Explainability in action: A metric-driven assessment of local explanations for healthcare tabular models

M. Atif Qureshi[1,2,3*], Abdul Aziz Noor[1,3], Awais Manzoor[1,3], Muhammad Deedahwar Mazhar Qureshi[2,3], Arjumand Younus[2,4], Wael Rashwan[5]

[1]ADAPT Centre, Ireland
[2]Research Ireland Centre for Research Training in Machine Learning (ML-Labs)
[3]eXplainable Analytics Group, Faculty of Business, Technological University Dublin
[4]School of Information and Communication Studies, University College Dublin
[5]Maynooth University, School of Business

*Corresponding author
Email:atif.qureshi@tudublin.ie (MAQ)

Abstract

Explainable AI (XAI) is essential in clinical machine learning, yet quantitative evaluation of explanation quality is rarely reported in a reproducible and comparable way. We address this gap with a reproducible, metric-driven evaluation framework for comparing XAI methods on healthcare tabular data, consolidating six established, family-specific quantitative metrics (fidelity, simplicity, consistency, robustness, precision, and coverage) into explicit equations, pairing them with a pre-specified focal-model selection protocol, and releasing open-source code together with a method–metric applicability map. All quantitative results are summarized as mean ± standard deviation over test instances to capture both average behavior and instance-level variability. Global summaries (e.g., aggregated SHAP importances, EBM main-effect shapes, or TabNet aggregated importances) are reported descriptively only. Using the framework, we evaluate five widely used approaches, LIME, SHAP, Anchors, EBM, and TabNet, across four healthcare tabular datasets spanning post-hoc feature attribution (LIME, SHAP), post-hoc rule extraction (Anchors), and inherently interpretable models (EBM, TabNet). For tree ensembles, we additionally report Random Forest global importances (Gini/MDI and permutation) as descriptive cross-checks alongside EBM/SHAP/TabNet global profiles. Empirically, SHAP (TreeSHAP) attains exact score fidelity (1.0) and near-perfect decision fidelity for tree ensembles; LIME yields simpler but less robust, lower-fidelity explanations with substantially higher instance-level variability in decision fidelity; TabNet most often produces the simplest explanations across thresholds but with high variance in some datasets; EBM and TabNet offer the most robust explanations under small perturbations; and Anchors returns high-precision, human-readable rules whose coverage decreases as precision thresholds tighten. LIME and SHAP show moderate-to-high agreement on salient features, and global profiles (reported descriptively) align with known risk factors. **Why this matters:** the framework enables apples-to-apples comparisons, reduces confounds, and turns narrative guidance into testable, quantitative practice, helping practitioners choose XAI methods by application priority (e.g., fidelity, robustness, rule precision/coverage). Although demonstrated in healthcare, the framework generalizes to other high-stakes tabular machine learning domains. Source code: `https://github.com/matifq/XAI_Tab_Health`.

Introduction

Machine learning (ML) continues to play a growing role in healthcare, supporting tasks such as diagnosis, prognosis, and treatment planning. However, the lack of transparency in many ML models remains a major barrier to clinical adoption [1]. In high-stakes domains like healthcare, explainability is essential not only for regulatory compliance and ethical deployment, but also for building trust among clinicians and patients [2].

Healthcare data are often tabular in nature, particularly in applications involving electronic health records (EHRs), clinical registries, and structured survey data. Despite the prevalence of tabular data in real-world clinical systems, much of the XAI literature has focused on image- and text-based modalities. While many studies advocate the use of XAI methods, comparatively few provide systematic empirical comparisons, and even fewer evaluate multiple explainers using clearly defined and practically meaningful quantitative metrics [1]. This gap limits the ability of practitioners and data scientists to make informed decisions about which XAI methods are appropriate for their specific use cases.

This article addresses this gap by benchmarking five widely used XAI methods, SHAP, LIME, Anchors, EBM, and TabNet, on four healthcare tabular datasets. The study is methodological in scope and does not aim to derive, validate, or recommend clinical decision rules; instead, it focuses on standardized quantitative evaluation of explainability methods. Based on formal definitions from the XAI literature, we operationalize and implement six established evaluation metrics: fidelity, simplicity, consistency, robustness, precision, and coverage. These metrics are applied systematically across methods and datasets to enable controlled comparison.

Study objective and scope. This study focuses on the methodological evaluation of explainable AI techniques for tabular healthcare models, rather than on addressing a specific clinical decision task or proposing new explainability methods. Specifically, it addresses a methodological gap in the XAI literature: the lack of a standardized, reproducible protocol for quantitatively evaluating and comparing explainability methods on tabular healthcare data. The novelty of the contribution lies at the framework level, through standardized operationalization of established metrics, explicit mapping of metrics to explainer families, control of model-induced confounds via a pre-specified focal-model protocol, and consistent application across multiple healthcare datasets with released code.

This work contributes to a growing body of benchmarking studies by explicitly aligning quantitative XAI evaluation with performance trade-offs commonly discussed in clinical tabular modeling contexts.

The contributions of this work are:

- A *standardized, metric-driven evaluation framework* for explainable AI in healthcare tabular data, enabling controlled and reproducible comparison across heterogeneous XAI method families.
- Explicit mathematical operationalization and open-source implementation of commonly cited XAI evaluation metrics, enabling auditable and repeatable benchmarking rather than narrative comparison.
- A pre-specified focal-model selection protocol that controls model-induced variance in post-hoc XAI evaluation, isolating explainer effects.
- A unified empirical study spanning post-hoc attribution, rule-based explainers, and inherently interpretable models across multiple public healthcare tabular datasets.

Positioning relative to prior work. In contrast to prior surveys, toolkits, and evaluation studies [3–8], this work (i) formalizes widely used XAI metrics using explicit equations and open-source implementations, (ii) pre-specifies a single post-hoc explainee to control model-induced variability, (iii) evaluates post-hoc and ante-hoc methods within a single protocol, and (iv) benchmarks performance across multiple clinical tabular datasets. A compact comparison is provided in Table 9 (Appendix).

Scope. This paper presents a reusable *framework* for applying auditable, quantitative XAI metrics to clinical tabular tasks, including metric definitions/equations, implementation details, and method–metric applicability. Exhaustive coverage of all explainer families and parameter regimes is left for future extensions.

Literature review

The role of explainability in clinical AI

XAI is increasingly vital in healthcare, where decisions must be transparent, auditable, and trustworthy. As ML supports diagnosis, prognosis, and treatment planning, interpretability[1] is essential for regulatory acceptance and practitioner trust [9, 10]. A wide spectrum of techniques exists, from instance-level feature attribution to global rule-based logic and visualization tools, aimed at rendering model behavior intelligible to clinicians [11–13]. XAI methods are often grouped into *post-hoc* and *ante-hoc* [14]. Post-hoc methods explain already-trained models; notable examples include SHAP (Shapley-value attributions), LIME (local surrogate explanations), and Anchors (precise local rules) [15–17]. Ante-hoc methods are inherently interpretable by design; examples include EBM (a GA^2M with learned interactions) and TabNet (an attentive neural architecture with per-sample feature selection) [18, 19].

XAI for tabular healthcare data

Despite their prevalence in clinical workflows (EHRs, registries, structured surveys), tabular datasets pose challenges such as missing values, class imbalance, and mixed data types [20]. Many healthcare XAI studies emphasize imaging and time series, or rely on proprietary data that limits reproducibility [1]. For tabular data specifically, comparative analyses include Hatwell et al. [21] (Ada-WHIPS, Anchors, LORE), Settouti et al. [22] (SHAP, LIME, permutation importance), and Alfeo et al. [23] (BoSCaR as a robust alternative to SHAP). Other works triangulate multiple explainers: Chadaga et al. [24] and Islam et al. [25] combine SHAP and LIME; Chun et al. [26], Khanna et al. [27], and Joseph et al. [28] cross-check attributions against known risk factors; Aiosa et al. [29] integrate SHAP with graphs; Peng et al. [30] pair SHAP with partial dependence. These examples illustrate growing use of XAI in clinical analytics, but most emphasize case-based interpretation rather than unified, metric-based comparisons across datasets.

Families. We distinguish (i) post-hoc feature-attribution methods, SHAP, which provides local Shapley values that can be aggregated to a global importance profile, and LIME, which is local by design; (ii) rule-based local explainers (Anchors); and (iii) inherently interpretable (ante-hoc) models (EBM, TabNet), both of which yield per-instance signals that can be aggregated to global views (e.g., EBM main-effect shapes; TabNet aggregated importances).

[1] or explainability

Ensembles in clinical tabular machine learning

A recurring result in the tabular-ML literature is that tree-based ensembles are strong baselines on low- to mid-sized tabular datasets. Comparative studies report that gradient boosting and random forests frequently match or outperform linear models, kernel methods, and recent deep architectures on such data regimes [31, 32]. Two properties are repeatedly cited as drivers of this performance: (i) flexible handling of heterogeneous feature types and non-linear interactions, and (ii) robustness to routine preprocessing choices such as standard imputation. These characteristics are well-suited to clinical records (EHRs, registries, structured surveys) with varying feature scales and missing data.

Within explainability research, this prevalence has practical implications. Many clinical tabular pipelines adopt a high-performing tree ensemble and then attach a post-hoc local explainer (e.g., SHAP, LIME, Anchors) for per-instance interpretability, while inherently interpretable models (e.g., EBM and TabNet) are used when a transparent global structure is required. The literature thus motivates studying both families on tabular healthcare tasks rather than focusing on a single paradigm [1].

Takeaway. This literature motivates our emphasis on tree ensembles as competitive tabular baselines and, by extension, on post-hoc explainers commonly paired with them in practice.

Quantitative XAI metrics

A recurring theme in the XAI literature is the call for *quantitative* criteria that move beyond narrative plausibility toward auditable evaluation [3, 4, 6, 33]. Several metrics have emerged as common currency, often defined conceptually in surveys and instantiated piecemeal in method papers or toolkits:

- **Fidelity.** Broadly, the agreement between an explanation and the behavior of the predictor being explained [4, 5]. Specific operationalizations include local accuracy for additive attributions (e.g., SHAP's additivity guarantees and TreeSHAP's exactness for tree ensembles) and surrogate–predictor agreement [15, 34].
- **Simplicity.** The cognitive compactness of the explanation, commonly proxied by sparsity (number of features/conditions) or length of the produced rationale [4, 35]. In rule learners, this often reduces to rule length; in attribution methods, thresholds on contribution magnitudes are frequently used.
- **Consistency (inter-explainer agreement).** Convergence across methods on what matters for the same instance/model, typically assessed via rank- or set-based concordance of salient features (e.g., correlations or overlap on top-k lists) [8].
- **Robustness.** Sensitivity of explanations to small, semantically irrelevant input changes; instantiated through input perturbations or sanity checks to ensure explanations reflect the model and data rather than artifacts [36, 37].
- **Precision and coverage for rule-based explanations.** For local if–then rationales (e.g., Anchors), empirical precision on covered instances and the proportion of instances covered are standard *rule-family evaluation metrics* [17].

While these metrics are widely recognized, their *operational forms* are scattered across individual papers and general-purpose libraries. Survey articles largely provide conceptual definitions rather than a unified, benchmark-ready set of equations [3, 4, 6, 33]. Libraries such as Quantus collect implementations and tests, but

are not tailored as end-to-end *clinical tabular* benchmarks with pre-specified explainee selection or a joint treatment of post-hoc and ante-hoc methods [7]. As a result, cross-paper comparisons often hinge on heterogeneous choices (datasets, predictors, thresholds, perturbation scales), making it difficult to assess trade-offs across methods.

Canonical definitions in prior work. Several metrics already have operational mathematics in the literature: (i) additive local accuracy for SHAP/TreeSHAP with exact reconstruction on tree ensembles [34]; (ii) local surrogate agreement for LIME (weighted regression fit to h in a locality kernel) [16]; (iii) Anchors' precision and coverage as rule-family metrics estimated against h [17]; (iv) rank-based agreement via Spearman/Kendall for top-k feature sets [8]; and (v) robustness via small input perturbations and explanation-distance sanity checks [36, 37].

Our instantiations vs. prior. Our contribution is not to introduce new metrics but to *standardize* one auditable instantiation per metric for clinical tabular settings and use them *side by side* in a single protocol: (1) *Fidelity* reported as both score- and decision-level; for SHAP/TreeSHAP we use exact additive reconstruction where applicable to avoid surrogate error. (2) *Simplicity* via a *relative* importance threshold (sparsity) so values are scale-free across methods/datasets. (3) *Consistency* as depth-averaged Spearman over top-k (we use k=5) to stabilize against ties and shallow lists. (4) *Robustness* as ℓ_1 change under small Gaussian perturbations with fixed σ to make comparisons reproducible. (5) *Precision/Coverage* restricted to rule-family (Anchors) to avoid cross-paradigm conflation. Where multiple forms coexist in prior work, we justify our single choice for reproducibility and cross-dataset comparability. Formal equations and the method–metric applicability appear in *Materials & Methods*.

Positioning relative to surveys and toolkits

Prior surveys synthesize concepts and taxonomies for explainability but stop short of a unified, auditable benchmarking recipe for clinical tabular settings [3, 4, 6, 33]. Method libraries such as Quantus assemble metric implementations and tests, yet they are not framed as end-to-end *benchmarks* with a pre-specified explainee selection protocol or a joint comparison of post-hoc and ante-hoc approaches on public clinical tabular data [7]. In contrast, we consolidate commonly cited metrics into explicit equations, release code, and apply them side-by-side across datasets within a single protocol that controls model-induced variance. A compact contrast with representative works appears in Table 9 (Appendix).

Study design overview

We adopt a simple, auditable protocol tailored to clinical tabular data. First, we benchmark five predictors (GBC, RF, XGB, EBM, TabNet) on four datasets and pre-specify a single-model rule: select the focal explainee by best average F1 rank across datasets (accuracy as tie-breaker). Second, we compare two families of explainability within one setup: post-hoc explainers (LIME, SHAP/TreeSHAP, Anchors) are applied only to the selected predictor, while inherently interpretable models (EBM, TabNet) are analyzed via their native interpretability[2]. Third, we assess explanation quality with six established metrics, fidelity, simplicity, consistency, robustness, precision, and coverage, formalized with explicit equations and released as open-source code. Method-specific

[2]We report SHAP both locally (per-instance) and as aggregated global importances (mean absolute SHAP values), whereas LIME is analyzed as a local explainer only. Anchors is local and rule-based. EBM and TabNet are evaluated via their inherent (ante-hoc) interpretability; TabNet also exposes per-sample feature masks (local), with optional aggregation to summarize global trends.

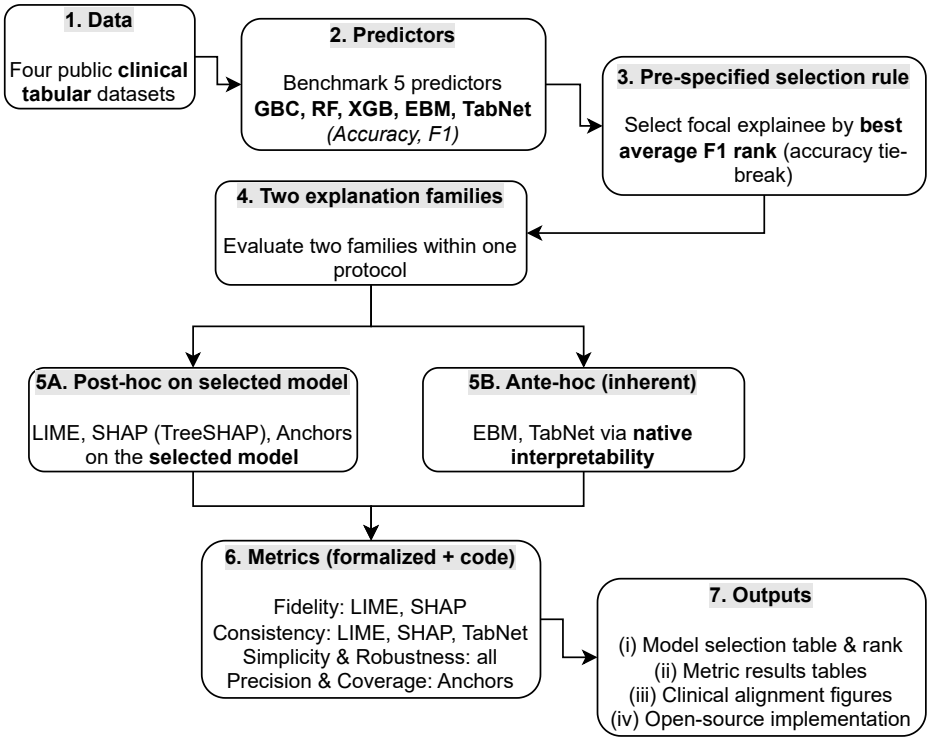


Fig 1. Study protocol at a glance. We benchmark five predictors on four clinical tabular datasets, pre-specify the focal explainee by best average F1 rank, apply post-hoc explainers (LIME, SHAP/TreeSHAP, Anchors) to that model, interpret ante-hoc models (EBM, TabNet) natively, and evaluate six metrics with method-specific applicability.

metric applicability is summarized in Table 1 (see *Materials and methods*); the overall protocol is illustrated in Fig. 1. We next detail datasets, models, and metric definitions in *Materials and methods*.

Materials and methods

Ethics statement

This study used publicly available, anonymized datasets and did not involve human subjects or personally identifiable data. No new data was collected, and all analyses prioritized research integrity and reproducibility. All datasets were accessed under their respective licences/terms of use.

Scope and rationale for method selection

We compare two complementary families of explainability on clinical tabular tasks: *local post-hoc* explainers applied to a fitted predictor (LIME, SHAP/TreeSHAP, Anchors) and *ante-hoc* models whose structure is itself interpretable (EBM, TabNet). These choices are guided by prior work [1, 4, 5] and by:

a) widespread adoption in healthcare ML pipelines for tabular data,

b) conceptual diversity (attribution, local rules, and glass-box models), and

c) coverage of established quantitative metrics (simplicity and robustness across families; fidelity for surrogate-style local attributions; consistency where ranked single-feature importances exist; precision and coverage for rule-based Anchors).

Concretely, we selected one representative per major explainer family commonly used in tabular healthcare pipelines (feature attribution, local rules) and paired them with two widely used interpretable-by-design baselines (EBM and TabNet) to cover both post-hoc and ante-hoc paradigms under a single protocol.

To avoid confounding explainer comparisons with model changes, all post-hoc methods are applied to a single, pre-specified focal predictor chosen by the best average F1 rank across datasets (accuracy as tie-breaker). Ante-hoc models are assessed via their native interpretability rather than as post-hoc targets. This design enables apples-to-apples evaluation where metrics are applicable; Table 1 makes the metric–method mapping explicit.

All quantitative evaluation metrics in this study are applied to *local* (per-instance) explanations; global aggregates (e.g., mean |SHAP|, EBM shape functions) are reported for context and are not included in metric scoring.

Global RF baselines. For the focal Random Forest, we also compute two standard global feature importances, permutation importance (PFI) and mean decrease in impurity (Gini/MDI). These provide conventional tree-model baselines to triangulate global rankings against SHAP/EBM/TabNet summaries (shown in the Appendix) and are not part of the local-metric scoring.

Alternatives and scope. The rule-based space includes other candidates (e.g., rule lists/sets). Interpretable modeling also spans linear/monotone GAM variants and decision sets. We do not claim exhaustiveness; rather, we provide a reproducible, metric-driven baseline with widely used representatives and release code to ease extension. Counterfactuals and example-based methods are orthogonal families we list for future work (see *Limitations*).

Datasets

Four publicly available healthcare tabular datasets were selected using the following inclusion criteria: each was publicly accessible, non-synthetic, suitable for binary classification, and either contained no missing values or had missing values already imputed in the publicly released version (i.e., no additional imputation by us).

1. **PDHS (C-section)**: Contains 26,284 rows and 24 columns. Derived from national health surveys; the publicly released version includes imputed values for originally missing entries. Used to predict Caesarean-section outcomes [25].

2. **Diabetes retinopathy**: Includes 1,151 rows and 19 columns. Used to assess the risk of diabetes retinopathy progression [21].

3. **Breast cancer**: Comprises 569 rows and 30 columns. Applied for binary classification of breast cancer diagnoses [21].

4. **ESDRPD (Early Stage Diabetes Prediction)**: Contains 520 rows and 16 columns. Targets the early detection of diabetes-related complications [28].

All tasks are binary classification using the dataset's default target variable.

Machine learning models

We evaluated five machine learning models: Gradient Boosting Classifier (GBC), Random Forest (RF), Extreme Gradient Boosting (XGB), TabNet, and Explainable Boosting Machines (EBM). Model performance was assessed using accuracy and F1 score to establish a predictive baseline across different methodological families.

Our emphasis on tree-based ensembles (RF, GBC, XGB) reflects both the structure of the datasets and converging evidence from the literature. All four datasets used in this study are low- to mid-sized tabular healthcare datasets (ranging from 520 to 26,284 rows, with mixed categorical and numerical features). For such data regimes, ensemble-based tree methods have repeatedly been shown to outperform linear models, kernel methods, and even recent deep learning architectures for tabular data [31, 32]. They handle heterogeneous predictors, capture non-linear interactions, and work well with standard imputation, properties that are particularly relevant to clinical data. The recent survey of explainable AI in healthcare similarly found that tree ensembles are among the most frequently reported "winner" models in comparative studies [1]. Evidence from other domains aligns with this conclusion, e.g., systematic churn-prediction reviews where ensembles dominate in tabular data settings [38].

To complement these ensembles, we also included two inherently interpretable models: EBM and TabNet. EBM is a generalized additive model with automatic interaction detection, inspired by boosting methods but designed as a glass-box approach in which feature effects and interactions can be inspected directly without post-hoc explanations [39]. TabNet is a deep learning architecture specifically developed for tabular datasets; it employs sequential attention masks to select salient features for each sample, thereby providing feature-level interpretability alongside competitive predictive performance [19]. Including these models ensured that both inherently interpretable and neural architectures were represented in the evaluation, not only ensemble-based methods.

XAI methods and evaluation metrics

Concrete implementation details, software packages, and environment specifications are summarized in the *Experimental setup* to support exact reproducibility.

Methods policy. We compute all metrics on *local* (per-instance) outputs with fixed parameters and seeds; apply each metric only where it is theoretically meaningful for the explainer family; and run all post-hoc explainers on a single, pre-specified focal predictor to ensure apples-to-apples comparability across datasets.

We evaluated five XAI methods spanning post-hoc and inherently interpretable approaches: LIME, SHAP[3], Anchors, EBM, and TabNet. To isolate differences attributable to explainers, all post-hoc methods (LIME, SHAP, Anchors) are applied to a single focal predictor chosen after benchmarking in the *Results* section. In contrast, EBM and TabNet are assessed via their inherent (ante-hoc) interpretability.

We use six established quantitative evaluation metrics from prior work: *fidelity*, *simplicity*, *consistency*, *robustness*, *precision*, and *coverage*. Although the metrics are not new, we formalize and implement their mathematical definitions for transparency and reproducibility [4, 5, 8]. Fidelity applies to post-hoc feature-attribution surrogates (LIME, SHAP); precision and coverage apply to rule-based explainers (Anchors); simplicity and robustness apply to all methods; and consistency is reported for methods

[3] TreeSHAP for tree ensembles; additive reconstruction is exact [34]. For non-tree models, KernelSHAP yields an additive approximation; the same fidelity equations apply, but we do not evaluate that variant here.

that yield ranked single-feature importances[4] (LIME, SHAP, TabNet). Table 1 summarizes this mapping.

Table 1. Applicability of Evaluation Metrics to XAI Methods. All quantitative metrics are computed on *local* (per-instance) explanations; global summaries (e.g., aggregated SHAP importances, EBM main-effect plots, TabNet aggregated importances) are reported descriptively only. Metrics are applied only where theoretically meaningful: fidelity for surrogate-style local attributions (LIME; SHAP via TreeSHAP on trees), precision/coverage for rule learners (Anchors), rank-based consistency for methods producing single-feature rankings (LIME, SHAP, TabNet), and simplicity/robustness across families.

Evaluation Metric	LIME	SHAP	Anchors	EBM	TabNet
Fidelity	✓	✓	-	-	-
Simplicity	✓	✓	✓	✓	✓
Consistency	✓	✓	-	-	✓
Robustness	✓	✓	✓	✓	✓
Precision	-	-	✓	-	-
Coverage	-	-	✓	-	-

Parameter choices and sensitivity. Several evaluation metrics require operational parameters, including the simplicity threshold τ, the top-k depth used for rank-based consistency, and the perturbation scale σ for robustness analysis. Rather than tuning these parameters to optimize results on a specific dataset, we adopt representative and conservative settings that are widely used in empirical XAI evaluation to characterize explanation behavior under standard operating conditions.

Specifically, simplicity thresholds $\tau \in \{0.10, 0.05, 0.01\}$ are used to examine explanation sparsity across coarse-to-fine attribution cutoffs. This follows prior work that evaluates simplicity using relative importance thresholds and reports sparsity trends across multiple operating points, reflecting the absence of a canonical sparsity cutoff and the context-dependent nature of explanation compactness [35, 42].

For rank-based consistency, we fix $k = 5$ and compute agreement over the top-ranked features only, focusing on the most salient elements of each explanation. This choice aligns with comparative XAI studies that restrict consistency analysis to a small number of dominant features to balance interpretability and stability, and to avoid noise introduced by low-importance feature tails [5, 8, 15, 16].

For robustness, we use a small perturbation scale $\sigma = 0.01$, introducing element-wise Gaussian noise as a controlled probe of explanation sensitivity to minor numerical variation (e.g., rounding or measurement noise), rather than as a model of semantically meaningful or clinically relevant change. This formulation is consistent with prior robustness and sanity-check analyses in explainable AI, where small-magnitude Gaussian jitter is used to assess local stability of explanations [7, 37, 43].

Partial sensitivity analysis is reflected in the reported results: simplicity is evaluated across multiple τ values, and Anchors are assessed across multiple precision thresholds. A more exhaustive sensitivity study over $\{k, \sigma\}$ and alternative perturbation models would be valuable but is beyond the scope of the present work and is identified as future research (see *Limitations and Future Work*).

[4] EBM (a GA^{2}M) provides per-feature shape functions and optional *pairwise* interactions [40]; our rank-based consistency metric is defined for single-feature lists. To avoid conflating main effects with interactions [41], we do not compute consistency for EBM, leaving interaction-aware consistency for future work.

Metric definitions and relation to prior work. Our fidelity follows the standard *agreement-based* notion in the XAI literature, i.e., how closely an explainer or surrogate reproduces the output of the predictive model for the same input [4, 5]: for LIME we compare surrogate outputs to the predictor; for SHAP we reconstruct the predictor's output from the additive explanation (local accuracy), which is exact for tree ensembles under TreeSHAP [34]. Our simplicity counts features above a relative-importance threshold [35]. Consistency uses Spearman correlation over top-k ranked features (agreement across explainers) [8]. Robustness measures the sensitivity of local explanations to small input perturbations [36]. For Anchors, we report rule precision and coverage as in the original work [17]. Our contribution is to operationalize these notions with unified equations and open-source code for comparative, multi-dataset evaluation in clinical tabular settings.

We now present metric definitions for evaluating the two classes of explainers: feature attribution–based (e.g., SHAP, LIME) and rule-based (e.g., Anchors). All metrics are defined over a dataset $X = \{x_1, x_2, \ldots, x_N\}$ of N instances, unless stated otherwise.

Common notation. Let $X = \{x_i\}_{i=1}^{N}$ be the dataset, F the set of features (with $|F|$ elements), h the trained predictor, and M_h an explainer instantiated with respect to h. We write $P_h(x)$ for the predictor's score and $P_{M_h}(x)$ for the explainer's surrogate/reconstruction score when defined (e.g., LIME, TreeSHAP). For local explanations on x_i we use $E(M_h, x_i)$ to denote the explanation object: (i) for attribution methods, $E(M_h, x_i) = f_i \in \mathbb{R}^{|F|}$ with entries f_{ij}; (ii) for rule-based methods, $E(M_h, x_i) = A_i \subseteq F$ is the set of feature conditions (literals) in the rule.

Evaluating feature importance-based XAI methods

We formalize established quantitative metrics for attribution-based explainers below.

Fidelity: Fidelity [44] evaluates how accurately an XAI method replicates the behavior of the original predictive model. It is computed in two complementary ways:

- **Score fidelity:** Measures the similarity between the model's probability and the explainer's reconstruction, operationalized as $1 - \left|P_h(x_i) - P_{M_h}(x_i)\right|$.
- **Decision fidelity:** Measures the agreement in final class predictions. It reflects how often the XAI method's predicted label matches the label of the original model.

The formal equations are:

$$\mathrm{Fid}_{\mathrm{score}}(M_h) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \left|P_h(x_i) - P_{M_h}(x_i)\right|\right) \tag{1}$$

$$\mathrm{Fid}_{\mathrm{decision}}(M_h) = \frac{1}{N}\sum_{i=1}^{N}\delta\big(\hat{y}_h(x_i),\, \hat{y}_{M_h}(x_i)\big) \tag{2}$$

Where:

- $P_h(x_i)$ is the probability (score) from the trained predictor h at x_i.
- $P_{M_h}(x_i)$ is the surrogate/reconstruction probability produced by the explainer M_h.

- $\hat{y}_h(x_i)$ and $\hat{y}_{M_h}(x_i)$ are the corresponding predicted class labels from h and M_h.
- $\delta(a, b)$ is the Kronecker delta, returning 1 if $a = b$ and 0 otherwise.

This dual view of fidelity allows both probabilistic closeness and class decision alignment to be assessed, highlighting whether an explanation preserves not just the model's outcome but its internal confidence as well.

A perfect score of 1.0 indicates that the explainer exactly replicates the model's predicted probability (score fidelity) or predicted label (decision fidelity); small deviations can arise from ties or probabilistic rounding. Scores closer to 0.0 signify poor alignment between the explainer and the model.

Simplicity (attribution-based): Simplicity [35] quantifies how concise and interpretable an explanation is, based on the number of features it includes. In high-stakes domains like healthcare, simpler explanations involving fewer features are often preferred to support user trust and rapid decision-making.

For attribution methods, we define simplicity as the *average number of features per instance* whose relative importance exceeds a threshold $\tau \in \{0.10, 0.05, 0.01\}$. Let $f_i = [f_{i1}, \ldots, f_{i|F|}]$ be the attribution vector for instance x_i. Features with $\frac{|f_{ij}|}{\max_{k \in \{1,\ldots,|F|\}} |f_{ik}|} > \tau$ are counted toward the explanation.

$$\mathrm{S}_{\mathrm{attr}}(M_h; \tau) \;=\; \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{|F|} \mathbb{I}\left(\frac{|f_{ij}|}{\max_{k \in \{1,\ldots,|F|\}} |f_{ik}|} > \tau \right) \tag{3}$$

Where:

- F: set of features (with cardinality $|F|$).
- f_{ij}: attribution score of feature j for instance x_i.
- $\max_{k \in \{1,\ldots,|F|\}} |f_{ik}|$: maximum absolute attribution in instance x_i.
- τ is the threshold value $\in \{0.10, 0.05, 0.01\}$.
- $\mathbb{I}(\cdot)$: indicator function, returning 1 if the condition is true, 0 otherwise.

The threshold parameter controls explanation sparsity. Higher thresholds (e.g., 0.1) retain only the most influential features, resulting in simpler, more focused explanations. Lower thresholds (e.g., 0.01) capture weaker attributions, producing denser explanations with more features.

This formulation enables consistent comparison of explanation conciseness across methods and datasets. Lower values indicate simpler explanations involving fewer features, which are generally easier for human interpretation. Higher values may capture more detail but reduce interpretability.

Consistency: Consistency [8] measures the agreement between different XAI methods in identifying influential features for the same model predictions. In clinical settings, consistency across explainers can build confidence and validate the reliability of explanations.

We compute pairwise consistency using Spearman's rank correlation across progressive feature depths. Specifically, for each instance, we compare the top n ranked features from two XAI methods for $n = 1$ to k, where k is the number of top-ranked features considered. In this study, we used $k = 5$ to capture the most influential part of the explanation while maintaining computational tractability. This stepwise comparison allows us to quantify how consistently two methods agree as more features are included.

The consistency score between two XAI methods $M_{1,h}$ and $M_{2,h}$ is defined as:

$$\mathrm{C}(M_{1,h}, M_{2,h}) \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{k}\sum_{n=1}^{k}\mathrm{Spearmanr}\big(\text{top-}n(M_{1,h}, x_i),\ \text{top-}n(M_{2,h}, x_i)\big)\right) \tag{4}$$

Where:

- top-$n(M_{1,h}, x_i)$, top-$n(M_{2,h}, x_i)$: the top-n ranked features for instance x_i from methods $M_{1,h}$ and $M_{2,h}$.
- $\mathrm{Spearmanr}(\cdot,\cdot)$: Spearman's rank correlation.
- k: the number of top-ranked features considered (we use $k = 5$).

This metric enables pairwise comparisons of local explanations, accounting for both rank order and alignment in feature attribution. The consistency score $\mathrm{C}(M_{1,h}, M_{2,h})$ lies in the range $[-1, 1]$, where 1.0 indicates perfect agreement across all top-k feature depths, 0 indicates no correlation, and negative values reflect inverse ranking alignment between the two XAI methods.

Robustness (attribution-based): Robustness [36] evaluates how stable a feature-based explanation remains when small input perturbations are introduced. This is essential in clinical workflows where slight variations (e.g., sensor noise or rounding) may occur.

We introduce Gaussian noise to each instance, drawn from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.01$, applied element-wise to all numerically encoded inputs (including integer-coded fields), with no model refitting, to probe small measurement/recording variation rather than clinically significant shifts. Robustness $\mathrm{R}_{\mathrm{attr}}(M_h)$ is computed as the average ℓ_1 change between explanation vectors before and after a small perturbation[5]:

$$\mathrm{R}_{\mathrm{attr}}(M_h) \;=\; \frac{1}{N}\sum_{i=1}^{N}\left\|E(M_h, x_i) - E(M_h, x_i + \epsilon)\right\|_1 \tag{5}$$

Where:

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$: element-wise perturbation.
- $E(M_h, x_i)$: explanation vector from method M_h for input x_i.

Lower values indicate higher robustness, i.e., explanations remain stable even under small input perturbations. A value of 0 means the explanation is entirely unchanged. This perturbation model is intended as a controlled sensitivity probe for explanation stability under minimal input variation, rather than as a simulation of clinically meaningful patient changes.

Transition to rule-based explainers: The metrics discussed thus far apply to explainers that yield *numeric, per-instance feature-importance vectors*. In contrast, rule-based explainers such as Anchors produce logical decision rules, necessitating a distinct set of metrics tailored to symbolic conditions and coverage-based evaluation. We now describe four such metrics used to evaluate rule-based explanations.

[5] We use the ℓ_1 norm; results are qualitatively similar with ℓ_2. We do not divide by $|F|$ so robustness reflects the *total* attribution shift per instance.

Evaluating rule-based XAI methods

Notation reminder (rule-based). In this subsection $E(M_h, x_i) = A_i \subseteq F$ denotes the set of feature conditions (literals) in the rule for x_i; primes (e.g., A_i') indicate the rule after a small perturbation of x_i.

Rule-based XAI methods, such as Anchors, generate explanations in the form of logical if-then rules composed of discrete feature conditions (e.g., "$Age > 40$ AND $Gender = Female$"). These rules differ fundamentally from numeric feature attributions and thus require tailored evaluation metrics. We define four such metrics: **precision**, **coverage**, **simplicity**, and **robustness**, detailed below.

Precision and coverage: *Precision* measures how often the black-box predictor agrees with the explained-instance prediction on the samples covered by the rule, and *coverage* is the fraction of instances to which the rule applies [17]. Tightening the precision threshold typically increases reliability while reducing coverage (the canonical precision–coverage trade-off).

Let $\hat{y}_h(x_i)$ be the prediction of h on the explained instance x_i, and let $\mathbb{I}[\cdot]$ be the indicator. Then

$$\text{Precision}(A_i) \;=\; \frac{1}{|\mathcal{C}(A_i)|} \sum_{z \in \mathcal{C}(A_i)} \mathbb{I}[\hat{y}_h(z) = \hat{y}_h(x_i)]\,, \tag{6}$$

$$\text{Coverage}(A_i) \;=\; \frac{|\mathcal{C}(A_i)|}{N}, \tag{7}$$

Where:

- z denotes a generic instance in the dataset X.
- $\mathcal{C}(A_i) = \{z \in X : A_i \text{ matches } z\}$.
- N is the dataset size used for coverage.

These metrics are computed using the Anchor package's inbuilt estimators and reflect how reliably and broadly each generated rule captures the model's behavior.

A perfect precision score is 1.0 (i.e., 100% correct on all covered instances). Coverage values range from 0 to 1, where 1 indicates full dataset applicability.

Higher thresholds (e.g., 0.99) enforce stricter rule precision, producing explanations with near-perfect reliability. However, this typically reduces coverage, as the rules apply to fewer data instances. This trade-off is particularly relevant in healthcare: while high-precision rules improve trust in local predictions, their narrow applicability must be weighed against the need for generalizability across diverse patient groups.

Simplicity (rule length): Simplicity equals rule length (number of literals)[6].

$$\text{S}_{\text{rule}}(M_h) \;=\; \frac{1}{N} \sum_{i=1}^{N} \big|\text{Features}(M_h, x_i)\big| \tag{8}$$

Where:

- $\text{Features}(M_h, x_i)$: the set of features used in the rule that explains x_i.

A smaller simplicity score may be desirable in clinical settings to support quick comprehension and decision-making.

[6] E.g., "$Age > 45$ AND $BMI > 30$" has length 2; adding "$Smoker = Yes$" gives length 3.

Robustness (rule-based): Robustness for rule-based explanations reflects the stability of the rule features under small perturbations. If different rules are generated from slightly altered inputs, it indicates lower robustness. Similar to the perturbation used for feature-attribution methods, we introduce small Gaussian noise to the input instance, drawn from a normal distribution $\mathcal{N}(0, \sigma^2)$, with $\sigma = 0.01$[7].

Following the same perturbation approach as attribution methods [36], we measure robustness via the Jaccard dissimilarity[8] between pre- and post-noise rule feature sets, as defined in Eq. (9).

Let A_i be the set of features used in the rule for instance x_i, and A'_i the rule after applying small Gaussian noise to the input x_i. Then:

$$\mathrm{R}_{\mathrm{rule}}(M_h) \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|A_i \cap A'_i|}{|A_i \cup A'_i|}\right) \tag{9}$$

Where:

- A_i is the set of feature conditions in the rule for x_i, and A'_i is the corresponding set after a small perturbation is applied to x_i (i.e, $x_i + \epsilon$).

Lower scores indicate higher robustness, i.e., greater stability in the rule structure when inputs are slightly perturbed. A robustness score of 0.0 means the rule remains identical under perturbation.

Framework components and extensibility. The framework comprises: (i) explicit equations for six established metrics; (ii) a pre-specified focal-model protocol to control model-induced variance; (iii) a metric–method applicability map; and (iv) open-source code. New explainers or metrics can be added by implementing the metric interface (inputs/outputs documented in the repository) and extending the applicability map; the protocol remains unchanged.

Experimental setup

Each dataset was partitioned into training and testing subsets by preserving original splits where available or applying a standard 70:30 split otherwise. The training sets were used to fit all five models (GBC, RF, XGB, TabNet, EBM). Predictive performance was evaluated using accuracy and F1 score on the test sets.

To control for model-induced variability in the post-hoc explainability study, we pre-specified a single-model protocol: after benchmarking, we select one tree-ensemble as the focal explainee using an aggregate performance rule. The rule chooses the model with the best *average rank across datasets* on F1 (primary) and accuracy (tie-breaker); if models are tied, we prefer the one with native support for exact local attributions to minimize explainer-approximation error. Post-hoc explainers (LIME, SHAP, Anchors) are then computed on the selected focal model, while EBM and TabNet are examined via their inherent interpretability.

Why a single explainee? Fixing one competitive predictor for post-hoc comparisons controls model-induced variance so that differences are attributable to the explainer rather than to changes in the underlying decision function [7, 37].

[7] i.e., applied element-wise to numerically encoded inputs (including integer-coded fields), with no refitting, before recomputing rules.

[8] We use Jaccard dissimilarity on rule feature sets in Eq. (9); other set distances (e.g., Hamming on binary literal vectors) produce qualitatively similar trends.

Hyperparameter policy and reproducibility. Because our primary goal is to benchmark *explainability*, not to maximize predictive accuracy, we adopt library defaults with a fixed global seed and avoid extensive per-dataset hyperparameter optimization. This standardization reduces model-induced variability so that observed differences are attributable to the explainers themselves, in line with recent guidance on rigorous and transparent XAI evaluation and reproducible ML workflows [7, 45, 46]. Moreover, explanation quality is known to be sensitive to modeling and procedural choices, which motivates controlling additional sources of variance introduced by heavy hyperparameter optimization [37]. For Random Forests specifically, prior work shows diminishing returns once the number of trees is sufficiently large and argues that treating the tree count as a major tuning parameter is often unnecessary; performance typically plateaus around ∼100–200 trees [47, 48]. Accordingly, we use defaults with a fixed seed for all models and a conventional tree count for RF.[9]

Software environment and packages. All experiments were conducted in Python 3.11.7. Explainability methods were implemented using established open-source libraries. Specifically, LIME explanations were generated using the `lime`, SHAP values were computed using the `shap`, rule-based explanations were generated using the `anchor-exp`, EBM were implemented via `interpret`, and TabNet models were trained using the `pytorch-tabnet` package. The full software environment, including all Python dependencies, was captured using `pip freeze` and is released as a `requirements.txt` file alongside the source code to support exact reproducibility.

Results

Classification performance

We benchmarked five classifiers (GBC, RF, XGB, TabNet, EBM) across the four healthcare datasets. Table 2 reports the accuracy and F1 scores for each model–dataset pair.

Role of this benchmark. Table 2 is the input to our pre-specified focal-model protocol: we report per-dataset performance so that ranks and the overall *average F1 rank* can be computed transparently.

Model selection rule (pre-specified). As outlined in the *Experimental setup*, the single focal explainee for post-hoc XAI is selected by *best average rank on F1 across datasets* (primary), with *average rank on accuracy* as a tie-breaker if needed.
Selection outcome. Applying the rule yields "Random Forest (RF)" as the focal post-hoc explainee (see Table 3): RF ranks first on two datasets (PDHS, ESDRPD), second on Breast cancer, and fifth on Diabetes retinopathy, giving the best average F1 rank (2.25). The accuracy tie-breaker was not required.

Implication for XAI. Having selected RF by a pre-specified, data-driven rule, we compute post-hoc explanations (SHAP, LIME, Anchors) on RF to control for model-induced variability. EBM and TabNet are reported via their inherent interpretability.

[9] Settings used in all experiments: `random_state`= 42 for all models; RF: `n_estimators`= 100, `n_jobs`= −1; GBC: scikit-learn defaults; XGB: library defaults; EBM: library defaults; TabNet: `seed`= 42, `verbose`= 0. Full configuration files and code are available.

Table 2. Classification performance (Accuracy, F1) across datasets; best per dataset in **bold** (primary: F1; tie-breaker: Accuracy).

Dataset	Model	Accuracy	F1 Score
PDHS (C-section)	GBC	0.9320	0.9276
	RF	**0.9849**	**0.9847**
	XGB	0.9817	0.9816
	TabNet	0.9507	0.9510
	EBM	0.9549	0.9539
Diabetes retinopathy	GBC	**0.7139**	**0.7273**
	RF	0.6792	0.6819
	XGB	0.6965	0.7075
	TabNet	0.5434	0.7019
	EBM	0.6908	0.7068
Breast cancer	GBC	0.9591	0.9677
	RF	0.9708	0.9772
	XGB	0.9708	0.9767
	TabNet	0.4327	0.5076
	EBM	**0.9766**	**0.9817**
ESDRPD	GBC	0.9872	0.9901
	RF	**0.9936**	**0.9951**
	XGB	0.9808	0.9851
	TabNet	0.3974	0.1754
	EBM	0.9744	0.9804

Table 3. Per-dataset F1 ranks (1 = best) and average rank used for model selection.

Model	PDHS (C-section)	Diabetes retinopathy	Breast cancer	ESDRPD	Avg. rank (F1)
RF	**1**	5	2	**1**	**2.25**
XGB	2	2	3	3	2.50
EBM	3	3	**1**	4	2.75
GBC	5	**1**	4	2	3.00
TabNet	4	4	5	5	4.50

Evaluation of XAI methods aggregated over local explanations

The effectiveness of XAI methods was assessed using predefined metrics: fidelity, simplicity, consistency, robustness, precision, and coverage. The results, averaged over all local instance explanations, provide insights into the performance and suitability of each method across the different datasets.

Evaluation of feature importance-based XAI methods

Fidelity: The fidelity scores for LIME and SHAP are reported in Table 4. In this context, fidelity is assessed using two complementary criteria: *Score* fidelity, which measures agreement in predicted probabilities, and *Decision* fidelity, which measures agreement in predicted class labels.

Table 4. Fidelity of LIME and SHAP explanations, reported as mean ± standard deviation over test instances. Best-performing values are highlighted with †. Standard deviations are not shown when they round to 0.000.

Dataset	Score Fidelity		Decision Fidelity	
	LIME	SHAP	LIME	SHAP
PDHS (C-section)	0.890 ± 0.152	1.000 †	0.932 ± 0.251	1.000 ± 0.016 †
Diabetes retinopathy	0.832 ± 0.118	1.000 †	0.538 ± 0.499	0.994 ± 0.076 †
Breast cancer	0.883 ± 0.076	1.000 †	0.983 ± 0.131	1.000 †
ESDRPD	0.791 ± 0.145	1.000 †	0.917 ± 0.276	1.000 †

SHAP, using TreeSHAP, achieved perfect score fidelity in approximating probability outputs across all datasets with zero variance confirming its theoretical assurance for tree-based models [15, 34]. Minor deviations in decision fidelity can occur due to thresholding/rounding effects when converting probabilities to class labels, despite near-identical probability reconstructions (e.g., 0.994 ± 0.076 for Diabetes retinopathy). These deviations reflect threshold artifacts[10] rather than genuine failures of explanation fidelity.

Conversely, LIME consistently exhibited lower fidelity and substantially higher variability across instances. For example, in the Diabetes retinopathy dataset, LIME achieved a score fidelity of 0.832 ± 0.118 and a markedly lower decision fidelity of 0.538 ± 0.499, indicating unstable local approximations at the instance level. Comparable patterns are observed in ESDRPD (0.791 ± 0.145 for score fidelity and 0.917 ± 0.276 for decision fidelity), highlighting LIME's sensitivity to complex or non-linear decision boundaries. These findings underscore that LIME's linear approximation of decision boundaries often sacrifices an accurate reflection of model behavior.

Simplicity: The simplicity of LIME, SHAP, EBM, and TabNet is measured by the number of features involved in the explanations, as reported in Table 5. Results are summarized using mean ± standard deviation over test instances to capture both the typical explanation size and its variability across instances.

The thresholds (0.1, 0.05, and 0.01) represent different levels of feature-importance cutoffs used to determine which features are included in an explanation. Specifically, the threshold indicates a fraction of the maximum absolute feature-attribution value in a given instance. A feature is included in the explanation if its absolute value exceeds the threshold multiplied by the maximum absolute value in that instance. Higher thresholds therefore result in fewer features and simpler explanations, while lower thresholds admit more features and yield more detailed explanations. Because the

[10] attributable to tie-breaking effects when predicted class probabilities are equal

Table 5. Simplicity of XAI methods, measured as the number of features in the explanation, reported as mean ± standard deviation over test instances. Best-performing (lowest) values are highlighted with †. Standard deviations are not shown when they round to 0.000.

Method	PDHS (C-section)	Diabetes retinopathy	Breast cancer	ESDRPD
Threshold = 0.1				
LIME	$1.82 \pm 0.39^{\dagger}$	9.86 ± 0.42	9.93 ± 0.37	9.95 ± 0.22
SHAP	3.86 ± 2.82	11.44 ± 2.52	13.86 ± 2.44	$8.93 \pm 1.99^{\dagger}$
EBM	7.68 ± 3.93	20.11 ± 5.95	28.40 ± 5.37	20.53 ± 2.12
TabNet	4.31 ± 1.38	$6.10 \pm 7.07^{\dagger}$	$6.21 \pm 9.56^{\dagger}$	11.53 ± 6.46
Threshold = 0.05				
LIME	$3.27 \pm 0.79^{\dagger}$	10.00 ± 0.05	10.000	10.000 †
SHAP	8.59 ± 3.37	14.23 ± 2.06	17.42 ± 2.83	12.28 ± 1.46
EBM	16.26 ± 7.08	27.16 ± 4.55	43.55 ± 5.20	27.74 ± 1.92
TabNet	4.85 ± 1.39	$7.34 \pm 7.53^{\dagger}$	$6.69 \pm 9.73^{\dagger}$	12.33 ± 5.74
Threshold = 0.01				
LIME	10.000	10.000	10.000	10.000 †
SHAP	18.09 ± 1.93	17.10 ± 1.14	25.46 ± 2.34	15.37 ± 0.75
EBM	35.96 ± 4.83	35.63 ± 1.79	56.08 ± 1.56	30.62 ± 0.98
TabNet	$5.46 \pm 1.46^{\dagger}$	$9.88 \pm 7.64^{\dagger}$	$7.37 \pm 9.84^{\dagger}$	13.49 ± 4.13

maximum absolute feature value varies across datasets and instances[11], the resulting explanation sizes also vary, as formalized in Eq. (3).

Across datasets and thresholds, TabNet frequently produced the simplest explanations, particularly at lower thresholds (0.05 and 0.01), using the fewest dominant features in most cases. LIME achieved the simplest outputs in only a few cases and ranked second overall in simplicity. At lower thresholds, it tended to saturate, often converging to a fixed number of selected features with near-zero deviation. SHAP explanations were consistently denser, while EBM generated the most feature-rich explanations across all thresholds.

The reported standard deviations highlight important differences in instance-level behavior. TabNet and SHAP exhibit higher variability in explanation size, reflecting adaptive feature selection that can change substantially across instances, whereas LIME often produces explanations of consistent size once the threshold constraint becomes binding. Overall, these results reinforce an interpretability–complexity trade-off: TabNet and LIME favor brevity, while SHAP and EBM provide more detailed but potentially less interpretable explanations.

Consistency: The consistency scores between LIME, SHAP, and TabNet are reported in Table 6. Results are summarized using mean ± standard deviation over test instances to capture both average agreement and instance-level variability in feature rankings.

LIME and SHAP exhibited moderate to high agreement in feature rankings across all datasets, indicating overlapping explanation patterns. These consistency scores were computed using a stepwise comparison over the top $k = 5$ ranked features per instance (see Eq. (4)), which captures agreement not only at the top rank but progressively across the most influential features, weighting higher ranks more strongly.

The reported standard deviations indicate that while the strength of agreement varies across instances, the LIME–SHAP relationship remains consistently positive. In contrast, both LIME and SHAP showed consistently low or near-zero correlations with

[11] i.e., practitioners assess which threshold appears most suitable based on the dataset

Table 6. Consistency between XAI methods measured using Spearman rank correlation, reported as mean ± standard deviation over test instances. Best-performing (highest) values are highlighted with †.

Dataset	LIME–SHAP	LIME–TabNet	TabNet–SHAP
PDHS (C-section)	$0.73 \pm 0.11^{\dagger}$	-0.06 ± 0.41	-0.04 ± 0.42
Diabetes retinopathy	$0.51 \pm 0.27^{\dagger}$	-0.04 ± 0.19	-0.08 ± 0.21
Breast cancer	$0.58 \pm 0.21^{\dagger}$	-0.05 ± 0.22	-0.06 ± 0.20
ESDRPD	$0.84 \pm 0.14^{\dagger}$	-0.02 ± 0.13	-0.03 ± 0.11

TabNet, with mean values close to zero and limited dispersion across datasets. This systematic divergence suggests that TabNet captures substantially different feature-ranking patterns.

Overall, these results highlight complementary roles for XAI methods: when consensus-driven interpretation is desired, LIME and SHAP provide aligned perspectives, whereas TabNet offers a more distinct and potentially exploratory view of feature importance.

Robustness: The robustness scores for LIME, SHAP, EBM, and TabNet are reported in Table 7. Results are summarized using mean ± standard deviation over test instances to capture both average sensitivity and variability of explanations under small input perturbations.

Table 7. Robustness of XAI methods measured as sensitivity of explanations to input perturbations, reported as mean ± standard deviation over test instances. Best-performing (lowest) values are highlighted with †. Standard deviations are not shown when they round to 0.000.

Method	PDHS (C-section)	Diabetes retinopathy	Breast cancer	ESDRPD
LIME	0.0477 ± 0.0338	0.0428 ± 0.0197	0.0422 ± 0.0202	0.0672 ± 0.0272
SHAP	0.0002 ± 0.0004	0.0132 ± 0.0219	0.0078 ± 0.0193	0.0002 ± 0.0009
EBM	$0.0000 \pm 0.0010^{\dagger}$	0.0021 ± 0.0103	0.0175 ± 0.0388	$0.0000^{\dagger}$
TabNet	0.0006 ± 0.0040	$0.0009 \pm 0.0106^{\dagger}$	$0.0000 \pm 0.0007^{\dagger}$	0.0201 ± 0.0343

EBM and TabNet demonstrated the highest robustness overall, achieving consistently low sensitivity scores across datasets. EBM in particular exhibits near-zero mean robustness with negligible variance, indicating highly stable explanations across instances. TabNet similarly produced robust explanations in most datasets, although with slightly higher variability in some cases (e.g., ESDRPD), suggesting limited dataset-dependent sensitivity.

SHAP generally showed low sensitivity, reflecting stable explanations under perturbations, but with moderate variability in certain datasets such as Diabetes retinopathy and Breast cancer. In contrast, LIME was consistently the most sensitive to noise and exhibited the largest standard deviations, indicating strong sensitivity to small input changes and substantial instance-level instability.

Overall, these findings reinforce the importance of robustness in clinical settings, where explanations must remain reliable under minor data variations. The inclusion of standard deviations further highlights that robustness differences are not only reflected in average behavior but also in the stability of explanations across individual patients.

Evaluation of rule-based XAI method

Precision, coverage, simplicity, and robustness: The quantitative results for Anchors are reported in Table 8, summarizing how rule-based explanations vary with different precision thresholds. All metrics are reported as mean ± standard deviation over test instances to capture both average behavior and instance-level variability. In Anchors, the threshold controls the minimum acceptable precision for a rule to be extracted. While higher thresholds increase rule reliability (precision), they reduce the number of data points covered (coverage) and typically increase rule complexity (simplicity), as more conditions are needed to meet the stricter criteria.

Table 8. Precision, coverage, simplicity, and robustness of Anchors across datasets, reported as mean ± standard deviation over test instances. Standard deviations are not shown when they round to 0.000.

Dataset	Threshold	Precision	Coverage	Simplicity	Robustness
PDHS	0.90	0.96 ± 0.04	0.29 ± 0.14	2.35 ± 4.05	0.39 ± 0.41
(C-section)	0.95	0.97 ± 0.04	0.24 ± 0.17	2.81 ± 4.68	0.40 ± 0.41
	0.99	0.99 ± 0.04	0.19 ± 0.19	3.65 ± 5.13	0.47 ± 0.38
Diabetes	0.90	0.91 ± 0.04	0.02 ± 0.03	11.01 ± 9.38	0.54 ± 0.24
retinopathy	0.95	0.94 ± 0.05	0.01 ± 0.01	13.89 ± 10.41	0.56 ± 0.20
	0.99	0.96 ± 0.06	0.01 ± 0.01	16.89 ± 10.79	0.56 ± 0.18
Breast	0.90	0.99 ± 0.02	0.22 ± 0.11	2.03 ± 0.17	0.72 ± 0.18
cancer	0.95	0.99 ± 0.01	0.22 ± 0.09	2.05 ± 0.25	0.72 ± 0.21
	0.99	1.00	0.18 ± 0.08	2.37 ± 0.57	0.70 ± 0.22
	0.90	0.93 ± 0.03	0.36 ± 0.14	2.12 ± 1.19	0.33 ± 0.38
ESDRPD	0.95	0.99 ± 0.02	0.21 ± 0.10	3.22 ± 1.91	0.45 ± 0.34
	0.99	1.00 ± 0.01	0.19 ± 0.11	3.60 ± 2.04	0.47 ± 0.34

Across all datasets, increasing the threshold improves rule precision, often reaching values of ≥ 0.99, but leads to a substantial reduction in coverage, in some cases explaining as little as 1% of instances (e.g., Diabetes retinopathy). This highlights that very precise rules typically apply only to a small subset of patients. Importantly, the relatively small standard deviations observed for precision indicate that Anchors reliably enforces the desired precision expectation level across instances once a threshold is fixed.

Coverage, in contrast, exhibits low mean values and modest variability across datasets, particularly for Diabetes retinopathy, where coverage remains consistently low across thresholds. This suggests that the narrow applicability of Anchors rules is not driven by a few outlier instances, but is a systematic consequence of enforcing high-precision constraints. In contrast, datasets such as PDHS and Breast cancer retain moderate coverage at lower thresholds, indicating that broader rules can be obtained when the underlying decision structure is simpler.

As thresholds increase, simplicity declines, with rules becoming longer and involving more feature conditions to satisfy stricter precision requirements. The reported standard deviations indicate substantial instance-level variability in rule length, particularly for datasets exhibiting higher heterogeneity such as Diabetes retinopathy and PDHS, where simplicity standard deviations are comparable to or larger than the mean. This indicates that while some patients can be explained using relatively compact rules, others require substantially more conditions to achieve the same precision level, which has implications for interpretability and clinician usability.

Robustness generally remains moderate and, in some datasets, worsens slightly at stricter precision thresholds (lower values indicate higher robustness against noise), suggesting increased sensitivity of longer rules to small perturbations. The moderate

standard deviations observed for robustness suggest that this sensitivity varies across instances, reflecting heterogeneity in how tightly rule conditions bind to individual patient profiles. Compared to the near-zero robustness observed for inherently interpretable models such as EBM, Anchors rules are less stable under perturbations, particularly at higher thresholds.

Overall, these results reinforce a fundamental trade-off in rule-based explainability: *precision versus coverage versus simplicity versus robustness*. Anchors excels at producing human-readable, high-precision rules for narrowly defined subpopulations, but at the cost of limited coverage, variable rule complexity, and reduced robustness under noise. Compared to attribution-based methods such as LIME and SHAP, Anchors offers more actionable (human-readable) rule-based explanations, but for a smaller and more selective set of cases.

Clinical relevance of global features across XAI methods

Scope and interpretation. This subsection presents *descriptive global explanation summaries* to contextualize model behavior and to examine qualitative alignment with well-established clinical risk factors reported in prior literature. The analysis is intended to provide transparent numerical summaries and triangulation across explainers and conventional tree importances, consistent with the framework-level focus of this study. Clinical validation and clinician-led qualitative review are outside the scope of the present work.

To contextualize model behavior, we examined global feature-importance outputs from EBM, SHAP, and TabNet across all four healthcare tabular datasets, and report qualitative alignment with well-established clinical risk factors in prior literature as a descriptive plausibility check rather than clinical validation. In addition, for the tree-based explainee we juxtapose conventional RF global importances (PFI and Gini/MDI) with SHAP/EBM/TabNet to check agreement on top signals (see Appendix, Figs. 3–10). Below, we summarize the top-ranked features per method and relate them to findings reported in prior clinical literature.

PDHS (C-section) dataset

EBM: Key features include *Births in the last five years & Had Previous C-section*, *Had Previous C-section*, *Number of household members & Mothers BMI*, and *Age of Mother & Number of ANC Visits*.

SHAP: Important features highlighted are *Had Previous C-section*, *Suffered by Domestic Violence*, *Number of ANC Visits*, and *Type of place of residence*.

TabNet: Significant features identified are *Births in the last five years*, *Reading Newspaper or Magazine*, *Had Previous C-section*, and *Household Toilet Facility*.

RF baseline (tree importances): Random Forest permutation (PFI) and Gini/MDI importances concur on *Had Previous C-section* as the dominant factor, with *Number of ANC Visits* and maternal-demographic signals (e.g., age/household context) also ranked highly. This aligns with EBM/SHAP, while SHAP further elevates socio-behavioral risk (*Domestic Violence*) not always captured at the very top by RF split statistics (Appendix, RF: Fig. 3; SHAP/EBM/TabNet: Fig. 4).

These features align with established maternal health research. For instance, prior C-sections and antenatal care are known predictors of surgical delivery outcomes, while domestic violence and household conditions reflect broader socioeconomic risks [25, 49, 50].

Diabetes retinopathy dataset

EBM, SHAP, TabNet: All three methods consistently identify variants of *ma* (microaneurysms) and *ex* (exudates) features, such as *ma0.5*, *ex0.5*, and *exm1*, as dominant.

RF baseline (tree importances): RF PFI and Gini/MDI highlight lesion features, *microaneurysms (ma*)* and *exudates (ex*)*, as primary drivers, mirroring the SHAP/EBM emphasis on *ma0.5*, *ex0.5*, *exm1*. The concordance across RF, SHAP, and EBM supports clinical plausibility for these retinal markers (Appendix, RF: Fig. 5; SHAP/EBM/TabNet: Fig. 6).

This coherence reinforces known clinical markers for diabetes retinopathy, where microaneurysms and retinal exudates serve as primary diagnostic indicators [51, 52].

Breast cancer dataset

EBM, SHAP, TabNet: All methods surfaced geometric and morphological features, including *Mean/Worst concave points*, *Texture*, *Area*, and *Perimeter*.

RF baseline (tree importances): RF importances (PFI and Gini/MDI) surface geometric/morphological variables, *concave points (mean/worst)*, *area*, *perimeter*, and texture-related measures, consistent with EBM/SHAP rankings. This agreement with oncological criteria strengthens face-validity of the global signals (Appendix, RF: Fig. 7; SHAP/EBM/TabNet: Fig. 8).

These features are widely accepted in breast cancer literature as strong predictors of malignancy, offering interpretability that resonates with established oncological criteria [53, 54].

ESDRPD dataset

EBM, SHAP, TabNet: Consistently high-ranking features include *Polyuria*, *Polydipsia*, *Gender*, and *Partial paresis*, with additional features like *Muscle stiffness* and *Sudden weight loss* emphasized by TabNet.

RF baseline (tree importances): RF PFI and Gini/MDI align with SHAP/EBM on classic symptoms, *Polyuria*, *Polydipsia*, *Gender*, and *Partial paresis*, as leading contributors, with occasional elevation of related neurologic/metabolic items. This matches the medically expected profile (Appendix, RF: Fig. 9; SHAP/EBM/TabNet: Fig. 10).

These symptoms correspond to hallmark indicators of early diabetes diagnosis and neurological complications in diabetic progression [55–57].

These findings indicate that global explanation summaries from EBM, SHAP, and TabNet are numerically coherent and qualitatively consistent with well-established clinical risk factors reported in prior studies, providing descriptive context rather than clinical validation. While SHAP and EBM often highlight similar top predictors, TabNet sometimes reveals distinct yet interpretable patterns. This convergence and divergence across methods emphasize the value of using diverse XAI techniques to triangulate insights, especially in complex clinical datasets. Figures supporting these analyses are presented in the Appendix (see Figs 3–10).

Discussion

This study presents a comparative, metric-based evaluation of five explainable AI (XAI) methods, SHAP, LIME, Anchors, EBM, and TabNet, on structured healthcare datasets. Using six evaluation metrics, fidelity, simplicity, consistency, robustness, precision, and coverage, we systematically assessed the strengths and limitations of these methods in tabular clinical settings (see Table 1). In addition, we juxtaposed conventional Random Forest (RF) global feature importances, permutation (PFI) and Gini/MDI, as contextual baselines against SHAP/EBM/TabNet global summaries. To probe sensitivity to the choice of explainee, the same post-hoc protocol can be re-run on alternative top-ranked predictors (e.g., GBC or XGB), as discussed in *Limitations and future work*.

Method-level interpretations and trade-offs

Fidelity: SHAP achieved perfect score fidelity in approximating the probability outputs of Random Forest models (mean = 1.000 with zero variance across datasets), and near-perfect performance for predicted class labels (e.g., 0.994 ± 0.076 in Diabetes retinopathy; Table 4) . These results empirically confirm TreeSHAP's theoretical assurances [15, 34]. Minor deviations in decision fidelity occur when class probabilities are equal and tie-breaking can flip labels marginally.

In contrast, LIME showed lower fidelity and markedly higher instance-level variability, especially in decision fidelity (e.g., 0.538 ± 0.499 in Diabetes retinopathy), consistent with the behavior of local linear surrogates under complex or non-linear decision boundaries (Table 4).

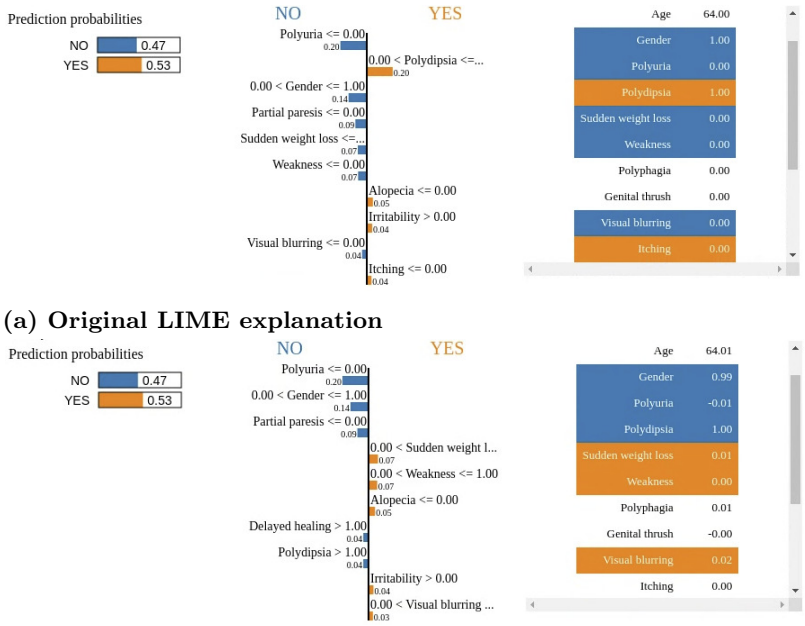
Simplicity: TabNet frequently produced the simplest explanations, particularly at lower thresholds (0.01 and 0.05), using the fewest dominant features across most datasets. However, the accompanying standard deviations (Table 5) show that explanation size can vary substantially across instances, especially for TabNet and SHAP in some datasets, indicating heterogeneous per-patient explanation complexity.

LIME achieved the simplest outputs in only a few cases and ranked second overall in simplicity, and in several settings its explanation size saturates with near-zero variance once the threshold constraint becomes binding. SHAP explanations were consistently denser, while EBM generated the most feature-rich outputs (see Table 5). Among rule-based explanations, Anchors produced short rules in several datasets (Table 8), but rule length can vary substantially across instances (large simplicity standard deviations in PDHS and Diabetes retinopathy), and its coverage remains limited, particularly at higher precision thresholds.

Consistency: SHAP and LIME showed moderate to high agreement in top feature rankings, especially in datasets like ESDRPD. TabNet, however, showed low (near-zero) mean agreement with both LIME and SHAP, with substantial standard deviations, indicating that its ranked feature attributions can differ substantially across instances (Table 6).

Robustness: EBM and TabNet exhibited strong robustness to input perturbations, maintaining low sensitivity scores across datasets (Table 7). Standard deviations further show that robustness is not uniform across patients: for example, SHAP and TabNet display higher variability in some datasets, while LIME has both higher mean sensitivity and the largest variance, indicating frequent instance-level instability. Anchors also shows non-trivial sensitivity under perturbation (Table 8), with robustness varying

across datasets and thresholds. Fig 2 demonstrates this instability using two LIME explanations for the same ESDRPD instance, once with the original input and once with a minor perturbation. While the predicted class remains the same, the explanations shift significantly in content and magnitude. This perturbation was applied by adding small Gaussian noise ($\mathcal{N}(0, 0.01^2)$) to the original input instance, as described in the *Materials and methods* section under Robustness. We emphasize that the robustness analysis is intended as a comparative sensitivity probe under minimal input perturbations, not as a simulation of clinically realistic patient changes; thus, the reported robustness scores should be interpreted as relative explanation stability across methods rather than absolute clinical reliability.



(a) Original LIME explanation

(b) Noisy LIME explanation

Fig 2. Original and perturbed/noisy LIME explanations for an ESDRPD instance (perturbation: element-wise Gaussian jitter, $\epsilon \sim \mathcal{N}(0, 0.01^2)$, on numerically encoded inputs; no model refitting).

In parallel, Anchors generated different rule sets for the same instance under noise:

- **Original Anchor Rule:** $Age \leq 47.50$ AND $Polyuria \leq 0.00$ AND $Visual_blurring \leq 0.00$ AND $Gender > 0.00$ AND $Itching > 0.00$
- **Noisy Anchor Rule:** $Polydipsia > 0.00$ AND $Irritability > 0.00$ AND $Obesity \leq 0.00$ AND $Age > 58.00$

These results show that post-hoc explainers relying on local sampling (LIME) or greedy rule induction (Anchors) can be fragile under perturbation, especially in datasets with closely correlated features or small decision boundaries.

Precision and coverage: Anchors excelled in producing precise rules with interpretable conditions, making it valuable in applications like patient subgroup identification. However, the cost of this precision is low coverage, limiting its utility for broad model auditing (Table 8). Reporting mean ± standard deviation clarifies that these behaviors are systematic rather than driven by a few outliers: precision shows

relatively small variability once a threshold is fixed, whereas coverage is consistently low (and only modestly variable), especially in Diabetes retinopathy. Rule complexity (simplicity) exhibits large instance-level dispersion in some datasets (e.g., PDHS and Diabetes retinopathy), indicating that some patients can be captured by compact rules while others require substantially longer rules to meet the same precision threshold. Robustness also shows moderate variability, suggesting heterogeneous stability under perturbations across instances. This trade-off, between narrow, high-confidence rules and broader applicability, should be considered when deploying Anchors in clinical decision support.

Triangulation with traditional tree importances. Across datasets, RF PFI and Gini/MDI broadly corroborate the global signals surfaced by SHAP/EBM/TabNet: (i) *PDHS (C-section)*, strong emphasis on *Had Previous C-section* with *ANC visits*/maternal-demographic factors also high; SHAP further lifts socio-behavioral risk (*Domestic Violence*). (ii) *Diabetes retinopathy*, lesion markers (*ma**, *ex**) dominate across RF and SHAP/EBM. (iii) *Breast cancer*, geometric/morphological variables (*concave points*, *area*, *perimeter*, texture) agree across methods. (iv) *ESDRPD*, classic symptoms (*Polyuria*, *Polydipsia*, *Gender*, *Partial paresis*) align. Where differences arise, they are typically in the ordering of secondary factors rather than wholesale disagreement, and can often be traced to known properties of impurity-based scores (e.g., bias toward high-cardinality splits) or permutation effects under correlated features. This triangulation increases confidence in the face-validity of the global patterns while motivating local metrics for fidelity/robustness.

Clinical relevance and scope of global explanations. The global explanation summaries discussed above (e.g., aggregated SHAP importances, EBM main-effect functions, and TabNet attention profiles) are intended as *descriptive, numerical context* rather than as statistically grounded feature selection, causal analysis, or clinical validation. Their role is to summarize model behavior and to support a qualitative alignment check against well-established risk factors reported in the clinical literature. No clinician-led qualitative review was conducted, and we do not claim clinical significance or medical decision validity based on these results alone. These summaries should therefore be interpreted as interpretability-oriented diagnostics that complement, rather than replace, domain-expert evaluation.

Best practices and practical recommendations

Based on our evaluation, we recommend the following XAI method selection guidelines tailored to typical healthcare scenarios. These recommendations are intended as practical heuristics for model debugging and interpretability workflows, rather than as guidance for clinical deployment.

- **For fidelity-critical applications (e.g., probability calibration):** Use SHAP with tree-based models (see Table 4) such as Random Forest or XGBoost. For example, in oncology risk stratification tools or ICU early warning systems, precise probability estimates are essential for triggering critical interventions like escalation of care. Misalignment between explanations and model outputs may result in incorrect risk thresholds.

- **For simplicity and rapid user-facing explanations:** Consider TabNet, LIME or Anchors (see Tables 5 and 8). These are suitable in resource-constrained or community settings, such as mobile-based screening tools for frontline triage

systems, where clinicians or health workers need fast, intuitive reasoning to communicate decisions effectively to patients.

- **For stable explanations under noisy conditions:** Use EBM or TabNet (see Table 7). These models are beneficial in environments like continuous patient monitoring or real-time clinical dashboards, where small perturbations due to sensor noise or missing data should not drastically alter the explanations provided.

- **For local, rule-based decision support:** Apply Anchors, noting its narrow applicability. Anchors is effective for tasks requiring precise rule extraction (see Table 8), such as auditing medical eligibility criteria, drug contraindication alerts, or subgroup-specific recommendations where interpretability outweighs global applicability.

- **When using tree ensembles, sanity-check with RF importances first, then escalate to SHAP when stakes are high:** PFI/MDI provide a fast global ranking for model debugging, but can misorder correlated or high-cardinality features; SHAP (TreeSHAP) offers faithful local attributions and more reliable global aggregation when decisions hinge on calibrated probabilities or safety-critical thresholds.

These recommendations are not limited to healthcare. The evaluation framework and insights can be applied to domains such as finance, insurance, fraud detection, marketing, and industrial systems, where tabular data is prevalent, and explainability is critical for regulatory compliance, operational trust, or model debugging.

Global alignment with clinical literature

Across datasets, SHAP, EBM, and TabNet surfaced global features that are qualitatively consistent with well-established risk factors reported in prior clinical studies, including maternal indicators for C-sections (Fig. 4), retinal lesion markers for diabetes retinopathy (Fig. 6), tumor morphology measures in breast cancer (Fig. 8), and early diabetes symptoms in ESDRPD (Fig. 10). These alignments are detailed in the *Results* section. We interpret this alignment as a descriptive plausibility check that contextualizes model behavior, rather than as feature validation, causal evidence, or clinical significance.

Conversely, lower-fidelity or less robust explainers (e.g., LIME and Anchors under perturbation) can produce explanations that vary substantially across instances, underscoring the value of metric-guided explainer selection and sanity checks in high-stakes tabular settings.

Limitations and future work

This study focuses on publicly available, tabular healthcare datasets and does not extend to imaging, time-series, or longitudinal modalities that are increasingly prominent in clinical AI. We evaluate six quantitative metrics; however, complementary qualitative dimensions lie outside our scope and warrant future user studies.

Human-centred evaluation. We assess explanation quality via quantitative, reproducible proxies, including fidelity, simplicity, inter-explainer consistency, robustness, and precision–coverage evaluated across datasets. Direct human-centred evaluation (e.g., clinician task accuracy, completion time, error rate, SUS/NASA–TLX, perceived trust or appropriateness) is beyond the scope of this study and planned as future work with domain experts. Direct clinician involvement in reviewing explanation

outputs and assessing medical plausibility was not part of this study and is identified as a necessary direction for future work.

Scope of coverage. Our emphasis is on a principled *evaluation framework* and transparent operationalization of quantitative XAI metrics, not exhaustiveness over explainer families. Metrics are reported only where theoretically appropriate to each explainer's output type (e.g., rank-based consistency for per-instance attribution vectors; precision/coverage for rule sets). This policy avoids cross-paradigm conflation and supports reuse; future extensions can broaden method families under the same protocol.

Design choices. Our post-hoc comparison is performed on a *single focal predictor* chosen *after* benchmarking via a pre-specified average-rank rule (F1 primary; accuracy tie-breaker). This controls model-induced variance and isolates explainer differences, but means we did not compute all post-hoc explainers on all predictors. In contrast, EBM and TabNet are assessed via their inherent (ante-hoc) interpretability, which is conceptually distinct from post-hoc surrogates. We also avoid extensive per-dataset hyperparameter tuning (fixed seeds; library defaults) to reduce confounding variability in explanation metrics. While this may affect absolute predictive scores, it improves comparability and reproducibility of explainer results.

Global tree importances. PFI and Gini/MDI are global summaries: they aid triangulation but neither guarantee local agreement with the model nor immunity to known biases (e.g., impurity-based bias toward high-cardinality features; permutation underestimation/instability with collinear predictors or leakage). Their sensitivity to data distribution and feature correlation further motivates our use of local, instance-level metrics (fidelity, robustness) for post-hoc explainers.

Perturbation model (robustness). Robustness was quantified using small, zero-mean, element-wise Gaussian perturbations as a *method-agnostic sensitivity probe* to assess explanation stability under minimal input variation, rather than to model realistic clinical or physiological changes. This mathematical probe isolates the sensitivity of explanations to minor measurement variability across mixed data types. Designing *clinically calibrated*, biomarker-aware perturbations (e.g., constrained to physiological ranges or valid codebooks) is valuable but beyond our present scope.

Method breadth. We study widely used representatives, post-hoc feature-attribution methods (SHAP, LIME), a rule-based local explainer (Anchors), and inherently interpretable models (EBM, TabNet), rather than an exhaustive catalog. Other rule-list/decision-set learners and glass-box variants could be included; our open code is structured to accommodate such extensions.

Metric scope. The reported metrics are computed deterministically on a fixed test set, and the study is designed for descriptive, instance-level comparison rather than population inference. Accordingly, we report dispersion via standard deviation to characterize variability and potential failure modes across patients.

We consolidate and operationalize commonly cited quantitative metrics (fidelity, simplicity, consistency, robustness, and precision/coverage for rule-based explanations). These do not exhaust all meaningful criteria (e.g., human-centered measures such as usefulness or actionability), and some metrics are family-specific by design (e.g., precision/coverage for rules; rank-based consistency for single-feature attribution lists). Where relevant, partial sensitivity is reported directly in the results (e.g., multiple

simplicity thresholds and Anchor precision levels), while broader sensitivity analyses are deferred to future work.

Future work. Future research should: (i) extend the benchmark to temporal and multimodal clinical datasets; (ii) run clinician-in-the-loop studies to assess perceived usefulness and cognitive effort; (iii) integrate additional explainer families, especially counterfactual and example-based methods, with appropriate, family-specific metrics (e.g., plausibility, feasibility, actionability); (iv) replicate the post-hoc analysis on alternative top-ranked predictors (e.g., boosted trees or a neural tabular model) to probe explainer–model interactions and assess cross-model generalizability; (v) harmonize evaluation between inherent and post-hoc explanations (e.g., restricting EBM consistency to main-effect functions or developing interaction-aware consistency measures); (vi) systematically characterize RF–SHAP divergence under controlled feature correlation and cardinality settings to identify regimes where traditional tree importances deviate most from faithful local attributions; and (vii) develop domain-calibrated perturbation schemes for robustness (biomarker- and codebook-aware ranges) and compare them to generic Gaussian jitter to assess clinical impact.

Conclusion

We presented a reproducible, metric-driven *evaluation framework* for explainable AI (XAI) on clinical tabular models, focusing on the standardized operationalization and comparative application of established quantitative metrics. Rather than proposing a new explainer or evaluation metric, we formalized six established quantitative metrics, fidelity, simplicity, consistency, robustness, precision, and coverage, into explicit equations, mapped their applicability across method families, and released open-source code with a pre-specified focal-model selection protocol to reduce confounds. Applying the framework to five widely used approaches (LIME, SHAP, Anchors, EBM, TabNet) over four healthcare datasets yields consistent patterns: SHAP (TreeSHAP) provides exact score fidelity for tree ensembles and near-perfect decision fidelity; LIME produces sparser (simpler) local rationales but with lower fidelity and robustness; TabNet most often yields the fewest dominant features across thresholds; EBM and TabNet consistently deliver the most robust explanations under small perturbations; and Anchors supplies high-precision, human-readable rules that exhibit the canonical precision–coverage trade-off as thresholds tighten. Reporting mean ± standard deviation reveals substantial instance-level variability for some explainers and metrics (e.g., LIME decision fidelity and TabNet/SHAP explanation sizes), highlighting that close mean values should be interpreted alongside dispersion rather than in isolation, particularly in the absence of inferential testing where stability and failure modes are the primary concern. LIME and SHAP show moderate-to-high agreement on top features, while global importance from EBM/SHAP/TabNet aligns with known risk factors.

Practical guidance. Match method to priority: (i) *Fidelity-critical* (auditability, risk): SHAP with tree ensembles; (ii) *Human-centered simplicity* (quick local insight): LIME; (iii) *Robustness* (noisy inputs, safety): EBM or TabNet; (iv) *Rule precision on subpopulations*: Anchors (accepting reduced coverage at higher precision).

Scope and next steps. The framework targets tabular healthcare and is readily transferable to other high-stakes tabular domains. It does not claim exhaustiveness of methods or metrics; rather, it offers a transparent baseline that can be extended. Future work includes adding temporal and multimodal datasets, clinician-in-the-loop evaluations of usefulness and cognitive load, integrating counterfactual and example-based families with appropriate metrics, and probing explainer–model

interactions by replicating the post-hoc analysis on alternative top-ranked predictors. Robustness analysis can be further strengthened by replacing generic Gaussian jitter with domain-calibrated perturbations tied to biomarker ranges and valid codebooks.

Finally, when tree ensembles are used, traditional global RF importances (Gini/MDI and permutation) are useful *descriptive cross-checks* but not substitutes for local fidelity-based evaluation; in our experiments, they qualitatively aligned with SHAP/EBM global patterns while SHAP additionally surfaced socio-behavioral risk factors that split-based importances may down-weight (see Appendix).

Overall, the framework turns narrative guidance into testable, quantitative practice, enabling apples-to-apples comparisons and more defensible XAI selection in regulated settings.

Author Contributions

Conceptualization: M. Atif Qureshi, Abdul Aziz Noor, Wael Rashwan, Arjumand Younus. **Data curation:** Abdul Aziz Noor, M. Atif Qureshi, Awais Manzoor. **Methodology - XAI Methods:** M. Atif Qureshi, Abdul Aziz Noor. **Methodology - XAI Metrics:** M. Atif Qureshi, Abdul Aziz Noor, Muhammad Deedahwar Mazhar Qureshi. **Software:** M. Atif Qureshi, Abdul Aziz Noor. **Formal analysis:** M. Atif Qureshi. **Investigation:** M. Atif Qureshi, Abdul Aziz Noor. **Supervision/Mentor of Student Authors:** M. Atif Qureshi, Wael Rashwan, Arjumand Younus. **Validation:** M. Atif Qureshi, Muhammad Deedahwar Mazhar Qureshi, Wael Rashwan, Arjumand Younus. **Visualization:** M. Atif Qureshi, Abdul Aziz Noor. **Writing – original draft:** M. Atif Qureshi, Abdul Aziz Noor. **Writing – review & editing:** M. Atif Qureshi, Muhammad Deedahwar Mazhar Qureshi, Awais Manzoor, Wael Rashwan, Arjumand Younus.

Data and Code Availability

All code for metric definitions and experiments is available at `https://github.com/matifq/XAI_Tab_Health`. All datasets used are publicly available as cited in the manuscript.

References

1. Noor AA, Manzoor A, Mazhar Qureshi MD, Qureshi MA, Rashwan W. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. WIREs Data Mining and Knowledge Discovery. 2025;15(2):e70018. doi:https://doi.org/10.1002/widm.70018.

2. Wang S, Qureshi MA, Miralles-Pechuán L, Huynh-The T, Gadekallu TR, Liyanage M. Explainable AI for 6G Use Cases: Technical Aspects and Research Challenges. IEEE Open Journal of the Communications Society. 2024;5:2490–2540. doi:10.1109/OJCOMS.2024.3386872.

3. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018;6:52138–52160.

4. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019;8(8):832.

5. Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics. 2021;10(5). doi:10.3390/electronics10050593.

6. Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artificial Intelligence Review. 2023;56(6):5261–5315.

7. Hedström A, Weber L, Krakowczyk D, Bareeva D, Motzkus F, Samek W, et al. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research. 2023;24(34):1–11.

8. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Comput Surv. 2023;55(13s). doi:10.1145/3583558.

9. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019;9(4):e1312.

10. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Consortium P. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020;20:1–9.

11. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences. 2021;11(11):5088.

12. Hossain MI, Zamzmi G, Mouton PR, Salekin MS, Sun Y, Goldgof D. Explainable AI for medical data: current methods, limitations, and future directions. ACM Computing Surveys. 2025;57(6):1–46.

13. Xu Q, Xie W, Liao B, Hu C, Qin L, Yang Z, et al. Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review. Journal of Healthcare Engineering. 2023;2023.

14. Retzlaff CO, Angerschmid A, Saranti A, Schneeberger D, Roettger R, Mueller H, et al. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. Cognitive Systems Research. 2024;86:101243.

15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

16. Ribeiro MT, 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; 2016. p. 1135–1144.

17. Ribeiro MT, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1). doi:10.1609/aaai.v32i1.11491.

18. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013. p. 623–631.

19. Arik SÖ, Pfister T. TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35(8):6679–6687. doi:10.1609/aaai.v35i8.16826.

20. Nayebi A, Tipirneni S, Reddy CK, Foreman B, Subbian V. WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values. Journal of Biomedical Informatics. 2023;144:104438.

21. Hatwell J, Gaber MM, Atif Azad RM. Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences. BMC Medical Informatics and Decision Making. 2020;20(1):1–25.

22. Settouti N, Saidi M. Preliminary analysis of explainable machine learning methods for multiple myeloma chemotherapy treatment recognition. Evolutionary Intelligence. 2023; p. 1–21.

23. Alfeo AL, Zippo AG, Catrambone V, Cimino MG, Toschi N, Valenza G. From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. Computer Methods and Programs in Biomedicine. 2023;236:107550.

24. Chadaga K, Prabhu S, Bhat V, Sampathila N, Umakanth S, Chadaga R. A Decision Support System for Diagnosis of COVID-19 from Non-COVID-19 Influenza-like Illness Using Explainable Artificial Intelligence. Bioengineering. 2023;10(4):439.

25. Islam MS, Awal MA, Laboni JN, Pinki FT, Karmokar S, Mumenin KM, et al. HGSORF: Henry Gas Solubility Optimization-based Random Forest for C-Section prediction and XAI-based cause analysis. Computers in Biology and Medicine. 2022;147:105671.

26. Chun MY, Park CJ, Kim J, Jeong JH, Jang H, Kim K, et al. Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. Frontiers in Aging Neuroscience. 2022;14:898940.

27. Khanna VV, Chadaga K, Sampathila N, Prabhu S, Chadaga R. A machine learning and explainable artificial intelligence triage-prediction system for COVID-19. Decision Analytics Journal. 2023; p. 100246.

28. Joseph LP, Joseph EA, Prasad R. Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. Computers in Biology and Medicine. 2022;151:106178.

29. Aiosa GV, Palesi M, Sapuppo F. EXplainable AI for Decision Support to Obesity Comorbidities Diagnosis. IEEE Access. 2023;.

30. Peng J, Zou K, Zhou M, Teng Y, Zhu X, Zhang F, et al. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. Journal of medical systems. 2021;45:1–9.

31. Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Information Fusion. 2022;81:84–90. doi:https://doi.org/10.1016/j.inffus.2021.11.011.

32. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems. vol. 35. Curran Associates, Inc.; 2022. p. 507–520. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf.

33. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W, editors. Explainable AI Methods - A Brief Overview. Cham: Springer International Publishing; 2022. p. 13–38. Available from: https://doi.org/10.1007/978-3-031-04083-2_2.

34. Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:180203888. 2018;.

35. Pawlicki M, Pawlicka A, Uccello F, Szelest S, D'Antonio S, Kozik R, et al. Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. Neurocomputing. 2024;602:128282.

36. Khalilpourazari S, Khalilpourazary S, Özyüksel Çiftçioğlu A, Weber GW. Designing energy-efficient high-precision multi-pass turning processes via robust optimization and artificial intelligence. Journal of Intelligent Manufacturing. 2021;32:1621–1647.

37. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018.Available from: https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

38. Manzoor A, Atif Qureshi M, Kidney E, Longo L. A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. IEEE Access. 2024;12:70434–70463. doi:10.1109/ACCESS.2024.3402092.

39. Körner A, Sailer B, Sari-Yavuz S, Haeberle HA, Mirakaj V, Bernard A, et al. Explainable Boosting Machine approach identifies risk factors for acute renal failure. Intensive Care Medicine Experimental. 2024;12(1):55. doi:https://doi.org/10.1186/s40635-024-00639-2.

40. InterpretML Team. Explainable Boosting Machine (EBM) — InterpretML documentation; 2025. Available from: https://interpret.ml/docs/ebm.html.

41. Nori H, Jenkins S, Koch P, Caruana R. Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:190909223. 2019;.

42. Hooker S, Erhan D, Kindermans PJ, Kim B. A benchmark for interpretability methods in deep neural networks. Advances in neural information processing systems. 2019;32.

43. Alvarez Melis D, Jaakkola T. Towards Robust Interpretability with Self-Explaining Neural Networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018.Available from: https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

44. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information fusion. 2023;99:101805.

45. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché Buc F, et al. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). Journal of machine learning research. 2021;22(164):1–20.

46. Bischl B, Casalicchio G, Feurer M, Gijsbers P, Hutter F, Lang M, et al. Openml benchmarking suites. arXiv preprint arXiv:170803731. 2017;.

47. Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? In: Perner P, editor. Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 154–168.

48. Probst P, Boulesteix AL. To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research. 2018;18(181):1–18.

49. Pun KD, Rishal P, Darj E, Infanti JJ, Shrestha S, Lukasse M, et al. Domestic violence and perinatal outcomes–a prospective cohort study from Nepal. BMC public health. 2019;19:1–9.

50. Sy I, Sandie AB, Sylla EM, Cissé B, Fall NA, Sow MO, et al. Spatial and Socioeconomic Inequalities in Cesarean Section Deliveries in Urban Settings in Dakar, Senegal. Journal of Urban Health. 2024; p. 1–11.

51. Kohner EM, Sleightholm M, Bergenstal R, Morse P, Rubenstein A, Bending J, et al. Does microaneurysm count reflect severity of early diabetic retinopathy? Ophthalmology. 1986;93(5):586–589.

52. Sopharak A, Uyyanonvara B, Barman S, Williamson TH. Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods. Computerized medical imaging and graphics. 2008;32(8):720–727.

53. Mangasarian OL, Street WN, Wolberg WH. Breast cancer diagnosis and prognosis via linear programming. Operations research. 1995;43(4):570–577.

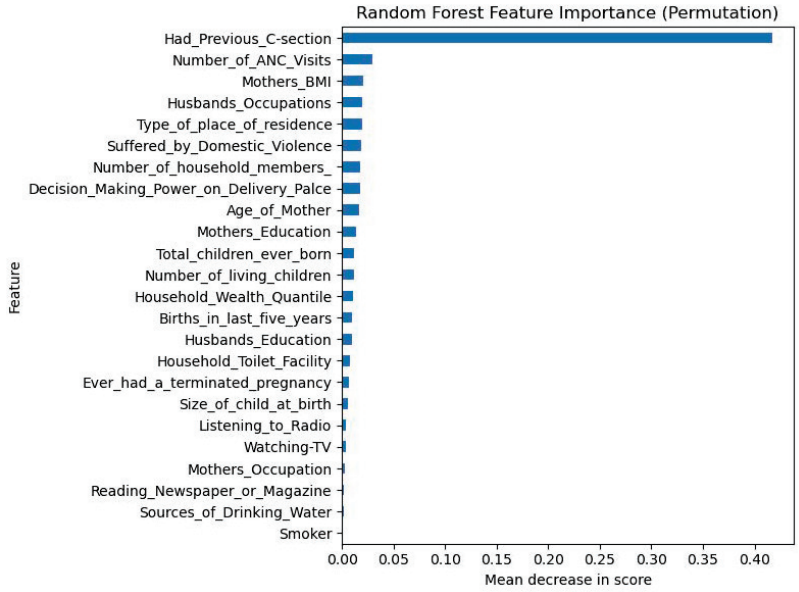
54. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In: Biomedical image processing and biomedical visualization. vol. 1905. SPIE; 1993. p. 861–870.

55. Fournier A. Diagnosing Diabetes: A Practitioner's Plea: Keep It Simple. Journal of General Internal Medicine. 2000;15(8):603.

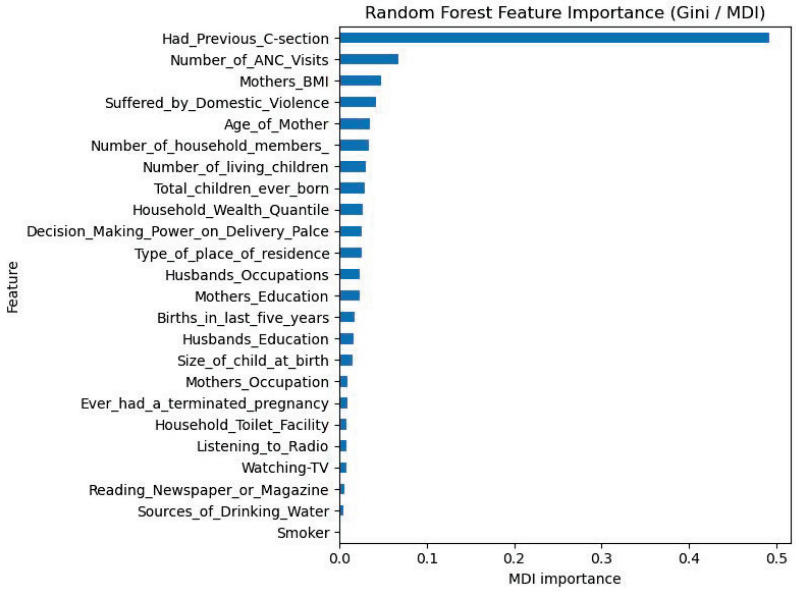
56. Konen JC, Curtis LG, Summerson JH. Symptoms and complications of adult diabetic patients in a family practice. Archives of Family Medicine. 1996;5(3):135.

57. Wyatt LH, Ferrance RJ. The musculoskeletal effects of diabetes mellitus. The Journal of the Canadian Chiropractic Association. 2006;50(1):43.

Appendix

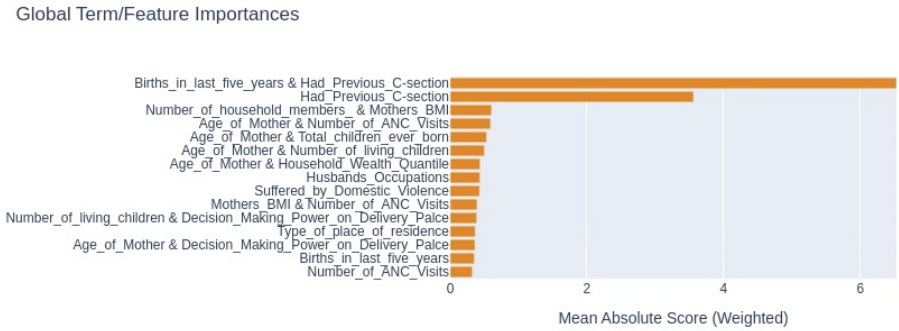


(a) RF Permutation Importance (PFI)

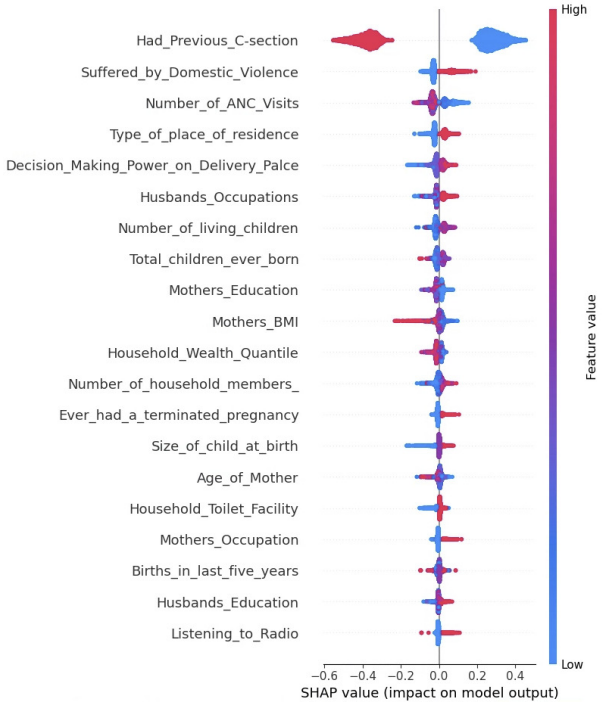


(b) RF Gini/MDI Importance

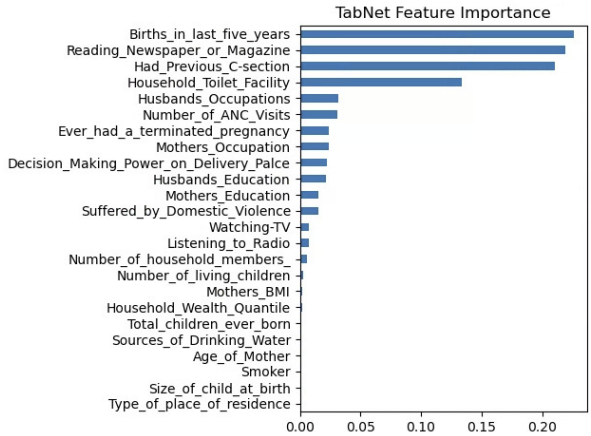
Fig 3. PDHS (C-section): RF global explanations.



(a) Global feature importance using EBM

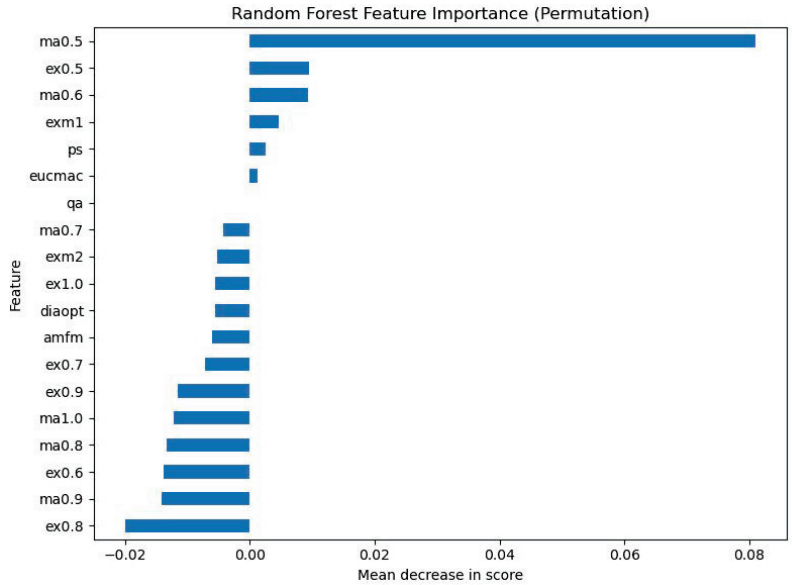


(b) Global feature importance using SHAP

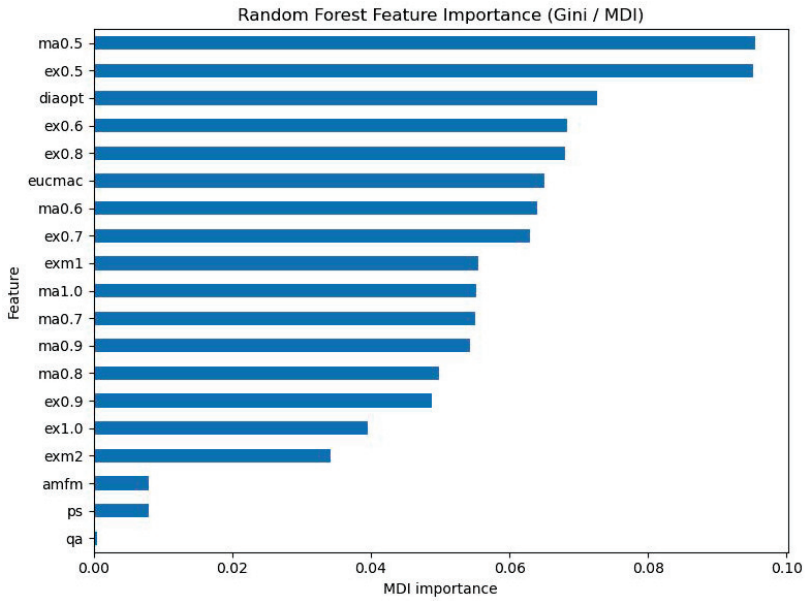


(c) Global feature importance using TabNet

Fig 4. Global feature importance for PDHS (C-section) dataset using EBM, SHAP, and TabNet

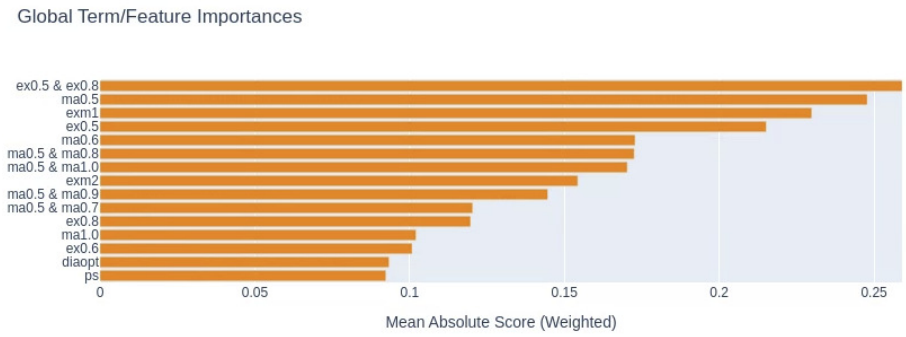


(a) RF Permutation Importance (PFI)

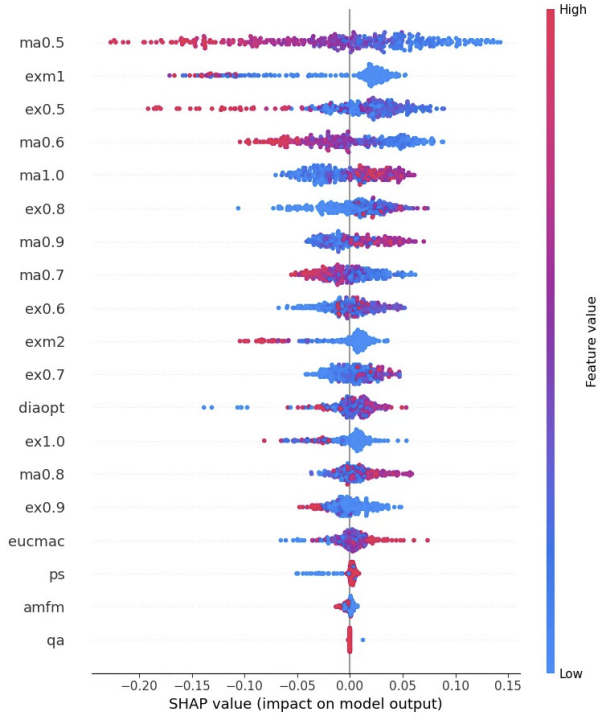


(b) RF Gini/MDI Importance

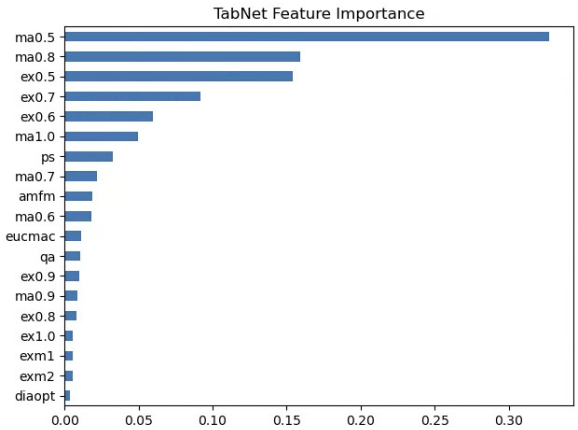
Fig 5. Diabetes retinopathy: RF global explanations.



(a) Global feature importance using EBM

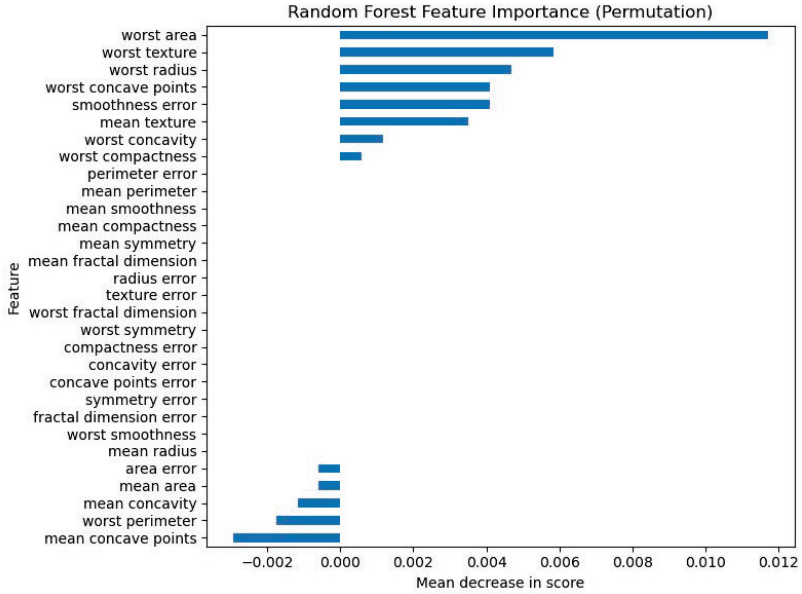


(b) Global feature importance using SHAP

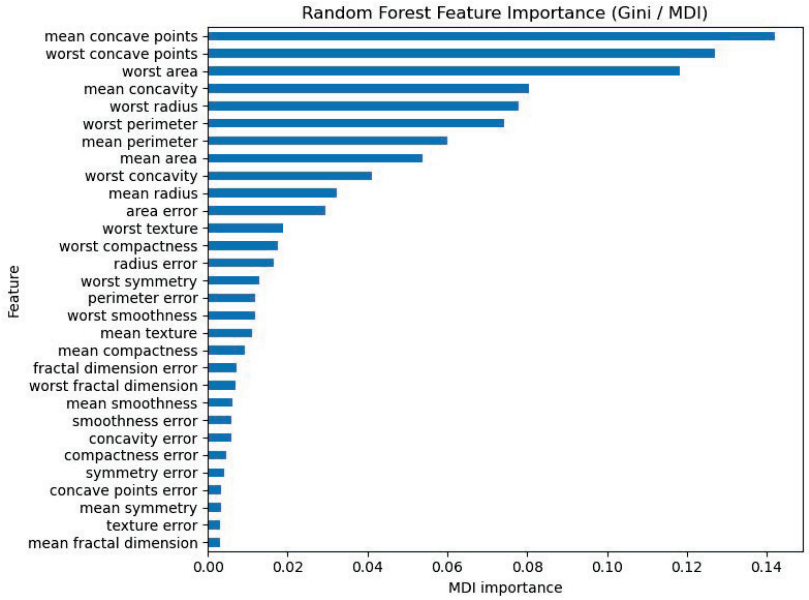


(c) Global feature importance using TabNet

Fig 6. Global feature importance analysis for diabetes retinopathy dataset using EBM, SHAP, and TabNet

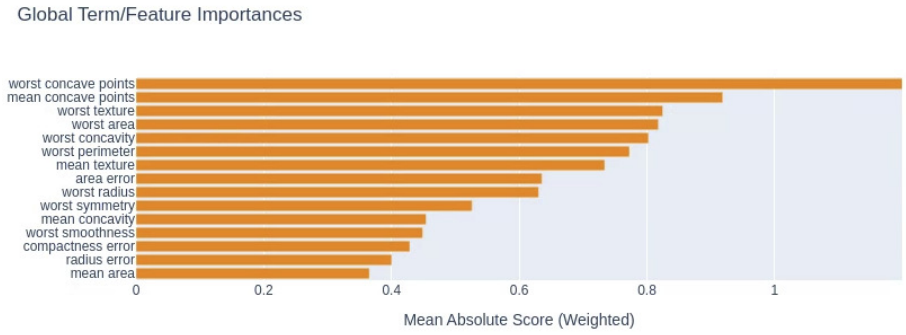


(a) RF Permutation Importance (PFI)

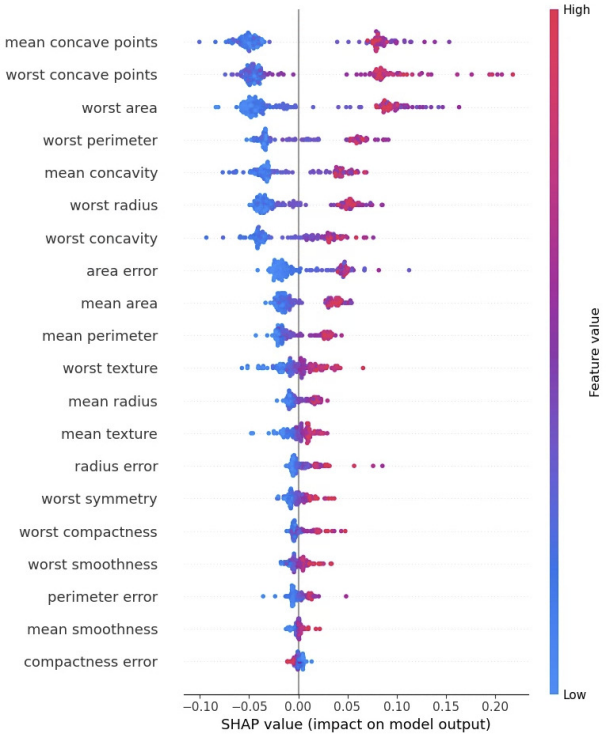


(b) RF Gini/MDI Importance

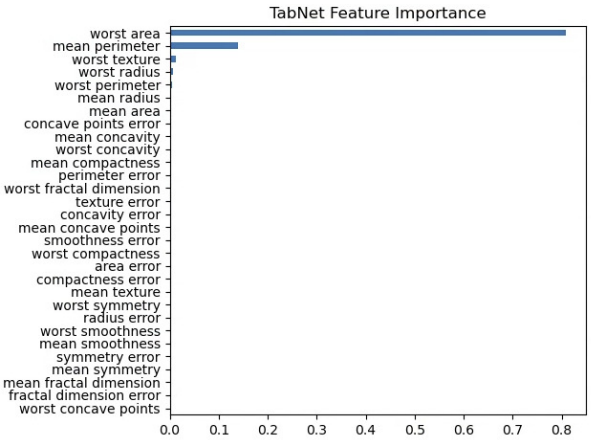
Fig 7. Breast cancer: RF global explanations.



(a) EBM feature importance

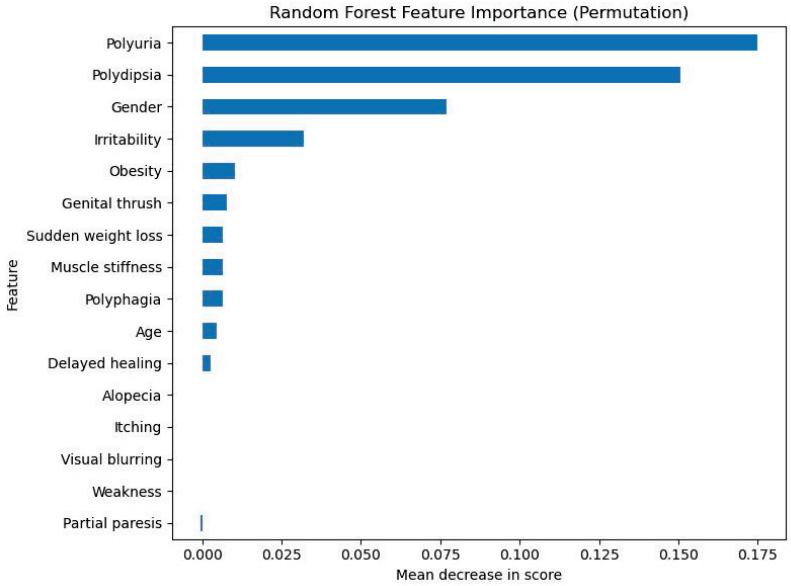


(b) SHAP feature importance

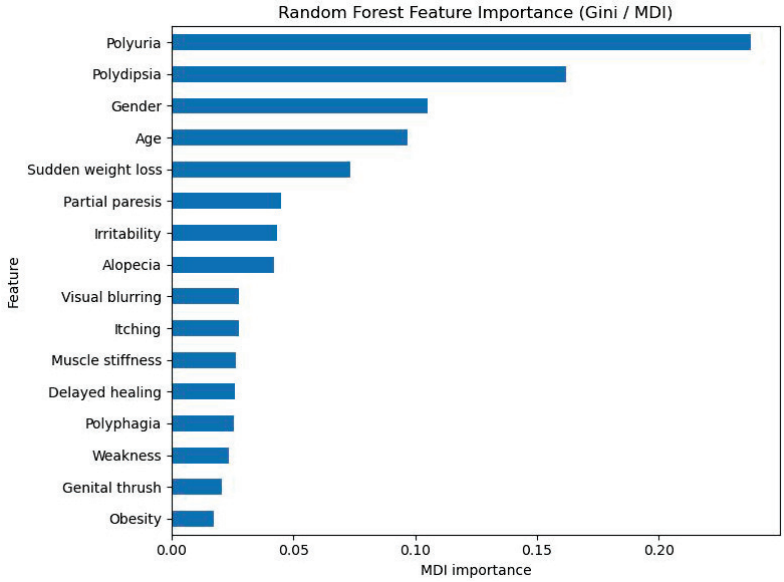


(c) TabNet feature importance

Fig 8. Global feature importance analysis for breast cancer dataset using EBM, SHAP, and TabNet

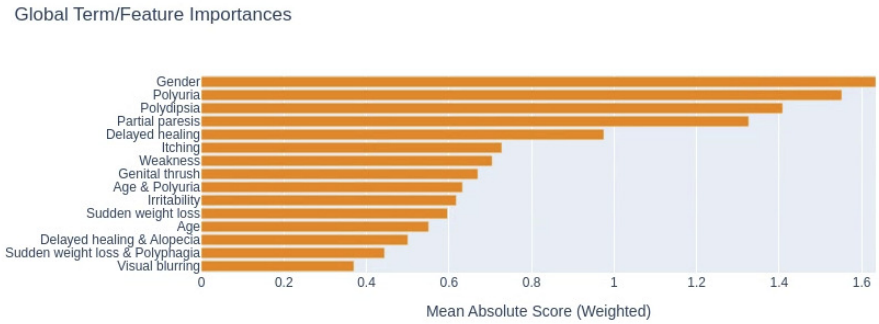


(a) RF Permutation Importance (PFI)

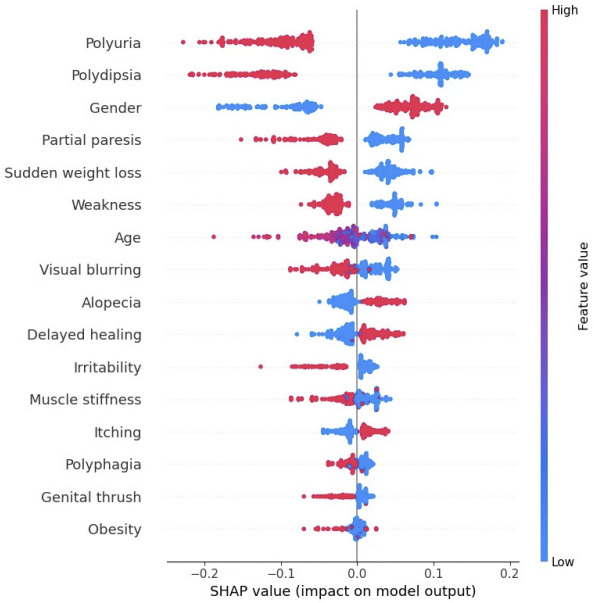


(b) RF Gini/MDI Importance

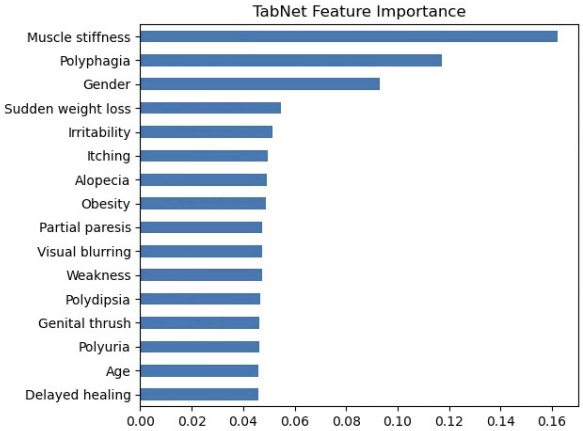
Fig 9. ESDRPD: RF global explanations.



(a) Global feature importance using EBM



(b) Global feature importance using SHAP



(c) Global feature importance using TabNet

Fig 10. Global feature importance analysis for ESDRPD dataset using EBM, SHAP, and TabNet

Table 9. What this work adds vs. representative surveys/toolkits (✓= provided; – = not a primary focus).

Survey/toolkit	Metric eqns	Open-source code	Pre-spec explainee	Post-hoc & ante-hoc	Multi-ds clin. tab.	Comment (vs. this work)
Adadi & Berrada [3]	–	–	–	Post-hoc focus	–	Conceptual survey; no code/benchmarks.
Carvalho et al. [4]	Concepts	–	–	Both (survey)	–	Narrative notions; no formal equations/impl.
Zhou et al. [5]	Concepts	–	–	Both (survey)	–	Evaluation principles; no unified benchmark
Di Martino & Delmastro [6]	Survey	–	–	Mixed	–	Healthcare overview; not metric-focused
Hedström et al. [7] (Quantus)	Library	✓	–	Post-hoc	–	NN-centric metric library (image/saliency emphasis); no pre-specified explainee; no ante-hoc/rule-based; no clinical-tabular benchmark.
Nauta et al. [8]	Consistency notion	–	–	Post-hoc	–	Defines/analyzes consistency; not multi-metric benchmark
This work	✓	✓	✓	✓	✓	**Unified equations + open code; pre-specified explainee; post-hoc & ante-hoc; multi-dataset clinical tabular benchmark**

Notes: "Metric eqns" = formal math definitions (not just verbal notions). "Pre-specified explainee" = rule-based protocol to fix a single post-hoc target model. "Multi-ds clin. tab." = Multi-dataset clinical tabular