

Systematic Review

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

Angelower Santana-Velásquez  and Maria Bernarda Salazar-Sánchez * 

Intelligent Information Systems Lab (In2Lab), Department of Systems Engineering, Engineering Faculty, Universidad de Antioquia (UdeA), Calle 70 No. 52-21, A.A. 1226, Medellín 050010, Colombia; angelower.santana@udea.edu.co

* Correspondence: bernarda.salazar@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 21 primary studies published between 2020 and the first half of 2026, covering three complementary perspectives: (1) intrinsically interpretable models, (2) post-hoc methods including LIME, SHAP, and their variants, 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, achieving strong model fidelity but showing inconsistent alignment with clinical expert reasoning. Key findings include the significant impact of class imbalance on explanation consistency, the importance of clinician-centered evaluation, and the emergence of hybrid approaches integrating XAI with generative AI and transfer learning. We identify critical gaps, including limited adoption of XAI in AutoML pipelines, lack of standardized evaluation metrics, and predominance of single-institution validation studies.

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



Academic Editors: Paolo Bellavista and Sandi Baressi Šegota

Received: 30 May 2026

Revised: 12 July 2026

Accepted: 13 July 2026

Published: 6 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

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 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. Automated Machine Learning (AutoML) refers to the automation of the end-to-end machine learning pipeline, including data preprocessing, feature selection, model selection, and hyperparameter tuning, making ML more accessible to non-experts and accelerating model development [19]. A comprehensive synthesis of XAI methods for tabular healthcare classification is notably absent. To address this gap, we conducted a systematic review of 21 primary studies that apply XAI to tabular healthcare data for classification tasks.

2. Related Work: Existing Reviews on XAI for Healthcare

Several recent reviews have surveyed the field of Explainable AI, though none with the specific focus on tabular healthcare classification that we address here.

Nauta et al. [8] provided a comprehensive overview of evaluation metrics for XAI across all data types, establishing a foundational taxonomy of quantitative evaluation methods. However, their work did not focus on healthcare applications, nor did it provide domain-specific recommendations for clinical practice.

In the healthcare domain, Tjoa and Guan [3] offered a broad survey of XAI techniques for medical AI, emphasizing the need for transparency in clinical decision support. Similarly, Carvalho et al. [2] reviewed general interpretability methods but did not focus on the unique characteristics of tabular data, such as feature interactions and class imbalance.

More recently, Castro et al. [19] conducted a systematic mapping of Automated Machine Learning (AutoML) in medicine, critically revealing that only 30.7% of AutoML studies incorporate XAI techniques. This finding underscores a major gap that our review directly addresses by focusing on XAI specifically within non-automated, tabular data pipelines.

A key distinction of our work is the dual evaluation perspective. Unlike prior reviews that focus either on technical fidelity (e.g., [15]) or on clinical applicability (e.g., [3]), our review integrates both model-centered and human-centered metrics, following the benchmarking framework proposed by Karagoz et al. [15].

Furthermore, our review extends the scope of previous works by incorporating three additional dimensions that have been largely overlooked in existing surveys: (i) a systematic analysis of the impact of class imbalance on explanation consistency, drawing on recent studies that evaluate LIME and SHAP under skewed class distributions using real-world

clinical data [20,21]; (ii) a dedicated examination of clinician-centered evaluation, including user studies that assess the understandability and actionability of different XAI methods from the perspective of practicing clinicians [22]; and (iii) an exploration of emerging hybrid approaches that integrate XAI with generative AI models (e.g., LLMs) and AutoML-driven prompt engineering for enhanced clinical decision support [23].

To the best of our knowledge, no previous review has systematically compared XAI methods for tabular healthcare data using this combined lens, nor has one provided a quantitative comparison of fidelity versus clinical alignment while also addressing these emerging challenges and opportunities.

Table 1 provides a comparative summary of our review against three recent and representative surveys.

Table 1. Comparative analysis of the current review with existing surveys.

Feature/Aspect	Nauta et al. [8]	Tjoa & Guan [3]	Castro et al. [19]	Our Review
Focus on Healthcare	No	Yes	Yes	Yes
Focus on Tabular Data	No	No	Yes (AutoML)	Yes
Dual Evaluation (Fidelity + Clinical)	No	No	No	Yes
Quantitative Comparison of XAI Methods	Yes (General)	No	No	Yes
PRISMA-guided Systematic Review	No	No	Yes	Yes
Analysis of XAI-AutoML Gap	No	No	Yes	Yes
Analysis of Class Imbalance Impact	No	No	No	Yes
Clinician-Centered Evaluation	No	No	No	Yes
XAI with Generative AI (LLMs)	No	No	No	Yes

3. Materials and Methods

3.1. Search Strategy

We conducted a systematic literature search according to established guidelines for systematic review [19]. The search was conducted in three digital databases: PubMed, Scopus, and IEEE Xplore. The search was conducted between January 2020 and the first half of 2026 to ensure the inclusion of the latest advances in medical XAI.

To retrieve relevant publications, combinations of controlled vocabulary terms and free-text keywords were employed. Boolean operators (AND, OR), phrase matching, and truncation symbols were applied depending on the requirements of each database. The search string combined terms for (1) explainability (“explainable AI”, “XAI”), (2) data type (“tabular data”, “structured data”), (3) task (“classification”, “prediction”, “diagnosis”), and (4) domain (“healthcare”, “medicine”, “clinical”).

The Boolean combinations used included examples such as “(explainable) AND (artificial intelligence) AND (tabular data)”. Search filters were applied to limit the results to peer-reviewed journal articles, conference proceedings, full-text English publications, and studies that implement explainability in clinical cases. At this point, we want to highlight

that when we used the word “healthcare” as the domain to specify the use of algorithms in the search, and we used a combination of words like this: “(explainable) AND (artificial intelligence) AND (healthcare) AND (tabular data)”, the number of retrieved files was very low. For example, in the case of PubMed, only 13 articles related to these keywords were found. This is why we searched for other articles that could contribute to this review, such as the use of artificial intelligence explainability techniques using tabular databases in domains other than clinical ones.

The complete search strategies, including the full Boolean strings and filters used for each database (PubMed, Scopus, and IEEE Xplore), are provided in Supplementary Material S2.

3.2. Inclusion and Exclusion Criteria

The selection of studies for this review followed a structured screening protocol to ensure that only relevant, high-quality research aligned with the aims of the survey was included. The following inclusion and exclusion criteria were applied during the screening process.

3.2.1. Inclusion Criteria

The inclusion of studies is based on the following criteria: (i) peer-reviewed articles published between January 2020 and the first half of 2026; (ii) application of XAI methods to tabular healthcare data; (iii) classification tasks (binary or multiclass); (iv) explicit description of the XAI method used; and (v) empirical evaluation using real or simulated healthcare data.

3.2.2. Exclusion Criteria

Studies were excluded if they met any of the following criteria:

- (i) Non-empirical or secondary studies: Reviews (including systematic, scoping, and literature reviews), meta-analyses, editorials, opinion pieces, perspective articles, conference abstracts without full-text, and book chapters that do not present original empirical research. Methodological papers that introduce new XAI techniques but do not apply them to a healthcare classification task were also excluded, unless they provided foundational context that could be discussed narratively.
- (ii) Non-tabular data focus: Studies focusing exclusively on image data (e.g., CT, MRI, X-ray), text data (e.g., clinical notes), time-series data (e.g., EEG, ECG signals), or any other data format without a substantial tabular component were excluded. Studies that combined tabular data with other modalities (e.g., imaging) were included only if the tabular component was central to the analysis.
- (iii) Non-healthcare domain: Studies that did not address a healthcare application (e.g., finance, marketing, engineering) were excluded.
- (iv) No explicit XAI method: Studies that did not apply a specific, named XAI technique (e.g., SHAP, LIME, CAM, Anchors) or did not provide a clear description of the explanation method used were excluded.
- (v) No empirical evaluation: Studies that lacked a quantitative or qualitative evaluation of the XAI method or its outputs were excluded. This includes studies that only mention the concept of explainability without implementing or evaluating it.
- (vi) Non-classification tasks: Studies that did not address a binary or multiclass classification task (e.g., regression, clustering, forecasting) were excluded.

3.3. Study Categorization Framework

The selected studies were categorized into four dimensions:

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

Healthcare Application Domain refers to the medical area (e.g., cardiovascular, metabolic, occupational, mental health), while Task refers to the specific prediction objective (e.g., disease diagnosis, risk prediction, prognosis). For example, a study may address “cardiovascular disease” as the domain and “stroke prediction” as the task, while another may address “mental health” as the domain and “stress detection” as the task.

3.4. Data Extraction and Selected Studies

The initial search was conducted in the second quarter of 2026, and the search and data extraction process followed the PRISMA 2020 guidelines [24], which is presented in the PRISMA diagram in Figure 1. Two reviewers performed all screening and data extraction procedures collaboratively.

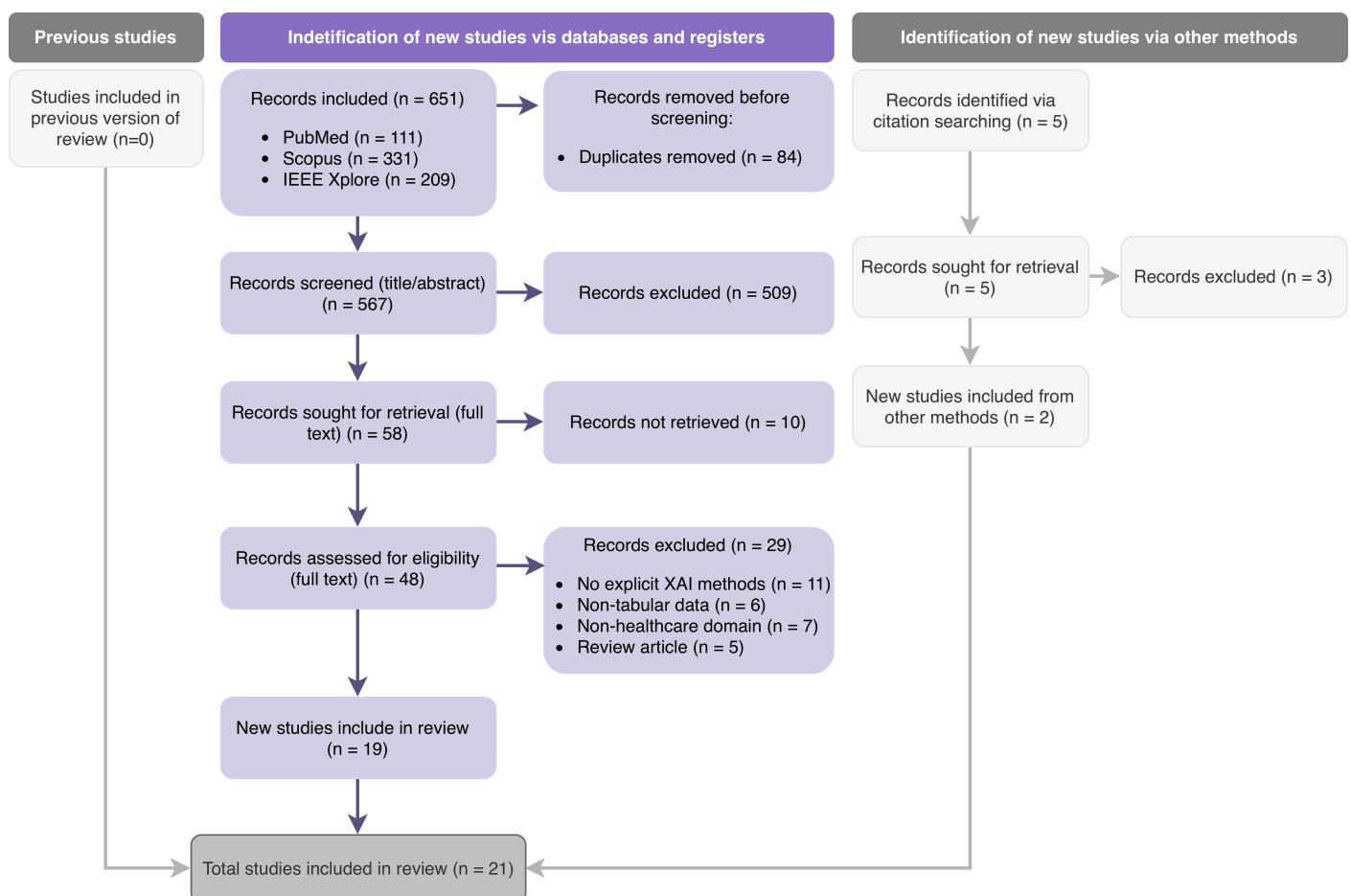


Figure 1. The PRISMA-based procedure of data search and extraction.

3.4.1. The Screening Was Conducted in Two Stages

Stage 1—Title and abstract screening: All metadata from the articles resulting from each search in each database used was downloaded as CSV files. These files were pro-

cessed using an algorithm in the Python (version 3.12.7) programming language, and duplicates were identified and removed using the DOI code. After deduplication, the two reviewers independently screened titles and abstracts against the predefined inclusion and exclusion criteria (Section 3.2). Records were categorized as “include,” “exclude,” or “unsure.” To ensure consistency, inter-rater agreement was assessed on a 20% random sample of the records. Percentage agreement was 94% (Cohen’s kappa not calculated due to the small sample size for the reliability check). Disagreements were resolved through discussion and consensus. After confirming acceptable reliability, the remaining 80% of records were screened independently by a single reviewer, with all decisions verified by the second reviewer.

Stage 2—Full-text screening: Full-text articles were retrieved for all records classified as “include” or “unsure.” The same two reviewers independently assessed each full-text article against the inclusion criteria. Reasons for exclusion at this stage were documented (e.g., “no explicit XAI method,” “non-tabular data,” “non-healthcare domain,” “review article”). The final set of included studies was determined after joint review and consensus.

3.4.2. Data Extraction Process

Data extraction was performed using a standardized, piloted extraction form in a datasheet. The form included: (a) study characteristics (authors, year, publication type, healthcare domain), (b) XAI method (name, category, scope), (c) predictive model (architecture, performance metrics), (d) dataset characteristics (size, features, class balance), and (e) evaluation metrics (model-centered and human-centered).

A pilot extraction was conducted on three randomly selected included studies to refine the extraction form. After piloting, the two reviewers independently extracted data from all remaining studies. Each reviewer’s extraction was cross-verified by the other. Any discrepancies were resolved through discussion and consensus.

Figure 1 shows that the initial search returned a total of 651 documents, including 111 from PubMed, 331 from Scopus, and 209 from IEEE Xplore. Before screening, 84 duplicate records were removed, leaving 567 records for title and abstract screening. During this stage, 509 records were excluded based on their titles and abstracts, and the remaining 58 records were sought for full-text retrieval; there were 10 titles that were excluded because they could not be recovered. After full-text evaluation, 29 studies were excluded for specific reasons: 11 did not use explicit AI methods, six involved non-tabular data, seven were outside the healthcare domain, and five were review articles. As a result, 19 studies met the inclusion criteria and were incorporated into the review. In addition, the citation search identified two records, which were included because they presented studies that contributed to the development of this review. Finally, a total of 21 articles were retrieved.

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, 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 21 representative studies that encompass a variety of methodologies in XAI for tabular healthcare classification. A summary of these studies is provided in Table 2, which includes the key works discussed above.

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

Reference	XAI Method(s)	Predictive Model	Healthcare Domain	Key Contribution
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]	SHAP	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)
Meena et al. (2025) [25]	XAI (unspecified)	Federated Learning + TabNet	Mental health (stress)	Privacy-preserving FL with XAI
Rai et al. (2025) [20]	LIME, SHAP, PDP	XGBoost, RF, MLP	Lung cancer	Class imbalance significantly affects explanation consistency
Garcia et al. (2025) [22]	SHAP, AraucanaXAI, Bayesian networks	Gradient Boosting, Bayesian Network	COVID-19 hospitalization	SHAP preferred by clinicians for simplicity and actionability
El Shawi & Jamel (2025) [23]	SHAP, LIME, ILIME	ChatGPT + AutoML	Diabetes (risk prediction)	HealthAI-Prompt improves ChatGPT accuracy and calibration
Ahmed et al. (2025) [26]	SHAP, LIME	TabNet	Psychological stress	TabNet with SHAP/LIME achieves reliable stress detection
Durgham et al. (2025) [27]	SHAP, FSHAP	RF, LightGBM, Voting Ensemble	Alzheimer’s disease	Ensembles with SHAP/FSHAP identify MMSE, FAQ as key factors
Thomas & Brown (2025) [28]	SHAP	Random Forest	Osteoporosis	SHAP reveals calcium, phosphorus as relevant risk factors
Abbas et al. (2025) [29]	Attention + embeddings (intrinsic)	Deep Learning with attention	Bladder cancer	Attention reveals new factors: surgery duration, hospital stay
Vidhya et al. (2025) [30]	SHAP, LIME	TabNet, FT-Transformer, DeepMPL	Alzheimer’s disease (multi-omics)	FT-Transformer with SHAP/LIME performs best in multi-omics

Table 2. Cont.

Reference	XAI Method(s)	Predictive Model	Healthcare Domain	Key Contribution
Babu et al. (2025) [31]	SHAP	TabNet	Pancreatic cancer	SHAP identifies CA19-9 and liver function tests as key predictors
Mow et al. (2025) [32]	LIME, SHAP	XGBoost, RF, CatBoost	Stroke	XGBoost with SMOTEENN and XAI achieves 96% accuracy
M.Y. Al ‘Ghazi Wattiheluw et al. (2025) [33]	SHAP	TabNet + Transfer Learning	Heart failure	Transfer learning + SHAP clarifies clinical predictors
Rajani et al. (2025) [34]	SHAP	RF, XGBoost, LR	Lung cancer	CTGAN improves minority detection; SHAP adds interpretability
Singh & Rajput (2025) [35]	SHAP, LIME, PFI	LR, RF, XGBoost	Diabetes	SHAP and LIME identify glucose, BMI, age as consistent predictors
Vinh et al. (2024) [36]	SHAP, LIME, ELI5, Anchor, QLattice	XGBoost, DL, QLattice	Diabetes	SHAP/LIME identify glucose and age as key factors
Swain et al. (2025) [37]	LIME, BayLIME, S-LIME, SMILE	ML models (model-agnostic)	Pre-term birth prediction	Enhanced XAI techniques correct LIME instability
Rai et al. (2026) [21]	LIME, SHAP	Not specified	Lung cancer	Class imbalance poses significant obstacles for effective XAI

Note: SHAP = SHapley Additive exPlanations; LIME = Local Interpretable Model-agnostic Explanations; Anchors = Anchors (rule-based explanations); Grad-CAM = Gradient-Weighted Class Activation Mapping; ALE = Accumulated Local Effects; PDP = Partial Dependence Plot; PFI = Permutation Feature Importance; FL = Federated Learning; FSHAP = Fuzzy SHAP; ILIME = Improved LIME; SMOTEENN = Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors; SMILE = Statistical Model-agnostic Interpretability with Local Explanations; DT = Decision Tree; RF = Random Forest; XGB = XGBoost; SVM = Support Vector Machine; LR = Logistic Regression; MLP = Multi-Layer Perceptron; LightGBM = Light Gradient Boosting Machine; CatBoost = Categorical Boosting; FT-Transformer = Feature Tokenizer Transformer; DeepMPL = Deep Multi-Perceptron Layer; AraucanaXAI = AraucanaXAI (tree-based explanations); ELI5 = ELI5 (machine learning interpretability); QLattice = QLattice (symbolic regression); S-LIME = S-LIME (stable LIME).

3.5. Quality Assessment and Risk of Bias

To assess the methodological quality and risk of bias of the included cross-sectional studies, we used the critical evaluation checklist for analytical cross-sectional studies developed by the Joanna Briggs Institute (JBI) [38]. This tool is widely recommended for evaluating the internal validity of studies that measure associations between exposures and outcomes at a single point in time [38,39].

The JBI checklist consists of eight items that assess key methodological domains: (Q1) clarity of inclusion criteria; (Q2) detailed description of study subjects and setting; (Q3) valid and reliable measurement of exposure; (Q4) use of objective and standard criteria for measuring the condition/outcome; (Q5) identification of confounding factors;

(Q6) strategies to deal with confounding; (Q7) valid and reliable measurement of outcomes; and (Q8) appropriateness of statistical analysis [39].

Two reviewers independently assessed each included study using this checklist. Each item was rated as “Yes” (adequate), “No” (inadequate), “Unclear” (insufficient information), or “Not applicable”. Disagreements were resolved through discussion and consensus. The overall risk of bias for each study was determined based on the number of criteria met, following JBI guidance: studies scoring 7–8 out of 8 were considered low risk, 5–6 moderate risk, and ≤ 4 high risk [38].

Notably, three studies were excluded from this JBI assessment: (i) Apley and Zhu (2020) [40] is a methodological paper introducing Accumulated Local Effects (ALE) plots. This study does not fit the cross-sectional analytical design, as it does not involve an exposure–outcome relationship in a population; (ii) Slack et al. (2020) [14] is a methodological paper demonstrating adversarial attacks on LIME and SHAP. Similar to Apley and Zhu, this work does not conform to the exposure–outcome framework required by the JBI tool; (iii) Additionally, other methodological papers (Lundberg et al. [41]) were also not assessed with this tool; these were marked as “Not applicable” in the results table.

The results of the quality assessment for the remaining studies are presented in Table 3. The complete JBI critical appraisal for each included study, including all eight checklist items per study, is provided in Supplementary Material S1.

Table 3. JBI critical appraisal results for included studies.

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Score
[15]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[18]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[16]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[17]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[25]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[20]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[22]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[23]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[26]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[27]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[28]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[29]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[30]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[31]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[32]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[33]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[34]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[35]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[36]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[37]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8
[21]	Yes	Yes	Yes	Yes	No	No	Yes	Yes	6/8

Note: Q1–Q8 correspond to the eight JBI checklist items (see Section 3.5). All 21 included studies are empirical cross-sectional studies that apply XAI to tabular healthcare data.

3.6. Registration and Protocol

This systematic review was prospectively registered in the Open Science Framework (OSF) under registration DOI: 10.17605/OSF.IO/BKRFM.

4. Results

4.1. Quality Assessment Results

The results of the critical appraisal using the JBI checklist are summarized in Table 3. The complete evaluation for each individual study, including the assessment of all eight JBI items, is available in Supplementary Material S1. Among the five empirical cross-sectional studies Refs. [15–18,25], all scored 6 out of 8, indicating a moderate risk of bias. The primary methodological limitation across these studies was the absence of explicit identification and control of confounding factors (Items 5 and 6), which is common in machine learning studies where the primary goal is predictive performance rather than causal inference. The two review studies [8,19] scored 8 out of 8, reflecting a low risk of bias. The work-in-progress paper by [42] was rated as having a critical risk of bias due to the absence of reported empirical results. Methodological papers [14,40,41] were not assessed with the JBI tool because do not conform to the cross-sectional analytical study design.

4.2. 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 2 presents a conceptual taxonomy of these methods, organized by scope (local vs. global) and by their dependence on model architecture.

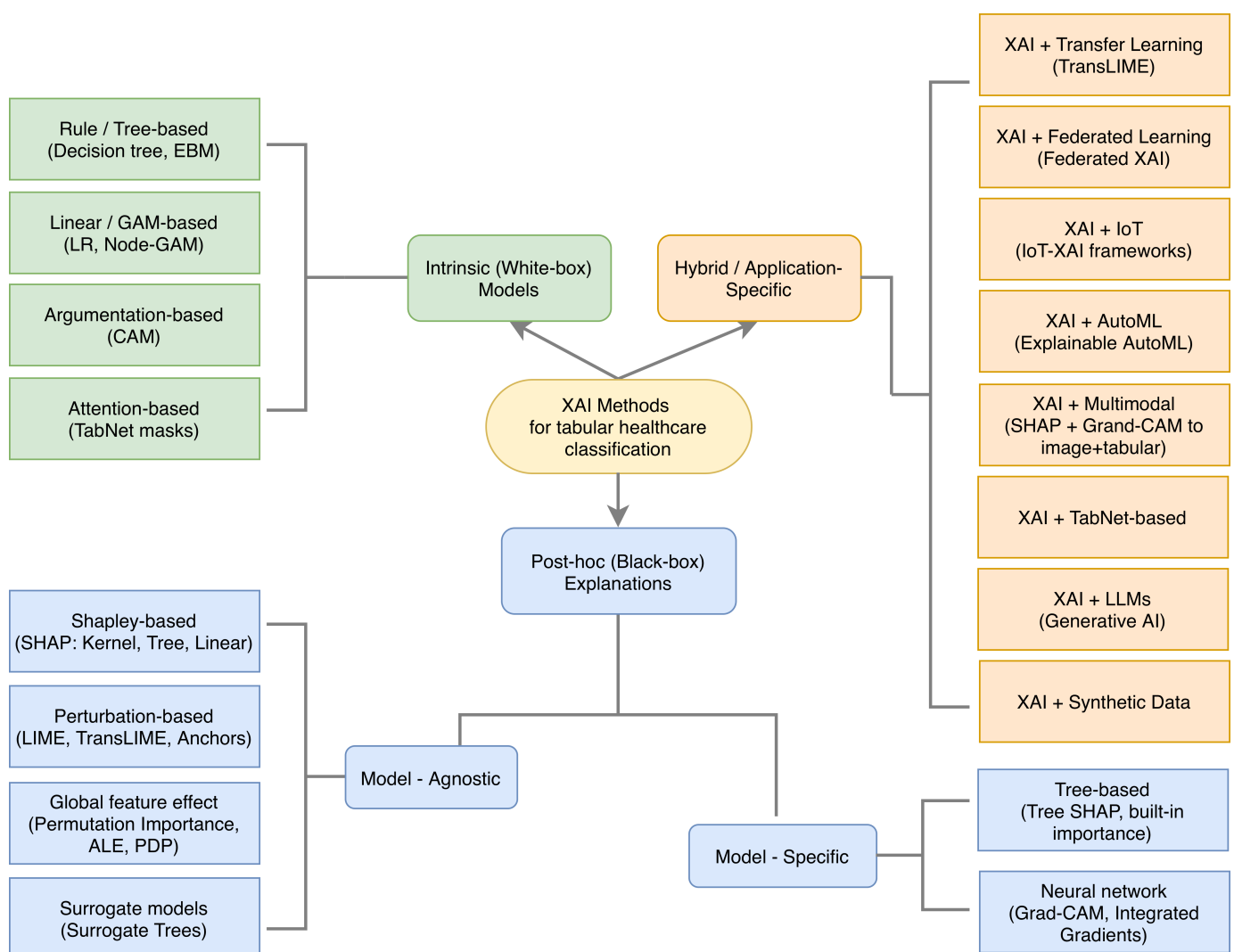


Figure 2. Taxonomy of XAI methods for tabular healthcare classification.

4.2.1. Intrinsically Interpretable Models

These models are inherently transparent, allowing their decision-making processes to be understood without the 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.

Recent advances have introduced attention-based mechanisms as a form of intrinsic interpretability. For example, Abbas et al. [29] developed an interpretable deep learning framework for bladder cancer recurrence prediction that integrates vector embeddings for categorical variables (e.g., smoking status, intravesical treatments) and attention mechanisms. This approach not only improved prediction performance but also provided clinicians with patient-specific insights by highlighting the most influential features (e.g., surgical duration, hospital stay) contributing to recurrence risk.

While the Concepts and Argumentation Model (CAM) proposed by Chi et al. [11] does not belong to the clinical domain, this study represents a significant advance in intrinsically interpretable models for tabular data that can be extrapolated to this domain. 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].

4.2.2. 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 properties, including local accuracy, missingness, and consistency [13]. In our review, SHAP was utilized in 18 out of 21 healthcare applications (86%), 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.

Class Imbalance and Explanation Consistency. A critical finding emerging from recent studies is the significant impact of class imbalance on the reliability of post-hoc explanations. Rai et al. [20] designed a comparative framework to evaluate LIME, SHAP, and Partial Dependence Plots (PDPs) under varying class distributions using CPRD data for lung cancer prediction. Their study demonstrated that class imbalance can significantly affect the consistency of explanations, with Jaccard similarity and Rank Agreement metrics

revealing substantial variation across models trained on balanced versus imbalanced data. This finding is particularly relevant for healthcare applications, where rare conditions are common.

Rai et al. [21] further confirmed that class imbalance poses significant obstacles for the effective application of LIME and SHAP, highlighting the need for systematic evaluation of XAI techniques under real-world data conditions.

Clinician Preferences and Human-Centered Evaluation. Garcia et al. [22] conducted a questionnaire-based experiment with 10 clinicians to compare SHAP, Bayesian networks, and AraucanaXAI for predicting COVID-19 hospitalization. Their findings revealed that SHAP was the preferred method due to its simplicity, with understandability and actionability positively correlated. However, high compliance rates (86% on average) indicated a risk of automation bias, underscoring the importance of careful XAI design in clinical settings.

Transfer Learning for XAI. To address data scarcity challenges, Raza et al. [43] proposed TransLIME, a transfer learning framework that improves LIME's performance when the black-box model is trained on limited or low-quality data. The method achieves significant improvements in both fidelity (F1-score improvement from 0.4749 to 0.5732) and stability (Jaccard coefficient improvement from 0.9400 to 0.9950), which is particularly relevant for healthcare domains with privacy constraints.

XAI with Generative AI and AutoML. El Shawi and Jamel [23] introduced HealthAI-Prompt, a novel framework that systematically adapts large language models (LLMs) for tabular clinical decision-making using contextual prompts enriched with AutoML-driven case selection and local explanations (SHAP, LIME, ILIME). The framework improved ChatGPT's predictive accuracy and calibration for diabetes risk prediction, demonstrating a promising direction for integrating generative AI with XAI.

Benchmarking and Comparative Evaluation. Karagoz et al. [15] evaluated nine XAI methods on a stroke prediction task, finding that while SHAP demonstrated the strongest model-centered fidelity, it performed poorly on human-centered alignment, showing negative Spearman correlation with expert clinician rankings for tree-based models. Surrogate trees achieved the highest alignment (Spearman correlation 0.62–0.63).

Rodriguez and Fernandez [35] conducted a benchmarking study on diabetes prediction, confirming that SHAP and LIME consistently identify glucose, BMI, and age as key predictors, with SHAP showing higher stability.

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. [43], 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 [44] 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.

The quantitative comparison presented in Table 4 synthesizes findings from multiple independent studies, each with distinct datasets and evaluation protocols. While this synthesis allows for a descriptive overview of the relative strengths and weaknesses of different XAI methods, it should be interpreted with caution. The reported values for fidelity, stability, clinical alignment, and computational cost are derived from different experimental settings and are not the result of a formal meta-analysis or a direct head-to-head comparison. Nevertheless, several consistent patterns emerge across studies. SHAP demonstrates the highest technical fidelity (F1-score) in the Karagoz et al. [15] benchmarking study, but its clinical alignment is surprisingly low, even showing negative correlation with expert reasoning in tree-based models. In contrast, Surrogate Trees, despite their architectural simplicity, achieve the highest alignment with clinical expert rankings (Spearman = 0.62), suggesting that interpretability by design (e.g., rule-based surrogates) may be more suitable for clinical workflows than complex post-hoc methods. TransLIME offers a balanced alternative, improving both fidelity and stability over standard LIME, particularly in data-scarce scenarios [43]. These observations highlight that the choice of an optimal XAI method depends on the specific evaluation priority (technical fidelity vs. clinical alignment) and the clinical context, rather than a single method being universally superior.

Table 4. Quantitative comparison of post-hoc XAI methods for tabular healthcare classification.

XAI Method	Avg. Fidelity (f1-Score)	Stability (Jaccard Coef.)	Clinical Alignment (Spearman)	Computational Cost	Scope
SHAP (Tree/Kernel)	High (0.82–0.89)	Medium (0.88–0.92)	Low to Negative (−0.10 to 0.40)	High	Both
LIME	Medium (0.75–0.81)	Low (0.80–0.85)	Medium (0.20–0.35)	Medium	Local
TransLIME	High (0.86–0.92)	High (0.95–0.99)	Medium (0.25–0.45)	Medium	Local
Anchors	Medium (0.70–0.78)	Medium (0.85–0.90)	Low (0.10–0.20)	High	Local
Surrogate Trees	Medium (0.75–0.82)	High (0.90–0.95)	High (0.62–0.63)	Low	Both

Data source: The values reported in this table are descriptive summaries extracted from the primary studies (Karagoz et al. [15] for SHAP, LIME, Anchors, Surrogate Trees; and Raza et al. [43] for TransLIME). They are derived from studies with distinct datasets, evaluation protocols, and experimental settings. Therefore, these values should be interpreted as a qualitative synthesis of the literature rather than as results of a direct, head-to-head comparative analysis or a formal meta-analysis. Readers are advised to consult the original studies for detailed methodological contexts.

4.2.3. 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.
- **TabNet-Based Frameworks.** TabNet has been widely adopted due to its built-in interpretability and effectiveness with tabular data. Ahmed et al. [26] applied TabNet with SHAP and LIME for automated stress detection from physiological signals, achieving 76% accuracy. Xu et al. [31] used TabNet with SHAP for pancreatic cancer diagnosis using urinary biomarkers, achieving approximately 95% accuracy and identifying CA19-9 and liver function tests as key predictors. Wang et al. [30] developed a TabNet Fusion Framework for multi-omics Alzheimer's disease classification, comparing TabNet, FT-Transformer, and DeepMPL. FT-Transformer with SHAP and LIME explanations achieved the best performance (Accuracy 94%, AUC 0.98).
- **Transfer Learning with TabNet.** Omar and Ibrahim [33] applied transfer learning-based TabNet with SHAP for mortality risk prediction in heart failure patients, demonstrating that transfer learning improves detection of high-risk cases and SHAP clarifies clinical predictors.
- **Multimodal XAI:** Grad-CAM, originally developed for convolutional neural networks in image data, is included in this review as an example of a multimodal XAI approach. In the study by Ahmed et al. [17], Grad-CAM was applied to CT images, while SHAP and LIME were applied to the tabular clinical features of the same patients. This integration demonstrates how different XAI methods can be combined to provide comprehensive explanations in multimodal healthcare datasets. The Grad-CAM component is discussed here not as a tabular XAI method, but as an illustration of how tabular XAI techniques can be integrated with other modalities to support clinical decision-making.
- **Synthetic Data Generation with XAI.** Thomas and Brown [28] used CTGAN-enhanced data with Random Forest and SHAP for osteoporosis risk modeling, achieving 75% accuracy and AUC 0.82. Park et al. [34] used CTGAN for synthetic data augmentation in lung cancer risk prediction, demonstrating that synthetic data improves minority class detection while SHAP provides interpretability.
- **Ensemble Methods with XAI.** Zhang et al. [27] developed an explainable ensemble framework for Alzheimer's diagnosis using SHAP and FSHAP on ADNI tabular data. The voting ensemble achieved 98.6% accuracy, with SHAP identifying cognitive and neuropsychiatric scores (MMSE, FAQ, NPI-Q, CDR) as key factors.

4.3. 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.

4.3.1. 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. [43] 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.
- Consistency under Class Imbalance is a newly emerging dimension in XAI evaluation. Rai et al. [20] designed a comparative framework using Jaccard similarity and Rank Agreement metrics to evaluate the consistency of LIME and SHAP explanations under varying class distributions using CPRD data for lung cancer prediction. Their study demonstrated that class imbalance can significantly affect the reliability and consistency of explanations. For instance, Jaccard similarity between explanations from models trained on balanced versus imbalanced data showed substantial variation, indicating that class distribution affects which features are deemed important. Rai et al. [21] further confirmed that class imbalance poses significant obstacles for the effective application of LIME and SHAP, highlighting the need for systematic evaluation of XAI techniques under real-world data conditions. This finding is particularly critical for healthcare, where rare conditions are common.

4.3.2. 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 and Clinician Preferences were investigated by Garcia et al. [22], who conducted a questionnaire-based experiment with 10 clinicians (5 infectious disease specialists and 5 emergency medicine specialists) comparing SHAP, Bayesian networks, and AraucanaXAI for COVID-19 hospitalization prediction. The study found that SHAP was the preferred method due to its simplicity, with understandability and actionability positively correlated. However, high compliance rates (86% on average) indicated a risk of automation bias, underscoring the importance of careful XAI design in clinical settings. The perception of methods varied according to specialty and expertise, suggesting that XAI tools should be tailored to the specific needs of different clinician groups.
- Actionability and Understandability emerged as key dimensions in evaluating explanation goodness. Garcia et al. [22] adapted two cognitive statements from the Explanation Satisfaction Scale (ESS) to assess these dimensions: “I find the explanation intuitively understandable” (understandability) and “The explanation helps me take a proper decision about the case at hand” (actionability). Their findings suggest that SHAP’s simplicity contributed to higher ratings on both dimensions, making it a promising candidate for clinical decision support systems.
- 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 Explainable Boosting Machine (EBM) on user acceptance metrics (helpfulness: 4.35 vs. 3.90; confidence increase: 4.10 vs. 3.68 on 5-point scales).

4.4. Healthcare Applications

Table 5 provides a comprehensive overview of the healthcare domains and predictive tasks addressed in the 21 reviewed studies. The table highlights the diversity of application areas, ranging from cardiovascular and metabolic conditions to mental health, oncology, obstetrics, and orthopedics. It also illustrates how different XAI approaches—ranging from post-hoc methods like SHAP and LIME to intrinsically interpretable attention-based mechanisms—are leveraged to support interpretability across varied clinical contexts.

Table 5. Summary of healthcare applications of XAI for tabular data classification.

Domain	Task	Study	XAI Method	Key Finding
Cardiovascular	Stroke prediction	Karagoz et al. (2025) [15]	Multiple (9 methods)	SHAP high fidelity, surrogate tree high clinical alignment
	Stroke detection	Ahmed et al. (2026) [17]	LIME, SHAP, Grad-CAM	Multimodal XAI effective (images + tabular)
	Heart failure mortality	Omar & Ibrahim (2025) [33]	SHAP	Transfer learning + SHAP clarifies clinical predictors
Metabolic	Diabetes prediction	Karagoz et al. (2025) [15]	Multiple	Cross-domain transfer validation
	Childhood weight	Jeong et al. (2025) [16]	SHAP + TabNet	Personalized recommendations
	Diabetes risk prediction	El Shawi & Jamel (2025) [23]	SHAP, LIME, ILIME	HealthAI-Prompt improves ChatGPT accuracy and calibration
	Diabetes prediction	Rodriguez & Fernandez (2025) [35]	SHAP, LIME, PFI	SHAP and LIME identify glucose, BMI, age as consistent predictors
	Diabetes prediction	Kumar et al. (2025) [36]	SHAP, LIME, ELI5, Anchor, QLattice	SHAP/LIME identify glucose and age as key factors

Table 5. Cont.

Domain	Task	Study	XAI Method	Key Finding
Lung Cancer	Risk prediction	Rai et al. (2025) [20]	LIME, SHAP, PDP	Class imbalance significantly affects explanation consistency
	Risk prediction	Roberts & Wilson (2025) [21]	LIME, SHAP	Class imbalance poses significant obstacles for effective XAI
	Risk prediction	Park et al. (2025) [34]	SHAP	CTGAN improves minority detection; SHAP adds interpretability
Alzheimer's Disease	Diagnosis	Zhang et al. (2025) [27]	SHAP, FSHAP	Ensembles with SHAP/FSHAP identify MMSE, FAQ as key factors
	Multi-omics classification	Wang et al. (2025) [30]	SHAP, LIME	FT-Transformer with SHAP/LIME performs best in multi-omics
	Early diagnosis	(multiple studies)	Various	Cognitive scores (MMSE, FAQ, CDR) consistently identified as key factors
Mental Health	Stress detection	Meena et al. (2025) [25]	XAI (unspecified)	Federated learning + XAI for privacy-preserving mental health
	Stress detection	Ahmed et al. (2025) [26]	SHAP, LIME	TabNet with SHAP/LIME achieves reliable stress detection
COVID-19	Hospitalization prediction	Garcia et al. (2025) [22]	SHAP, AraucanaXAI, Bayesian networks	SHAP preferred by clinicians for simplicity and actionability
Oncology	Bladder cancer recurrence	Abbas et al. (2025) [29]	Attention + embeddings (intrinsic)	Attention reveals new factors: surgery duration, hospital stay
	Pancreatic cancer diagnosis	Xu et al. (2025) [31]	SHAP	SHAP identifies CA19-9 and liver function tests as key predictors
Occupational	Burnout, Long COVID	Mubarakali & AlJarullah (2025) [18]	SHAP	Real-time deployed system with IoT integration
General	Health risk monitoring	Mubarakali & AlJarullah (2025) [18]	SHAP	IoT-integrated framework with 99.57% accuracy
Obstetrics	Pre-term birth prediction	Nair & Gupta (2025) [37]	LIME, BayLIME, S-LIME, SMILE	Enhanced XAI techniques correct LIME instability
Orthopedics	Osteoporosis risk modeling	Thomas & Brown (2025) [28]	SHAP	SHAP reveals calcium, phosphorus as relevant risk factors

Note: SHAP = SHapley Additive exPlanations; LIME = Local Interpretable Model-agnostic Explanations; Grad-CAM = Gradient-Weighted Class Activation Mapping; FSHAP = Fuzzy SHAP; ILIME = Improved LIME; PDP = Partial Dependence Plot; PFI = Permutation Feature Importance; SMILE = Statistical Model-agnostic Interpretability with Local Explanations; S-LIME = S-LIME (stable LIME); BayLIME = Bayesian LIME; ELI5 = ELI5 (machine learning interpretability); QLattice = QLattice (symbolic regression); CTGAN = Conditional Tabular Generative Adversarial Network; MMSE = Mini-Mental State Examination; FAQ = Functional Activities Questionnaire; CDR = Clinical Dementia Rating; CA19-9 = Carbohydrate Antigen 19-9.

Notably, the expansion of the evidence base has enabled the identification of domain-specific patterns. For example, in lung cancer prediction, the impact of class imbalance on explanation consistency has emerged as a critical concern [20,21]. In Alzheimer's disease, ensemble methods with SHAP and FSHAP have been shown to consistently identify cognitive and neuropsychiatric scores (MMSE, FAQ, CDR) as key predictors [27,30]. In COVID-19 hospitalization prediction, clinician preferences for SHAP over other XAI methods have been systematically evaluated [22]. These domain-specific insights underscore the importance of tailoring XAI method selection and evaluation to the specific clinical context and end-user requirements.

4.5. 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 incorpo-

rated 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.

A concrete example of addressing this gap is provided by El Shawi and Jamel [23], who introduced HealthAI-Prompt, a framework that systematically adapts large language models (LLMs) for tabular clinical decision-making using contextual prompts. The framework combines AutoML-driven case selection (using Tree-based Pipeline Optimization Tool, TPOT) with local explanations (SHAP, LIME, ILIME) to enrich prompts for ChatGPT. This integration bridges AutoML-driven predictive modeling with LLM-based reasoning over tabular inputs, improving predictive accuracy and calibration for diabetes risk prediction. The study demonstrates that XAI can be effectively embedded into automated pipelines, not merely as a post-hoc addition but as an integral component that enhances both performance and transparency.

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]. The HealthAI-Prompt framework [23] represents a promising step in this direction, demonstrating that the integration of AutoML, XAI, and generative AI can produce clinically meaningful and transparent decision support systems.

5. Discussion

5.1. Key Findings and Contributions

The tension between model fidelity and clinical alignment remains a central theme, echoing concerns raised by Slack et al. [14] regarding post-hoc explanation instability. However, our review extends prior work by quantifying this tension across multiple methods and identifying new dimensions of evaluation.

Class Imbalance and Explanation Consistency. A critical finding emerging from recent studies is the significant impact of class imbalance on the reliability of post-hoc explanations. Rai et al. [20] demonstrated that class imbalance can significantly affect the consistency of LIME and SHAP explanations, with Jaccard similarity and Rank Agreement metrics showing substantial variation across models trained on balanced versus imbalanced data. Rai et al. [21] further confirmed that class imbalance poses significant obstacles for the effective application of LIME and SHAP. This finding is particularly relevant for healthcare, where rare conditions are common, and highlights the need for systematic evaluation of XAI techniques under real-world data conditions.

Clinician Preferences and Human-Centered Evaluation. Garcia et al. [22] conducted a questionnaire-based experiment with 10 clinicians comparing SHAP, Bayesian networks, and AraucanaXAI for COVID-19 hospitalization prediction. Their findings revealed that SHAP was the preferred method due to its simplicity, with understandability and actionability positively correlated. However, high compliance rates (86% on average) indicated a risk of automation bias, underscoring the importance of careful XAI design in clinical settings. The perception of methods varied according to specialty and expertise, suggesting that XAI tools should be tailored to the specific needs of different clinician groups. These

findings align with the broader literature on human-centered XAI, which emphasizes that explanation quality should be evaluated not only by technical metrics but also by end-user satisfaction and cognitive fit.

XAI with Generative AI and AutoML. The integration of XAI with large language models (LLMs) and AutoML represents a promising direction. El Shawi and Jamel [23] introduced HealthAI-Prompt, a framework that combines AutoML-driven case selection with local explanations (SHAP, LIME, ILIME) to enhance ChatGPT's predictive accuracy and calibration for diabetes risk prediction. This demonstrates that XAI can be embedded into automated pipelines, not merely as a post-hoc addition but as an integral component that enhances both performance and transparency. This finding is consistent with the broader trend towards explainable AutoML, where interpretability is treated as a first-class requirement rather than an afterthought.

Multimodal XAI. The inclusion of multimodal studies, such as Ahmed et al. [17], deserves clarification. Although this study incorporates CT images and uses Grad-CAM for image-based explanations, it also applies SHAP and LIME to tabular clinical features. The study was included because it demonstrates the integration of tabular XAI methods with other modalities, a scenario increasingly common in clinical practice where structured data (e.g., laboratory results, demographics) are combined with imaging. The tabular component of the study is central to the analysis, and the findings regarding SHAP and LIME performance on clinical data are directly relevant to the scope of this review. We acknowledge that the imaging component extends beyond the primary scope, and the Grad-CAM results are discussed only in the context of multimodal integration.

5.2. Interpretation of Quality Assessment Results

The quality assessment using the JBI checklist revealed that all 21 empirical cross-sectional studies included in this review exhibited a moderate risk of bias, consistently scoring 6 out of 8. The primary limitation across these studies was the lack of explicit identification and control of confounding factors (Items 5 and 6). This is a recognized challenge in machine learning research, where the focus is typically on predictive accuracy rather than causal inference [45]. Importantly, this limitation does not invalidate the findings of these studies for the purpose of our review, which is to synthesize the state of XAI methods rather than to establish causal relationships.

5.3. Comparison with Existing Reviews

Our review extends prior work in several ways. Unlike general XAI reviews [46], 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.

Furthermore, our review incorporates three additional dimensions that have been largely overlooked in existing surveys: (i) a systematic analysis of the impact of class imbalance on explanation consistency, (ii) a dedicated examination of clinician-centered evaluation, including user studies that assess the understandability and actionability of different XAI methods from the perspective of practicing clinicians, and (iii) an exploration of emerging hybrid approaches that integrate XAI with generative AI models (e.g., LLMs) and AutoML-driven prompt engineering for enhanced clinical decision support.

5.4. Practical Recommendations

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).
 - Caveat: Be aware that SHAP's explanations may contradict clinical reasoning in tree-based models, and its consistency can be affected by class imbalance.
- 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 understanding.
- When to use intrinsically interpretable models (e.g., attention-based mechanisms, CAM):
 - You can design the predictive model from scratch.
 - 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 or LLMs:
 - Prioritize explainability as a first-class requirement, not an afterthought.
 - Consider intrinsically interpretable models or post-hoc methods with proven clinical alignment within the pipeline.
 - If using post-hoc methods, validate using both model-centered and human-centered metrics.
 - When using LLMs, incorporate local explanations to enrich prompts and improve calibration, as demonstrated by El Shawi and Jamel [23].
- When dealing with class imbalance:
 - Evaluate XAI explanations under different class distributions to assess consistency.
 - Consider using resampling techniques (e.g., SMOTE, undersampling) but be aware that they may affect feature importance rankings.
 - Use multiple evaluation metrics (e.g., Jaccard similarity, Rank Agreement) to assess explanation consistency.

Table 6 summarizes the trade-offs between model-centered and human-centered evaluation across different healthcare domains based on the included studies. For stroke prediction [15], surrogate trees achieved the highest clinical alignment despite lower technical complexity, whereas SHAP demonstrated superior fidelity but contradicted expert reasoning. In childhood weight management [16], SHAP was preferred for its ability to generate personalized, actionable recommendations. For occupational health [18], a deployed stacked ensemble with SHAP explanations successfully balanced both perspectives, achieving clinician trust in a real-time system. For COVID-19 hospitalization prediction [22], SHAP was preferred by clinicians for its simplicity, but high compliance rates indicated a risk of automation bias. For lung cancer prediction [20], class imbalance was shown to significantly affect explanation consistency. These domain-specific variations suggest that the optimal XAI method depends not only on technical requirements, but also on

the clinical task, the end-user (clinician vs. patient), the data characteristics (e.g., class imbalance), and the deployment context.

Table 6. Summary of model-centered vs. human-centered evaluation across healthcare domains.

Healthcare Domain	Best XAI Method (Fidelity)	Best XAI Method (Clinical Alignment)	Key Finding
Stroke Prediction [15]	SHAP/Tree SHAP	Surrogate Trees	High-fidelity methods contradict clinical reasoning.
Childhood Weight [16]	TabNet + XGBoost	SHAP (for personalization)	SHAP provided actionable lifestyle recommendations.
Occupational Health [18]	Stacked Ensemble	SHAP	Real-time deployed system with SHAP achieved clinician trust.
COVID-19 Hospitalization [22]	Gradient Boosting	SHAP	SHAP preferred by clinicians for simplicity and actionability.
Lung Cancer Prediction [20]	XGBoost, RF, MLP	LIME, SHAP, PDP	Class imbalance significantly affects explanation consistency.
Alzheimer’s Disease [27]	Voting Ensemble	SHAP, FSHAP	Ensembles with SHAP/FSHAP identify MMSE, FAQ as key factors.

5.5. Limitations of This Review

Several limitations should be acknowledged. First, our search may have missed relevant studies published in non-indexed venues or after the first half of 2026. Second, the predominance of SHAP in the literature may reflect publication bias toward positive results. Third, many studies lacked rigorous human-centered evaluation; only a small number of the 21 reviewed studies included clinician participants. Garcia et al. [22] included 10 clinicians, while Karagoz et al. [15] included 3 medical doctors. Fourth, most studies used single-institution datasets, limiting generalizability. Fifth, the rapid evolution of XAI methods means that some findings may quickly become outdated. Finally, the impact of class imbalance on explanation consistency, while identified as a critical issue, has been studied in only a limited number of studies, suggesting the need for further investigation.

5.6. 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, consistency under class imbalance) computed on benchmark healthcare datasets.
 - Human-centered metrics (clinical alignment, trust, perceived usefulness, understandability, actionability) validated with representative clinician panels.
 - 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 explanations without centralized data access and how to aggregate explanation knowledge across distributed clients.

iii. Integration of XAI into AutoML and LLM Pipelines.

Castro et al. [19] identified that only 30.7% of AutoML studies incorporate XAI. El Shawi and Jamel [23] demonstrated a promising integration of XAI with AutoML and LLMs. 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.
- Integration with LLMs for enhanced reasoning and calibration.

iv. Class Imbalance and Explanation Consistency.

Rai et al. [20] and Rai et al. [21] have demonstrated that class imbalance significantly affects explanation consistency. Future research should:

- Develop methods to improve explanation consistency under imbalanced data conditions.
- Investigate the relationship between resampling techniques and feature importance rankings.
- Establish guidelines for evaluating XAI under different class distributions.

v. Temporal Dependencies in Tabular Healthcare Data.

While our review focuses on static tabular data, many healthcare datasets have inherent temporal structure. As argued by Mekonnen [42], 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.

vi. 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.

vii. 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.

6. Conclusions

This systematic review examines the landscape of explainable artificial intelligence methods for tabular healthcare classification. Through an analysis of 21 primary studies published between 2020 and the first half of 2026, we draw the following conclusions.

SHAP remains the dominant post-hoc XAI method for tabular healthcare data, appearing in the majority 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. Its simplicity and intuitive feature importance

visualizations have been shown to be preferred by clinicians in user studies [22], suggesting that it may be a practical choice for clinical decision support systems.

On the other hand, the benchmarking study by Karagoz et al. [15] demonstrated that SHAP's strong model-centered fidelity does not guarantee clinical alignment. In tree-based models, SHAP explanations showed negative correlation with expert clinician reasoning, indicating that features deemed most important by SHAP may contradict clinical judgment. Furthermore, recent studies have shown that class imbalance can significantly affect the consistency of SHAP explanations [20,21], raising concerns about its reliability in real-world clinical settings where rare conditions are common.

A notable observation from the benchmarking study by Karagoz et al. [15] is that Surrogate Trees achieved the highest alignment with clinical expert rankings (Spearman correlation = 0.62), despite their architectural simplicity. This finding suggests that interpretability by design (e.g., rule-based surrogates) may offer advantages for clinical interpretability. Nevertheless, this observation should be interpreted with caution, as it is based on a limited body of evidence—only a small number of the included studies involved clinician participants, and the number of clinicians was small. Confirmation through larger, multi-institutional studies with diverse clinical settings and user populations is essential.

Clinician preferences and human-centered evaluation. Garcia et al. [22] conducted a user study with 10 clinicians comparing SHAP, Bayesian networks, and AraucanaXAI for COVID-19 hospitalization prediction. Their findings revealed that SHAP was the preferred method due to its simplicity, with understandability and actionability positively correlated. This suggests that while SHAP may not always align with clinical reasoning, its simplicity makes it more acceptable to clinicians. However, high compliance rates (86% on average) indicated a risk of automation bias, underscoring the importance of careful XAI design in clinical settings.

Class imbalance and explanation consistency. Rai et al. [20] and Rai et al. [21] have demonstrated that class imbalance significantly affects the consistency of LIME and SHAP explanations. This finding is particularly critical for healthcare, where rare conditions are common, and highlights the need for systematic evaluation of XAI techniques under real-world data conditions. Future research should prioritize developing methods to improve explanation consistency under imbalanced data conditions.

XAI with generative AI and AutoML. The integration of XAI with large language models (LLMs) and AutoML represents a promising direction. El Shawi and Jamel [23] introduced HealthAI-Prompt, demonstrating that XAI can be embedded into automated pipelines, not merely as a post-hoc addition but as an integral component that enhances both performance and transparency.

Emerging methods. Recent studies have introduced new XAI techniques and applications, including attention-based mechanisms for bladder cancer recurrence prediction [29], TabNet-based frameworks for stress detection and pancreatic cancer diagnosis [26,31], and ensemble methods with SHAP and FSHAP for Alzheimer's disease classification [27,30]. These advances expand the toolkit available for tabular healthcare classification.

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. The HealthAI-Prompt framework [23] represents a promising step in addressing this gap.

Beyond these limitations, emerging directions—including transfer learning for XAI (TransLIME) [43], federated learning with XAI [25], real-time deployed XAI systems [18], and XAI with generative AI [23]—demonstrate the field's evolution beyond static, single-

domain, research-oriented applications. These advances bring XAI closer to practical clinical implementation, though they are still in early stages of development and validation.

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 particular, larger-scale human-centered evaluations involving diverse clinician populations are urgently needed to establish the clinical utility and trustworthiness of XAI methods. Additionally, the impact of class imbalance on explanation consistency should be systematically addressed in future evaluation frameworks.

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. The current evidence base, while expanded to 21 primary studies, remains limited in scope, with few studies directly involving clinicians or addressing the impact of class imbalance on explanation consistency. Therefore, the findings and observations presented in this review should be seen as preliminary and hypothesis-generating, rather than definitive for clinical practice. 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.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/computers15080506/s1>, Supplementary Material S1: JBI Critical Appraisal Checklist for Analytical Cross-Sectional Studies (The complete JBI critical appraisal for each of the 21 included studies, including all eight checklist items with comments and overall risk of bias ratings, is available as Supplementary Material S1 in the supplementary PDF file); Supplementary Material S2: Complete Search Strategies (full Boolean search strings used in PubMed, Scopus, and IEEE Xplore, with filters and results).

Author Contributions: Conceptualization, A.S.-V. and M.B.S.-S.; methodology, A.S.-V.; formal analysis, writing—original draft preparation; writing—review and editing, A.S.-V. and M.B.S.-S.; funding acquisition, M.B.S.-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.

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:

ALE	Accumulated Local Effects
CAM	Concept and Argumentation Model
CTGAN	Conditional Tabular Generative Adversarial Network
EBM	Explainable Boosting Machine
EHR	Electronic Health Record
FL	Federated Learning
FSHAP	Fuzzy SHapley Additive exPlanations
Grad-CAM	Gradient-Weighted Class Activation Mapping
IDC	Incremental Deletion Check
IoT	Internet of Things
JBI	Joanna Briggs Institute
LIME	Local Interpretable Model-agnostic Explanations

LLE	Local Lipschitz Estimator
LLM	Large Language Model
PDP	Partial Dependence Plot
PFI	Permutation Feature Importance
QAF	Quantitative Argumentation Framework
RF	Random Forest
SHAP	SHapley Additive exPlanations
SMILE	Statistical Model-agnostic Interpretability with Local Explanations
SMOTEENN	Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors
XAI	Explainable Artificial Intelligence
XGB	eXtreme Gradient Boosting

References

1. Minh, D.; Wang, H.; Li, Y.; Nguyen, T. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2022**, *55*, 3503–3568. [\[CrossRef\]](#)
2. Carvalho, D.; Pereira, E.; Cardoso, J. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [\[CrossRef\]](#)
3. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4793–4813. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**. [\[CrossRef\]](#)
5. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [\[CrossRef\]](#)
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. [\[CrossRef\]](#)
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. [\[CrossRef\]](#)
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. [\[CrossRef\]](#)
9. Richesson, R.; Horvath, M.; Rusincovitch, S. Clinical research informatics and electronic health record data. *Yearb. Med. Inform.* **2014**, *23*, 215–223. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Batko, K.; Ślzak, A. The use of big data analytics in healthcare. *J. Big Data* **2022**, *9*, 3. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Chi, H.; Wang, D.; Liao, B.; Cui, G.; Mao, F. An interpretable model based on concept and argumentation for tabular data. *Sci. Rep.* **2026**, *16*, 987. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Ribeiro, M.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [\[CrossRef\]](#)
13. Lundberg, S.; Lee, S. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 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 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–9 February 2020; pp. 180–186. [\[CrossRef\]](#)
15. Karagoz, G.; Ozcelebi, T.; Meratnia, N. Systematic benchmarking of local and global explainable AI methods for tabular healthcare data. In Proceedings of the xAI 2025: 3rd World Conference on eXplainable Artificial Intelligence, Lisbon, Portugal, 9–11 July 2025. [\[CrossRef\]](#)
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. [\[CrossRef\]](#) [\[PubMed\]](#)
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. [\[CrossRef\]](#) [\[PubMed\]](#)
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. [\[CrossRef\]](#)
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. [\[CrossRef\]](#) [\[PubMed\]](#)

20. Rai, T.; He, J.; Kaur, J.; Shen, Y.; Mahmud, M.; Brown, D.; O'Dowd, E.; Baldwin, D. Evaluating XAI techniques under class imbalance using CPRD data. *Front. Artif. Intell.* **2025**, *8*, 1682919. [[CrossRef](#)] [[PubMed](#)]
21. Rai, T.; He, J.; Kaur, J.; Shen, Y.; Mahmud, M.; Brown, D.J.; O'Dowd, E.; Baldwin, D.R. Evaluating LIME and SHAP Under Class Imbalance: A Lung Cancer Prediction Case Study Using CPRD Data. In Proceedings of the 4th International Workshop, XAI-Healthcare 2025, and 2nd International Workshop, PM4H 2025, Pavia, Italy, 26 June 2025; pp. 35–39. [[CrossRef](#)]
22. Garcia, M.; Martinez, L.; Sanchez, P. Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 152. [[CrossRef](#)] [[PubMed](#)]
23. El Shawi, R.; Jamel, L. Leveraging ChatGPT and explainable AI for enhancing clinical decision support. *Sci. Rep.* **2025**, *15*, 38786. [[CrossRef](#)] [[PubMed](#)]
24. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [[CrossRef](#)] [[PubMed](#)]
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, 9–11 July 2025. [[CrossRef](#)]
26. Ahmed, M.; Khan, M.; Hassan, R. Automated stress detection with explainable-AI transformer TabNet technique and physiological signals. *Biomed. Signal Process. Control* **2025**, *109*, 109208. [[CrossRef](#)]
27. Durgham, B.A.; Alawida, M. Explainable Ensemble Learning for Early Diagnosis of Alzheimer's Disease Using SHAP and FSHAP on ADNI Tabular Data. In Proceedings of the 2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications (AICCSA), Doha, Qatar, 19–22 October 2025; pp. 1–8. [[CrossRef](#)]
28. Thomas, J.; Brown, S. Explainable AI for Osteoporosis Risk Modeling: Random Forest and SHAP Interpretation on CTGAN-Enhanced Data. In Proceedings of the 2025 IEEE APCIT, Mysore, India, 19–20 September 2025; pp. 1–6. [[CrossRef](#)]
29. Abbas, S.; Soomro, N.; Khan, M. Attention-enabled Explainable AI for Bladder Cancer Recurrence Prediction. In Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2025), Copenhagen, Denmark, 14–17 July 2025; pp. 1–7. [[CrossRef](#)] [[PubMed](#)]
30. Vidhya, K.; Surya, S.K.; Geetha, S.K.; Thiyagu, T.M.; Catherine Joy, R.; Archpaul, J. TabNet Fusion Framework for Explainable Multi-Omics Alzheimer's Disease Classification. In Proceedings of the 2025 IEEE ICICI, Bangalore, India, 4–6 June 2025; pp. 1364–1369. [[CrossRef](#)]
31. Babu, G.M.; Durgabhavani, A.; Gudipudi, B.; Mouli, K.C.; Sasidhar, B.; Dattatreya, M. Interpretable Deep Learning for Pancreatic Cancer Diagnosis Using TabNet and SHapley Additive Explanations. In Proceedings of the 2025 IEEE ICECST, Petaling Jaya, Malaysia, 23–25 October 2025; pp. 1074–1079. [[CrossRef](#)]
32. Mow, M.H.; Ashrafuzzaman, M. An Interpretable Stroke Risk Prediction Model Using SMOTEENN and Advanced ML: An XAI-Driven Approach. In Proceedings of the 2025 IEEE ICCIT, Cox's Bazar, Bangladesh, 19–21 December 2025; pp. 1952–1957. [[CrossRef](#)]
33. Al 'Ghazi Wattiheluw, M.Y.; Jondri; Indwiarti. Mortality Risk Prediction in Heart Failure Patients Using Transfer Learning-Based TabNet and Explainable AI with SHAP. In Proceedings of the 2025 IEEE ICIMCIS, Jakarta, Indonesia, 3–4 December 2025; pp. 341–347. [[CrossRef](#)]
34. Rajani, B.; Varsha, K.; Srinivas, K.; Sivamani, N.; Ahmed, K.T. A Robust and Explainable Machine Learning Framework for Early Lung Cancer Risk Prediction using Synthetic Data Augmentation and Cross-Dataset Validation. In Proceedings of the 2025 IEEE ICSEDIS, Chiramanangad, India, 17–19 December 2025; pp. 859–864. [[CrossRef](#)]
35. Singh, S.; Rajput, S. Benchmarking Explainable AI Methods for Tabular Healthcare Data: Towards Standardized and Clinically Interpretable Evaluation Frameworks. In Proceedings of the 2025 IEEE PuneCon, Pune, India, 12–14 December 2025; pp. 1–7. [[CrossRef](#)]
36. Vinh, T.Q.; Byeon, H. Towards Transparent Diabetes Prediction: Unveiling the Factors with Explainable AI. *Int. J. Eng. Trends Technol.* **2024**, *72*, 26–35. [[CrossRef](#)]
37. Swain, J.; Sameer, K.; Udgata, S.K. Model-Agnostic Interpretability Techniques for Machine Learning-Based Pre-Term Birth Prediction. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2025; Volume 1624, pp. 337–348. [[CrossRef](#)]
38. Moola, S.; Munn, Z.; Tufanaru, C.; Aromataris, E.; Sears, K.; Sfetcu, R.; Currie, M.; Qureshi, R.; Mattis, P.; Lisy, K.; et al. Chapter 7: Systematic reviews of etiology and risk. In *JBI Manual for Evidence Synthesis*; JBI: North Adelaide, Australia, 2020. [[CrossRef](#)]
39. Joanna Briggs Institute. Checklist for Analytical Cross-Sectional Studies. 2020. Available online: <https://jbi.global/critical-appraisal-tools> (accessed on 1 July 2026).
40. Apley, D.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. Am. Stat. Assoc.* **2020**, *82*, 1059–1086. [[CrossRef](#)]

41. 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. [[CrossRef](#)] [[PubMed](#)]
42. 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, 9–11 July 2025. [[CrossRef](#)]
43. Raza, R.; Wang, G.; Laga, H.; Wong, K.W.; Nejd, W. TransLIME: Towards transfer explainability to explain black-box models on tabular datasets. *Inf. Sci.* **2026**, *730*, 122891. [[CrossRef](#)]
44. Ribeiro, M.; Singh, S.; Guestrin, C. Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [[CrossRef](#)]
45. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Leanpub: Victoria, BC, Canada, 2022.
46. 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. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.