RESEARCH PAPER

A Comprehensive Review of Explainable Generative AI for Healthcare Interpretability, Clinical Reliability, and Prognostic Applications

Jignasha Kapadiya[1], Dr. Bhupendra Ramani[2]

[1-2]Faculty of Engineering & Technology, Parul University, Vadodara, India
Email: [1,*]23230041360004@paruluniversity.ac.in, [2]bhupendra.ramani2817@paruluniversity.ac.in

Abstract: The rapid evolution of Generative Artificial Intelligence (GenAI) has significantly transformed the landscape of intelligent healthcare by enabling advanced capabilities in disease prognosis, clinical risk stratification, synthetic medical data generation, and personalized decision support. Despite these advancements, the opaque and probabilistic nature of generative architectures—including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, and Transformer-based large language models—continues to raise critical concerns regarding interpretability, reliability, fairness, and clinical trustworthiness in high-stakes healthcare environments. Consequently, the integration of Explainable Artificial Intelligence (XAI) with GenAI has emerged as a pivotal research direction aimed at bridging the gap between predictive performance and transparent clinical reasoning. This review presents a comprehensive and critical analysis of recent developments in explainable generative AI frameworks for healthcare risk assessment and prognosis prediction. The study systematically examines state-of-the-art generative architectures and their applications across diverse medical domains, including oncology, cardiology, radiology, neurology, dermatology, and intensive care systems. Furthermore, the review investigates contemporary explainability approaches—such as SHAP, LIME, counterfactual reasoning, saliency visualization, attention mechanisms, and intrinsically interpretable architectures—and evaluates their effectiveness in enhancing the transparency and clinical interpretability of generative outputs. In addition, this work critically analyzes existing integration strategies between GenAI and XAI, highlighting key challenges associated with black-box decision-making, multimodal data fusion, explanation fidelity, algorithmic bias, fairness preservation, uncertainty quantification, and real-time clinical interpretability. Particular emphasis is placed on the limitations of current post hoc explanation methods when applied to probabilistic and high-dimensional generative outputs, including synthetic electronic health records, medical imaging, and longitudinal prognosis trajectories. The review further explores emerging concerns related to ethical governance, privacy preservation, regulatory compliance, and accountability under healthcare frameworks such as HIPAA, GDPR, and FDA AI guidelines. A major contribution of this review is the identification of unresolved research gaps, including the lack of standardized evaluation benchmarks for explainability quality, insufficient clinician-centered validation, limited deployment in real-world healthcare environments, and the absence of scalable, domain-adaptive GenXAI frameworks capable of integrating heterogeneous clinical modalities. Based on these findings, the paper proposes a future research roadmap emphasizing hybrid interpretable architectures, fairness-aware generative systems, multimodal explainability pipelines, human-in-the-loop validation strategies, and regulation-ready AI governance mechanisms. Overall, this review provides a consolidated foundation for researchers, clinicians, policymakers seeking to develop trustworthy, interpretable, and clinically deployable generative AI systems that not only deliver accurate predictions but also

provide transparent and medically meaningful justifications to support informed healthcare decision-making.

Keywords: Generative AI, Explainable AI (XAI), Risk Assessment, Prognosis Prediction, Trustworthy AI

How to cite this article: Kapadiya J, Ramani B. A Comprehensive Review of Explainable Generative AI for Healthcare Interpretability, Clinical Reliability, and Prognostic Applications. Int J Drug Deliv Technol. 2026;16(53s): 305-322. DOI: 10.25258/ijddt.16.53s.35

Source of support: Nil.

Conflict of interest: None.

I. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) technologies has catalyzed transformative changes in the healthcare sector, particularly in areas like disease diagnosis, prognosis prediction, and risk stratification. Among these advancements, Generative AI (GenAI) models—such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Transformer-based architectures—have demonstrated strong capabilities in synthesizing patient data, enhancing predictive accuracy, and modeling complex clinical patterns across diverse datasets [1], [2]. However, the black-box nature of many generative models poses a substantial challenge to their adoption in real-world clinical settings. Physicians and healthcare professionals often find it difficult to trust AI predictions without clear insight into the rationale behind them [3]. This lack of transparency and interpretability can hinder clinical decision-making, particularly in high-stakes applications involving patient risk assessment and prognostic analysis [4].

To address this limitation, the integration of Explainable AI (XAI) techniques has become a key focus area in medical AI research. XAI frameworks such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and attention-based visualization methods have been adapted to help clinicians interpret and validate AI-driven decisions [5]. When combined with generative models, these techniques can make outputs more transparent, ethically aligned, and clinically trustworthy—ultimately supporting improved patient outcomes and shared decision-making processes [6]. Despite the growing body of work at the intersection of GenAI and XAI, the landscape remains fragmented, with disparate methodologies, insufficient validation on real-world clinical datasets, and a lack of standardized evaluation metrics. Moreover, there is no consolidated framework or comprehensive review that thoroughly examines how explainable generative models are being applied specifically for risk assessment and prognosis in healthcare.

The aim of this review is to address this gap by providing a comprehensive synthesis of current techniques that integrate explainability into generative models for healthcare applications. We analyze key architectural innovations, XAI integration strategies, challenges in clinical deployment, and emerging trends. This review also outlines future research directions necessary to develop transparent, ethical, and trustworthy generative AI systems for critical medical decision support.

1.1 Background

The convergence of artificial intelligence (AI) and healthcare has unlocked new avenues for data-driven clinical support. Among various AI paradigms, Generative AI (GenAI) stands out for its ability to learn the underlying distribution of complex medical datasets and generate synthetic, predictive, or augmented information. Applications range from synthetic patient record creation to predictive modeling of disease trajectories, all of which have the potential to support diagnosis, treatment planning, and early risk detection [1], [2].

Yet, while these models deliver strong performance, they often lack interpretability—a core requirement in healthcare decision-making. Clinicians must understand, question, and trust the recommendations made by AI systems, especially when the consequences impact patient outcomes. As a result, Explainable AI (XAI) has emerged as a pivotal area of research. XAI methods aim to elucidate the logic and reasoning behind AI predictions, using techniques like SHAP, LIME, saliency maps, and attention mechanisms [5], [6]. The intersection of GenAI and XAI is particularly promising. If generative models can be made explainable, they can not only predict risks or simulate outcomes but also justify those predictions in a human-understandable format. This capability is essential to ensure ethical, fair, and regulatory-compliant deployment of AI tools in medicine [3], [4].

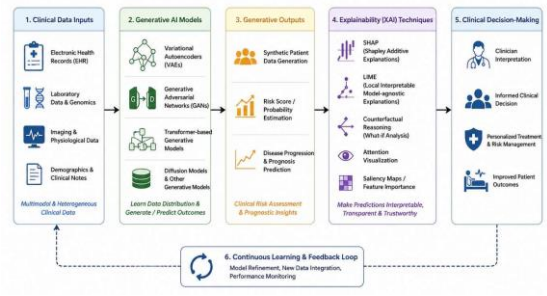

Figure 1. *Overview of Generative AI and Explainability integration for clinical risk assessment and decision-making*

1.2 Research Gap

Despite the rapid advancement of both Generative Artificial Intelligence (GenAI) and Explainable Artificial Intelligence (XAI), their convergence within healthcare risk assessment and prognosis prediction remains fragmented, insufficiently validated, and methodologically immature. Current research predominantly emphasizes predictive performance and synthetic data generation capabilities, while comparatively limited attention has been directed toward ensuring interpretability, clinical transparency, and regulatory trustworthiness in generative healthcare systems [3], [5]. As a result, most state-of-the-art generative architectures—including GANs, VAEs, diffusion models, and Transformer-based frameworks—continue to function as opaque black-box systems, thereby restricting their practical adoption in high-stakes clinical environments where explainability and accountability are essential [1], [6].

A major research gap lies in the incompatibility between existing XAI techniques and modern generative architectures. Most explainability approaches, such as SHAP, LIME, saliency mapping, and feature attribution methods, were originally designed for discriminative learning models including classifiers and regressors [4], [5]. However, generative models produce probabilistic, multimodal, and high-dimensional outputs such as synthetic medical images, clinical narratives, longitudinal patient trajectories, and disease progression simulations, making conventional explanation methods inadequate for capturing the underlying reasoning processes [3], [8]. Consequently, current XAI frameworks often fail to provide explanations that are faithful, stable, clinically interpretable, and actionable for healthcare professionals.

Another critical limitation in existing literature is the absence of unified and

standardized evaluation frameworks for GenXAI systems. Most prior studies evaluate model effectiveness primarily using predictive accuracy or image quality metrics while neglecting explanation fidelity, robustness, fairness, uncertainty quantification, and clinical usability [5], [14]. Furthermore, there is currently no consensus regarding how "explainability quality" should be measured in healthcare applications, particularly for probabilistic outputs such as survival predictions, synthetic radiological scans, or personalized prognosis trajectories. This lack of standardized benchmarking significantly limits reproducibility, cross-study comparison, and clinical reliability.

The literature also reveals insufficient clinical validation and limited real-world deployment of explainable generative models. Existing studies are largely theoretical, experimental, or domain-specific, with minimal large-scale validation across heterogeneous healthcare environments [2], [10]. Very few investigations compare AI-generated explanations with expert physician reasoning or assess clinician trust, interpretability satisfaction, and workflow integration in hospital settings. Moreover, the majority of current systems rely heavily on post hoc explanation techniques, while intrinsically interpretable generative architectures and real-time explainability mechanisms remain underexplored [3], [8].

From an ethical and regulatory standpoint, substantial gaps persist in fairness-aware modeling, accountability, and compliance with healthcare regulations such as HIPAA, GDPR, and emerging FDA AI governance frameworks [7], [16]. Generative models trained on biased or demographically imbalanced datasets risk amplifying healthcare disparities and producing clinically unsafe recommendations. Although some recent studies have introduced fairness-aware generative approaches, robust integration of transparency, fairness auditing, bias mitigation, and privacy preservation into a unified explainable generative framework remains largely unresolved [2].

Additionally, current research demonstrates limited focus on multi-modal and domain-adaptive GenXAI systems capable of integrating heterogeneous healthcare data sources such as electronic health records (EHRs), medical imaging, genomic information, wearable sensor streams, and clinical text. Most existing models are modality-specific and lack scalable architectures that can simultaneously provide accurate predictions, transparent reasoning, and clinically contextualized explanations across diverse medical specialties [6], [12].

These unresolved gaps collectively highlight the urgent need for a comprehensive and systematic review that critically synthesizes the evolving landscape of explainable generative AI in healthcare. Such an investigation is necessary to identify existing limitations, evaluate integration strategies between GenAI and XAI, examine clinical applicability, and establish future research directions toward the development of trustworthy, interpretable, ethically aligned, and regulation-ready AI systems for healthcare risk assessment and prognosis prediction.

1.3 Objectives

This review aims to bridge the identified research gaps by presenting a comprehensive, focused, and systematically structured analysis of the evolving intersection between Generative Artificial Intelligence (GenAI) and Explainable Artificial Intelligence (XAI) in healthcare. The primary objectives of this review are as follows:

- To critically examine and synthesize state-of-the-art generative AI

architectures employed in healthcare applications, particularly for clinical risk assessment, disease prognosis, patient stratification, and predictive decision support.

- To classify, compare, and evaluate existing explainability techniques—including post hoc and intrinsic XAI methods—applicable to generative models, with special emphasis on their interpretability, transparency, and clinical relevance.
- To investigate current integration strategies that combine GenAI and XAI frameworks for enhancing trustworthy, transparent, and human-centered clinical decision-making processes.
- To analyze key technical, ethical, and operational challenges associated with deploying explainable generative models in healthcare environments, including issues related to fairness, bias mitigation, data privacy, robustness, accountability, and regulatory compliance.
- To explore the role of explainable generative systems in supporting personalized medicine, real-time prognosis prediction, synthetic clinical data generation, and multimodal healthcare analytics.
- To identify limitations in current research, such as the lack of standardized evaluation frameworks, insufficient clinical validation, and limited real-world deployment of GenXAI systems across diverse healthcare domains.
- To propose a future research roadmap for developing reliable, scalable, and clinically trustworthy explainable generative AI systems that can support safe, ethical, interpretable, and evidence-driven medical decision-making.

II. LITERATURE REVIEW

2.1 Generative AI in Healthcare

Generative AI has emerged as a powerful paradigm in medical AI, with applications that range from synthesizing high-dimensional clinical data to predicting disease trajectories. Popular generative architectures include Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Transformer-based models. These models can generate synthetic electronic health records (EHRs), simulate patient outcomes, and aid in augmenting limited datasets, particularly in rare disease scenarios [1], [2].

For instance, Bhuyan et al. [1] emphasized the role of generative AI in clinical efficiency and personalized care, highlighting how synthetic data generation can mitigate privacy concerns and enhance machine learning performance. Similarly, Bt-GAN by Ramachandranpillai et al. [2] demonstrated how fairness-aware generative models could generate unbiased synthetic health data across different demographic groups.

Recent studies have extended GenAI to high-resolution image generation, cancer histopathology, and time-series patient monitoring, proving its utility in both structured and unstructured modalities. However, these models remain largely opaque, raising concerns about their clinical trustworthiness.

2.2 Explainable AI (XAI) Techniques in Clinical Settings

Explainable AI (XAI) seeks to provide transparency to black-box AI models by offering human-interpretable justifications for predictions. Techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are widely adopted to understand the influence of input features on model outputs [5]. In medical domains, these tools have been used to enhance trust in diagnostic systems, support clinical

audits, and meet ethical and legal requirements.

Chang et al. [4] applied XAI to a sepsis mortality prediction model, showing that interpretable outputs improved clinicians' ability to validate AI-generated risk scores. Other methods like counterfactual explanations, saliency maps, and attention visualizations have also proven effective in interpreting complex models, particularly in medical imaging and natural language processing (NLP) contexts [6].

Despite their promise, most XAI methods have been developed for discriminative models like classifiers or regressors. Their extension to generative models—which often output data distributions rather than scalar predictions—presents unique methodological challenges [3].

2.3 Integrating XAI with Generative Models

The intersection of GenAI and XAI remains relatively nascent. Schneider [3] identifies this integration as a crucial but underdeveloped research frontier. Most efforts focus on post hoc explanations, which attempt to explain generative models after training, often using surrogate models or feature attribution. However, these approaches can misrepresent the true logic of generative architectures, reducing their reliability in clinical use.

Some studies have proposed intrinsically interpretable generative models, such as those incorporating disentangled latent variables or attention-based conditioning. Yet, these models often trade off performance for transparency, and few are validated on real-world clinical datasets.

Xiong et al. [5] propose a data mining framework to evaluate explanation quality, fairness, and semantic consistency in generative outputs. Meanwhile, Huang et al. [6] argue that there is still a long way to go before deep generative models can produce interpretable outcomes that align with clinicians' reasoning processes.

2.4 Gaps in Existing Literature

Although the promise of explainable generative AI in healthcare is clear, several key gaps persist:

- Limited clinical validation of explainable generative models on real-world patient data.
- Lack of unified evaluation frameworks for measuring interpretability, accuracy, and fairness jointly.
- Absence of regulatory alignment with explainability standards in FDA, GDPR, or HIPAA contexts.
- Insufficient personalization in generated outputs for individualized prognosis or risk profiling.
- Over-reliance on post hoc explanations, which may offer limited reliability in clinical practice.

Table 1. *Comparison of recent (2023–2025) research studies on the integration of Generative AI and Explainable AI in healthcare, highlighting their objectives, methodologies, limitations, and proposed future directions.*

Study (Year)	Research Objective	Methodology Used	Limitations	Future Scope
Bhuyan et al. (2025)	Leverage Gen AI for enhancing clinical efficiency	GenAI models for patient data synthesis; exploratory	Limited implementation of real-time explainability;	Integration with clinical workflows and real-time expla

	and personalized care with transparent outputs.	use of XAI for explainability.	lacks external validation.	nation systems.
Ramachandranpillai et al. (2024)	Develop a fairness-aware GAN for generating unbiased synthetic health data.	Bias-transforming GAN architecture; evaluated on demographic fairness.	Post hoc fairness explainability only; does not address interpretability of outcomes.	Expand to other diseases and embed intrinsic interpretability in GANs.
Schneider (2024)	Survey and conceptualize the field of explainable generative AI.	Comprehensive literature survey; taxonomy and conceptual road map proposed.	Conceptual but not experimentally validated.	Implementation of GenXAI models across diverse healthcare domains.
Chang et al. (2024)	Apply XAI to sepsis	SHAP applied to mortality	Focuses only on one doma	Develop doma in-specific
	mortality prediction for improving clinical interpretability.	prediction using clinical features.	in (sepsis); lacks generalizability.	interpretability metrics for clinical tasks.
Huang et al. (2024)	Assess current deep learning models in healthcare for explainability gaps.	Review and critique of XAI techniques in NLP-based medical models.	Primarily focused on NLP, not applicable to imaging or structured data.	Adapt methodologies for multi-modal data and real-time outputs.
Xiong et al. (2024)	Propose a framework to evaluate explanation quality in generative models.	Data mining-based semantic alignment for XAI evaluation.	The framework is theoretical; lacks clinical trials or dataset deployment.	Apply the framework to longitudinal clinical datasets for validation.
Okonji et al. (2024)	Examine algorithmic, ethic	Analytical review of GenAI	Lacks empirical validation; focus	Develop robust regulatory

	al, legal, and societal considerations in Gen AI applications in healthcare.	applications in medical imaging and text analysis.	es on theoretical aspects.	frameworks and conduct empirical studies.
Hou et al. (2024)	Survey self-explainable AI methods for medical image analysis.	Review of S-XAI techniques integrating explainability into model training.	Limited to image modalities; lacks application to other data types.	Extend S-XAI approaches to multi-modal healthcare data.
Al Amin et al. (2024)	Develop an explainable AI framework for AIoMT applications in healthcare.	Integration of SHAP, LIME, and Grad-CAM in ensemble CNN models.	Focused on brain tumor detection; needs broader application.	Apply framework to diverse AIoMT healthcare scenarios.
Shokrollahi et al. (2023)	Provide a comprehensive review of Gen AI applications in healthcare.	Analysis of transformer and diffusion models in various medical domains.	Does not focus on explainability aspects.	Incorporate XAI techniques into GenAI applications.
John Snow Labs (2025)	Explore use cases, benefits, and challenges of Gen AI in healthcare.	Industry-focused analysis of GenAI trends and applications.	Lacks academic rigor and empirical data.	Conduct systematic studies to validate industry observations.
Nature Digital Medicine (2024)	Review practical models for generating synthetic health records.	Scoping review of DL models for synthetic medical text, time series, and longitudinal data.	Limited discussion on explainability of generated data.	Integrate XAI methods to enhance transparency of synthetic data generation.

The literature review reveals a growing body of work at the intersection of Generative AI (GenAI) and Explainable AI (XAI), with researchers increasingly recognizing the need for transparency in high-stakes clinical applications. While GenAI models such as GANs, VAEs, and Transformers are being employed for synthetic data generation, disease prediction, and risk stratification, the integration of explainability remains inconsistent and largely underdeveloped. Most studies employ post hoc methods like SHAP and LIME, with limited exploration of intrinsic interpretability or real-time explanation mechanisms. Several works, such as those by Ramachandranpillai et al. and Huang et al., highlight fairness and modality-specific concerns, while others like Schneider and Xiong emphasize the need for systematic evaluation frameworks and conceptual clarity. Notably, the majority of current efforts are theoretical or domain-specific, with minimal empirical validation across diverse clinical settings. This fragmentation underscores the urgent need for standardized methodologies, multi-modal explainability approaches, and rigorous clinical testing to ensure the safe and trustworthy deployment of GenAI systems in healthcare.

While the preceding sections highlighted the gaps in current literature, it is equally important to illustrate how explainable generative AI methods can be aligned with real-world clinical challenges. Table X provides a mapping between major clinical problems, the GenXAI methods applied, and the expected benefits, thereby demonstrating the translational potential of these approaches in healthcare. This practical perspective helps bridge the gap between technical innovation and clinical applicability.

Table 2. *Mapping of clinical problems to GenXAI methods and their expected benefits in healthcare risk assessment and prognosis*

Clinical Problem	GenXAI Methods Applied	Expected Benefits
Sepsis Mortality Prediction	Transformer-based generative models + SHAP, counterfactual explanations [4]	Transparent risk scores, improved clinician trust in ICU prognosis, early intervention strategies
Oncology (Cancer Prognosis & Treatment Response)	GANs for synthetic histopathology + attention-based VAEs with saliency maps [6]	Data augmentation for rare cancers, interpretable tumor progression modeling, personalized treatment planning
Cardiology (Arrhythmia & Heart Failure Risk)	Variational Autoencoders (VAEs) with SHAP/LIME applied to ECG/EHR data [1]	Improved interpretability of risk predictions, patient-specific monitoring, reduced false alarms in cardiac care
Neurology (Alzheimer's Disease Progression)	Diffusion models + counterfactual reasoning + attention heatmaps [2,5]	Visual simulation of disease trajectories, explainable biomarkers, better prognosis communication with patients/families
Dermatology (Skin Lesion & Hair	GANs/ViTs for image synthesis	Enhanced diagnosis/prognosis from trichoscopy

Growth Prognosis)	+ Grad-CAM + SHAP [3]	and dermoscopy, transparent treatment outcome predictions, patient-centered visual explanations
Radiology (CT/MRI Prognosis & Risk Assessment)	Multimodal Transformers (imaging + EHR) + hybrid GenXAI frameworks [1,6]	Explainable diagnostic imaging, integration with clinical workflows, regulatory-compliant reporting

As shown in Table 2, the integration of GenXAI methods across diverse medical specialties highlights both the breadth of ongoing applications and the depth of opportunities for further exploration. Oncology and radiology demonstrate relatively mature adoption of generative modeling combined with explainability, largely due to the availability of annotated imaging datasets and established clinical workflows. In contrast, areas such as dermatology and neurology remain underexplored, despite their strong potential for prognosis-driven decision support. Sepsis and cardiology applications show promise in early risk stratification but still require rigorous validation on longitudinal, real-world datasets. Collectively, these mappings emphasize that while GenXAI holds significant potential to deliver accurate and interpretable predictions, the degree of clinical readiness varies across domains, underscoring the importance of standardized evaluation and domain-specific validation in future research.

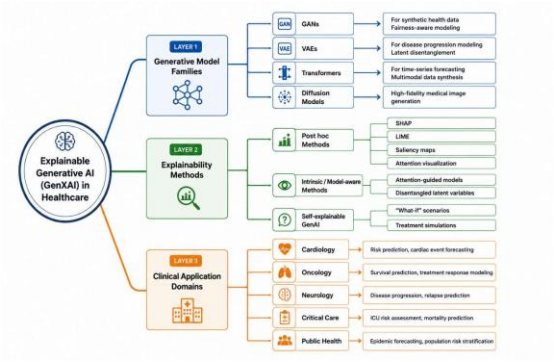

Figure 2. *Conceptual taxonomy of Explainable Generative AI (GenXAI) in healthcare, illustrating the relationship between generative models, explainability methods, and clinical application domains*

The conceptual taxonomy presented in Figure 2 synthesizes the current landscape by linking generative model families, explainability techniques, and their clinical application domains. This framework illustrates how different strands of research converge toward the goal of transparent and trustworthy AI for healthcare risk assessment and prognosis. However, while the taxonomy highlights promising directions, it also exposes the fragmentation of existing approaches—most applications remain domain-specific, rely on post hoc methods, and lack consistent validation across clinical settings. Building on this synthesis, the next section examines the key challenges that hinder the integration of GenAI and XAI into clinical workflows and reviews emerging techniques designed to overcome these barriers.

III. CHALLENGES AND INTEGRATION TECHNIQUES

The mapping of clinical problems to GenXAI methods (Table X) highlights the translational potential of explainable generative models across multiple medical domains. However, the practical realization of these applications is hindered by several unresolved challenges. While oncology and radiology have demonstrated relatively

mature adoption of GenXAI approaches, other domains such as dermatology, neurology, and cardiology reveal significant gaps in clinical validation, scalability, and explainability integration. These disparities underscore that the promise of GenXAI is not yet matched by consistent real-world deployment. To advance toward trustworthy clinical adoption, it is essential to critically examine the technical, clinical, and regulatory barriers that limit current progress. This section categorizes the key challenges into technical limitations of GenXAI architectures, domain-specific clinical constraints, and integration strategies that have been proposed to address them. By analyzing these dimensions, we aim to provide a structured understanding of the roadblocks in this field and the emerging techniques that attempt to bridge the gap between generative power and clinical interpretability.

3.1 Technical Challenges in Integrating GenAI and XAI

Integrating Explainable Artificial Intelligence (XAI) into Generative Artificial Intelligence (GenAI) architectures introduces several complex technical challenges, particularly in healthcare applications where transparency and reliability are essential for clinical decision-making. Advanced generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, and Transformer-based architectures operate within high-dimensional latent spaces that are inherently abstract and non-interpretable. As a result, understanding how these models derive predictions or generate synthetic outputs remains difficult for clinicians and researchers alike [2], [3]. Unlike traditional discriminative AI models that generate interpretable outputs such as class labels or probability scores, generative models produce synthetic samples, reconstructed patient records, clinical narratives, or probabilistic data distributions, thereby complicating the extraction of meaningful and clinically understandable explanations [6], [10].

Another major challenge lies in the adaptation of existing XAI methods to generative contexts. Widely adopted post hoc explanation techniques such as SHAP, LIME, saliency maps, and attention visualization were originally designed for classification and regression tasks, where outputs are comparatively straightforward to interpret [4], [5]. However, when applied to generative systems that produce complex outputs such as medical images, longitudinal time-series data, or synthetic electronic health records (EHRs), these techniques often fail to provide stable, faithful, and clinically relevant explanations [3], [8]. Furthermore, generative architectures are susceptible to issues such as training instability, mode collapse, hallucinated outputs, latent feature entanglement, and non-deterministic generation behavior, all of which negatively impact explanation consistency and reproducibility [2], [10].

In healthcare environments, these limitations are further amplified by the need for clinically trustworthy and regulation-compliant AI systems. Explanations generated by current XAI approaches frequently remain too technical, abstract, or visually ambiguous for practical use by healthcare professionals [13], [14]. For instance, heatmaps or feature attribution scores may indicate influential regions in medical imaging, yet they often fail to provide causal or clinically actionable reasoning behind predictions. Similarly, counterfactual explanation methods, while promising, still struggle to model realistic physiological variations and patient-specific clinical scenarios effectively [15]. As emphasized by Schneider [3], the future of GenXAI research requires moving beyond treating generative systems as opaque black-box models toward the

development of intrinsically interpretable architectures capable of providing transparent, human-centered, and medically meaningful reasoning processes.

Additionally, the lack of standardized evaluation frameworks for explainability in generative healthcare models remains a critical concern. Existing studies primarily focus on predictive performance while neglecting explanation quality, fairness, robustness, and clinician usability [5], [16]. This imbalance creates barriers to regulatory approval and real-world clinical adoption under healthcare standards such as HIPAA, GDPR, and FDA guidelines [7], [16]. Consequently, there is an urgent need for domain-adaptive explanation mechanisms, hybrid interpretable architectures, and clinically validated evaluation metrics that can ensure trustworthy deployment of explainable generative AI systems in healthcare risk assessment and prognosis prediction.

3.2 Clinical and Domain-Specific Challenges

From a clinical perspective, one of the most significant barriers to the adoption of Generative Artificial Intelligence (GenAI) in healthcare is the issue of trust and interpretability. In high-stakes medical environments, clinicians require AI systems that not only provide accurate predictions but also generate explanations that are clinically meaningful, transparent, and aligned with established medical reasoning processes. However, many existing Explainable AI (XAI) techniques remain technically oriented and often fail to deliver domain-specific insights that healthcare professionals can readily interpret and apply in real-time clinical decision-making [4], [13]. For example, explanation methods such as saliency maps, feature attribution scores, or attention heatmaps may highlight influential regions or variables, yet they frequently lack contextual medical reasoning and actionable interpretability necessary for critical applications such as intensive care monitoring, oncology prognosis, radiology, and emergency medicine [6], [14].

Another major concern is the presence of algorithmic bias and fairness-related challenges in GenAI systems. Since generative models are heavily dependent on training data distributions, datasets that are imbalanced, incomplete, or demographically skewed can lead to biased predictions and synthetic outputs that disproportionately affect underrepresented patient populations [2], [7]. Such biases may inadvertently reinforce existing healthcare disparities related to ethnicity, gender, age, socioeconomic status, or geographic accessibility. Ramachandranpillai et al. [2] attempted to address this issue through bias-transforming Generative Adversarial Networks (Bt-GANs), which aim to generate fairer synthetic healthcare data while preserving utility. Nevertheless, only a limited number of current GenAI frameworks actively integrate fairness-aware learning mechanisms alongside explainability and transparency objectives.

Clinical deployment is further complicated by challenges related to reliability, robustness, and generalizability across healthcare institutions. Models trained on specific datasets or hospital environments often fail to generalize effectively to diverse patient populations, resulting in reduced predictive consistency and lower clinician confidence [1], [10]. Moreover, healthcare data are inherently heterogeneous and multi-modal, involving imaging data, electronic health records (EHRs), genomic profiles, wearable sensor streams, and clinical narratives. Integrating these diverse modalities into a unified explainable generative framework remains a technically demanding task [8], [12].

Regulatory and ethical compliance also represent critical domain-specific barriers.

Healthcare regulations such as HIPAA, GDPR, and emerging FDA guidelines increasingly demand AI systems that are transparent, auditable, privacy-preserving, and capable of providing explainable reasoning behind automated decisions [1], [7], [16]. However, the opaque and probabilistic nature of many GenAI architectures limits their ability to satisfy these regulatory expectations. Existing models often lack traceability, reproducibility, and clinically verifiable justification mechanisms, making them difficult to validate in real-world healthcare settings [3], [16]. Furthermore, concerns regarding patient privacy, informed consent, accountability, and ethical use of synthetic medical data continue to hinder broader clinical acceptance of generative AI technologies.

In addition, there remains a substantial gap between technical explainability and practical clinical usability. While researchers frequently evaluate explanation quality using computational metrics, relatively few studies conduct clinician-centered validation or assess whether explanations genuinely improve diagnostic confidence, treatment planning, or patient outcomes [5], [15]. This disconnect highlights the urgent need for collaborative, interdisciplinary research involving clinicians, AI researchers, ethicists, and policymakers to develop domain-adaptive GenXAI systems that are not only technically robust but also clinically trustworthy, ethically responsible, and operationally deployable across diverse healthcare environments.

3.3 Integration Techniques for GenAI + XAI

Several techniques have been explored to bridge the gap between generative capabilities and interpretability:

- Post hoc XAI Techniques: These include adaptations of SHAP and LIME to generative contexts, surrogate models that approximate the behavior of GenAI systems, and saliency-based visualization methods. Chang et al. [4], for instance, successfully applied SHAP explanations to a sepsis mortality prediction model, improving clinical interpretability.
- Intrinsic Explainability: Some models embed explainability into the architecture itself. Examples include attention mechanisms in transformer-based generators, disentangled latent representations in VAEs, and conditional generation paths where latent variables are tied to observable features [5].
- Counterfactual Explanations: Counterfactuals offer actionable insights by showing minimal changes needed in input data to alter a prediction or outcome. When applied to GenAI, they help simulate alternative clinical scenarios and identify decision boundaries, improving clinician understanding.
- Hybrid Pipelines: A growing trend is to design pipelines that combine generative modeling with interpretability modules. For instance, Xiong et al. [5] proposed a semantic evaluation framework that aligns generated outputs with clinically meaningful interpretations using data mining tools.
- Multi-modal Fusion with Explanation Layers: Advanced systems now combine text, images, and EHR data into multi-modal GenAI models, augmented by explanation layers that reveal which modalities or features influenced the synthetic outputs [6].

3.4 Open Questions and Limitations

Although significant progress has been made in integrating Generative Artificial Intelligence (GenAI) with Explainable Artificial Intelligence (XAI), several unresolved questions and practical

limitations continue to hinder their widespread clinical adoption. These challenges are not only technical in nature but also involve issues related to clinical reliability, evaluation standardization, ethical governance, and real-world usability. Figure X conceptually summarizes the major unresolved concerns surrounding explainable generative healthcare systems.

A. Lack of Standardized Explainability Benchmarks

One of the most critical limitations in current GenXAI research is the absence of universally accepted evaluation benchmarks for assessing explanation quality in healthcare applications. Existing studies primarily focus on predictive performance metrics such as accuracy, sensitivity, and F1-score, while explanation fidelity, interpretability consistency, fairness, and clinician usability remain insufficiently evaluated. As a result, there is currently no standardized framework capable of objectively measuring whether a generated explanation is clinically meaningful, trustworthy, or actionable across different medical domains.

B. Ambiguity in Defining "Sufficient Explainability"

Another unresolved issue is the lack of consensus regarding what constitutes "sufficient explainability" in clinical AI systems. In healthcare settings, explanations must satisfy diverse stakeholders including physicians, radiologists, hospital administrators, regulators, and patients. However, the level of explanation detail required often varies across applications. For instance, probabilistic outputs such as survival curves, synthetic CT scans, disease progression simulations, or risk stratification scores may be mathematically interpretable yet remain difficult for clinicians to validate intuitively. This creates uncertainty regarding how much transparency is necessary to establish clinical trust and regulatory acceptance.

C. Limited Clinical Validation and Human-Centered Evaluation

Most current explainable generative models are evaluated in controlled experimental environments rather than real-world healthcare settings. Very few studies conduct comparative analysis between AI-generated explanations and expert physician reasoning processes. Similarly, usability studies involving clinicians, radiologists, or intensive care specialists remain limited. Consequently, it is still unclear whether existing explanation techniques genuinely improve clinical confidence, diagnostic efficiency, treatment planning, or patient outcomes in practice. The absence of longitudinal hospital-based validation significantly limits the translational readiness of GenXAI systems.

D. Challenges in Real-Time Explainability

Real-time explainability represents another major open challenge. Modern healthcare environments often require immediate and high-stakes decision-making, particularly in emergency medicine, intensive care units (ICUs), oncology, and surgical support systems. Current XAI techniques are generally computationally intensive and are not optimized for delivering instant, interactive, and context-aware explanations during live clinical workflows. Developing systems capable of providing just-in-time reasoning, dynamic visualization, and clinician-interactive explanations remains an unresolved research problem.

E. Ethical, Regulatory, and Accountability Concerns

The growing adoption of GenAI in healthcare also introduces unresolved

ethical and regulatory questions related to accountability, transparency, patient consent, and data governance. Existing generative systems frequently operate as opaque black-box architectures, making it difficult to determine responsibility in cases of incorrect predictions, biased outputs, or harmful clinical recommendations. Furthermore, healthcare regulations such as HIPAA, GDPR, and emerging FDA AI guidelines increasingly demand explainable, auditable, and privacy-preserving AI systems, yet most current GenAI frameworks still fall short of these expectations.

F. Future Need for Clinically Grounded GenXAI Systems

Collectively, these limitations emphasize the urgent need for next-generation GenXAI systems that are clinically grounded, ethically aligned, regulation-ready, and validated in real-world healthcare environments. Future research must focus on developing standardized interpretability benchmarks, physician-centered evaluation frameworks, fairness-aware generative architectures, and scalable real-time explanation mechanisms. Addressing these challenges will be essential to ensure the safe, transparent, and trustworthy deployment of explainable generative AI systems for healthcare risk assessment, prognosis prediction, and clinical decision support..

IV. FUTURE RESEARCH DIRECTIONS

Despite notable advancements in Generative AI (GenAI) and Explainable AI (XAI), their integration for clinical risk assessment and prognosis is still in its infancy. Unlocking their full potential in healthcare requires addressing several technical, ethical, and operational challenges identified in recent studies. Future work should emphasize the development of domain-specific explainable generative architectures, tailored for specialties such as oncology, cardiology, neurology, and dermatology. Unlike general-purpose models, these systems must embed medical reasoning pathways and structured clinical knowledge to improve interpretability, with conditional GANs, attention-guided VAEs, and diffusion models adapted to incorporate clinical hierarchies and ontologies. Another critical priority is the standardization of evaluation frameworks that can jointly assess prediction accuracy, explanation quality, fairness, and clinical usability. This includes benchmarking datasets, interpretability scoring metrics, and human-in-the-loop validation pipelines. Without consistent evaluation standards, reproducibility and trustworthiness remain limited. Hybrid models that combine the accuracy of GenAI with the transparency of symbolic reasoning or probabilistic frameworks also warrant further exploration, enabling systems that balance high-fidelity generative outputs with simplified, clinically meaningful explanations.

Emerging research should also investigate federated learning and privacy-preserving GenXAI approaches, which allow collaborative training across multiple institutions without sharing sensitive patient data. Such methods can address data scarcity, improve generalization across diverse populations, and meet stringent privacy regulations such as GDPR and HIPAA. Another promising direction is the advancement of multi-modal GenXAI frameworks capable of fusing heterogeneous data types—including imaging, text (clinical notes), genomics, and sensor-derived time series—within a single interpretable generative architecture. These systems could provide richer, patient-specific insights by reflecting the true complexity of medical decision-making.

Additionally, the growing role of large language models (LLMs) in medicine, such as GPT-4 and Med-PaLM 2, highlights the urgent need for explainability in generative outputs. LLMs are increasingly used for clinical note generation, triage recommendations, and summarization of patient histories. However, without transparent justification mechanisms, their predictions risk perpetuating errors or biases. Future work must explore integrating XAI with LLM-based clinical assistants, ensuring that generated recommendations are not only fluent but also clinically verifiable and ethically accountable. Finally, long-term success requires rigorous clinical validation and regulatory readiness. GenXAI systems must undergo prospective clinical trials, external validation, and continual fairness audits to meet the standards of regulatory bodies such as the FDA, EMA, and WHO. Collaborations with healthcare providers, ethics committees, and policymakers will be critical to move these systems from laboratory prototypes to trusted clinical tools.

V. CONCLUSION

Generative AI (GenAI) has emerged as a transformative technology in healthcare, capable of generating synthetic data, predicting patient trajectories, and enhancing diagnostic insights. However, its widespread adoption in critical clinical tasks such as risk assessment and prognosis is hindered by a fundamental lack of transparency. In this context, Explainable AI (XAI) plays a vital role in bridging the gap between powerful generative capabilities and the trust requirements of clinical decision-making. This review has presented a comprehensive overview of the recent advancements in GenAI and XAI, highlighting their complementary potential and the technical, ethical, and clinical challenges that arise when integrating them. Through an analysis of state-of-the-art literature from 2024–2025, we identified that most current systems rely heavily on post hoc explanation techniques, are domain-limited, and often lack empirical validation in real-world settings. Although hybrid and intrinsically explainable models show promise, significant gaps remain in evaluation standardization, user-centric design, regulatory compliance, and clinical trust. To move forward, future research must focus on developing domain-adapted architectures, real-time interpretability tools, fairness-aware generative pipelines, and robust evaluation benchmarks. Furthermore, interdisciplinary collaboration with clinicians, data scientists, ethicists, and policymakers will be essential in translating explainable generative models from research prototypes to dependable tools in clinical environments. By advancing the synergy between GenAI and XAI, this field holds the potential to not only enhance predictive performance but also ensure that AI-driven healthcare remains ethical, interpretable, and ultimately aligned with the goal of improving patient outcomes.

REFERENCES

[1] Bhuyan, S. S., Sateesh, V., Mukul, N., et al. (2025). Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *Journal of Medical Systems, 49*(1), 10. https://doi.org/10.1007/s10916-024-02136-1

[2] Ramachandranpillai, R., Sikder, M. F., Bergström, D., & Heintz, F. (2024). Bt-GAN: Generating Fair Synthetic Health Data via Bias-transforming Generative Adversarial Networks. *arXiv preprint*, arXiv:2404.13634. https://arxiv.org/abs/2404.13634

[3] Schneider, J. (2024). Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda. *Artificial Intelligence Review, 57*, 289. https://doi.org/10.1007/s10462-024-10916-x

[4] Chang, C.-H., Wang, X., & Yang, C. C. (2024). Explainable AI for Fair Sepsis Mortality Predictive Model. *arXiv preprint*, arXiv:2404.13139. https://arxiv.org/abs/2404.13139

[5] Xiong, H., Zhang, X., Chen, J., Sun, X., Li, Y., Sun, Z., & Du, M. (2024). Towards Explainable Artificial Intelligence (XAI): A Data Mining Perspective. *arXiv preprint*, arXiv:2401.04374. https://arxiv.org/abs/2401.04374

[6] Huang, G., Li, Y., Jameel, S., Long, Y., & Papanastasiou, G. (2024). From Explainable to Interpretable Deep Learning for Natural Language Processing in Healthcare: How Far from Reality? *arXiv preprint*, arXiv:2403.11894. https://arxiv.org/abs/2403.11894

[7] Okonji, O. R., Yunusov, K., & Gordon, B. (2024). Applications of Generative AI in Healthcare: Algorithmic, Ethical, Legal and Societal Considerations. *arXiv preprint*, arXiv:2406.10632. https://arxiv.org/abs/2406.10632

[8] Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., & Chen, H. (2024). Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks. *arXiv preprint*, arXiv:2410.02331. https://arxiv.org/abs/2410.02331

[9] Al Amin, K., Hasan, K., Zein-Sabatto, S., Chimba, D., Ahmed, I., & Islam, T. (2024). An Explainable AI Framework for Artificial Intelligence of Medical Things. *arXiv preprint*, arXiv:2403.04130. https://arxiv.org/abs/2403.04130

[10] Shokrollahi, Y., Yarmohammadtoosky, S., Nikahd, M. M., Dong, P., Li, X., & Gu, L. (2023). A Comprehensive Review of Generative AI in Healthcare. *arXiv preprint*, arXiv:2310.00795. https://arxiv.org/abs/2310.00795

[11] John Snow Labs. (2025). Generative AI in Healthcare: Use Cases, Benefits, and Challenges. https://www.johnsnowlabs.com/generative-ai-healthcare/

[12] Nature Digital Medicine. (2024). A review on generative AI models for synthetic medical text, time series, and longitudinal data. https://www.nature.com/articles/s41746-024-01409-w

[13] Elsharkawy, M., El Sayed, A. A., & El-Khatib, M. (2024). Explainable Artificial Intelligence in Radiology: Current Trends and Future Perspectives. *The Egyptian Journal of Radiology and Nuclear Medicine, 55*, 73. https://doi.org/10.1186/s43055-024-01356-2

[14] Teoh, J., Ho, J., & Liu, Y. (2024). Explainable AI for Medical Applications: A Review of Techniques and Challenges. *Heliyon, 10*(4), e15768. https://doi.org/10.1016/j.heliyon.2024.e15768

[15] Shahin, A., Dey, L., & Ghosh, S. (2024). Counterfactual Explanations in Medical AI: A Practical Framework. *PLOS Digital Health, 3*(1), e11048122. https://doi.org/10.1371/journal.pdig.0000123

[16] Hassan, M., Sharma, D., & Agrawal, R. (2024). Generative AI for Healthcare: Model Validation and

Regulatory Alignment. *Journal of Medical Internet Research, 26*, e53008. https://doi.org/10.2196/53008