Metaheuristic Optimization Review

♦ ISSN: 2692-4048 Vol: 04, No. 01, 2025 ♦ PP. 05–08

Article DOI: https://doi.org/10.54216/MOR.040102

REVIEW ARTICLE

A Review of Adversarial Deep Learning Models in Neuroscience Research and Clinical Practice

Khaled Sh. Gaber[1,*] **Ehsan Khodadadi**[2]

[1] Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA

[2] Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, AR 72701, USA

Emails: khsherif@jcsis.org · Ehsank@uark.edu

Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025 ⋆ Corresponding author

ABSTRACT

Adversarial deep learning has become one of the key research focus areas in neuroscience, revealing both opportunities and drawbacks in the use of deep learning models for neuroimaging and diagnostic tasks. This review examines model weaknesses under adversarial attacks, which can severely affect diagnosis and patient care. Slight disturbances in EEG signals, for example, may confuse deep learning algorithms used for epilepsy identification and lead to serious diagnostic mistakes. In addition, generative adversarial networks (GANs) play a dual role: they generate realistic neuroimaging data that can improve diagnostic processes while also producing adversarial images that expose the deficits of current models. This duality highlights the need to defend models against such risks and to employ adversarial training and bio-mimetic resilient neural network techniques. These discoveries reveal the necessity of improving safety in the clinical use of deep learning techniques. By addressing these weaknesses, this review aims to improve diagnostic systems and expand understanding of how adversarial deep learning may affect health, well-being, and patient safety in neuroscience.

Keywords: Adversarial deep learning ▪ Neuroimaging ▪ Diagnostic accuracy ▪ Generative adversarial networks ▪ Model robustness ▪ Patient safety

1. INTRODUCTION

Adversarial deep learning is an evolving area within machine learning, offering significant potential for neuroscience, particularly in neuroimaging and diagnostic processes. This field focuses on designing models that can improve data interpretation while simultaneously identifying their weaknesses, specifically with respect to adversarial attacks. Neuroscience applications require precise and reliable models, and adversarial deep learning helps refine them by simulating conditions that test their robustness [1].

In neuroscience, adversarial attacks—deliberate slight disturbances in data—can cause model misinterpretations with critical implications for patient outcomes. Slight modifications in EEG signals can disrupt models for epilepsy detection, potentially leading to severe diagnostic errors. These challenges underscore the necessity of safeguarding deep learning models against such risks to ensure they serve effectively in clinical practice [2].

The vulnerabilities of deep learning models to adversarial attacks highlight the risk they pose in diagnostic contexts. Misclassifications can result from minimal data alterations, which, in neuroscience, can lead to incorrect diagnoses and compromised patient care. Addressing these vulnerabilities is essential because even minor inaccuracies in neurological diagnosis can have profound implications for treatment and patient safety [3].

GANs serve as a data generation tool for neuroscience re-

search and can also be used as quantitative models to explain findings. They can generate accurate neuroimaging data for diagnosis and model improvement. At the same time, adversarial images obtained from GANs indicate current model shortcomings and the need to enhance stability and robustness, which are decisive factors for applying AI models in clinical practice [4].

Recent progress has introduced bio-mimetic concepts to enhance deep learning frameworks. Scientists are leveraging adversarial disturbances to train robust neural networks capable of withstanding perturbing noise. This approach has the potential to bridge the performance gap between simulated and real-world environments, a significant advancement in deep learning [5].

Applying adversarial deep learning in clinical activities brings security to the forefront. As these models become integral to patient care, protection from interference by malicious sources becomes paramount. Robust defense mechanisms ensure patient data security and maintain data accuracy, both of which are critical in healthcare [6]. Diagnostic development would encourage confidence in application-based healthcare systems.

Another issue concerns the values involved in employing adversarial deep learning models in healthcare. Organizational concerns include accountability, particularly when model defects contribute to diagnostic errors. Appreciating and managing the risks associated with these models is necessary for proper implementation in neuroscience-related healthcare.

This review paper consolidates knowledge of adversarial deep learning applications in neuroscience and highlights existing challenges and advancements. By examining the dual roles of these models in enhancing diagnostic accuracy and exposing vulnerabilities, the paper emphasizes the importance of secure and resilient models and the potential of adversarial deep learning to advance neuroscience research.

Ultimately, this paper underscores the importance of ongoing research in adversarial deep learning for neuroscience. Enhancing model robustness and ethical deployment are technological goals essential for protecting patient welfare. Through an in-depth exploration of adversarial challenges and innovations, this review supports future advancements in neuroscience AI models and promotes safety and efficacy in clinical practice.

2. LITERATURE REVIEW

Deep learning has become a robust approach in neuroscience, especially in neuroimaging and diagnosis. However, many examinations show that these models are vulnerable to adversarial attacks, limiting their practical use. This literature review examines adversarial deep learning, vulnerabilities in current models, robust engineering techniques, and implications for diagnostic systems.

Adversarial deep learning has emerged as a critical research focus in neuroscience, uncovering potential and limitations in neuroimaging and diagnostic applications. Even minor data disturbances, such as slight changes in EEG signals, can lead to significant errors in deep learning models used for epilepsy detection, raising concerns about diagnostic accuracy and patient safety [7]. GANs play a dual role by creating synthetic neuroimaging data that enhances diagnostic training while simultaneously revealing model weaknesses. Adversarial training and bio-inspired resilient networks offer promising approaches to mitigate these vulnerabilities by mimicking the brain's ability to handle noisy inputs. Ethical considerations are also crucial because accountability for diagnostic errors stemming from adversarial weaknesses must be addressed to maintain trust in AI-driven healthcare systems.

Deep learning has achieved impressive performance across various tasks, including medical image processing. Deep neural networks are notably vulnerable to small adversarial perturbations, which can disrupt model accuracy even with minimal image alterations. Anatomical context has been shown to protect deep learning from adversarial perturbations in medical imaging, suggesting that domain knowledge can improve model robustness [8].

Reproducibility is another important requirement in neuroimaging. ClinicaDL provides an open-source deep learning software framework for reproducible neuroimaging processing, highlighting the need for standardized pipelines, transparent evaluation, and repeatable experimental settings in clinical AI research [9].

In epileptic seizure detection, adversarial representation learning has been applied to build robust patient-independent models. Such work addresses the variability between subjects and the sensitivity of EEG-based models to noise and distribution shifts. These approaches are especially important because diagnostic systems must operate reliably across different patients and acquisition conditions [10].

Generative adversarial neural networks and deep learning have shown successful cases and advanced approaches across applied research. These methods increase efficiency through calibrated data samples, algorithmic improvements, and hybrid optimization techniques, underscoring the transformative role of AI in modern research [11].

Data diversity is essential for training effective deep learning models. Medical imaging datasets are often imbalanced due to the rarity of pathological findings, creating substantial challenges for model training. Synthetic MRI images with brain tumors generated by GANs can support data augmentation and anonymization. This improves tumor segmentation performance and may reduce privacy barriers because models trained on synthetic data can achieve comparable segmentation results to those trained on real subject data [12].

GANs have also been combined with light microscopy to enable deep learning super-resolution imaging across large fields of view. By leveraging prior microscopy data in adversarial training, neural networks can reconstruct high-resolution images from single low-resolution measurements. This approach has been validated across sample types, including resolution targets, pathological slides, fluorescence-labeled cells, and deep tissues in transgenic mouse brains [13].

The successful training of deep neural networks is commonly believed to require large volumes of annotated data, which are costly and challenging to acquire in biomedical imaging. GAN-based data augmentation with a coarse-to-fine generator architecture can capture the training data manifold and produce high-quality augmented images. Experiments

on MRI images demonstrated improved Dice coefficients and state-of-the-art segmentation performance in brain tumor segmentation tasks [14].

Functional magnetic resonance imaging (fMRI) is essential for analyzing cognitive brain function. Small sample sizes often lead to overfitting in fMRI classification tasks. Enhanced deep learning GANs using Wasserstein distance and double-class distance constraints have been proposed to augment fMRI functional connectivity data for classifying attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD), improving accuracy over existing classifiers [15].

Deep learning has been extensively explored in brain image analysis for diagnosing neurological conditions such as Alzheimer's disease. Brain status transferring GANs can generate corresponding healthy brain images for patients, enabling the decoding of individualized brain atrophy. The residual between generated and input images quantifies pathological change and supports personalized diagnosis and precision medicine [16].

GANs have also been used to improve segmentation reliability of multi-scanner brain images. Multi-scanner variability can reduce model generalization, but adversarial image harmonization and segmentation strategies can make models more reliable across acquisition protocols and clinical settings [17].

Although deep learning applications provide powerful capabilities for big data analytics, they introduce important challenges related to data scale, privacy, computational resources, and interpretability [18]. These issues are particularly important in neuroimaging, where datasets are high-dimensional and clinical decisions require transparent evidence.

Deep generative-discriminative learning has further supported multimodal representation in imaging genetics by integrating imaging and genetic information. Such multimodal approaches can improve the discovery of disease-relevant patterns, but they also require robust learning systems that resist perturbations and preserve clinical meaning [19].

Understanding adversarial attacks on deep learning-based medical image analysis systems remains essential. Attacks can alter medical images in ways that are nearly imperceptible to humans yet highly disruptive to models. This creates serious risks for diagnostic workflows, automated triage, and clinical decision support [20].

Adversarial learning has also been applied to brain stroke lesion segmentation using consistent perception GANs. These systems demonstrate the value of GAN-based learning for segmentation tasks, but clinical use requires careful validation across patient groups, imaging devices, and pathological variability [21].

Adversarial time-to-event modeling extends adversarial learning into survival analysis and clinical outcome prediction. This direction is relevant to neuroscience because patient prognosis often depends on longitudinal patterns, treatment response, and time-dependent risk [22].

Recent studies on deep learning for neuroimaging continue to highlight applications in brain disease diagnosis and research. These developments show transformative implications for improving diagnostic precision and patient-specific treatment. However, two issues remain open: model robustness and ethical concerns in AI healthcare applications. Further studies should develop deep learning systems that are resistant to adversarial attacks and interpretable enough to improve patient outcomes [23].

3. CONCLUSION

Adversarial deep learning is a significant innovation for neuroscience applications, especially in neuroimaging and diagnostic systems. This review shows both the opportunities and issues that may arise from using such models in clinical practice. Although techniques such as GANs enable the creation of synthetic data that can improve diagnostic accuracy and expose model weaknesses, data perturbations may still lead to diagnostic inaccuracies. Mitigating these risks is critical because seemingly small changes in EEG signals or other neuroimaging data can deceive models and erode patient safety and diagnostic accuracy.

Adversarial training and biology-inspired neural approaches provide potential solutions for improving the robustness of deep learning systems. When adversarial scenarios are used in model training, researchers can develop systems better suited for real-world clinical practice and capable of handling disturbances in data. Bio-mimetic approaches mirror the brain's ability to cope with noisy signals and help bridge the gap between controlled laboratory environments and volatile clinical contexts, increasing the reliability of AI systems used in diagnosis and care.

However, using adversarial deep learning in healthcare still raises ethical and security concerns. Persistent threats of adversarial attacks, even in clinical situations, emphasize the importance of defense mechanisms that protect patient data and diagnostic results. Questions of accountability, especially when adversarial weaknesses cause mistakes, must also be answered to preserve public confidence in AI health systems. Security, transparency, and willingness to share validated methods will be critical as clinical neuroscience applications of adversarial deep learning progress.

The ongoing research and advancement of adversarial deep learning have the potential to bring revolutionary change to neuroscience investigation and practice. By enhancing model robustness, ensuring ethical application, and strengthening patient security, this field can reshape diagnostics and patient care. By balancing innovation with preventive safeguards, adversarial deep learning may pave the way for superior and safer AI solutions in healthcare.

REFERENCES

[1] Y. Guo, J. Zhang, B. Sun, and Y. Wang, "Adversarial deep transfer learning in fault diagnosis: Progress, challenges, and future prospects," *Sensors*, vol. 23, no. 16, p. 7263, 2023.

[2] N. Ghaffari Laleh *et al.*, "Adversarial attacks and adversarial robustness in computational pathology," *Nature Communications*, vol. 13, no. 1, pp. 1–10, 2022.

[3] R. Wang *et al.*, "Applications of generative adversarial networks in neuroimaging and clinical neuroscience," *NeuroImage*, vol. 269, p. 119898, 2023.

[4] G. S. Hong *et al.*, "Overcoming the challenges in the development and implementation of artificial intelligence in radiology: A comprehensive review of solutions beyond supervised learning," *Korean Journal of Radiology*, vol. 24, no. 11, pp. 1061–1080, 2023.

[5] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust ai applications," *Artificial Intelligence Review*, vol. 58, no. 1, pp. 1–107, 2024.

[6] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in healthcare," *Annual Review of Biomedical Data Science*, vol. 4, p. 123, 2021.

[7] H. Arabi, G. Zeng, G. Zheng, and H. Zaidi, "Novel adversarial semantic structure deep learning for mri-guided attenuation correction in brain pet/mri," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 46, no. 13, pp. 2746–2759, 2019.

[8] Y. Li, H. Zhang, C. Bermudez, Y. Chen, B. A. Landman, and Y. Vorobeychik, "Anatomical context protects deep learning from adversarial perturbations in medical imaging," *Neurocomputing*, vol. 379, pp. 370–378, 2020.

[9] E. Thibeau-Sutre *et al.*, "Clinicadl: An open-source deep learning software for reproducible neuroimaging processing," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106818, 2022.

[10] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, "Adversarial representation learning for robust patient-independent epileptic seizure detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2852–2859, 2020.

[11] O. Striuk and Y. Kondratenko, "Generative adversarial neural networks and deep learning: Successful cases and advanced approaches," *International Journal of Computing*, vol. 20, no. 3, pp. 339–349, 2021.

[12] H. C. Shin *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Lecture Notes in Computer Science*, vol. 11037, 2018, pp. 1–11.

[13] Y. Yang *et al.*, "High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network," *Biomedical Optics Express*, vol. 10, no. 3, pp. 1044–1063, 2019.

[14] T. C. W. Mok and A. C. S. Chung, "Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks," in *Lecture Notes in Computer Science*, vol. 11383, 2019, pp. 70–80.

[15] Q. Yao and H. Lu, "Brain functional connectivity augmentation method for mental disease classification with generative adversarial network," in *Lecture Notes in Computer Science*, vol. 11857, 2019, pp. 444–455.

[16] X. Gao, H. Liu, F. Shi, D. Shen, and M. Liu, "Brain status transferring generative adversarial network for decoding individualized atrophy in alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 10, pp. 4961–4970, 2023.

[17] K. Niu *et al.*, "Improving segmentation reliability of multi-scanner brain images using a generative adversarial network," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 3, p. 1775, 2022.

[18] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, pp. 1–21, 2015.

[19] W. Ko, W. Jung, E. Jeon, and H. I. Suk, "A deep generative-discriminative learning for multimodal representation in imaging genetics," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2348–2359, 2022.

[20] X. Ma *et al.*, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2021.

[21] S. Wang, Z. Chen, S. You, B. Wang, Y. Shen, and B. Lei, "Brain stroke lesion segmentation using consistent perception generative adversarial network," *Neural Computing and Applications*, vol. 34, no. 11, pp. 8657–8669, 2022.

[22] P. Chapfuwa *et al.*, "Adversarial time-to-event modeling," in *Proceedings of Machine Learning Research*, 2018. [Online]. Available: https://proceedings.mlr.press/v80/chapfuwa18a.html

[23] V. Sovann and C. Thach, "Deep learning for neuroimaging: Applications in brain disease diagnosis and research," *Asian American Research Letters Journal*, vol. 1, no. 3, 2024. [Online]. Available: https://aarlj.com/index.php/AARLJ/article/view/41