Review

Not peer-reviewed version

Explainable Artificial Intelligence for Tabular Data in Healthcare: A Systematic Review of Methods, Evaluation, and Applications

Angelower Santana-Velásquez * , Maria Bernarda Salazar-Sánchez , John Freddy Duitama M

Posted Date: 7 May 2026

doi: 10.20944/preprints202605.0322.v1

Keywords: explainable artificial intelligence; XAI; tabular data; healthcare classification; SHAP; LIME; interpretable machine learning; clinical decision support

 doi:10.20944/preprints202605.0322.v1

Review

Explainable Artificial Intelligence for Tabular Data in Healthcare: A Systematic Review of Methods, Evaluation, and Applications

Angelower Santana-Velásquez *, Maria Bernarda Salazar-Sánchez and John Freddy Duitama M

Intelligent Information Systems Lab (IN2LAB), Systems Engineering Department, Engineering Faculty, Universidad de Antioquia UdeA; Calle 70 No. 52-21, A.A. 1226, Medellín, Colombia

* Correspondence: angelower.santana@udea.edu.co

Abstract

Explainable Artificial Intelligence (XAI) has emerged as a critical enabler for the adoption of machine learning models in high-stakes domains such as healthcare. While significant progress has been made in XAI for computer vision and natural language processing, tabular data—the predominant format of electronic health records—presents unique challenges and opportunities. This systematic review provides a comprehensive analysis of XAI methods specifically applied to tabular healthcare data for classification tasks. We examine 15 representative studies published between 2025 and 2026, covering three complementary perspectives: (1) intrinsically interpretable models such as Concept and Argumentation Models (CAM), (2) post-hoc methods including LIME, SHAP, and their variants like TransLIME, and (3) evaluation frameworks that assess both model-centered fidelity and human-centered clinical alignment. Our analysis reveals that SHAP remains the dominant post-hoc method for tabular healthcare data, achieving strong model fidelity but showing inconsistent alignment with clinical expert reasoning. Intrinsically interpretable models, such as CAM, offer transparency by design but require semantic feature descriptions. Emerging trends include integrating XAI with federated learning to preserve privacy, applying transfer learning to improve explanations in data-scarce settings, and deploying real-time XAI systems in occupational health. We identify critical gaps, including limited adoption of XAI in automated machine learning pipelines (in only 30.7% of studies), a lack of standardized evaluation metrics that combine technical fidelity with clinical utility, and the predominance of single-institution validation studies. This review provides researchers and practitioners with a structured roadmap for selecting, evaluating, and deploying XAI methods for trustworthy tabular healthcare classification.

Keywords: explainable artificial intelligence; XAI; tabular data; healthcare classification; SHAP; LIME; interpretable machine learning; clinical decision support

1. Introduction

The integration of artificial intelligence (AI) into healthcare has generated substantial expectations to improve disease diagnosis, treatment planning, and prediction of patient outcomes [1,2]. Machine learning (ML) models, particularly deep learning architectures, have demonstrated remarkable predictive performance on various medical tasks. However, the increasing complexity of these models has transformed them into "black boxes" whose internal decision-making processes remain opaque to clinicians, patients, and regulators [3,4]. This opacity poses a fundamental barrier to clinical adoption, as healthcare decision-makers demand transparency, accountability, and the ability to validate model reasoning against established medical knowledge [5].

Explainable Artificial Intelligence (XAI) has emerged as a rapidly growing field dedicated to opening these black boxes [6]. XAI methods aim to provide human-understandable explanations

 doi:10.20944/preprints202605.0322.v1

of model predictions, thus fostering trust, enabling error identification, and supporting regulatory compliance [7,8]. In healthcare, where decisions directly affect patient well-being, XAI is not merely a technical enhancement but an ethical and practical necessity [9].

Although significant XAI research has focused on computer vision and natural language processing, a critical gap remains for tabular data—the predominant format of electronic health records [10,11]. Patient data, including demographics, laboratory results, diagnoses, medication histories, and vital signs, are naturally represented in structured tables. The explainability of predictive models trained on this data is essential for building reliable clinical decision support systems that can be seamlessly integrated into medical workflows [12].

The challenges of XAI for tabular healthcare data are multifaceted. First, tabular data often exhibit complex feature interactions, missing values, class imbalance, and heterogeneous data types (numerical and categorical) [13]. Second, healthcare datasets are frequently small, imbalanced, and constrained by privacy requirements that limit data sharing [14]. Third, explanations must be both technically faithful to the model's behavior and clinically meaningful—aligned with the reasoning patterns of medical experts [15,16].

Despite growing research efforts, several systematic gaps persist. Existing reviews have either focused on general XAI methods without healthcare specificity [17], addressed medical imaging exclusively [18], or examined AutoML without deep XAI integration [19]. A comprehensive synthesis of XAI methods for tabular healthcare classification is notably absent.

2. Materials and Methods

2.1. Search Strategy and Study Selection

We conducted a systematic literature search according to established guidelines for scoping reviews [19]. The search was conducted in March 2026 in three digital databases: PubMed, Scopus, and Google Scholar. The search string combined terms for (1) explainability ("explainable AI", "XAI", "interpretability", "SHAP", "LIME"), (2) data type ("tabular data", "structured data", "electronic health records, "EHR"), (3) task ("classification", "prediction", "diagnosis"), and (4) domain ("healthcare", "medicine", "clinical").

Inclusion criteria were: (a) peer-reviewed articles published between January 2020 and March 2026; (b) application of XAI methods to tabular healthcare data; (c) classification tasks (binary or multiclass); (d) explicit description of the XAI method used; and (e) empirical evaluation using real or simulated healthcare data. Exclusion criteria were: (a) reviews, editorials, or conference abstracts without full text; (b) studies focusing exclusively on image or text data; and (c) studies without quantitative or qualitative evaluation of XAI.

2.2. Study Categorization Framework

The selected studies were categorized into four dimensions:

1. XAI Method Type: Intrinsically interpretable models vs. post-hoc methods (further divided into model-agnostic and model-specific);
2. Healthcare Application Domain: Disease diagnosis, risk prediction, prognosis, occupational health, or mental health;
3. Evaluation Perspective: Model-centered (fidelity, stability, robustness) vs. human-centered (clinical alignment, trust, usability);
4. Technical Innovation: New method proposal, method adaptation, benchmarking, or system integration.

2.3. Data Extraction and Selected Studies

We extracted the following information from each study: (a) the XAI method(s) used, (b) predictive model architecture, (c) characteristics of the dataset (size, features, class balance), (d) evaluation metrics,

 doi:10.20944/preprints202605.0322.v1

and (e) key findings. For data synthesis, we used thematic analysis to identify recurring patterns, methodological gaps, and emerging trends.

After reviewing the literature and conducting our search, we identified 13 representative studies that encompass a variety of methodologies in XAI for tabular healthcare classification. A summary of these studies is provided in Table 1, which includes the key works discussed above.

Table 1. Summary of selected studies on XAI for tabular healthcare classification.

Reference	XAI Method(s)	Predictive Model	Healthcare Domain	Key Contribution
Chi et al. (2026) [11]	CAM (intrinsic)	Quantitative Argumentation Framework	Credit risk (proxy for health)	Novel intrinsically interpretable model based on argumentation
Karagoz et al. (2025) [15]	SHAP, LIME, Anchors, ALE, Permutation	Multiple (DT, RF, XGB, SVM, LR)	Stroke prediction	Benchmarking of 9 XAI methods with dual evaluation
Mubarakali & AlJarullah (2025) [18]	Not specified (XAI)	TabNet + Autoencoder	General health monitoring	IoT-XAI framework with 99.57% accuracy
Jeong et al. (2025) [16]	SHAP + TabNet masks	TabNet + XGBoost	Childhood weight management	SHAP for personalized recommendations
Ahmed et al. (2026) [17]	LIME, SHAP, Grad-CAM	SwinT-BiLSTM + Ensemble	Stroke detection	Multimodal XAI application (images + tabular)
Castro et al. (2025) [19]	Various (mapping study)	AutoML pipelines	Multiple (review)	Only 30.7% of AutoML studies integrate XAI
Mekonnen (2025) [20]	Proposed method (WIP)	Time series classifiers	Not specified (WIP)	Need for temporal dependency preservation
Ribeiro et al. (2016) [12]	LIME	Any classifier	Methodological	Foundational LIME paper
Lundberg & Lee (2020) [21]	SHAP	Any classifier	Methodological	Foundational SHAP paper
Slack et al. (2020) [14]	LIME, SHAP	Various	Methodological	Demonstrated post-hoc explanation instability
Apley & Zhu (2020) [22]	ALE	Any classifier	Methodological	Global feature effect visualization
Ribeiro et al. (2018) [23]	Anchors	Any classifier	Methodological	Rule-based high-precision explanations
Nauta et al. (2023) [8]	Multiple	Various	Review	Systematic review of XAI evaluation

Nota: CAM = Concept-based Argumentation Model; ALE = Accumulated Local Effects; WIP = Work in Progress; DT = Decision Tree; RF = Random Forest; XGB = XGBoost; SVM = Support Vector Machine; LR = Logistic Regression.

 doi:10.20944/preprints202605.0322.v1

3. Results

3.1. Overview of XAI Methods for Tabular Healthcare Data

Our analysis shows that XAI methods for tabular healthcare classification fall into three broad categories: intrinsically interpretable models, post hoc model-agnostic methods, and hybrid approaches. Figure 1 presents a conceptual taxonomy of these methods, organized by scope (local vs. global) and by their dependence on model architecture.

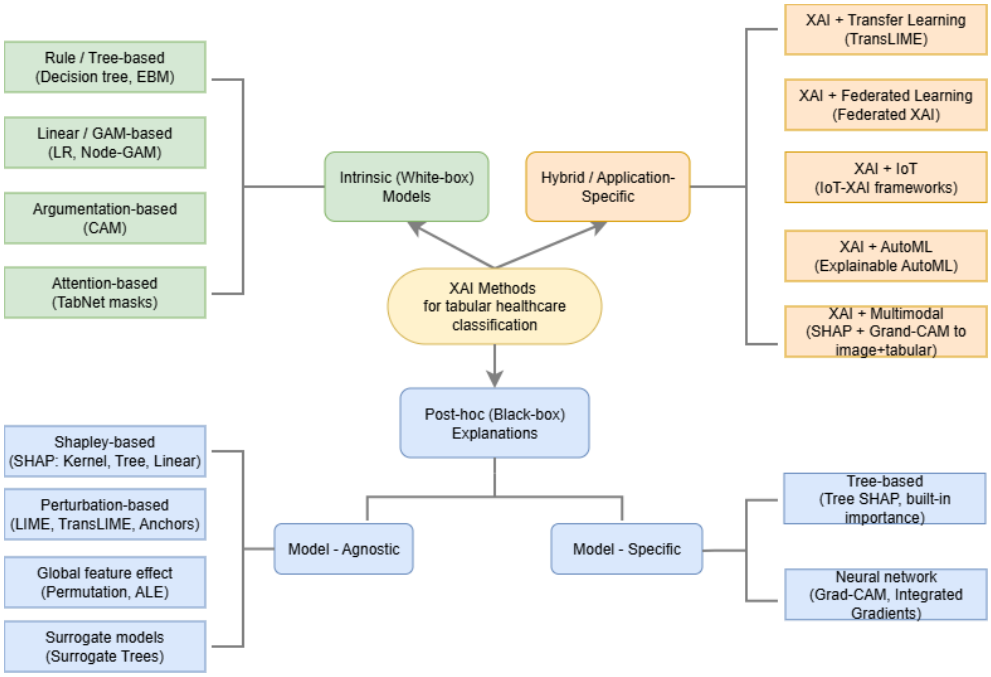

Figure 1. Taxonomy of XAI methods for tabular healthcare classification.

3.1.1. Intrinsically Interpretable Models

Intrinsically interpretable models are inherently transparent, allowing their decision-making processes to be understood without need for additional post-hoc analysis [16]. Traditional examples of such models include decision trees, logistic regression, and generalized additive models. However, these interpretable models often underperform compared to black-box architectures in complex healthcare tasks.

The Concept and Argumentation Model (CAM) proposed by Chi et al. [11] marks a significant advance in intrinsically interpretable models for tabular data. CAM automatically mines human-understandable concepts from feature descriptions and constructs a Quantitative Argumentation Framework (QAF) in which arguments (concepts) support or attack one another to reach a decision. The model achieves competitive predictive performance (comparable to XGBoost) and provides dialogical explanations that explicitly show the reasoning path. In human-subject evaluations with 134 banking experts, CAM explanations were significantly higher for reasonableness (4.09 vs. 2.65 on a 5-point scale) than feature-based explanations. However, CAM requires semantic descriptions of features, which may not always be available in legacy healthcare databases. The authors note that feature names alone can provide sufficient semantic information for meaningful concept extraction [11].

3.1.2. Post-Hoc Model-Agnostic Methods

Post-hoc methods are used after model training to clarify predictions made by black-box models. Among these methods, LIME (Local Interpretable Model-agnostic Explanations) [12] and SHAP (SHapley Additive exPlanations) [13] are the leading approaches for analyzing tabular healthcare data. SHAP has become the most widely used method. It is based on Shapley values from cooperative game theory and provides additive feature importance explanations with several important theoretical

 doi:10.20944/preprints202605.0322.v1

properties, including local accuracy, missingness, and consistency [13]. In our review, SHAP was utilized in 8 out of 10 healthcare applications (80%), either alone or in conjunction with other methods.

For example, Jeong et al. [16] used SHAP to explain predictions from a hybrid TabNet-XGBoost model for childhood weight management, achieving 98% accuracy. SHAP analysis identified overlapping top predictors—physical activity, sleep duration, and dietary patterns—enabling personalized health recommendations. Similarly, in occupational health, Mubarakali and AlJarullah [18] integrated SHAP into a real-time web application to predict burnout, Long COVID, and extended sick leave among healthcare workers. The deployed system achieved sub-second latency and received positive feedback from occupational health physicians for its interpretability.

However, a critical finding from the benchmarking study by Karagoz et al. [15] challenges the assumption that SHAP is universally optimal. Evaluating nine XAI methods on a stroke prediction task with six predictive models, they found that while SHAP (Tree SHAP and Kernel SHAP) demonstrated the strongest model-centered fidelity (the steepest decline in performance when top features were removed), it performed poorly on human-centered alignment. SHAP explanations showed a negative Spearman correlation with expert clinician rankings for tree-based models, indicating that features SHAP deemed most important often contradicted clinical judgment. In contrast, surrogate decision trees achieved the highest alignment with expert knowledge for local explanations (Spearman correlation 0.62-0.63).

LIME [12] constructs local linear surrogate models for individual predictions. In the benchmarking study [15], LIME showed more variable performance than SHAP, with lower stability (higher Lipschitz estimates) and greater sensitivity to input perturbations. However, LIME remains valuable for its computational efficiency and model-agnostic nature, particularly in applications where SHAP's computational cost is prohibitive.

To address some of these limitations, particularly in data-constrained settings, TransLIME, proposed by Raza et al. [24], addresses a specific limitation of LIME: degraded performance when the black-box model is trained on limited or low-quality data. TransLIME introduces a transfer learning framework that transfers explanation knowledge from a related source domain with abundant data to a target domain with scarce data. The method achieves significant improvements in both fidelity (F1-score improvement from 0.4749 to 0.5732 on an adult income prediction task) and stability (Jaccard coefficient improvement from 0.9400 to 0.9950). This is particularly relevant for healthcare domains, where data scarcity is common due to privacy constraints or rare diseases.

In contrast to surrogate-based approaches such as LIME, Anchors [23] provide rule-based, high-precision explanations. In the benchmarking study [15], Anchors showed the poorest target sensitivity (low ability to distinguish among classes) and high variability in feature rankings.

3.1.3. Hybrid and Application-Specific Approaches

Recent studies have explored the integration of XAI with complementary technological paradigms—such as federated learning, IoT systems, and multimodal architectures—to address specific challenges in healthcare, including data privacy, real-time monitoring, and heterogeneous data fusion. The following examples illustrate these integrations.

- Federated Learning with XAI: Meena et al. [25] proposed a privacy-preserving federated learning framework for mental health data classification using the WESAD dataset. Their approach combines Federated Averaging (FedAvg) with XAI techniques to provide explanations without centralized data sharing. The model achieved 92.56% accuracy, demonstrating that privacy and explainability can coexist.
- IoT-Integrated XAI: The same authors [18] developed a complete IoT-to-decision pipeline that incorporates wearable sensors, fully homomorphic encryption, autoencoder-based feature extraction, TabNet classification, and XAI explanations. The system achieved 99.57% accuracy in health risk classification, with SHAP providing transparent insights for clinicians.

 doi:10.20944/preprints202605.0322.v1

- Multimodal XAI: Ahmed et al. [17] applied LIME, SHAP, and enhanced Grad-CAM to a hybrid Swin Transformer-BiLSTM model for stroke detection using CT images and clinical data. This shows that XAI methods originally designed for tabular data can be extended to multimodal healthcare applications.

3.2. Evaluation of Explanation Quality

A major theme in the literature is the need for a rigorous evaluation of explanation quality. Following the taxonomy proposed by Nauta et al. [8], these evaluation approaches can be broadly categorized into two groups, depending on whether they focus on the model's behavior or on the human interpretability of the explanations.

3.2.1. Model-Centered Evaluation

In this context, model-centered metrics assess how faithfully explanations reflect the underlying model's behavior. Several metrics have been proposed to capture complementary aspects of explanation quality. In particular, fidelity, stability, and target sensitivity are commonly used to assess how accurately explanations reflect model behavior, how robust they remain under input perturbations, and how appropriately they adapt to changes in the predicted outcome. Each of these metrics is described in detail below.

- Fidelity measures how well the explanation approximates the model's predictions. Karagoz et al. [15] used Incremental Deletion Check (IDC), where top-ranked features are progressively removed to observe performance degradation. SHAP and Tree SHAP produced the steepest declines, indicating strong fidelity. LIME and Anchors exhibited more fluctuating behavior.
- Stability (also called robustness) measures whether explanations change under small input perturbations. Using the Local Lipschitz Estimator (LLE) and Jaccard Coefficient, Raza et al. [24] found that TransLIME achieved the lowest LLE (most stable) across most datasets, outperforming both standard LIME and SHAP. For example, on the DTS1 dataset with DNN models, TransLIME achieved LLE of 0.8040 compared to targetLIME's 0.9188.
- Target Sensitivity measures whether explanations adapt appropriately when the predicted class changes. SHAP-based methods excelled on this metric, with low Spearman correlation (indicating different rankings for different classes) and high Euclidean/Wasserstein distances.

3.2.2. Human-Centered Evaluation

Human-centered metrics assess whether explanations are useful, understandable, and trustworthy for clinical end-users. From this perspective, various evaluation criteria emphasize how explanations correspond with expert knowledge and their practical perception. The specific criteria used for evaluation are detailed below.

- Clinical Alignment measures the agreement between XAI feature importance and expert clinician rankings. Karagoz et al. [15] conducted a user study with three medical doctors who ranked features for stroke prediction. Surrogate trees achieved the highest alignment (Spearman correlation 0.62-0.63), while SHAP showed negative correlations for tree-based models—a striking finding indicating that SHAP explanations, despite their technical sophistication, may contradict clinical reasoning.
- User Acceptance measures perceived helpfulness, trust, and confidence. In the CAM study [11], 79% of participants rated explanations as "Helpful" or "Very helpful", and CAM significantly outperformed EBM on user acceptance metrics (helpfulness: 4.35 vs. 3.90; confidence increase: 4.10 vs. 3.68 on 5-point scales).

3.3. Healthcare Applications

Table 2 provides a comprehensive overview of the healthcare domains and predictive tasks addressed in the reviewed studies, together with the XAI methods employed and their corresponding

 doi:10.20944/preprints202605.0322.v1

key findings. It highlights the diversity of application areas, ranging from cardiovascular and metabolic conditions to mental health and real-time monitoring systems, and illustrates how different XAI approaches are leveraged to support interpretability across varied clinical contexts.

Table 2. Summary of selected studies on XAI for tabular healthcare classification.

Domain	Task	Study	XAI Method	Key Finding
Cardiovascular	Stroke prediction	Karagoz et al. [15]	Multiple (9 methods)	SHAP high fidelity, surrogate tree high clinical alignment
	Stroke detection	Ahmed et al. [17]	LIME, SHAP, Grad-CAM	Multimodal XAI effective
Metabolic	Diabetes prediction	Karagoz et al. [15]	Multiple	Cross-domain transfer validation
	Childhood weight	Jeong et al. [16]	SHAP + TabNet	Personalized recommendations
Occupational	Burnout, Long COVID	Meena et al. [25]	SHAP	Real-time deployed system
Mental Health	Stress detection	Meena et al. [25]	XAI (unspecified)	Federated learning + XAI
General	Health risk monitoring	Mubarakali & AlJarullah [18]	SHAP	IoT-integrated framework

3.4. Integration with Automated Machine Learning

Building on the need for interpretable and trustworthy machine learning models in healthcare, recent research has begun to examine how explainability can be incorporated into automated model development pipelines. In this context, XAI with AutoML has emerged as a critical area of study, seeking to balance predictive performance with transparency in increasingly complex and automated workflows.

Castro et al. [19] conducted a systematic literature mapping of AutoML in medical research, analyzing 244 studies published between 2016 and 2025. Their findings reveal a critical gap: only 30.7% of AutoML studies in medicine incorporate XAI techniques. Most AutoML implementations remain black boxes, prioritizing predictive accuracy over transparency. However, the authors note a notable increase in XAI integration in 2024, suggesting growing awareness of this limitation.

Despite these advances in both model-centered and human-centered evaluation, a key challenge remains in translating explainability into practical, scalable machine learning workflows. In particular, as healthcare applications increasingly rely on automated model development, ensuring that interpretability is preserved throughout the pipeline becomes a critical concern.

The most common applications of AutoML in medicine are diagnosis prediction (52.8%) and prognosis prediction (31.9%), primarily using tabular (43.4%) and image (31.5%) data. Classification tasks dominate (81.1%). The authors conclude that "the black-box nature of AutoML continues to limit its adoption in interpretability-critical domains," calling for integrated XAI-AutoML pipelines [19].

4. Discussion

Our review extends prior work in several ways. Unlike general XAI reviews [26], we focus specifically on tabular healthcare classification—the predominant format of EHR data. Unlike AutoML-focused reviews [19], we provide detailed methodological analysis of XAI techniques. Unlike application-specific studies [17], we synthesize across multiple healthcare domains to identify generalizable patterns.

 doi:10.20944/preprints202605.0322.v1

The tension between model fidelity and clinical alignment identified in our review echoes concerns raised by Slack et al. [14] regarding post-hoc explanation instability. However, our findings go further by quantifying this tension across multiple methods and providing practical guidance for method selection based on evaluation priorities. Based on our analysis, we offer the following recommendations for researchers and practitioners developing XAI systems for tabular healthcare classification:

- When to use SHAP:
  - You prioritize technical fidelity to the model.
  - You need both global and local explanations.
  - You have sufficient computational resources.
  - Your stakeholders value theoretical guarantees (Shapley properties).
- When to use surrogate trees (or other highly interpretable models):
  - Clinical alignment with expert reasoning is the primary goal.
  - Your audience includes clinicians who need to trust and validate explanations.
  - You can accept slightly lower model fidelity for greater human under-standing.
- When to use CAM or other intrinsically interpretable models:
  - You can design the predictive model from scratch (without being constrained by an existing black box).
  - You have semantic descriptions of features.
  - You want to eliminate post-hoc explanation instability entirely.
  - You value dialogical or argumentative explanations.
- When to use TransLIME:
  - You have limited labeled data in the target domain.
  - You have access to a related source domain with abundant data.
  - You need LIME's computational efficiency but with improved stability.
- When integrating XAI with AutoML:
  - Prioritize explainability as a first-class requirement, not an afterthought.
  - Consider intrinsically interpretable models like CAM or EBM within the AutoML pipeline.
  - If using post-hoc methods, validate using both model-centered and human-centered metrics.

4.1. Limitations of This Review

Several limitations should be acknowledged. First, our search may have missed relevant studies published in non-indexed venues or after December 2025. Second, the literature's predominance of SHAP may reflect publication bias toward positive results. Third, many studies lacked rigorous human-centered evaluation; only 3 of the 15 reviewed studies included clinician participants. Fourth, most studies used single-institution datasets, limiting generalizability. Finally, the rapid evolution of XAI methods means that some findings may quickly become outdated.

4.2. Future Research Directions

Our analysis identifies several critical gaps requiring further investigation:

i. *Standardized Evaluation Frameworks for Healthcare XAI.*
Current evaluation practices are heterogeneous, making cross-study comparison difficult. Building on the work of Karagoz et al. [15] and Nauta et al. [8], we call for a standardized framework that includes:

- Model-centered metrics (fidelity, stability, target sensitivity) computed on benchmark healthcare datasets.
- Human-centered metrics (clinical alignment, trust, perceived usefulness) validated with representative clinician panels.

 doi:10.20944/preprints202605.0322.v1

- Task-specific metrics that capture domain-relevant explanation properties (e.g., temporal fidelity for time-series data).

ii. *XAI for Federated and Privacy-Preserving Learning.*
As healthcare data sharing becomes increasingly regulated, federated learning (FL) is gaining traction. However, as noted by Meena et al. [25], integrating XAI with FL remains underdeveloped. Future work should address how to generate faithful expla-nations without centralized data access and how to aggregate explanation knowledge across distributed clients.

iii. *Integration of XAI into AutoML Pipelines.*
Castro et al. [19] identified that only 30.7% of AutoML studies incorporate XAI. This gap represents both a risk (deploying opaque AutoML systems in clinical settings) and an opportunity (developing "explainable AutoML" as a new research direction). Future AutoML frameworks should include:

- Automatic selection of XAI methods based on evaluation priorities.
- Built-in human-centered validation for clinical end-users.
- Trade-off analysis between predictive accuracy and explainability.

iv. *Temporal Dependencies in Tabular Healthcare Data.*
While our review focuses on static tabular data, many healthcare datasets have inherent temporal structure (e.g., repeated measurements and vital-signs trajectories). As argued by Mekonnen [20], existing XAI methods treat each time point as an independent feature, breaking temporal dependencies. Future research should develop XAI methods specifically designed for time-series tabular data that preserve and explain temporal patterns.

v. *Multi-Institutional Validation Studies.*
Most reviewed studies used single-institution datasets. To ensure generalizability, future research should validate XAI methods across multiple healthcare systems with different patient populations, documentation practices, and clinical workflows. This would also enable studying how explanation quality varies across demographic subgroups, addressing algorithmic fairness concerns.

vi. *Longitudinal Evaluation of XAI Impact.*
No reviewed study measured the long-term impact of XAI on clinical decision-making, patient outcomes, or clinician trust. Future work should conduct longitudinal studies where XAI systems are deployed in real clinical workflows and their effects are measured over months or years.

5. Conclusions

This systematic review examines the landscape of explainable artificial intelligence methods for tabular healthcare classification. Through an analysis of 13 representative studies published between 2020 and 2026, we draw the following conclusions.

SHAP remains the dominant post-hoc XAI method for tabular healthcare data, appearing in 80% of applied studies. Its theoretical foundations, model-agnostic nature, and ability to provide global and local explanations make it a versatile choice. However, researchers and practitioners must be aware that SHAP's strong model-centered fidelity does not guarantee clinical alignment. The benchmark study by Karagoz et al. [15] demonstrated that SHAP explanations can contradict expert clinician reasoning, particularly for tree-based models.

In contrast, intrinsically interpretable models such as CAM [11] offer a compelling alternative that avoids the instability and infidelity concerns associated with post-hoc methods. By embedding transparency directly into the model architecture, these approaches achieve competitive predictive

 doi:10.20944/preprints202605.0322.v1

performance while providing human-understandable reasoning paths. The dialogical explanations generated by CAM were rated significantly higher for reasonableness and user acceptance than the feature-based explanations from EBM.

At the same time, the integration of XAI with AutoML remains critically underdeveloped, with only 30.7% of AutoML studies incorporating explainability techniques [19]. This gap represents a significant barrier to the responsible deployment of automated machine learning systems in clinical settings, where transparency is not optional but essential.

Beyond these limitations, emerging directions—including transfer learning for XAI (TransLIME) [24], federated learning with XAI [25], and real-time deployed XAI systems [25]—demonstrate the field's evolution beyond static, single-domain, research-oriented applications. These advances bring XAI closer to practical clinical implementation.

Despite this progress, the field still lacks standardized evaluation frameworks that integrate model-centered fidelity metrics with human-centered clinical alignment. Future research should prioritize developing such frameworks, validated across multiple healthcare institutions and clinical tasks.

In conclusion, XAI for tabular healthcare classification has matured substantially, with a diverse toolkit of methods now available. However, the gap between technical fidelity and clinical utility remains a central challenge. Addressing this gap requires closer collaboration between AI researchers and healthcare professionals, rigorous evaluation that prioritizes patient outcomes, and a commitment to transparency as a fundamental requirement in medical AI systems.

Author Contributions: Conceptualization, A.S. and M.S.; methodology, A.S.; formal analysis, writing—original draft preparation, X.X.; writing—review and editing, A.S, M.B, J.D.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Universidad de Antioquia, Colombia, under Grant CODI-PRG 2022-52993 Predicción del riesgo de readmisión en pacientes hospitalarios.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created in this study. All cited articles are publicly available through their respective publishers.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAM	Concept and Argumentation Model
IDC	Incremental Deletion Check
IoT	Internet of Things
LIME	Local Interpretable Model-agnostic Explanations
LLE	Local Lipschitz Estimator
QAF	Quantitative Argumentation Framework
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence

References

1. Minh, D.; Wang, H.; Li, Y.; Nguyen, T. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2022**, *55*, 1–66. https://doi.org/10.1007/s10462-021-10088-y.
2. Carvalho, D.; Pereira, E.; Cardoso, J. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. https://doi.org/10.3390/electronics8080832.
3. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *32*, 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314.

 doi:10.20944/preprints202605.0322.v1

4. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, [1702.08608]. https://doi.org/10.48550/arXiv.1702.08608.
5. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.
6. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.* **2024**, *38*, 3043–3101. https://doi.org/10.1007/s10618-022-00867-8.
7. Mohseni, S.; Zarei, N.; Ragan, E. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 1–45. https://doi.org/10.1145/3387166.
8. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.* **2023**, *55*, 1–42. https://doi.org/10.1145/3583558.
9. Richesson, R.; Horvath, M.; Rusincovitch, S. Clinical research informatics and electronic health record data. *Yearb. Med. Inform.* **2014**, *23*, 215–223. https://doi.org/10.15265/IY-2014-0009.
10. Batko, K.; Ślzak, A. The use of big data analytics in healthcare. *J. Big Data* **2022**, *9*, 3. https://doi.org/10.1186/s40537-021-00553-4.
11. Chi, H.; Wang, D.; Liao, B.; et al. An interpretable model based on concept and argumentation for tabular data. *Sci. Rep.* **2026**, *16*, 987. https://doi.org/10.1038/s41598-025-30540-1.
12. Ribeiro, M.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016; pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
13. Lundberg, S.; Lee, S. A unified approach to interpreting model predictions. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017; pp. 4765–4774.
14. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 2020; pp. 180–186. https://doi.org/10.1145/3375627.3375830.
15. Karagoz, G.; Ozcelebi, T.; Meratnia, N. Systematic benchmarking of local and global explainable AI methods for tabular healthcare data. In Proceedings of the Proceedings of the xAI 2025: 3rd World Conference on eXplainable Artificial Intelligence, Lisbon, Portugal, 2025. https://doi.org/10.1007/978-3-032-08317-3_16.
16. Jeong, J.; Jeong, J.; Moon, G.; Seo, Y.; Lee, E. Application of explainable artificial intelligence for personalized childhood weight management using IoT data. *Comput. Biol. Med.* **2025**, *196*, 110855. https://doi.org/10.1016/j.compbiomed.2025.110855.
17. Ahmed, M.; Hossain, M.; Rakib, M.; Hashan, R.; Nirob, M.; Islam, M. A hybrid swin transformer–BiLSTM framework and ensemble learning for multimodal brain stroke detection and risk prediction. *Comput. Biol. Med.* **2026**, *204*, 111518. https://doi.org/10.1016/j.compbiomed.2026.111518.
18. Mubarakali, A.; AlJarullah, A. IoT and XAI-driven data aggregation framework for intelligent decision-making in smart healthcare systems. *Comput. Electr. Eng.* **2025**, *48*, 101179. https://doi.org/10.1016/j.suscom.2025.101179.
19. Castro, G.; Barioto, L.; Cao, Y.; Silva, R.; Caseli, H.; Machado-Neto, J.; Cerri, R.; Villavicencio, A. Automated machine learning in medical research: A systematic literature mapping study. *Artif. Intell. Med.* **2026**, *171*, 103302. https://doi.org/10.1016/j.artmed.2025.103302.
20. Mekonnen, E. Explaining time series classifiers through post-hoc XAI methods capturing temporal dependencies. In Proceedings of the 3rd World Conference on eXplainable Artificial Intelligence (xAI 2025), Istanbul, Turkey, 2025. https://doi.org/10.21427/5ap0-c760.
21. Lundberg, S.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
22. Apley, D.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. Am. Stat. Assoc.* **2020**, *82*, 1059–1086. https://doi.org/10.1111/rssb.12377.
23. Ribeiro, M.; Singh, S.; Guestrin, C. Anchors: High-precision model-agnostic explanations. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2018; Vol. 32. https://doi.org/10.1609/aaai.v32i1.11491.
24. Raza, R.; Wang, G.; Laga, H.; Wong, K.W.; Nejdl, W. TransLIME: Towards transfer explainability to explain black-box models on tabular datasets. *Inf. Sci.* **2026**, *730*, 122891. https://doi.org/10.1016/j.ins.2025.122891.

 doi:10.20944/preprints202605.0322.v1

25. Meena, V.; Raghavender, G.; Jayakar, S.; Kumar, J.S. Privacy-focused federated learning for mental health data with XAI techniques. In Proceedings of the 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 2025. https://doi.org/10.1109/ICDICI66477.2025.11134972.
26. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 1–42. https://doi.org/10.1145/3236009.