

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE: A MULTI-CRITERIA BENCHMARK OF EXPLANATION FAITHFULNESS, STABILITY AND CONSENSUS IN CLINICAL RISK PREDICTION

ARULMURUGAN R.*, TESHOME D.A.

*Arulmurugan Ramu - PhD, associate professor, Department of information technology university, Mattu University, Mattu, Ethiopia

E-mail: arulmr@gmail.com, <https://orcid.org/0000-0003-1030-9565>

Teshome Debushe Adugna – M. Sc, head of department IT, Mattu University, College of engineering and technology department of information technology, Mattu, Ethiopia

E-mail: teshome.debushe@meu.edu.et, <https://orcid.org/0000-0002-4316-8474>

Abstract. Explainable artificial intelligence (XAI) is routinely proposed as the mechanism that will make clinical machine learning trustworthy, yet the explanations themselves are almost never validated: in the applied healthcare literature a SHAP or LIME plot is presented as evidence rather than tested as a claim. This study asks whether post-hoc explanations of clinical risk models are faithful, stable and reproducible, and whether these properties can be measured routinely. We propose CLARA (Clinical Explanation Reliability Assessment), a four-axis evaluation protocol that scores an explainer on faithfulness (area over the perturbation curve for progressive feature deletion and insertion, benchmarked against a random-attribution control), stability (top-k set overlap under input perturbation), consensus (rank agreement between independent explainers) and computational cost, and aggregates these into an Explanation Quality Index. CLARA was applied to three public clinical cohorts (Cleveland heart disease, $n = 297$; Pima diabetes, $n = 768$; Wisconsin diagnostic breast cancer, $n = 569$), four model families (logistic regression, random forest, gradient boosting, multilayer perceptron) and two local explainers (SHAP and a local surrogate following the LIME formulation), producing 14,400 instance-level explanations across 24 configurations. Both explainers were far more faithful than random feature orderings, and SHAP was significantly more faithful in six of twelve model-dataset cells (Wilcoxon signed-rank, $p < 0.01$) while the surrogate prevailed in two. The main finding is that agreement between explainers is governed by the model rather than by the explainer: mean rank correlation fell from 0.91 for logistic regression to 0.54 for gradient boosting, whereas hold-out AUROC was uncorrelated with explanation quality ($\rho = 0.08$, $p = 0.70$). Faithfulness and consensus should therefore be reported alongside discrimination whenever an explained model is proposed for clinical use.

Key words: explainable artificial intelligence, clinical decision support, SHAP, LIME, explanation faithfulness, explanation stability, machine learning in medicine, model evaluation.

Introduction

The transition of machine learning from retrospective analysis to deployed clinical decision support has been rapid. Risk scores for sepsis, unplanned readmission, cardiovascular events and malignancy are now embedded in hospital information systems, and the models behind them are increasingly ensembles of trees or neural networks rather than the regression equations that clinicians were trained to read. The regulatory environment has adjusted accordingly. The European Union's Artificial Intelligence Act classifies medical decision-support software as a high-risk system and requires that its operation be sufficiently transparent for deployers to interpret and use the output appropriately [1]. Comparable expectations are emerging in the digital-health strategies of Kazakhstan and other states that are building national health data platforms. In practice this means that a clinician is asked to act on a number produced by a model whose internal reasoning is not directly inspectable.

The dominant technical response has been post-hoc explanation. Rather than restricting the model class, one trains the most accurate model available and then attaches a separate procedure that attributes the prediction to individual input features. SHapley Additive exPlanations (SHAP) [2] and Local Interpretable Model-agnostic Explanations (LIME) [3] have become the de facto standard for tabular clinical data, with gradient-based saliency and class-activation mapping playing the same role in medical imaging [4; 5]. Systematic reviews of the field confirm this concentration: Loh et al. surveyed ninety-nine high-quality healthcare studies and found SHAP, LIME and Grad-CAM to

account for the large majority of applications [6], and a comparable distribution is reported by Antoniadi et al. for clinical decision support specifically [7].

A critical literature has grown alongside this practice. Rudin argues that for high-stakes decisions the post-hoc approach is misconceived, because an explanation that is not the computation being explained can, by construction, be wrong [8]. Ghassemi, Oakden-Rayner and Beam describe current explainability in health care as a false hope, noting that local explanations are frequently unstable and that clinicians tend to accept them uncritically [9]. Slack et al. demonstrate that a deliberately biased classifier can be constructed whose LIME and SHAP explanations conceal the bias entirely [10], and Alvarez-Melis and Jaakkola show that widely used interpretability methods are not robust: inputs that are imperceptibly different receive materially different explanations [11]. DeGrave, Janizek and Lee provide a clinical illustration, showing that models for radiographic COVID-19 detection latch onto acquisition artefacts rather than pathology, a failure that surfaced only when someone examined the explanations instead of displaying them [12].

The two positions leave a methodological gap, and that gap motivates this work. In the applied clinical literature an explanation is almost always presented rather than tested. A SHAP summary plot is included, its top features are observed to be clinically plausible, and plausibility is treated as validation. Yet plausibility and faithfulness are different properties: an attribution can agree with medical intuition while bearing little relation to what the model actually computes, which is the failure mode that makes post-hoc explanation risky. Recent reviews make this omission explicit. Abbas, Jeong and Lee, in a meta-analysis of sixty-two clinical decision-support studies published between 2018 and 2025, find almost no work that assessed explanation fidelity, clinician trust or usability in realistic settings [13]. Zhang et al. reach the same conclusion and call for an integrated evaluation system balancing technical fidelity against clinical utility [14].

This article addresses that gap with four contributions. First, we define CLARA (Clinical Explanation Reliability Assessment), a compact four-axis protocol that measures faithfulness, stability, cross-method consensus and computational cost of any local feature-attribution method applied to a clinical risk model, and combines them into a single Explanation Quality Index (EQI). Second, we execute the protocol on a full factorial design of three public clinical cohorts, four model families and two local explainers, generating 14,400 instance-level explanations across twenty-four configurations, each benchmarked against a random-attribution control that is routinely omitted in applied work. Third, we report an empirical result with direct practical consequences: the agreement between two independent explainers is governed by the model being explained rather than by the explainers themselves, falling from a mean rank correlation of 0.91 for logistic regression to 0.54 for gradient boosting, while predictive discrimination carries no information about explanation quality. Fourth, we argue from these measurements that faithfulness and consensus belong in the reporting checklist for any clinical model whose deployment is justified by its explainability, and we release the complete experimental code to make such reporting inexpensive.

Foundations of post-hoc attribution. The modern attribution literature begins with two model-agnostic proposals. LIME approximates the decision surface of an arbitrary classifier in the neighbourhood of a single instance by fitting a sparse linear surrogate to perturbed samples weighted by proximity, and reports the surrogate coefficients as the explanation [3]. SHAP reformulates attribution as a cooperative-game problem, assigning each feature the Shapley value of its marginal contribution to the prediction and thereby obtaining the unique attribution satisfying local accuracy, missingness and consistency [2]. The two are related, in that kernel SHAP is a particular weighting of the LIME regression, but they differ in what they estimate, so their disagreement carries information about the model. For tree ensembles, Lundberg et al. derived a polynomial-time exact algorithm and extended it to global summaries, making Shapley attribution practical for the gradient-boosted models that dominate tabular clinical prediction [15]. Parallel work in deep learning produced gradient-based attributions with axiomatic guarantees [4] and class-activation mapping for convolutional networks [5], which together dominate the imaging branch of clinical XAI [16].

Explainable AI in clinical practice. Surveys of the healthcare application literature are numerous and consistent in their findings. Payrovnaziri et al. scoped explainable models built on real-

world electronic health record data and observed a strong preference for post-hoc, model-agnostic attribution over inherently interpretable design [17]. Markus, Kors and Rijnbeek set out a detailed terminology and design taxonomy, distinguishing explainability from interpretability and cataloguing evaluation strategies, while noting that quantitative evaluation is rarely performed in practice [18]. Antoniadi et al. found that tabular clinical decision-support systems are the most common XAI application and that user studies with clinicians are almost entirely absent [7]. In imaging, van der Velden et al. catalogue interpretability methods for deep networks and reach a similar conclusion about the scarcity of evaluation [16]. More recent syntheses shift the emphasis from methods to integration, documenting how technically sound explanations fail when they are misaligned with clinical workflow. Van Dort et al., reviewing sixteen qualitative studies of hospital professionals, report that SHAP plots were frequently found difficult to interpret and that clinicians used explanations chiefly to validate model outputs [19].

The case against explanation-as-reassurance. Rudin's position is that in high-stakes domains one should train an interpretable model rather than explain an opaque one, since a post-hoc explanation is by definition an approximation of a different function [8]. The empirical support for this concern is substantial. Alvarez-Melis and Jaakkola quantify the local Lipschitz instability of saliency and surrogate methods [11]; Slack et al. construct scaffolded classifiers whose LIME and SHAP explanations are arbitrary [10]; Caruana et al. showed a decade ago, in a pneumonia risk model, that a genuinely intelligible model exposes dangerous learned patterns that an opaque model conceals [20]. Amann et al., analysing a deployed cardiac-arrest decision-support system, conclude that whether explainability adds value depends on the clinical context, the validation status of the explanation and the intended role of the system in the decision, and cannot be assumed [21]. Ghassemi et al. sharpen the argument to a recommendation: explanations should not be used as a substitute for rigorous prospective validation of the model itself [9].

Quantitative evaluation of explanations. A smaller methodological literature has developed the measurements this article uses. Doshi-Velez and Kim proposed the standard taxonomy of functionally-grounded, human-grounded and application-grounded evaluation, and argued that functionally-grounded proxies must come first because they are cheap and falsifiable [22]. Perturbation-based faithfulness metrics operationalise this: deletion and insertion curves, introduced for image explanations by Petsiuk, Das and Saenko, progressively remove or restore the features an explanation ranks highest and record the response of the model [23]. Hooker et al. showed that naive deletion confounds attribution quality with distribution shift and proposed retraining-based benchmarking as a remedy [24]. Comparative studies applying such metrics to clinical tabular data remain rare, and where they exist they typically evaluate one model family, omit a random-attribution control, or report faithfulness without stability. Table 1 positions the present study against representative prior work along these axes.

Table 1. Positioning of the present study against representative prior work on the evaluation of feature attributions

Study	Data domain	Models compared	Explainers compared	Faithful-ness	Stability	Consensus / random control
Petsiuk et al. [23]	Natural images	1 CNN	RISE, Grad-CAM, LIME	Yes	No	No / no
Hooker et al. [24]	Natural images	CNNs	4 attribution methods	Yes	No	No / yes
Alvarez-Melis, Jaakkola [11]	Tabular, text	Several	LIME, SHAP, saliency	No	Yes	No / no
Slack et al. [10]	Tabular (fairness)	Adversarial classifier	LIME, SHAP	No	No	Partial / no
Loh et al. [6] (review)	Healthcare	Many	Many	Narrative	No	No / no
Antoniadi et	Clinical DSS	Many	Many	Narrative	No	No / no

al. [7] (review)						
Abbas et al. [13] (meta-analysis)	Clinical DSS	Many	Many	Reported as a gap	No	No / no
This study	3 clinical cohorts	4 families	SHAP, LIME-type surrogate	Yes	Yes	Yes / yes

Two gaps follow from this comparison. Faithfulness, stability, consensus and cost are almost never measured together, although a method that is faithful but unstable and a method that is stable but unfaithful fail a clinical deployment in different ways and call for different remedies. The more consequential gap is that the literature treats explanation quality as a property of the explainer, comparing SHAP with LIME as though the answer were a single ranking, when the model being explained may be the dominant factor. The design described below was built to test that possibility directly.

Materials and methods of research

Study design. The study is a full factorial computational experiment with three factors: clinical cohort (three levels), predictive model family (four levels) and local explanation method (two levels), giving twenty-four configurations. Every configuration is scored on the four CLARA axes and on the composite index. The overall design is shown in Figure 1. Stage 1 prepares the cohorts; Stage 2 fits and audits the risk models; Stage 3 generates explanations; Stage 4 computes the evaluation metrics that are aggregated into the Explanation Quality Index.

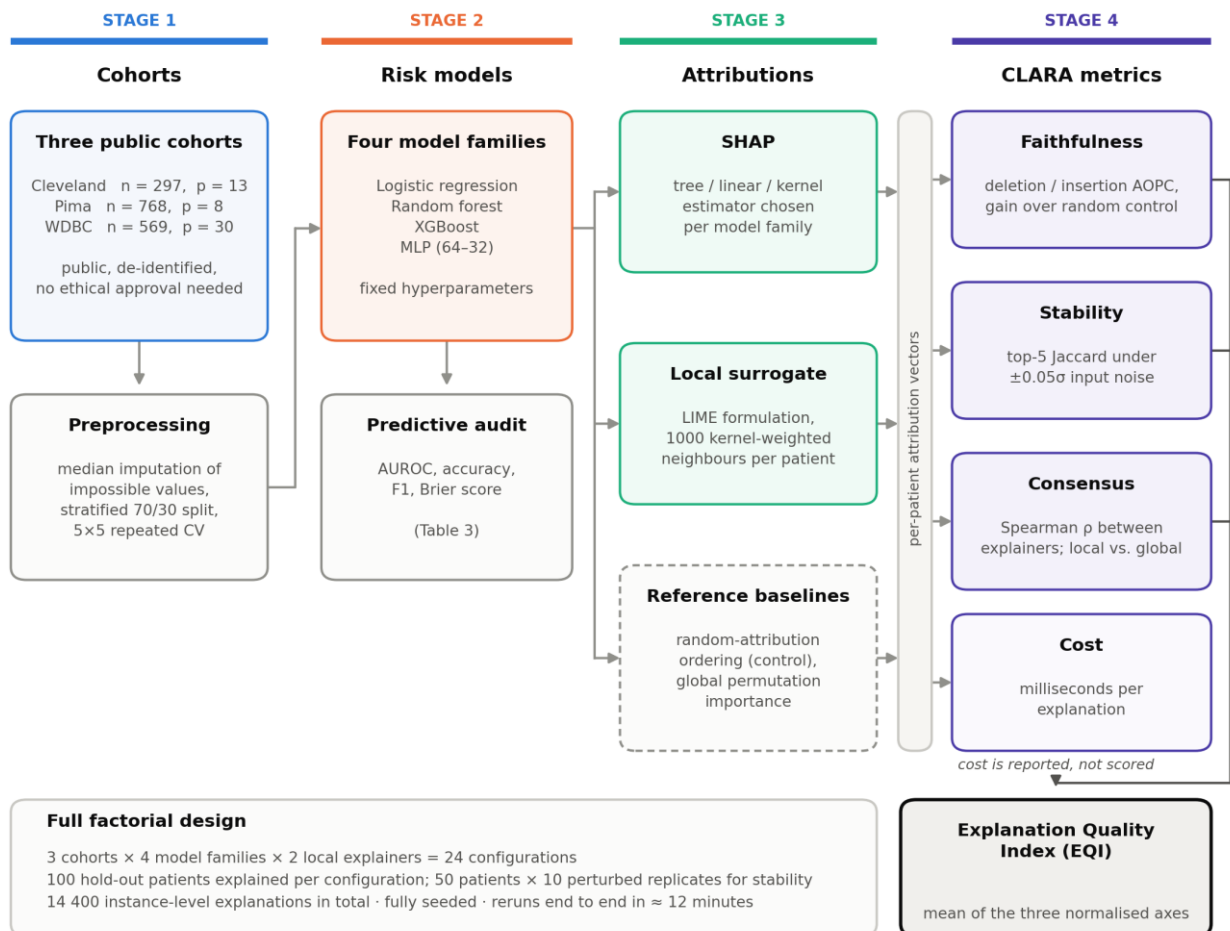


Figure 1. Research design of the CLARA evaluation protocol. Every model family is explained by both local methods and by the two reference baselines; the Explanation Quality Index aggregates the faithfulness, stability and consensus axes only, while computational cost is reported separately

Data. Three public, de-identified clinical datasets were used, chosen because they span the range of dimensionality and difficulty typical of tabular clinical risk problems and because their continued public availability makes the study reproducible without data-access approvals. The Cleveland subset of the UCI heart disease database contains 303 records collected at the Cleveland Clinic Foundation and described by Detrano et al. [25]; six records with missing values were removed, leaving 297 patients with thirteen attributes and a binary target formed by dichotomising the angiographic disease status. The Pima Indians diabetes dataset, introduced with the ADAP learning algorithm by Smith et al. [26], contains 768 records of female patients of Pima heritage aged at least twenty-one, with eight attributes and a binary diagnosis of diabetes within five years. The Wisconsin Diagnostic Breast Cancer dataset, derived from digitised fine-needle aspirates by Street, Wolberg and Mangasarian [27], contains 569 samples described by thirty morphometric features, with malignancy as the positive class. Table 2 summarises the three cohorts.

Table 2. Characteristics of the three clinical cohorts

Cohort	Source	n	p	Positive class	Prevalence	Missing-value handling
Heart disease (Cleveland)	UCI, Cleveland Clinic Foundation [25]	297	13	Angiographic disease present	46.1 %	6 incomplete records removed
Diabetes (Pima)	NIDDK / UCI [26]	768	8	Diabetes within five years	34.9 %	Impossible zeros replaced by the median
Breast cancer (WDBC)	UCI, University of Wisconsin [27]	569	30	Malignant tumour	37.3 %	None required

Preprocessing. In the Pima cohort, values of zero in plasma glucose, diastolic blood pressure, triceps skinfold thickness, serum insulin and body mass index are physiologically impossible and encode missingness; these were replaced by the column median, which is the standard treatment for this dataset. No other imputation was performed. Features were left in their original units for the tree ensembles and standardised inside the model pipeline for logistic regression and the neural network, so that scaling is fitted on training folds only and cannot leak information across the split. For all explanation experiments the data were divided once into a stratified 70/30 training and hold-out partition with a fixed random seed; predictive performance was estimated separately by five-times-repeated stratified five-fold cross-validation on the complete cohort.

Predictive models. Four model families were selected to span the interpretability spectrum encountered in clinical practice. Logistic regression with L2 regularisation is the transparent baseline that clinical risk scores are traditionally built on. Random forest [28], with 300 trees and a minimum of two samples per leaf, covers bagged ensembles. Gradient boosting, implemented with XGBoost [29] using 300 trees of maximum depth four, a learning rate of 0.08, subsampling of 0.9 and L2 regularisation of 1.0, is the model class that currently wins most tabular clinical prediction tasks. A multilayer perceptron with two hidden layers of sixty-four and thirty-two units, early stopping and a maximum of 2,000 iterations stands for dense neural networks. Hyperparameters were fixed a priori at conventional values rather than tuned, because the object of study is the explanation of a model rather than the model's optimality; tuning each family per cohort would have confounded explanation quality with search effort.

Explanation methods. Two local attribution methods were compared. SHAP was computed with the estimator appropriate to each model family: the exact interventional tree estimator for the random forest and the gradient-boosted ensemble, the linear estimator for logistic regression, and kernel SHAP with a k-means-summarised background of twenty reference points and 150 coalition samples for the neural network [2; 15]. The second method is a local surrogate explainer (LSE) implementing the LIME formulation for tabular data [3]: for an instance x , one thousand neighbours are sampled from a Gaussian distribution centred on x with the per-feature training standard deviation, each neighbour is weighted by an exponential kernel of its standardised distance from x with width

$0.75 \cdot \sqrt{p}$, and a weighted ridge regression is fitted to the model's predicted probability. The reported attribution is the product of the surrogate coefficient and the standardised deviation of the instance from the training mean, so that the surrogate decomposes additively in the same contribution units as SHAP and the two are directly comparable. Permutation feature importance computed on the hold-out set with twenty repeats and the area under the receiver operating characteristic curve as the scoring function provides a global reference against which the aggregated local attributions are compared.

Faithfulness. Faithfulness asks whether the features an explanation ranks highest are in fact the features the model uses. Let $f(x)$ denote the model's predicted probability for the class it assigns to x , let π be the ordering of features by decreasing absolute attribution, and let x_k denote x with the k highest-ranked features replaced by their training-set median. The area over the perturbation curve for deletion over K steps is

$$AOPC_{del} = (1/K) \cdot \sum_{k=1 \dots K} [f(x) - f(x_k)], \quad (1)$$

and the corresponding insertion measure starts from the all-median reference vector and restores the k highest-ranked features,

$$AOPC_{ins} = (1/K) \cdot \sum_{k=1 \dots K} [f(r_k) - f(r)], \quad (2)$$

where r is the all-median reference vector and r_k is r with the k highest-ranked features of x restored. K was set to $\min(p, 10)$. Because the magnitude of both quantities depends on the cohort and the model rather than on the explanation alone, every configuration was additionally evaluated with attributions drawn from a standard normal distribution, averaged over five random draws, and we report the faithfulness gain, defined as the difference between the deletion score of the explanation and that of the random control. Without that control, a large deletion score shows only that removing features changes the prediction, not that the explanation picked the right features to remove.

Stability. An explanation that changes materially when the input changes imperceptibly cannot support a clinical decision [11]. For each of fifty hold-out instances per configuration, ten perturbed copies were generated by adding Gaussian noise with standard deviation equal to 0.05 of each feature's training standard deviation, and the explanation was recomputed for each copy. Stability is reported as the mean Jaccard index of the top- k feature sets,

$$J = (1/M) \cdot \sum_{m=1 \dots M} |S_k(x) \cap S_k(x_m)| / |S_k(x) \cup S_k(x_m)|, \quad (3)$$

with $k = 5$ and $M = 10$, complemented by the mean Spearman correlation between the full attribution vectors of the original and perturbed instances.

Consensus. If two attribution methods with different assumptions converge on the same explanation of the same prediction, a clinician has grounds to treat the explanation as a property of the model; if they diverge, the explanation is a property of the method. Consensus is measured per instance as the Spearman rank correlation between the absolute attribution vectors of the two explainers,

$$\rho_{cons} = (1/N) \cdot \sum_{i=1 \dots N} \rho_S(|a_{SHAP}(x_i)|, |a_{LSE}(x_i)|), \quad (4)$$

with the top-5 Jaccard overlap and the signed rank correlation reported alongside it. Cost is the mean wall-clock time required to produce one explanation.

Composite index. To summarise a configuration in one number, the faithfulness gain g , the stability index J and the consensus correlation ρ_{cons} were each min-max normalised across all twenty-four configurations by the operator $n(\cdot)$ and averaged with equal weights,

$$EQI = (1/3) \cdot [n(g) + n(J) + n(\rho_{cons})]. \quad (5)$$

Equal weighting is a deliberate simplification: the index is intended for ranking configurations

within a study, not as an absolute score, and any deployment would weight the axes according to whether the explanation is used for auditing, for bedside reasoning or for regulatory documentation.

Statistical analysis. Predictive performance is reported as the mean and standard deviation over the twenty-five cross-validation folds. Faithfulness of the two explainers is compared per configuration with the Wilcoxon signed-rank test on paired per-instance deletion scores, with the rank-biserial correlation as the effect size; no correction across the twelve model-cohort cells is applied, and the exact p-values are reported so that any correction can be applied by the reader. Relationships between predictive discrimination, model complexity and explanation quality are quantified with Spearman correlations at the level of the twenty-four configurations.

Implementation and reproducibility. All experiments were run in Python 3.11 with scikit-learn 1.8 [30], XGBoost 3.2 [29], SHAP 0.51 [15], NumPy 2.4 and SciPy 1.17 on a two-core Linux container. Every random operation is seeded, and the complete pipeline (data preparation, model fitting, explanation, evaluation and figure generation) runs end to end in under fifteen minutes, so reporting these metrics routinely costs very little. The datasets are obtained directly from their public repositories at run time; no proprietary or patient-identifiable data were used, and no ethical approval was required.

Results and its discussion

Predictive performance. Table 3 reports discrimination and calibration for the twelve model-cohort combinations. The pattern is the one familiar from tabular clinical data: the differences between model families are small, and the transparent baseline is not outperformed. Logistic regression attains the highest or joint-highest AUROC on all three cohorts (0.898 on Cleveland, 0.836 on Pima, 0.995 on WDBC), and gradient boosting, the most complex tabular learner in the comparison, is the weakest on two of the three. This matters for the interpretation of everything that follows: in these cohorts the opaque models are not purchasing accuracy with their opacity, so any deficit in explanation quality that they incur is not compensated elsewhere.

Table 3. Predictive performance, five-times-repeated stratified five-fold cross-validation (mean $\pm$ standard deviation)

Cohort	Model	Accuracy	AUROC	F1-score	Brier score
Heart disease (Cleveland)	LR	0.830 $\pm$ 0.047	0.898 $\pm$ 0.044	0.808 $\pm$ 0.059	0.126
	RF	0.826 $\pm$ 0.051	0.902 $\pm$ 0.050	0.802 $\pm$ 0.067	0.127
	XGB	0.803 $\pm$ 0.052	0.878 $\pm$ 0.050	0.781 $\pm$ 0.062	0.154
	MLP	0.805 $\pm$ 0.053	0.883 $\pm$ 0.042	0.781 $\pm$ 0.060	0.148
Diabetes (Pima)	LR	0.768 $\pm$ 0.033	0.836 $\pm$ 0.030	0.626 $\pm$ 0.059	0.157
	RF	0.768 $\pm$ 0.028	0.836 $\pm$ 0.028	0.647 $\pm$ 0.046	0.156
	XGB	0.749 $\pm$ 0.032	0.812 $\pm$ 0.031	0.633 $\pm$ 0.049	0.181
	MLP	0.750 $\pm$ 0.038	0.826 $\pm$ 0.055	0.571 $\pm$ 0.112	0.167
Breast cancer (WDBC)	LR	0.978 $\pm$ 0.012	0.995 $\pm$ 0.004	0.970 $\pm$ 0.018	0.020
	RF	0.956 $\pm$ 0.020	0.991 $\pm$ 0.008	0.941 $\pm$ 0.028	0.033
	XGB	0.967 $\pm$ 0.014	0.994 $\pm$ 0.004	0.956 $\pm$ 0.019	0.025
	MLP	0.926 $\pm$ 0.032	0.983 $\pm$ 0.011	0.892 $\pm$ 0.052	0.062

Faithfulness. Figure 2 shows deletion curves for the gradient-boosted model on all three cohorts. Both explainers separate clearly from the random-attribution control. That comparison is the minimum evidence needed before an attribution is shown to a clinician, and most applied papers omit it. The separation is largest on WDBC, where removing the ten features SHAP ranks highest reduces the predicted class probability by 0.33 while a random ordering of the same length reduces it by 0.06. Complete results for all twenty-four configurations are given in Table 4. Averaged over

configurations, SHAP achieves a faithfulness gain of 0.105 against 0.084 for the local surrogate, and a higher insertion score (0.219 against 0.204).

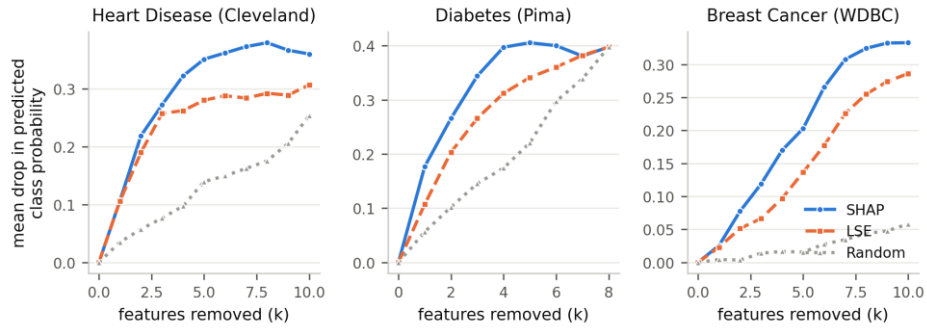


Figure 2. Deletion curves for the gradient-boosted model on the three cohorts. The mean drop in the predicted class probability is shown as the k highest-ranked features are progressively replaced by their training median

Table 4. CLARA results for all twenty-four configurations. AOPCdel and AOPCrand are the deletion scores of the explanation and of the random control; the gain is their difference. Stability is the mean top-5 Jaccard index; consensus is the mean Spearman rank correlation between the two explainers.

Cohort	Model	Expl.	AOPCdel	AOPCrand	Gain	AOPCins	Stability	Consensus	EQI
Cleveland	LR	SHAP	0.239	0.129	0.110	0.270	0.958	0.958	0.833
		LSE	0.238	0.129	0.109	0.269	0.833	0.958	0.726
	RF	SHAP	0.161	0.087	0.074	0.181	0.927	0.657	0.569
		LSE	0.141	0.087	0.054	0.156	0.697	0.657	0.340
	XGB	SHAP	0.311	0.151	0.160	0.299	0.741	0.378	0.410
		LSE	0.256	0.151	0.104	0.255	0.726	0.378	0.295
Pima	MLP	SHAP	0.133	0.073	0.060	0.133	0.914	0.852	0.644
		LSE	0.129	0.073	0.055	0.127	0.932	0.852	0.650
	LR	SHAP	0.171	0.103	0.067	0.172	0.971	0.940	0.755
		LSE	0.172	0.103	0.069	0.173	0.935	0.940	0.729
	RF	SHAP	0.178	0.104	0.074	0.212	0.899	0.704	0.573
		LSE	0.171	0.104	0.067	0.202	0.857	0.704	0.525
WDBC	XGB	SHAP	0.346	0.192	0.155	0.392	0.781	0.516	0.513
		LSE	0.296	0.192	0.105	0.362	0.834	0.516	0.466
	MLP	SHAP	0.045	0.032	0.013	0.040	0.938	0.672	0.475
		LSE	0.053	0.032	0.021	0.047	0.916	0.672	0.470
	LR	SHAP	0.116	0.017	0.099	0.272	0.901	0.828	0.690
		LSE	0.102	0.017	0.085	0.265	0.572	0.828	0.390
	RF	SHAP	0.185	0.049	0.136	0.219	0.950	0.736	0.746
		LSE	0.155	0.049	0.106	0.194	0.730	0.736	0.507
	XGB	SHAP	0.216	0.021	0.195	0.263	0.951	0.718	0.846
		LSE	0.159	0.021	0.139	0.218	0.622	0.718	0.467
	MLP	SHAP	0.141	0.031	0.110	0.178	0.575	0.663	0.344
		LSE	0.128	0.031	0.098	0.174	0.843	0.663	0.545

The paired tests in Table 5 show that this aggregate advantage is neither universal nor an artefact of averaging. SHAP is significantly more faithful in six of the twelve cells, with the largest effect on WDBC with gradient boosting (rank-biserial 0.72, $p = 5.0 \cdot 10^{-11}$) and on Cleveland with gradient boosting and random forest (0.47 and 0.44, $p < 10^{-4}$). The local surrogate is significantly more faithful in two cells, most clearly for the neural network on Pima (rank-biserial -0.48 , $p = 3.1 \cdot 10^{-7}$). In the

remaining four cells, three of which involve logistic regression, the two methods are statistically indistinguishable, as they should be: for a linear model both estimate the same quantity. The choice of explainer therefore matters most for boosted ensembles and neural networks, which are also the models hardest to reason about, and no single method dominates. For the neural network on Pima, the widely recommended default would have been the less faithful choice.

Table 5. Paired comparison of explanation faithfulness (Wilcoxon signed-rank test on per-instance deletion scores, $n = 100$ instances per cell). A positive rank-biserial correlation favours SHAP

Cohort	Model	AOPCdel SHAP	AOPCdel LSE	W	p-value	Rank-biserial r
Cleveland	LR	0.239	0.238	1988	0.953	-0.01
	RF	0.161	0.141	1033	$4.5 \cdot 10^{-5}$	+0.44
	XGB	0.311	0.256	912	$4.9 \cdot 10^{-6}$	+0.47
	MLP	0.133	0.129	1610	0.078	+0.20
Pima	LR	0.171	0.172	781	0.017	-0.16
	RF	0.178	0.171	2161	0.273	-0.01
	XGB	0.346	0.296	1259	$1.3 \cdot 10^{-5}$	+0.36
	MLP	0.045	0.053	1037	$3.1 \cdot 10^{-7}$	-0.48
WDBC	LR	0.116	0.102	2484	0.888	-0.10
	RF	0.185	0.155	1295	$2.4 \cdot 10^{-5}$	+0.34
	XGB	0.216	0.159	614	$5.0 \cdot 10^{-11}$	+0.72
	MLP	0.141	0.128	1612	$1.7 \cdot 10^{-3}$	+0.18

Stability and cost. The centre panel of Figure 3 shows that SHAP is on average more stable than the local surrogate under input perturbation (mean top-5 Jaccard 0.876 against 0.791), but again with an instructive exception: for the neural network the surrogate is the more stable of the two (0.897 against 0.809), and for logistic regression on the thirty-feature WDBC cohort the surrogate degrades sharply to 0.572, reflecting the sampling variance of a local regression fitted in a high-dimensional neighbourhood. Costs differ by two orders of magnitude across configurations. Tree-structured SHAP is effectively free, at 1.1 to 23.7 ms per explanation, whereas the surrogate requires one thousand model evaluations per instance and reaches 87 to 91 ms per explanation for the random forest. For a bedside decision-support system generating one explanation per patient this is irrelevant; for a batch audit of a year of admissions it turns a job of minutes into one of hours.

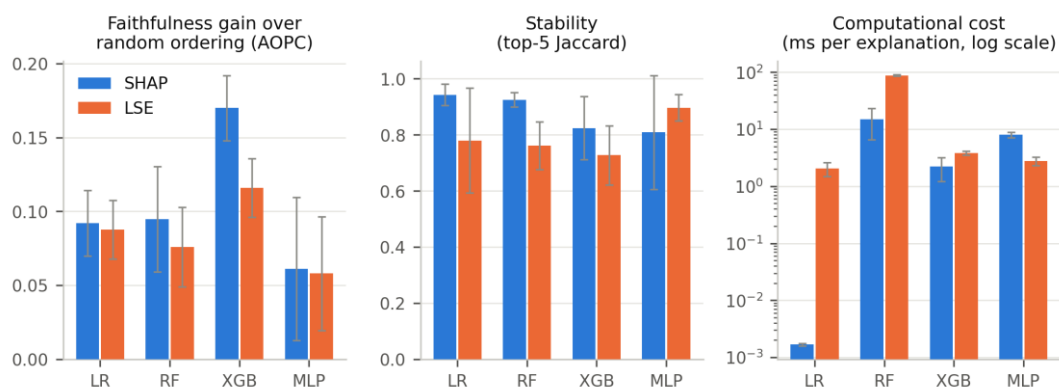


Figure 3. Faithfulness gain, stability and computational cost by model family and explanation method, averaged over the three cohorts (error bars: standard deviation across cohorts)

Consensus is a property of the model, not of the explainer. Figure 4a and the consensus columns of Table 4 carry the main result of the study. Averaged over cohorts, the two explainers agree almost perfectly on logistic regression (mean Spearman $\rho = 0.908$), moderately on the neural network (0.729) and the random forest (0.699), and poorly on gradient boosting (0.538, with a minimum of 0.378 on

Cleveland). Across the twenty-four configurations, a simple ordinal complexity proxy correlates negatively with consensus ($\rho = -0.444$, $p = 0.030$). The top-5 overlap tells the same story in a form directly relevant to bedside display: for logistic regression the two methods agree on 78% of the five features a clinician would be shown, but for gradient boosting only on 52%. For the model class most likely to be deployed, then, about half of the features shown as the reason for a patient's risk score depend on which explanation library the vendor chose.

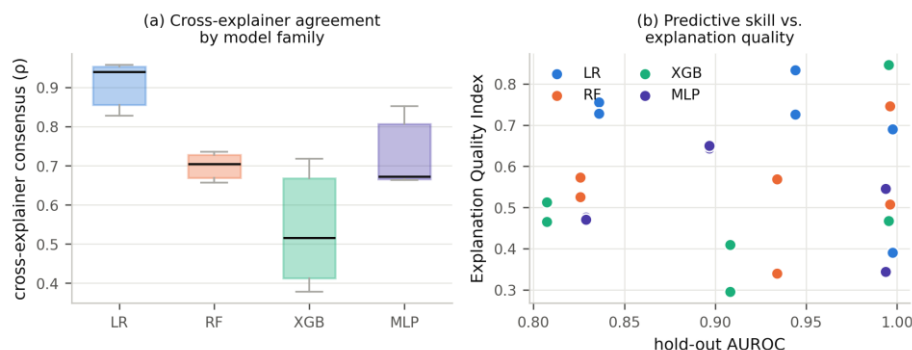


Figure 4. (a) Cross-explainer consensus by model family; (b) hold-out discrimination against the Explanation Quality Index for all twenty-four configurations

The finding changes how the usual comparison between methods should be read. The question 'is SHAP better than LIME?' has no cohort-independent answer, but the question 'how much does the answer matter for this model?' does, and it is answerable cheaply. A high consensus score licenses treating the explanation as a property of the model; a low one is a warning that the explanation is partly an artefact of the estimator, and that a second method should be run before the output is shown to a clinician. The agreement between the aggregated local attributions and global permutation importance (Table 6) declines along the same axis, from about 0.86 for logistic regression to about 0.62 for gradient boosting, indicating that the local and global views of the same opaque model also diverge.

Table 6. Agreement between explanation methods and between local and global attributions, averaged over the three cohorts

Model	Consensus, Spearman ρ	Consensus, top-5 Jaccard	PFI vs SHAP, ρ	PFI vs LSE, ρ
Logistic regression	0.908	0.780	0.868	0.845
Random forest	0.699	0.632	0.687	0.672
Gradient boosting	0.538	0.524	0.599	0.647
Multilayer perceptron	0.729	0.662	0.678	0.677

Predictive skill does not predict explanation quality. Figure 4b plots the composite index against hold-out discrimination for all twenty-four configurations. The relationship is absent (Spearman $\rho = 0.084$, $p = 0.697$). Faithfulness gain alone is weakly and positively associated with AUROC ($\rho = 0.405$, $p = 0.050$). That is intelligible: a model that discriminates strongly has a sharper response surface, so removing genuinely important features moves the prediction further. Stability ($\rho = -0.206$, $p = 0.335$) and consensus ($\rho = 0.287$, $p = 0.174$) show no such association. A model selected on AUROC alone, which is standard clinical practice, therefore carries no guarantee about the quality of the explanations that will accompany it, and those explanations cannot be treated as a free by-product of good modelling.

Clinical plausibility of the attributions. Figure 5 shows the global SHAP importances of the gradient-boosted model on each cohort. The rankings are clinically coherent: thallium stress-test result, chest-pain type and the number of major vessels coloured by fluoroscopy dominate the Cleveland model; plasma glucose concentration, body mass index and the diabetes pedigree function

dominate the Pima model; and the WDBC model is driven by concavity and perimeter descriptors of the cell nuclei, consistent with the morphometric basis of the original diagnostic system [27]. Applied papers usually rest their claim to explainability on an observation of this kind. The results above show why it is not enough. The same models whose global attributions look clinically sensible produce local explanations on which two standard methods agree on barely half of the displayed features, and plausibility of the aggregate is compatible with substantial unreliability of the individual explanation that a clinician actually sees.

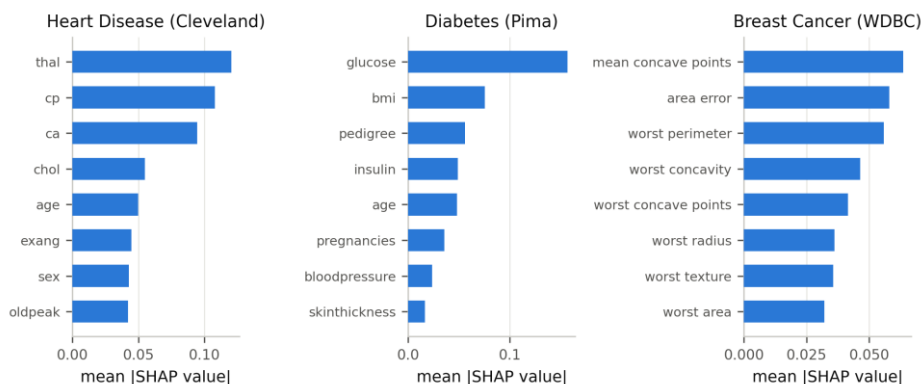


Figure 5. Global SHAP feature importances of the gradient-boosted model: the eight highest-ranked features in each cohort

Relation to prior work. Our results support the concerns raised by Rudin [8] and by Ghassemi et al. [9] without endorsing the strongest form of either. Post-hoc explanations were not arbitrary: on every configuration both methods were substantially more faithful than random attributions, which contradicts the reading that they carry no signal. But their reliability degrades where transparency is most needed, as the instability results of Alvarez-Melis and Jaakkola [11] and the adversarial constructions of Slack et al. [10] predict. Our observation that logistic regression is simultaneously the most accurate, the most stable and the most consensual model in these cohorts is a concrete instance of Rudin's argument that in many tabular high-stakes problems the accuracy sacrificed by choosing an interpretable model is negligible. The finding also gives a measurable form to the complaint recorded by Van Dort et al., that clinicians find SHAP output difficult to interpret at the bedside [19]. Part of that difficulty is substantive rather than presentational, because the explanations of opaque models really are less reproducible.

Implications for practice and policy. Three recommendations follow from these measurements. Any clinical model whose deployment is justified by its explainability should report a faithfulness gain against a random-attribution control, in the same way that a classifier reports discrimination against chance, at a computational cost of a few seconds per model. Consensus between two independent explainers should also be reported for the model, not for the explainer, and a low value should be treated as a contraindication for presenting single-method explanations to clinicians. And where an interpretable model matches the opaque alternative on discrimination, as it did on all three cohorts here, the interpretable model should be preferred, since it removes the explanation-reliability question instead of answering it. These recommendations are compatible with the transparency obligations of the EU Artificial Intelligence Act [1], which require that outputs be interpretable by deployers but say nothing about how the interpretation itself should be validated.

Limitations. Four constraints bound these conclusions. First, the cohorts are small, tabular and well-curated; high-dimensional, temporally structured electronic health record data may behave differently, and the deletion metric is known to be confounded with distribution shift when features are replaced by out-of-distribution values, a problem that retraining-based protocols address at much higher cost [24]. Second, the comparison covers two local attribution families; counterfactual, rule-based, concept-based and inherently interpretable additive models were not evaluated. Third, faithfulness, stability and consensus are functionally-grounded proxies in the sense of Doshi-Velez

and Kim [22]; they establish necessary conditions for a trustworthy explanation but not sufficient ones, since an explanation may be faithful and stable and still be useless to a clinician, which only a human-grounded study can determine. Fourth, the equal weighting of the composite index is a convention rather than a finding, and the ranking it produces should be read as a summary of the three underlying measurements rather than as a score in its own right.

Conclusion

This study asked whether the post-hoc explanations that justify the clinical deployment of opaque machine learning models can be trusted, and whether their trustworthiness can be measured as routinely as predictive accuracy. Applying the proposed CLARA protocol to twenty-four configurations spanning three public clinical cohorts, four model families and two local explainers, and generating 14,400 instance-level explanations, produced three findings. Post-hoc attributions do carry real signal about model behaviour: in every configuration both explainers were far more faithful than random feature orderings, and SHAP was the more faithful method in six of twelve model-cohort cells, though the local surrogate prevailed in two, so no method can be recommended unconditionally. Explanation reliability is governed by the model rather than by the explanation method: agreement between two independent explainers fell from a mean rank correlation of 0.91 for logistic regression to 0.54 for gradient boosting, and the two methods agreed on only 52% of the top-five features for the boosted model, meaning that for the model class most likely to be deployed the reasons displayed to a clinician depend substantially on the software chosen. Predictive performance carries no information about explanation quality ($\rho = 0.084$, $p = 0.697$), so a model selected on discrimination alone offers no guarantee about the explanations that accompany it.

The practical conclusion is that explainability should be treated as a measurable property to be reported, not as an attribute conferred by attaching an explanation library. Faithfulness against a random control and cross-method consensus are inexpensive to compute, since the whole protocol runs in minutes on a two-core machine, and they belong in the model card of any clinical decision-support system. Where an interpretable model matches an opaque one on discrimination, as logistic regression did on all three cohorts studied here, it should be preferred, because the explanation-reliability problem then does not arise at all. Future work will extend the protocol to longitudinal electronic health record data, to counterfactual and concept-based explanation families, and to human-grounded evaluation with clinicians, in order to test whether the configurations that score well on these functional axes are also the ones that support better clinical decisions.

References

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [Electronic resource] // Official Journal of the European Union. – OJ L, 2024/1689, 12.07.2024. – Access mode: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> (date of access: 01.09.2026).
2. Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems 30 (NeurIPS 2017). – Long Beach, 2017. – P. 4765-4774.
3. Ribeiro M.T., Singh S., Guestrin C. «Why should I trust you?»: Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – San Francisco, 2016. – P. 1135-1144. <https://doi.org/10.1145/2939672.2939778>
4. Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks // Proceedings of the 34th International Conference on Machine Learning, PMLR. – Sydney, 2017. – Vol. 70. – P. 3319-3328. URL: <https://proceedings.mlr.press/v70/sundararajan17a.html>
5. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – Venice, 2017. – P. 618-626. <https://doi.org/10.1109/ICCV.2017.74>
6. Loh H.W., Ooi C.P., Seoni S., Barua P.D., Molinari F., Acharya U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022).

Computer Methods and Programs in Biomedicine, 2022. Vol. 226. 107161. <https://doi.org/10.1016/j.cmpb.2022.107161>

7. Antoniadis A.M., Du Y., Guendouz Y., Wei L., Mazo C., Becker B.A., Mooney C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Applied Sciences, 2021. Vol. 11. Issue 11. 5088. <https://doi.org/10.3390/app11115088>

8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019. Vol. 1. Issue 5. 206-215. <https://doi.org/10.1038/s42256-019-0048-x>

9. Ghassemi M., Oakden-Rayner L., Beam A.L. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 2021. Vol. 3. Issue 11. e745-e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)

10. Slack D., Hilgard S., Jia E., Singh S., Lakkaraju H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. – New York, 2020. – P. 180-186. <https://doi.org/10.1145/3375627.3375830>

11. Alvarez-Melis D., Jaakkola T.S. On the robustness of interpretability methods [Electronic resource]. – arXiv:1806.08049, 2018. – Access mode: <https://arxiv.org/abs/1806.08049> (date of access: 01.09.2026).

12. DeGrave A.J., Janizek J.D., Lee S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 2021. Vol. 3. 610-619. <https://doi.org/10.1038/s42256-021-00338-7>

13. Abbas Q., Jeong W., Lee S.W. Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. Healthcare, 2025. Vol. 13. Issue 17. 2154. <https://doi.org/10.3390/healthcare13172154>

14. Zhang K., Wang D., Lin F., Xie J., Zhou W. A comprehensive review of explainable artificial intelligence in healthcare: methods, evaluation, and clinical integration. iScience, 2026. Vol. 29. Issue 3. 115026. <https://doi.org/10.1016/j.isci.2026.115026>

15. Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020. Vol. 2. Issue 1. 56-67. <https://doi.org/10.1038/s42256-019-0138-9>

16. van der Velden B.H.M., Kuijff H.J., Gilhuijs K.G.A., Viergever M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 2022. Volume 79. 102470. <https://doi.org/10.1016/j.media.2022.102470>

17. Payrovnaziri S.N., Chen Z., Rengifo-Moreno P., Miller T., Bian J., Chen J.H., Liu X., He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association, 2020. Vol. 27. Issue 7. 1173-1185. <https://doi.org/10.1093/jamia/ocaa053>

18. Markus A.F., Kors J.A., Rijnbeek P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, 2021. Vol. 113. 103655. <https://doi.org/10.1016/j.jbi.2020.103655>

19. Van Dort B.A., Engelsma T., Sneekes P., Peute L., Medlock S. Explainable AI in hospital clinical decision support systems: A scoping review of healthcare professionals' perspectives. PLOS Digital Health, 2026. Vol. 5. Issue 5. e0001417. <https://doi.org/10.1371/journal.pdig.0001417>

20. Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission // Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – Sydney, 2015. – P. 1721-1730. <https://doi.org/10.1145/2783258.2788613>

21. Amann J., Vetter D., Blomberg S.N., Christensen H.C., Coffee M., Gerke S., et al. To explain or not to explain? Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health, 2022. Vol. 1. Issue 2. e0000016. <https://doi.org/10.1371/journal.pdig.0000016>

22. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning

[Electronic resource]. – arXiv:1702.08608, 2017. – Access mode: <https://arxiv.org/abs/1702.08608> (date of access: 01.09.2026).

23. Petsiuk V., Das A., Saenko K. RISE: Randomized input sampling for explanation of black-box models // Proceedings of the British Machine Vision Conference (BMVC). – Newcastle, 2018. – arXiv: URL: <https://arxiv.org/abs/1806.07421>.

24. Hooker S., Erhan D., Kindermans P.-J., Kim B. A benchmark for interpretability methods in deep neural networks // Advances in Neural Information Processing Systems 32 (NeurIPS 2019). – Vancouver, 2019. – arXiv: URL: <https://arxiv.org/abs/1806.10758>.

25. Detrano R., Janosi A., Steinbrunn W., Pfisterer M., Schmid J.-J., Sandhu S., Guppy K.H., Lee S., Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 1989. Vol. 64. Issue 5. 304-310. [https://doi.org/10.1016/0002-9149\(89\)90524-9](https://doi.org/10.1016/0002-9149(89)90524-9)

26. Smith J.W., Everhart J.E., Dickson W.C., Knowler W.C., Johannes R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus // Proceedings of the Annual Symposium on Computer Application in Medical Care. – Washington, 1988. – P. 261-265. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2245318/>

27. Street W.N., Wolberg W.H., Mangasarian O.L. Nuclear feature extraction for breast tumor diagnosis // IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology, Biomedical Image Processing and Biomedical Visualization. – San Jose, 1993. – Vol. 1905. – P. 861-870. <https://doi.org/10.1117/12.148698>

28. Breiman L. Random forests. Machine Learning, 2001. Vol. 45. Issue 1. 5-32. <https://doi.org/10.1023/A:1010933404324>

29. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – San Francisco, 2016. – P. 785-794. <https://doi.org/10.1145/2939672.2939785>

30. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011. Vol. 12. P. 2825-2830. URL: <https://www.jmlr.org/papers/v12/pedregosa11a.html>

Әдебиеттер тізімі

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [Electronic resource] // Official Journal of the European Union. – OJ L, 2024/1689, 12.07.2024. – Access mode: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> (date of access: 01.09.2026).

2. Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems 30 (NeurIPS 2017). – Long Beach, 2017. – P. 4765-4774.

3. Ribeiro M.T., Singh S., Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – San Francisco, 2016. – P. 1135-1144. <https://doi.org/10.1145/2939672.2939778>

4. Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks // Proceedings of the 34th International Conference on Machine Learning, PMLR. – Sydney, 2017. – Vol. 70. – P. 3319-3328. URL: <https://proceedings.mlr.press/v70/sundararajan17a.html>

5. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – Venice, 2017. – P. 618-626. <https://doi.org/10.1109/ICCV.2017.74>

6. Loh H.W., Ooi C.P., Seoni S., Barua P.D., Molinari F., Acharya U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). Computer Methods and Programs in Biomedicine, 2022. Vol. 226. 107161. <https://doi.org/10.1016/j.cmpb.2022.107161>

7. Antoniadi A.M., Du Y., Guendouz Y., Wei L., Mazo C., Becker B.A., Mooney C. Current

challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences*, 2021. Vol. 11. Issue 11. 5088. <https://doi.org/10.3390/app11115088>

8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019. Vol. 1. Issue 5. 206-215. <https://doi.org/10.1038/s42256-019-0048-x>

9. Ghassemi M., Oakden-Rayner L., Beam A.L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 2021. Vol. 3. Issue 11. e745-e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)

10. Slack D., Hilgard S., Jia E., Singh S., Lakkaraju H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods // *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. – New York, 2020. – P. 180-186. <https://doi.org/10.1145/3375627.3375830>

11. Alvarez-Melis D., Jaakkola T.S. On the robustness of interpretability methods [Electronic resource]. – arXiv:1806.08049, 2018. – Access mode: <https://arxiv.org/abs/1806.08049> (date of access: 01.09.2026).

12. DeGrave A.J., Janizek J.D., Lee S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021. Vol. 3. 610-619. <https://doi.org/10.1038/s42256-021-00338-7>

13. Abbas Q., Jeong W., Lee S.W. Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare*, 2025. Vol. 13. Issue 17. 2154. <https://doi.org/10.3390/healthcare13172154>

14. Zhang K., Wang D., Lin F., Xie J., Zhou W. A comprehensive review of explainable artificial intelligence in healthcare: methods, evaluation, and clinical integration. *iScience*, 2026. Vol. 29. Issue 3. 115026. <https://doi.org/10.1016/j.isci.2026.115026>

15. Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2020. Vol. 2. Issue 1. 56-67. <https://doi.org/10.1038/s42256-019-0138-9>

16. van der Velden B.H.M., Kuijff H.J., Gilhuijs K.G.A., Viergever M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 2022. Volume 79. 102470. <https://doi.org/10.1016/j.media.2022.102470>

17. Payrovnaziri S.N., Chen Z., Rengifo-Moreno P., Miller T., Bian J., Chen J.H., Liu X., He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 2020. Vol. 27. Issue 7. 1173-1185. <https://doi.org/10.1093/jamia/ocaa053>

18. Markus A.F., Kors J.A., Rijnbeek P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 2021. Vol. 113. 103655. <https://doi.org/10.1016/j.jbi.2020.103655>

19. Van Dort B.A., Engelsma T., Sneekes P., Peute L., Medlock S. Explainable AI in hospital clinical decision support systems: A scoping review of healthcare professionals' perspectives. *PLOS Digital Health*, 2026. Vol. 5. Issue 5. e0001417. <https://doi.org/10.1371/journal.pdig.0001417>

20. Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission // *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. – Sydney, 2015. – P. 1721-1730. <https://doi.org/10.1145/2783258.2788613>

21. Amann J., Vetter D., Blomberg S.N., Christensen H.C., Coffee M., Gerke S., et al. To explain or not to explain? Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, 2022. Vol. 1. Issue 2. e0000016. <https://doi.org/10.1371/journal.pdig.0000016>

22. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning [Electronic resource]. – arXiv:1702.08608, 2017. – Access mode: <https://arxiv.org/abs/1702.08608> (date of access: 01.09.2026).

23. Petsiuk V., Das A., Saenko K. RISE: Randomized input sampling for explanation of black-

box models // Proceedings of the British Machine Vision Conference (BMVC). – Newcastle, 2018. – arXiv: URL: <https://arxiv.org/abs/1806.07421>.

24. Hooker S., Erhan D., Kindermans P.-J., Kim B. A benchmark for interpretability methods in deep neural networks // Advances in Neural Information Processing Systems 32 (NeurIPS 2019). – Vancouver, 2019. – arXiv: URL: <https://arxiv.org/abs/1806.10758>.

25. Detrano R., Janosi A., Steinbrunn W., Pfisterer M., Schmid J.-J., Sandhu S., Guppy K.H., Lee S., Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 1989. Vol. 64. Issue 5. 304-310. [https://doi.org/10.1016/0002-9149\(89\)90524-9](https://doi.org/10.1016/0002-9149(89)90524-9)

26. Smith J.W., Everhart J.E., Dickson W.C., Knowler W.C., Johannes R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus // Proceedings of the Annual Symposium on Computer Application in Medical Care. – Washington, 1988. – P. 261-265. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2245318/>

27. Street W.N., Wolberg W.H., Mangasarian O.L. Nuclear feature extraction for breast tumor diagnosis // IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology, Biomedical Image Processing and Biomedical Visualization. – San Jose, 1993. – Vol. 1905. – P. 861-870. <https://doi.org/10.1117/12.148698>

28. Breiman L. Random forests. Machine Learning, 2001. Vol. 45. Issue 1. 5-32. <https://doi.org/10.1023/A:1010933404324>

29. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – San Francisco, 2016. – P. 785-794. <https://doi.org/10.1145/2939672.2939785>

30. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011. Vol. 12. P. 2825-2830. URL: <https://www.jmlr.org/papers/v12/pedregosa11a.html>

ДЕНСАУЛЫҚ САҚТАУДАҒЫ ТҮСІНДІРІЛЕТІН ЖАСАНДЫ ИНТЕЛЛЕКТ: КЛИНИКАЛЫҚ ҚАУІПТІ БОЛЖАУДАҒЫ ТҮСІНДІРМЕЛЕРДІҢ ДӘЛДІГІН, ТҰРАҚТЫЛЫҒЫН ЖӘНЕ КЕЛІСІМДІЛІГІН КӨПКРИТЕРИЙЛІ БАҒАЛАУ

АРУЛМУРУГАН Р.*, ТЕШОМЕ Д.А.

*Арулмуруган Раму — PhD, қауымдастырылған профессор, Ақпараттық технологиялар кафедрасы, Матту университеті, Матту қ., Эфиопия.

E-mail: arulmr@gmail.com, <https://orcid.org/0000-0003-1030-9565>

Тешоме Дебуше Адугна – Магистр (M.Sc.), ақпараттық технологиялар кафедрасының меңгерушісі, Матту университеті, Ақпараттық технологиялар кафедрасы, Инженерия және технология колледжі, Матту қ., Эфиопия.

E-mail: teshome.debushe@meu.edu.et, <https://orcid.org/0000-0002-4316-8474>

Аңдатпа. Түсіндірілетін жасанды интеллект (ХАІ) клиникалық машиналық оқытуды сенімді ететін тетік ретінде жиі ұсынылады, алайда түсіндірмелердің өзі іс жүзінде тексерілмейді: қолданбалы медициналық әдебиетте SHAP немесе LIME графигі тексерілуге тиіс тұжырым емес, дайын дәлел ретінде беріледі. Бұл зерттеуде клиникалық қауіп модельдерінің апостериорлық түсіндірмелері дәл, тұрақты және қайталанатын бола ма және осы қасиеттерді күнделікті өлшеуге бола ма деген сұрақ қойылады. CLARA (Clinical Explanation Reliability Assessment) — түсіндіру әдісін дәлдігі (белгілерді дәйекті түрде алып тастау және қосу кезіндегі бұзылу қисығының үстіндегі аудан, кездейсоқ бақылаумен салыстырғанда), тұрақтылығы (кіріс деректері бұзылған кездегі top-k жиынтықтарының сәйкестігі), келісімділігі (тәуелсіз әдістердің рангілік келісімі) және есептеу құны бойынша бағалайтын, сосын оларды түсіндірме сапасының индексіне біріктіретін төрт өлшемді бағалау хаттамасы ұсынылды. Хаттама үш ашық клиникалық жиынтыққа (Кливленд жүрек ишемиясы, n = 297; Пима қант диабеті, n = 768; Висконсин сүт безі қатерлі ісігі, n = 569), төрт модель тобына (логистикалық регрессия, кездейсоқ орман, градиенттік бустинг, көпқабатты перцептрон) және екі жергілікті түсіндіру әдісіне (SHAP және LIME тұжырымдамасы бойынша жергілікті суррогат модель) қолданылды; 24 конфигурацияда 14 400 жеке түсіндірме алынды. Екі әдіс те белгілерді кездейсоқ реттеуден едәуір дәлірек болды; SHAP «модель — жиынтық» он екі үйлесімнің алтауында статистикалық маңызды дәлірек болды (Уилкоксон критерийі, p < 0,01), суррогат модель екеуінде басым түсті. Негізгі нәтиже: түсіндірмелердің келісімділігі әдіспен емес, модельмен анықталады — орташа рангілік келісім логистикалық регрессия үшін 0,91-ден градиенттік бустинг үшін 0,54-ке

дейін төмендейді, ал бақылау жиынтығындағы AUROC түсіндірме сапасымен байланысты емес ($\rho = 0,08$, $p = 0,70$).

Түйін сөздер: түсіндірілетін жасанды интеллект, клиникалық шешімдерді қолдау, SHAP, LIME, түсіндірме дәлдігі, түсіндірме тұрақтылығы, медицинадағы машиналық оқыту, модельдерді бағалау.

ОБЪЯСНИМЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ЗДРАВООХРАНЕНИИ: МНОГОКРИТЕРИАЛЬНАЯ ОЦЕНКА ДОСТОВЕРНОСТИ, УСТОЙЧИВОСТИ И СОГЛАСОВАННОСТИ ОБЪЯСНЕНИЙ В ЗАДАЧАХ КЛИНИЧЕСКОГО ПРОГНОЗИРОВАНИЯ РИСКА

АРУЛМУРУГАН Р.*, ТЕШОМЕ Д.А.

*Арулмуруган Раму — PhD, ассоциированный профессор, кафедра информационных технологий, Университет Матту, г. Матту, Эфиопия

E-mail: arulmr@gmail.com, <https://orcid.org/0000-0003-1030-9565>

Тешоме Дебуше Адугна – Магистр (M.Sc.), заведующий кафедрой информационных технологий, Университет Матту, кафедра информационных технологий, Колледж инженерии и технологий, Университет Матту, г. Матту, Эфиопия.

E-mail: teshome.debushe@meu.edu.et, <https://orcid.org/0000-0002-4316-8474>

Аннотация. Объяснимый искусственный интеллект (XAI) регулярно предлагается в качестве механизма, который сделает клиническое машинное обучение заслуживающим доверия, однако сами объяснения почти никогда не проверяются: в прикладной медицинской литературе график SHAP или LIME представляется как доказательство, а не проверяется как утверждение. В настоящем исследовании ставится вопрос о том, являются ли апостериорные объяснения клинических моделей риска достоверными, устойчивыми и воспроизводимыми и можно ли измерять эти свойства в рутинном порядке. Предложен протокол CLARA (Clinical Explanation Reliability Assessment) — четырёхосевая процедура оценки, которая характеризует метод объяснения по достоверности (площадь над кривой возмущения при последовательном удалении и добавлении признаков в сравнении со случайным контролем), устойчивости (совпадение множеств top-k при возмущении входных данных), согласованности (ранговое согласие независимых методов объяснения) и вычислительной стоимости, а затем объединяет их в индекс качества объяснений. Протокол применён к трём общедоступным клиническим выборкам (Кливлендская выборка ишемической болезни сердца, $n = 297$; диабет Пима, $n = 768$; Висконсинская диагностическая выборка рака молочной железы, $n = 569$), четырём семействам моделей (логистическая регрессия, случайный лес, градиентный бустинг, многослойный перцептрон) и двум локальным методам объяснения (SHAP и локальная суррогатная модель в формулировке LIME); получено 14 400 объяснений на уровне отдельных наблюдений в 24 конфигурациях. Оба метода существенно превзошли случайное упорядочение признаков; SHAP оказался достоверно точнее в шести из двенадцати сочетаний «модель — выборка» (критерий Уилкоксона, $p < 0,01$), суррогатная модель — в двух. Основной результат состоит в том, что согласованность объяснений определяется моделью, а не методом объяснения: среднее ранговое согласие снижается с 0,91 для логистической регрессии до 0,54 для градиентного бустинга, тогда как AUROC на отложенной выборке не связан с качеством объяснений ($\rho = 0,08$, $p = 0,70$).

Ключевые слова: объяснимый искусственный интеллект, поддержка клинических решений, SHAP, LIME, достоверность объяснений, устойчивость объяснений, машинное обучение в медицине, оценка моделей.