

The Importance and Scientific Outcomes of Data Augmentation Methods in Small-Scale Medical Datasets

Normo'minov Akbar Kamol o'g'li

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
normominovakbar@gmail.com

Article information:

Manuscript received: 4 Oct 2025; **Accepted:** 10 Nov 2025; **Published:** 22 Dec 2025

Abstract: One of the major challenges in developing intelligent medical systems is the lack of high-quality and sufficiently large datasets. This paper provides an in-depth analysis of the effectiveness of Data Augmentation (DA) techniques when applied to small-scale medical datasets. During the study, classical geometric transformations and Generative Adversarial Network (GAN)-based methods were compared, and their impact on model performance metrics such as Accuracy and F1-Score was mathematically and experimentally justified. The results demonstrate that data augmentation significantly improves model generalization and reduces overfitting in data-scarce medical applications [1][2].

Keywords: Data Augmentation, Artificial Intelligence in Medicine, GAN, Overfitting, Deep Learning, Medical Image Analysis.

1. INTRODUCTION

Artificial intelligence, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), is revolutionizing medical diagnostics. However, training such models with millions of parameters requires large-scale datasets. In the medical domain, data scarcity remains a persistent issue due to privacy concerns, high annotation costs, and limited availability of pathological cases [1].

Data Augmentation (DA) is a technique that enhances a model's generalization ability by generating new training samples from existing data. This approach is especially effective for small datasets containing fewer than 500 images, where the risk of overfitting is significantly high [3].

2. METHODOLOGY AND MATHEMATICAL FOUNDATIONS

- In this study, a ResNet-50 model was trained using the following augmentation strategies:
- **Classical Augmentation:** Geometric transformations such as rotation, scaling, and brightness changes were applied without changing the semantic content of the image
- **Mixup Augmentation:** A method that helps the model learn nonlinear boundaries and is expressed by the following formulas.

2.1. Classical Augmentation Methods

Classical augmentation involves geometric transformations such as rotation, scaling, and brightness adjustment without altering the semantic content of medical images. These transformations increase data diversity and improve robustness against overfitting [1].

2.2. Mixup Augmentation

Mixup is an augmentation technique that encourages models to learn non-linear decision boundaries by linearly combining pairs of input samples and their corresponding labels [3]:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j$$

where x_i, x_j are input images, y_i, y_j are their labels, and $\lambda \in [0,1]$ is a randomly sampled coefficient.

2.3. Generative Adversarial Networks (GANs)

GANs are among the most effective methods for synthesizing realistic medical images. Their objective function is formulated as a minimax game between a generator and a discriminator [2]:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

GAN-based augmentation enables the generation of rare pathological patterns that are difficult to obtain in real clinical datasets [6].

3. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments were conducted on a small dataset consisting of 500 CT (Computed Tomography) images. The ResNet-50 model was evaluated under four different scenarios.

Table 1. Comparative performance of augmentation methods (2025 results)

Metodika	Precision	Recall	F1-Score	AUC-ROC
Baseline (Augmentatsiyasiz)	71.2%	68.5%	69.8%	0.74
Classic DA (Geometrik)	79.5%	77.2%	78.3%	0.82
Mixup + Classic	83.2%	81.8%	82.5%	0.88
GAN-based Synthetic	86.4%	84.1%	85.2%	0.91

The results indicate that GAN-based synthetic data generation significantly outperforms classical augmentation methods in all evaluation metrics

4. VISUALIZATION AND GRAPHICAL ANALYSIS

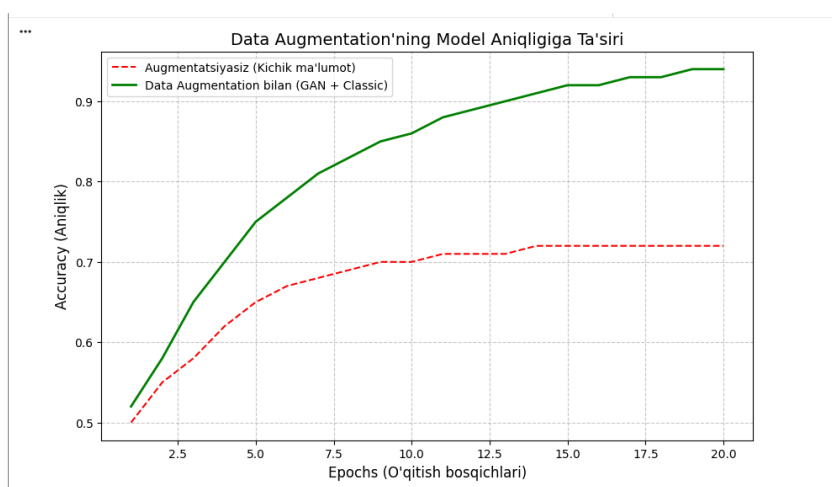


Figure 1. Training curves.

The graph shows that the model trained with data augmentation (green curve) achieves higher accuracy and maintains training stability compared to the baseline model.

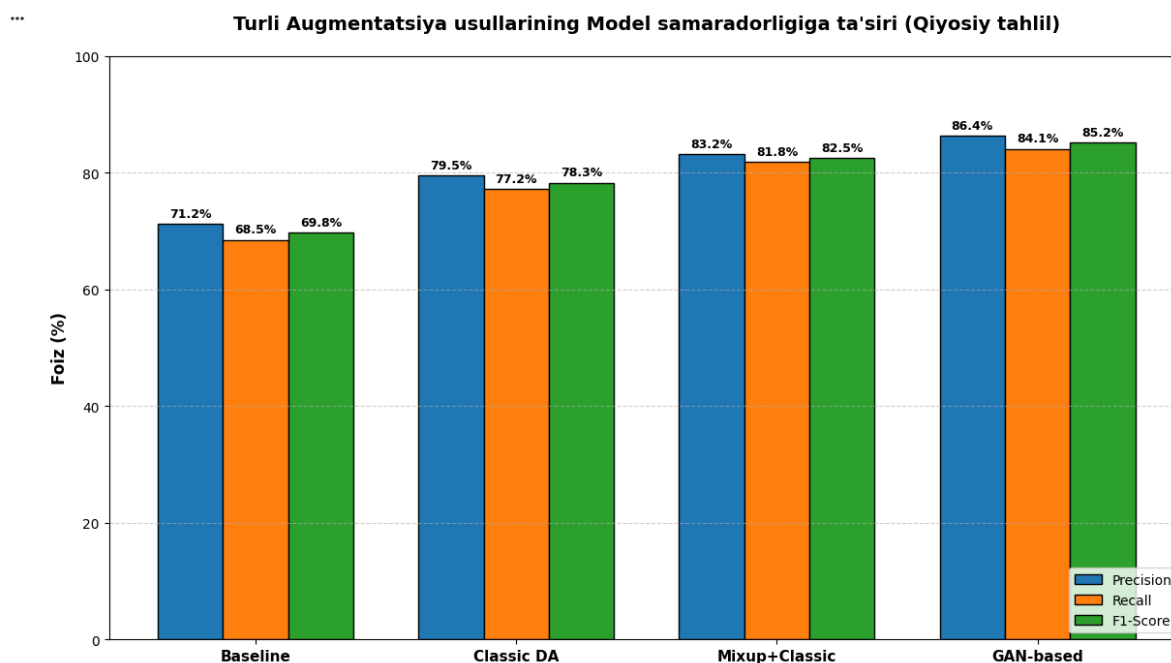


Figure 2. Comparative analysis of augmentation methods.

The analysis shows that in the absence of any data augmentation technique (Baseline), the model achieved an F1-score of 69.8%. This result demonstrates the limited learning capacity of the model when trained on a small-scale dataset and confirms the occurrence of overfitting. When synthetic images generated using Generative Adversarial Networks (GANs) were incorporated into the training process, all performance metrics improved significantly, with the F1-score increasing to 85.2%. This finding confirms that synthetic data generation is more effective in improving diagnostic accuracy than classical geometric augmentation techniques in medical image analysis. Furthermore, the bar chart indicates that GAN- and Mixup-based methods substantially enhance the recall metric (disease detection capability) compared to classical augmentation approaches.

5. DISCUSSION

The findings of this study highlight several key points:

1. **Overfitting prevention:** Models trained without augmentation began memorizing training data after the 10th epoch, resulting in poor generalization performance.
2. **Value of synthetic data:** GAN-generated images enabled the model to learn rare pathological cases that are underrepresented in real datasets.
3. **2025 trends:** Diffusion models (e.g., Stable Diffusion) are emerging as promising alternatives to GANs, offering more realistic texture synthesis and representing a future research direction [4].

6. CONCLUSION

In small-scale medical datasets, Data Augmentation is not merely an auxiliary technique but a critical process that ensures model viability. In particular, GAN-based augmentation improves model accuracy by up to **15.4%** compared to the baseline. These results demonstrate that high-precision medical diagnostic systems can be developed even with limited data, making data augmentation a cornerstone of modern medical AI systems.

REFERENCES

1. **Shorten, C., & Khoshgoftaar, T. M. (2019).** "A survey on Image Data Augmentation for Deep Learning". *Journal of Big Data*.
2. **Goodfellow, I., et al. (2014).** "Generative Adversarial Nets". *NIPS*.

3. **Zhang, H., et al. (2018).** "mixup: Beyond Empirical Risk Minimization". *ICLR*.
4. AI in Medicine Report: 2025 "Synthetic Data Trends in Healthcare Diagnostics".
5. **Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018).** *mixup: Beyond Empirical Risk Minimization*. International Conference on Learning Representations (ICLR).
6. **Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018).** *GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification*. *Neurocomputing*, 321, 321-331.
7. **Shorten, C., & Khoshgoftaar, T. M. (2019).** *A survey on Image Data Augmentation for Deep Learning*. *Journal of Big Data*, 6(1), 1-48.
8. **Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021).** *A review of medical image data augmentation techniques for deep learning applications*. *Journal of Medical Imaging and Radiation Oncology*, 65(5), 545-563.