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

Abstract

Adversarial deep learning has, therefore, been tabled as one of the key research focus areas in neurosciences, and both the opportunities and drawbacks for the operation of deep learning models on neuroimaging and diagnostic jobs have been unveiled. This review examines these models' weaknesses from adversarial attacks, which can severely affect diagnosis and patient care. For example, it has been shown that slight disturbances in the level of EEG signals can confuse more profound learning algorithms employed for the identification of epilepsy, which can lead to severe diagnostic mistakes. In addition, GANs have the dual role of generating realistic neuroimaging data that can improve diagnostic processes while at the same time using adversarial images that expose the deficits of current models. This duality highlights the need to securely defend models against such risks and employ adversarial training and bio-mimic-based resilient neural network techniques. The consequence of these discoveries should not be underestimated because they reveal the necessity of showing further safety in using deep learning techniques in clinical practices. In addressing these weaknesses, the principle goal of this research is not only to help improve the diagnostic systems but also to expand the knowledge on how adversarial deep learning might affect the health, well-being and safety of patients 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 regarding 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. For example, 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].

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

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, as even minor inaccuracies in a neurological diagnosis can have profound implications for treatment and patient safety [3].

GANs serve as a data generation tool for neuroscience research and can be used as quantitative models to explain findings. They can generate accurate neuroimaging data that can be used to diagnose and improve the models. At the same time, the adversarial images obtained from the GANs indicate the current shortcomings within the models and the necessity to enhance their stability and robustness, which is the decisive factor 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, the need for protection from interference by malicious sources becomes paramount. A robust defense mechanism ensures patient data security and maintains data accuracy, a critical consideration in healthcare [6].

Diagnostic development would encourage people to have confidence in application-based healthcare systems.

Another issue that should be mentioned regarding values is the employment of adversarial deep learning models in healthcare. Some are organizational concerns, such as accountability, mainly when the model defects contribute to diagnostic errors. Appreciating and managing the risks that come with these models is something that needs to be accomplished in order to achieve proper implementation of these technologies in neuroscience-related healthcare.

This review paper aims to consolidate the knowledge of adversarial deep learning applications in neuroscience and highlight existing challenges and advancements. By examining the dual roles of these models in enhancing diagnostic accuracy and exposing vulnerabilities, the paper not only emphasizes the importance of advancing secure and resilient models but also underscores the potential of adversarial deep learning to inspire a sense of optimism and motivation in the audience.

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 seeks to support future advancements in neuroscience AI models, promoting safety and efficacy in clinical practice.

2. Literature Review

Hence, deep learning has spread over the years as a robust approach, especially in neuroscience, which affects neuroimaging and diagnosis. However, the results of the examinations prove that these models are vulnerable to adversarial attacks and, therefore, their practical usage could be better. This literature review examines adversarial deep learning, its weaknesses in current models, how robust techniques can be engineered, and the implications of using deep learning for diagnostic systems.

Adversarial deep learning has emerged as a critical research focus in neuroscience, uncovering potential and limitations in neuroimaging and diagnostic applications. As discussed in [7], 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. Generative Adversarial Networks (GANs) serve a dual role by creating synthetic neuroimaging data that enhances diagnostic training while simultaneously revealing the weaknesses of current models, emphasizing the importance of advancing model resilience. Moreover, adversarial training and bio-inspired resilient networks offer promising approaches to mitigate these vulnerabilities, mimicking the brain's ability to handle noisy inputs and bolstering the robustness of AI applications in clinical settings. Ethical considerations are also crucial, as accountability for

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

diagnostic errors, stemming from adversarial weaknesses must be addressed to maintain trust in AI-driven healthcare systems. The ongoing research in this field aims to ensure that adversarial deep learning models achieve the necessary security and reliability, ultimately supporting patient well-being and expanding the safe use of AI in neuroscience.

Deep learning has achieved impressive performance across various tasks, including medical image processing. According to the findings in [8], deep neural networks are notably vulnerable to small adversarial perturbations, which can disrupt model accuracy even with minimal image alterations. This research examines the impact of these adversarial perturbations in predicting age from 3D MRI brain images, evaluating both a conventional deep neural network and a hybrid model incorporating anatomical features. Results indicate that subtle, imperceptible noise can introduce substantial errors in age predictions, even when applied across large batches of images using a single perturbation. Notably, the hybrid model demonstrates greater robustness against adversarial perturbations than the conventional model, underscoring a critical limitation in current deep learning techniques within clinical contexts and suggesting that integrating anatomical context may enhance model resilience against adversarial attacks.

Deep learning faces a reproducibility crisis and methodological flaws in neuroimaging studies. It is for this reason that safer tools are needed to guide researchers in avoiding the potential pitfalls that could be behind their work. In the article referred to as [9], the authors propose Clinical DL. This software tool should help deep learning users, especially those who deal with neuroimaging data, avoid such problems as data leakage and increase the model's reproducibility. Clinical DLK works with the BIDS format applied in neuroimaging, making cooperation with different datasets easier. Combined with its companion project, Clinica, it also provides an end-to-end neuroimaging data analysis workflow that starts with raw data download, preprocessing, QC, and modeling. To tackle common issues that many practicing researchers face—formatting of neuroimaging data, contamination in the assessment process, and lack of reproducibility—ClinicaDL envisions enhancing the efficiency and significance of profound learning studies in neuroimaging.

Epilepsy, a chronic neurological disorder affecting about one percent of the global population, is characterized by spontaneous seizures, making reliable detection crucial for patient care. As outlined in [10], most seizure detection methods depend heavily on patient history, which limits their effectiveness in diagnosing new patients. To address this, the study proposes a robust, explainable model for epilepsy detection that distinguishes seizure-specific patterns in EEG signals through adversarial training, minimizing inter-patient noise. The model leverages a deep neural network that processes raw, non-invasive EEG data without manual feature engineering, enhancing efficiency for broader clinical use. Additionally, an attention mechanism is developed to identify the significance of individual EEG channels in diagnosis, improving interpretability for clinical insights. Evaluation of the Temple University Hospital EEG (TUH EEG) dataset shows that this model surpasses current state-of-the-art approaches with low latency and provides fine-grained pathological insights, positioning it as a promising tool for scalable, patient-independent seizure detection.

Cross-domain artificial intelligence (AI) frameworks are essential for accelerating scientific progress by integrating advanced deep learning methodologies across various fields. As discussed in [11], cutting-edge AI approaches provide unprecedented opportunities to retrieve, optimize, and enhance diverse data types, enabling refined performance across applied sciences. Recent advances in generative adversarial networks (GANs) and deep learning significantly improve the quality of graphic samples generated by research tools, affecting fields like observational astronomy, healthcare, and materials science. These advancements can be integrated into a unified academic and technological pipeline, fostering rapid scientific and technological innovation. The study explores successful GAN and deep learning applications across several scientific domains. It evaluates methods to increase efficiency through calibrated data samples, algorithmic improvements, and hybrid optimization techniques, underscoring AI's transformative role in applied research.

Data diversity is essential for training effective deep learning models. However, medical imaging datasets often need to be more balanced due to the rarity of pathological findings, creating substantial challenges for model training. In the study referenced in [12], the authors address this issue by developing a method to generate synthetic MRI images with brain tumors using a generative adversarial network trained on two publicly available brain MRI datasets. This approach yields two main benefits: improved tumor segmentation performance through data augmentation with synthetic images and the potential for anonymization, as models

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

trained on synthetic data achieved comparable tumor segmentation results to those trained on accurate subject data. These findings present a promising solution to two primary obstacles in medical imaging: the limited availability of abnormal cases and the strict regulations on sharing patient data.

A generative adversarial network (GAN) is combined with light microscopy to enable deep learning super-resolution imaging across large fields of view (FOV). As outlined in [13], this method leverages prior microscopy data in adversarial training, allowing the neural network to reconstruct high-resolution images from single low-resolution measurements. The model's effectiveness is validated through various sample types, including resolution targets, human pathological slides, fluorescence-labeled cells, and deep tissues in transgenic mouse brains, imaged using wide-field and light-sheet microscopy techniques. An image-degrading model is also introduced to generate low-resolution training images, eliminating the need for complex image registration in data preparation. Once trained, the network can achieve gigapixel, multi-color reconstructions with enhanced resolution (~1.7 μm) over a large FOV (~95 mm^2) at high speed (within 1 second) without altering existing microscope setups. This GAN-based approach demonstrates the potential of deep learning for efficient, high-resolution imaging in microscopy.

The successful training of deep neural networks is commonly believed to require large volumes of annotated data, which is often costly and challenging to acquire, particularly in biomedical imaging. In the publication [14], the authors address this issue by proposing a novel automatic data augmentation technique using generative adversarial networks (GANs) to create augmented data that enhances model learning from limited annotated samples. This GAN-based approach features a coarse-to-fine generator architecture to capture the training data's manifold and produce high quality, generic augmented images. In experiments on MRI images, this method achieved a 3.5% improvement in the Dice coefficient on the BRATS15 Challenge dataset compared to traditional augmentation methods. Additionally, this augmentation technique boosted a .shared segmentation network's performance, achieving state-of-the-art results on the BRATS15 Challenge, demonstrating the approach's effectiveness in advancing segmentation accuracy with limited annotated data.

Functional magnetic resonance imaging (fMRI) is essential for studying and analyzing cognitive brain function. In the study referenced as [15], the authors address a limitation in fMRI classification research, where small sample sizes often lead to overfitting in classification tasks. To overcome this, the study proposes an enhanced deep learning generative adversarial network (GAN) to augment fMRI functional connectivity data. This GAN utilizes Wasserstein distance and a double-class distance constraint to augment data from subject and control groups, enhancing classifier training. The augmented data was applied to improve classification performance for two brain disorders: attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD). The results demonstrated substantial improvements in classification accuracy compared to existing classifiers, and the proposed GAN model outperformed other standard deep network data generation methods, highlighting its potential to mitigate overfitting and strengthen classification in small fMRI datasets.

Deep learning has been extensively explored in brain image analysis for diagnosing neurological conditions such as Alzheimer's (AD). The analysis conducted in [16] highlights that traditional methods generally use group-wise analysis to build end-to-end models for feature learning, limiting their ability to detect subject-specific pathological changes critical for personalized diagnosis and precision medicine. To address this, the study proposes a novel generative adversarial network, Brunstetters-GAN, designed to generate corresponding healthy brain images for patients, enabling the decoding of individualized brain atrophy. BrainStatTrans-GAN incorporates a generator, discriminator, and status discriminator. Initially, a normative GAN generates healthy brain images based on normal controls. However, this approach alone cannot create healthy versions of diseased images due to the absence of paired healthy and diseased data. The authors solve this by introducing a status discriminator within an adversarial learning framework, enabling the network to produce healthy images from diseased inputs. Calculating the residual between generated and input images can quantify pathological changes in the brain. Additionally, a residual-based multi-level fusion network (RMFN) enhances diagnostic precision. Experimental evaluations on T1-weighted MRI data from 1,739 subjects across three datasets demonstrate the method's capacity to model individualized brain atrophy, thereby advancing disease diagnosis and interpretation in clinical practice.

The spatially localized atlas network tiles-27 (SLANT-27) deep learning model was employed to develop an automatic segmentation module using a large, multi-center dataset of 1,917 three-dimensional (3D) T1-weighted MR images. As outlined in [17], this led to the creation of the Qbrain framework, which integrates

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

a generative adversarial network (GAN) image transfer module with the SLANT-27 segmentation module for enhanced image processing. On the same day, a separate 3D T1-weight MRI dataset, consisting of 48 participants scanned across three different MRI scanners (1.5T Siemens Avanta, 3T Siemens Trio Tim, and 3T Philips Ingenia), was used to train and validate Brains. Comparative analysis was conducted on volumetric T1-weighted images processed with Qbrain, SLANT-27, and FreeSurfer (FS) to evaluate segmentation consistency across scanners. The reliability of automatic segmentation was assessed based on test-retest variability (TRV), offering insights into the robustness of the Brains model for consistent multi-scanner segmentation.

Deep learning models typically require large datasets to extract complex patterns effectively, but in brain disease research, omics data often need more patient samples, creating challenges for reliable biomarker prioritization. In the research presented in [18], the authors address this limitation by developing a generative adversarial network (GAN) model to improve disease gene prediction with RNA-seq data. The model integrates a denoising auto-encoder (DAE) as the generator and a multilayer perceptron (MLP) as the discriminator, with prediction residual errors backpropagated to refine the DAE's probability distribution. This GAN-based framework successfully generated samples with similar distributions to the original dataset, enhancing prediction accuracy and robustness. The results improved the identification of disease genes beyond existing approaches. They identified new disease-related genes and pathways in the brain, offering valuable insights into molecular mechanisms underlying brain disease phenotypes.

Imaging genetics, a rapidly advancing field within medical imaging, focuses on uncovering the relationships between neuroimaging and genetic data. As outlined in [19], while deep learning has been integrated into imaging genetics, current methods face limitations, including simplistic approaches for joint learning of phenotypic and genotypic features, limited extension to biomedical applications such as brain disease diagnosis, and inadequate data analysis from scientific perspectives. To address these gaps, the study proposes a deep learning framework capable of effectively representing neuroimaging and genetic data, achieving state-of-the-art results in diagnosing Alzheimer's disease and identifying mild cognitive impairment. Unlike existing approaches, this framework nonlinearly learns imaging-genetic associations without relying on prior neuroscientific assumptions, enhancing its applicability and interpretive power. Experimental validation on a public dataset demonstrated the framework is potential to yield new insights and perspectives in deep learning-based imaging genetics.

Deep neural networks (DNNs) have gained traction in medical image analysis, especially for cancer diagnosis and lesion detection. In the study referenced as [20], the authors examine a critical vulnerability of these systems: susceptibility to adversarial attacks, where subtle, imperceptible modifications can significantly alter model outcomes. This vulnerability raises substantial safety concerns for DNN deployment in clinical environments. The research finds that DNN models in medical imaging are, in fact, more susceptible to adversarial attacks than models used for natural images. However, it also reveals an important insight—medical adversarial attacks can be detected with high accuracy, achieving over 98% detection AUC using simple detectors. This detectability is attributed to intrinsic feature differences between typical and adversarial examples in medical images. This discovery could inform the development of more secure and interpretable deep-learning models for medical applications.

State-of-the-art deep learning methods have achieved remarkable success in segmentation tasks. However, they require large volumes of manually labeled data, which can be costly and time-intensive to collect. In the article denoted as [21], the authors propose a novel consistent perception generative adversarial network (CPGAN) designed for semi-supervised stroke lesion segmentation, aiming to reduce dependency on fully labeled datasets. The CPGAN model incorporates a similarity connection module (SCM) that captures multi-scale feature information, selectively aggregating features across positions using a weighted sum for enhanced segmentation accuracy. A consistent perception strategy is also implemented to improve stroke lesion predictions on unlabeled data. The model also includes an assistant network to guide the discriminator in learning meaningful feature representations, which can diminish during training. The assistant network and discriminator collaborate to distinguish between natural and synthetic segmentation outputs. Evaluated on the Anatomical Tracings of Lesions after Stroke (ATLAS) dataset, CPGAN demonstrated superior segmentation performance, surpassing fully supervised approaches using only two-fifths of labeled samples, making it highly effective for semi-supervised segmentation tasks.

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

Modern health data science applications benefit from extensive molecular and electronic health data, enabling machine learning to develop statistical models that enhance clinical practice. The research presented in [22] addresses a central challenge in time-to-event analysis, a critical statistical model in health data science, by introducing a deep-network-based approach utilizing adversarial learning for nonparametric estimation of event-time distributions. This approach includes a principled cost function that effectively incorporates information from censored events beyond the observation period. Distinctly, the model prioritizes estimating time-to-event distributions rather than simple time ordering, offering a nuanced perspective on event prediction. Validated on benchmark and real-world datasets, the proposed model demonstrates substantial performance improvements over parametric alternatives, supporting its application in complex clinical data scenarios.

Neuroimaging methods have revolutionized how brain structure and function are assessed, especially with new developments in deep learning that have enhanced diagnostic efficacy, speed and precision in neuroimaging. The authors of [23] provide revelation into the uses of deep learning in neuroimaging, emphasizing the diagnosis of brain diseases and research enhancement. The paper reviews the limitations and obstacles associated with using deep learning for neuroimaging and presents avenues of progress. Moreover, it outlines future directions for improving the impact and utility of deep learning to support neuroimaging applications in clinical and investigational settings.

Table 1 provides an extended description of the studies reviewed, including the study's focus, research methods, and significant outcomes. In general, the table shows the vast potential of adversarial deep learning for the broad realm of neuroscience, including diagnostics and data quality and understanding functionality and disease. The combination of all these studies strongly indicates that adversarial deep learning has the potential to solve some of the most pressing issues in neuroimaging and neuroinformatics.

Table 1: Summary of Literature Review

Study	Key Focus	Methodology	Key Findings
[7]	Adversarial Attacks on Deep Learning for Epilepsy Detection	Analysis of EEG signal perturbations	Minor data disturbances can significantly affect model performance, highlighting the need for robust models.
[8]	Adversarial Perturbations in Age Prediction from 3D MRI Brain Images	Evaluation of deep neural network and hybrid model	Subtle adversarial noise can lead to substantial errors in age prediction, emphasizing the importance of model robustness.
[9]	Clinical DL: A Tool for Reproducible and Reliable Deep Learning in Neuroimaging	Development of a software tool for neuroimaging data analysis	Clinical DL aims to address data leakage and reproducibility issues in neuroimaging studies.
[10]	Robust and Explainable Model for Epilepsy Detection	Adversarial training and attention mechanism for EEG signal analysis	The proposed model achieves state-of-the-art performance in seizure detection with low latency and provides interpretable insights.
[11]	Cross-Domain AI Frameworks for Accelerating Scientific Progress	Integration of deep learning and GANs across various fields	AI can significantly enhance the quality of generated data and accelerate scientific research.
[12]	Generating Synthetic MRI Images with Brain Tumors Using GANs	Data augmentation for improving tumor segmentation	Synthetic data generation can improve model performance and address data scarcity in medical imaging.
[13]	Deep Learning Super-Resolution Imaging Using GANs and Light Microscopy	Enhancing image resolution with GANs	GANs can be used to reconstruct high-resolution images from low-resolution microscopy data.

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

[14]	Automatic Data Augmentation Using GANs for Biomedical Image Analysis	Data augmentation for enhancing model performance	GAN-based data augmentation can improve the performance of deep learning models with limited annotated data.
[15]	Enhancing fMRI Classification with GAN-Generated Data	Data augmentation for improving classification accuracy	GAN-generated fMRI data can improve classification performance in small datasets.
[16]	Personalized Brain Atrophy Modeling with GANs	Modeling individual brain changes for Alzheimer's disease diagnosis	GAN-based methods can generate personalized healthy brain images to quantify pathological changes.
[17]	Qbrain: A Framework for Multi-Scanner Segmentation Consistency	Integrating GANs and deep learning for image processing	Qbrain improves image segmentation consistency across different MRI scanners.
[18]	Improving Disease Gene Prediction with GANs and RNA-seq Data	Data augmentation for gene prediction	GAN-based data augmentation can enhance the identification of disease-related genes and pathways.
[19]	Deep Learning Framework for Imaging Genetics	Integrating neuroimaging and genetic data for disease diagnosis	Nonlinear learning of imaging-genetic associations can improve disease diagnosis.
[20]	Adversarial Attacks on Deep Learning Models for Medical Image Analysis	Evaluating the vulnerability of medical image analysis models	Medical image analysis models are susceptible to adversarial attacks, but these attacks can be detected with high accuracy.
[21]	Consistent Perception GAN for Semi-Supervised Stroke Lesion Segmentation	Semi-supervised learning for medical image segmentation	CPGAN achieves state-of-the-art performance with limited labeled data.
[22]	Deep Learning for Nonparametric Time-to-Event Analysis in Health Data Science	Nonparametric estimation of event-time distributions	Deep learning can improve the accuracy of time-to-event analysis in complex clinical data.
[23]	Deep Learning in Neuroimaging: Opportunities and Challenges	Review of deep learning applications in neuroimaging	Deep learning has the potential to revolutionize neuroimaging, but challenges such as data quality and interpretability need to be addressed.

Combining adversarial deep learning and neuroscience has enormously transformative implications for improving diagnostic precision and patient-specific treatment. However, two points remain open: model robustness and ethical concerns in AI healthcare applications. Further studies should be conducted to develop deep learning that is resistant to antagonistic attacks and brilliant for human understanding to improve patient health outcomes.

3. Conclusion

Consequently, adversarial deep learning is a significant innovation for neuroscience applications, especially in neuroimaging and diagnostic systems. This review also shows the opportunities and issues that may arise from using such models in clinical practice. Despite techniques like GANs allowing for the creation of synthetic data that, when integrated with diagnostics improves diagnostic accuracy and indicates model weaknesses, there are inherent dangers of data perturbations that lead to diagnosis inaccuracy. Mitigating these risks is critical because seemingly small changes in the patient's EEG signal or other neuroimaging data can deceive the models or erode the patient's safety and diagnostic accuracy.

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

Additionally, learning adversarial training and biology-inspired neural approaches provide potential solutions to the robustness of deep learning systems. In addition, exploring adversarial training and bio-inspired neural network methods provides potential solutions for improving the resilience of deep learning systems. When adversarial scenarios are used in model training, the researchers develop systems well suited for real-world clinical practice in handling disturbances in the data. Understanding the brain's mechanisms to address noisy signals is possible with the help of such bio-mimetic approaches mirror the ability to cope with the noise; it helps to traverse from the highly controlled environment of laboratories to a more volatile clinical context, increasing the reliability of the AI systems to diagnose the patient and take care about them.

However, using adversarial deep learning in healthcare still raises ethical and security concerns. The constant threats of adversarial attacks, even in clinical situations, raise the importance of defense against threats that continue to compromise the patient's data and diagnostic results. Questions of accountability, especially concerning problems such as adversarial weaknesses that caused mistakes, must also be answered to restore public confidence in AI health systems. A concern with security and a willingness to share will be critical as clinical neuroscience and adversarial profound learning applications progress in clinical practice.

The ongoing research and advancement of adversarial deep learning, as detailed in this study, have the potential to usher in a revolutionary change in neuroscience investigation and practice. As a field that strives to enhance model robustness, ensure ethical application, and bolster patient security, this development is poised to reshape the approaches to diagnostics and patient care. By striking a balance between innovation and preventive measures, adversarial deep learning could pave the way for superior and safer AI solutions in healthcare.

Reference

[1] Y. Guo, J. Zhang, B. Sun, and Y. Wang, "Adversarial Deep Transfer Learning in Fault Diagnosis: Progress, Challenges, and Future Prospects," *Sensors (Basel)*, vol. 23, no. 16, p. 7263, Aug. 2023, doi: 10.3390/S23167263.

[2] N. Ghaffari Laleh et al., "Adversarial attacks and adversarial robustness in computational pathology," *Nature Communications*, vol. 13, no. 1, pp. 1–10, Sep. 2022, doi: 10.1038/s41467-022-33266-0.

[3] R. Wang et al., "Applications of generative adversarial networks in neuroimaging and clinical neuroscience," *Neuroimage*, vol. 269, Apr. 2023, doi: 10.1016/J.NEUROIMAGE.2023.119898.

[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 J Radiol*, vol. 24, no. 11, pp. 1061–1080, Nov. 2023, doi: 10.3348/KJR.2023.0393.

[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, Nov. 2024, doi: 10.1007/S10462-024-11005-9.

[6] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical Machine Learning in Healthcare," *Annu Rev Biomed Data Sci*, vol. 4, p. 123, May 2021, doi: 10.1146/ANNUREV-BIODATASCI-092820-114757.

[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," *Eur J Nucl Med Mol Imaging*, vol. 46, no. 13, pp. 2746–2759, Dec. 2019, doi: 10.1007/S00259-019-04380-X/TABLES/4.

[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, Feb. 2020, doi: 10.1016/J.NEUCOM.2019.10.085.

[9] E. Thibeau-Sutre et al., "ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing," *Comput Methods Programs Biomed*, vol. 220, p. 106818, Jun. 2022, doi: 10.1016/J.CMPB.2022.106818.

[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 J Biomed Health Inform*, vol. 24, no. 10, pp. 2852–2859, Oct. 2020, doi: 10.1109/JBHI.2020.2971610.

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025

[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, doi: 10.47839/ijc.20.3.2278.

[12] H. C. Shin et al., "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," *Lecture Notes in Computer Science*, vol. 11037, pp. 1–11, 2018, doi: 10.1007/978-3-030-00536-8_1/TABLES/1.

[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, Mar. 2019, doi: 10.1364/BOE.10.001044.

[14] T. C. W. Mok and A. C. S. Chung, "Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks," *Lecture Notes in Computer Science*, vol. 11383, pp. 70–80, 2019, doi: 10.1007/978-3-030-11723-8_7/TABLES/3.

[15] Q. Yao and H. Lu, "Brain functional connectivity augmentation method for mental disease classification with generative adversarial network," *Lecture Notes in Computer Science*, vol. 11857, pp. 444–455, 2019, doi: 10.1007/978-3-030-31654-9_38/TABLES/4.

[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 J Biomed Health Inform*, vol. 27, no. 10, pp. 4961–4970, Oct. 2023, doi: 10.1109/JBHI.2023.3304388.

[17] K. Niu et al., "Improving segmentation reliability of multi-scanner brain images using a generative adversarial network," *Quant Imaging Med Surg*, vol. 12, no. 3, p. 1775, Mar. 2022, doi: 10.21037/QIMS-21-653.

[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," *J Big Data*, vol. 2, no. 1, pp. 1–21, Dec. 2015, doi: 10.1186/S40537-014-0007-7/METRICS.

[19] W. Ko, W. Jung, E. Jeon, and H. Il Suk, "A Deep Generative-Discriminative Learning for Multimodal Representation in Imaging Genetics," *IEEE Trans Med Imaging*, vol. 41, no. 9, pp. 2348–2359, Sep. 2022, doi: 10.1109/TMI.2022.3162870.

[20] X. Ma et al., "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognit*, vol. 110, p. 107332, Feb. 2021, doi: 10.1016/J.PATCOG.2020.107332.

[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 Comput Appl*, vol. 34, no. 11, pp. 8657–8669, Jun. 2022, doi: 10.1007/S00521-021-06816-8/METRICS.

[22] P. Chapfuwa et al., "Adversarial Time-to-Event Modeling," *PMLR*, Jul. 03, 2018. Accessed: Nov. 12, 2024. [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, Apr. 2024. Accessed: Nov. 12, 2024. [Online]. Available: https://aarlj.com/index.php/AARLJ/article/view/41

DOI: https://doi.org/10.54216/MOR.040102
Received: January 27, 2025 Revised: March 03, 2025 Accepted: May 05, 2025