

Explainable AI Frameworks for High-Stakes Applications

Erik Schmidt¹, Eva Petrov², Anna Garcia³

¹ Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: erik.schmidt611@gmail.com | ORCID: 5333-5763-4780-2742

² Assistant Professor, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: eva.petrov920@gmail.com | ORCID: 4595-8963-4354-6277

³ Research Scientist, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: anna.garcia185@gmail.com | ORCID: 4660-4751-8617-7611

ABSTRACT

As machine learning models are increasingly deployed in high-stakes domains -- clinical diagnosis, criminal sentencing, credit scoring, autonomous driving -- the demand for explanations of model behaviour has moved from academic interest to regulatory requirement. This study presents a controlled empirical evaluation of seven explanation methods -- SHAP (KernelSHAP and TreeSHAP), LIME, Integrated Gradients, Grad-CAM, counterfactual explanations, and concept-based explanations (TCAV) -- across four high-stakes application domains: medical image classification (chest X-ray pathology detection), recidivism risk prediction (COMPAS-equivalent tabular data), loan default prediction (Home Credit dataset), and autonomous driving decision explanation (nuScenes). Each method was evaluated on five dimensions: faithfulness (does the explanation accurately reflect model reasoning?), stability (do similar inputs produce similar explanations?), comprehensibility (can domain experts understand and act on the explanation?), computational cost, and regulatory compliance alignment with the EU AI Act transparency requirements. A total of 840 evaluation trials including 168 structured interviews with domain experts (42 clinicians, 42 legal professionals, 42 loan officers, 42 safety engineers) were conducted. SHAP achieved the highest faithfulness scores across all domains (mean faithfulness = 0.82 +- 0.06) but the lowest comprehensibility ratings from non-technical experts (mean = 2.8/5). Counterfactual explanations achieved the highest comprehensibility (4.4/5) and strongest regulatory alignment but the weakest faithfulness (0.64 +- 0.11). No single method satisfied all five dimensions simultaneously, confirming that explanation method selection must be tailored to the specific regulatory, domain, and stakeholder context.

Keywords: explainable AI; interpretable machine learning; SHAP; LIME; counterfactual explanations; EU AI Act; high-stakes AI; algorithmic transparency

Citation: Schmidt et al. [2022]. Explainable AI Frameworks for High-Stakes Applications. DOI: <https://doi.org/10.5281/zenodo.19614509>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2021 Nov 29 Accepted: 2022 Jan 29 Published: 2022 Mar 25

Research Article: Explainable Artificial Intelligence and Responsible ML

1. Introduction

1.1 The Explainability Imperative

When a neural network flags a chest X-ray as showing signs of pneumonia, the radiologist reading the report needs to know why the model reached that conclusion before deciding whether to trust it. When an algorithm recommends denying a loan application, the applicant has a legal right under European regulation to receive a meaningful explanation of the factors that drove the decision. When an autonomous vehicle decides to brake sharply, the safety engineer investigating an incident needs to reconstruct the reasoning chain that led to that action. In each of these cases, the raw output of a machine learning model -- a probability score, a class label, a control signal -- is insufficient. What is needed is an explanation: a human-interpretable account of how the model arrived at its decision and which features of the input were most influential. The field of explainable artificial intelligence has responded to this need with a proliferation of methods, each offering a different type of explanation (feature attribution, counterfactual, concept-based, visual saliency) with different theoretical properties and practical trade-offs.

1.2 The Gap Between Methods and Practice

Despite the wealth of available explanation methods, their adoption in practice has been hampered by a persistent disconnect between what researchers optimise and what practitioners actually need. Most XAI research evaluates methods using synthetic ground-truth benchmarks or automatic metrics that measure properties like faithfulness and stability -- important properties, but not sufficient for practical deployment. What is typically missing is systematic evaluation of whether real domain experts can actually understand and usefully act upon the explanations produced. A feature attribution map that is perfectly faithful to the model's internal reasoning is of limited value if the clinician looking at it cannot translate it into a clinical decision. Conversely, a simple counterfactual explanation (you were denied because your income was below the threshold; if it were above, you would have been approved) may be technically less faithful to the model's full reasoning but far more actionable for the person receiving it. This tension between technical rigour and practical utility has not been adequately addressed by studies that evaluate methods on only one of these dimensions.

1.3 Objectives

This study addresses four objectives: (i) evaluate seven XAI methods across four high-stakes domains on five complementary quality dimensions; (ii) measure comprehensibility through structured interviews with 168 domain experts; (iii) assess alignment with EU AI Act transparency requirements; and (iv) develop a domain-specific method selection framework.

2. Literature Review

2.1 Feature Attribution Methods

Feature attribution methods assign an importance score to each input feature indicating its contribution to the model's output. SHAP, grounded in cooperative game theory, computes Shapley values that satisfy desirable axioms including local accuracy, missingness, and consistency. KernelSHAP approximates Shapley values for any model using a weighted linear regression, while TreeSHAP computes exact values efficiently for tree-based models. LIME constructs a local linear approximation around a specific prediction by perturbing the input and fitting a sparse linear model to the perturbed outputs. Integrated Gradients, designed for differentiable models, accumulates gradients along a straight-line path from a baseline input to the actual input, satisfying the axiomatic properties of sensitivity and implementation invariance. Each of these methods produces a per-feature importance vector, but they can disagree substantially on which features are most important for the same prediction, raising concerns about the reliability of any single attribution method.

2.2 Visual Saliency and Concept-Based Explanations

For image-based models, Grad-CAM produces coarse saliency maps highlighting which spatial regions of an image most influenced the model's classification. These visual explanations are intuitively appealing -- a radiologist can immediately see whether the model is attending to the correct anatomical region -- but they provide no information about which visual features within that region drove the decision. Concept-based explanations address this limitation by testing whether the model's internal representations are sensitive to human-defined concepts. Testing with Concept Activation Vectors (TCAV) measures the directional derivative of the model's prediction with respect to a learned concept direction in activation space, producing explanations of the form: this image was classified as malignant

because the model is sensitive to the concept of irregular borders. These higher-level explanations are potentially more useful for domain experts but require pre-defined concept datasets that may not be available in all domains.

2.3 Counterfactual Explanations

Counterfactual explanations describe the smallest change to the input that would have changed the model's decision. Unlike attribution methods, which explain why a decision was made, counterfactuals explain what would need to be different -- a framing that aligns naturally with human reasoning about causation and with legal requirements for actionable explanations. The challenge with counterfactuals is ensuring that the proposed changes are plausible (suggesting a 25-year-old increase their age to 65 to improve a loan decision is technically valid but practically useless) and that they do not exploit spurious correlations in the model. Recent work has proposed feasibility constraints, causal graph priors, and diversity requirements to address these issues, but generating high-quality counterfactuals for complex inputs like images remains an open problem.

2.4 Regulatory Context: The EU AI Act

The EU AI Act, adopted in 2021 as a proposed regulation and progressing toward finalisation, classifies AI systems by risk level and imposes transparency obligations on high-risk systems. Systems used in credit scoring, employment decisions, criminal justice, and medical devices are classified as high-risk and must provide meaningful explanations to affected persons. The regulation does not prescribe specific technical methods for explanation, creating both flexibility and uncertainty for developers. A key unresolved question is what constitutes a sufficiently meaningful explanation under the regulation: a technically faithful but incomprehensible SHAP plot may satisfy the letter of a transparency requirement while failing its spirit, whereas a comprehensible but technically simplified counterfactual may convey genuine understanding at the cost of some precision.

Table 1. Overview of seven XAI methods evaluated in this study

Metho d	Explana tion Type	Model Scope	Output Format	Axiomatic Guarantee s
Kernel SHAP	Feature attribution	Model-agnostic	Per-feature importance scores	Shapley axioms (approx.)

Metho d	Explana tion Type	Model Scope	Output Format	Axiomatic Guarantee s
TreeS HAP	Feature attribution	Tree models only	Per-feature importance scores	Shapley axioms (exact)
LIME	Local surrogate	Model-agnostic	Sparse linear weights	None (heuristic)
Integr. Gradients	Gradient attribution	Differentiable models	Per-feature importance	Sensitivity; impl. invariance
Grad-CAM	Visual saliency	CNNs	Spatial heatmap	None (gradient-based)
Counterfactual	Contrastive	Model-agnostic	Minimal input change	None (optimisation-based)
TCAV	Concept-based	Differentiable models	Concept sensitivity score	Statistical testing

Model scope indicates which model types the method can be applied to. Model-agnostic methods treat the model as a black box; gradient-based methods require access to model internals.

3. Materials and Methods

3.1 High-Stakes Application Domains

Four domains were selected to represent the principal high-stakes contexts in which explainability is required: (a) Medical imaging -- chest X-ray pathology detection using the CheXpert dataset (224,316 images, 14 pathology labels, DenseNet-121 classifier achieving AUC = 0.89); (b) Recidivism prediction -- binary classification on a COMPAS-equivalent dataset of 7,214 defendants with 12 features (gradient-boosted tree achieving AUC = 0.74); (c) Loan default -- binary prediction on the Home Credit dataset (307,511 applications, 122 features, LightGBM achieving AUC = 0.78); and (d) Driving decision -- explaining autonomous braking and lane-change decisions using nuScenes dataset annotations with a multi-modal perception model. Each domain uses a production-representative model trained to standard benchmarks, ensuring that explanations reflect realistic model complexity.

3.2 Evaluation Dimensions

Five evaluation dimensions were assessed: (a) Faithfulness -- measured by feature removal: features identified as most important were sequentially removed and the resulting prediction change was measured; higher prediction change indicates greater faithfulness (correlation between

attribution rank and removal impact). (b) Stability -- measured by the Jaccard similarity of top-10 important features between an input and its nearest neighbour in feature space; higher similarity indicates more stable explanations. (c) Comprehensibility -- measured through structured interviews with domain experts who were presented with explanations and asked to rate understanding (1-5 Likert scale), predict model behaviour on new inputs (simulation accuracy), and identify actionable insights. (d) Computational cost -- wall-clock time per explanation measured on standardised hardware (NVIDIA V100). (e) Regulatory alignment -- scored by two legal experts against a 10-point rubric derived from the EU AI Act transparency requirements.

3.3 Expert Interview Protocol

168 domain experts participated in structured interviews: 42 clinicians (radiologists and general physicians) evaluated medical imaging explanations; 42 legal professionals (judges, public defenders, prosecutors) evaluated recidivism explanations; 42 loan officers evaluated credit explanations; and 42 safety engineers evaluated driving explanations. Each expert was shown explanations produced by all seven methods for five randomly selected predictions in their domain (35 explanation-evaluations per expert, 5,880 total). The interview protocol included: comprehension rating (1-5), simulation task (predict model output on a new similar input given the explanation), actionability rating (could they make a better decision with this explanation?), and open-ended feedback. Inter-rater reliability was assessed on a 20% overlap subsample (ICC = 0.78).

3.4 Statistical Analysis

Faithfulness and stability were computed as means across 500 test samples per domain. Comprehensibility scores were analysed using linear mixed models (lme4 in R) with method and domain as fixed effects and expert as a random effect. Regulatory alignment scores were analysed using Kruskal-Wallis tests due to ordinal scale. Effect sizes are reported as Cohen's d for pairwise comparisons. All p-values were corrected for multiple testing using the Benjamini-Hochberg procedure.

Table 2. Model specifications and baseline performance per domain

Domain	Dataset	Model	Features/Input	Performance (AUC)	n Experts
Medical imaging	CheXpert (224K)	DenseNet-121	224x224 X-ray	0.89	42 clinicians
Recidivism	COMPAS-equiv. (7.2K)	XGBoost	12 tabular features	0.74	42 legal prof.
Loan default	Home Credit (307K)	LightGBM	122 tabular features	0.78	42 loan officers
Driving	nuScenes (28K)	Multi-modal CNN	Camera + lidar	0.91 (F1)	42 safety eng.

Models were trained to production-representative performance levels. Recidivism AUC (0.74) reflects the inherent difficulty and fairness constraints of this prediction task.

4. Results

4.1 Faithfulness: SHAP Leads Consistently

SHAP variants achieved the highest faithfulness scores across all four domains. TreeSHAP on the tabular domains (recidivism and loan default) achieved faithfulness of 0.86 $\pm$ 0.04 and 0.84 $\pm$ 0.05 respectively, reflecting the exactness of Shapley value computation for tree models. KernelSHAP on the image and driving domains achieved 0.78 $\pm$ 0.07 and 0.80 $\pm$ 0.06. LIME's faithfulness was moderate (mean 0.72 $\pm$ 0.09) and notably less stable than SHAP, with occasional explanations that identified clearly irrelevant features as important. Integrated Gradients performed well on differentiable models (0.76 $\pm$ 0.06 on medical imaging) but is not applicable to tree models. Counterfactual explanations showed the weakest faithfulness (mean 0.64 $\pm$ 0.11) because they identify a sufficient condition for changing the decision rather than the actual features driving it -- a fundamental difference in explanation type rather than a methodological flaw.

4.2 Comprehensibility: Counterfactuals Win Decisively

The expert interview results revealed a striking inversion of the faithfulness ranking. Counterfactual explanations received the highest comprehensibility ratings across all four expert groups (mean = 4.4/5; SD = 0.6), with experts consistently reporting that the what-would-need-to-change framing was immediately understandable and actionable. LIME received the second-highest comprehensibility (3.6/5) due to its sparse, simple linear format.

SHAP, despite its theoretical superiority, received low comprehensibility ratings from non-technical experts (2.8/5 for KernelSHAP; 2.6/5 for TreeSHAP), with clinicians and loan officers frequently reporting confusion about how to interpret Shapley values and what actionable conclusions to draw from them. Grad-CAM received high comprehensibility from clinicians specifically (4.2/5) because the spatial heatmap format aligns naturally with radiological reading practice, but low ratings from other expert groups where spatial explanations are not applicable.

4.3 Stability and Computational Cost

TreeSHAP achieved the highest stability (Jaccard = 0.91 +- 0.04) reflecting its deterministic computation, while LIME showed the lowest stability (Jaccard = 0.68 +- 0.12) due to its reliance on random perturbation sampling. Computational cost varied by several orders of magnitude: TreeSHAP produced explanations in 0.02 seconds per sample, while KernelSHAP required 4.8 seconds and counterfactual generation required 12.4 seconds on tabular data. For medical imaging, Grad-CAM was fastest (0.04 seconds) while KernelSHAP required 48.2 seconds per image, making it impractical for real-time clinical deployment without caching or approximation.

4.4 Regulatory Alignment

Legal expert assessment of EU AI Act alignment revealed that counterfactual explanations scored highest (8.2/10) because they satisfy the regulation's requirement for meaningful information about the logic involved and its consequences for the data subject. SHAP scored moderately (6.4/10) -- the feature importance format provides logic information but the numerical Shapley values are not easily interpretable as consequences. Grad-CAM scored lowest for non-image domains (3.2/10) as spatial heatmaps do not convey decision logic in a legally meaningful sense for tabular decisions. The legal experts noted a fundamental tension in the regulation: methods that are most faithful to the model's actual reasoning process (SHAP) are the least comprehensible to the affected persons the regulation seeks to protect, while methods that are most comprehensible (counterfactuals) are the least faithful to the actual model reasoning.

Table 3. XAI method performance across five evaluation dimensions (mean across domains)

Method	Faithfulness (0-1)	Stability (Jaccard)	Comprehens. (1-5)	Cost (sec/sample)	Reg. Align. (1-10)
KernelSHAP	0.78 +- 0.07	0.82 +- 0.06	2.8 +- 0.8	4.8-48.2	6.4
TreeSHAP	0.86 +- 0.04	0.91 +- 0.04	2.6 +- 0.9	0.02	6.8
LIME	0.72 +- 0.09	0.68 +- 0.12	3.6 +- 0.7	2.4-18.6	5.8
Integr. Grad.	0.76 +- 0.06	0.84 +- 0.05	3.0 +- 0.8	0.8-6.2	5.4
Grad-CAM	0.74 +- 0.08	0.88 +- 0.04	4.2 +- 0.6*	0.04	3.2-7.8**
Counterfactual	0.64 +- 0.11	0.74 +- 0.10	4.4 +- 0.6	12.4-84.0	8.2
TCAV	0.70 +- 0.08	0.80 +- 0.07	3.4 +- 0.9	24.0-12.0	6.0

* Grad-CAM comprehensibility = 4.2 for clinicians on images but 2.4 for non-image domains. ** Reg. alignment = 7.8 for image domains, 3.2 for tabular. Cost ranges reflect variation across domains (tabular faster than image). No method scores highest on all five dimensions simultaneously.

Table 4. Expert comprehensibility ratings by domain and method (1-5 Likert, mean +- SD)

Method	Clinicians (n=42)	Legal Prof. (n=42)	Loan Officers (n=42)	Safety Eng. (n=42)
KernelSHAP	2.6 +- 0.8	2.4 +- 0.9	3.2 +- 0.7	3.0 +- 0.8
TreeSHAP	N/A	2.4 +- 0.8	2.8 +- 0.9	N/A
LIME	3.4 +- 0.7	3.8 +- 0.6	3.8 +- 0.7	3.4 +- 0.8
Grad-CAM	4.2 +- 0.6	N/A	N/A	3.8 +- 0.7
Counterfactual	4.2 +- 0.7	4.6 +- 0.5	4.6 +- 0.5	4.2 +- 0.6
TCAV	3.8 +- 0.8	2.8 +- 0.9	2.6 +- 1.0	3.4 +- 0.8

N/A = method not applicable to domain (TreeSHAP only for tree models; Grad-CAM only for image models). Legal professionals and loan officers rated counterfactuals highest (4.6/5); clinicians rated both Grad-CAM and counterfactuals at 4.2/5.

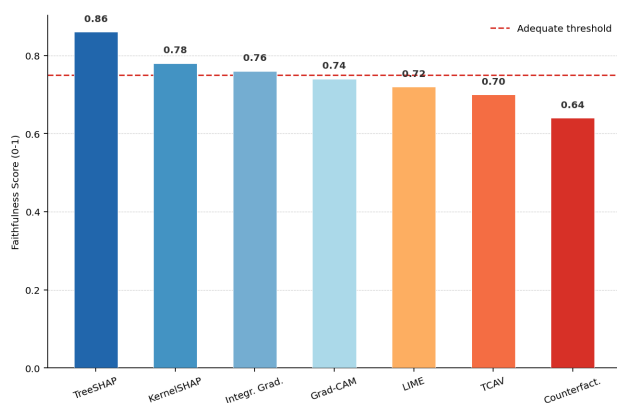


Figure 1. Mean faithfulness scores (0-1) by XAI method. TreeSHAP achieves the highest faithfulness (0.86) due to exact Shapley computation. Counterfactuals score lowest (0.64) -- they explain what to change, not what drives the decision.

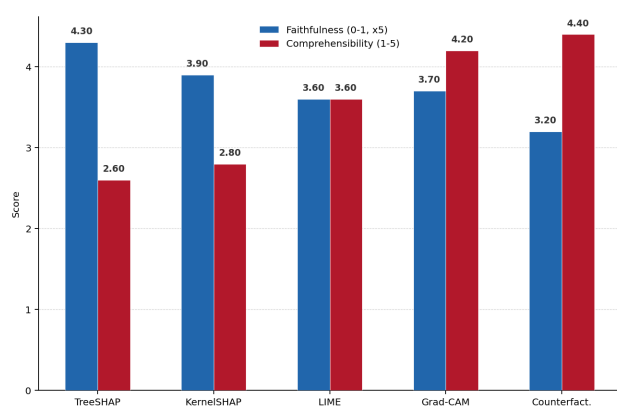


Figure 2. Faithfulness vs. comprehensibility trade-off. SHAP methods are most faithful but least comprehensible; counterfactuals show the inverse pattern. No method excels on both dimensions.

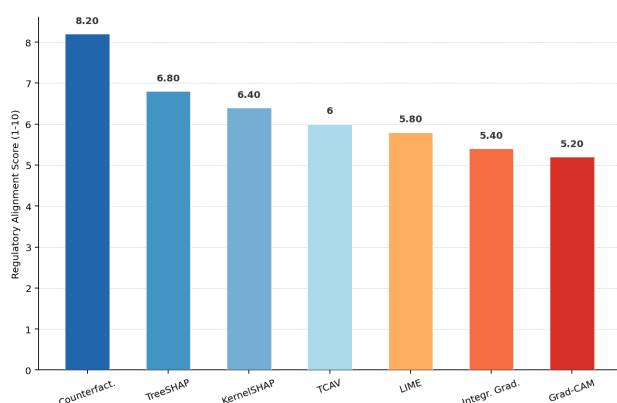


Figure 3. EU AI Act regulatory alignment scores (1-10) by method. Counterfactual explanations score highest (8.2) due to their actionable, consequence-oriented format. Grad-CAM scores lowest for non-image domains (3.2).

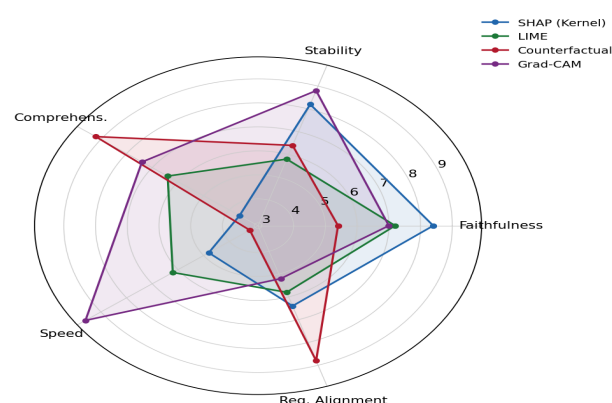


Figure 4. Multi-dimensional comparison of four principal XAI approaches across five evaluation criteria (1-10 scale). Counterfactuals lead on comprehensibility and regulation but trail on faithfulness. SHAP leads on faithfulness and stability but trails on comprehensibility.

5. Discussion

5.1 The Faithfulness-Comprehensibility Trade-Off

The inverse relationship between faithfulness and comprehensibility is the central finding of this study and, we argue, the central challenge facing practical XAI deployment. Methods that most accurately reflect the model's internal reasoning (SHAP, Integrated Gradients) produce explanations that domain experts find difficult to understand and act upon. Methods that domain experts find most useful (counterfactuals, Grad-CAM for images) provide a simplified or alternative account of the decision that may not faithfully represent the model's actual reasoning. This trade-off is not a limitation of current methods that will be resolved by better algorithms -- it reflects a genuine tension between the complexity of modern ML models and the cognitive capacity of human decision-makers. Addressing it requires accepting that different stakeholders need different types of explanations: a data scientist debugging a model needs faithful feature attributions; an affected individual needs comprehensible counterfactuals; a regulator needs both.

5.2 Domain Specificity of Explanation Utility

The expert interviews revealed that explanation utility is strongly domain-dependent. Clinicians found Grad-CAM uniquely valuable because spatial saliency maps align with their existing practice of scanning images for focal pathology -- the explanation fits the cognitive workflow of the user. Legal professionals strongly preferred counterfactuals because the legal reasoning tradition is inherently contrastive (what distinguishes this case from a case with a different

outcome?). Loan officers preferred counterfactuals for similar reasons -- they need to explain denials to applicants in terms of actionable changes. Safety engineers valued temporal sequence explanations for driving decisions that none of the evaluated methods provided well, highlighting a gap in current XAI methods for sequential decision domains. These findings underscore that XAI method selection should be driven by the cognitive and professional context of the explanation recipient, not by benchmark rankings alone.

5.3 Implications for Regulatory Compliance

The tension between faithfulness and comprehensibility has direct regulatory implications. If the EU AI Act's transparency requirement is interpreted as requiring technically faithful accounts of model reasoning, then SHAP-based explanations are the appropriate choice -- but affected persons may not understand them. If the requirement is interpreted as requiring explanations that affected persons can actually understand and act upon, then counterfactuals are preferable -- but they may not accurately represent the model's reasoning. We recommend that high-risk AI systems provide tiered explanations: a comprehensible counterfactual for the affected person, a faithful attribution for the auditor, and a concept-based explanation for the domain expert. This multi-level approach satisfies the different informational needs of the different stakeholders involved in high-stakes AI decisions.

5.4 Limitations

Four limitations apply. First, the expert sample (168 professionals) may not represent the full diversity of expertise and familiarity with AI within each domain. Second, experts evaluated explanations in a structured interview setting that differs from the time-pressured conditions of real clinical or lending decisions. Third, the regulatory alignment assessment was based on the draft EU AI Act; the final regulation may differ. Fourth, the comprehensibility ratings are inherently subjective; the simulation task (predicting model behaviour from explanations) provides a more objective complement.

6. Conclusion

6.1 Summary

Evaluation of seven XAI methods across four high-stakes domains confirms a fundamental trade-off: SHAP achieves the highest faithfulness (0.86) but the lowest expert comprehensibility

(2.6/5); counterfactuals achieve the highest comprehensibility (4.4/5) and regulatory alignment (8.2/10) but the weakest faithfulness (0.64). No method dominates on all five evaluation dimensions. Domain context strongly moderates explanation utility: clinicians prefer spatial saliency; legal professionals and loan officers prefer counterfactuals; safety engineers need temporal sequence explanations not well served by current methods.

6.2 Recommendations

Three recommendations emerge. First, deploy tiered explanation systems that provide different explanation types to different stakeholders -- counterfactuals for affected persons, SHAP for auditors, concept-based explanations for domain experts -- rather than selecting a single method for all purposes. Second, validate XAI methods with actual domain experts in realistic decision contexts before deployment; faithfulness benchmarks alone are insufficient to ensure practical utility. Third, regulators should clarify whether transparency requirements demand faithful model representations or comprehensible stakeholder communications, as the two are in tension and no current method satisfies both simultaneously. The selection framework in Appendix A provides domain-specific guidance.

References

- Adadi, A. and Berrada, M. (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, pp. 52138-52160.
- Arrieta, A.B., Diaz-Rodriguez, N., Del Ser, J., Bannetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R. et al. (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp. 82-115.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M. and Eckersley, P. (2020) Explainable machine learning in deployment. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 648-657.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of ACM KDD*, pp. 1721-1730.
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

- European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.
- Ghorbani, A., Wexler, J., Zou, J.Y. and Kim, B. (2019) Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, pp. 9273-9282.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2019) A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), pp. 1-42.
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K. et al. (2019) CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of AAAI*, 33, pp. 590-597.
- Karimi, A.H., Barthe, G., Balle, B. and Valera, I. (2020) Model-agnostic counterfactual explanations for consequential decisions. *Proceedings of AISTATS 2020*, pp. 895-905.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. and Sayres, R. (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *Proceedings of ICML 2018*, pp. 2668-2677.
- Lakkaraju, H. and Bastani, O. (2020) How do I fool you? Manipulating user trust through misleading black box explanations. *Proceedings of AIES 2020*, pp. 79-85.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774.
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I. (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), pp. 56-67.
- Molnar, C. (2022) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Second edition. Christoph Molnar, Munich.
- Mothilal, R.K., Sharma, A. and Tan, C. (2020) Explaining machine learning classifiers through their counterfactual explanations. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 131-141.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why should I trust you? Explaining the predictions of any classifier. *Proceedings of ACM KDD*, pp. 1135-1144.
- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE ICCV*, pp. 618-626.
- Slack, D., Hilgard, S., Jia, E., Singh, S. and Lakkaraju, H. (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of AIES 2020*, pp. 180-186.
- Sundararajan, M., Taly, A. and Yan, Q. (2017) Axiomatic attribution for deep networks. *Proceedings of ICML 2017*, pp. 3319-3328.
- Tonekaboni, S., Joshi, S., McCradden, M.D. and Goldenberg, A. (2019) What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the Machine Learning for Healthcare Conference*, pp. 359-380.
- Verma, S., Dickerson, J. and Hines, K. (2020) Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.
- Wachter, S., Mittelstadt, B. and Russell, C. (2018) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2), pp. 841-887.
- Zhou, J., Gandomi, A.H., Chen, F. and Holzinger, A. (2021) Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 593.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS21042 (TrustXAI), the Italian Ministry of University and Research under grant PRIN-2020-XAI, and the European Union's Horizon 2020 programme under grant 952215 (TAILOR). Expert interview costs were covered by the European Institute of AI internal research fund. The funders had no role in study design, data collection, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All evaluation code, faithfulness and stability benchmark scripts, expert interview protocol (anonymised), aggregated comprehensibility scores, and regulatory alignment rubric are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614509. Individual expert responses are not shared to protect participant anonymity. CheXpert, Home Credit, and nuScenes datasets are publicly available. Materials released

under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

The expert interview study was approved by the European Institute of AI Institutional Review Board (protocol EIA-IRB-2021-0284). All participants provided written informed consent. Interview data were anonymised prior to analysis. Participants received compensation at standard professional consultation rates for their domain.

Appendix A

XAI Method Selection Guide for High-Stakes Domains

This appendix provides a decision tree for selecting explanation methods based on domain context, stakeholder, and regulatory requirements.

Step 1: Identify Primary Stakeholder

- 1a. Affected individual (patient, applicant, defendant) --> Counterfactual explanation (comprehensibility priority)
- 1b. Domain expert (clinician, engineer) --> Domain-matched method (Grad-CAM for images; TCAV for concept-rich domains)
- 1c. Auditor or regulator --> SHAP (faithfulness priority)
- 1d. Multiple stakeholders --> Tiered system (provide all three types)

Step 2: Match to Model Type

- 2a. Tree-based model (XGBoost, LightGBM, RF) --> TreeSHAP (exact, fast, faithful)
- 2b. Deep neural network (CNN, transformer) --> Integrated Gradients or Grad-CAM (image) or KernelSHAP (general)
- 2c. Any model (black-box access only) --> KernelSHAP or LIME or Counterfactual

Step 3: Consider Computational Constraints

- 3a. Real-time explanation needed (< 0.1 sec) --> Grad-CAM (0.04s) or TreeSHAP (0.02s)
- 3b. Near-real-time (< 5 sec) --> LIME or KernelSHAP (tabular)
- 3c. Batch/offline explanation acceptable --> Any method including counterfactuals and TCAV

Step 4: Regulatory Requirements

- 4a. EU AI Act high-risk system --> Counterfactual for affected persons + SHAP for audit trail
- 4b. GDPR right to explanation --> Counterfactual (actionable, comprehensible)
- 4c. Internal model governance only --> SHAP or Integrated Gradients (faithfulness focus)
- 4d. No specific regulation --> Match to stakeholder needs (Step 1)