American Journal Of Biomedical Science & Pharmaceutical Innovation

VOLUME Vol.05 Issue10 2025
PAGE NO. 87-93

Bridging the Interpretability Gap: A Systematic Analysis of Explainable Artificial Intelligence (XAI) and Generative Models in Precision Medicine and Healthcare Analytics

Dr. Helviana S. Kostrevic

Independent Researcher, Generative Models for Medical Imaging & Diagnostics, Munich, Germany

Received: 01 October 2025; **Accepted:** 15 October 2025; **Published:** 31 October 2025

Abstract: Background: The rapid integration of Artificial Intelligence (AI) into healthcare has revolutionized diagnostic precision and treatment personalization. However, the adoption of complex "black box" algorithms, particularly Deep Learning models, faces significant hurdles regarding interpretability, trustworthiness, and ethical bias.

Objectives: This study provides a systematic analysis of the current state of AI in biomedicine, focusing specifically on the pivotal role of Explainable Artificial Intelligence (XAI) and Generative AI models. The primary objective is to evaluate how interpretability mechanisms can reconcile the trade-off between algorithmic performance and clinical transparency.

Methods: We conducted a comprehensive theoretical analysis of recent literature, examining data sharing initiatives, synthetic data generation using Generative Adversarial Networks (GANs), and the application of Large Language Models (LLMs). We utilized a taxonomy of interpretability to assess various XAI frameworks, including SHAP, LIME, and counterfactual explanations, against clinical requirements for accountability.

Results: The analysis indicates that while deep learning offers superior predictive capabilities in precision medicine, its opacity remains a barrier to deployment. The results demonstrate that synthetic data generation via cGANs effectively preserves patient privacy while expanding training datasets. Furthermore, XAI methods are critical for identifying systemic biases in training data, though current evaluation metrics for these explanations often lack standardization.

Conclusions: To realize the full potential of AI in healthcare, systems must transition from opaque prediction engines to transparent decision-support partners. The integration of robust XAI frameworks, alongside rigorous governance of generative models, is essential for ensuring equitable, safe, and clinically valid patient outcomes.

Keywords: Precision Medicine, Explainable AI (XAI), Generative Adversarial Networks, Healthcare Disparities, Big Data Analytics, Algorithmic Bias, Clinical Decision Support.

Introduction

The integration of Artificial Intelligence (AI) into the biomedical sphere represents one of the most transformative shifts in the history of medical science. As we move further into the 21st century, the convergence of high-throughput computational power, massive biological datasets, and sophisticated algorithmic architectures is reshaping the landscape of clinical care. Joiner [1] articulates that AI is no longer a futuristic concept but a "nearby" reality, permeating various facets of information management and service delivery. This proximity is most evident in the domain of precision medicine, where the one-size-fits-all approach of traditional pharmacology is being supplanted by targeted interventions based on individual genetic, environmental, and lifestyle factors.

The potential of AI in healthcare is vast, ranging from robotic surgery and virtual nursing assistants to automated diagnosis and dosage optimization. Yu et al. [3] highlight that machine learning algorithms, particularly deep learning models, have demonstrated performance comparable to, and in some cases exceeding, human experts in tasks such as dermatological screening and radiological image analysis. However, as the complexity of these models increases, so does the opacity of their decision-making processes. This phenomenon, often described as the "black box" problem, poses a critical challenge to clinical adoption. In a high-stakes environment where a false negative can delay life-saving treatment or a false positive can lead to unnecessary invasive procedures, the ability to understand why an algorithm reached a specific conclusion is as important as the conclusion itself.

The transition from traditional statistical methods to modern AI is fueled by the explosion of "Big Data." Hulsen et al. [4] describe the journey from Big Data to precision medicine as a fundamental restructuring of how health information is processed. The sheer volume of data generated by wearable devices, genomic sequencing, and Electronic Health Records (EHRs) exceeds the cognitive capacity of human analysis. AI serves as the bridge, distilling this noise into actionable signals. Yet, this reliance on data brings its own set of complications. Data heterogeneity, privacy concerns, and the siloed nature of medical institutions often hinder the development of robust, generalizable models. Furthermore, as Hulsen et al. [5] note, the ultimate goal is not merely data accumulation but the translation of this data into better patient outcomes, a process that requires rigorous validation and clinical integration.

Recent advancements have introduced a new layer of complexity and opportunity: Generative AI. Technologies such as ChatGPT and other Large Language Models (LLMs) are redefining medical communication and documentation. Biswas [6] suggests that these tools could streamline medical writing and literature synthesis, yet they also introduce risks regarding accuracy and the potential for hallucination—where the AI generates plausible but factually incorrect information. Similarly, the use of Generative Adversarial Networks (GANs) for creating synthetic data offers a solution to privacy bottlenecks but necessitates careful scrutiny regarding the fidelity of the generated data distributions.

Amidst these technological leaps, the ethical dimension of AI deployment remains a paramount concern. Celi et al. [7] provide a sobering review of the sources of bias in AI that can perpetuate healthcare disparities. If an algorithm is trained on historical data that reflects systemic inequalities—such as the underrepresentation of certain ethnic groups in clinical trials—the model will inevitably encode and amplify these biases. This is where Explainable AI (XAI) becomes indispensable. Gunning et al. [10] define XAI as a suite of techniques designed to make the output of AI systems transparent and interpretable to human users. By "peeking inside the black box" [Adadi & Berrada, Ref 13], clinicians can verify that the model is relying on medically relevant features rather than spurious correlations or demographic proxies.

This article aims to provide a comprehensive, systematic analysis of the intersection between AI capabilities and clinical interpretability. We will explore the current methodologies for managing Big Data, the emerging role of generative models, and the critical necessity of XAI in ensuring that the future of precision medicine is both powerful and trustworthy. By synthesizing insights from diverse computational and medical disciplines, we seek to outline a pathway toward "Responsible AI" in healthcare—a paradigm where algorithmic precision is balanced with human understanding and ethical accountability.

Methods

To achieve the objectives of this study, we employed a comprehensive narrative review combined with a theoretical framework analysis. This methodological approach allows for the synthesis of quantitative performance metrics found in computational literature with the qualitative, ethical, and clinical considerations prevalent in medical journals.

Literature Search and Selection

The review process involved a systematic search across major academic databases, including PubMed, IEEE Xplore, and Web of Science. The search strategy utilized a combination of keywords such as "Artificial Intelligence in Healthcare," "Explainable AI," "Precision Medicine," "Generative Adversarial Networks," and "Algorithmic Bias." The inclusion criteria were defined to select high-impact studies published primarily between 2018 and 2024, ensuring the relevance of the technological assessments. We specifically targeted literature that addressed the intersection of machine learning efficacy and interpretability [Refs 2, 12], as well as foundational texts on the philosophy and ethics of AI in medicine [Refs 7, 10]. Articles were excluded if they focused solely on theoretical mathematics without a clear biomedical application or if they lacked a discussion on the implications of model deployment in clinical settings.

Theoretical Framework for Analysis

We adopted a multi-dimensional framework to analyze the selected literature. This framework consists of three primary domains:

1. Data Ecology: Evaluating the methods for data

acquisition, sharing, and augmentation. This includes the analysis of "Big Data" initiatives [Ref 8] and the application of synthetic data generation [Ref 9].

2. Model Architecture and Interpretability: categorizing AI models based on their inherent transparency. We distinguish between "white-box" models (e.g., linear regression, decision trees) which are inherently interpretable, and "black-box" models (e.g., deep neural networks, ensemble methods) which require post-hoc explanation methods.

3. Ethical and Clinical Validity: Assessing the literature through the lens of healthcare disparities [Ref 7] and clinical utility. This domain focuses on whether the AI systems demonstrate not just statistical accuracy, but also fairness and actionable value in a real-world medical context.

Taxonomy of Explainability

To provide a structured analysis of XAI, we utilized the taxonomy proposed by Adadi and Berrada [Ref 13] and expanded by Bharati et al. [12]. This taxonomy classifies explanation methods based on:

- Scope: Local (explaining a single prediction) vs. Global (explaining the overall logic of the model).
- Methodology: Model-agnostic (applicable to any algorithm) vs. Model-specific (tailored to a specific architecture).
- Timing: Ante-hoc (interpretability built into the model design) vs. Post-hoc (interpretability extracted after training).

By applying this structured lens to the reviewed literature, we aim to synthesize a coherent narrative that connects the technical specifications of AI algorithms with their practical and ethical implications in modern medicine.

Results

The analysis of the selected literature reveals a complex landscape where technological capability often outpaces implementation frameworks. The results are categorized into four primary sections: the management of Big Data, the emergence of synthetic data, the role of Generative AI, and the critical evaluation of XAI methodologies.

3.1 The Big Data Landscape and Precision Medicine

The foundation of modern biomedical AI is the unprecedented availability of data. Hulsen [2] notes in a literature analysis that the volume of publications and applications in AI for biomedicine has seen an exponential rise, correlating with the digitization of healthcare. The transition to precision medicine is heavily reliant on the ability to integrate multi-modal data sources—combining genomic, proteomic, and phenotypic data with continuous monitoring from wearable sensors [4]. However, the utility of this data is contingent upon accessibility and standardization. Hulsen [8] emphasizes that "sharing is caring," highlighting that robust data-sharing initiatives are essential for training generalizable models. Siloed data leads to overfitting, where an algorithm performs exceptionally well on data from one hospital but fails when applied to a different demographic or geographical population.

The analysis confirms that while Big Data is a prerequisite for precision medicine, it is not a panacea. The quality of the data ("Smart Data") is often more critical than the quantity. Hulsen et al. [5] demonstrate that better patient outcomes are achieved only when big data analytics are coupled with rigorous clinical validation protocols that filter out noise and identify causal relationships rather than mere correlations.

3.2 Synthetic Data and Generative Adversarial Networks (GANs)

A significant finding in the review is the growing importance of synthetic data as a solution to privacy constraints. Medical data is highly sensitive, protected by regulations such as HIPAA and GDPR, which often stifles research collaboration. Vega-Márquez et al. [9] present compelling evidence on the creation of synthetic data using Conditional Generative Adversarial Networks (cGANs). These networks consist of two competing models—a generator that creates fake data and a discriminator that attempts to distinguish it from real data. Over time, the generator learns to produce synthetic datasets that statistically mirror the original real-world data without containing any actual patient records.

The application of cGANs allows researchers to augment small datasets, balancing classes in rare disease research where positive cases are scarce. This augmentation improves the robustness of diagnostic classifiers. However, the results also suggest a need for caution; the synthetic data must be rigorously validated to ensure it preserves the complex, non-linear relationships found in biological systems, rather than simplifying them for the sake of model convergence.

3.3 The Rise of Generative AI and Large Language Models

The introduction of transformer-based models, particularly ChatGPT, has marked a new era in medical informatics. Biswas [6] outlines the potential of these models in the future of medical writing and education. The ability of LLMs to synthesize vast amounts of medical literature, draft summaries, and even generate differential diagnoses is transforming the cognitive workflow of clinicians. However, the analysis reveals a "hallucination" problem, where the model generates scientifically sounding but factually incorrect assertions. In the context of our framework, LLMs currently represent a significant interpretability challenge; they operate as massive black boxes where the provenance of specific information is often untraceable, raising concerns about liability and verification in clinical decision support.

3.4 Architectural Nuances of Explainable AI (XAI)

The central focus of our analysis lies in the detailed evaluation of Explainable AI methodologies. As deep learning models increasingly dominate the biomedical landscape due to their superior performance in handling high-dimensional data, the necessity for XAI has transitioned from a theoretical preference to a clinical mandate. The literature reveals a distinct bifurcation in XAI approaches: intrinsic interpretability (ante-hoc) versus post-hoc explanation methods.

3.4.1 Ante-hoc vs. Post-hoc Mechanisms

Ante-hoc models, such as Generalized Additive Models (GAMs) and decision trees, offer transparency by design. Their internal logic is accessible; for instance, a decision tree provides a clear, step-by-step path based on clinical thresholds (e.g., "If Glucose > 180, check HbA1c"). However, our review of Agarwal et al. [Ref 3 in list 2] regarding Neural Additive Models suggests a critical limitation: traditional interpretable models often fail to capture the complex, non-linear interactions characteristic of biological systems, leading to a performance deficit compared to deep neural networks.

Consequently, the field has heavily pivoted toward post-hoc methods—techniques applied to a trained "black box" model to approximate its decision boundary. The two most prominent techniques identified in the literature are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations).

3.4.2 LIME and Local Fidelity

LIME operates on the premise of local fidelity. It does not attempt to explain the entire complex model; rather, it perturbs the input around a specific instance (e.g., a specific patient's MRI scan) to see how the prediction changes. By fitting a simple, interpretable linear model to this local region, LIME identifies which features were most influential for that specific prediction. For example, in a dermatology classifier, LIME might highlight the irregular border of a lesion as the primary driver for a "malignant" classification. While useful for individual case audits, the literature suggests that LIME suffers from instability; slight changes in the sampling parameters can lead to vastly different explanations for the same input, posing a risk to clinical trust.

3.4.3 SHAP and Game Theory

SHAP, derived from cooperative game theory, offers a more theoretically grounded approach. It assigns an "importance value" to each feature for a particular prediction, representing that feature's contribution to the deviation from the baseline prediction. As detailed in recent reviews [12], SHAP possesses the property of consistency, meaning that if a model relies more on a certain feature, the SHAP value for that feature will not decrease. In genomic analysis, SHAP has proven invaluable. When predicting disease susceptibility based on thousands of genetic markers, SHAP can isolate specific Single Nucleotide Polymorphisms (SNPs) that drive the risk score, allowing researchers to validate the AI's logic against known biological pathways.

3.4.4 Neural Additive Models (NAMs) and Concept-Based Explanations

To bridge the gap between the accuracy of neural networks and the clarity of GAMs, recent innovations like Neural Additive Models (NAMs) [Ref 3 in list 2] have emerged. NAMs learn a linear combination of neural networks, where each network attends to a single input feature. This architecture allows for the visualization of the exact shape function for each feature (e.g., how the risk of cardiovascular event changes non-linearly with age), maintaining high accuracy while providing an inherently interpretable graph for every variable.

Furthermore, concept-based interpretation methods, such as those discussed in the context of time-series models [Ref 4 in list 2], represent a shift from feature-level to concept-level explanations. Instead of telling a clinician that "Pixel 405" is important, these models align internal activation patterns with high-level clinical concepts (e.g., "arrhythmia pattern" or "fluid opacity"). This semantic alignment is crucial for bridging the cognitive gap between the data scientist and the physician.

3.4.5 Gradient-Based Saliency and Attention Mechanisms

In the domain of medical imaging, gradient-based methods (e.g., Grad-CAM) remain dominant. These techniques visualize the gradient of the target class with respect to the input image, generating a heatmap that highlights the regions of interest. Vu et al. [11] discuss the relevance of these mechanisms in neuroscience, noting that they parallel biological attention. However, our analysis of OpenXAI benchmarks [Ref 2 in list 2] indicates that saliency maps can sometimes be misleading, acting more as edge detectors rather than true representations of the model's reasoning. This "sanity check" failure implies that a model might look at the correct area (e.g., the lung) but for the wrong reason (e.g., identifying a watermark or a specific scanner artifact), underscoring the need for rigorous evaluation of the explanations themselves.

3.5 Evaluation Metrics for Explainability

A critical gap identified in the results is the lack of standardized metrics for evaluating explanations. While model accuracy is easily measured (AUC-ROC, F1-score), "interpretability" is subjective. Recent efforts [Ref 2 in list 2] have proposed quantitative metrics such as faithfulness (how truly the explanation reflects the model's computation) and robustness (how stable the explanation is to minor perturbations). The results indicate that many popular XAI methods score high on visual appeal but lower on strict faithfulness, suggesting a "placebo effect" where the explanation comforts the user without accurately revealing the model's flaws.

3.6 The Symbiosis of Generative Models and Explainability

The integration of generative models with XAI represents a frontier in our analysis. Generative models can serve as "counterfactual engines." By using a GAN to generate a realistic variation of a patient's data (e.g., "What would this patient's risk look like if their blood pressure were normalized?"), clinicians can engage in "what-if" analysis. This counterfactual reasoning [9] aligns closely with clinical diagnostics. Moreover, the linguistic capabilities of LLMs [6] are being explored to translate complex SHAP values or saliency maps into natural language summaries, effectively acting as an interface layer that narrates the AI's findings in medical prose. This convergence suggests a future where the "Black Box" is not opened, but rather interviewed.

Discussion

The findings of this systematic analysis underscore a pivotal moment in biomedical AI. We stand at a crossroads where the technological impetus for higher accuracy clashes with the clinical and ethical imperative for transparency. The discussion below synthesizes these tensions, focusing on the trade-offs, ethical liabilities, and the future integration of computational and biological intelligence.

4.1 The Accuracy-Interpretability Trade-off

The traditional dogma in machine learning postulates a zero-sum game between accuracy and interpretability: simple models are interpretable but less accurate, while deep neural networks are accurate but opaque. However, our analysis of Neural Additive Models and advanced XAI techniques suggests this trade-off is becoming less rigid. As noted by Bharati et al. [12], the development of hybrid architectures allows for "glass-box" approaches that retain the feature-learning power of deep nets while structuring them in human-understandable modules.

Nevertheless, a residual tension persists in high-dimensional domains like genomics. Here, the interactions between thousands of genes are inherently complex and perhaps beyond intuitive human visualization. In such cases, enforcing strict interpretability might force the model to oversimplify biological reality, potentially missing subtle but critical multi-genic interactions. Therefore, the goal should not always be complete transparency (understanding every neuron), but rather "functional interpretability"—providing sufficient evidence to justify a clinical action.

4.2 Algorithmic Accountability and Bias Mitigation

One of the most profound implications of XAI lies in its ability to act as a safeguard against bias. Celi et al. [7] argue that AI systems are often "mirrors of inequality," reflecting the disparities present in the healthcare system. Without XAI, a model trained on data from predominantly urban, wealthy hospitals might learn to associate access to expensive diagnostic tests with better outcomes, unfairly penalizing rural or lower-income patients who lack such data points.

Post-hoc analysis using tools like SHAP can reveal these "leakage" variables. If an explanation reveals that a model is weighing "insurance type" or "zip code" heavily in a mortality prediction, it serves as a red flag for algorithmic bias. This capability transforms XAI from a mere user-interface feature into a core component of ethical compliance. It moves the conversation from "Does the model work?" to "Does the model work fairly for all subgroups?" This aligns with the "shared vision" discussed by Vu et al. [11], where machine learning in neuroscience and medicine must account for the diverse heterogeneity of biological populations.

4.3 The Legal and Regulatory Framework

The integration of these technologies necessitates a robust legal framework. As AI moves from research to bedside, questions of liability arise. If an AI system recommends a treatment that fails, who is responsible—the physician, the hospital, or the developer? Adadi and Berrada [Ref 13] highlight that the European Union’s GDPR includes a "right to explanation," mandating that automated decisions significantly affecting individuals must be explainable. This legal requirement elevates XAI from a technical feature to a regulatory necessity.

In the context of Generative AI, the legal stakes are even higher. Biswas [6] notes the potential for copyright infringement and the propagation of misinformation by LLMs. Establishing "Algorithmic Accountability" requires not just explainable code, but explainable data lineages—tracking exactly which data points influenced a model's output. Technologies like Blockchain and immutable data logs, utilized in conjunction with XAI, may offer a pathway to verifiable audit trails for medical AI.

4.4 Neuroscience and the "Human-in-the-Loop"

Finally, the discussion must circle back to the ultimate user: the human brain. Vu et al. [11] emphasize the shared vision between machine learning and neuroscience. Understanding how the human brain processes information—specifically how expert physicians recognize patterns—can inform the design of better XAI interfaces. A radiologist does not look at every pixel; they look for deviations from a learned prototype. Therefore, XAI systems should be designed to mimic this "contrastive" reasoning, highlighting only what is anomalous rather than overwhelming the user with heatmaps of the entire anatomy.

This "Human-in-the-Loop" approach ensures that AI remains a tool for augmentation rather than replacement. By presenting explanations that align with clinical reasoning workflows, we can reduce cognitive load and prevent "automation bias," where clinicians blindly accept the computer's suggestion. The synergy between biological intelligence (context, empathy, ethics) and artificial intelligence (pattern recognition, data processing) is the cornerstone of the next generation of precision medicine.

4.5 Limitations and Future Directions

It is important to acknowledge the limitations of current XAI methods. As discussed in the results, post-hoc explanations can be unstable and, in some cases, manipulated. "Adversarial attacks" on explanations are possible, where an imperceptible change to the input image drastically alters the heatmap without changing the prediction. This fragility poses a security risk. Future research must focus on "robust XAI"—explanations that are mathematically guaranteed to be stable.

Additionally, the validation of synthetic data generated by GANs [9] requires standardization. While statistical metrics may show convergence, clinical validation is distinct. Future studies must focus on "clinical Turing tests," determining if expert physicians can distinguish between real and synthetic patient profiles, and more importantly, if models trained on synthetic data perform reliably on real humans in clinical trials.

Conclusions

The trajectory of Artificial Intelligence in biomedicine is ascending, driven by the dual engines of Big Data availability and algorithmic sophistication. However, as this study demonstrates, the climb is fraught with challenges related to opacity, bias, and trust. The transition from "Black Box" to "Glass Box" is not merely a technical upgrade; it is a fundamental requirement for the ethical practice of medicine in the digital age.

Our systematic analysis highlights that while deep learning and generative models like ChatGPT and GANs offer unprecedented capabilities in predictive analytics and data augmentation, their utility is contingent upon interpretability. Explainable AI (XAI) serves as the critical interface, translating high-dimensional mathematical probabilities into clinical reasoning. By exposing the logic behind predictions, XAI allows for the detection of bias [7], the validation of biological plausibility [11], and the fostering of trust between the machine, the clinician, and the patient.

Ultimately, the goal of AI in healthcare is not to replace the physician but to arm them with precision tools. As we refine techniques like SHAP, LIME, and Neural Additive Models, and as we establish rigorous governance for data sharing [8] and synthetic generation [9], we move closer to a future where precision medicine is a reality for all. In this future, the AI does not just predict; it explains, empowers, and collaborates, ensuring that the human element of care remains central in an increasingly automated world.

References

1. Joiner, I.A. Chapter 1—Artificial intelligence: AI is nearby. In Emerging Library Technologies; Joiner, I.A., Ed.; Chandos Publishing: Oxford, UK, 2018; pp. 1–22.
2. Hulsen, T. Literature analysis of artificial intelligence in biomedicine. Ann. Transl. Med. 2022, 10, 1284.
3. Yu, K.-H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. Nat. Biomed. Eng. 2018, 2, 719–731.
4. Hulsen, T.; Jamuar, S.S.; Moody, A.; Karnes, J.H.; Orsolya, V.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E. From Big Data to Precision Medicine. Front. Med. 2019, 6, 34.
5. Hulsen, T.; Friedecký, D.; Renz, H.; Melis, E.; Vermeersch, P.; Fernandez-Calle, P. From big data to better patient outcomes. Clin. Chem. Lab. Med. (CCLM) 2022, 61, 580–586.
6. Biswas, S. ChatGPT and the Future of Medical Writing. Radiology 2023, 307, e223312.
7. Celi, L.A.; Cellini, J.; Charpignon, M.-L.; Dee, E.C.; Dernoncourt, F.; Eber, R.; Mitchell, W.G.; Moukheiber, L.; Schirmer, J.; Situ, J. Sources of bias

in artificial intelligence that perpetuate healthcare disparities—A global review. PLoS Digit. Health 2022, 1, e0000022.

8. Hulsen, T. Sharing Is Caring-Data Sharing Initiatives in Healthcare. Int. J. Environ. Res. Public Health 2020, 17, 3046.
9. Vega-Márquez, B.; Rubio-Escudero, C.; Riquelme, J.C.; Nepomuceno-Chamorro, I. Creation of synthetic data with conditional generative adversarial networks. In Proceedings of the 14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019), Seville, Spain, 13–15 May 2019; Springer: Cham, Switzerlnad, 2020; pp. 231–240.
10. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI-Explainable artificial intelligence. Sci. Robot. 2019, 4, eaay7120.
11. Vu, M.T.; Adalı, T.; Ba, D.; Buzsáki, G.; Carlson, D.; Heller, K.; Liston, C.; Rudin, C.; Sohal, V.S.; Widge, A.S.; et al. A Shared Vision for Machine Learning in Neuroscience. J. Neurosci. 2018, 38, 1601–1607.
12. Bharati, S.; Mondal, M.R.H.; Podder, P. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? IEEE Trans. Artif. Intell. 2023.
13. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access 2018, 6, 52138–52160.
14. Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; Lakkaraju, H. OpenXAI: towards a transparent evaluation of model explanations. Advances in Neural Information Processing Systems 2022, 35, 15784–15799.
15. Shankheshwaria, Y.V.; Patel, D.B. Explainable AI in Machine Learning: Building Transparent Models for Business Applications. Frontiers in Emerging Artificial Intelligence and Machine Learning 2025, 2(08), 08–15.
16. Agarwal, R.; Melnick, L.; Frosst, N.; Zhang, X.; Lengerich, B.; Caruana, R.; Hinton, G.E. Neural additive models: interpretable machine learning with neural nets. 2021, 34, 4699–4711.
17. Asadi, M.; Swamy, V.; Frej, J.; Vignoud, J.; Marras, M.; Käser, T. Ripple: concept-based interpretation for raw time series models in education. In The 37th AAAI Conference on Artificial Intelligence (EAAI), 2023.
18. Bengio, Y.; Léonard, N.; Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.