International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

https://gssrr.org/index.php/JournalOfBasicAndApplied/index

Generative AI and Explainable AI in the Field of the Health Industry: Application, Ethical Considerations and Future Challenges

Mitha Rachel Jose[a]*, Tom Tuunainen[b], Olli Isohanni[c]

[a]Laurea University of Applied Sciences, Department of Business Information Technology, Vanha maantie 9, 02650 Espoo, Finland

[b,c]Centria University of Applied Sciences, Research and Development Operations, Talonpojankatu 2, 67100 Kokkola, Finland

[a]Email: mitha.jose@laurea.fi,[b]Email: tom.tuunainen@centria.fi,[c]Email: olli.isohanni@centria.fi

Abstract

The emergence of Generative Artificial Intelligence (GenAI) and Explainable Artificial Intelligence (XAI) represents a paradigm shift in the digital transformation of the healthcare industry. GenAI models, including generative adversarial networks (GANs), variational autoencoders (VAEs) and large language models (LLMs), can generate synthetic yet realistic medical data. This facilitates innovations in medical imaging, drug discovery and personalized treatment planning. XAI focuses on ensuring the transparency, interpretability and trustworthiness of AI models. The most important factors in clinical applications are where human lives are at risk. The integration of GenAI and XAI has the potential to improve clinical decision support systems, optimize workflows, and enhance healthcare outcomes, while addressing critical issues such as data scarcity, bias, and patient privacy. However, combining them also presents substantial technical, ethical, and regulatory challenges, including the need for interpretable generative models, adherence to data protection laws such as the General Data Protection Regulation (GDPR), and ensuring fairness and accountability in automated decision-making. This paper explains a comprehensive analysis of the applications and implementation frameworks of GenAI and XAI in healthcare, emphasizing their role in fostering trustworthy, transparent and patient-centered AI solutions.

Keywords: GENAI; XAI; LLM.

Received: 12/21/2025
Accepted: 2/21/2026
Published: 3/2/2026

----

* *Corresponding author.*

1. Introduction

The healthcare industry has been at the forefront of the most promising applications of artificial intelligence (AI), which has seen a rapid advancement. According to [1], generative AI presents both opportunities and regulatory challenges on governance and society and [2] emphasize that XAI is essential for bridging the gap between model performance and human understanding, particularly in high-stakes decision-making systems. Generative AI is a class of machine learning model that can create new content, such as synthetic medical images, patient data, text, and molecular structures, based on patterns learned from existing datasets. Since the advent of advanced models such as LLMs and GANs, GenAI has demonstrated remarkable potential in automating medical documentation, assisting diagnostics, supporting clinical communication, and accelerating drug discovery. These applications are reshaping the healthcare landscape by improving efficiency, reducing administrative burden, and fostering personalized and predictive care.

These advancements also introduce significant challenges concerning accuracy, bias, data privacy and clinical safety. The complexity of these models raises ethical and legal concerns, especially when AI-generated outputs influence critical medical decisions which are called black box nature. The black box nature of AI systems refers to the fact that they are so complex that it is difficult to understand how they work. This makes it difficult for clinicians to understand how specific recommendations or predictions are made, potentially undermining trust and accountability. This is where XAI becomes truly important. The objective of XAI is the making of the decision-making processes of AI systems transparent and interpretable, with the enabling of verification, validation, and justification of algorithmic outcomes by healthcare professionals. When doctors use technology in their work, it is important that they can understand how it works. This is not just a technical choice, but something that is also needed for ethical and legal reasons. [3] explains that to keep patients safe, the professional responsibility of doctors and nurses, and following rules such as the EU AI Act and the GDPR [4] impose legal and ethical constraints on AI systems, influencing model design, data governance, and algorithmic transparency.

The growing recognition of XAI's importance means that healthcare systems continue to face considerable obstacles in integrating it. These obstacles include balancing interpretability with predictive performance, embedding XAI tools into clinical workflows, and ensuring adherence to strict regulatory and ethical standards. The intersection of GenAI and XAI represents an emerging research frontier that combines the creative power of generative models with the transparency of explainable systems could enable the development of trustworthy, human-centred, and legally compliant AI solutions. These advancements aim to bridge the gap between cutting-edge AI innovation and the practical, accountable deployment required for real-world medical decision-making.

This paper provides a detailed analysis of the applications, ethical considerations and potential future challenges of GenAI and XAI in the healthcare sector. It examines how these technologies can jointly advance clinical decision-making, data management, and medical innovation while addressing crucial issues such as model transparency, accountability, and bias mitigation. The study also explores new perspectives on responsible and ethical AI, highlighting emerging trends and offering strategic insights for researchers, practitioners, and policymakers. By bridging innovation with interpretability, this study aims to contribute to the ongoing transformation toward transparent, reliable, and patient-centred AI-driven healthcare systems.

2. Methodology

This literature review aimed to identify and summarise relevant studies, with a particular focus on those examining the reliability of AI systems in healthcare. The concept of trustworthy AI extends beyond technical performance. It includes principles of transparency, reliability, safety, fairness, accountability, and ethical alignment. The establishment of trust is a fundamental prerequisite when implementing both GenAI and XAI in clinical contexts, where AI outputs can bring a direct influence on diagnostic accuracy, treatment decisions, and the ultimate patient outcomes.

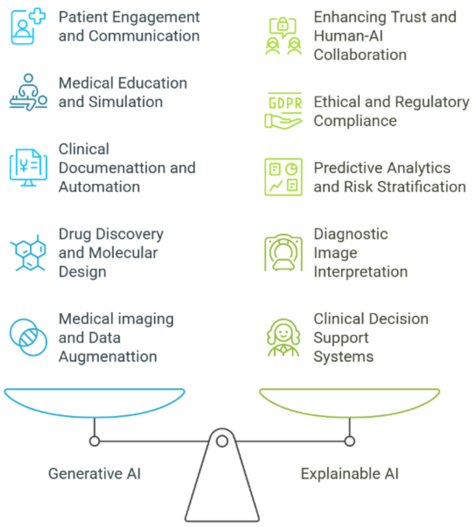
Patient Engagement and Communication
Enhancing Trust and Human-AI Collaboration
Medical Education and Simulation
GDPR
Ethical and Regulatory Compliance
Clinical Documenattion and Automation
Predictive Analytics and Risk Stratification
Drug Discovery and Molecular Design
Diagnostic Image Interpretation
Medical imaging and Data Augmenattion
Clinical Decision Support Systems
Generative AI
Explainable AI

Figure 1: Integration of Generative AI and Explainable AI in the field of the health industry

In GenAI, trustworthiness is primarily linked to the authenticity and validity of generated content. For example, the use of synthetic medical data and AI-generated diagnostic images require rigorous methods to ensure that the produced results are both clinically accurate and free from bias or misrepresentation. This study emphasizes the importance of preserving data integrity, privacy protection, and the implementation of explainable data generation processes, enabling healthcare professionals to confidently utilize AI-generated information without compromising patient safety or ethical standards. XAI plays a crucial role in enhancing interpretability and human understanding of complex AI models. Trustworthy XAI systems help clinicians to understand the reasons behind a particular decision or prediction by providing clear explanations. This not only facilitates more informed clinical judgments but also supports regulatory compliance and ethical accountability in line with international standards, including the EU AI Act, GDPR, and the World Health Organization (WHO) guidance on ethics and governance of AI in health.

By integrating the innovative potential of GenAI with the transparency and interpretability of XAI, trustworthy AI in healthcare emerges as a multidimensional framework balancing innovation with responsibility. It ensures that AI-driven systems are technically robust and scientifically valid, but also socially acceptable and ethically sustainable. This synthesis of generative capability and explainable reasoning forms the foundation of responsible AI adoption, fostering a healthcare ecosystem where AI acts as a reliable partner to clinicians and patients, rather than as an opaque or autonomous entity.

This study ensures the critical importance of fostering trustworthy AI within both GenAI and XAI applications in healthcare, which highlights the need to develop technologies that are transparent, ethically responsible, and human-centred, that are essential prerequisites for establishing and maintaining long-term trust in AI-driven medical systems.

3. Different GenAI and XAI applications

GenAI and XAI are increasingly shaping innovation in the healthcare sector. GenAI supports clinical and research workflows by generating new insights and data patterns, as well as potential therapeutic solutions. Meanwhile, XAI ensures that these AI-driven decisions remain transparent, interpretable and trustworthy for medical professionals. The article provides an explanation of some of the instances in the different areas.

3.1. GenAI applications in healthcare

A subset of AI models that can create new, realistic data or content from existing datasets is known as Generative AI. These models, powered by architectures such as GANs, VAEs, and LLMs (e.g., GPT-4, Med-PaLM), have shown significant potential in enhancing data-driven medical research and clinical practice.

Medical imaging and data augmentation

GenAI is widely used for synthetic medical image generation, helping to overcome data scarcity and class imbalance in training datasets [5]. For example, GANs can generate realistic MRI, CT, or X-ray images to augment existing datasets, especially for rare diseases, improving diagnostic model accuracy without additional patient data collection [6]. In radiology, a GAN-based model [7] can synthesize new lung CT scans resembling real patient scans, which can then be used to train diagnostic AI systems to detect early signs of pulmonary fibrosis or tumours. [8] explains GANs are a useful tool for creating medical images and increasing the amount of data available, which can help solve problems related to having limited data and a lack of diversity.

Drug discovery and molecular design

GenAI is transforming drug discovery and molecular design by creating new molecular structures with specific biological properties. It can rapidly generate drug candidates that can effectively bind to target proteins by using advanced techniques such as variational autoencoders and diffusion models. This capability is evident in platforms such as DeepMind's AlphaFold and Insilico Medicine's GenAI systems, which employ generative modelling to predict protein folding patterns and design potential therapeutic compounds. This substantially reduces research and development time, as well as associated costs [6,7].

Clinical documentation and automation

Large language models such as GPT-4 and Med-PaLM are also transforming clinical documentation and automation by generating summaries of patient records, discharge reports and clinical letters from structured and unstructured data within electronic health records, thereby streamlining administrative and reporting tasks. Physicians, for example, can dictate notes into a voice-to-text system integrated with an LLM. The LLM then processes, summarises and formats the content automatically, producing a standardised medical report, thereby improving accuracy and efficiency [49].

Medical education and simulation

In medical education and simulation, generative AI facilitates the creation of interactive learning platforms and virtual patient scenarios, providing medical trainees with realistic, adaptive, case-based exercises. For instance, AI systems can simulate diverse clinical situations, such as cardiac emergencies, enabling students to practise diagnostic reasoning in a safe, controlled digital environment [50].

Patient engagement and communication

GenAI helps patients by talking to them and giving them information. Treatment plans can be explained by these AI-driven chatbots, medication reminders can be sent, and patient queries can be responded to, thereby fostering continuous and informed interaction. One such example is a virtual health assistant that has been developed for diabetic patients [12].

3.2. XAI applications

XAI includes methods that make AI models easier to understand, so human experts can see how decisions are made. In healthcare, this is important for building trust and ensuring accountability, as well as meeting ethical and legal requirements. XAI transforms AI systems from opaque 'black box' models into more transparent ones. While simple models such as decision trees or linear regression are easy to interpret, complex models such as deep learning are more difficult to explain and only partially transparent [13,14,15]. The more transparent the process, the better people understand how models reach specific conclusions, thereby increasing trust. This trust can manifest at two distinct levels. Trust in the model itself, which is of primary importance to data scientists, and trust in the model’s predictions, which is crucial for clinicians and patients [14]. These two dimensions of trust require tailored approaches. The provision of explanations for individual predictions serves to enhance prediction-level trust, while the examination of multiple explanations across cases serves to reinforce trust in the model’s overall reliability. [15] categorise XAI methods can be divided into two main types: transparent and post-hoc models. Transparent models, like linear regression, decision trees, and Bayesian models, are easy to understand because their structure is clear. The post-hoc methods are used to explain complex black-box models, such as neural networks and deep learning systems. These models need extra layers of interpretation to make it clear how they work. These post-hoc approaches employ techniques such as textual explanations, visual interpretations, example-based reasoning, simplification, and feature relevance mapping to enhance interpretability. [16] identified four foundational principles for effective XAI systems:

- Explanation – providing evidence and rationale for outcomes and processes,
- Meaningfulness – ensuring explanations are understandable to specific users,

- Explanation accuracy – ensuring explanations accurately represent model behaviour, and
- Knowledge limits – ensuring the model operates within its designed boundaries and confidence levels [16].

However, [17] that post-hoc explanations may not always guarantee fairness or reliability in critical areas such as healthcare. They should be used alongside thorough validation and ethical checks, rather than as a replacement for them.

3.3. Collaboration between humans and AI

Effective healthcare decision-making requires a combination of human expertise and AI intelligence, rather than replacing human judgment with algorithms [18]. While AI systems may outperform clinicians in specific diagnostic or predictive tasks, human professionals contribute indispensable domain expertise, contextual understanding, and ethical reasoning [14].

In clinical practice, the combination of human decision-making and AI support commonly referred to as AI-assisted decision-making, often yields the most accurate and reliable outcomes. This collaboration improves diagnostic accuracy and increases transparency, allowing clinicians to understand and verify AI recommendations using their expertise [18]. The explainability in clinical decision-support systems (CDSS) enables developers to identify system limitations and ensures clinicians maintain confidence in AI-augmented decisions. Similarly, placing too much reliance on opaque, non-explainable systems could result in patients becoming passive recipients of algorithmic outputs, thereby undermining ethical and participatory medical practice. Therefore, it is essential to promote trust, accountability, and shared decision-making in the healthcare ecosystem by using explainable and collaborative AI frameworks.

Clinical decision support systems

XAI enhances the transparency of AI-powered decision support systems by providing reasoning explanations behind recommendations. [38] explains how XAI is used in healthcare, focusing on areas like disease diagnosis, predictive analytics, and personalized treatment. It examines different XAI methods, including model-agnostic tools like LIME and SHAP, interpretable deep learning models, and healthcare-specific applications. It also considers ethical issues, such as accountability and reducing bias, which are important for fair healthcare. XAI helps clinicians and AI work together, making AI decisions easier to understand. The study shows that XAI is key to using AI in healthcare, turning complex AI models into clear tools that help decision-making, build trust, and support ethical practices. The study also evaluates the ethical implications associated with AI in healthcare, such as accountability and bias mitigation, highlighting how XAI can contribute to addressing these issues.

Diagnostic image interpretation

Explainable models help clinicians interpret AI-generated imaging results. [39] shows that XAI techniques like Grad-CAM help make deep learning models in medical diagnostics more transparent. Grad-CAM highlights the parts of an image that most influence the model’s prediction, turning a “black box” into an understandable tool. For example, in pneumonia detection from X-rays, Grad-CAM creates a heatmap showing the lung areas the AI focuses on, allowing radiologists to check if the model's attention matches clinically important regions. This improves trust, accountability, and diagnostic confidence. [40] improved brain tumor detection from MRI scans by combining a CNN model (Xception-based) with Grad-CAM and SHAP. The model classified images into four categories: glioma, meningioma, pituitary tumor, and non-tumor, making predictions easier to interpret.

Grad-CAM was used to generate heat maps that localize the image regions most influential in each prediction. SHAP was used to quantify feature contributions, giving insight into which image-based features were driving decisions. The model achieved very high accuracies; ~99.95% on training, ~99.08% on validation, and ~98.78% on test sets. The dual XAI approach (Grad-CAM + SHAP) was shown to enhance transparency and help address potential bias, thereby increasing trustworthiness of the model for potential clinical use. All these methods substantially increase the transparency and trustworthiness of the AI system in a clinical diagnostic setting. By analysing misclassified cases, the study identifies avenues for improvement (e.g., handling noise and ambiguous feature overlap) and underscores that interpretability may serve as a key enabler for real-world adoption of AI in healthcare.

Predictive analytics and risk stratification

XAI provides interpretability in AI models used for disease prediction or patient risk assessment. [41] present a machine-learning and data-analytics framework for patient risk stratification designed to support proactive clinical decision-making. The authors employ a structured pipeline combining data preprocessing, feature engineering from electronic health records (EHR) and demographic sources, and supervised learning algorithms to classify patients into risk tiers (e.g., low, medium, high). The study evaluates multiple classifier models, comparing performance metrics such as accuracy, Area Under the ROC Curve (AUC), sensitivity, and specificity. The authors also discuss how analytic insights can aid healthcare providers in identifying high-risk individuals earlier, optimising resource allocation, and informing preventive interventions. The results demonstrate that the proposed approach achieves strong discriminative ability across risk categories, suggesting its potential for deployment in clinical or hospital-system settings.

The authors highlight how data analytics and machine learning integration can contribute to improved health outcomes and system efficiency. [39] carried out a thorough review of attribution-based explainability methods in the context of computer-vision applications in the field of medical imaging. The review focuses on how deep learning models used for tasks like disease classification, segmentation, and detection can be made interpretable through attribution-based explainable computer vision (X-CV) methods. The paper examines methods such as saliency maps, Grad-CAM, Integrated Gradients, SHAP, and concept-vector approaches, and evaluates their suitability, limitations, and clinical implications in medical imaging contexts. The study also considers how these methods map model decisions to visually or conceptually meaningful features and discusses the gap between purely predictive performance and the interpretability or trustworthiness required for healthcare deployment.

The review emphasises the need for more robust evaluation of explanation methods, better alignment with clinical needs, and increased attention to how explanations affect trust, accountability, and safety in medical AI settings. The explanation methods must be evaluated not just for visualization quality, but also for clinical relevance, repeatability, and the trust impact on users. There's a need to integrate explanation mechanisms early in model design ("explainability-by-design") rather than as purely post-hoc add-ons. The performance interpretability trade-off remains real, highly complex models may produce better predictions but be harder to explain, and many explanation tools themselves may mislead or oversimplify. [43] explains that the department of oncology generates a large amount of data, making it well-suited for big data models and predictive analytics. It reviews how predictive modelling has been used to assess patient risk, grouping applications into three areas: (1) population health management, (2) radiomics (quantitative imaging analysis), and (3) pathology. The study also explores opportunities for predictive analytics in clinical decision support and genomic risk assessment. The key challenges include limited data availability, lack of prospective validation, and the risk of spreading bias from existing datasets. If these issues are addressed, predictive methods could improve cancer risk stratification. The study also gives an idea of different use cases for their data sources, such as electronic health records, imaging, and biopsy data. The analysis of the study showcased how predictive algorithms have been developed, often leveraging large-scale EHR, claims, genomic, imaging, and other data sets and highlighted factors such as model performance metrics, operational integration, and the scope of risk stratification (e.g., mortality, hospitalisation, acute utilisation).

The further review constraints such as missing data, limited prospective validation, data interoperability challenges, and bias. If these challenges are addressed, computational methods will rapidly advance care for patients with cancer by enabling earlier identification of risk, more personalized intervention strategies, and improved outcomes.

Ethical and regulatory compliance

XAI supports transparency and accountability, aligning healthcare AI applications such as the EU AI Act, GDPR, and the WHO's AI ethics guidelines. [44] addresses the growing imperative of XAI within the healthcare domain, particularly focusing on the interplay between explainability and regulatory compliance. It argues that while AI systems in healthcare offer significant benefits, their black-box nature poses ethical, legal, and operational risks especially around accountability, transparency, bias, and patient safety. The study proposes that bridging the gap between technical explainability and regulatory frameworks is essential for ethical deployment of AI in clinical settings. Through analysis of explainability techniques, regulatory landscapes (including privacy laws, medical device regulation, and AI-specific legislation), and stakeholder implications, the study outlines strategies for implementing XAI in a compliant and ethically grounded manner. It highlights the role of interdisciplinary collaboration, user-centric design, and ongoing governance to ensure AI systems are trustworthy, transparent, and aligned with healthcare values. [45] focuses on the provenance, methods and potential of XAI within healthcare, particularly in building clinician trust, improving patient understanding, and satisfying regulatory and ethical obligations. It covers techniques such as LIME, SHAP, and Grad-CAM, discusses their roles in medical imaging and decision-making contexts, and analyses frameworks like GDPR and the U.S. Health Insurance Portability and Accountability Act (HIPAA) in relation to transparent and compliant AI systems. The study advocates that XAI should serve as a foundational principle for safe, ethical development of healthcare AI. The study also reviews the ethical and regulatory frameworks (e.g., GDPR, HIPAA) to map how transparency and explanation requirements intersect with legal compliance. The approach involves summarising existing methods, benchmarking their relative strengths or weaknesses, and discussing trends and future directions.

Enhancing trust and human–AI collaboration

XAI builds trust between humans and AI systems by providing clinicians and patients with clear insights into how these systems work. It makes predictive models in healthcare more transparent, helping doctors and patients to understand AI-based decisions [33]. The improvement of accuracy and the addressing of ethical concerns, which is essential for trust, is achieved by the comparison of AI outputs with clinical expertise [34]. The integration of XAI is expected to lead to enhanced diagnosis and treatment strategies, improved patient engagement, and provide deeper insights into algorithmic decision-making. These advancements collectively contribute to a more trustworthy and understandable healthcare AI landscape [35,36]. [37] highlights the critical need for transparent and understandable AI in healthcare to build trust and accountability. It proposes a conceptual framework linking trust to explanation characteristics and points towards future research directions to achieve trustworthy XAI. The goal is to create clear and understandable decision-making systems in healthcare, where trust is built through XAI features. This aims to improve accountability, trust, and acceptance of AI in areas affecting human well-being. The study focuses on the concept of trust within AI healthcare systems and proposes a framework that outlines the relationship between trust and the characteristics of explanations provided by AI.

4. The synergy of GenAI and XAI for trustworthy healthcare AI

This study explores how AI can be developed and implemented in a safe, ethical and trustworthy manner within the healthcare sector. It examines the regulatory frameworks of the EU and the US, addressing ethical issues such as bias, fairness, transparency, accountability and liability in healthcare applications. The study focuses on the lifecycle approach, emphasising the importance of managing AI systems throughout their entire lifespan from design and development to deployment and monitoring while distinguishing between the responsibilities of developers and implementers. To operationalise ethical and regulatory compliance, the authors propose two structured questionnaires to guide developers and healthcare implementers through key legal, ethical and organisational considerations. The study confirms several hypotheses, including the feasibility of developing such tools based on regulatory analysis, the differing obligations of developers and implementers, and the necessity of a multidisciplinary approach integrating ethical, legal, and technical perspectives. The questionnaires are presented as practical instruments for use in health technology assessment, public procurement, accreditation and professional training. The authors also observe that existing regulatory regimes, such as the EU AI Act, are evolving but remain incomplete in areas such as liability and accreditation. The future directions emphasise integrating ethical and legal principles into AI design from the outset, enhancing stakeholder training, and promoting the broader adoption of system-level frameworks for trustworthy AI in healthcare. This study successfully bridges the gap between high-level regulatory and ethical guidance and practical implementation by providing a structured approach to the safer and more reliable deployment of AI in healthcare systems [46].

Reference [47] explores how trust can be established, maintained and evaluated in the context of GenAI, which offers theoretical and empirical insights into the mechanisms that support trustworthy AI systems. It identifies emerging trends and addresses the technical, ethical and organisational challenges involved in the responsible deployment of GenAI. Crucial issues covered in the discussion include algorithmic bias, privacy protection, the opacity of AI models, the accountability of different stakeholders, and alignment with societal values. Trust is examined in terms of transparency, explainability, reliability, fairness and ethical governance, paying particular attention to how these aspects manifest in generative applications such as creative content production and conversational systems. The study describes that building trust in GenAI requires a collaborative, interdisciplinary effort in which developers, users, regulators and organisations share responsibility for embedding trust mechanisms throughout the entire system lifecycle. The consistent patterns are observed, including the growing demand for transparency and fairness, and the recognition that regulatory and governance structures are not yet fully keeping pace with technological advances [42]. Therefore, establishing robust infrastructures for auditability, traceability, and oversight is essential, as this highlights that trust is not only a technical concern, but also a socio-technical and organisational one. The future developments should prioritise the empirical validation of trust frameworks in real-world contexts and explore how trust evolves as AI systems become more autonomous and interactive. There is a need for greater attention to be given to governance models, liability structures, and policy standards to ensure responsible innovation. The broader scope includes applications across sectors such as education, the creative industries, business and public services, underscoring the importance of cross-disciplinary collaboration [47].

Reference [48] states that the integration of model-informed drug development (MIDD) and AI are an approach to speed up pharmaceutical innovation. It demonstrates how the combination of mathematical modelling techniques employed in pharmacokinetics, pharmacodynamics and clinical development, alongside AI methods such as machine learning, deep learning and GenAI, can improve predictive accuracy, optimise candidate selection and streamline dosage and trial strategies. The discussion outlines how MIDD simulates complex biological and clinical processes while AI identifies data patterns, designs novel molecules and enables virtual trials. These approaches strengthen the reliability of drug discovery and development by merging mechanistic and data-driven insights. The study also addresses key challenges, including data quality, model interpretability, algorithmic bias, and the need for standardisation and regulatory adaptation. The findings suggest that this integration could transform drug development by improving efficiency, reducing costs, and advancing personalised medicine. Despite its potential, the field is still in its infancy, and further work is needed to establish validation protocols, harmonise data standards, and develop regulatory frameworks that support AI-driven methodologies. It will be very important for experts in pharmacometrics, AI, clinical research, and regulatory science to work together, which will help to make most use of this approach. It will create more effective therapies, virtual trials, and new ways to develop drugs. These will benefit both industry and patients.

5. Challenges and ethical considerations

The use of GenAI and XAI in healthcare presents many challenges and ethical issues that impact the reliability and trustworthiness of AI systems. GenAI raises concerns about the authenticity of data, the propagation of bias, and the spread of misinformation, particularly when synthetic medical data or AI-generated diagnostic content is used without adequate validation [9,10,11]. XAI has problems balancing transparency and model complexity [13]. If the model is too simple, it may not work well. Then, if it is too complex, it can be difficult to understand and use. The problems like patient privacy, data ownership, fairness, and informed consent make using AI in healthcare more difficult. Therefore, ensuring trustworthy AI requires adherence to the principles of transparency Reference [19], reliability, safety and human oversight, in line with global frameworks such as the EU AI Act, the GDPR and the WHO's ethical guidelines for AI in health. It is very important to deal with these problems to make the public believe in AI-driven healthcare systems and ensure that they are strong and fair. A few of the challenges and ethical considerations are explained as follows.

5.1. Legal and regulatory compliance

XAI plays a vital role in ensuring compliance with global regulatory frameworks that mandate transparency and accountability in AI systems [9]. Within the European Union, the GDPR [4] establishes transparency as a key principle in data processing [3]. However, scholars such as [19] argue that transparency alone is insufficient for fostering trust, emphasizing instead a relational model where transparency is contextual and dependent on trustworthiness between users and system providers. The EU AI Act, which is soon to be implemented, will categorise AI systems as either high-risk or low risk. Those deemed high-risk, including medical AI systems, must adhere to stricter regulations, such as ensuring proper documentation, traceability [11], human oversight [10] and reliability.

Similar privacy frameworks exist globally, including in the HIPAA [30], which promotes openness in health data handling [31], and China's Personal Information Protection Law (PIPL) [32], which mandates transparency in data usage and algorithmic processes. These regulatory developments show the growing importance of XAI in achieving lawful, ethical, and transparent AI integration in healthcare.

5.2. Privacy and security

XAI improves the safety and reliability of AI systems by enhancing error detection and interpretability, it also introduces new privacy and security challenges [28]. For example, excessive transparency may inadvertently expose sensitive data or system vulnerabilities, which could facilitate model manipulation or adversarial attacks. To balance explainability with confidentiality, researchers advocate the integration of privacy-preserving machine learning methods, such as Federated Learning (FL) and differential privacy. The emerging concept of Explainable Security (XSec) [29] applies XAI principles to cybersecurity, emphasising stakeholder collaboration and resilience against input manipulation, bias and inaccurate predictions. In healthcare, explainability aids the detection of such vulnerabilities and enhances clinicians' ability to identify bias, improve fairness, and safeguard the integrity of patient data.

5.3. Explainability and trust

The primary goal of XAI is to enhance user trust, however, explainability does not always directly translate to increased confidence. [26] reveals that multi-modal explanations combining graphical, contextual, and narrative insights can significantly improve trust, yet contextual factors also influence user perception. In remote AI systems Reference [27], where users interact via APIs, the reliability of explanations may diminish, as systems could misrepresent reasoning processes. Within healthcare, achieving the right level of trust is critical, under-trust can lead clinicians to dismiss accurate outputs, while over-trust can result in excessive dependence on flawed recommendations. Transparent education on AI capabilities and limitations is therefore essential to calibrate appropriate trust among healthcare professionals and patients.

5.4. Balancing explainability and accuracy

There is a trade-off between transparency and accuracy. The simple models like linear regression might be less accurate, while complex models like deep neural networks are more accurate but harder to understand [24]. In healthcare contexts, where precision can directly impact patient safety, accuracy is frequently prioritized over interpretability. [25] explains that policymakers and system designers must engage with public and professional stakeholders to determine context-specific balances between explainability and performance, recognizing that these preferences may vary across medical applications and societal expectations.

5.5. Measuring explainability

Metrics such as sensitivity or AUC can be used to quantify accuracy, but explainability is not so straightforward. This is because it is inherently subjective and context dependent. Several frameworks have been proposed to operationalize its measurement. [20] introduced Explainability Fact Sheets assessing functionality, usability, and safety. [21] focuses on simulatability, decomposability, and algorithmic transparency, while [22] propose evaluating users' "mental models" of system understanding. [23] extends these approaches by assessing comprehensibility, granularity, faithfulness, and user relevance, offering a multidimensional basis for benchmarking XAI systems. These frameworks emphasize that explainability should be evaluated from both human and technical perspectives.

6. Future directions and conclusion

The future of GenAI and XAI in healthcare lies in creating transparent, ethical and human-centred systems that enhance clinical decision-making. It is anticipated that forthcoming studies will concentrate on combining the creative capabilities of GenAI with the interpretability of XAI. This integration will result in intelligent systems that can produce novel medical insights and present substantiated, evidence-based rationales for their conclusions. The development of regulation-compliant, bias-resilient, and privacy-preserving AI frameworks is set to be a key focus in future directions. It will be very important to make sure that it follows important principles of AI, such as fairness, accountability, and transparency. The increased collaboration among clinicians, data scientists, and policymakers will be crucial for translating AI innovations into safe, explainable, and patient-centred healthcare solutions. The combination of GenAI and XAI based on trust and ethical responsibility has the potential to change the healthcare landscape, leading to a future where intelligent systems work as reliable partners in clinical excellence and compassionate care.

References

[1] E.N. Cawood, M. Vespe, A. Kotsev and R. Bavel. Generative AI outlook report – Exploring the intersection of technology, society, and policy. Publications Office of the European Union, 2025, https://data.europa.eu/doi/10.2760/1109679

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Ser, N. Díaz-Rodríguez and F. Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, Volume 99, 2023, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2023.101805

[3] M. Kritikos. The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. European Parliamentary Research Service, 2020, ISBN 978-92-846-6771-0, https://doi.org/10.2861/293

[4] Intersoft consulting. General Data Protection Regulation (GDPR). 2025, https://gdpr-info.eu/

[5] F. Garcea, A. Serra, F. Lamberti and L. Morra. Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine, Volume 152, 2023, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2022.106391

[6] G. M. Harshvardhan, M. K. Gourisaria, M. Pandey and S. S. Rautaray. A comprehensive survey and analysis of generative models in machine learning. Computer Science Review, Volume 38, 2020, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2020.100285

[7] X. Wang and H. Liu. Data supplement for a soft sensor using a new generative model based on a variational autoencoder and Wasserstein GAN. Journal of Process Control, Volume 85, 2020, Pages 91-99, ISSN 0959-1524, https://doi.org/10.1016/j.jprocont.2019.11.004

[8] D. Paulson and L. Victor. Generative Adversarial Networks (GANs) for Medical Image Synthesis and Data Augmentation. Preprints, 2025, https://doi.org/10.20944/preprints202506.1310.v1

[9] I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel and B. Ogbuokiri. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. Informatics in Medicine Unlocked, Volume 51, 2024, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2024.101587

[10] C. Giorgetti, G. Contissa and G. Basile. Healthcare AI, explainability, and the human-machine relationship: a (not so) novel practical challenge. Front Med (Lausanne), 2025, https://doi.org/10.3389/fmed.2025.1545409

[11] L. Weber, S. Lapuschkin, A. Binder and W. Samek. Beyond explaining: Opportunities and challenges of XAI-based model improvement. Information Fusion, Volume 92, 2023, Pages 154-176, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2022.11.013

[12] S. A. Rabbani, M. El-Tanani, S. Sharma, S. S. Rabbani, Y. El-Tanani, R. Kumar and M. Saini. Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. BioMedInformatics, 2025, https://doi.org/10.3390/biomedinformatics5030037

[13] T. Hulsen. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. AI, 2023, pp. 652-666, https://doi.org/10.3390/ai4030034

[14] M.T. Ribeiro, S. Singh and C. Guestrin. “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144, https://doi.org/10.48550/arXiv.1602.04938

[15] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina and R. Benjamins. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 2020, 58, pp. 82–115, https://doi.org/10.48550/arXiv.1910.10045

[16] P.J. Phillips, C.A. Hahn, P.C. Fontana, D.A. Broniatowski and M.A. Przybocki. Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology, Gaithersburg, MD, USA, Volume 18, 2020, https://doi.org/10.6028/NIST.IR.8312

[17] D. Vale, A. El-Sharif and M. Ali. Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI Ethics 2, pp. 815–826, 2022, https://doi.org/10.1007/s43681-022-00142-y

[18] J. Amann, A. Blasimme, E. Vayena, D. Frey and V. I. Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20, 310, 2020, https://doi.org/10.1186/s12911-020-01332-6

[19] H. Felzmann, E.F Villaronga, C. Lutz and A. Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 2019, https://doi.org/10.1177/2053951719860542

[20] K. Sokol and P. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Association for Computing Machinery, New York, USA, pp. 56–67, 2020, https://doi.org/10.1145/3351095.3372870

[21] Z. C. Lipton. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. Queue 16, 2018, https://doi.org/10.1145/3236386.3241340

[22] R. R. Hoffman, S. T. Mueller, G. Klein and J. Litman. Metrics for explainable AI: Challenges and prospects. Computer Science, 2019, https://doi.org/10.48550/arXiv.1812.04608

[23] K. Fauvel, V. Masson and E. A. Fromont. A performance-explainability framework to benchmark machine learning methods: Application to multivariate time series classifiers. Computer Science, Machine Learning, 2021, https://doi.org/10.48550/arXiv.2005.14501

[24] Y. Guang, Y. Qinghao and X. Jun. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, Volume 77, Pages 29-52, 2022, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2021.07.016

[25] S. N. Veer, L. Riste, S. Cheraghi-Sohi, D. L. Phipps, M. P. Tully, K. Bozentko, S. Atwood, A. Hubbard, C. Wiper, M. Oswald et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. Journal of the American Medical Informatics Association, Volume 28, Issue 10, pp. 2128–2138, 2021, https://doi.org/10.1093/jamia/ocab127

[26] J. Druce, M. Harradon and J. Tittle. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. Computer Science, Artificial Intelligence, 2021, https://doi.org/10.48550/arXiv.2106.03775

[27] E. Merrer and G. Trédan. Remote explainability faces the bouncer problem. Nature Machine Intelligence, 2020, https://doi.org/10.1038/s42256-020-0216-z

[28] G. A. Kaissis, M. R. Makowski, D. Rückert and R. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2020, https://doi.org/10.1038/s42256-020-0186-1

[29] S. Saifullah, D. Mercier, A. Lucieri, A. Dengel and S. Ahmed. Privacy Meets Explainability: A Comprehensive Impact Benchmark. Computer Science, Machine Learning, 2022, https://doi.org/10.48550/arXiv.2211.04110

[30] HHS Office for Civil Rights. Standards for privacy of individually identifiable health information – Final rule. pp. 53181–53273, 2002, https://aspe.hhs.gov/standards-privacy-individually-identifiable-health-information

[31] HHS Office for Civil Rights. The HIPAA Privacy Rule and Electronic Health Information Exchange in a Networked Environment – Openness and Transparency. pp. 1-4, 2025, https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/special/healthit/opennesstransparency.pdf

[32] R. Creemers and G. Webster. Translation: Personal Information Protection Law of the People's Republic of China. Effective 1 November 2021, https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/

[33] R. K. Yekollu, T. B. Ghuge, S. S. Biradar, S. V. Haldikar and O. F. M. A. Kader. Explainable AI in Healthcare: Enhancing Transparency and Trust in Predictive Models. pp. 1660–1664, 2024, https://doi.org/10.1109/icesc60852.2024.10690121

[34] A. S. Albahri. A systematic review of trustworthy and explainable artificial intelligence in Healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion, vol. 96, pp. 156–191, 2023, https://doi.org/10.1016/j.inffus.2023.03.008

[35] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez and F. Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion, Volume 99, 2023, ISSN 1566-2535,

https://doi.org/10.1016/j.inffus.2023.101805

[36] J. Amann, A. Blasimme, E. Vayena, D. Frey and V. I. Madai. Explainability for Artificial Intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, vol. 20, no. 1, 2020, https://doi.org/10.1186/s12911-020-01332-6

[37] R. Larasati and A. DeLiddo. Building a Trustworthy Explainable AI in Healthcare. Human Computer Interaction and Emerging Technologies, 2020, https://doi.org/10.18573/book3.ab

[38] A. A. Adeniran, A. P. Onebunne and P. William. Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. World Journal of Advanced Research and Reviews, 23(3), pp. 2447–2658, 2024, https://doi.org/10.30574/wjarr.2024.23.3.2936

[39] K. N. Alam, P. B. Zadeh and A. Sheikh-Akbari. Attribution-Based Explainability in Medical Imaging: A Critical Review on Explainable Computer Vision (X-CV) Techniques and Their Applications in Medical AI. Electronics, 2025, https://doi.org/10.3390/electronics14153024

[40] N. Gharaibeh. Enhancing interpretability in brain tumor detection: Leveraging Grad-CAM and SHAP for explainable AI in MRI-based cancer diagnosis. Applied Computer Science, 21(3), pp. 182–197, 2025, https://doi.org/10.35784/acs_7375

[41] V. Sankaradass, V. K. M. Manish, R. Anandhan, R. Velmurugan and S. Sakthivel. A Machine Learning and Data Analytics Approach to Patient Risk Stratification. 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, pp. 1-6, 2025, https://doi.org/10.1109/ICDSAAI65575.2025.11011598

[42] N. Bernard, and K. Balog. A Systematic Review of Fairness, Accountability, Transparency, and Ethics in Information Retrieval. ACM Computer Survey, Vol. 57, Article 136, 2025, https://doi.org/10.1145/3637211

[43] R.B. Parikh, A. Gdowski, D. A. Patt et al. Using Big Data and Predictive Analytics to Determine Patient Risk in Oncology. American Society of Clinical Oncology Educational book, American Society of Clinical Oncology, Annual Meeting, 2019, https://doi.org/10.1200/edbk_238891

[44] U. M. K. Moorthy, A. M. J. Muthukumaran, V. Kaliyaperumal, S. Jayakumar and K. A. Vijayaraghavan. Explainability and Regulatory Compliance in Healthcare: Bridging the Gap for Ethical XAI Implementation. Explainable Artificial Intelligence in the Healthcare Industry, pp. 521-561, 2025, https://doi.org/10.1002/9781394249312.ch23

[45] S. Daram. Explainable AI in Healthcare: Enhancing Trust, Transparency, and Ethical Compliance in Medical AI Systems. pp. 11-20, 2025, https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I2P102

[46] S. Jenko, E. Papadopoulou, V. Kumar, S. S. Overman, K. Krepelkova, J. Wilson, E. L. Dunbar, C. Spice and

T. Exarchos. Artificial Intelligence in Healthcare: How to Develop and Implement Safe, Ethical and Trustworthy AI Systems. AI, 6(6), 2025, https://doi.org/10.3390/ai6060116

[47] M. Mądra-Sawicka, J. Gołuchowski and J. Paliszkiewicz. Trust-Building in the Generative AI – Future Perspectives and Emerging Trends. Routledge, 1st Edition, 2025, https://doi.org/10.4324/9781003586944-2

[48] K. Raman, R. Kumar, C. J. Musante and S. Madhavan. Integrating Model-Informed Drug Development With AI: A Synergistic Approach to Accelerating Pharmaceutical Innovation. Clinical and Translational Science, 18(1), 2025, https://doi.org/10.1111/cts.70124

[49] A. Taheri, A. Farhadi, and A. Zamanifar. Application of GenAI in Clinical Administration Support. Application of Generative AI in Healthcare Systems, Springer, pp. 91-117, 2025, ISBN 978-3-031-82962-8, https://doi.org/10.1007/978-3-031-82963-5_4

[50] K. Zheng, Z. Shen, Z. Chen et al. Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education. BMC Medical Education, 2024, ISSN 1472-6920, https://doi.org/10.1186/s12909-024-05977-z