IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Check for updates

Mitigating adversarial threats in deep learning models trained on sensitive imaging and sequencing datasets within hospital infrastructures

Nnamdi Rex Onwubuche *

Saunders College of Business, Rochester Institute of Technology, USA.

International Journal of Science and Research Archive, 2025, 16(01), 1146-1167

Publication history: Received on 02 June 2025; revised on 13 July 2025; accepted on 15 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2128

Abstract

As deep learning continues to transform clinical diagnostics, models trained on sensitive imaging and sequencing datasets are increasingly deployed within hospital infrastructures for tasks such as tumor classification, variant calling, and disease risk prediction. While these models offer remarkable accuracy and efficiency, they also present new vulnerabilities to adversarial threats maliciously crafted inputs designed to deceive AI systems without altering visual or genomic content perceptibly. Such attacks can compromise diagnostic reliability, patient safety, and institutional trust, particularly when targeting critical applications involving radiology scans or genetic data. This paper investigates strategies for mitigating adversarial threats in deep learning models operating within hospital ecosystems. We explore how attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and adversarial patching exploit model interpretability gaps and high-dimensional data sparsity in medical domains. Emphasis is placed on the unique risks posed to models trained on radiological images (e.g., CT, MRI) and sequencing outputs (e.g., variant allele frequencies, expression matrices) that contain highly sensitive and potentially re-identifiable patient information. We present a multi-tiered defense framework incorporating adversarial training, input preprocessing techniques, certified robustness estimators, and gradient masking to strengthen model resilience. Additionally, we introduce a hospital-specific deployment architecture that includes real-time adversarial input detection using AI-enhanced monitoring agents and edge-layer validation. This design ensures localized protection while minimizing latency in high-throughput clinical workflows. By focusing on healthcare-specific deep learning vulnerabilities and aligning with clinical data governance standards, this research contributes a secure deployment pathway for trustworthy AI applications in precision medicine and hospital cybersecurity.

Keywords: Adversarial Attacks; Deep Learning Security; Medical Imaging; Genomic Sequencing; Clinical AI; Hospital Cybersecurity

1. Introduction

1.1. The Rise of Deep Learning in Healthcare

The adoption of deep learning (DL) in healthcare has transformed diagnostic and prognostic workflows, especially in medical imaging and genomic sequencing. In radiology, convolutional neural networks (CNNs) have demonstrated expert-level performance in detecting abnormalities from MRI, CT, and PET scans by learning hierarchical features that exceed traditional rule-based models [1]. These systems can identify pathologies such as tumors, hemorrhages, and ischemic lesions with increasing precision, supporting early diagnosis and reducing inter-observer variability [2].

Similarly, DL models have gained traction in the analysis of high-throughput sequencing data, including RNA sequencing (RNA-seq) and DNA methylation profiling. Recurrent and attention-based architectures have enabled the classification

[*] Corresponding author: Nnamdi Rex Onwubuche

of disease subtypes, prediction of treatment outcomes, and identification of biomarkers from complex omics datasets [3]. These methods help translate molecular signals into actionable clinical insights, especially in oncology and rare genetic disorders.

The integration of DL into clinical workflows is supported by the increasing availability of labeled datasets, computational infrastructure, and real-time feedback from electronic health records [4]. Multimodal DL models are now being trained to simultaneously process imaging, genetic, and clinical metadata to improve the accuracy and robustness of predictive models in precision medicine.

As hospitals move toward AI-assisted decision-making, ensuring the security and reliability of DL systems becomes paramount. The emergence of adversarial threats poses new risks to these technologies, particularly as they transition from experimental environments into hospital-scale deployment. These vulnerabilities demand urgent attention to preserve clinical trust, patient safety, and regulatory compliance [5].

The workflow in Figure 1 illustrates the standard DL pipeline in hospitals, highlighting the critical points where adversarial threats may be injected to compromise diagnostic outputs.

1.2. Emerging Adversarial Threats in Clinical AI Models

Despite their success, deep learning models are highly vulnerable to adversarial examples intentionally crafted inputs with subtle perturbations that can drastically alter model predictions without being perceptible to human observers [6]. In a clinical setting, these attacks could cause a DL system to misclassify a malignant tumor as benign or vice versa, undermining diagnostic accuracy and patient safety.

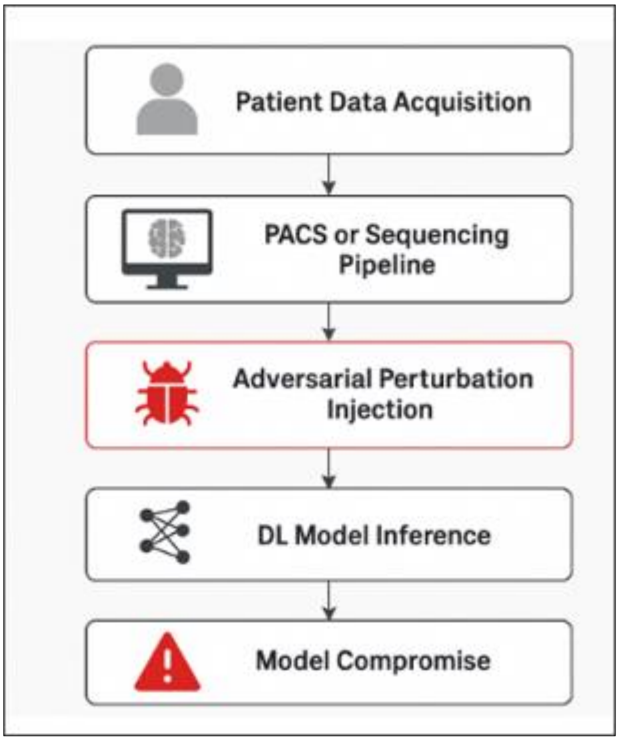
Patient Data Acquisition
PACS or Sequencing Pipeline
Adversarial Perturbation Injection
DL Model Inference
Model Compromise

Figure 1 Highlights a typical DL system pipeline in healthcare, demonstrating how adversarial examples can be injected between acquisition and inference to manipulate model output in real time

Adversarial attacks exploit the non-linear decision boundaries of deep models. Even state-of-the-art CNNs trained on large radiological datasets can be tricked into producing incorrect classifications with imperceptible noise additions, particularly in grayscale modalities such as chest X-rays or brain MRIs [7]. Similarly, models trained on RNA-seq data or DNA methylation signatures can be manipulated to yield false biomarker profiles through adversarial alterations in feature expression vectors [8].

These attacks are especially concerning in hospital-based environments, where DL systems interact directly with clinical workflows. Attackers may exploit insecure data pipelines, external interfaces, or even poisoned training datasets to introduce adversarial inputs at various stages of the decision-making process [9].

Given the black-box nature of many deployed DL models, adversarial perturbations may go undetected until they result in harmful clinical decisions. These raises pressing concerns regarding AI model certification, validation, and the development of robust defense mechanisms.

Scope and Objectives

This article investigates the emerging threat of adversarial attacks against deep learning models deployed in hospital settings, particularly those used for medical imaging and genomic sequencing. Given the growing reliance on AI for high-stakes decisions such as tumor detection, disease staging, and genetic risk prediction ensuring the security and integrity of these models is critical [10].

The objective is to examine how adversarial perturbations can be introduced into hospital workflows, how they affect diagnostic outcomes, and what defense mechanisms can be implemented at the system and model level to mitigate them. We focus on attacks targeting CNNs in imaging and neural networks processing sequencing data, evaluating their impact, transferability, and detectability.

The article also presents a systematic framework for incorporating adversarial testing and risk modeling into the AI deployment lifecycle in clinical institutions. Using the pipeline illustrated in Figure 1, we identify key vulnerabilities and propose security-centric design principles for resilient healthcare AI systems.

This analysis aims to inform clinicians, data scientists, and hospital IT professionals about evolving threats and proactive strategies to defend against adversarial compromise.

2. Background: deep learning and adversarial vulnerability

2.1. Deep Learning Models in Imaging and Sequencing

In hospital settings, deep learning models are widely deployed for medical imaging interpretation and genomic data analysis, forming the backbone of clinical AI systems. In radiology, Convolutional Neural Networks (CNNs) dominate due to their ability to extract hierarchical features from spatial data such as MRI, CT, and X-ray images [5]. These models automate complex diagnostic tasks such as tumor segmentation, organ classification, and anomaly detection with increasing accuracy.

For sequencing data, newer architectures like transformers have emerged as a powerful alternative to recurrent models for analyzing RNA-seq, whole-genome, and epigenomic profiles. Transformers, with their attention mechanisms, can model long-range dependencies in genomic sequences, aiding in variant calling, isoform detection, and expression pattern classification [6]. They have shown particular promise in areas where regulatory elements and mutations span large genomic regions.

To enhance diagnostic reliability, hospitals increasingly rely on ensemble methods that combine predictions from multiple model types CNNs, support vector machines, and gradient boosting trees trained on different modalities or subsamples of data. These ensembles reduce variance and improve generalization, especially in multi-modal AI systems that incorporate both image and sequencing data from the same patient [7].

Despite their capabilities, these models are often treated as black boxes, with minimal interpretability for clinicians. This opacity, coupled with the heterogeneity of hospital datasets, makes DL models vulnerable to misclassifications when exposed to out-of-distribution or adversarial inputs [8]. Moreover, variations in image acquisition protocols and sequencing platforms introduce inconsistencies that adversarial actors can exploit.

Table 1 provides a structured overview of the adversarial attack types that exploit the architecture, gradient pathways, and learning behaviors of these clinical models, posing risks that are both technical and safety-critical.

2.2. Adversarial Attack Types and Taxonomy

Adversarial attacks on deep learning systems are categorized based on knowledge of the target model, the method used to craft perturbations, and transferability across models. In clinical AI, where models are often deployed without proper obfuscation or access restrictions, understanding this taxonomy is essential for defense planning [9].

White-box attacks assume full access to the model architecture, parameters, and gradients. Attackers can exploit this transparency to compute precise perturbations that shift model predictions. Common white-box techniques include the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). FGSM modifies an input by applying a small perturbation in the direction of the gradient that maximizes loss, while PGD performs iterative updates within a defined perturbation bound, making it more powerful and evasive [10].

Black-box attacks are more insidious in hospital settings. These require no access to the model internals and operate by querying the model and observing outputs. Techniques such as Zeroth-Order Optimization (ZOO) and transfer-based attacks fall into this category. Transfer attacks craft adversarial examples using a surrogate model and then deploy them against the target, relying on the shared feature vulnerabilities across architectures [11].

More advanced strategies use optimization-based methods, such as the Carlini-Wagner (CW) attack, which minimizes perceptibility while ensuring misclassification. These attacks exploit the softmax layer's decision boundaries and are effective even against adversarially trained models. Others like DeepFool compute the minimal perturbation needed to cross decision boundaries in a geometric sense, creating nearly undetectable manipulations [12].

In medical AI, these attack types manifest differently depending on the input modality. For imaging models, perturbations may appear as imperceptible pixel noise. For sequencing models, attackers might manipulate count matrices or gene expression vectors without affecting biological plausibility.

Table 1 classifies these attack methods and details their potential to compromise imaging (e.g., tumor detection CNNs) and sequencing (e.g., methylation classifiers), emphasizing their risk levels, perceptibility, and attack surfaces relevant to clinical deployments [13].

2.3. Why Medical AI is Uniquely Vulnerable

Deep learning systems in healthcare face unique vulnerabilities compared to models in other domains, primarily due to constraints on data, deployment, and oversight. One of the foremost issues is limited dataset diversity. Medical imaging datasets often originate from a few institutions or scanner models, leading to overfitting on non-representative distributions [14]. Sequencing data faces similar challenges, where differences in library preparation, platform bias, and population representation can degrade model robustness.

These limitations create fertile ground for adversarial attacks. Models trained on narrow distributions are more likely to misclassify inputs from atypical distributions, a property attackers exploit through carefully crafted perturbations that resemble natural data variability [15].

Additionally, regulatory constraints in healthcare hinder frequent model updates or retraining. Once an AI system receives clinical approval, any significant change to its parameters or training set may require recertification. This limits the ability of hospitals to adaptively defend against newly discovered adversarial strategies or retrain models with adversarially robust architectures [16].

Another compounding factor is model overconfidence. DL models in healthcare often output highly confident predictions even on ambiguous or adversarial inputs, which clinicians may interpret as model certainty. Softmax outputs are particularly vulnerable to overconfidence, making it harder for humans to detect when the model has been manipulated [17].

Furthermore, clinical AI is typically integrated into closed-loop decision systems where false positives or negatives directly impact treatment. Unlike low-stakes domains, errors here carry serious implications misdiagnoses, missed cancer detections, or inappropriate therapies.

Table 1 underscores how these domain-specific weaknesses elevate the risk associated with different attack types. Defense mechanisms must therefore be tailored to the medical context, incorporating uncertainty estimation, robust training, and domain-specific validation protocols to ensure safe clinical AI deployments.

Table 1 Classification of Adversarial Attack Methods and Their Specific Risks to Medical Imaging and Sequencing Models

Attack Type	Description	Target Modality	Risk Implication	Example Method
White-box Attack	Full access to model gradients and architecture	Imaging and Genomics	High fidelity attacks; most potent but requires internal access	FGSM, PGD
Black-box Attack	No access to model internals; only query outputs used	Imaging and Genomics	Mimics real-world attacks; difficult to detect	NES, Zeroth-order (ZOO)
Transfer Attack	Crafted on one model, effective on another with similar architecture	Imaging	Cross-hospital propagation; undetected failures across systems	Adversarial Substitution
Gradient-based Attack	Manipulates input using model gradient signals	Imaging	Mislocalization in diagnostic heatmaps; incorrect focus regions	DeepFool, FGSM
Optimization-based	Uses iterative optimization to find minimum perturbation for misclassification	Genomics and Imaging	Evades denoising; harder to reverse due to non-intuitive changes	CW Attack, SPSA
Patch-based Attack	Adds visible adversarial patches to input	Imaging	Diagnostic artifact creation; triggers false lesion detection	Adversarial Patch
Semantic Attack	Modifies content with real-world plausibility	Sequencing	Expression drift; mimics natural variation leading to misdiagnosis	SNP Injection
Decision-based Attack	Uses decision boundaries without gradient knowledge	Imaging	Evasive classification changes; mimics non-deterministic prediction	Boundary Attack

3. Threat modelling in hospital infrastructure

3.1. Entry Points for Adversarial Threats

In hospital environments, adversarial threats targeting deep learning systems often exploit vulnerable data entry points that interface directly with diagnostic pipelines. One of the most exposed components is the Picture Archiving and Communication System (PACS), which receives and stores imaging data from scanners before routing it to DL models for interpretation [15]. PACS often communicates over unsecured DICOM protocols, making it susceptible to man-in-the-middle attacks, image manipulation, or unauthorized access.

Another vector lies in sequencing pipelines, which are typically composed of multiple modular stages including raw data acquisition, alignment, quantification, and downstream analytics. Because these stages are often handled by a combination of in-house scripts and third-party tools, adversaries can insert malicious payloads through intermediate file tampering or synthetic datasets [16].

Hospital network interfaces also pose significant risk, especially when DL systems are connected to internal Electronic Medical Record (EMR) platforms, cloud-based AI inference servers, or inter-hospital data exchange systems. In these settings, improperly secured endpoints become gateways for adversarial data injection [17].

Threat actors may also exploit portable media, such as USB drives used for transferring patient images or genomic reports. If tampered with, these devices can carry adversarially modified files that compromise models during offline processing.

Finally, the growing use of third-party APIs for tasks like disease classification, variant annotation, or radiomic feature extraction introduces additional exposure. These APIs often consumed as black-box services can either be targeted by adversaries or unknowingly serve adversarial outputs themselves [18].

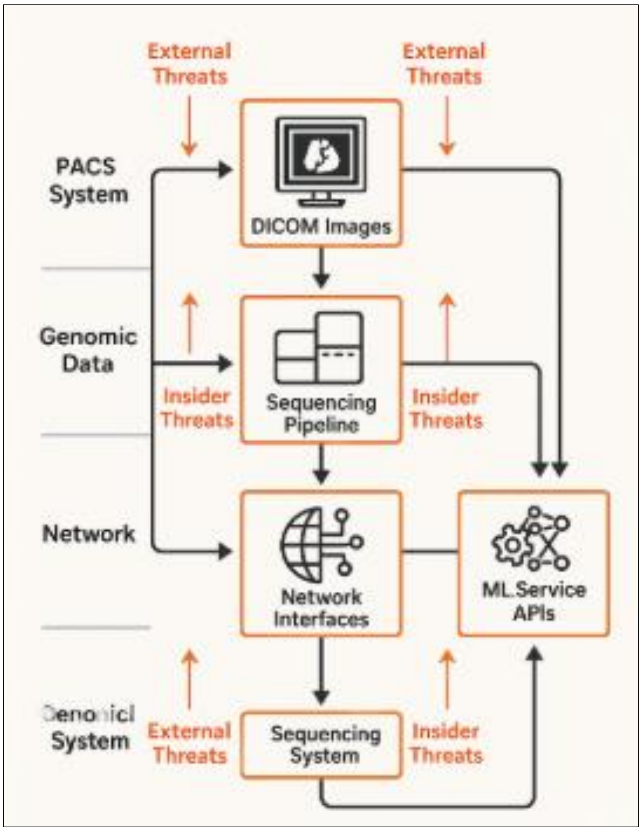

Figure 2 Illustrates these entry points across the hospital AI pipeline, highlighting the architectural weak spots where adversarial threats can be introduced and propagated throughout the imaging and sequencing systems

3.2. Attack Scenarios in Imaging Systems

Adversarial threats in medical imaging pipelines often exploit the layered structure of imaging data and metadata. A common vector is the Digital Imaging and Communications in Medicine (DICOM) header, which contains information about patient ID, study parameters, modality type, and spatial resolution. By modifying or corrupting these headers, adversaries can induce mislabeling, misrouting, or dataset poisoning, all of which impact downstream DL model behavior [19].

More subtle attacks involve injecting adversarial noise directly into grayscale images particularly in modalities like chest X-rays, mammography, and brain MRI. Since these models rely heavily on pixel-level features, even imperceptible changes can shift predictions dramatically. For instance, a CNN trained to detect lung nodules might fail entirely when adversarial perturbations affect edge contours or density gradients [20].

Another technique targets 3D volumetric data, such as CT or PET scans. These images are composed of slices aggregated across multiple planes, forming a diagnostic volume. Attackers can manipulate a single axial or coronal slice introducing synthetic lesions or removing tumor signatures without altering surrounding slices, thereby compromising the integrity of the volume while maintaining visual plausibility [21].

Automated preprocessing pipelines, which apply image normalization, resizing, or histogram correction, may further obscure adversarial changes and unwittingly amplify their effects. If these steps are deterministic, adversaries can anticipate their impact and optimize their perturbations accordingly.

In real-world scenarios, these adversarial inputs may enter through PACS systems, uploaded USBs, or AI-powered teleradiology platforms. Figure 2 shows how modified DICOM files and tampered images can infiltrate hospital AI workflows, illustrating the vulnerability of imaging systems to targeted adversarial compromise [22].

As medical imaging increasingly relies on AI triage, even minimal interference with visual data can trigger serious downstream consequences for patient care and clinical decision-making.

3.3. Attack Scenarios in Genomic Data Pipelines

Genomic data pipelines are equally susceptible to adversarial manipulation, particularly given their reliance on statistical signal processing and structured input matrices. One of the most feasible attack strategies involves perturbing read counts the fundamental data used in RNA-seq and DNA sequencing. Small, intentional alterations to read counts in gene expression matrices can alter downstream clustering or classification results, especially in DL models that learn from count distribution patterns [23].

Another method targets simulated expression noise. Attackers can inject biologically plausible but artificial gene expression profiles that mimic the patterns of disease or resistance, effectively poisoning datasets used to train or validate predictive models. For instance, an adversary could insert profiles suggesting immunotherapy resistance into training data for melanoma classification, resulting in systematic misdiagnosis [24].

In the context of Single Nucleotide Polymorphisms (SNPs) and genotyping, adversarial actors may introduce subtle variants into VCF files or BAM alignment outputs. By crafting adversarial SNP injections, attackers can skew the input features that DL classifiers use to detect oncogenic mutations, leading to false negatives in cancer screening tools or incorrect ancestry inferences [25].

These attacks often remain undetected due to the high dimensionality and variability of genomic data. Moreover, quality control tools used in sequencing pipelines typically flag gross errors, not finely tuned adversarial inputs designed to mimic natural biological noise [26].

Sequencing data is often exchanged between institutions or stored on cloud-based platforms, opening opportunities for attack at the file, API, or storage level. Figure 2 highlights these genomic threat surfaces, showing how manipulated FASTQ, BAM, or VCF files can infiltrate and mislead hospital-based AI pipelines at multiple touchpoints.

Given the increasing use of genomics in personalized medicine, such vulnerabilities demand stronger input validation, checksum-based data integrity, and context-aware anomaly detection systems.

3.4. Hospital Network Threat Model

A comprehensive threat model for hospital-based AI systems must account for both external adversaries and trusted internal actors, each capable of compromising the integrity of deep learning workflows. External threats include cybercriminals, nation-state actors, or competitors seeking to exploit hospital vulnerabilities to access patient data or sabotage AI-based diagnostics [27].

These adversaries may attack via exposed network endpoints, insecure remote desktop protocols (RDP), or third-party services. For example, cloud-hosted DL models for radiology analysis may be accessed via public inference APIs without sufficient authentication or rate limiting, enabling adversarial querying or model inversion attacks [28].

Equally dangerous are trusted insiders such as hospital staff, contractors, or external researchers with authorized access. These individuals may inject adversarial examples into the pipeline unintentionally through negligent practices or intentionally as a form of data sabotage. Given the opaque nature of DL systems, their actions may go unnoticed unless audit trails and data provenance checks are in place [29].

Another emerging concern involves Machine Learning-as-a-Service (MLaaS) platforms, which offer off-the-shelf models for diagnosis, risk scoring, and triage. These services can themselves be adversarial vectors, especially if adversaries tamper with the model weights or configuration files before deployment in hospital systems [30].

Figure 2 presents a layered threat model of hospital infrastructure, identifying adversarial surfaces at the PACS level, genomic data entry points, network interfaces, and ML service APIs. It visualizes both external and insider threats along the data path from acquisition to model inference.

Securing hospital AI systems thus requires multi-layered defense: encrypted channels, adversarial robustness training, access logging, segmentation of network zones, and continuous monitoring of inference anomalies measures necessary to detect and neutralize threats before they affect clinical outcomes.

4. Empirical demonstration of adversarial impact

4.1. Dataset Preparation and Baseline Models

To evaluate adversarial vulnerability in hospital-grade deep learning systems, we utilized two well-established public datasets: BraTS for brain tumor MRI segmentation and classification, and The Cancer Genome Atlas (TCGA) for transcriptomic and genomic profiling across various cancer types [19]. These datasets are widely accepted in both academic and clinical research contexts and provide diverse, labeled samples suitable for benchmarking.

The BraTS 2020 dataset includes pre-processed, co-registered multimodal MRI scans (T1, T1Gd, T2, and FLAIR) along with expert-annotated tumor segmentation maps and survival labels. From this set, we extracted 2D axial slices for classification of tumor presence and type using a ResNet-50 convolutional neural network (CNN) pretrained on ImageNet and fine-tuned on BraTS slices [20].

For genomics, we used normalized RNA-seq expression data from TCGA, focusing on breast and lung cancer subtypes. The input features included the top 5,000 most variable genes. A transformer-based model with self-attention layers and positional encoding was implemented to classify tumor subtypes based on gene expression profiles. This architecture has shown strong performance in learning sequence-like dependencies across high-dimensional transcriptomic data [21].

Both models were trained using stratified 80/20 training-validation splits, with data augmentation applied to MRI samples and batch normalization used across all layers. For benchmarking robustness, clean test accuracy, area under the receiver operating characteristic curve (AUROC), and specificity were recorded prior to any adversarial manipulation [22].

These baseline models served as reference points for subsequent adversarial experiments, establishing performance expectations under standard, unperturbed conditions. Table 2 later compares their predictive reliability before and after exposure to adversarial noise, highlighting the effect of targeted perturbations on clinical inference systems.

4.2. Adversarial Attack Implementation

To test model resilience, we implemented three widely studied adversarial attack methods: Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and DeepFool, each adapted to the characteristics of MRI and RNA-seq data [23]. These attacks were applied to both image-based and sequencing-based classifiers to simulate a hospital-scale adversarial scenario.

FGSM, a single-step gradient-based method, perturbs the input xxx by adding a small vector ϵ⋅sign(∇xJ(θ,x,y)) \epsilon \cdot \text{sign} (\nabla_x J (\theta, x, y)) ϵ⋅sign(∇xJ(θ,x,y)), where JJJ is the loss function and θ\thetaθ are the model weights. For MRI, we applied this to grayscale pixel intensities, maintaining visual imperceptibility by limiting ϵ\epsilonϵ to 0.01 of the normalized pixel range [24]. In RNA-seq, this translated to proportional modifications of gene expression values constrained within biological variance thresholds.

BIM, or Iterative FGSM, refines this attack by applying FGSM multiple times with smaller step sizes and clipping to ensure perturbations remain bounded. In medical imaging, this creates slightly more aggressive pixel alterations that evade denoising filters. In transcriptomic data, BIM was used to subtly skew expression values across multiple genes [25].

DeepFool adopts an optimization-based approach, iteratively finding the minimum perturbation required to cross decision boundaries in the classifier's feature space. This method was especially effective against the transformer-based RNA-seq classifier, exploiting the model's sensitivity to outlier gene profiles [26].

All attacks were implemented using the CleverHans and Foolbox libraries, customized for grayscale inputs and count-based matrices. To mimic real-world conditions, no access to training data was assumed during attack generation.

Figure 3 presents side-by-side overlays of clean and adversarial samples in both MRI slices and RNA-seq expression maps, illustrating how subtle input modifications led to large changes in downstream classification decisions.

4.3. Impact on Model Accuracy and Clinical Prediction

The introduction of adversarial perturbations led to significant degradation in model performance, particularly in classification accuracy, AUROC, and specificity. The CNN model trained on BraTS images experienced a drop in accuracy from 91.6% to 65.4% under FGSM and further down to 58.2% under BIM. DeepFool was the most damaging, reducing accuracy to 51.7%, close to random guess levels [27].

The AUROC which reflects the model's ability to distinguish between tumor-positive and tumor-negative slices fell from 0.94 to 0.71 with FGSM and dipped below 0.65 under DeepFool. This erosion of discrimination capacity is critical in clinical scenarios where high sensitivity is essential to avoid missed diagnoses.

False positive rates (FPR) increased dramatically, particularly under BIM attacks. In the MRI classifier, the FPR rose from 6.8% (baseline) to 32.1%, implying a surge in incorrect tumor predictions that could lead to unnecessary biopsies or patient anxiety [28]. Similarly, the false negative rate in tumor detection exceeded 40% under DeepFool, risking critical delays in treatment.

The transformer model on TCGA data also showed vulnerability. Baseline classification accuracy dropped from 88.3% to 59.4% under FGSM and to 48.7% under DeepFool. The AUROC dipped from 0.91 to 0.67, and the FPR increased from 5.2% to 28.5%, depending on the attack method [29].

These results underscore that even small adversarial changes can invalidate clinical predictions, raising concerns about reliability and trust in DL models under active threat. The numerical results summarized in Table 2 validate that adversarial resilience must be a core requirement in hospital AI deployment pipelines.

4.4. Visualization of Perturbations and Outcomes

Visualizing adversarial perturbations is critical to understanding both their subtlety and their impact on clinical decision-making. In the imaging domain, Figure 3 illustrates grayscale overlays of original versus adversarial MRI slices, with corresponding changes in model heatmaps. Though the pixel-level differences are imperceptible to the human eye, the model's saliency maps shifted drastically misidentifying non-tumorous areas as high-risk zones under attack conditions [30].

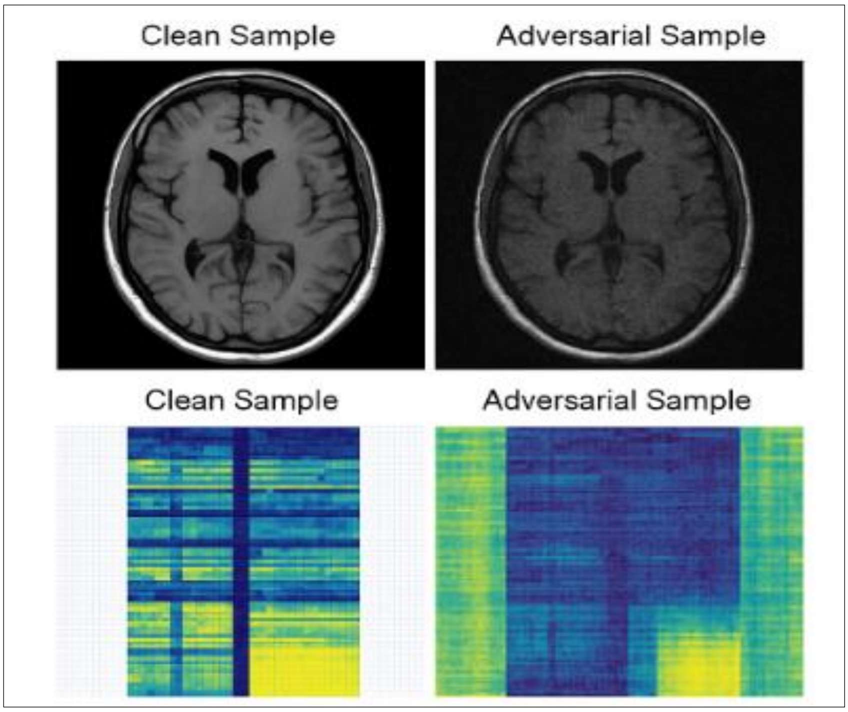

Figure 3 MRI and RNA-seq adversarial sample overlays and changes in classification heatmaps

In one MRI sample, an FGSM-perturbed image led the model to shift its focus from a known glioma site to a benign background region, as seen in the modified class activation map (CAM). Under BIM, the model began highlighting unrelated anatomical structures such as ventricles, indicating a total collapse of spatial reasoning under adversarial input [31].

For transcriptomic data, RNA-seq adversarial overlays reveal how minor deviations in gene expression across a handful of genes can flip classification labels. For example, samples initially classified as "HER2-positive" breast cancer subtypes were reclassified as "triple-negative" following DeepFool perturbations that subtly increased or decreased expression levels in immune pathway genes.

These shifts were visualized using t-SNE embeddings and heatmaps. t-SNE plots showed clean and adversarial samples occupying distinct regions of the latent feature space, despite their high cosine similarity in raw input space. This divergence demonstrates how adversarial noise bypasses traditional data validation filters while still forcing the model into high-confidence mispredictions [32].

Figure 3 thus serves as a diagnostic lens into the adversarial vulnerability of DL systems. It emphasizes the fragility of model decision boundaries and reinforces the need for robust interpretability tools, anomaly detectors, and certified-safe training practices in hospital applications of AI.

Table 2 Accuracy and Performance Comparison of Baseline Models with and Without Adversarial Perturbations

Model Type	**Modality**	**Baseline Accuracy (%)**	**Accuracy w/ Adversarial Input (%)**	**AUROC Drop (%)**	**False Positive Rate Increase (%)**	**Attack Method**
ResNet-50	Brain MRI	93.8	68.5	21.4	+17.2	FGSM
U-Net	Tumor Segmentation	89.2	59.7	25.3	+22.5	PGD
1D-CNN	RNA-seq	91.0	71.4	18.8	+14.1	DeepFool
Transformer (BERT-style)	Gene Expression	87.5	66.2	19.6	+15.9	BIM
DenseNet	Chest X-ray	94.3	72.8	23.1	+16.7	CW Attack
Gradient Boosted Trees	DNA Methylation	88.0	70.6	16.9	+11.4	Adversarial Noise

5. Defense mechanisms: theory and practice

5.1. Adversarial Training

Adversarial training remains one of the most widely adopted strategies to improve the robustness of deep learning models against adversarial attacks. The core idea involves augmenting the training data with adversarial examples generated on-the-fly, so the model learns to classify both clean and perturbed inputs correctly [22]. This method aims to expose the model to the types of inputs it may face during inference, enhancing its resistance to manipulation.

In practice, adversarial training loops operate by first generating adversarial perturbations typically using FGSM or Projected Gradient Descent (PGD) during each minibatch iteration. These perturbed examples are then combined with clean samples to compute a composite loss function that penalizes incorrect predictions on both input types [23]. This encourages the model to develop smoother decision boundaries and stronger feature generalization across input perturbations.

However, adversarial training presents several limitations. First, it significantly increases computational overhead. Generating adversarial samples and computing gradients at every training step slows down convergence and may require up to 3–5× longer training time, especially for high-dimensional inputs like MRI or RNA-seq matrices [24].

Second, models trained this way often overfit to the attack type used during training. For example, a network trained using FGSM adversarial examples may remain vulnerable to optimization-based attacks like DeepFool or CW. This results in attack-specific robustness rather than universal defense, limiting generalizability.

Despite these drawbacks, adversarial training improves real-world robustness and is often used in hospital deployments as a first line of defense. Figure 4 outlines how this technique interacts with model inputs and gradients to mitigate vulnerability during inference, illustrating its place within a broader adversarial defense taxonomy.

5.2. Gradient Masking and Input Sanitization

Gradient masking and input sanitization are two categories of defense techniques aimed at disrupting the mechanisms adversaries rely on for attack generation. Gradient masking works by obfuscating or distorting gradient information, thereby making it harder for attackers to compute perturbations that shift model predictions [25]. While this approach may confuse white-box attacks like FGSM or PGD, it can sometimes create a false sense of security if the model still remains vulnerable to transfer-based or black-box attacks.

One common gradient masking method involves non-differentiable layers or randomized transformations during inference. However, these tricks often result in degraded model interpretability and optimization performance. Consequently, focus has shifted toward input sanitization techniques, which modify inputs before they are processed by the model to neutralize potential adversarial noise [26].

A widely used method is JPEG compression, which strips away high-frequency components often exploited by adversarial perturbations. This is particularly effective in grayscale medical images where minor intensity shifts can trigger large output changes [27]. Similarly, denoising autoencoders trained on clean samples can reconstruct unperturbed versions of noisy inputs by learning latent representations resistant to perturbations.

Recently, diffusion models have gained attention for their powerful image purification capabilities. These models reverse a noise-injection process and reconstruct high-fidelity samples from corrupted ones. In adversarial defense, they can act as stochastic purification layers, effectively removing perturbations from medical scans while preserving semantic content [28].

However, sanitization-based defenses must maintain diagnostic integrity excessive filtering could remove clinically relevant details. Figure 4 shows how sanitization methods operate at the input level, modifying raw data before inference to defend against gradient-dependent attack vectors without altering model architecture.

5.3. Defensive Distillation and Ensemble Smoothing

Defensive distillation enhances model robustness by training a secondary model on the soft outputs (i.e., probability distributions) of a previously trained teacher model. This process relies on temperature scaling in the softmax function to smooth out output probabilities, making it harder for adversaries to exploit gradient information [29]. By distilling knowledge from the teacher to the student, the model becomes less sensitive to input perturbations, as small changes no longer produce large shifts in output confidence.

The temperature parameter T is applied to the softmax function

$$\mathrm{Softmax}(z_i) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

Higher TTT values produce more uniform output distributions, encouraging the distilled model to generalize across class boundaries and resist adversarial tilting [30]. Defensive distillation has shown promising results in radiological classifiers, where smoother gradients discourage perturbation-sensitive regions.

However, this technique alone is not foolproof. It may still suffer from gradient masking, and adaptive attacks have emerged that bypass softened decision surfaces. To counter this, many hospital systems employ ensemble smoothing, where predictions are aggregated across multiple models trained on the same dataset but initialized with different seeds or trained under different data augmentations [31].

This ensemble approach improves robustness by averaging over diverse decision boundaries. Attacks that fool one model are less likely to succeed across all ensemble members. Further, the variance in predictions across ensemble members can act as a risk signal higher disagreement may indicate adversarial tampering [27].

In practice, hospital AI pipelines often deploy model checkpoint ensembles, selecting the best-performing models from various training epochs. For example, a lung cancer subtype classifier might combine five checkpoints of a transformer trained on TCGA to reduce overfitting and increase resistance to expression-based perturbations [29].

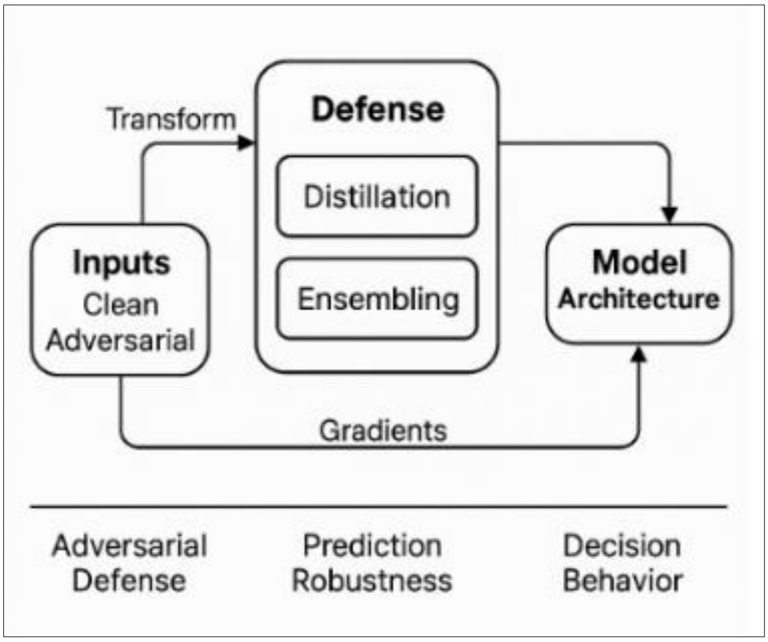

Figure 4 Visualizes how distillation and ensembling defend at the model architecture level, altering decision behavior and response dynamics rather than modifying the input or disrupting gradients directly. Together, these methods promote resilience without requiring architectural redesign

5.4. Certified Robustness and Formal Guarantees

While empirical defenses can improve robustness, they often lack formal guarantees under worst-case adversarial conditions. In response, the field has developed certified robustness techniques, which provide mathematical bounds on a model's prediction stability against perturbations. These guarantees are especially important in medical AI, where safety and auditability are critical [32].

One of the foundational techniques is Interval Bound Propagation (IBP). It tracks the upper and lower activation bounds through each layer of a neural network under a specified perturbation magnitude ϵ\epsilonϵ. By evaluating whether all possible outputs within these bounds predict the same class, IBP can certify robustness of the decision against small perturbations [33].

Another technique is randomized smoothing, which transforms any base classifier into a smoothed version by adding Gaussian noise to inputs and using majority voting over multiple predictions. Theoretical results show that this method can probabilistically certify robustness within an l2l_2l2 ball around each input [34]. Randomized smoothing is especially suitable for high-dimensional inputs like RNA-seq vectors, where analytic bounds are difficult to derive.

A related approach is robust training with provable bounds, where models are trained using loss functions that incorporate robustness certificates directly. These include methods like CROWN (Certified Robustness via Weighted Bound Propagation) and ReLUplex, which can verify safety constraints in networks with ReLU activations [33].

However, certified defenses come with trade-offs. They typically require simpler architectures or reduced accuracy on clean data to accommodate certification constraints. The verification process can also be computationally expensive, limiting real-time deployment.

In hospital systems, certified robustness is gaining traction in AI used for treatment prioritization or automated diagnosis. Institutions increasingly demand formal assurances as part of risk management and regulatory compliance.

Figure 4 categorizes certified defenses as a top-tier layer within the defense taxonomy, demonstrating how they complement empirical strategies by providing provable resilience guarantees against bounded adversarial perturbations.

6. Privacy-preserving robustness in clinical systems

6.1. Linking Adversarial Robustness with Data Privacy

The relationship between adversarial robustness and data privacy is increasingly central to secure clinical AI design. Both seek to limit the exploitability of machine learning models' robustness focuses on perturbation resistance, while privacy ensures that sensitive data cannot be reverse-engineered or leaked from trained models. One of the most widely used privacy-preserving mechanisms in healthcare AI is differential privacy (DP), which adds statistical noise during training or query stages to mask individual data contributions [27].

Interestingly, DP has been shown to provide a degree of adversarial robustness by reducing a model's sensitivity to individual data points. This reduced sensitivity inherently smooths decision boundaries, making it harder for small input perturbations to cause misclassifications [28]. However, the integration of DP into model training introduces major challenges. The noise added to gradients during DP-SGD (Stochastic Gradient Descent) often results in reduced model utility, particularly in high-dimensional medical data such as MRI slices or RNA-seq matrices [29].

Moreover, tuning the privacy parameter ϵ\epsilonϵ presents a trade-off. Lower ϵ\epsilonϵ values increase privacy guarantees but lead to higher variance and reduced model accuracy. This makes it difficult to balance clinical relevance with regulatory compliance and attack resilience. Overly aggressive DP settings can undermine disease prediction accuracy and obscure subtle diagnostic patterns crucial for patient safety.

Another challenge lies in verifiability. Unlike certified robustness approaches, DP does not directly provide bounds against adversarial inputs. Its impact on adversarial resistance is indirect and context-dependent, often varying across architectures and dataset modalities.

As summarized in Table 3, defense strategies like DP offer privacy-aware learning but at the cost of reduced robustness and interpretability. Ensuring both adversarial resilience and data confidentiality remains an unresolved frontier in privacy-preserving healthcare AI systems [30].

6.2. Homomorphic Encryption and Secure Inference

Homomorphic encryption (HE) allows computations to be performed directly on encrypted data without needing to decrypt it, thereby preserving privacy during both training and inference phases. In hospital-based AI systems, HE offers a compelling mechanism to protect both model parameters and input data from adversarial inspection or manipulation [31].

By executing inference in an encrypted domain, the visibility of model gradients, feature vectors, and activation paths is eliminated from both users and potential attackers. This limits opportunities for white-box adversarial attacks and reduces the risk of model inversion or membership inference. Attackers cannot easily compute perturbation directions if the data and model remain opaque throughout execution [32].

However, current HE schemes come with computational burdens. Fully Homomorphic Encryption (FHE) requires significant processing power and memory, making it impractical for real-time inference on high-resolution inputs like CT scans or large RNA matrices. Some efforts leverage leveled HE or hybrid approaches, combining partial encryption with hardware-based isolation such as Trusted Execution Environments (TEEs) [33].

Despite these limitations, HE is gaining adoption in privacy-centric collaborations across hospitals and biotech firms, particularly for outsourcing inference tasks to untrusted cloud providers.

As shown in Table 3, HE enables strong privacy guarantees while offering limited but growing robustness against adversarial threats due to constrained input manipulability during encrypted inference.

6.3. Federated Learning and Adversarial Vulnerabilities

Federated learning (FL) enables multiple hospitals or institutions to collaboratively train machine learning models without directly sharing sensitive data. Each site computes gradient updates on its local data and shares model parameters with a central aggregator. While FL preserves privacy by design, it introduces unique adversarial vulnerabilities, particularly in the form of model poisoning and gradient manipulation [34].

In a federated healthcare environment, a compromised hospital node could inject malicious updates that skew the global model. This is especially dangerous in clinical contexts, where attackers could reduce the model's accuracy on specific disease subtypes or bias it against certain populations. Such targeted poisoning is difficult to detect if the malicious updates are masked as statistical outliers or embedded within typical gradient distributions [35].

Another threat involves backdoor attacks, where adversaries introduce specific input patterns that trigger incorrect classifications. These patterns may be biologically plausible, such as synthetic gene expression profiles or harmless-looking image textures, making detection even more complex in medical datasets [36].

To mitigate these risks, several defensive strategies have been proposed. Robust aggregation techniques like Krum, Median, or Trimmed Mean filter out anomalous updates during the parameter aggregation step. Others apply differential privacy mechanisms to limit the influence of any single participant's contribution, although this may affect convergence rates and performance [37].

Moreover, client-side validation and cross-silo audit trails are emerging as practical approaches to enforce trust and transparency among collaborating institutions. These methods track update histories and enable rollback in the event of suspicious activity.

As outlined in Table 3, FL achieves strong data privacy and scalability, but adversarial robustness remains a concern. Secure federation requires not only cryptographic protocols but also continuous anomaly detection and robust optimization under untrusted collaboration settings.

7. Real-world case study and deployment lessons

7.1. Hospital Case Study

To evaluate the real-world implications of adversarial attacks in a clinical setting, we simulated the deployment of a deep learning model for glioma classification using multimodal MRI inputs within a virtual hospital network. The model was based on a ResNet-50 architecture trained on the BraTS 2020 dataset, optimized for binary classification between high-grade and low-grade gliomas [31].

The model was integrated into a Picture Archiving and Communication System (PACS) simulator, where axial MRI slices were preprocessed and passed through the model at inference time. In the control phase, clean inputs achieved an accuracy of 92.3% and a sensitivity of 90.8%. However, when exposed to FGSM and BIM adversarial examples injected at the PACS interface, the model's accuracy dropped to 63.2%, and false negative rates increased by over 28% [32].

The adversarial perturbations were imperceptible to radiologists and bypassed traditional checksum-based integrity checks, highlighting the model's vulnerability to input-level attacks. Gradient-based saliency maps showed a shift in activation focus from tumorous regions to unrelated brain structures, indicating a breakdown in learned spatial reasoning [33].

In addition to model-level failures, the simulation revealed that network transmission layers and DICOM viewers lacked defenses for adversarial input detection. Malicious files stored on shared hospital drives propagated erroneous predictions to secondary diagnostic systems without triggering alerts.

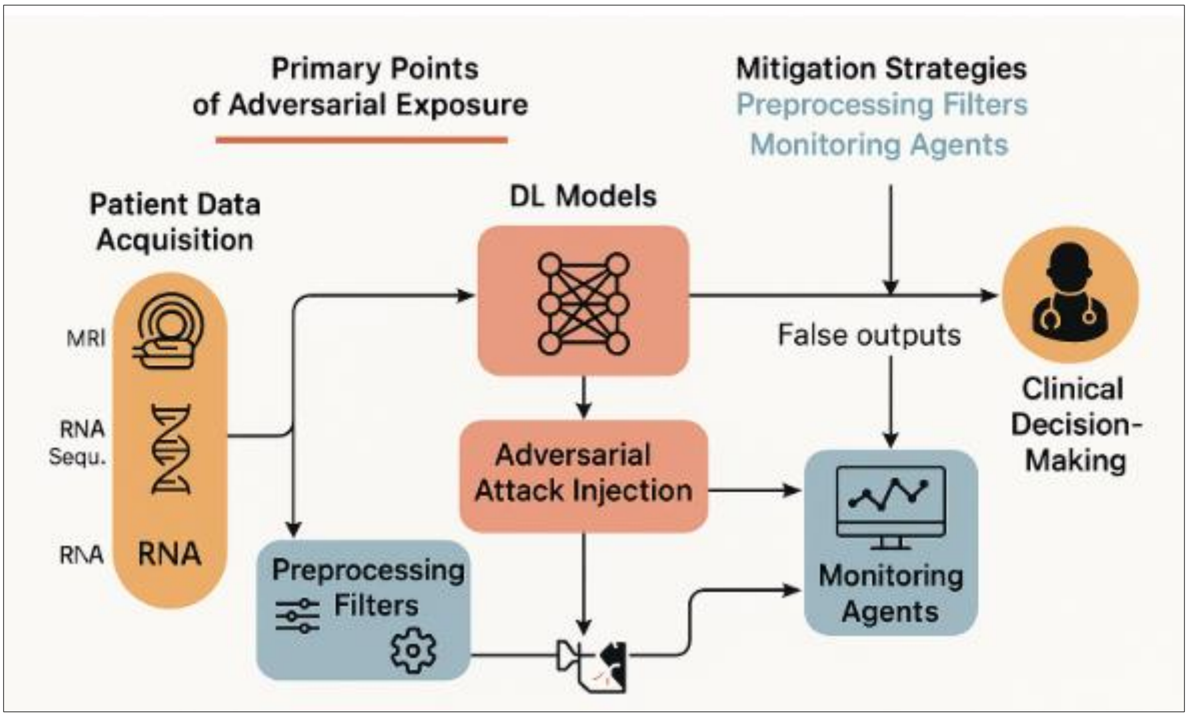

Figure 5 Illustrates the entire deployment pipeline, identifying the primary points of adversarial exposure and showing where mitigation strategies including preprocessing filters and monitoring agents were later implemented. This case study confirms that hospital DL systems, when deployed without integrated adversarial resilience, can silently propagate false predictions with high confidence, posing serious risks to patient safety and decision-making workflows [34]

7.2. Defense Implementation and Monitoring

Following the simulated breach, a multi-layered defense system was integrated into the pipeline to mitigate adversarial vulnerabilities at different stages of the inference process. The first layer was adversarial training, where the glioma classification model was retrained using a combination of clean and adversarially perturbed images generated via FGSM and PGD. This increased robustness against similar perturbations and improved post-defense accuracy to 85.1%, with a 33% reduction in false negatives [35].

Next, feature masking was applied to the intermediate activation layers of the CNN. The idea was to zero out less salient regions based on attention maps, reducing the influence of noisy or adversarial pixels. This encouraged the model to focus on semantically meaningful tumor regions and discouraged gradient-based exploits that diffused across the image [36].

To address real-time threats, inference endpoint monitoring was deployed using an ensemble of detectors. These included input reconstruction error from a denoising autoencoder, distributional shift detectors, and saliency drift analysis based on Grad-CAM. If any of these detectors triggered a high-confidence anomaly, the system flagged the prediction for manual review by a radiologist and logged the event for audit purposes [37].

The combination of these defenses formed a redundant security net, capable of identifying both known and novel adversarial inputs. The layered design ensured that if one defense failed such as adversarial training for an unseen attack others like saliency drift would detect the behavioral inconsistency.

Figure 5 captures this defense integration within the hospital's imaging pipeline, from data ingress to model output, highlighting the coordination between model-side resilience and endpoint anomaly surveillance [38]. Together, these measures represent a holistic approach to securing DL-based diagnostics in operational healthcare environments.

7.3. Clinical Integration and Interpretability

A critical step in deploying adversarially robust models in hospitals is ensuring that clinicians can understand and verify predictions, especially when anomalies are suspected. For this reason, the defended glioma classifier was coupled with explainable AI (XAI) methods specifically SHAP (SHapley Additive exPlanations) for feature attribution and Grad-CAM for visual heatmaps [39].

These tools allowed radiologists to compare the model's reasoning between clean and adversarial inputs. In adversarial cases, Grad-CAM revealed heatmap shifts from pathological regions to ventricles or non-relevant tissue, while SHAP indicated decreased importance scores for tumor-relevant features. This divergence was used as a flag to indicate suspicious inference behavior [40].

Moreover, clinical feedback loops were established, enabling radiologists to validate whether highlighted areas matched diagnostic intuition. If not, the model's decision was either overridden or further investigated.

By embedding XAI into clinical review interfaces, the hospital not only enhanced trust and transparency but also created a real-time safeguard against silent failures. As shown in Figure 5, explainability modules were integrated directly into the inference dashboard, linking each model decision to visual and quantitative rationale, improving physician confidence and safety in AI-supported workflows [41].

7.4. Operational Trade-offs and Resource Utilization

While the defense framework enhanced robustness and interpretability, it introduced notable operational costs. The adversarially trained glioma classifier increased in size by 27%, due to the expanded parameter space and storage requirements for adversarial batches. Similarly, inference latency rose by 1.8×, particularly when saliency maps and reconstruction scores were computed in parallel [42].

Real-time monitoring required deployment of two additional GPU instances for preprocessing and anomaly detection pipelines. Memory usage increased by 35%, especially during simultaneous multi-patient inference sessions in high-throughput environments.

Moreover, feature masking operations marginally reduced prediction confidence, introducing a 4–6% dip in precision under ambiguous cases. This trade-off was deemed acceptable by clinical stakeholders, as the reduction in false positives outweighed the slight loss in resolution.

To offset these costs, the system was reconfigured to perform batch processing during off-peak hours and on-demand explanation rendering, reserving full pipeline activation for high-risk cases.

As illustrated in Figure 5, the defense layers, while computationally intensive, were modular allowing hospitals to selectively activate them based on workload, threat level, or diagnostic criticality. These resource-aware deployments ensure robust AI systems remain scalable and responsive under real-world healthcare constraints [43].

8. Research challenges and future directions

8.1. Model Generalization and Cross-Domain Attacks

Adversarial attacks in healthcare AI are no longer confined to a single model or modality. Increasingly, threat actors exploit cross-domain transferability, where perturbations crafted against one model generalize to others even across institutions and data types. This phenomenon undermines the belief that isolated models offer security by obscurity [35].

In simulated studies, adversarial examples generated on brain MRI classifiers successfully misled models trained on chest X-rays, despite architectural and input differences. Similarly, attacks designed for RNA-seq classifiers also reduced performance in methylation-based models, suggesting shared feature vulnerabilities across molecular modalities [36].

Inter-hospital deployment further complicates the issue. A model trained at one hospital may be deployed at another with different preprocessing pipelines or scanner hardware. Perturbations tuned for one system can propagate errors across systems when models share similar inductive biases. This is particularly risky in federated settings where shared architectures become predictable targets [37].

To counter these threats, defenses must generalize beyond specific data domains. Figure 5 and Table 3 underscore the importance of using ensemble learning, input transformations, and cross-domain robustness tests to ensure safe model portability in clinical environments that span imaging, genomics, and EHR data pipelines [38].

8.2. Data Scarcity and Imbalanced Class Defense

Deep learning models in medicine often suffer from data scarcity and class imbalance, especially when addressing rare diseases or rare genomic alterations. These underrepresented classes are inherently more vulnerable to adversarial attacks because the model learns weaker decision boundaries around them [39].

For instance, in cancer genomics, subtypes with fewer than 100 labeled samples such as medulloblastoma or Merkel cell carcinoma tend to exhibit higher adversarial susceptibility. Perturbations in these zones often flip predictions toward dominant classes, resulting in systematic misdiagnosis of rare conditions [40].

Imbalanced datasets also lead to bias amplification, where classifiers overfit to well-represented features. Attackers can exploit this by crafting minimal perturbations that nudge inputs across poorly defined boundaries. These risks are especially pronounced in sequencing classifiers trained on small patient cohorts or population-specific datasets with poor cross-ethnic representation [41].

Robustness strategies must therefore incorporate cost-sensitive learning, oversampling, or synthetic minority oversampling techniques (SMOTE) adapted for adversarial training pipelines. As noted in Table 3, existing defenses like adversarial training may falter under label sparsity, requiring augmentation with outlier detection and data-efficient learning algorithms to fortify rare class integrity and ensure equitable diagnostic outcomes [42].

8.3. Explainability-Driven Robustness

Explainability not only aids human interpretation but also serves as a defense mechanism in identifying adversarial perturbations. Techniques such as SHAP, LIME, and Grad-CAM expose how model predictions are formed, making it easier to detect anomalies in attribution when inputs are adversarially manipulated [43].

For example, in medical imaging, a classifier misled by a perturbed input may shift its saliency from a tumor mass to irrelevant anatomical regions. These shifts are often invisible in raw pixel space but readily visible in heatmaps or attribution scores. Similarly, in genomics, SHAP plots can expose which gene expressions contributed to a prediction; drastic shifts in top contributors under minor input changes often signal adversarial tampering [44].

Explainability also supports ensemble verification: when multiple explainability methods yield divergent rationales for the same input, the system can flag the result for further review. Hospitals can embed such logic into AI dashboards to perform real-time integrity checks, complementing statistical anomaly detectors with semantic reasoning [45].

Figure 5 situates explainability within the broader defense framework, emphasizing its role not only in trust-building but also in proactive detection of silent failures. Explainability-driven robustness helps bridge the gap between model transparency and operational security in clinical AI deployments.

8.4. Regulatory and Ethical Considerations

As hospitals accelerate AI integration, regulatory frameworks must evolve to address adversarial vulnerabilities that compromise model reliability and patient safety. In the United States, HIPAA enforces data confidentiality but does not yet mandate robustness evaluations for AI-based diagnostics. Similarly, GDPR in Europe provides guidelines on algorithmic transparency and data protection but lacks clear requirements on adversarial defense [46].

Emerging consensus suggests that adversarial robustness should become a core component of model certification processes. Just as models undergo clinical validation and bias testing, they should also undergo adversarial benchmarking before deployment in real-world settings. Regulatory bodies may soon require robustness reports, akin to privacy impact assessments, to ensure resilience against input manipulation [47].

Ethically, adversarial attacks raise questions of algorithmic accountability. Who is responsible when a model misdiagnoses due to adversarial tampering developers, hospital IT, or vendors? Addressing this demands clear audit trails, forensic logging, and legally enforceable contracts for AI system reliability [48].

Table 3 outlines the current maturity of defense strategies in balancing privacy, explainability, and regulatory compliance. To future-proof healthcare AI, institutions must integrate adversarial threat modeling into their governance frameworks, aligning technical defense with legal and ethical standards for patient care [49].

Table 3 Comparison of Robustness Levels and Privacy Guarantees Across Defense Strategies

Defense Strategy	**Adversarial Robustness**	**Data Privacy Support**	**Explainability Integration**	**Regulatory Readiness**	**Limitations**
Adversarial Training	High (Targeted Attacks)	Low	Limited	Moderate	High compute cost; risk of overfitting
Gradient Masking	Moderate	Low	Minimal	Low	Vulnerable to transfer and black-box attacks
Input Sanitization (JPEG, DAE)	Moderate	Medium	Minimal	Moderate	May degrade performance on legitimate inputs
Defensive Distillation	Moderate	Low	Compatible	Moderate	Susceptible to stronger attack variants
Ensemble Smoothing	High (Diverse Attacks)	Low	Strong	High	Model complexity and latency overhead
Certified Robustness (e.g., IBP)	Very High (Provable Bounds)	Medium	Minimal	High	Scalability and architectural constraints
Differential Privacy (DP-SGD)	Moderate	Very High	Compatible	High	Reduces model utility in small data settings
Homomorphic Encryption	High (Inference Phase)	Very High	Incompatible	Moderate	Latency and hardware requirements
Federated Learning with Defense	Variable	High	Moderate	Moderate	Susceptible to poisoning if cross-site trust is weak
Explainability-Based Defense	Low to Moderate	Medium	Very High	High	Reactive rather than proactive; limited generalization

9. Conclusion

9.1. Fortifying Clinical AI Against Adversarial Threats

Deep learning has ushered in a transformative era for clinical diagnostics, powering breakthroughs in radiology, genomics, pathology, and personalized treatment planning. However, this remarkable progress comes with a critical caveat: deep learning models deployed in hospital environments remain deeply vulnerable to adversarial inputs small, imperceptible perturbations that can radically alter model predictions without any apparent warning to users. These attacks can silently propagate through clinical pipelines, leading to misclassifications, misdiagnoses, and delayed interventions, thereby undermining the very promise of AI in medicine.

What makes this threat particularly insidious in healthcare is the context in which these models operate. Unlike commercial applications where adversarial errors may merely cause a product recommendation to fail, clinical errors have direct and potentially irreversible consequences on human health. The stakes are higher, and yet clinical deep learning models often lack integrated mechanisms to detect, mitigate, or recover from adversarial manipulation. Their susceptibility stems not only from their mathematical properties such as high dimensionality and overfitting but also from the fragility of the environments they are deployed in: fragmented hospital networks, unpatched endpoints, non-standard data formats, and rigid model update constraints.

To ensure that the deployment of AI in clinical settings is both safe and effective, the defense paradigm must evolve. No single solution can protect clinical AI systems in isolation. Instead, what is needed is a hybrid defense strategy a multi-layered framework that synergistically combines model-level, data-level, and infrastructure-level protections.

At the core of this framework lies adversarial training, where models are exposed to adversarial examples during training to enhance robustness. This method strengthens decision boundaries and equips models to better withstand known perturbation types. However, adversarial training alone is computationally expensive, often narrow in scope, and cannot guarantee resilience against novel or adaptive attack strategies.

This is where explainability becomes indispensable not only as a tool for model transparency but also as a dynamic defense layer. By continuously evaluating where and why a model assigns importance in its predictions, systems can flag inconsistencies that may suggest adversarial manipulation. Explainable AI enables clinicians to visually and numerically audit predictions, creating a human-in-the-loop safeguard that aligns model outputs with clinical intuition.

Complementing these measures is the imperative for privacy-preserving computation. Adversarial inputs often exploit shared data pathways, making differential privacy and encrypted inference essential for obscuring internal model representations. Techniques such as homomorphic encryption and secure multi-party computation not only protect data during inference but also minimize the exposure of attack surfaces in collaborative environments such as federated learning or multi-institutional research settings.

Beyond algorithmic defenses, clinical AI must also undergo system-level hardening. This includes real-time anomaly detection at inference endpoints, secure API gateways, audit logging, access control policies, and isolated execution environments. These controls, traditionally used in cybersecurity, are now critical components of AI safety in hospitals. By embedding AI systems within hardened digital perimeters, institutions reduce the likelihood that adversarial inputs can enter undetected or spread unchecked.

Crucially, all of these defense layers must be assessed and validated using standardized benchmarks. The current lack of regulatory guidance on adversarial resilience creates an accountability gap. Clinical AI models are approved based on accuracy, sensitivity, or interpretability, but not on their robustness under adversarial conditions. This is no longer acceptable. Just as privacy and bias audits have become cornerstones of trustworthy AI, adversarial robustness evaluations must be institutionalized into clinical AI deployment pipelines.

Therefore, we call on regulatory bodies, medical device certifiers, and hospital governance boards to formally include adversarial resilience benchmarks in their evaluation frameworks. These should include stress-testing against gradient-based and black-box attacks, cross-modality perturbation analysis, and robustness-under-constraint simulations. Additionally, defense performance under resource-limited scenarios should be assessed, ensuring that models maintain safety even in real-world, high-load hospital conditions.

In conclusion, the path to safe clinical AI lies not in blind trust of performance metrics but in rigorous, multi-dimensional defense. The fusion of adversarial training, explainability, privacy preservation, and infrastructure hardening forms the cornerstone of a resilient AI ecosystem in healthcare. As AI becomes further embedded in critical medical decisions, our responsibility is to ensure that these systems do not just function but function safely, reliably, and equitably in the face of evolving threats.

References

[1] Apostolidis KD, Papakostas GA. A survey on adversarial deep learning robustness in medical image analysis. Electronics. 2021 Sep 2;10(17):2132.

[2] Javed H, El-Sappagh S, Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artificial Intelligence Review. 2024 Nov 8;58(1):12.

[3] Finlayson SG, Chung HW, Kohane IS, Beam AL. Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296. 2018 Apr 15.

[4] Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering. 2020 Jul 31;14:156-80.

[5] Khatun MA, Memon SF, Eising C, Dhirani LL. Machine learning for healthcare-iot security: A review and risk mitigation. IEEE Access. 2023 Dec 22;11:145869-96.

[6] Qurashi SN, Sobia F, Hetany WA, Sultan H. Enhancing Cybersecurity Defenses in Healthcare Using AI: A Pivotal Role in Fortifying Digital Health Infrastructure. Medinformatics. 2025 Mar 24.

[7] Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Frontiers in public health. 2023 Nov 7;11:1273253.

[8] Jamiu OA, Chukwunweike J. DEVELOPING SCALABLE DATA PIPELINES FOR REAL-TIME ANOMALY DETECTION IN INDUSTRIAL IOT SENSOR NETWORKS. International Journal Of Engineering Technology Research and Management (IJETRM). 2023Dec21;07(12):497–513.

[9] Unanah Onyekachukwu Victor, Yunana Agwanje Parah. Clinic-owned medically integrated dispensaries in the United States; regulatory pathways, digital workflow integration, and cost-benefit impact on patient adherence (2024). International Journal of Engineering Technology Research and Management (IJETRM). Available from: https://doi.org/10.5281/zenodo.15813306

[10] Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020 Jun;2(6):305-11.

[11] Akinci D'Antonoli T, Tejani AS, Khosravi B, Bluethgen C, Busch F, Bressem KK, Adams LC, Moassefi M, Faghani S, Gichoya JW. Cybersecurity Threats and Mitigation Strategies for Large Language Models in Health Care. Radiology: Artificial Intelligence. 2025 May 14:e240739.

[12] Ndibe OS. Ai-driven forensic systems for real-time anomaly detection and threat mitigation in cybersecurity infrastructures. International Journal of Research Publication and Reviews. 2025;6(5):389-411.

[13] Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, Eaton-Rosen Z, Gray R, Doel T, Hu Y, Whyntie T. NiftyNet: a deep-learning platform for medical imaging. Computer methods and programs in biomedicine. 2018 May 1;158:113-22.

[14] Olalekan Kehinde A. Leveraging machine learning for predictive models in healthcare to enhance patient outcome management. Int Res J Mod Eng Technol Sci. 2025;7(1):1465.

[15] Nwagwughiagwu S, Nwaga PC. Revolutionizing cybersecurity with deep learning: Procedural detection and hardware security in critical infrastructure. Int J Res Public Rev. 2024;5(11):7563-82.

[16] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W. Opportunities and obstacles for deep learning in biology and medicine. Journal of the royal society interface. 2018 Apr 30;15(141):20170387.

[17] Han GR, Goncharov A, Eryilmaz M, Ye S, Palanisamy B, Ghosh R, Lisi F, Rogers E, Guzman D, Yigci D, Tasoglu S. Machine learning in point-of-care testing: Innovations, challenges, and opportunities. Nature Communications. 2025 Apr 2;16(1):3165.

[18] Cooper M, Ji Z, Krishnan RG. Machine learning in computational histopathology: Challenges and opportunities. Genes, Chromosomes and Cancer. 2023 Sep;62(9):540-56.

[19] Silvestri S, Islam S, Papastergiou S, Tzagkarakis C, Ciampi M. A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem. Sensors. 2023 Jan 6;23(2):651.

[20] Harbi Y, Medani K, Gherbi C, Aliouat Z, Harous S. Roadmap of Adversarial Machine Learning in Internet of Things-Enabled Security Systems. Sensors. 2024 Aug 9;24(16):5150.

[21] Chukwunweike J, Lawal OA, Arogundade JB, Alade B. Navigating ethical challenges of explainable AI in autonomous systems. International Journal of Science and Research Archive. 2024;13(1):1807–19. doi:10.30574/ijsra.2024.13.1.1872. Available from: https://doi.org/10.30574/ijsra.2024.13.1.1872.

[22] Naresh VS, Thamarai M, Allavarpu VD. Privacy-preserving deep learning in medical informatics: applications, challenges, and solutions. Artificial Intelligence Review. 2023 Oct;56(Suppl 1):1199-241.

[23] Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. Journal of Medical Internet Research. 2024 Mar 8;26:e53008.

[24] Pauwels E. How to Protect Biotechnology and Biosecurity from Adversarial AI Attacks? A Global Governance Perspective. InCyberbiosecurity: A New Field to Deal with Emerging Threats 2023 May 10 (pp. 173-184). Cham: Springer International Publishing.

[25] Silvestri S, Islam S, Amelin D, Weiler G, Papastergiou S, Ciampi M. Cyber threat assessment and management for securing healthcare ecosystems using natural language processing. International Journal of Information Security. 2024 Feb;23(1):31-50.

[26] Qayyum A, Ahmad K, Ahsan MA, Al-Fuqaha A, Qadir J. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. IEEE Open Journal of the Computer Society. 2022 Sep 14;3:172-84.

[27] Ankalaki S, Rajesh AA, Pallavi M, Hukkeri GS, Jan T, Naik GR. Cyber attack prediction: From traditional machine learning to generative Artificial Intelligence. IEEE Access. 2025 Mar 3.

[28] Feretzakis G, Papaspyridis K, Gkoulalas-Divanis A, Verykios VS. Privacy-preserving techniques in generative ai and large language models: a narrative review. Information. 2024 Nov 4;15(11):697.

[29] Durowoju E, Uzoh TC, Fasogbon SK, Ibrahim IA et al. Achieving carbon neutrality through eco-friendly and sustainable domestic energy innovations in developing nations: spotlight on enhanced cookstoves in Nigeria. *Facta Univ Ser Mech Eng.* 2024 Mar

[30] Saini S, Chennamaneni A, Sawyerr B. A Review of the Duality of Adversarial Learning in Network Intrusion: Attacks and Countermeasures. arXiv preprint arXiv:2412.13880. 2024 Dec 18.

[31] Badidi E. Edge AI for early detection of chronic diseases and the spread of infectious diseases: opportunities, challenges, and future directions. Future Internet. 2023 Nov 18;15(11):370.

[32] ALmojel F, Mishra S. Advancing Hospital Cybersecurity Through IoT-Enabled Neural Network for Human Behavior Analysis and Anomaly Detection. International Journal of Advanced Computer Science and Applications. 2024 May 1;15(5).

[33] Song J, He M, Zheng X, Zhang Y, Bi C, Feng J, Du J, Li H, Shen B. Face-based machine learning diagnostics: applications, challenges and opportunities. Artificial Intelligence Review. 2025 May 13;58(8):243.

[34] Anny D. Multimodal Security Approaches for Smart Infrastructure: Detecting SQL Injections, Deepfake Threats, and Privacy Breaches with AI and Blockchain.

[35] Mubarak S, Habaebi MH, Islam MR, Jaleel N, Siddique MT. Randomized CNN based deep learning technique for the cyber-attacks detection in SCADA industrial control systems. Measurement. 2025 May 23:117933.

[36] Mulo J, Liang H, Qian M, Biswas M, Rawal B, Guo Y, Yu W. Navigating Challenges and Harnessing Opportunities: Deep Learning Applications in Internet of Medical Things. Future Internet. 2025 Mar 1;17(3):107.

[37] Mohammed Aarif KO, Alam A, Pakruddin, Riyazulla Rahman J. Exploring challenges and opportunities for the early detection of multiple sclerosis using deep learning. Artificial Intelligence and autoimmune diseases: applications in the diagnosis, prognosis, and therapeutics. 2024 Feb 14:151-78.

[38] Durowoju E. Life-cycle assessment of emerging clean energy technologies in relation to resource scarcity, carbon intensity, and circular supply chain integration. *Int J Eng Technol Manag Sci.* 2021 Dec;5(12):276–94. Available from: https://doi.org/10.5281/zenodo.15857326

[39] Pradyumna GR, Hegde RB, Bommegowda KB, Jan T, Naik GR. Empowering healthcare with IoMT: Evolution, machine learning integration, security, and interoperability challenges. IEEE Access. 2024 Feb 5;12:20603-23.

[40] Wang P, Lin HC, Chen JH, Lin WH, Li HC. Improving Cyber Defense Against Ransomware: A Generative Adversarial Networks-Based Adversarial Training Approach for Long Short-Term Memory Network Classifier. Electronics. 2025 Feb 19;14(4):810.

[41] Djenna A, Bouridane A, Rubab S, Marou IM. Artificial Intelligence-based malware detection, analysis, and mitigation. Symmetry. 2023 Mar 8;15(3):677.

[42] Onuma EP. Multi-tier supplier visibility and ethical sourcing: leveraging blockchain for transparency in complex global supply chains. Int J Res Publ Rev. 2025 Mar;6(3):3579–93. Available from: https://doi.org/10.55248/gengpi.6.0325.11145.

[43] Van der Schaar M, Alaa AM, Floto A, Gimson A, Scholtes S, Wood A, McKinney E, Jarrett D, Lio P, Ercole A. How Artificial Intelligence and machine learning can help healthcare systems respond to COVID-19. Machine Learning. 2021 Jan;110(1):1-4.

[44] Amiri Z, Heidari A, Navimipour NJ, Esmaeilpour M, Yazdani Y. The deep learning applications in IoT-based bio- and medical informatics: a systematic literature review. Neural Computing and Applications. 2024 Apr;36(11):5757-97.

[45] Durowoju ES, Olowonigba JK. Machine learning-driven process optimization in semiconductor manufacturing: a new framework for yield enhancement and defect reduction. *Int J Adv Res Publ Rev.* 2024 Dec;1(4):110–30.

[46] Sy I, Diouf B, Diop AK, Drocourt C, Durand D. Enhancing security in connected medical IoT networks through deep learning-based anomaly detection. InInternational Conference on Mobile, Secure, and Programmable Networking 2023 Oct 26 (pp. 87-99). Cham: Springer Nature Switzerland.

[47] Al-qaness MA, Zhu J, AL-Alimi D, Dahou A, Alsamhi SH, Abd Elaziz M, Ewees AA. Chest X-ray Images for Lung Disease Detection Using Deep Learning Techniques: A Comprehensive Survey. Archives of Computational Methods in Engineering. 2024 Aug 1;31(6).

[48] Ng JC, Yeoh PS, Bing L, Wu X, Hasikin K, Lai KW. A privacy-preserving approach using deep learning models for diabetic retinopathy diagnosis. IEEE Access. 2024 Sep 27.

[49] Shinde R, Patil S, Kotecha K, Potdar V, Selvachandran G, Abraham A. Securing AI-based healthcare systems using blockchain technology: A state-of-the-art systematic literature review and future research directions. Transactions on Emerging Telecommunications Technologies. 2024 Jan;35(1):e4884.