JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR HEALTHCARE DATA ENGINEERING: A COMPREHENSIVE SURVEY OF ARCHITECTURES, CHALLENGES, AND FUTURE DIRECTIONS

Shireesha Mangannagari
Lead Data Engineer – Government & Healthcare Operations
Choctaw Nation, Durant, *OK, 74701, United States*

Abstract: *Artificial intelligence (AI) and machine learning (ML) models are increasingly embedded in healthcare data pipelines to support diagnosis, risk stratification, and treatment planning. However, the predictive strength of complex models such as deep neural networks and ensemble learners is frequently accompanied by a loss of interpretability, which creates barriers to clinical trust, regulatory approval, and patient safety. Explainable Artificial Intelligence (XAI) has emerged as a discipline that seeks to reconcile predictive performance with human-understandable justification. This paper presents a survey of XAI in the context of healthcare data engineering. We review the major architectural families of explainability methods, including feature-attribution techniques, saliency and attention-based approaches, counterfactual reasoning, and intrinsically interpretable models, and examine how each can be embedded within the stages of a healthcare data engineering pipeline, from ingestion and terminology standardization through feature engineering, model serving, and audit logging. We further discuss the technical, evaluation, human-factors, and regulatory challenges that limit the practical adoption of XAI in clinical settings, and outline future research directions, including causal explainability, federated explanation generation, and large language model-assisted explanation delivery.*

Index Terms – *Explainable Artificial Intelligence, XAI, Healthcare Data Engineering, Interpretability, Clinical Decision Support, SHAP, LIME, Electronic Health Records, Trustworthy AI*

I. Introduction

Modern healthcare systems generate data at a volume and variety that few other industries match. Electronic health records (EHR), medical imaging archives, laboratory information systems, insurance claims, genomic sequencing platforms, and an expanding ecosystem of wearable and remote-monitoring devices together produce a heterogeneous, high-dimensional, and often longitudinal data landscape. Artificial intelligence and machine learning models trained on this data are now used to support diagnosis, risk stratification, treatment recommendation, and operational decision-making across the care continuum [1], [2].

The predictive accuracy gains achieved by deep neural networks, gradient-boosted ensembles, and other complex model families are frequently accompanied by a corresponding loss of transparency. In most other application domains this trade-off is an inconvenience; in healthcare it becomes a safety and accountability concern. A clinician who cannot understand why a model recommended a particular course of action is poorly positioned to catch an error, and a health system that cannot demonstrate how a model reached a decision is poorly positioned to defend that decision to a regulator, an auditor, or a court. The consequences of unexamined model behavior are not hypothetical: widely cited work has shown that a commercially deployed population-health algorithm systematically under-referred Black patients for additional care because it used healthcare cost, rather than clinical need, as a proxy label for illness burden [12].

Explainable Artificial Intelligence (XAI) is the body of methods and design practices developed to address this gap by producing human-understandable justifications for model behavior, either by constructing models that are transparent by design or by generating post-hoc explanations for models that are not. The literature draws a useful distinction between

interpretability, the degree to which a human can understand the cause of a model's decision without additional tooling, and explainability, the broader set of techniques – including post-hoc explanation methods applied to otherwise opaque models – used to make model behavior comprehensible [3], [4].

This paper surveys XAI specifically through the lens of healthcare data engineering: the pipelines, standards, and infrastructure that move raw clinical data into a form usable by predictive models. Rather than treating explainability as a technique bolted onto a finished model, we argue that it is best treated as a property that must be engineered into the data pipeline itself – from how data is ingested and standardized, through feature construction, to how predictions and their justifications are served, logged, and audited. The remainder of the paper is organized as follows. Section II provides background on the motivation for XAI in clinical settings. Section III reviews the stages of a typical healthcare data engineering pipeline. Section IV presents a taxonomy of XAI architectures. Section V proposes a reference architecture for integrating explainability into the data engineering lifecycle. Section VI surveys representative applications. Section VII discusses open challenges, and Section VIII outlines future research directions.

II. Background and Motivation

A. The Growing Footprint of AI in Clinical Practice

Machine learning models are now deployed across nearly every stage of clinical care, from triage and early-warning scores computed on vital signs and laboratory trends, to convolutional networks that flag suspicious regions on radiographs and pathology slides, to natural language processing pipelines that mine unstructured clinical notes for adverse events and coding support [1], [2]. This proliferation has been enabled by the digitization of health records and by advances in deep learning architectures capable of handling the scale and heterogeneity of clinical data. It has also raised the stakes for interpretability, since a model error in a high-volume screening or triage context can propagate across an entire patient population rather than affecting a single decision.

B. Why Explainability Matters in Healthcare

Three overlapping motivations drive the demand for explainability in clinical AI. First, clinician trust and adoption: physicians are understandably reluctant to act on a recommendation they cannot evaluate against their own clinical reasoning, and studies of clinician expectations consistently find that usable explanations are a precondition for integrating model output into workflow rather than a nice-to-have addition [8]. Second, regulatory and medico-legal accountability: health systems must be able to reconstruct the basis for a model-influenced decision when it is challenged, whether by an internal quality review, a regulator, or a malpractice proceeding. Third, error and bias auditing: explanation methods that surface which features drove a prediction give data science teams a mechanism to detect when a model has learned a spurious or inequitable pattern, such as relying on a proxy variable correlated with race or socioeconomic status rather than genuine clinical need [12].

C. Interpretability versus Explainability

It is useful to separate models that are intrinsically interpretable – such as linear and logistic regression, generalized additive models, decision trees, and compact rule lists, whose internal logic can be inspected directly – from models that are explained post-hoc through a secondary technique applied after training. Rudin has argued that for high-stakes decisions, practitioners should prefer intrinsically interpretable models wherever they can match the accuracy of a black-box alternative, since post-hoc explanations are themselves approximations and can misrepresent the true reasoning of the underlying model [4]. In practice, healthcare data engineering teams operate along a spectrum: some use cases (e.g., a mortality risk score built on a modest number of structured features) are well served by interpretable models, while others (e.g., image-based diagnosis) currently require deep architectures for acceptable accuracy and therefore depend on post-hoc explanation.

III. Healthcare Data Engineering Pipeline: An Overview

A. Data Sources and Standards

Healthcare data engineering pipelines typically draw on several categories of source systems: structured EHR data exchanged via HL7 v2 messages or the newer FHIR (Fast Healthcare Interoperability Resources) standard; medical imaging stored and transmitted through DICOM and PACS systems; laboratory results coded against LOINC; diagnoses and procedures coded against ICD-10 and SNOMED CT; insurance claims data; genomic data in formats such as VCF; and increasingly, continuous streams from wearable and remote-monitoring devices. Each source system carries its own vocabulary, granularity, and update cadence, which is the first source of the heterogeneity that downstream models, and their explanations, must contend with.

B. Pipeline Stages

A representative pipeline moves through ingestion, in which raw records are extracted from source systems; standardization, in which local codes and free-text values are mapped onto common terminologies; cleaning and imputation, in which missing values, duplicate encounters, and outliers are addressed; feature engineering, in which raw events are aggregated into model-ready features such as rolling statistics over vital signs or embeddings of diagnosis codes; storage in a feature store or curated analytic dataset; model training; model and prediction serving; and ongoing monitoring for data and model drift. Explainability considerations touch every one of these stages, not only the final modeling step, because an explanation is only as trustworthy as the data and feature semantics that feed it.

C. Data Quality Challenges Relevant to Explainability

Clinical data poses data quality challenges that are somewhat distinct from other domains and that directly affect the reliability of downstream explanations. Missingness in EHR data is frequently informative rather than random – a laboratory test that was never ordered may itself signal something about a patient's clinical trajectory – so imputation choices can silently alter what an explanation attributes importance to. Data is heterogeneous across source systems and care settings, with the same concept sometimes represented differently across departments or institutions. Clinical encounters are temporally irregular, complicating attribution methods that assume fixed-length or evenly spaced inputs. Finally, labels derived from billing codes or retrospective chart review can be noisy, and a model that has learned to fit label noise will produce explanations that faithfully, but misleadingly, attribute importance to whatever features correlate with that noise.

Table 1: Healthcare Data Engineering Pipeline Stages and Associated XAI Touchpoints

Pipeline Stage	Primary Engineering Task	XAI Touchpoint
Ingestion	Extracting data from EHR/FHIR, DICOM, claims, and device feeds	Provenance logging to support later attribution traceability
Standardization	Mapping local codes to ICD-10, SNOMED CT, LOINC	Ensures attributions reference clinically meaningful, standardized concepts
Cleaning & Imputation	Handling missing values, outliers, duplicate encounters	Flagging imputed values so explanations can separate observed from inferred evidence
Feature Engineering	Temporal aggregation, embeddings, derived clinical scores	Human-readable feature naming and versioning enables interpretable attribution labels
Model Training	Fitting ML/DL models on curated features	Choice between intrinsically interpretable and post-hoc-explained architectures
Model & Explanation Serving	Real-time or batch inference	Generating and returning per-prediction explanations alongside outputs
Monitoring & Audit	Tracking data and model drift, downstream outcomes	Logging explanation drift as an early warning signal and regulatory audit evidence

IV. Taxonomy of XAI Architectures

A. Dimensions of Explainability

XAI methods can be organized along three largely independent dimensions [3]. The first is scope of intrinsic transparency versus post-hoc explanation, discussed above. The second is local versus global explanation: a local explanation accounts for a single prediction on a single patient, whereas a global explanation characterizes the overall behavior of a model across a population. The third is model-agnostic versus model-specific: model-agnostic methods treat the underlying model as a black box and can be applied to any predictor, whereas model-specific methods exploit the internal structure of a particular architecture, such as the gradients of a neural network.

B. Feature Attribution Methods

Feature attribution methods assign a numeric contribution to each input feature for a given prediction. SHAP (SHapley Additive exPlanations) grounds these contributions in Shapley values from cooperative game theory, distributing a prediction's deviation from a baseline fairly across the contributing features, and provides both local, per-patient attributions and aggregated global summaries [5]. LIME (Local Interpretable Model-agnostic Explanations) instead fits a simple, interpretable surrogate model, typically a sparse linear model, in the local neighborhood of a single prediction to approximate the black-box model's behavior near that point [6]. Integrated Gradients and related gradient-based attribution methods are model-specific variants suited to differentiable architectures such as deep neural networks.

C. Saliency and Attention-Based Methods

For image-based clinical models, Grad-CAM and related saliency techniques use the gradients flowing into a convolutional layer to produce a coarse heatmap over the input image, highlighting the regions that most influenced a

classification decision [7]. This is particularly relevant in radiology and digital pathology, where a heatmap overlaid on the original image gives a reviewing clinician a direct visual check on whether the model attended to a clinically plausible region. In clinical natural language processing, the attention weights of transformer-based architectures are sometimes presented as an explanation of which tokens in a note influenced a prediction, although the literature cautions that attention weights are not always a faithful indicator of the true basis for a model's decision and should be validated rather than assumed to be explanatory.

D. Counterfactual and Example-Based Explanations

Counterfactual explanations describe the minimal change to a patient's input features that would have altered a model's prediction – for example, the reduction in a particular lab value that would move a patient below a risk threshold. This framing is attractive in healthcare because it can translate directly into an actionable recommendation rather than a static importance score. Example-based or prototype methods take a complementary approach, explaining a prediction by presenting the most similar historical cases on which the model was confident, which mirrors the case-based reasoning that clinicians already use.

E. Intrinsically Interpretable Models

Where achievable without a material loss of accuracy, intrinsically interpretable model families – generalized additive models, scoring systems, and compact rule lists – avoid the approximation risk inherent in post-hoc explanation altogether, since the model's decision logic is the explanation [4]. These architectures remain widely used for structured, moderate-dimensional clinical risk scores, even as deep learning dominates image- and text-based tasks.

Table 2: Comparison of Major XAI Method Families

Method Family	Scope	Model Dependency	Typical Healthcare Use Case	Relative Cost
SHAP	Local & Global	Model-agnostic (optimized variants for trees/deep nets)	Attribution for tabular risk models (sepsis, readmission)	Moderate–High
LIME	Local	Model-agnostic	Explaining individual ensemble predictions on structured EHR data	Low–Moderate
Grad-CAM / saliency maps	Local	Model-specific (CNNs)	Localizing regions of interest in radiology/pathology images	Moderate
Attention visualization	Local	Model-specific (transformers/RNNs)	Token-level explanation for clinical NLP tasks	Low
Counterfactual explanations	Local	Model-agnostic	Actionable, patient-specific risk-reduction guidance	Moderate–High
Intrinsically interpretable models	Local & Global	Model-specific (by design)	High-stakes structured scoring where transparency is prioritized	Low

V. Integrating XAI into the Healthcare Data Engineering Lifecycle

Explainability is most effective when it is treated as a first-class artifact of the data engineering pipeline rather than an afterthought computed once a model is already in production. Figure 1 sketches a reference architecture in which the explanation-generation module sits alongside model training and inference as a peer component rather than a downstream add-on. Ingested data from EHR/FHIR feeds, imaging archives, and other sources passes through standardization and curation into a shared feature store, which supplies both the model-training process and the explanation-generation module with a consistent, versioned view of features. This shared lineage matters because an explanation that references a feature definition different from the one the model actually saw at inference time is not merely unhelpful but actively misleading.

At serving time, the model and explanation-serving layer returns a prediction together with its corresponding local explanation through a single API contract, so that a clinician-facing user interface can present both the recommendation and its justification in one view rather than requiring a separate lookup. Every prediction-explanation pair is written to an audit trail and monitoring log, which supports two functions: regulatory traceability, since a health system can reconstruct why a specific decision was made months after the fact, and drift detection, since a sustained shift in the distribution of feature attributions can signal that the underlying data or population has changed even before overall model accuracy visibly degrades. This audit log closes a feedback loop back into the feature store and training process, informing when a model should be recalibrated or retrained.

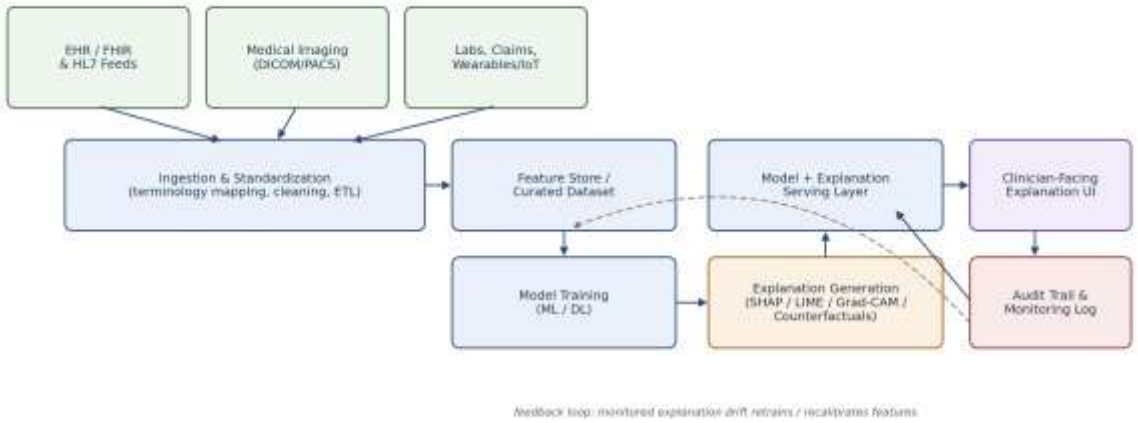

Figure 1: Reference architecture for integrating explanation generation into the healthcare data engineering lifecycle

VI. Applications of XAI in Healthcare

A. Clinical Risk Prediction

Structured risk-prediction models for sepsis onset, 30-day readmission, and inpatient mortality are among the most mature applications of clinical XAI. SHAP-based attribution is commonly used to surface the labs, vitals, and comorbidities that most strongly influenced a given risk score, giving a bedside clinician a concrete rationale rather than an opaque probability.

B. Medical Imaging

Convolutional neural networks used for radiology and pathology classification are typically explained through Grad-CAM or similar saliency techniques, which overlay a heatmap on the source image to indicate the region the model relied on. This provides a direct visual sanity check: a heatmap centered on an anatomically implausible region is a strong signal that the model may be relying on an imaging artifact rather than genuine pathology.

C. Clinical Natural Language Processing

NLP pipelines applied to unstructured clinical notes, for tasks such as summarization or adverse-event detection, increasingly present attention-weight visualizations over the source text as a form of explanation, allowing a reviewer to see which phrases the model weighted most heavily, while being mindful of the caveats around attention faithfulness noted in Section IV.

D. Genomics and Precision Medicine

In genomics, feature-attribution methods are applied to variant pathogenicity and drug-response prediction models to surface which genomic annotations drove a classification, supporting the kind of case-by-case scrutiny that precision-medicine decisions require.

VII. Challenges

A. Technical Challenges

Healthcare data is high-dimensional, multimodal, and temporally irregular, which complicates the direct application of attribution methods developed on simpler, static tabular benchmarks. Sampling-based methods such as SHAP can be computationally expensive when applied to deep architectures at the scale required for real-time clinical serving. A further, more fundamental concern is the trade-off between faithfulness and plausibility: an explanation can look convincing to a human reviewer while not accurately reflecting the model's actual decision process, and several authors have argued that this gap is under-appreciated in clinical deployments [10].

B. Evaluation of Explanation Quality

There is not yet a standardized, widely agreed set of metrics for evaluating whether an explanation is good, in the sense of being both faithful to the underlying model and useful to the intended clinical audience. Existing surveys catalogue a range of proposed metrics for fidelity, stability, and comprehensibility, but consistent adoption across the clinical XAI literature remains limited [14].

C. Human Factors and Clinical Workflow

An explanation that is technically accurate but poorly matched to a clinician's workflow and mental model can contribute to alert fatigue rather than improving decision-making. Work directly surveying clinician expectations emphasizes that explanations need to be contextualized to the specific decision at hand and delivered with minimal added cognitive burden,

which argues for user-centered design processes rather than treating explanation-interface design as an implementation detail [8], [11].

D. Regulatory and Ethical Considerations

Regulatory frameworks are increasingly explicit about explainability expectations for clinical AI. In the United States, the FDA's action plan for AI/ML-based Software as a Medical Device outlines expectations around transparency and monitoring across the product lifecycle [15]. In the European Union, AI systems used in healthcare are generally classified as high-risk under the EU AI Act, with corresponding transparency obligations, while the GDPR's provisions on automated decision-making have been read by many commentators as implying a right to explanation for affected individuals. Alongside these formal requirements, fairness auditing remains an essential use of explanation methods: the widely discussed finding that a population-health algorithm used healthcare cost as a proxy for clinical need, and thereby under-identified Black patients for additional care, was itself uncovered through exactly this kind of attribution analysis, underscoring both the risk of unexamined models and the value of explainability as a bias-detection tool [12]. A final, less-discussed concern is that explanation outputs can themselves leak sensitive information about a patient's features, which means explanation-serving infrastructure needs to be brought within the same privacy and access-control boundary as the underlying clinical data.

VIII. Future Directions

Several directions appear promising for advancing XAI in healthcare data engineering. Causal explainability seeks to move beyond correlational feature attribution toward estimates of causal effect, which would allow an explanation to support a genuinely actionable clinical recommendation rather than a description of statistical association. Federated explanation generation would allow multiple institutions to contribute to shared model development and validation of explanation methods without centralizing protected health information, extending the existing federated-learning paradigm to the explanation layer. Large language models are increasingly being explored as a means of translating quantitative attribution output into natural-language narratives that are easier for a busy clinician to consume, though this direction requires careful guardrails to ensure the generated narrative does not embellish or misrepresent the underlying quantitative explanation. Finally, the field would benefit from standardized, clinically grounded benchmarks and evaluation protocols for explanation quality, together with participatory design processes that involve clinicians directly in shaping how explanations are surfaced within their workflow.

IX. Conclusion

As machine learning becomes further embedded in clinical decision-making, explainability can no longer be treated as an optional add-on layered onto a finished model. This survey has argued that XAI is best approached as a property engineered throughout the healthcare data pipeline, from data ingestion and terminology standardization through feature construction, model training, prediction serving, and audit logging. We reviewed the major families of explanation methods and their respective trade-offs, proposed a reference architecture for embedding explanation generation into the data engineering lifecycle, and surveyed representative clinical applications. Significant technical, evaluative, human-factors, and regulatory challenges remain, and addressing them will require sustained collaboration between data engineers, machine learning researchers, clinicians, and regulators. Progress on causal explainability, federated explanation generation, and clinician-centered explanation interfaces represents a promising path toward AI systems in healthcare that are not only accurate, but trustworthy and accountable.

References

[1] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347–1358.

[2] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018). Opportunities and obstacles for deep learning in biomedicine. Journal of The Royal Society Interface, 15(141), 20170387.

[3] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608.

[4] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.

[5] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS).

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[7] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. IEEE International Conference on Computer Vision (ICCV).

[8] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Proceedings of Machine Learning Research, 106, 359–380.

[9] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310.

[10] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.

[11] Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey. Journal of Biomedical Informatics, 113, 103655.

[12] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

[13] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4).

[14] Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics, 10(5), 593.

[15] U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.

