Eksplorium *p*-ISSN 0854-1418
Volume 46 No. 1, May 2025: 104–133 *e*-ISSN 2503-426X

Explainable AI & Model Interpretability in Healthcare: Challenges & Future Directions

Puneet Garg[1], Gunjan Beniwal[2], Priya Dalal[3], Monika[4], Meeta chaudhry[5]

[1]Associate Professor, KIET Group of Institutions, Delhi NCR, Ghaziabad, India

[2,3]Assistant Professor, Maharaja Surajmal Institute of Technology, New Delhi, India

[4]Assistant Professor, SAITM, Gurugram, Delhi NCR, India

[5]Associate Professor, KIET Group of Institutions, Delhi NCR, Ghaziabad, India

puneetgarg.er@gmail.com[1], gunjanbeniwal@msit.in[2], priya@msit.in[3], monikadhiman20@gmail.com[4], chaudhrymeeta1@gmail.com[5]

Abstract: Explainable Artificial Intelligence (XAI) in healthcare seeks to make the behavior and reasoning of complex machine learning models transparent to stakeholders. As AI systems become increasingly prevalent in clinical decision support, their "black-box" nature raises concerns about trust, safety, and ethical use. This paper presents a comprehensive overview of explainability and model interpretability in healthcare, emphasizing theoretical foundations, key challenges, and emerging solutions. We begin by defining XAI and its importance in the medical context, outlining how interpretability can enhance clinician and patient trust without sacrificing model performance. We then review the broad applications of XAI across healthcare domains, illustrating its growing adoption. Next, we delve into key challenges that impede the integration of XAI into clinical workflows: the need for trust and transparency, the complexity of state-of-the-art models, ethical and regulatory requirements for explainability, data privacy constraints, and practical barriers to deployment in healthcare settings. This future entails interdisciplinary collaboration, standardized evaluation metrics for explanations, and regulatory frameworks that encourage safe, transparent AI in medicine. By addressing current challenges and leveraging emerging methods, XAI can foster appropriate trust in AI-driven healthcare and ultimately improve decision-making and patient outcomes.

Keywords: Explainable AI; Model Interpretability; Healthcare AI; Trustworthy AI; Transparency; Federated Learning; Attention Mechanisms; Counterfactual Explanations; Clinical Decision Support; Ethical AI.

1. INTRODUCTION

Artificial Intelligence (AI) is reshaping healthcare by delivering powerful data-driven tools for diagnosis, prognosis, and treatment recommendation. Complex machine learning models – particularly deep neural networks – have achieved remarkable accuracy in tasks such as medical image analysis, outcome prediction, and therapy planning. However, a critical barrier to routine clinical adoption of these AI systems is the lack of explainability or interpretability of their decision-making processes. Healthcare professionals and patients are often reluctant to trust a "black-box" model whose internal logic remains opaque, especially in high-stakes medical decisions that demand accountability and justification [1][2]. This concern has catalyzed intense interest in Explainable AI (XAI) for healthcare, which aims to make AI's reasoning transparent and understandable to humans [8][9].

Research in explainable and interpretable ML has grown rapidly to meet these needs. **Figure 1** illustrates the surge in XAI-focused publications over the past several years, highlighting the burgeoning interest post-2016 and especially in the 2020s [2]. This trend reflects both advances in XAI techniques and the recognized necessity of explanations for AI systems

deployed in sensitive domains like healthcare. Indeed, explainability has become a policy priority and a research frontier as stakeholders realize that improving model transparency is essential for clinical acceptance [1][5][6]. Despite this progress, significant challenges remain in achieving explainable AI that clinicians will trust and routinely use. Explainability methods must navigate a delicate balance: providing meaningful insights into complex model behavior without overwhelming users or compromising the model's accuracy [4][3][7]. In the healthcare context, explanations need to satisfy multiple stakeholders – from data scientists validating model behavior, to physicians seeking clinical rationale, to patients demanding understandable reasoning for personal health decisions. Each stakeholder may require a different depth and form of explanation (e.g., a clinician might ask *"Why did the model predict this patient is high-risk?"* whereas a patient might ask *"What factors led to this diagnosis?"*). Moreover, explanations must be delivered under constraints of privacy (sensitive patient data), regulation, and the time-critical nature of clinical workflows [13][14].

This paper provides a comprehensive review of explainable AI and model interpretability in healthcare, focusing on overarching challenges and future directions rather than disease-specific case studies. We generalize across medical domains to identify common themes in how explainability can support clinical AI systems. First, we discuss the theoretical foundations of XAI in healthcare, clarifying what it means for an AI model to be "explainable" or "interpretable" and why those properties are crucial in medicine. We then outline the range of current applications of XAI across healthcare, demonstrating its relevance to diverse clinical scenarios. Next, we delve into the key challenges that must be addressed to integrate XAI into routine healthcare: establishing appropriate trust and accountability, achieving transparency with highly complex models, meeting ethical and regulatory standards, protecting patient privacy, and ensuring that explainable models fit seamlessly into clinical workflows. We organize these challenges into distinct categories (Section 3) and illustrate each with recent examples or evidence from the literature. In Section 4, we survey emerging solutions and techniques designed to tackle these challenges – including privacy-preserving learning strategies (like federated learning), model design innovations (like attention mechanisms and inherently interpretable models), advanced explanation techniques (such as counterfactual reasoning), visualization and interaction tools for clinicians, and hybrid approaches that combine knowledge-based and data-driven methods. Two summary tables are provided: Table 1 catalogues the major challenges and their implications, and Table 2 maps the emerging solutions to the challenges they aim to address. Throughout the paper, we include conceptual diagrams (to clarify theoretical ideas) and empirical figures (to present evidence of trends and outcomes in XAI research). Finally, we conclude (Section 5) by synthesizing the insights and highlighting future directions – emphasizing the need for interdisciplinary collaboration, standardized evaluation of explainability, and human-centered design in the next generation of explainable AI systems for healthcare.

The ultimate goal is to inform researchers and practitioners about the current state of explainable AI in healthcare, and to guide efforts toward AI tools that are not only accurate but also transparent, trustworthy, and aligned with clinical needs. By addressing the identified

challenges and leveraging emerging innovations, the community can move closer to AI systems that clinicians readily adopt and patients confidently accept, thereby unlocking the full potential of AI-driven improvements in healthcare outcomes.

2. EXPLAINABLE AI IN HEALTHCARE: CONCEPTS AND APPLICATIONS

2.1 Defining Explainability and Interpretability

In the context of AI, *explainability* refers to the ability of a model to provide reasons or mechanisms for its decisions in a way that humans can understand [2]. A closely related term, *interpretability*, often denotes the degree to which a human can intuitively comprehend the model's internal workings (some literature uses these terms interchangeably, while others draw subtle distinctions). In practical terms, an explainable AI system can answer questions like "*Why did the model make this prediction?*" by highlighting relevant features or providing a human-comprehensible justification [2]. Interpretability is especially critical in medicine, where understanding *why* a recommendation was made can be as important as the recommendation itself for establishing clinical credibility and accountability [10][11][12].

There are two broad approaches to achieving explainability in AI models [2]:

- **Intrinsic Interpretability (Transparent Models):** Using models that are inherently interpretable due to their simple structure. Examples include decision trees, rule-based systems, linear or logistic regression models, and generalized additive models. These models are often called "white-box" because their decision process can be followed step-by-step by humans. In healthcare, an intrinsically interpretable model might be a small decision tree for triage or a logistic regression scoring system for disease risk – systems that clinicians can manually inspect and verify. Intrinsic interpretability is advantageous because it provides *built-in* explanations (for instance, a decision tree yields human-readable paths like "IF age > 60 and smoker = yes THEN risk = high"), but these models may lack the accuracy of more complex models on high-dimensional medical data [15][16].
- **Post-hoc Explainability (Explainability for Black-Box Models):** Applying techniques to explain the decisions of an already-trained complex model without altering its internal architecture. These methods treat the AI model as a black box and produce **auxiliary explanations** for its outputs. Common post-hoc approaches include feature attribution methods (e.g., SHAP values or LIME) that assign importance scores to input features for a given prediction, visualization methods (e.g., saliency maps in medical imaging that highlight influential regions of an image), simplified surrogate models that approximate the black-box model's behavior, and counterfactual explanations that indicate how changing an input would change the output. Post-hoc methods allow clinicians to reap the benefits of complex models (like deep neural networks) while still obtaining some explanation of *why* a particular output was produced. However, these explanations are often approximations and may not perfectly reflect the model's true reasoning [17][18]. For example, a linear approximation via LIME around a specific case provides insight into local decision factors, but it doesn't capture the model's full global logic. Most explainability methods rely on simpler proxy models or visualizations, meaning they can offer only an approximation of the black-box's internal

rationale – not true transparency – and thus have limitations in faithfully improving user understanding and trust [19][20].

Explainability can also be considered at different stages of the AI pipeline. Some models are designed to be interpretable by structure (intrinsic, as above), whereas in many cases we apply explainability *after* the model is trained (post-hoc). Explainability methods can even assist in data understanding (for example, uncovering which input features are most influential during model development). Regardless of stage, the goal of XAI in healthcare is to align the model's decision-making with *human cognitive frameworks*. This means the explanation should ideally connect to medical concepts that clinicians find meaningful (symptoms, lab values, imaging findings, risk factors, etc.) [21][22][23]]. For example, an ideal explanation for a machine-learned diagnostic model might be: *"The AI predicts pneumonia with 90% confidence because it identified an opacity in the lower left lung field on the X-ray and the patient's white blood cell count is elevated,"* linking the model's internal features to established clinical reasoning. Achieving this level of intuitive explanation remains challenging, but research is progressing toward bridging the gap between complex algorithmic logic and the domain knowledge of clinicians. [24][25]

2.2 Why Explainability Matters in Healthcare

In general AI applications, explainability is often touted for increasing user trust and aiding in system debugging. In healthcare, these benefits are magnified and joined by additional critical motivations:

- **Trust and Adoption:** Doctors are more likely to trust and adopt AI tools if they can verify the reasoning behind predictions [3]. Clinical decisions directly impact patient lives; thus, a clinician will be understandably hesitant to base a diagnosis or treatment on an algorithm's output without understanding how it was derived. A recent systematic review of XAI's impact on clinician trust found that providing explanations (either visual or textual) generally increased their confidence in AI recommendations, particularly when the explanations were clear and relevant to clinical context [3]. For example, in several studies XAI visualizations or narratives improved clinicians' trust in an AI's diagnosis compared to using the AI with no explanation at all [3]. On the other hand, misleading or overly complicated explanations can erode trust – thus the quality of explanation is key (simply having an explanation is not a panacea). It is worth noting that some studies found no significant effect of XAI on trust, indicating that the presence of an explanation does not automatically improve trust if the explanation is not useful or understandable [3] [26][27].
- **Transparency and Accountability:** In medicine, decisions often need to be justified to multiple parties – hospital administrators, patients, or even courts in cases of malpractice. Explainable models provide a **traceable justification** for each output, which can be crucial for accountability. For example, if an AI system recommends sending a patient to the ICU, an explanation allows the care team to communicate the rationale (e.g., *"the system detected early signs of sepsis in the vitals and labs"*) which helps everyone involved understand and agree with the action. Moreover, transparency is ethically important: patients have a right to an explanation for decisions affecting their care, aligning with principles of informed consent

and shared decision-making in medicine. The ethical requirement for explainability is increasingly discussed in literature, with some arguing that opaque AI systems should not be used in clinical practice precisely because they undermine informed consent and the physician's duty to explain decisions to the patient [7]. Conversely, a lack of transparency makes it difficult to assign responsibility when errors occur and can impede learning from mistakes (if nobody understands the AI's reasoning, it's hard to fix it or to know when to override it) [28][29][30].

- **Safety and Error Detection:** Explainability can act as a safety net by allowing humans to catch errors or biases in the model's reasoning. If an AI mispredicts because it was paying attention to an irrelevant artifact, a human reviewer is more likely to discover this through explanation tools. For instance, an explainability analysis in one case revealed that a supposedly high-performing skin lesion classifier was focusing on surgical marking stickers in the images rather than the lesion itself – a spurious correlation that a human could detect once the model's attention was visualized. In healthcare, where biases in data or unusual input conditions can lead to dangerous errors, having insight into the model's decision basis is critical for validation and calibration of trust [31].
- **Ethical and Regulatory Compliance:** As mentioned, regulations and guidelines increasingly demand explainability for high-risk AI applications. In healthcare, organizations like the FDA now consider the interpretability of AI/ML tools as part of the evaluation process. The European Union's AI Act will classify medical AI as high-risk and likely mandate appropriate explainability or transparency features [7]. From an ethical standpoint, explainability ties into principles of beneficence and non-maleficence: one must ensure AI benefits patients and does not inadvertently cause harm. Explanations allow detection of potential harm or bias (e.g., uncovering that a predictive model systematically underestimates risk for a certain minority group – an insight that might surface through examining explanation outputs across cohorts) [32].
- **Improving Model Performance and Human–AI Teamwork:** Interestingly, explainability can also improve the *effective* performance of an AI when used in tandem with human experts. By understanding the model's strengths and weaknesses via explanations, clinicians can calibrate when to trust the model and when to be cautious or override it [4]. For example, if an AI imaging tool's explanation highlights an unusual region as key evidence for a tumor, a radiologist might scrutinize that region more carefully, potentially catching something they would have missed or confirming a subtle finding. Explanations can also highlight when the model is unsure (e.g., revealing contradictory evidence, or showing a near-balance of factors) which can prompt humans to gather more information. Studies have begun to show that human–AI teams make better decisions when the AI provides helpful explanations, especially if the explanation format is intuitive to the human user [33].

2.3 XAI Techniques and Approaches in the Medical Context

To ground the discussion, we briefly overview the types of XAI techniques commonly employed in healthcare AI and how they map onto practical needs:

- **Feature Attribution Methods:** These methods, like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations), are widely used to explain ML models in medicine. They typically output a set of feature contributions for a given prediction. For instance, a SHAP explanation for a mortality risk model might show that *"low oxygen saturation"* and *"elevated lactate"* were major contributors to a high-risk prediction, whereas *"younger age"* and *"no comorbidities"* pushed the risk down. Such quantitative attributions help clinicians see which factors the model considered most important. SHAP and related methods have been applied to numerous clinical models (for example, explaining ICU sepsis predictions by showing which vital signs and labs most influenced the risk score). Feature attributions are popular because they are model-agnostic and relatively easy to compute, but care must be taken in interpretation – they explain the model's output, not necessarily the true causal relationships [34][35][36].
- **Saliency Maps and Attention Maps:** In medical imaging applications of AI (radiology, pathology, dermatology, etc.), visual explanation techniques are common. These include saliency maps (e.g., gradient-based or Class Activation Maps) and attention maps that highlight the regions of an image most influential for a CNN's prediction. For example, a saliency map on a chest X-ray processed by a pneumonia-detection model might highlight the area of the lungs with opacities that led to the "pneumonia" prediction, aligning with what a radiologist would consider evidence. Such visual overlays allow clinicians to verify that the model is "looking" at medically relevant features (e.g., lung fields, not an irrelevant part of the image) and can increase trust in the model's decision if the highlighted regions make sense clinically [3]. Many recent healthcare AI systems incorporate attention mechanisms in neural networks specifically to provide this kind of interpretability – the network learns to weight certain parts of the input, and those weights can be visualized as an explanation [37][38][39].
- **Rule-Based and Example-Based Explanations:** Some XAI methods provide explanations in forms closer to human reasoning, such as decision rules or similar case examples. *Rule extraction* algorithms attempt to derive logical rules from a trained model. A rule-based explanation might be: *"System predicts high stroke risk because: IF (age > 75) AND (atrial fibrillation present) THEN high risk."* This is essentially mimicking a clinical rule (similar to CHADS2 stroke risk criteria in this example) and is highly interpretable [40][41].
- **Self-Explaining Models and Attention-Based Models:** Modern deep learning models sometimes incorporate architectural features that lend themselves to interpretation. As noted, attention mechanisms in models (especially in NLP and in emerging vision transformers) naturally provide weights that indicate what the model focused on. There are also *self-explainable neural networks* that are designed to output not just a prediction but also an explanation. For example, some networks for clinical data are structured to first produce intermediate concepts (like "evidence of heart failure = yes/no") and then make a prediction based on those concepts. The intermediate outputs serve as an explanation because they correspond to clinical findings [42][43].

- **Interactive Visualization Tools:** Beyond algorithmic methods, a practical aspect of XAI is how explanations are delivered to users. There are tools and platforms (some standalone, some integrated into clinical software) that allow users to explore model predictions and explanations. For example, an interactive dashboard might let a doctor adjust a patient's risk factors (age, lab results, etc.) and see how the AI's predicted risk changes – a "what-if" analysis interface. Another tool might overlay AI explanations on medical images and allow the user to toggle them on or off. Big tech companies and research groups have released general XAI toolkits (like Google's What-If Tool, IBM's AI Explainability 360) which can be tailored to healthcare applications. These tools can present explanations in multiple forms (text, charts, images) and at various levels of detail. In a clinical deployment, one might envision an EHR system where, next to an AI-generated risk score, there is a button or snippet that, when clicked, expands to show the top reasons for that score. Studies have found that providing such explanation interfaces can improve clinicians' understanding and appropriate use of AI recommendations [3][44].

Table 1 below summarizes key challenges in making AI in healthcare explainable, which we will explore in Section 3, along with their implications for XAI methods. Each challenge creates certain requirements for explainability (e.g., the trust challenge requires that explanations be easily understandable

Table 1. Key Challenges for Explainable AI in Healthcare and Their Implications

Challenge	Description	Implications for XAI
Trust and Acceptance	Clinicians and patients must have appropriate trust in AI recommendations. Black-box models can undermine trust or lead to over-reliance.	Explanations need to be clear, clinically relevant, and accurate to foster trust and understanding [3]. Without trust, clinicians may reject AI advice even if accurate; conversely, over-trust must be mitigated by conveying model confidence and limitations.
Transparency ("Black-Box" Problem)	Many high-performing models (e.g., deep neural networks, ensembles) operate as "black boxes" with opaque decision logic.	XAI methods must make model reasoning visible (through feature attributions, saliency maps, etc.)[1]. Surrogate models or visualizations can offer insight, but they provide only an *approximation* of the true logic[1]. Fully opening the black box remains challenging, so XAI strives for **useful** transparency even if complete transparency is unattainable.
Model Complexity vs. Interpretability	There is often a trade-off between model complexity and predictive accuracy versus human interpretability. Highly complex models tend to be less	Need techniques to explain complex models without significantly sacrificing performance. This may involve **hybrid approaches** or constraints to retain

	interpretable.	interpretability[5]. In some cases, a slightly less complex (but interpretable) model may be chosen for deployment if its performance is sufficient for clinical needs.
Ethical and Regulatory Requirements	Ethical guidelines and laws demand explainability for AI in healthcare (e.g., EU AI Act requires explanations for high-risk AI). Accountability and fairness are key concerns.	XAI is often a **requirement** for deployment[7]. Systems should provide reasons that clinicians and patients can understand, enabling informed consent and auditability. Explainable models help identify biases and ensure decisions can be justified, which is crucial for meeting legal and ethical standards.
Data Privacy and Security	Patient data is sensitive and protected by privacy regulations. Centralizing data or revealing patient-specific details in explanations can violate privacy.	Approaches like **federated learning** allow training and explaining models across silos without sharing raw data[8]. Explanations themselves must be privacy-preserving (avoiding disclosure of protected health information). This imposes constraints on XAI techniques – for instance, using aggregated or de-identified exemplars in example-based explanations.
Integration into Clinical Workflow	AI tools must fit within clinical workflows and time constraints; explanations should aid rather than impede decision-making.	Explanations must be concise, context-specific, and delivered through user-friendly interfaces (e.g., embedded in the EHR or PACS)[10]. They should not overwhelm users with information or cause alert fatigue. Human-centered design and clinician training are needed so that XAI tools complement the decision process instead of disrupting it.

2.4 Applications of XAI Across Healthcare Domains

Explainable AI techniques are being applied in a wide array of medical domains and tasks, underscoring their broad relevance. Rather than focusing on specific case studies, we highlight a few general application areas and the role of XAI in each:

- **Diagnostic Decision Support:** AI models for assisting diagnosis (e.g., detecting diseases from medical images, lab results, or patient symptoms) benefit greatly from XAI. In medical imaging, for example, deep learning models can identify subtle patterns; XAI methods like saliency maps and concept annotations help radiologists see *what* the model identified as abnormal. If an AI flags a chest X-ray for pneumonia, an accompanying heatmap

highlighting lung infiltrates provides confidence that the model's prediction is based on relevant imaging features. In pathology, CNNs analyzing slides can highlight regions with malignant features, effectively pointing the pathologist to areas of interest [2]. For diagnostic models using electronic health records (EHR) data, explanations might list the top contributing symptoms and risk factors for a given diagnosis [45][46].

- **Risk Stratification and Prognosis:** Many healthcare AI systems provide risk scores (e.g., risk of 30-day readmission, surgical complication, disease progression). XAI is crucial here so that clinicians know *why* a patient is at high risk and can take appropriate action. For example, a risk model for cardiac events might explain that a patient's risk is high due to factors like uncontrolled diabetes and hypertension and a history of smoking. Such an explanation allows the care team to verify that the risk prediction aligns with known risk factors and also to communicate to the patient which factors are most contributing to their risk (potentially motivating lifestyle changes or adherence to therapy). If a risk model were a black box, clinicians might distrust or ignore a high-risk flag, but with explanations, they can integrate the model's insights into care planning (e.g., scheduling closer follow-up for a patient because the model identified concerning trends in their lab results [47][48].
- **Therapy Recommendation and Treatment Planning:** AI is increasingly explored for suggesting treatments (for example, recommending personalized chemotherapy regimens based on tumor genomics, or optimal insulin dosing for diabetics based on continuous glucose monitoring). In such settings, interpretability is essential for the AI to be accepted as part of the clinical decision-making team. If an AI recommends Treatment A over Treatment B, it should provide reasoning: e.g., *"Treatment A is suggested because the patient's tumor has biomarkers X and Y which were associated with better response to A in clinical studies"*. This can be achieved by integrating clinical knowledge databases with the model's output (a hybrid approach) [49].
- **Patient-facing Applications:** While much XAI research focuses on clinicians, patients themselves can benefit from explainable AI tools. For instance, symptom-checker apps or chronic disease management apps use AI to give users advice (like "possible condition X, consider doing Y or seeking care"). Providing explanations in lay terms (e.g., *"You reported symptom A and B, which often suggest condition X"*) can increase a patient's trust in the app and help them make informed choices. In mental health apps, if an AI chatbot flags a user as high-risk for depression relapse, an explanation might be delivered as part of the feedback (for example, pointing out that certain mood questionnaire answers or inactivity patterns triggered the concern). Patients generally have a right to an explanation when AI influences their care, and patient-facing XAI can improve engagement and adherence (people are more likely to follow health advice if they understand the rationale) [50][51]. Achieving patient-friendly explanations is an area of ongoing development, likely involving natural language generation techniques to translate model insights into simple language [52].
- **Healthcare Operations and Triage:** Outside of direct patient care, XAI aids in administrative or operational AI systems. For example, algorithms that predict no-show appointments or optimize hospital bed management can provide explanations that help

administrators act. If a model predicts a patient is likely to miss an appointment, an explanation could be *"Prediction based on the patient's past attendance pattern and long travel distance to clinic."* This might lead the staff to reach out with a reminder or arrange transportation – an actionable use of the explanation. In emergency department triage, AI systems that prioritize patients could show the vital sign or symptom that drove a high-acuity classification, which nurses can double-check. This can serve as a second set of eyes; if the explanation highlights something the triage nurse overlooked (e.g., subtle low oxygen level), it can improve care, whereas if it highlights something irrelevant, the nurse knows the model may have erred. By making AI suggestions interpretable, healthcare providers can combine their expertise with AI insights to improve workflow efficiency and patient outcomes [53].

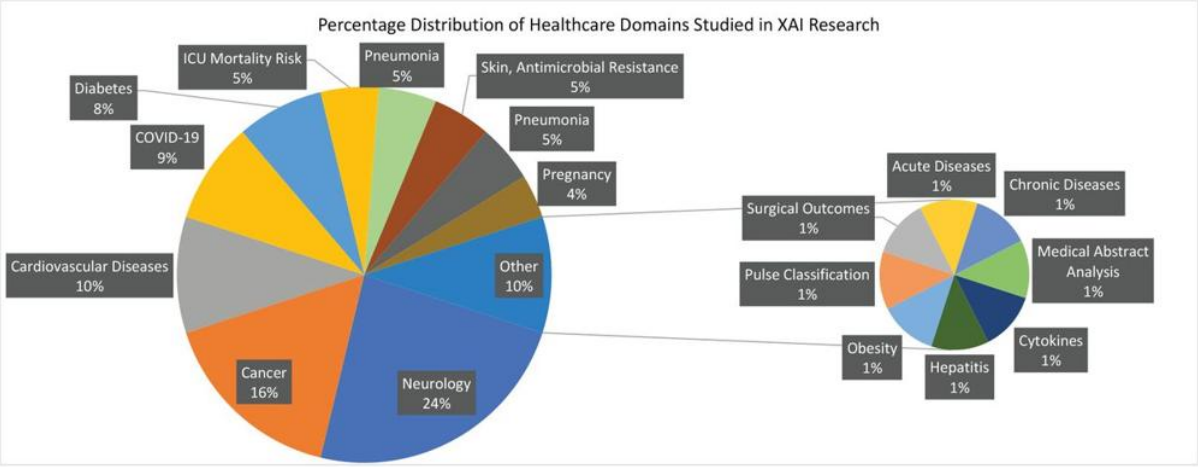

Figure 2: Percentage distribution of healthcare domains studied in recent XAI research. Neurology (24%) and Cancer (16%) are the most common focus areas, followed by Cardiovascular (10%), COVID-19 (9%), Diabetes (8%), ICU outcomes (5%), and others. Even domains like obstetrics, surgery outcomes, and rare diseases (grouped as "Other") appear in the XAI literature, indicating the wide scope of explainable AI applications in healthcare[6]. (Data source: Analysis of 89 XAI-in-healthcare publications, 2020–2024[6])

Figure 2 illustrates how XAI research is being applied across a variety of healthcare domains. It shows the distribution of recent XAI-in-healthcare studies by clinical area, demonstrating that explainability is a concern not just in one specialty but in many – from neurology and oncology to cardiology and beyond [6]. This broad applicability reinforces that the challenges and solutions we discuss in the next sections have wide relevance.

3. KEY CHALLENGES IN EXPLAINABLE AI FOR HEALTHCARE

Despite significant progress in XAI methods, several core challenges hinder the seamless integration of explainable AI into healthcare practice. These challenges stem from both technical limitations and human/organizational factors. We identify six major challenges [54]:

(1) Cultivating appropriate trust in AI (and avoiding under- or over-reliance),

(2) Achieving transparency with complex "black-box" models,

(3) Managing the trade-off between model complexity and interpretability,

(4) Meeting ethical and regulatory requirements,

(5) Ensuring data privacy, and

(6) Integrating XAI systems into clinical workflows.

These are not entirely independent – they interact in various ways – but we discuss them separately for clarity. Addressing each challenge is essential to move XAI in healthcare from research to real-world impact [55].

3.1 Building Trust without Overreliance

Trust is frequently cited as both a goal and a challenge for AI in healthcare. Clinicians must have *calibrated trust* in an AI system – meaning they trust it when it's correct and useful, but remain appropriately skeptical when it might be wrong. Achieving this balance is tricky. A lack of trust can lead to AI tools being ignored, whereas blind overreliance can be dangerous if clinicians defer to AI even when it contradicts their own judgment or has made an error [3][4].

To manage trust, XAI systems often convey not just an explanation but also some indication of confidence or uncertainty. This helps prevent over-trusting a wrong prediction. Moreover, training and user experience play a role: clinicians need to learn when and how to rely on the AI. Explanations can aid this learning by highlighting cases where the AI is on shaky ground. For example, an explanation that shows *"Few similar cases in training data; model extrapolating"* might alert a doctor to be extra cautious. Over time, as clinicians see the AI perform and explain itself, they develop a mental model of its reliability [56][57][58].

3.2 The Transparency Challenge (Opening the "Black Box")

Modern AI models, especially deep learning architectures, are often described as "black boxes" because their internal decision processes are not readily interpretable by humans. This opacity directly conflicts with the medical demand for transparency in decision-making. Clinicians and regulators alike want to know *how* an AI arrived at its conclusion, but with complex models it can be extremely difficult to provide a complete answer. This challenge is essentially at the heart of XAI: how to provide insight into black-box models [59].

Deep neural networks with millions of parameters or ensemble models with hundreds of decision trees do not lend themselves to simple interpretation. Most explainability methods in use (like feature importance rankings or saliency maps) provide partial transparency. They might highlight one aspect of the model's reasoning (e.g., which features were most influential for one prediction), but they don't reveal the full conditional logic or interactions inside the model. For example, a saliency map can show areas of an MRI that influenced a tumor detection model, but it won't tell the clinician *why those pixel patterns correspond to a tumor*. The clinician must infer that from their own knowledge. Similarly, feature attribution might say a sepsis prediction was 80% influenced by blood pressure and lactate, but the *interplay* of various factors in the model's latent space remains hidden [60].

One straightforward approach to the transparency issue is to use inherently interpretable models when possible. For certain problems, simpler models (like logistic regression with a handful of features) might suffice and are directly transparent. Indeed, some experts argue that for high-stakes decisions, one should use the simplest effective model and avoid unnecessary complexity – "*do not use a black box when a light grey box will do*" [5]. In practice, however, it's known that for many complex healthcare tasks (image analysis, high-dimensional genomics, etc.), simpler models cannot achieve the accuracy of complex models. Clinicians also demand high accuracy; they may not accept an interpretable model that has significantly lower performance in diagnosing cancer compared to a black-box model. This creates a tension between the desire for transparency and the need for top-notch performance (as discussed in the next subsection) [61][62].

3.3 Balancing Model Complexity and Interpretability

This challenge is closely tied to transparency but focuses on the tension between **model complexity vs. interpretability** and the associated trade-offs in performance. In many healthcare AI problems, the most accurate models are very complex (e.g., deep learning models with millions of parameters), while simpler, interpretable models might not achieve the same accuracy. This raises a fundamental question: How do we balance the need for high accuracy (to benefit patients and clinical outcomes) with the need for interpretability (to ensure understanding and trust) [63] ?

In scenarios where complex models significantly outperform simpler ones, abandoning complexity altogether could mean sacrificing clinical efficacy. For example, in radiology, deep CNNs have proven extremely powerful in image interpretation, whereas linear models or shallow decision trees would perform poorly on raw pixel data. In such cases, outright choosing an interpretable model over a black box might not be viable if accuracy is paramount (e.g., missing cancers). On the other hand, if a simpler model can achieve nearly the same accuracy, many argue it should be favored for its transparency – an embodiment of **"**as simple as possible, but no simpler" [64].

One approach to this challenge is to develop methods that make complex models more interpretable (thus shifting the trade-off curve). For instance, attention mechanisms (as discussed, they add some interpretability to neural networks), or post-hoc distillation where a complex model's behavior is approximated by a simpler model. Model distillation might create a simplified surrogate (say a decision tree or rule set) that mimics the complex model on the training data. If this surrogate is reasonably accurate, it can serve as an interpretable representation for users to understand general patterns in the complex model's decisions. However, if the complex model is highly nonlinear, any single surrogate might be too approximate. Nonetheless, such distilled models can sometimes achieve surprisingly good fidelity and have been used to explain complex models like ensembles in medical prognosis tasks [65].

3.4 Ethical and Regulatory Challenges

Healthcare is a highly regulated domain with strict ethical standards, and this poses unique challenges for AI explainability. Beyond the pragmatic issues of trust and transparency discussed above, there are specific ethical and legal expectations that AI systems must meet – many of which explicitly or implicitly require explainability [66].

One major impetus is the impending *regulatory requirements* for AI in healthcare. For instance, the European Union's AI Act (expected to be enacted around 2024) will classify most medical AI systems as "high-risk" and impose requirements such as transparency, traceability, and human oversight [7]. In practice, this means developers will need to ensure their AI systems can produce understandable explanations of their outputs, among other safeguards, to get approval in the EU market. Likewise, the FDA in the United States has published guidelines (e.g., the Proposed Regulatory Framework for AI/ML-based Software as a Medical Device) that emphasize the importance of algorithm transparency and the ability to audit AI decision-making. Explainability is becoming part of the regulatory checklist that AI developers must satisfy to deploy in healthcare settings.

There's also an *ethical mandate* for explainability in medicine. Clinicians have professional and moral obligations to make decisions in the best interest of patients and to be able to justify those decisions. If a clinician relies on an AI, one could argue that the clinician should understand the basis of the AI's recommendation to ensure it aligns with standard of care and is free of bias. The concept of "accountability" is crucial: if an AI system causes harm, how do we assign responsibility? Without explanations, it becomes hard to investigate errors or biases – was it a data issue, a model flaw, or misuse by a clinician? XAI helps by creating a record of *why* a certain recommendation was made, which is essential for post-hoc analysis of adverse events or discrepancies. From an ethical lens, some scholars argue there is a "right to an explanation" for patients when AI is involved in their care, analogous to how patients have a right to be informed about the reasoning behind medical decisions [67].

3.5 Data Privacy and Security

Data privacy is a critical concern in healthcare, and it presents a unique challenge for AI explainability. Healthcare datasets are rich and often needed in large quantity to train robust AI models, but patient information is protected by laws and ethical norms. Two main issues arise: *Training Data Privacy* and *Explanation Content Privacy [68]*.

Firstly, gathering and centralizing the huge datasets for training complex models can violate privacy or be blocked by regulations. Hospitals and clinics are often reluctant or legally unable to send patient data to a central location for AI development. This challenge has given rise to *Federated Learning* (FL) approaches, where models are trained in a distributed fashion (see Section 4.1). FL keeps data on-site and only shares model parameters or gradients, which is more privacy-friendly. Federated learning has been applied in healthcare scenarios like multi-hospital training of medical imaging models [8]. While FL addresses the training data privacy to a large extent, it introduces complexity in how to generate explanations. If the model is distributed, one must ensure that collecting information for an explanation (which

might involve querying the model or its components) does not inadvertently leak data from a particular site. For example, if an explanation method requires model gradients, those gradients might carry some information about local data. Researchers have begun exploring *Privacy-Preserving XAI* – ensuring that the act of explaining a prediction doesn't compromise privacy. One study noted that aggregation of interpretability metrics across federated nodes diluted some local patterns, but still provided useful global insights, highlighting a trade-off between privacy and granularity of explanation [69][70].

Secondly, even if a model is trained, the *explanation content* itself could pose privacy risks. Consider example-based explanations: if an AI says *"Patient X is similar to patient Y in the database who had outcome Z,"* and if patient Y could be identified from that info, that's a privacy breach. In a small pool of patients, even saying "this prediction was influenced by the patient's rare genetic mutation" might effectively reveal that patient's identity to someone with knowledge of who has that mutation. Thus, XAI methods must be careful not to disclose sensitive attributes that are not already known or necessary to explain the decision. Techniques such as using representative *synthetic* patient profiles or de-identified aggregate examples are ways to mitigate this risk [71].

3.6 Integration into Clinical Workflow

Even the most well-designed explainable AI system can fail to have impact if it is not effectively integrated into the clinical workflow. Healthcare professionals operate under intense time pressure, using established processes and tools. Introducing a new AI with additional information (explanations) can either streamline decisions or, if poorly integrated, become a distraction or burden. Thus, usability and workflow fit are crucial challenges [72].

One aspect is user interface and accessibility. Explanations should be presented in a manner that clinicians can quickly interpret. If a physician has to log into a separate system or dig through multiple screens to find an AI explanation, they might ignore it altogether. Ideally, explainable AI outputs should be embedded in the systems clinicians already use – for example, the Electronic Health Record (EHR) interface, radiology viewer, or patient monitoring dashboard. A practical instance: some EHR vendors have begun integrating risk scores with one-click access to a brief explanation. If a primary care doctor sees a pop-up that a patient has a "high risk of hospitalization" and can hover to see "Key factors: uncontrolled diabetes, frequent ED visits," that's immediate and useful. On the other hand, if the doctor just sees a number or an alert without context, they may not know what to do with it or may mistrust it [73][74].

Another issue is *time and cognitive load*. Clinicians often have mere minutes with each patient or in reviewing each image. Explanations must therefore be concise. Long-winded or overly detailed explanations can be counterproductive. For instance, a model might analyze 100 lab values – but telling the doctor all 100 contributed is not helpful. Instead, highlighting the top three contributors or an overall pattern is more effective. This is where careful design is needed: perhaps using graphical elements (like color coding high-impact features) or simple natural language summaries. There have been attempts to generate automatic brief

text explanations (like "The AI suspects sepsis mainly due to rising lactate and low blood pressure over 2 hours"). Such summaries can be read at a glance [75].

Alert fatigue is a serious concern in clinical environments. If AI systems produce too many alerts or explanations too frequently, they can become just another source of noise. It's essential that explanatory alerts are delivered only when significant and actionable. For example, if an AI constantly explains trivial predictions or things the clinician already knows, they'll tune it out. Integration strategy might involve setting thresholds for when an explanation is shown (e.g., only when the model's confidence is low, or when the risk score crosses a high threshold). Additionally, personalizing the explanation interface to the user's role can help. A nurse might want different information than a physician; a junior clinician might appreciate more detailed rationale, whereas an experienced specialist might only want the bottom-line [76].

4. EMERGING SOLUTIONS AND FUTURE DIRECTIONS

Researchers are actively developing innovative techniques to make AI in healthcare more explainable and address the challenges discussed in Section 3. In this section, we highlight several key emerging solutions:

(i) **Federated Learning** for privacy-preserving model training and explanation,

(ii) **Attention mechanisms and interpretable model architectures** that provide built-in explanations,

(iii) **Counterfactual explanations** that offer insight into how to change outcomes,

(iv) **Visualization and interactive tools** to present explanations effectively to users, and

(v) **Hybrid models** that combine data-driven learning with knowledge-based reasoning for better interpretability. Many of these approaches are complementary and can be combined in an AI system.

Table 2 below maps these solutions to the challenges they primarily target.

Table 2. Emerging XAI Solutions in Healthcare and the Challenges They Address

Solution	Description	Addresses Challenges
Federated Learning (FL)	Distributed training of models across multiple institutions' data without centralizing patient data. The global model is shared and can be used with explainability methods, while raw data remains private.	*Privacy (3.5):* Preserves patient confidentiality by keeping data on-site during training[8].
		Regulatory (3.4): Facilitates multi-center AI development in compliance with data protection laws.
		Workflow (3.6): Allows institutions to benefit from larger datasets and more robust models without altering local data governance or IT infrastructure, smoothing integration.

Attention Mechanisms & Self-Interpretable Models	Incorporating model components (e.g., attention layers in neural networks, prototype learning, sparse feature models) that provide interpretability as part of the model's operation. Designing models that output human-understandable intermediate features or explanations.	*Transparency (3.2):* Makes deep models more transparent by design, highlighting which parts of input the model is focusing on[5].
		Complexity vs Interpretability (3.3): Retains much of the accuracy of complex models while offering more interpretable reasoning (e.g., attention weights on clinical text, prototypical cases for comparison).
		Trust (3.1): By revealing model attention or logic in human-comprehensible terms, clinicians can better trust the model's focus aligns with clinical cues.
Counterfactual Explanations	Generating "what-if" explanations that show how a model's output would change if certain input features were different. These highlight minimal changes needed to alter a prediction (e.g., showing how a patient's risk would drop if a particular risk factor were modified).	*User Understanding & Trust (3.1):* Provides insight in a causal, action-oriented way, helping clinicians and patients understand crucial factors and potential interventions[12].
		Transparency (3.2): Reveals aspects of the model's decision boundary (what conditions would lead to a different outcome).
		Ethical Use (3.4): Aligns with informed decision-making by illustrating reasons and possible changes – useful for patient communications and ethical justification of decisions.
Visualization & Interactive Tools	User interfaces and tools that integrate explanations into clinical workflows – including dashboards with visual annotations (graphs, heatmaps), interactive plots, or natural-language summaries. Users can explore or query the model's reasoning (for example, adjust input values to see predicted outcome changes).	*Workflow Integration (3.6):* Improves usability by presenting explanations in a concise, intuitive format within existing clinical software (EHRs, imaging viewers)[10].
		Trust (3.1): Enhances trust by making the AI's reasoning process visible and tangible (e.g., a highlighted tumor region on a scan, or a trend graph explaining a risk score)[3].
		Transparency & Debugging (3.2): Allows users to interact with the model (simulate scenarios, drill down into features), facilitating deeper understanding and error analysis.

Hybrid Models (Knowledge + ML)	Combining interpretable knowledge-based components (like rule-based systems or causal models) with machine learning models. Examples include: rule-augmented neural networks, models that follow clinical guidelines unless data suggests otherwise, or distilling a complex model into an interpretable surrogate.	*Interpretability vs Accuracy (3.3):* Aims to achieve high accuracy of ML while maintaining interpretability via human-understandable logic[5]. The knowledge component can catch known patterns, while ML handles exceptions.
		Ethical/Regulatory (3.4): Knowledge-driven parts provide transparent reasoning (e.g., "alarm triggered because BP < 90 per clinical protocol"), satisfying demands for clear justification.
		Trust (3.1): Clinicians trust systems that incorporate medical knowledge and guidelines, and are more likely to accept AI recommendations that come with rule-based explanations or familiar reasoning steps.

These emerging approaches are being actively researched and in some cases implemented in prototype or even commercial systems. In the rest of this section, we delve into each solution area, providing examples of how they work and citing recent studies that demonstrate their potential in healthcare.

4.1 Federated Learning and Privacy-Preserving XAI

Federated learning (FL) has emerged as a powerful approach to train AI models on distributed healthcare data while addressing privacy concerns. In a federated setup, multiple hospitals or institutions collaboratively train a shared model but each institution's data remains on its local servers. Only model updates (like gradient information or trained weights) are exchanged and aggregated. This approach has been successfully demonstrated in scenarios such as multi-hospital medical imaging classification, digital health wearable data analysis, and outcome prediction models that use data from different clinics [8]. By design, FL tackles the privacy challenge (3.5) because patient data never leaves the source institution in identifiable form [77].

The good news for explainability is that once a federated model is trained, it can be used with most of the same XAI techniques as a centrally trained model. For example, if a federated model is deployed at Hospital A, clinicians at that hospital can input their patient's data and get a prediction and an explanation (feature importance, etc.) just as they would with a model trained on centralized data. The explanation is based on a model that has learned from a much broader dataset (all hospitals participating), so it may actually be more reliable and comprehensive than a model trained only on Hospital A's data [78].

One key consideration is ensuring that the explanation process itself does not leak information about other institutions' data. Generally, if one is just explaining a single

patient's prediction with respect to the final model, there's no direct privacy issue – the model weights might contain some traces of other data, but nothing more than what the prediction already uses. However, if a user at Hospital A tried to query the model's behavior on inputs outside the distribution of Hospital A (perhaps to infer something about Hospital B's data patterns), that could be problematic. Techniques like *secure aggregation* and differential privacy are sometimes integrated into FL to ensure that even model updates don't reveal individual data points [8]. When it comes to explanations, researchers are looking at whether certain explanation outputs (like gradients for a particular feature across all hospitals) could inadvertently expose differences that correlate with private info. So far, approaches to mitigate this include sharing only aggregated or non-sensitive explanation-related information. For instance, Lopez-Ramos *et al.* (2024) mapped out publications on FL+XAI and noted that only a minority explicitly tackled how FL influences explanations; one insight was that overly granular interpretability metrics might reveal site-specific patterns, so focusing on global patterns (which is usually enough for interpretation) maintains privacy[79][80].

A practical application of federated learning with XAI is in medical imaging: suppose 10 hospitals train a federated model for detecting a certain rare cancer on MRI scans. Each hospital gets the final model. A radiologist at Hospital X uses it on a patient's scan and it says "high likelihood of tumor." Using XAI (like Grad-CAM), the radiologist sees a heatmap highlighting a lesion area. This process didn't require any other hospital's data at inference or explanation time – the model was already trained. The radiologist benefits from a model effectively trained on data from 10 hospitals (maybe thousands of MRIs), which is much more robust, and still gets a clear explanation (highlighted region) just as they would if it were a model trained at their own site. From the patient's perspective, their data stayed at Hospital X and only the model (which is not patient-specific data) was shared and returned, so privacy is intact [81].

4.2 Attention Mechanisms and Interpretable Model Design

To tackle the transparency and interpretability challenges, a promising direction is to design AI models that are more **self-explanatory**. Instead of treating explainability as an afterthought, these approaches bake interpretability into the model architecture. Two key trends here are the use of *attention mechanisms* in deep learning models and the development of inherently interpretable or constrained models.

Attention mechanisms: In neural networks, attention modules allow the model to weight different parts of the input when making a decision. For example, in a sequence model analyzing a patient's history, an attention layer might assign higher weight to recent events (e.g., a spike in heart rate) and lower weight to older or less relevant events. The resulting attention weights can be visualized as a heat map over the timeline, showing which time points influenced the prediction of, say, a septic shock onset. Clinicians can look at that and often it makes intuitive sense (e.g., "the model paid most attention to the moment when blood pressure dropped and lactate rose," which aligns with clinical reasoning for sepsis). In NLP applications like clinical text analysis, attention can highlight the words in a doctor's note

that led an AI to flag possible diagnosis. For instance, if an AI reading an ER triage note predicts a high risk of stroke, the attention weights might highlight terms like "slurred speech" and "left arm weakness." This provides an immediate clue to the clinician why the model thought of stroke, essentially surfacing *which parts of the input the model considered important*.

Attention mechanisms have been successfully used in models like RETAIN (Reverse Time Attention Network) for healthcare, which provides interpretable attention on past visits for risk prediction, and in various clinical NLP models for cohort selection or outcome prediction where they highlight key symptoms in text [5]. A caveat: attention is not a perfect explanation – there's an ongoing debate in ML about whether attention weights always correlate with feature importance. But in practice, they often improve interpretability and, importantly, allow some user control (e.g., a clinician might say: if the model isn't attending to symptom X that I consider important, maybe the model is flawed or the documentation was incomplete) [82].

Beyond attention, there are model architectures aimed at **interpretability by design**. Examples include:

- **Generalized Additive Models with splines or neural components (GA^2M)**: These models learn a contribution function for each feature (or pair of features) that is human-reviewable. One such model was deployed to predict pneumonia risk and provided doctors with graphs per feature like "risk vs age" which were monotonic and made clinical sense (they could see risk rises sharply after age 65, etc.). Interpretable models like this have been used in some healthcare settings where transparency is paramount, like predicting readmission risk using a few key features and showing those contributions [5].
- **Prototype and case-based models**: As mentioned, some deep models learn "prototypes" of each class. In medical imaging, a prototype might be a typical example of a certain lesion. When diagnosing a new image, the model can say "this image is classified as melanoma because it contains a patch that looks 90% similar to prototype #7 (a known melanoma image)." Sun *et al.* (2022) surveyed such approaches in medical image analysis where the internal representation is aligned with concepts or examples that humans can recognize[5]. This gives a very tangible explanation: actual similar past images. In text, a prototype could be an exemplar patient case description that the new case resembles.
- **Sparse or rule-based neural networks**: Some research tries to make neural networks mimic decision trees or rule sets. One method is to enforce sparsity (so only a small number of neurons fire significantly for a given input), effectively creating a kind of rule: "if these specific pattern neurons activate, then outcome = X." These can sometimes be interpreted post-hoc as rules.
- **Causal or disentangled representations**: There's a push to structure models so that internal features have meanings. For example, a model might be trained in a multi-task way to predict not just an outcome but also intermediate clinical concepts (e.g., "Is there lung opacity? Yes/No" as an auxiliary output when detecting pneumonia). By doing so, the model has to

explicitly compute that concept, which can be extracted as an explanation (and also validated against radiologist annotations). This is being seen in medical imaging: networks that output not only the predicted diagnosis but also a set of human-labelable features (like tumor size, location, presence of certain markers), thereby providing a conceptual basis for the decision.

Interpretable models are particularly appealing to regulators (addressing challenge 3.4) because one can often *formally verify* certain properties (like monotonicity with risk factors, which ensures no bizarre inverse correlations). They also often generalize better in scenarios where data is limited, by incorporating domain knowledge (for instance, an interpretable model might enforce that some features only add positive risk, which could reduce overfitting).

However, inherently interpretable models may require more effort to design for each application. They might not be as plug-and-play as a standard deep learning model. There is also sometimes a small loss in raw accuracy compared to an unconstrained model. But as discussed in the trade-off section, this loss may be offset by gains in human-AI team performance.

4.3 Counterfactual Explanations and Causal Insights

Counterfactual explanations have gained attention as an intuitive and useful form of explainability, particularly in domains like healthcare where one often asks "what could have been done differently?" A counterfactual explanation describes how an outcome would change if certain input factors were different. In other words, it answers questions of the form: *"If X (input) had not been true (or had been true), what would the model prediction be?"* This provides insight into the model's decision-making in a way that is naturally aligned with human reasoning about cause and effect.

For example, consider an AI system that predicts a high risk of hospital readmission for a patient. A counterfactual explanation might be: *"If the patient had someone to assist at home (social support feature changed) and attended a follow-up within 7 days, the model's predicted readmission risk would drop from 20% to 5%."* This kind of explanation is powerful: it not only tells which factors are important (lack of social support, no early follow-up) but also indicates potential actions (arrange home support, schedule an early follow-up appointment) that could change the outcome. Such information is incredibly valuable to clinicians and care managers, aligning the AI's insight with actionable steps [83].

Counterfactual explanations directly address trust and understanding because they frame the model's reasoning in terms of real-world changes and consequences. Clinicians often think in terms of counterfactuals: *"Would this patient still have deteriorated if we had started antibiotics earlier?"* If an AI can provide input on that (based on learned patterns), it engages with the clinician's mode of thinking. Patients, too, may find counterfactuals easier to understand: *"What can I change to reduce my risk?"* is a common patient question. An AI-driven explanation like "If you lose 10 kg, your predicted diabetes risk drops significantly" speaks directly to that.

From a technical perspective, generating counterfactual explanations can be challenging because the system must find a plausible input change that alters the outcome. There are methods like DiCE (Diverse Counterfactual Explanations) that use optimization to find small changes in input features that would flip the model's prediction [12]. In healthcare, one must ensure these changes are *feasible* and *realistic*. Changing "age" by -10 years is not feasible, so age wouldn't be a useful counterfactual suggestion. But changing "blood pressure" or "smoking status" is plausible (blood pressure via medication, smoking by cessation). Therefore, counterfactual generation methods often have constraints to suggest only actionable feature changes. In clinical settings, domain knowledge is used to filter counterfactuals (e.g., only suggest things like lab value improvements, medication adherence, lifestyle factors, or timely interventions, and not things like altering immutable traits or introducing impossible scenarios).

A concrete research example: an intensive care predictive model for septic shock was enhanced with counterfactual explanation capability. For a given patient with a high predicted risk, it could output: *"If fluid intake in the next hour were increased by 500 ml, the risk would decrease by X%."* This was derived from the model recognizing volume depletion as a risk factor, and the counterfactual is simulating the effect of an intervention (fluid bolus). This kind of explanation, if reliable, could actually assist clinicians in deciding interventions (though caution is needed to trust such suggestions, it's an emerging area for decision support).

Counterfactuals are also useful for auditing fairness: for instance, one can examine if the counterfactual changes differ systematically by protected attributes. If, say, for minority patients the model's counterfactual often is "if you were of a different race, outcome changes," that would expose a serious bias. Ideally, counterfactual reasons should revolve around clinical factors, not things like race or gender (unless medically relevant). In a fair model, changing race or ethnicity should *not* alter the prediction in the counterfactual sense – a point some studies check to ensure model fairness.

4.4 Visualization and Interaction Tools for Explainability

While sophisticated algorithms under the hood are crucial, the end utility of XAI in healthcare often comes down to how well the information is conveyed to the user – typically a clinician or sometimes a patient. This is where visualization and interactive tools play a pivotal role. A well-designed interface can make complex explanations immediately understandable, whereas a poor interface can render even a good explanation technique useless.

Visual explanation tools present model insights in intuitive formats. For example, in medical imaging, the standard approach is to overlay heatmaps or contour highlights on the image to show regions the model found important. Many FDA-approved AI tools for radiology (for detection of nodules, fractures, etc.) provide such visual cues – these serve as a form of explanation, giving radiologists confidence about *where* the AI is "looking." Similarly, in pathology AI, tools highlight cells or regions of interest. These visual explanations are part of

workflow: a pathologist might scan slide thumbnails with AI highlights to decide which region to examine first under the microscope.

For non-image data, visualization might mean charts or graphs. For instance, a risk prediction system could show a bar chart of the top five contributing factors to a risk score (with lengths proportional to their contribution). If "elevated HbA1c" has a long bar for a diabetes complication risk, a clinician instantly knows glycemic control is a major issue for this patient. Some systems use *waterfall plots* or *force plots* (as in SHAP library outputs) to break down how various features push the prediction up or down from the baseline. While these require some training to interpret, once users learn them, they can parse the information quickly. A study found that presenting clinicians with a simple bar chart of feature importances for each prediction helped them identify when the model might be off (e.g., if the chart showed something nonsensical as top factor, they knew to be skeptical) and increased their acceptance when the chart matched their own reasoning[3].

Interactive dashboards take it a step further by allowing the user to query and manipulate the model's reasoning process. An example is the "What-If Tool" (by Google PAIR) applied to a clinical dataset: a clinician can adjust input values via sliders and see how the model's prediction changes. This not only provides counterfactual insights (as discussed) but also engages the user in exploring the model. It demystifies the model as something they can poke at and understand, rather than a one-way output generator. Interactive tools can also allow filtering: e.g., a doctor might filter similar past patients from the database to see outcomes ("Show me past cases similar to this one with high risk – what happened to them?"). If integrated with a hospital's data, this becomes a kind of case-based explanation combined with model prediction.

Natural language explanations are another front. There are prototypes where AI models generate a brief text explanation along with a prediction. For example, a cardiology AI might output: "Predicted risk: 85%. Explanation: patient's age (76 years) and history of atrial fibrillation indicate high stroke risk according to learned patterns." This reads almost like a doctor's note or an excerpt from guidelines. Using language-generation models or templating, these explanations can be made quite fluent. Some recent works use large language models (LLMs) to "explain" the outputs of a separate prediction model – effectively translating the model's logic into a narrative. Caution is needed because if not grounded, an LLM might hallucinate reasons. But when constrained by the model's actual feature attributions, it can produce human-like rationales that are easier for clinicians to absorb quickly, as shown in some early studies on AI-aided report writing.

From a workflow standpoint, integration means these visual or interactive explanation elements should appear in the software the clinician already uses. For instance, some hospital EHR systems have incorporated risk scores for things like sepsis (e.g., a sepsis early warning system). Initially, many just showed a numeric score or color code. Newer iterations are beginning to include a drop-down or pop-up that lists top factors contributing to that sepsis score (like "WBC high, HR high, hypotension present"). This is an improvement influenced

by user feedback and XAI research. It means the physician or nurse doesn't have to guess why the alert fired – they see key evidence at a glance.

4.5 Hybrid Models and Knowledge-Guided AI

Hybrid modeling refers to AI systems that combine traditional, knowledge-based decision logic with modern machine learning. The idea is to get the best of both worlds: the *interpretability* and *prior knowledge* of expert systems or causal models, and the *flexibility and performance* of statistical learning. In healthcare, a vast amount of domain knowledge exists (clinical guidelines, pathophysiological principles, etc.), so ignoring it and purely relying on brute-force learning may not be optimal or necessary. Hybrid approaches can embed this knowledge to guide the model and provide human-understandable rationales.

One simple form of hybrid model is a *Rule-Based Scaffold* around an ML model. For example, a hospital might use a set of if-then rules (derived from policy or experience) to handle very critical conditions – ensuring the AI doesn't inadvertently override obvious clinical protocol. For instance, a hospital's sepsis alert system might be hybrid: it follows a known screening rule (e.g., SIRS criteria and lactate level thresholds) to trigger an alert, but also incorporates an ML model that fine-tunes the risk prediction. The rule-based part guarantees that no high-risk case is missed according to standard criteria (and provides a clear rationale: "triggered due to low BP and high HR"), while the ML part improves specificity by analyzing patterns beyond the rule (with its own explanation). Clinicians thus get an alert that might say: *"Sepsis Alert (Rule-based criteria met); ML model concurs with high risk due to worsening SOFA score."* This dual explanation (rule justification + model insight) is very transparent.

5. CONCLUSION

In conclusion, explainable AI has moved from a theoretical ideal to a practical necessity in healthcare AI systems. Through techniques like federated learning, attention-based models, counterfactual reasoning, visualization tools, and hybrid designs, we are beginning to unlock the "black boxes" of medical AI. These innovations enable AI systems to provide not just predictions, but also the reasoning behind them – a development that is essential for aligning AI with the values and workflows of medicine. The future of AI in healthcare will likely be one in which clinicians routinely interact with AI as a partner: querying its reasoning, getting clarifications, and combining its learned knowledge with their own expertise. Such a future promises AI tools that are *transparent*, *trustworthy*, and *human-aligned*, ultimately leading to better and more accountable patient care.

REFERENCES

[1] Sagona, M., Dai, T., Macis, M., & Darden, M. (2025). Trust in AI-assisted health systems and AI's trust in humans. *npj Health Systems*, *2*(1), 10.

[2] Ferguson, S., Aoyagui, P. A., Rizvi, R., Kim, Y. H., & Kuzminykh, A. (2024). The Explanation That Hits Home: The Characteristics of Verbal Explanations That Affect Human Perception in Subjective Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1-37.

[3] Rosenbacke, R. (2024). How Explainable Artificial Intelligence Can Increase or Decrease. *COGNITIVE CHALLENGES IN HUMAN-AI COLLABORATION*, 75.

[4] Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., ... & Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, *4*(1), 31.

[5] Kamei, K., & Yuge, K. (2022). Canonical Nonlinearity for Coupled Linear Systems. *arXiv preprint arXiv:2210.11609*.

[6] Khanam, R., Hussain, M., Hill, R., & Allen, P. (2024). A comprehensive review of convolutional neural networks for defect detection in industrial applications. *IEEE Access*.

[7] Freyer, N., Groß, D., & Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. *BMC Medical Ethics*, *25*(1), 104.

[8] Lopez-Ramos, L. M., Leiser, F., Rastogi, A., Hicks, S., Strümke, I., Madai, V. I., ... & Hilbert, A. (2024). Interplay between Federated Learning and Explainable Artificial Intelligence: a Scoping Review. *arXiv preprint arXiv:2411.05874*.

[9] P. Garg, A. Dixit and P. Sethi, "Ml-fresh: novel routing protocol in opportunistic networks using machine learning," *Computer Systems Science and Engineering*, vol. 40, no.2, pp. 703–717, 2022. DOI:10.32604/csse.2022.019557

[10] Yadav, P. S., Khan, S., Singh, Y. V., Garg, P., & Singh, R. S. (2022). A Lightweight Deep Learning-Based Approach for Jazz Music Generation in MIDI Format. *Computational Intelligence and Neuroscience*, *2022*. DOI:10.1155/2022/2140895

[11] Hossain, M. I., Zamzmi, G., Mouton, P. R., Salekin, M. S., Sun, Y., & Goldgof, D. (2025). Explainable AI for medical data: current methods, limitations, and future directions. *ACM Computing Surveys*, *57*(6), 1-46.

[12] Xu, R., Yu, Y., Zhang, C., Ali, M. K., Ho, J. C., & Yang, C. (2022, November). Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health* (pp. 259-278). PMLR.

[13] Soni, E., Nagpal, A., Garg, P., & Pinheiro, P. R. (2022). Assessment of Compressed and Decompressed ECG Databases for Telecardiology Applying a Convolution Neural Network. *Electronics, 11*(17), 2708. DOI: 10.3390/electronics11172708

[14] Pustokhina, I. V., Pustokhin, D. A., Lydia, E. L., Garg, P., Kadian, A., & Shankar, K. (2021). Hyperparameter search based convolution neural network with Bi-LSTM model for intrusion detection system in multimedia big data environment. *Multimedia Tools and Applications*, 1-18. https://doi.org/10.1007/s11042-021-11271-7

[15] Wells, P. S., Anderson, D. R., Rodger, M., Ginsberg, J. S., Kearon, C., Gent, M., ... & Hirsh, J. (2000). Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thrombosis and haemostasis*, *83*(03), 416-420.

[16] Stein, P. D., Hull, R. D., Saltzman, H. A., & Pineo, G. (1993). Strategy for diagnosis of patients with suspected acute pulmonary embolism. *Chest*, *103*(5), 1553-1559.

[17] Khanna, A., Rani, P., Garg, P., Singh, P. K., & Khamparia, A. (2021). An Enhanced Crow Search Inspired Feature Selection Technique for Intrusion Detection Based Wireless Network

System. *Wireless Personal Communications*, 1-18. https://doi.org/10.1007/s11277-021-09144-1

[18] Garg, P., Dixit, A., Sethi, P., & Pinheiro, P. R. (2020). Impact of node density on the qos parameters of routing protocols in opportunistic networks for smart spaces. *Mobile Information Systems, 2020*. https://doi.org/10.1155/2020/8868842

[19] Upadhyay, D., Garg, P., Aldossary, S. M., Shafi, J., & Kumar, S. (2023). A Linear Quadratic Regression-Based Synchronised Health Monitoring System (SHMS) for IoT Applications. Electronics, 12(2), 309. https://doi.org/10.3390/electronics12020309

[20] Saini, P., Nagpal, B., Garg, P., & Kumar, S. (2023). CNN-BI-LSTM-CYP: A deep learning approach for sugarcane yield prediction. *Sustainable Energy Technologies and Assessments*, *57*, 103263. https://doi.org/10.1016/j.seta.2023.103263

[21] Saini, P., Nagpal, B., Garg, P., & Kumar, S. (2023). Evaluation of Remote Sensing and Meteorological parameters for Yield Prediction of Sugarcane (Saccharumofficinarum L.) Crop. Brazilian Archives of Biology and Technology, 66, e23220781. DOI: 10.1590/1678-4324-errata-2024999903

[22] Beniwal, S., Saini, U., Garg, P., & Joon, R. K. (2021). Improving performance during camera surveillance by integration of edge detection in IoT system. *International Journal of E-Health and Medical Communications (IJEHMC)*, *12*(5), 84-96. DOI: 10.4018/IJEHMC.20210901.oa6

[23] Garg, P., Dixit, A., & Sethi, P. (2019). Wireless sensor networks: an insight review. *International Journal of Advanced Science and Technology*, *28*(15), 612-627. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/1838

[24] Sharma, N., & Garg, P. (2022). Ant colony based optimization model for QoS-Based task scheduling in cloud computing environment. *Measurement: Sensors*, 100531. https://doi.org/10.1016/j.measen.2022.100531

[25] Pawan Kumar, Rakesh Kumar, Puneet Garg. (2020). Hybrid Crowd Cloud Routing Protocol For Wireless Sensor Networks. *International Journal of Advanced Science and Technology*, *29*(12s), 766 - 775. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/22539

[26] Raj, G. ., Verma, A. ., Dalal, P. ., Shukla, A. K. ., & Garg, P. . (2023). Performance Comparison of Several LPWAN Technologies for Energy Constrained IOT Network. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(1s), 150–158. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/2487

[27] Garg, P. ., Sharma, N. ., Sonal, & Shukla, B. . (2023). Predicting the Risk of Cardiovascular Diseases using Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(2s), 165 –. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/2520

[28] Patil, S. C. ., A. Mane, D. ., Singh, M. ., Garg, P. ., Desai, A. B. ., & Rawat, D. . (2023). Parkinson's Disease Progression Prediction Using Longitudinal Imaging Data and Grey Wolf Optimizer-Based Feature Selection. *International Journal of Intelligent Systems and*

Applications in Engineering, *12*(3s), 441–451. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3725

[29] Gudur, A. ., Pati, P. ., Garg, P., & Sharma, N. . (2023). Radiomics Feature Selection for Lung Cancer Subtyping and Prognosis Prediction: A Comparative Study of Ant Colony Optimization and Simulated Annealing. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 553–565. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3735

[30] Dixit, A., Garg, P., Sethi, P., & Singh, Y. (2020, April). TVCCCS: Television Viewer's Channel Cost Calculation System On Per Second Usage. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012046). IOP Publishing. Doi:10.1088/1757-899X/804/1/012046

[31] Sethi, P., Garg, P., Dixit, A., & Singh, Y. (2020, April). Smart number cruncher–a voice based calculator. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012041). IOP Publishing. doi:10.1088/1757-899X/804/1/012041

[32] S. Rai, V. Choubey, Suryansh and P. Garg, "A Systematic Review of Encryption and Keylogging for Computer System Security," *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 2022, pp. 157-163, doi: 10.1109/CCiCT56684.2022.00039.

[33] L. Saraswat, L. Mohanty, P. Garg and S. Lamba, "Plant Disease Identification Using Plant Images," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 79-82, doi: 10.1109/CCiCT56684.2022.00026.

[34] L. Mohanty, L. Saraswat, P. Garg and S. Lamba, "Recommender Systems in E-Commerce," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 114-119, doi: 10.1109/CCiCT56684.2022.00032.

[35] C. Maggo and P. Garg, "From linguistic features to their extractions: Understanding the semantics of a concept," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 427-431, doi: 10.1109/CCiCT56684.2022.00082.

[36] N. Puri, P. Saggar, A. Kaur and P. Garg, "Application of ensemble Machine Learning models for phishing detection on web networks," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 296-303, doi: 10.1109/CCiCT56684.2022.00062.

[37] R. Sharma, S. Gupta and P. Garg, "Model for Predicting Cardiac Health using Deep Learning Classifier," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 25-30, doi: 10.1109/CCiCT56684.2022.00017.

[38] Varshney, S. Lamba and P. Garg, "A Comprehensive Survey on Event Analysis Using Deep Learning," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 146-150, doi: 10.1109/CCiCT56684.2022.00037.

[39] Dixit, A., Sethi, P., Garg, P., & Pruthi, J. (2022, December). Speech Difficulties and Clarification: A Systematic Review. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 52-56). IEEE. DOI: 10.1109/SMART55829.2022.10047048

[40] Garg, P., Dixit, A., Sethi, P., & Pruthi, J. (2023). *Strengthening Smart City with Opportunistic Networks: An Insight*. https://doi.org/10.1109/icacctech61146.2023.00117

[41] Rana, S., Chaudhary, R., Gupta, M., & Garg, P. (2023, December). Exploring Different Techniques for Emotion Detection Through Face Recognition. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 779-786). IEEE. DOI: 10.1109/ICACCTech61146.2023.00128

[42] Mittal, K., Srivastava, K., Gupta, M., & Garg, P. (2023, December). Exploration of Different Techniques on Heart Disease Prediction. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 758-764). IEEE. **DOI:** 10.1109/ICACCTech61146.2023.00125

[43] Gautam, V. K., Gupta, S., & Garg, P. (2024, March). Automatic Irrigation System using IoT. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 100-103). IEEE. **DOI:** 10.1109/AUTOCOM60220.2024.10486085

[44] Ramasamy, L. K., Khan, F., Joghee, S., Dempere, J., & Garg, P. (2024, March). Forecast of Students' Mental Health Combining an Artificial Intelligence Technique and Fuzzy Inference System. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 85-90). IEEE. https://doi.org/10.1109/AUTOCOM60220.2024.10486194

[45] Rajput, R., Sukumar, V., Patnaik, P., Garg, P., & Ranjan, M. (2024, March). The Cognitive Analysis for an Approach to Neuroscience. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 524-528). IEEE. **DOI:** 10.1109/AUTOCOM60220.2024.10486081

[46] Dixit, A., Sethi, P., Garg, P., Pruthi, J., & Chauhan, R. (2024, July). CNN based lip-reading system for visual input: A review. In *AIP Conference Proceedings* (Vol. 3121, No. 1). AIP Publishing. https://doi.org/10.1063/5.0221717

[47] Bose, D., Arora, B., Srivastava, A. K., & Garg, P. (2024, May). A Computer Vision Based Framework for Posture Analysis and Performance Prediction in Athletes. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)* (pp. 942-947). IEEE. DOI: 10.1109/IC3SE62002.2024.10593041

[48] Singh, M., Garg, P., Srivastava, S., & Saggu, A. K. (2024, April). Revolutionizing Arrhythmia Classification: Unleashing the Power of Machine Learning and Data Amplification for Precision Healthcare. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 516-522). IEEE. **DOI:** 10.1109/CCICT62777.2024.00086

[49] Kumar, R., Das, R., Garg, P., & Pandita, N. (2024, April). Duplicate Node Detection Method for Wireless Sensors. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 512-515). IEEE. **DOI:** 10.1109/CCICT62777.2024.00085

[50] Bhardwaj, H., Das, R., Garg, P., & Kumar, R. (2024, April). Handwritten Text Recognition Using Deep Learning. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 506-511). IEEE. **DOI:** 10.1109/AIST55798.2022.10065348

[51] Gill, A., Jain, D., Sharma, J., Kumar, A., & Garg, P. (2024, May). Deep learning approach for facial identification for online transactions. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 715-722). IEEE. **DOI:** 10.1109/INNOCOMP63224.2024.00123

[52] Mittal, H. K., Dalal, P., Garg, P., & Joon, R. (2024, May). Forecasting Pollution Trends: Comparing Linear, Logistic Regression, and Neural Networks. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 411-419). IEEE. DOI: 10.1109/INNOCOMP63224.2024.00074

[53] Malik, T., Nandal, V., & Garg, P. (2024, May). Deep Learning-Based Classification of Diabetic Retinopathy: Leveraging the Power of VGG-19. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 645-651). IEEE. DOI: 10.1109/INNOCOMP63224.2024.00111

[54] Srivastava, A. K., Verma, I., & Garg, P. (2024, May). Improvements in Recommendation Systems Using Graph Neural Networks. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 668-672). IEEE. **DOI:** 10.1109/INNOCOMP63224.2024.00115

[55] Aggarwal, A., Jain, D., Gupta, A., & Garg, P. (2024, May). Analysis and Prediction of Churn and Retention Rate of Customers in Telecom Industry Using Logistic Regression. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 723-727). IEEE. **DOI:** 10.1109/INNOCOMP63224.2024.00124

[56] Mittal, H. K., Arsalan, M., & Garg, P. (2024, May). A Novel Deep Learning Model for Effective Story Point Estimation in Agile Software Development. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 404-410). IEEE. DOI: 10.1109/INNOCOMP63224.2024.00073

[57] Chaudhary, A., & Garg, P. (2014). Detecting and diagnosing a disease by patient monitoring system. *International Journal of Mechanical Engineering And Information Technology*, *2*(6), 493-499.

[58] Malik, K., Raheja, N., & Garg, P. (2011). Enhanced FP-growth algorithm. *International Journal of Computational Engineering and Management*, *12*, 54-56.

[59] Garg, P., Dixit, A., & Sethi, P. (2021, May). Link Prediction Techniques for Opportunistic Networks using Machine Learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*. http://dx.doi.org/10.2139/ssrn.3842849

[60] Garg, P., Dixit, A., & Sethi, P. (2021, April). Opportunistic networks: Protocols, applications & simulation trends. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*. http://dx.doi.org/10.2139/ssrn.3834099

[61] Garg, P., Dixit, A., & Sethi, P. (2021). Performance comparison of fresh and spray & wait protocol through one simulator. *IT in Industry*, *9*(2). https://doi.org/10.17762/itii.v9i2.369

[62] Malik, M., Singh, Y., Garg, P., & Gupta, S. (2020). Deep Learning in Healthcare system. *International Journal of Grid and Distributed Computing*, *13*(2), 469-468.

[63] Gupta, M., Garg, P., Gupta, S., & Joon, R. (2020). A Novel Approach for Malicious Node Detection in Cluster-Head Gateway Switching Routing in Mobile Ad Hoc Networks. *International Journal of Future Generation Communication and Networking*, *13*(4), 99-111.

[64] Gupta, A., Garg, P., & Sonal, Y. S. (2020). Edge Detection Based 3D Biometric System for Security of Web-Based Payment and Task Management Application. *International Journal of Grid and Distributed Computing*, *13*(1), 2064-2076.

[65] Pawan Kumar, Rakesh Kumar, Puneet Garg. (2020). Hybrid Crowd Cloud Routing Protocol For Wireless Sensor Networks. *International Journal of Advanced Science and Technology*, *29*(12s), 766 - 775. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/22539

[66] Garg, P., & Raman, P. K. Broadcasting Protocol & Routing Characteristics With Wireless ad-hoc networks.

[67] Garg, P., Arora, N., & Malik, T. Capacity Improvement of WI-MAX In presence of Different Codes WI-MAX: Speed & Scope of future.

[68] Garg, P., Saroha, K., & Lochab, R. (2011). Review of wireless sensor networks-architecture and applications. IJCSMS International Journal of Computer Science & Management Studies, 11(01), 2231-5268.

[69] Yadav, S., &Garg, P. Development of a New Secure Algorithm for Encryption and Decryption of Images.

[70] Dixit, A., Sethi, P., & Garg, P. (2022). Rakshak: A Child Identification Software for Recognizing Missing Children Using Machine Learning-Based Speech Clarification. International Journal of Knowledge-Based Organizations (IJKBO), 12(3), 1-15. https://doi.org/10.4018/IJKBO.299968

[71] Shukla, N., Garg, P., & Singh, M. (2022). MANET Proactive and Reactive Routing Protocols: A Comparison Study. International Journal of Knowledge-Based Organizations (IJKBO), 12(3), 1-14. https://doi.org/10.4018/IJKBO.299970

[72] Arya, A., Garg, P., Vellanki, S., Latha, M., Khan, M. A., & Chhbra, G. (2024). Optimisation Methods Based on Soft Computing for Improving Power System Stability. *Journal of Electrical Systems*, *20*(6s), 1051-1058. https://doi.org/10.52783/jes.2837

[73] Chauhan, S., Singh, M., & Garg, P. (2021). Rapid Forecasting of Pandemic Outbreak Using Machine Learning. *Enabling Healthcare 4.0 for Pandemics: A Roadmap Using AI, Machine Learning, IoT and Cognitive Technologies*, 59-73. https://doi.org/10.1002/9781119769088.ch4

[74] Gupta, S., & Garg, P. (2021). An insight review on multimedia forensics technology. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 27. DOI: 10.1515/9783110677478

[75] Shrivastava, P., Agarwal, P., Sharma, K., & Garg, P. (2021). Data leakage detection in Wi-Fi networks. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 215. https://doi.org/10.1515/9783110677478-010

[76] Meenakshi, P. G., & Shrivastava, P. (2021). Machine learning for mobile malware analysis. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 151. https://doi.org/10.1515/9783110677478-008

[77] Garg, P., Pranav, S., & Prerna, A. (2021). Green Internet of Things (G-IoT): A Solution for Sustainable Technological Development. In *Green Internet of Things for Smart Cities* (pp. 23-46). CRC Press. https://doi.org/10.1201/9781003032397

[78] Nanwal, J., Garg, P., Sethi, P., & Dixit, A. (2021). Green IoT and Big Data: Succeeding towards Building Smart Cities. In *Green Internet of Things for Smart Cities* (pp. 83-98). CRC Press. https://doi.org/10.1201/9781003032397

[79] Gupta, M., Garg, P., & Agarwal, P. (2021). Ant Colony Optimization Technique in Soft Computational Data Research for NP-Hard Problems. In *Artificial Intelligence for a Sustainable Industry 4.0* (pp. 197-211). Springer, Cham. https://doi.org/10.1007/978-3-030-77070-9_12

[80] Magoo, C., & Garg, P. (2021). Machine Learning Adversarial Attacks: A Survey Beyond. *Machine Learning Techniques and Analytics for Cloud Security*, 271-291. https://doi.org/10.1002/9781119764113.ch13

[81] Garg, P., Srivastava, A. K., Anas, A., Gupta, B., & Mishra, C. (2023). Pneumonia Detection Through X-Ray Images Using Convolution Neural Network. In *Advancements in Bio-Medical Image Processing and Authentication in Telemedicine* (pp. 201-218). IGI Global. DOI: 10.4018/978-1-6684-6957-6.ch011

[82] Gupta, S., & Garg, P. (2023). Code-based post-quantum cryptographic technique: digital signature. *Quantum-Safe Cryptography Algorithms and Approaches: Impacts of Quantum Computing on Cybersecurity*, 193. https://doi.org/10.1515/9783110798159-014

[83] Prakash, A., Avasthi, S., Kumari, P., & Rawat, M. (2023). Puneet Garg 18 Modern healthcare system: unveiling the possibility of quantum computing in medical and biomedical zones. Quantum-Safe Cryptography Algorithms and Approaches: Impacts of Quantum Computing on Cybersecurity, 249. https://doi.org/10.1515/9783110798159