Explainability in Action: A Metric-Driven Assessment of Five XAI Methods 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[3◕]

1 ADAPT Centre
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

◕Faculty Author.
‡Student Author.
* atif.qureshi@tudublin.ie

Abstract

As explainable AI (XAI) becomes increasingly important in healthcare machine learning (ML) applications, there is a growing need for reproducible frameworks that quantitatively assess the quality of explanations. In this study, we conduct a comparative evaluation of five widely used XAI methods, LIME, SHAP, Anchors, EBM, and TABNET, on multiple healthcare tabular datasets using six well-established metrics: fidelity, simplicity, consistency, robustness, precision, and coverage. While the metrics are derived from existing literature, we formalize and implement them mathematically, providing open-source code to support standardized benchmarking. Empirically, our experiments confirm that SHAP (with TreeSHAP) achieves perfect fidelity in approximating probability outputs for tree-based models, consistent with its theoretical design. LIME offers simpler explanations but sacrifices fidelity. EBM and TABNET demonstrate strong robustness to input perturbations, while Anchors produces precise rule-based explanations with limited data coverage. These results offer practical guidance for selecting XAI methods based on application priorities such as fidelity, robustness, or simplicity. Our open-source framework enables reproducible, quantitative evaluation of XAI techniques in clinical ML workflows. Although evaluated in a clinical context, the proposed framework and metrics are broadly applicable and generalizable to other domains involving tabular data. The source codes are available at `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 significant barrier to clinical adoption [1]. In high-stakes domains like healthcare, explainability is essential, not only for regulatory compliance and ethical deployment but also to build trust among clinicians and patients [2].

Another aspect of healthcare data is that it is often tabular, particularly in applications involving electronic health records (EHRs), clinical registries, or structured survey data. Despite the prominence of tabular data in real-world systems, most XAI

research has focused on image and text modalities. While many studies suggest using XAI methods, few provide empirical comparisons, and none have thoroughly evaluated various XAI methods using distinct and useful quantitative metrics [1]. This gap limits the ability of healthcare practitioners and data scientists to make informed choices about which XAI methods are best suited to their tasks.

This article addresses this gap by benchmarking five widely used XAI methods, SHAP, LIME, Anchors, EBM, and TABNET, on four healthcare tabular datasets. Rather than proposing a new explainer method, this study introduces a novel contribution in the form of a reproducible, metric-driven evaluation framework. Based on their formal definitions in the literature, we operationalize and implement six established evaluation metrics: fidelity, simplicity, consistency, robustness, precision, and coverage. These metrics are systematically applied to five widely used XAI methods across diverse healthcare tabular datasets.

To our knowledge, this is the first structured benchmarking effort that aligns XAI method selection with quantitative performance trade-offs relevant to clinical decision-making.

The contributions of this work are threefold:

- A multi-metric benchmarking of five widely used XAI methods applied to healthcare tabular data.
- Formal implementation and open-sourcing of reproducible XAI evaluation metric formulations to assess explainability in a consistent way.
- Evidence-based guidance for XAI method selection based on application needs, such as fidelity, robustness, or rules based on precision and coverage.

Literature Review

The Role of Explainability in Clinical AI

XAI is increasingly vital in healthcare, where clinical decisions must be transparent, accountable, and trustworthy. As ML is applied to tasks like diagnosis, treatment planning, and outcome prediction, interpreting model outputs is essential for regulatory acceptance and practitioner trust [3, 4]. To address this, researchers have developed a range of explanation techniques, from instance-level feature attribution to global rule-based logic and visualization tools, that aim to make model decisions more intelligible [5–7].

XAI methods can be divided into post-hoc and ante-hoc [8]. Post-hoc methods interpret the decisions of already-trained models. Notable examples include SHAP (which uses Shapley values for consistent attribution), LIME (which approximates local behavior using interpretable surrogates), and Anchors (which derive precise decision rules for local regions) [9–11]. In contrast, ante-hoc methods are inherently interpretable by design. EBM, a generalized additive model with boosted feature interactions, and TABNET, a neural network with attention-driven feature selection, exemplify this class [12, 13].

XAI for Tabular Healthcare Data

Despite the growing adoption of XAI in healthcare, most applications focus on imaging or time-series data. Often more structured and standard in clinical records, tabular datasets pose distinct challenges, including missing values, class imbalance, and high dimensionality [14]. Moreover, many published studies rely on proprietary datasets, limiting reproducibility and generalization [1].

Several works have investigated XAI methods for tabular data. Some focus on comparative analysis: Hatwell et al. [15] evaluated Ada-WHIPS, Anchors, and LORE; Settouti et al. [16] benchmarked SHAP, LIME, and permutation importance; and Alfeo et al. [17] proposed BoSCaR as a more robust alternative to SHAP for physiological data.

Other studies adopt multiple explainers to triangulate results. Chadaga et al. [18] and Islam et al. [19] used SHAP and LIME to identify key predictors for COVID-19 severity and C-section delivery, respectively. Chun et al. [20], Khanna et al. [21], and Joseph et al. [22] cross-validated explanations from multiple methods against known clinical risk factors. Aiosa et al. [23] integrated SHAP with graph-based methods to explore chronic disease associations, while Peng et al. [24] used SHAP and partial dependence plots to interpret hepatobiliary outcomes.

These examples highlight the increasing use of XAI in clinical analytics. However, most rely on case-based interpretation rather than systematic evaluation across multiple datasets or metrics.

Gaps in Metric-Based Comparative Evaluations

Comparative evaluations of XAI methods remain underdeveloped, especially for tabular healthcare models [5, 25]. As also pointed out by Di Martino and Delmastro [26], there is a lack of comprehensive sets of evaluations in healthcare settings when it comes to tabular and time-series data. Moreover, existing studies often lack unified benchmarks or rely on qualitative assessments, making it difficult to compare methods across settings [1].

Metrics such as fidelity, simplicity, consistency, robustness, precision, and coverage offer principled ways to assess explanation quality [27, 28]. For example, fidelity evaluates how well an explainer approximates model predictions [29], while simplicity focuses on explanation length and interpretability [30]. Robustness assesses explanation stability under input noise [31], and consistency gauges agreement across explainers [32]. Anchors-style rule-based methods are often evaluated using precision and coverage, which quantify rule accuracy and applicability scope [11].

Despite their importance, these metrics are rarely applied consistently across methods and datasets. This study addresses that gap by implementing and empirically applying a suite of formally defined metrics to five widely used XAI methods across four publicly available healthcare tabular datasets. By releasing the evaluation framework as open-source code, our goal is to support reproducibility and to guide the evidence-based selection of XAI methods for clinical machine learning.

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.

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 contained no missing values or included imputations.

1. **PDHS (C-Section)**: Contains 26,284 rows and 24 columns. Derived from national health surveys, this dataset was preprocessed to impute missing values and used to predict Caesarean section outcomes [19].

2. **Diabetic Retinopathy**: Includes 1,151 rows and 19 columns. Used to assess the risk of diabetic retinopathy progression [15].

3. **Breast Cancer**: Comprises 569 rows and 30 columns. Applied for binary classification of breast cancer diagnoses [15].

4. **ESDRPD (Early Stage Diabetes Prediction)**: Contains 520 rows and 16 columns. Targets the early detection of diabetes-related complications [22].

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.

RF[1] was chosen as the primary model for XAI evaluation due to its consistently strong performance and compatibility with TreeSHAP, enabling theoretically exact feature attributions.

XAI Methods and Evaluation Metrics

We evaluated five XAI methods, LIME, SHAP, Anchors, EBM, and TABNET, to assess the interpretability of RF predictions. These methods span both post-hoc (LIME, SHAP, Anchors) and inherently interpretable (EBM, TABNET) approaches. While LIME and SHAP provide instance-level feature importance, Anchors offers rule-based local explanations. EBM and TABNET, as ante-hoc models, natively expose their decision logic via global and local interpretability mechanisms.

To evaluate these methods, we adopted six established metrics from prior literature: **fidelity**, **simplicity**, **consistency**, **robustness**, **precision**, and **coverage**. While none of the metrics are novel, we formalized and implemented their mathematical definitions to ensure transparency and reproducibility. Table 1 outlines the applicability of each metric to the five XAI methods.

Table 1. Applicability of Evaluation Metrics to XAI Methods

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

Table 1 summarises the applicability of evaluation metrics across the five XAI methods. Fidelity applies only to post-hoc feature importance-based surrogate methods (LIME and SHAP), which aim to approximate the model's behavior. Simplicity is universally applicable and refers to the number of features or rules involved in the explanation. Consistency is evaluated for methods that produce feature importance

[1] hyperparameters were held constant across all datasets; $n_estimators = 100, n_jobs = -1, random_state = 42$.

rankings, while robustness assesses stability under input perturbation and can be calculated for all methods. Precision and coverage are specific to rule-based methods like Anchors, which define local decision rules applicable to subgroups of data instances.

Evaluating Feature Importance-Based XAI Methods

We now introduce the proposed equations for XAI metrics used to evaluate feature importance-based XAI methods. These proposed equations were inferred and defined using the philosophical definitions of these metrics from the research literature.

Fidelity: Fidelity [29] 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 probability outputs of the original model and those reconstructed by the XAI method. This captures the closeness of predicted class probabilities, typically via a 1 minus absolute error formulation.
- **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:

$$Fid_{\text{score}}(M_F) = \frac{1}{|X|} \sum_{i \in X} \left(1 - |P_{\text{model}}(x_i) - P_{M_F}(x_i)|\right) \quad (1)$$

$$Fid_{\text{decision}}(M_F) = \frac{1}{|X|} \sum_{i \in X} \mathbb{I}\left(\hat{y}_{\text{model}}(x_i) = \hat{y}_{M_F}(x_i)\right) \quad (2)$$

Where:

- X is the set of instances.
- $P_{\text{model}}(x_i)$ is the probability predicted by the original model.
- $P_{M_F}(x_i)$ is the surrogate probability from the XAI method.
- $\hat{y}$ represents class labels.
- $\mathbb{I}$ is the indicator function that returns 1 if both class predictions match.

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). Scores closer to 0.0 signify poor alignment between the explainer and the model.

Simplicity: Simplicity [30] 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 feature attribution–based XAI methods, we define simplicity as the *average number of features per instance* whose relative importance exceeds a specified threshold.

Let threshold $\in \{0.1, 0.05, 0.01\}$ be a proportion of the maximum feature importance for a given instance. Features with absolute importance above this threshold are counted toward the explanation.

Let $f_i = [f_{i1}, f_{i2}, \ldots, f_{i|F|}]$ be the vector of feature attributions for instance x_i, where $|F|$ is the total number of features. The simplicity score $S(M_F)$ for explainer M_F is defined as:

$$S(M_F) = \frac{1}{|X|} \sum_{i \in X} \sum_{j=1}^{|F|} \mathbb{I}\left(\frac{|f_{ij}|}{\max(|f_i|)} > \text{threshold} \right) \tag{3}$$

where:

- X: set of input instances, and $|X|$ denotes its cardinality.
- F: set of features.
- f_{ij}: attribution score of feature j for instance i.
- $\max(|f_i|)$: maximum absolute attribution in instance i.
- $\mathbb{I}(\cdot)$: 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 [32] 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 features under evaluation. In our experiments, we set $k = 5$, which captures the most influential portion of the feature rankings. 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_{F1} and M_{F2} is defined as:

$$C(M_{F1}, M_{F2}) \quad = \quad \frac{1}{|X|} \sum_{i \in X} \left(\frac{1}{k} \sum_{n=1}^{k} \text{Spearmanr}(M_{F1}(i,n), M_{F2}(i,n)) \right) \tag{4}$$

Where:

- X: the set of instances.
- $M_{F1}(i,n)$, $M_{F2}(i,n)$: top n features in instance i from methods M_{F1}, M_{F2}.
- Spearmanr: Spearman rank correlation function.
- k: number of top-ranked features considered (we used $k = 5$ in this study).

This metric enables pairwise comparisons of local explanations, accounting for both rank order and alignment in feature attribution. A consistency score of 1.0 indicates perfect agreement across all top-k depths; lower or negative scores reflect divergence in the prioritised features between methods.

Robustness: Robustness [31] 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 a normal distribution $\mathcal{N}(0, \sigma^2)$, where $\sigma = 0.01$. Robustness $R(M_F)$ is computed as the average absolute difference between explanation vectors before and after perturbation:

$$R(M_F) \quad = \quad \frac{1}{|X|} \sum_{i \in X} |E(M_F, x_i) - E(M_F, x_i + \text{noise})| \tag{5}$$

Where:

- X: the set of instances.
- x_i: the original input instance.
- noise $\sim \mathcal{N}(0, \sigma^2)$: element-wise perturbation.
- $E(M_F, x)$: explanation vector from method M_F for input x.

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.

Transition to Rule-Based Explainers: The evaluation metrics discussed thus far apply to feature attribution–based methods (e.g., SHAP, LIME, TABNET, EBM), which quantify feature contributions numerically. However, rule-based explainers like 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.

Evaluating Rule-Based XAI Methods

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 the accuracy of a rule in predicting the model's output on the instances it covers [11]. A high precision score means the rule reliably explains the model's decisions[2].

Coverage quantifies the proportion of the dataset for which the rule is applicable [11]. A higher coverage indicates broader generalizability, while lower coverage reflects a rule that applies only to narrow subpopulations.

[2]For example, a precision of 0.97 means that 97% of the instances are correctly classified according to the rule generated by the XAI method.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6}$$

$$\text{Coverage} = \frac{|\{x_i \in X : \text{rule matches } x_i\}|}{|X|} \tag{7}$$

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: Simplicity is defined as the number of feature conditions in a rule. Fewer conditions result in simpler and more interpretable explanations. For example:

- Rule A: "$Age > 45$ AND $BMI > 30$" $\rightarrow Simplicity = 2$
- Rule B: "$Age > 45$ AND $BMI > 30$ AND $Smoker = Yes$" $\rightarrow Simplicity = 3$

$$S(M_R, x) = |\text{Features}(M_R, x)| \tag{8}$$

Where:

- M_R: the rule-based explainer.
- x: the input instance.
- $\text{Features}(M_R, x)$: the set of features used in the rule that explains x.

A simplicity score of 1–3 may be desirable in clinical settings to support quick comprehension and decision-making.

Robustness: 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.

Following a similar spirit to robustness for feature-attribution methods [31], we measure the change in rule structure by computing the average dissimilarity between the sets of rule features generated before and after noise.

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:

$$R(M_R) = \frac{1}{|X|} \sum_{i \in X} \left(1 - \frac{|A_i \cap A'_i|}{|A_i \cup A'_i|}\right) \tag{9}$$

Where:

- X: the set of input instances.

- 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.

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.

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 train all five ML models: GBC, RF, XGB, TABNET, and EBM. Model performance was evaluated using accuracy and F1 score on the test sets.

We focused on the RF model for explainability assessment due to its consistently strong performance and compatibility with TreeSHAP. The RF predictions were interpreted using the five XAI methods, and each method's output was evaluated using the six predefined metrics: fidelity, simplicity, consistency, robustness, precision, and coverage.

This experimental setup enables a reproducible and comparative assessment of XAI methods across diverse clinical tabular datasets. It also supports standardized benchmarking that is aligned with real-world healthcare decision-support needs.

Results

Classification Performance

The classification performance of multiple models, GBC, RF, XGB, TABNET, and EBM, was evaluated across the selected datasets. Table 2 summarizes the accuracy and F1 scores for each model on each dataset. The RF model was selected for further XAI analysis due to its consistent and competitive performance across all datasets, aligning with existing literature that positions RF as a robust choice for feature importance analysis in healthcare contexts [1].

Evaluation of XAI Methods Based on Aggregation of Local Instances

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 shown in Table 3. In this context, fidelity is assessed using two approaches: "Score" and "Decision".

SHAP, using TreeSHAP, achieved perfect fidelity in approximating probability outputs across all datasets, confirming its theoretical assurance for tree-based models [9, 33]. Slight deviations in label fidelity (e.g., 0.9942 for Diabetic Retinopathy) stem from tie-breaking when class probabilities are exactly equal, these are inconsequential artifacts of classification thresholds, not fidelity failures.

Conversely, LIME consistently exhibited lower fidelity. For instance, in the case of Diabetic Retinopathy, it achieved a probability fidelity score of 0.8334 and only 0.5520 for predicted labels, hinting at unreliable local estimations. Comparable declines in

Table 2. Classification Performance of Different Models

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. Fidelity of LIME and SHAP Explanations

Dataset	LIME (Score)	SHAP (Score)	LIME (Decision)	SHAP (Decision)
PDHS (C-Section)	0.8897	**1.0000**	0.9323	**0.9997**
Diabetes Retinopathy	0.8334	**1.0000**	0.5520	**0.9942**
Breast Cancer	0.8809	**1.0000**	0.9708	**1.0000**
ESDRPD	0.7889	**1.0000**	0.9167	**1.0000**

ESDRPD (with scores of 0.7889 and decisions at 0.9167) emphasize LIME's vulnerability to intricate or nonlinear decision boundaries. These findings underscore that LIME's straightforwardness in approximating 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 shown in Table 4.

The thresholds (0.1, 0.05, and 0.01) represent different levels of feature importance cutoff used to determine which features are included in the explanation. Specifically, the threshold indicates the fraction of the maximum absolute feature value in the instance. A feature is included in the explanation if its absolute value exceeds the threshold multiplied by the maximum absolute value in the instance. A higher threshold results in fewer features being included, making the explanation simpler, while a lower threshold results in more features being included, making the explanation more detailed. The maximum absolute value of the most significant feature sets the cutoff (multiplied by the chosen threshold) for feature inclusion, resulting in varying values for different datasets[3], as shown in equation 3.

Table 4. Simplicity of XAI Methods (Average Number of Features)

Method	PDHS (C-Section)	Diabetes Retinopathy	Breast Cancer	ESDRPD
Threshold=0.1				
LIME	**1.82**	9.87	9.93	9.82
SHAP	3.86	11.44	13.86	**8.93**
EBM	7.68	20.11	28.40	20.53
TABNET	3.32	**6.10**	**6.21**	11.53
Threshold=0.05				
LIME	**3.29**	10.00	10.00	**10.00**
SHAP	8.59	14.23	17.42	12.28
EBM	16.26	27.16	43.55	27.73
TABNET	3.85	**7.34**	**6.69**	12.33
Threshold=0.01				
LIME	10.00	10.00	10.00	**10.00**
SHAP	18.09	17.10	25.46	15.37
EBM	35.96	35.63	56.08	30.62
TABNET	**4.69**	**9.88**	**7.37**	13.49

LIME consistently provided the simplest explanations, involving fewer dominant features, particularly at higher thresholds. TABNET struck a balance between detail and conciseness, often yielding shorter explanations than SHAP. EBM, by contrast, returned the most feature-rich outputs. These patterns reflect a clear trade-off: while LIME optimizes for interpretability, SHAP and TABNET balance clarity with coverage, and EBM prioritizes completeness over simplicity.

Consistency: The consistency scores between LIME, SHAP, and TABNET are presented in Table 5.

LIME and SHAP showed moderate to high agreement in feature rankings, suggesting overlapping explanation patterns. These consistency scores were computed using a stepwise comparison over the top $k = 5$ ranked features per instance, see equation 4. This approach captures agreement not just at the top rank but progressively across the most influential features, weighting higher ranks more.

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

Table 5. Consistency Between XAI Methods

Method Pair	PDHS (C-Section)	Diabetes Retinopathy	Breast Cancer	ESDRPD
LIME, SHAP	**0.73**	**0.51**	**0.58**	**0.84**
LIME, TABNET	0.10	-0.03	-0.05	-0.02
TABNET, SHAP	0.10	-0.08	-0.06	-0.03

In contrast, both LIME and SHAP diverged significantly from TABNET, whose outputs were consistently less aligned. This variation underscores the need to tailor XAI method selection: for consensus-driven interpretation, LIME or SHAP may be preferable; for exploratory or diverse perspectives, TABNET offers an alternative view on feature importance.

Robustness: The robustness scores for LIME, SHAP, EBM, and TABNET are shown in Table 6.

Table 6. Robustness of XAI Methods

Method	PDHS (C-Section)	Diabetes Retinopathy	Breast Cancer	ESDRPD
LIME	0.0472	0.0444	0.0448	0.0690
SHAP	0.0002	0.0139	0.0080	0.0002
EBM	**0.0000**	0.0022	0.0177	**0.0000**
TABNET	0.0007	**0.0012**	**0.0001**	0.0191

TABNET and EBM demonstrated the highest robustness, maintaining stable explanations despite input perturbations. In contrast, LIME and SHAP were more sensitive to noise. Robustness is particularly critical in clinical workflows, where even small data variations must not compromise interpretability.

Evaluation of Rule-Based XAI Method

Precision, Coverage, Simplicity, and Robustness: The quantitative results for Anchors are presented in Table 7, showing how its explanations vary with different precision thresholds. 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.

Across all datasets, increasing the threshold improves rule precision (often reaching => 0.99) but causes coverage to drop significantly, sometimes to as low as 1% of the data (e.g., Diabetic Retinopathy). In practice, this means that very precise rules can only explain a small subset of instances. For example, in high-stakes clinical settings like surgery planning or rare disease diagnostics, such high-precision, low-coverage rules may still be valuable for providing *high-confidence, case-specific recommendations*. Conversely, broader screening applications may require more generalizable rules, warranting slightly lower thresholds (e.g., 0.9–0.95) that cover more patients, albeit with some trade-off in precision.

As thresholds rise, simplicity declines, rules become longer, involving more feature conditions to maintain stricter decision boundaries. This impacts interpretability and clinician usability. For example, PDHS rules grow from 2.3 to 3.6 features as the threshold increases from 0.9 to 0.99. Robustness also slightly declines[4], suggesting that

[4] as the value increases

Table 7. Precision and Coverage of Anchors Across Datasets

Dataset	Threshold	Precision	Coverage	Simplicity	Robustness
PDHS	0.9	0.96	0.29	2.31	0.38
(C-Section)	0.95	0.97	0.24	2.79	0.40
	0.99	0.99	0.19	3.64	0.47
Diabetes	0.9	0.91	0.02	10.94	0.58
Retinopathy	0.95	0.94	0.01	13.86	0.57
	0.99	0.96	0.01	17.09	0.57
Breast	0.9	0.99	0.22	2.02	0.70
Cancer	0.95	0.99	0.22	2.06	0.70
	0.99	1.00	0.18	2.39	0.72
	0.9	0.93	0.36	2.15	0.38
ESDRPD	0.95	0.99	0.21	3.30	0.47
	0.99	1.00	0.19	3.66	0.47

more complex rules are marginally more sensitive to small input perturbations.

These patterns reinforce a key trade-off in rule-based explainability: *precision vs coverage vs interpretability*. Anchors performs best in settings where narrow, reliable rules are acceptable, but is less suitable when broad applicability or robustness under noise is required. Unlike LIME and SHAP, which provide widespread scope, Anchors offers human-readable, actionable rules, albeit for a smaller subset of cases.

Clinical Relevance of Global Features Across XAI Methods

To assess whether the explanations produced by global XAI methods align with established domain knowledge, we examined feature importance outputs from EBM, SHAP, and TABNET across all four healthcare datasets. The goal is to determine whether the highlighted features correspond to known clinical risk factors, thus reinforcing the credibility and utility of these methods in medical decision-making. 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*.

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 [19, 34, 35].

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.

This coherence reinforces known clinical markers for diabetic retinopathy, where microaneurysms and retinal exudates serve as primary diagnostic indicators [36, 37].

Breast Cancer Dataset

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

These features are widely accepted in breast cancer literature as strong predictors of malignancy, offering interpretability that resonates with established oncological criteria [38, 39].

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.

These symptoms correspond to hallmark indicators of early diabetes diagnosis and neurological complications in diabetic progression [40–42].

These findings confirm that different XAI methods, EBM, SHAP, and TABNET, identify clinically meaningful features that align with established medical knowledge across datasets. 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. 2–5).

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.

Method-Level Interpretations and Trade-offs

Fidelity: SHAP achieved perfect fidelity in approximating the probability outputs of Random Forest models, and near-perfect performance for class prediction labels. These results empirically confirm TreeSHAP's theoretical assurances [9, 33]. Minor deviations, such as 0.9942 in Diabetic Retinopathy, occurred when class probabilities were equal, leading to random tie-breaking. LIME consistently showed lower fidelity, particularly in Diabetic Retinopathy (score: 0.8334, predict: 0.5520), due to its reliance on locally approximated surrogate models.

Simplicity: LIME produced the most concise explanations across thresholds, which can be useful in clinical scenarios requiring immediate human interpretability. Anchors also generated simple, rule-based explanations, although its coverage was limited. EBM yielded a more complex but inherently interpretable model, while SHAP and TABNET offered intermediate simplicity, with explanations that balance detail and interpretability.

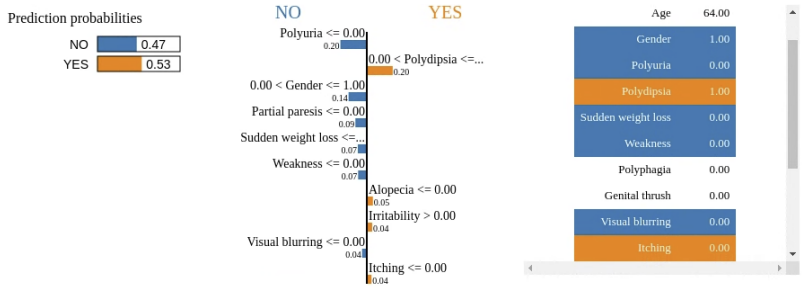
Consistency: SHAP and LIME showed moderate to high agreement in top feature rankings, especially in datasets like ESDRPD. TABNET, however, diverged significantly from both, highlighting a difference in how its attention-based mechanism attributes importance, suggesting it may surface novel or domain-specific patterns that other methods overlook.

Robustness: EBM and TABNET exhibited strong robustness to input perturbations, maintaining consistent feature importance outputs across noise injections. In contrast, LIME and Anchors were more sensitive. Figure 1 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.

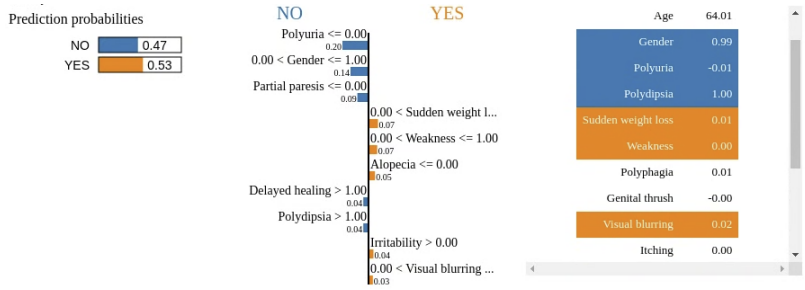
In parallel, Anchors generated different rule sets for the same instance under noise:

- **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.



(a) Original LIME Explanation



(b) Noisy LIME Explanation

Fig 1. Comparison of Original and Noisy LIME Explanations for an instance from the ESDRPD Dataset

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 in global model auditing. This trade-off, between narrow, high-confidence rules and

broader applicability, should be carefully considered in any domain requiring interpretable decision-making.

Best Practices and Practical Recommendations

Based on our evaluation, we recommend the following XAI method selection guidelines tailored to typical healthcare scenarios:

- **For fidelity-critical applications (e.g., probability calibration):** Use SHAP with tree-based models 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 LIME or Anchors. These are suitable in resource-constrained or community settings, such as mobile-based screening tools for diabetic retinopathy or 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. 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, such as auditing medical eligibility criteria, drug contraindication alerts, or subgroup-specific recommendations where interpretability outweighs global applicability.

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 consistently surfaced features aligned with established clinical risk factors, such as maternal health indicators for C-sections, tumor morphology in breast cancer, and diabetic symptoms like polyuria and polydipsia. These alignments, detailed in the Results section, reinforce the clinical plausibility of the model explanations and strengthen their case for deployment in decision-support settings. Conversely, lower-fidelity or less robust methods, such as LIME and Anchors, occasionally excluded known predictors, highlighting the need for careful explainer selection guided by appropriate evaluation metrics. This finding extends beyond medicine: in any domain where subject-matter knowledge exists, explanations can and should be validated against expert understanding. Our framework allows for such validation by surfacing interpretable, ranked, and consistent feature attributions.

Limitations and Future Work

While the analysis covers representative healthcare tabular datasets, it does not extend to time series, imaging, or longitudinal data domains that are increasingly relevant in

clinical AI. Additionally, while six quantitative metrics were used, qualitative aspects such as user trust, cognitive load, and domain-specific validation remain outside the scope.

Future research should:

- Extend this framework to multimodal datasets and temporal health records.
- Conduct clinician-facing user studies on explanation usefulness.
- Integrate additional XAI approaches, such as counterfactuals or example-based explanations.
- Explore dynamic explainer ensembles that combine global and local reasoning.

These steps will help translate explainability benchmarks into actionable clinical AI guidance.

Conclusion

This study presents a reproducible, metric-driven evaluation of five widely used explainable AI (XAI) methods, LIME, SHAP, Anchors, EBM, and TABNET, applied to healthcare tabular datasets. Using six established evaluation metrics (fidelity, simplicity, consistency, robustness, precision, and coverage), we formalized each mathematically and implemented them in open-source code to support transparency and standardized benchmarking.

Our findings reinforce known strengths and limitations: SHAP (via TreeSHAP) delivers perfect fidelity for tree-based models; LIME provides simple, interpretable explanations with reduced fidelity; EBM and TABNET offer robust, noise-tolerant insights; and Anchors generates high-precision, rule-based explanations with narrow coverage.

This work enables evidence-based XAI method selection, allowing practitioners to align explainability tools with task-specific requirements. For instance:

- **High-fidelity needs:** Use SHAP with Random Forest or XGBoost.
- **Human-centered simplicity:** Use LIME for quick, interpretable outputs.
- **Robustness under noisy data:** Prefer EBM or TABNET in safety-critical settings.
- **Targeted rule-based insight:** Apply Anchors when precision in subpopulations is key.

While our experiments focus on healthcare, the proposed framework and evaluation metrics are not domain-specific. These can be readily applied to any machine learning application involving tabular data, such as finance, logistics, fraud detection, and policy analysis. The emphasis on transparency, reproducibility, and quantitative trade-offs makes the framework suitable for both research and operational settings.

In future work, we aim to extend this framework to additional data modalities (e.g., time series, imaging), incorporate qualitative assessments from domain experts, and explore integrated explainer strategies that combine global and local perspectives.

By supporting principled evaluation of explainability tools, this study contributes toward more trustworthy, interpretable, and context-aware AI systems in healthcare and beyond..

Acknowledgments

This publication has emanated from research conducted with the support of Research Ireland under award nos. GOIPG/2022/660 & GOIPG/2021/1354 and grant number 18/CRT/6183 from Research Ireland Centre for Research Training in Machine Learning (ML-Labs) at Technological University Dublin with the financial support of Research Ireland under grant no. 13/RC/2106_P2 at the ADAPT Research Centre at Technological University Dublin.

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.

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. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

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

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

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

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

24. 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.

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

26. 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.

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

28. 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`.

29. 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.

30. 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.

31. 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.

32. 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 Computing Surveys. 2023;55(13s):1–42.

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

34. 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.

35. 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.

36. 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.

37. 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.

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

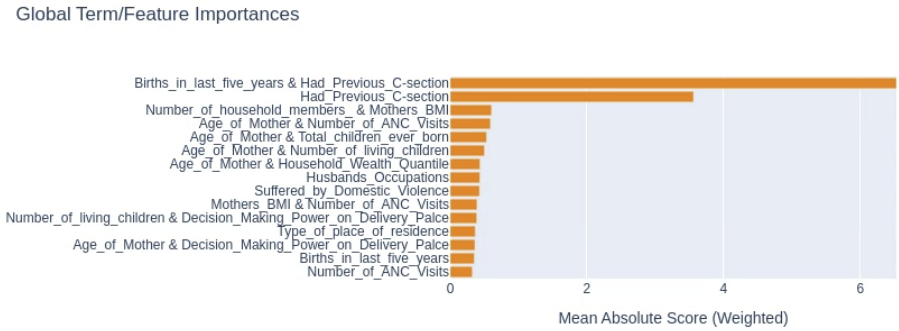
39. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In: Biomedical image processing and biomedical visualization. vol. 1905. SPIE; 1993. p. 861–870.

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

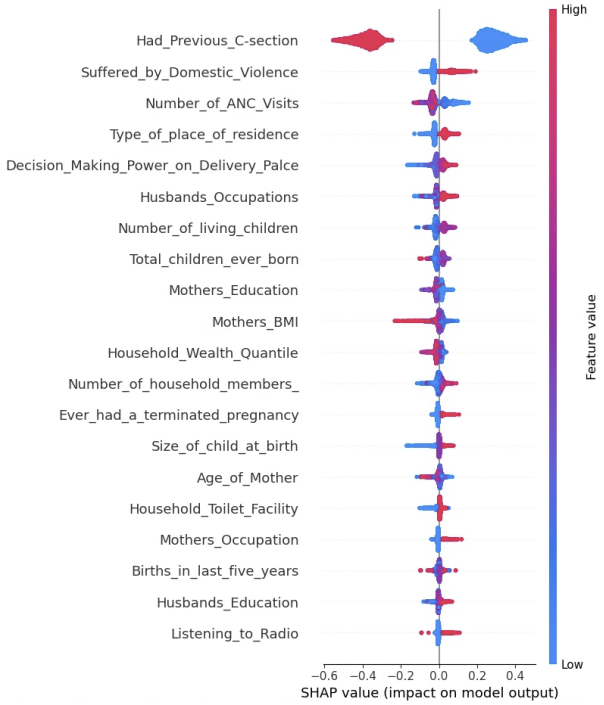
41. Konen JC, Curtis LG, Summerson JH. Symptoms and complications of adult diabetic patients in a family practice. Archives of Family Medicine. 1996;5(3):135.

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

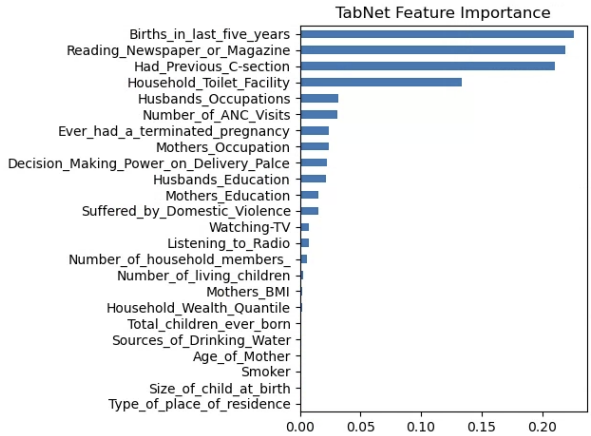
Appendix



(a) Global Feature Importance using EBM

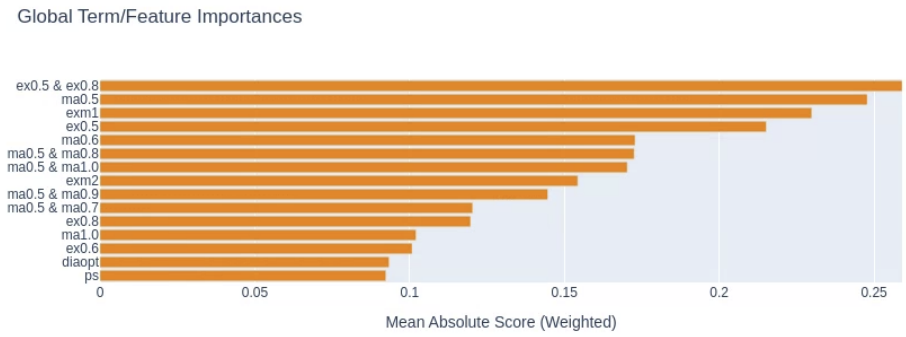


(b) Global Feature Importance using SHAP

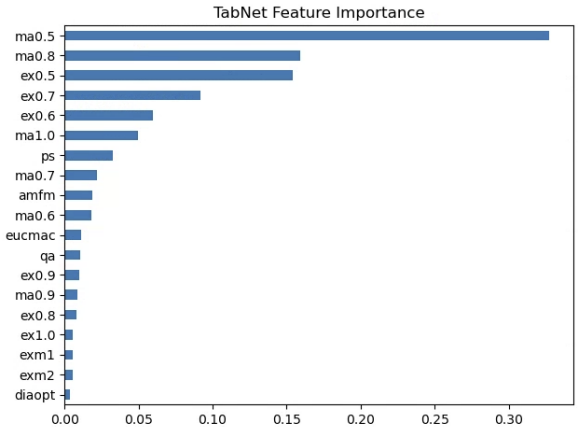


(c) Global Feature Importance using TABNET

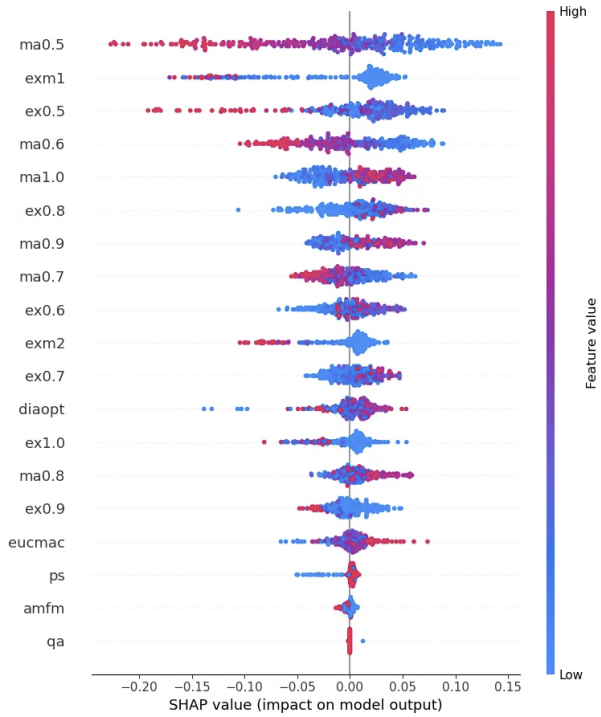
Fig 2. Global Feature Importance for PDHS (C-Section) Dataset using EBM, SHAP, and TABNET.



(a) Global Feature Importance using EBM

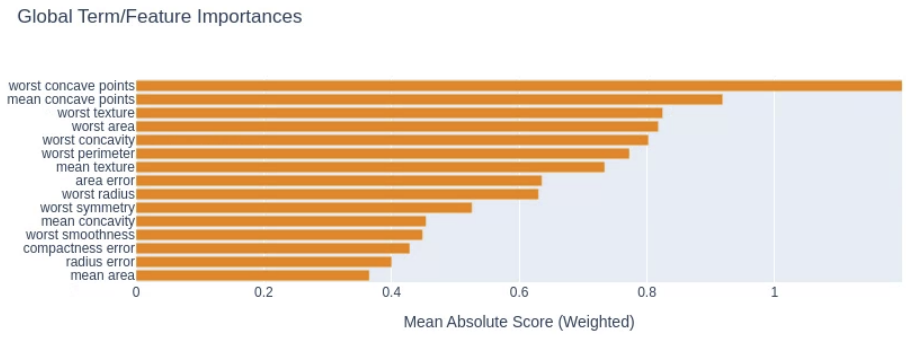


(b) Global Feature Importance using TABNET

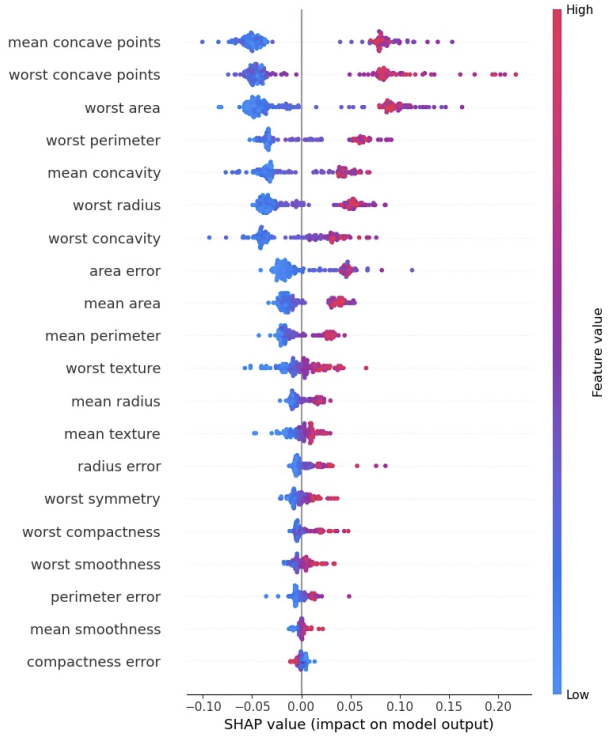


(c) Global Feature Importance using SHAP

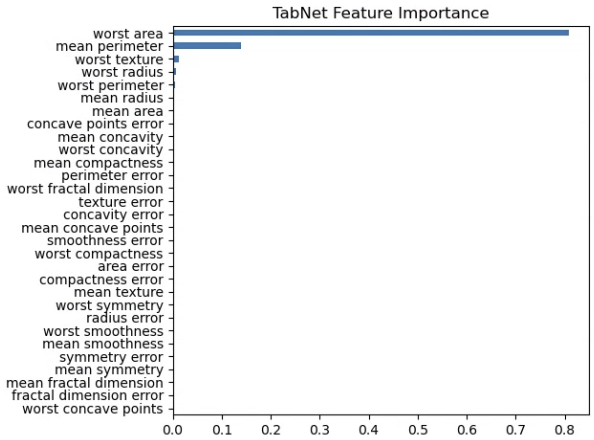
Fig 3. Global Feature Importance Analysis for Diabetes Retinopathy Dataset using EBM, SHAP, and TABNET.



(a) EBM Feature Importance

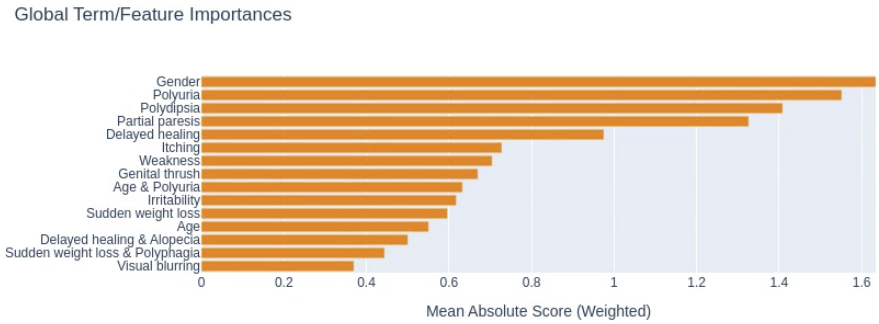


(b) SHAP Feature Importance

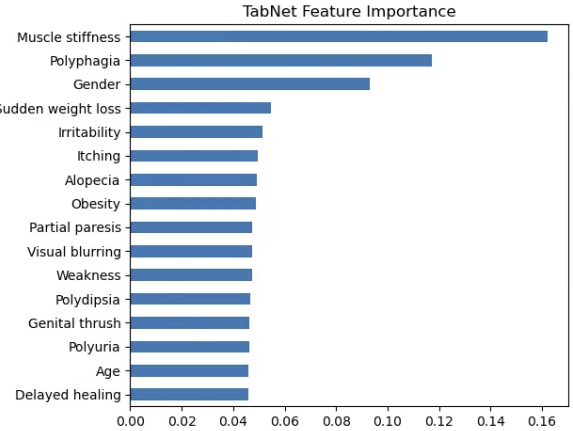


(c) TABNET Feature Importance

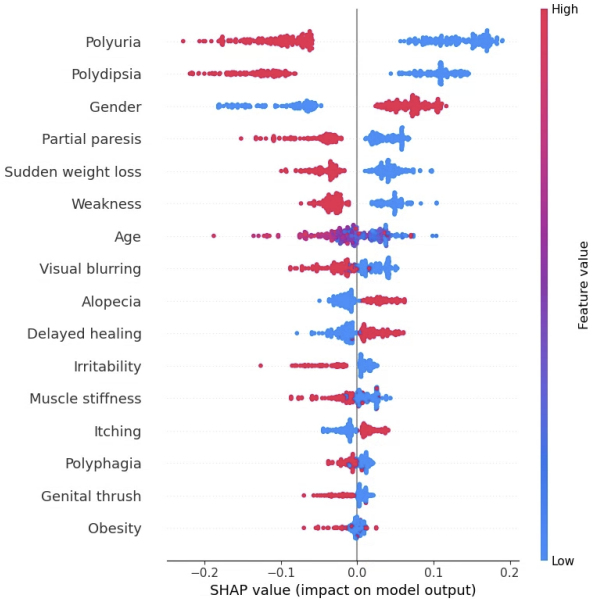
Fig 4. Global Feature Importance Analysis for Breast Cancer Dataset using EBM, SHAP, and TABNET.



(a) Global Feature Importance using EBM



(b) Global Feature Importance using TABNET



(c) Global Feature Importance using SHAP

Fig 5. Global Feature Importance Analysis for ESDRPD Dataset using EBM, SHAP, and TABNET.