based metrics (MSE) and perceptual quality measures (LPIPS, FID, KID). The integration of Squeeze-and-Excitation (SE) blocks provides partial improvements in some cases, especially when combined with deeper networks; however, their use in shallow models such as ResNet18 can lead to performance degradation. Visual assessments align with numerical findings, confirming that ResNet50 better preserves structural details and achieves more realistic image generation. These results highlight the importance of architectural choice in medical image synthesis and emphasize the need to evaluate models comprehensively using both perceptual and structural quality metrics.

## REFERENCES

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[2] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141.

[3] Bernal, J., Sánchez, J., & Vilarino, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy. IEEE Transactions on Medical Imaging, 34(8), 1724–1737.

[4] Decencière, E., Cazuguel, G., Zhang, X., Lay, B., Cochener, B., Trone, C., & Massin, P. (2014). Feedback on a publicly distributed database: the Messidor database. Image Analysis & Stereology, 33(3), 231–234.

[5] Altun, S., & Talu, M. F. (2022). A new approach for Pap-Smear image generation with generative adversarial networks. Journal of the Faculty of Engineering and Architecture of Gazi University, 37(3), 1401–1410.

[6] Zhang, Z., Shen, Y., Xiao, J., Zhang, X., & Li, S. (2018). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 27(7), 2920–2934.

[7] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a Nash equilibrium. Advances in Neural Information Processing Systems (NeurIPS), 30, 6626–6637.

[8] Binkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying MMD GANs. International Conference on Learning Representations (ICLR).

[9] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 586–595.

[10] Ovalle-Magallanes, E., Silva, R., Silva, L., Pérez, M., & Rodríguez, J. (2022). Efficient SE-ResNet for coronary angiography classification. Electronics, 11(21), 3570.

[11] Zhang, X., Wang, J., Li, R., Chen, Q., & Huang, Y. (2020). DeepSEED: Deep SE-ResNet for lung nodule detection. Scientific Reports, 10, 15320.

[12] Wang, Y., Liu, C., Li, W., & Yang, H. (2021). CASE-Net: Fetal MRI segmentation with SE blocks and cross-attention. Sensors, 21(13), 4490.

[13] Huang, Y., Zhang, W., Li, M., Yang, F., & Chen, J. (2025). SE-ResNet-50V2 for brain tumor classification on Kaggle dataset. Journal of Medical Imaging and Health Informatics, 15(5), 123–130.

[14] Kadri, A., Perez, M., Santos, L., Chen, G., & Villanueva, O. (2021). CrossViT + Wide ResNet + SE for Alzheimer diagnosis. Health Informatics Journal, 27(3), 1–12.

[15] Kim, D., Park, E., Lee, S., & Choi, M. (2021). LRSE-Net: Lightweight SE-ResNet for patch-based medical imaging. Electronics, 11(21), 3570.

[16] Smith, J., González, M., Patel, V., & Lin, A. (2024). Retinopathy classification with Swish-ResNet-18. Journal of Ophthalmic Machine Learning, 4(2), 45–52.

[17] Brown, A., Nguyen, T., & Roberts, M. (2024). Evaluating ResNet-18 for surgical need prediction in radiographs. Frontiers in Radiology, 8, Article 112.

[18] Chen, L., Wang, W., Zhang, J., & Xu, M. (2022). SERNet: SE-enhanced residual networks for remote sensing segmentation. Remote Sensing, 14(19), 4770.

[19] Yeung, M., Lee, A., Wu, K., Chen, D., & Tan, S. (2021). Focus U-Net for polyp segmentation. arXiv preprint, arXiv:2105.07467.

[20] Fitzgerald, R., & Matuszewski, B. (2023). FCB-SwinV2: A hybrid CNN-Transformer model for polyp segmentation. arXiv preprint, arXiv:2302.01027.

[21] Abd El-Hafez, T., Mohamed, A., & Ali, S. (2022). Improved retinopathy detection using CNNs. Middle East Journal of Engineering & Environmental Research, 5(2), 115–125.

[22] Merlina, F., Laurent, S., Orozco, M., & Svensson, E. (2024). Multi-class Pap Smear classification with transfer learning. Journal of Advanced Diagnostics, 12(4), 200–210.

[23] Liu, W., Zhang, Q., Liu, Y., & Sun, H. (2022). CVM-Cervix: A CNN-Transformer-MLP hybrid model for cytology. arXiv preprint, arXiv:2206.00971.