УДК 004.8; 004.8.032.26

EVALUATING VULNERABILITIES AND PROTECTIONS OF AI METHODS IN COMPUTER-ASSISTED IMAGE-BASED DIAGNOSIS

V.A. Kovalev
Leading researcher, Dept. Biomedical Image Analysis, United Institute of Informatics Problems, National Academy of Sciences of Belarus, PhD
vassili.kovalev@gmail.com

E.V. Snezhko
Head of Dept. Biomedical Image Analysis, United Institute of Informatics Problems, National Academy of Sciences of Belarus, PhD
eduard.snezhko@gmail.com

D.S. Karpenko
Junior researcher, trainee, Dept. Biomedical Image Analysis, United Institute of Informatics Problems, National Academy of Sciences of Belarus
karpenko.dima.s11@gmail.com

A.G. Varvashevich
Junior researcher
varvashevichangelina@gmail.com

V.A. Kovalev
Graduated from the Tomsk Polytechnic University. Research interests: image analysis, image generation, AI security

E.V. Snezhko
Graduated from the Belarusian State University. Research interests: image analysis, AI security.

D.S. Karpenko
Graduated from the Belarusian State University. Research interests: generative neural networks, federated learning.

A.G. Varvashevich
Graduated from the Belarusian State University. Research interests:image analysis, Adversarial Attacks.

Abstract. This paper examines three safety aspects of AI in medical image analysis: vulnerability to adversarial attacks, patient data protection via generative models, and federated learning without data sharing. Experiments on over 260,000 chest X-rays, CT slices, and histological slides show that white-box adversarial attacks (FGSM, AutoAttack, Carlini-Wagner) can reduce classification accuracy to 0% when no defense is applied. Among three defense strategies, a high-level task-driven denoiser proved most effective, restoring accuracy up to 100%. Generative models (DC-GAN, Pro-GAN, diffusion) produced realistic synthetic images; using them for data augmentation led to only a modest accuracy drop (2.2-3.5% for deep learning). Federated learning (FedAVG) succeeded only for homogeneous datasets (chest X-

rays) but failed for highly variable histological images. The paper concludes that adversarial attacks pose critical threats, generative models enable privacy-preserving augmentation, and federated learning requires careful modality-specific adaptation.

Keywords: AI safety, medical imaging, adversarial attacks, deep learning, generative models, federated learning, data privacy

Introduction. Recently, artificial intelligence (AI) methods have seen significant development and proliferation in various application areas, from medicine and finance to solving technical problems in industry [1].

In healthcare, AI-based methods significantly facilitate the analysis of medical images used in computerized disease diagnosis [2, 3, 4].

The use of AI methods and software tools allows for more effective detection of various diseases, increases the standardization of diagnostic processes, and helps physicians make more informed decisions. However, the intensive processes of developing and implementing AI solutions into practice are not yet supported by appropriate technical solutions for the protection of patient personal data and AI software, which differ significantly from existing software [5, 6].

The aim of this paper is to examine the results of experiments assessing vulnerabilities associated with adversarial attacks on deep neural networks, the possibilities of patient identification from medical images, as well as the effectiveness of federated learning of neural networks, which does not require the transfer of patient personal data between its participants.

Attacks on Deep Neural Networks. During research and development of various deep neural network architectures, it has been discovered that under certain, deliberately created conditions, neural networks can become extremely unstable. For example, a well-trained and seemingly very reliable neural network recognizing cancerous tumors and normal tissue in histological biopsy images (the gold standard in oncology) may classify tumor tissue as normal and areas of normal tissue as malignant neoplasms. The reason is that special algorithms have been discovered and implemented that minimally alter the input image, after which, for some reason, it is misclassified by the neural network. These image modifications are so minor that they are often indistinguishable to the human eye.

The process by which such an image is generated and fed to the network is called an adversarial attack [7].

Obviously, this problem creates a serious security breach in neural networks when solving image recognition and classification tasks in medicine and other areas of AI application.

Computational experiments were conducted on a large sample of more than 260,000 medical images of the following three types:

- Digital chest X-rays of people who underwent regular examinations within a computerized telemedicine system operating on the basis of medical institutions in Minsk (15,600 people aged 18 and older);
- Computed tomography (CT) images of patients with pulmonary tuberculosis undergoing treatment at one of the specialized medical centers in Belarus (3D tomograms of 414 patients divided into 53,677 2D images of axial slices of size 512×512 pixels each);
- Histological images of patients from one of the oncology centers in Belarus undergoing treatment for thyroid cancer and ovarian cancer (tissue samples routinely stained with hematoxylin and eosin, as well as special immunohistochemical preparations such as D2-40, Ki6, CD-105, FRES, and CD-31; total 192,000 images of size 256×256 pixels).

Examples of original images are shown in Fig. 1.

a) b) c)

Figure 1. Examples of original images: a) X-ray; b) CT; c) histological.

In the computational experiments, the EfficientNet-B3 network [8] was chosen as the target neural network. Only "white-box" attacks were performed, i.e., all elements of the target network architecture were assumed known to the attacker [9]. For each of the three image types, a corresponding version of the EfficientNet-B3 neural network was trained to classify input images into one of two conditionally designated classes: Normal and Pathology.

The classification accuracy under normal conditions, i.e., before the attack, was recorded for subsequent statistical analysis.

Attack effectiveness was evaluated by the difference in image classification accuracy before and after the attack. Three independent series of experiments were conducted for three different types of attacks, including gradient attacks of the FGSM type [9], AutoAttack proposed by Croce and Hein [10], and Carlini-Wagner attacks [11].

In all three attacks, the attack strength (magnitude) parameters were chosen such that the presence of pixel modifications to the original images could not be reliably detected visually.

Examples of original CT (top row) and histological (bottom row) images and their modified adversarial versions are shown in Fig. 2.

After the attacks, the adversarial images underwent a defense procedure. Currently, researchers propose various methods of defense against adversarial attacks on neural networks. In this work, an experimental evaluation of the effectiveness of three types of defense, presented below, was carried out.

Adversarial training of neural networks. This type of defense involves training networks using mixed training datasets containing both normal images and images that have already been maliciously modified [9].

Denoiser driven by high-level task data. This defense is based on the idea that malicious image modifications can be considered as a type of "noise" that should be removed [12].

To perform this function, an auxiliary neural network is trained that tracks the relationships between local image elements and high-level data, such as Normal/Pathology image class labels, etc. *Specialized MagNet neural network.*

Similar to the previous defense method, the MagNet network treats adversarial image modifications as special types of noise and is trained to localize and remove them using the principle of a filtering auto-encoder [13].

The experimental results are presented in the Table 1.

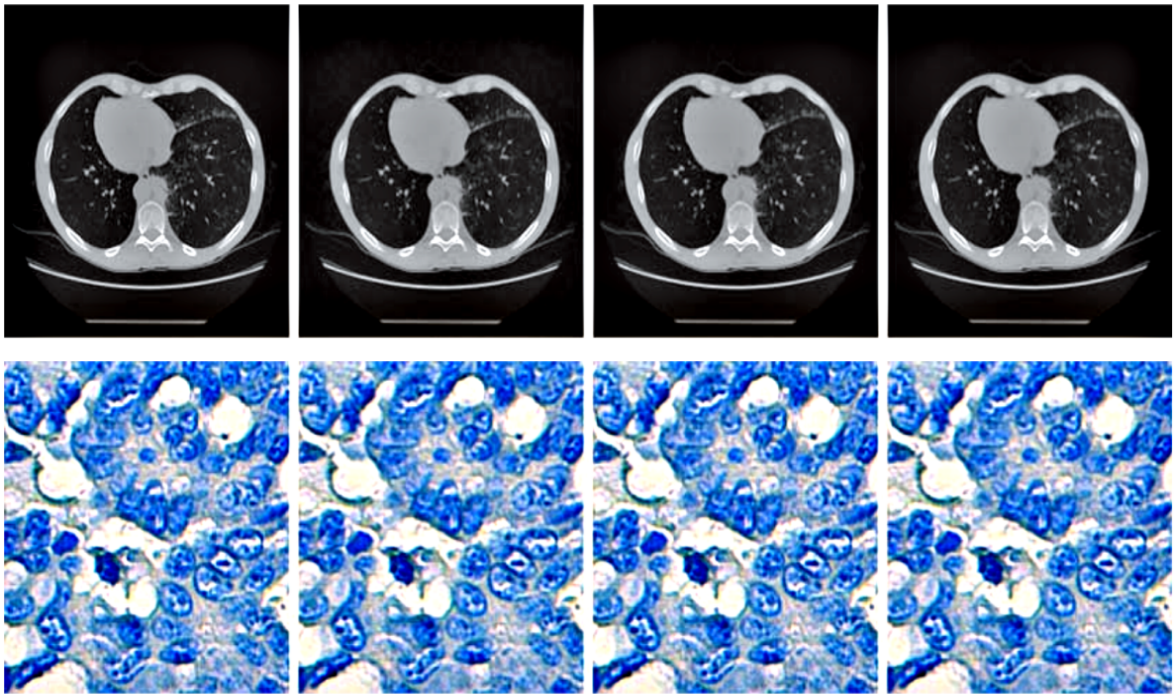


Figure 2. Examples of CT and histological images including original images (column 1) and their adversarial versions (columns 2–4)

As can be seen from the table, the strongest attacks are AutoAttack and Carlini-Wagner attacks, while the Denoiser provides the best defense.

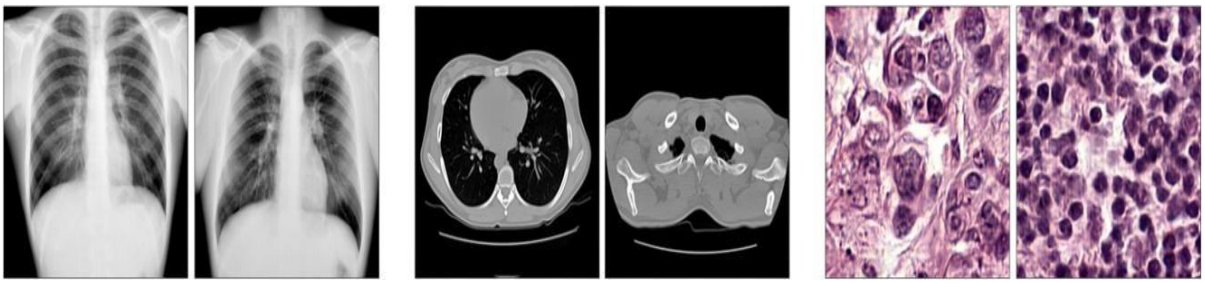
As a general conclusion, it can be noted that adversarial attacks can significantly reduce the accuracy of correct medical image classification, up to completely substituting Normal for Pathology and, what is even more dangerous, suggesting a decision of complete absence of disease when it is actually present. Therefore, the development of methods and software tools for protecting AI systems against adversarial attacks in medical and other critical tasks is, in our view, absolutely necessary.

Security of Patient Personal Data. One way to ensure the security of patient personal data contained in their medical images is to replace real patient images with images generated using modern generative neural network models. Such replacement is possible in various scenarios of development, implementation, and use of decision support information systems in medical diagnostics and allows resolving a number of legal, ethical, and purely technical issues. The most obvious example is the use of realistic artificial images visually indistinguishable from real ones in the preparation of information and advertising materials, development of thematic websites, production of printed and video products, as well as other cases of promoting and implementing relevant subsystems, blocks, and computing services of computer-aided disease diagnostic systems. Research and development in the field of generative models has been conducted by the authors of this work since 2017. As generative models, DC-GAN type models [14] consisting of a generator and a discriminator competing with each other according to the scheme of mathematical models of antagonistic game theory, as well as generators with progressively increasing image resolution (Pro-GAN) [15] and diffusion models of the UNET + Gaussian denoiser type [16] (see Fig. 3), were used.

To assess the quality of the generated images, a professional test was conducted involving experienced radiologists (5 groups of 1–3 physicians each). The task was to separate a sample of 200 randomly mixed X-ray images, 100 of which were images of real people and the other 100 were generated. The results of the blind test showed that correctly separating these image groups is a fairly difficult task even for professionals.

Table 1. Classification accuracy in % for different attack and defense types

Attack Type	**Defense Type**			
	No Defense	**Adversarial Training**	**Denoiser**	**MagNet**
X-ray images				
No attack	92.8	–	–	–
FGSM gradient attack	60.2	82.8	93.4	73.8
AutoAttack	00.0	89.6	93.6	73.6
Carlini-Wagner attacks	00.0	84.2	92.9	73.9
CT images				
No attack	100.0	–	–	–
FGSM gradient attack	57.9	98.4	100.0	98.8
AutoAttack	00.0	97.5	100.0	99.0
Carlini-Wagner attacks	00.0	98.3	100.0	98.9
Histological images				
No attack	95.8	–	–	–
FGSM gradient attack	17.8	89.4	91.4	71.3
AutoAttack	00.0	93.1	94.2	72.7
Carlini-Wagner attacks	10.2	82.7	92.9	73.0



a) b) c)

Figure 3. Examples of generated images: a) X-ray; b) CT; c) histological.

Another important direction for the use of generative models is their use as an "intelligent" *augmentation* service, i.e., expanding the *quantity* of training images by including not only real but also generated images when there is a shortage. It is known that a lack of training images is one of the central practical problems in developing AI systems.

Conducted research and experiments have shown that even a complete replacement of real images with generated ones leads to a drop in classification accuracy; the decrease caused by such a replacement of training data ranges from 2.2% to 3.5% for deep learning models and from 5.5% to 13.25% for conventional methods such as LBP + Random Forests.

Federated Learning of Neural Networks. Another way to solve the problem of a limited set of training images is to combine image databases from multiple centers. However, combining and sharing medical images belonging to different institutions can be problematic for legal, social, and ethical reasons.

These issues can be overcome by using the so-called Federated Learning technology [17]. In this work, we present the results of an experimental study of the effectiveness of the FedAVG variant of federated learning for DC-GAN type generative neural networks.

The computational experiments assessed the joint federated training of generative models designed to synthesize images at two spatial resolutions: 64×64 and 128×128 pixels. The resulting patterns were identical for both resolutions. The analysis covered three types of medical images: chest X-rays, computed tomography slices, and histological specimens.

The convergence dynamics of the federated learning process were examined under three distinct aggregation regimes performed by the server. Weight averaging of the neural network was carried out after clients had completed 1, 3, or 5 training epochs on their private data. The datasets employed

differed in their degree of heterogeneity. The most homogeneous was the set of chest X-ray images (10,000 healthy individuals).

To a non-specialist, the majority of these images appear nearly identical, differing only in subtle visual features related to age, sex, and sometimes overall body constitution. When training on small images (64×64 pixels), the federated approach demonstrated stable and consistent convergence across all three aggregation regimes. However, with larger images (128×128 pixels), the curves reflecting the consistency of the results (as measured by the FID score [18]) became noisier.

When aggregation was performed only after every 5 epochs, the training process failed to converge altogether. The learning trajectory for computed tomography images broadly resembled that of X-ray images, but the training was considerably less stable in all cases. This was evident from the behavior of the curves both for the individual clients' training processes and for the server's aggregated results. Histological images are characterized by high morphological variability. In a typical whole-slide biopsy specimen, the same cellular patterns rarely repeat.

For this image type, the FedAVG approach behaved identically at both resolutions: throughout most training cycles, the model that aggregated the results from two clients consistently performed worse than the models trained locally by each client. In other words, the Fréchet distance between real and generated images remained large, and the use of federated learning technology led to a high degree of dissimilarity between synthetic and real images. Consequently, under these specific conditions, federated training is not advisable.

A key finding is that the potential utility of the horizontal FedAVG approach strongly depends on the natural homogeneity of the image datasets involved. Among the three examined modalities, chest X-rays showed the greatest promise for federated training, whereas standard hematoxylin-eosin-stained histological images proved unsuitable.

Two-dimensional computed tomography slices occupied an intermediate position, exhibiting unstable behavior during training. Furthermore, the aggregation period on the federated server should be kept reasonably short; it is recommended to perform aggregation after every 1 to 3 local training epochs completed by the clients on their respective image datasets.

Conclusion. This paper has systematically examined three critical safety aspects of artificial intelligence methods in medical diagnostics: vulnerability to adversarial attacks, protection of patient personal data via generative models, and the feasibility of federated learning for collaborative model training without data sharing.

Experimental results on three large-scale medical image datasets (chest X-rays, CT scans, and histology slides) demonstrate that deep neural networks (EfficientNet-B3) are highly susceptible to adversarial attacks. Under white-box conditions, gradient-based FGSM attacks, AutoAttack, and Carlini-Wagner attacks caused dramatic drops in classification accuracy, in many cases reducing it to 0% for AutoAttack and Carlini-Wagner attacks on X-ray and CT images without defense. Even more alarmingly, such attacks can induce misclassifications that swap "Normal" and "Pathology" labels, potentially leading physicians to conclude that a disease is absent when it is actually present. Among the three defense strategies evaluated – adversarial training, a high-level task-driven denoiser, and the MagNet auto-encoder – the denoiser consistently provided the best protection, restoring accuracy to 92.9–100% across all attack types and imaging modalities. These findings underscore the absolute necessity of integrating robust defense mechanisms into AI systems intended for critical medical applications. The study on generative models for patient data security explored replacing real patient images with synthetically generated ones (DC-GAN, Pro-GAN, and diffusion models) to circumvent legal, ethical, and technical issues associated with personal data handling. A blind test involving experienced radiologists revealed that distinguishing real from generated chest X-rays is surprisingly difficult, confirming the high realism of the synthetic images. Furthermore, the authors investigated using generative models for data augmentation. Replacing real training images entirely with generated ones led to only a modest decline in classification accuracy (2.2–3.5% for deep learning models and 5.5–13.25% for conventional methods like LBP + Random Forests), suggesting that generative augmentation is a viable solution when real data are scarce. As an alternative

approach to address limited data availability across multiple institutions without sharing sensitive patient information, the FedAVG variant of federated learning was tested on DC-GANs. The experiments involved two partners with equally split datasets, with model aggregation performed after 1, 2, or 5 local training epochs.

The results showed that the success of federated learning heavily depends on the homogeneity of the image sets. Chest X-ray images, being relatively uniform, demonstrated strong potential for federated learning. In contrast, histological images proved unsuitable for the FedAVG approach due to their high heterogeneity. Axial 2D CT slices exhibited intermediate behavior.

Thus, federated learning is not a universal solution; its applicability must be assessed on a per-modality and per-dataset basis.

The study confirms that while AI methods offer immense benefits for medical diagnostics, they introduce novel vulnerabilities and privacy risks that are not adequately addressed by traditional security measures.

Adversarial attacks can completely subvert diagnostic decisions, and effective defenses (such as the high-level denoiser) are essential.

Generative models provide a promising pathway for anonymizing patient data and augmenting limited datasets. Federated learning offers a collaborative framework but requires careful handling of data heterogeneity.

Future work should focus on developing hybrid protection schemes that combine adversarial robustness, generative privacy preservation, and adaptive federated strategies tailored to the specific characteristics of medical imaging data.

Acknowledgments

This work was partly supported by State Program for Scientific Research, project И104.

References

[1] Azad A., Banu A. Publication trends in artificial intelligence conferences: The rise of super prolific authors //arXiv preprint arXiv:2412.07793. – 2024.

[2] Hou Q. et al. A clinical-oriented multi-level contrastive learning method for disease diagnosis in low-quality medical images //International Conference on Medical Image Computing and Computer-Assisted Intervention. – Cham : Springer Nature Switzerland, 2024. – С. 13-23.

[3] Wu J., Xu M. One-prompt to segment all medical images //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2024. – С. 11302-11312.

[4] Queiroz D. et al. Fair foundation models for medical image analysis: Challenges and perspectives //ACM Transactions on Computing for Healthcare. – 2025.

[5] Chang H., Shokri R. On the privacy risks of algorithmic fairness //2021 IEEE European Symposium on Security and Privacy (EuroS&P). – IEEE, 2021. – С. 292-303.

[6] Zhu Y. et al. Privacy-preserving in medical image analysis: A review of methods and applications //International Conference on Parallel and Distributed Computing: Applications and Technologies. – Singapore : Springer Nature Singapore, 2024. – С. 166-178.

[7] Finlayson S. G. et al. Adversarial attacks on medical machine learning //Science. – 2019. – Т. 363. – №. 6433. – С. 1287-1289.

[8] Tan M., Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks //International conference on machine learning. – PMLR, 2019. – С. 6105-6114.

[9] Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.

[10] Croce F., Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks //International conference on machine learning. – PMLR, 2020. – С. 2206-2216.

[11] Carlini N., Wagner D. Towards evaluating the robustness of neural networks //2017 ieee symposium on security and privacy (sp). – Ieee, 2017. – С. 39-57.

[12] Liao F. et al. Defense against adversarial attacks using high-level representation guided denoiser //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 1778-1787.

[13] Meng D., Chen H. Magnet: a two-pronged defense against adversarial examples //Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. – 2017. – С. 135-147.

[14] Radford A., Metz L., Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks //arXiv preprint arXiv:1511.06434. – 2015.

[15] Karras T. et al. Progressive growing of gans for improved quality, stability, and variation //arXiv preprint arXiv:1710.10196. – 2017.

[16] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – Т. 33. – С. 6840-6851.

[17] McMahan B. et al. Communication-efficient learning of deep networks from decentralized data //Artificial intelligence and statistics. – PMLR, 2017. – С. 1273-1282.

[18] Heusel M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium //Advances in neural information processing systems. – 2017. – Т. 30.

Author's contribution

Kovalev Vassili – problem statement, development of key approaches and algorithms.

Snezhko Eduard – problem statement, selection and testing of neural network architectures.

Karpenka Dmitry – development of software tools and conducting computational experiments on federated learning.

Varvashevich Angelina – development of software tools and conducting computational experiments on Adversarial Attacks.

ОЦЕНКА УЯЗВИМОСТЕЙ И МЕТОДОВ ЗАЩИТЫ ИИ В КОМПЬЮТЕРИЗИРОВАННОЙ ДИАГНОСТИКЕ ПО ИЗОБРАЖЕНИЯМ

В.А. Ковалев
Ведущий научный сотрудник, лаборатория анализа биомедицинских изображений, Объединенный Институт Проблем Информатики Национальной Академии Наук Беларуси, кандидат технических наук
vassili.kovalev@gmail.com

Э.В. Снежко
Заведующий лаборатории анализа биомедицинских изображений, Объединенный Институт Проблем Информатики Национальной Академии Наук Беларуси, кандидат технических наук
eduard.snezhko@gmail.com

Д.С. Карпенко
Младший научный сотрудник, стажер, лаборатория анализа биомедицинских изображений, Объединенный Институт Проблем Информатики Национальной Академии Наук Беларуси
karpenko.dima.s11@gmail.com

А.Г. Варвашевич
Младший научный сотрудник
varvashevichangelina@gmail.com

Аннотация. В данной статье рассматриваются три аспекта безопасности ИИ при анализе медицинских изображений: уязвимость к состязательным атакам, защита данных пациентов с помощью генеративных моделей и федеративное обучение без обмена данными. Эксперименты на более чем 260 000 рентгеновских снимков грудной клетки, КТ-срезов и гистологических препаратов показывают, что атаки «белого ящика» (FGSM, AutoAttack, Карлини – Вагнера) могут снизить точность классификации до 0% при отсутствии защиты. Среди трёх стратегий защиты Denoiser, управляемый высокоуровневыми данными задачи, оказался наиболее эффективным, восстанавливая точность до 100%. Генеративные модели (DC-GAN, Pro-GAN, диффузионные модели) создавали реалистичные синтетические изображения; их использование для аугментации данных приводило лишь к незначительному снижению точности (2,2–3,5% для глубокого обучения). Федеративное обучение (FedAVG) оказалось успешным только для однородных наборов данных (рентгеновские снимки грудной клетки), но неэффективным для гистологических изображений. В статье делается вывод, что состязательные атаки представляют собой критическую угрозу, генеративные модели обеспечивают аугментацию с сохранением конфиденциальности, а федеративное обучение требует тщательной адаптации к конкретному типу изображений.

Ключевые слова: безопасность ИИ, медицинские изображения, атаки на нейронные сети, глубокое обучение, генеративные модели, федеративное обучение, конфиденциальность данных.