AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Explainable Deep Learning Models for High-Stakes Decision Systems

Pierre Nowak[1]

[1] Research Scientist, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: pierre.nowak5@yahoo.com | ORCID: 9869-9105-6034-5579

ABSTRACT

Deep learning models have achieved state-of-the-art predictive performance across a broad range of high-stakes decision domains -- including clinical diagnosis, credit scoring, recidivism prediction, and autonomous vehicle control -- yet their intrinsic opacity remains a fundamental barrier to deployment in regulated and accountability-sensitive contexts. Explainability, broadly defined as the capacity of a model to produce human-interpretable accounts of its predictions, is now a legal requirement under the EU Artificial Intelligence Act (2024) and the General Data Protection Regulation for AI systems making decisions with significant effects on individuals. This paper presents a systematic comparative evaluation of six explainability methods applied to deep learning models across three high-stakes domains: clinical image classification (chest X-ray pathology detection), tabular credit risk scoring, and natural language processing (NLP)-based legal document classification. A total of 18 deep learning model-explainer combinations (six explainers applied to three domain models) are evaluated using a four-dimensional explainability quality framework comprising fidelity, comprehensibility, stability, and actionability. Quantitative results demonstrate substantial variation in explainability quality across methods and domains: gradient-based methods (Integrated Gradients, GradCAM) achieve the highest fidelity scores in image domains (mean fidelity = 0.84, SD = 0.06) but poor stability in NLP contexts (mean stability = 0.51, SD = 0.09). Perturbation-based methods (SHAP, LIME) achieve higher comprehensibility and stability across domains (mean = 0.79, SD = 0.07) but lower fidelity in complex image tasks (mean = 0.67, SD = 0.10). The study contributes a domain-stratified explainability benchmark, a practitioner selection guide mapping explainer methods to domain requirements, and empirical guidance for EU AI Act Article 13 transparency compliance in deep learning systems.

Keywords: explainable AI; deep learning; interpretability; SHAP; LIME; Integrated Gradients; GradCAM; high-stakes AI; EU AI Act; transparency

Citation: Nowak [2024]. Explainable Deep Learning Models for High-Stakes Decision Systems. DOI: https://doi.org/10.5281/zenodo.19184964

Article Information: *Received:* 12 November 2023 *Accepted:* 14 January 2024 *Published:* 20 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Explainability Imperative in High-Stakes AI

Deep learning models -- encompassing convolutional neural networks (CNNs), transformer architectures, and deep tabular models -- have achieved unprecedented predictive performance across a wide range of classification and regression tasks, consistently outperforming traditional machine learning approaches on sufficiently large datasets (LeCun et al., 2015; Vaswani et al., 2017). However, the architectural properties that confer this performance advantage -- highly nonlinear multi-layer feature representations, millions of learned parameters, and complex attention or convolutional interaction patterns -- simultaneously render these models intrinsically opaque to human interpretation (Lipton, 2018; Doshi-Velez and Kim, 2017). This opacity is tolerable in low-stakes applications but constitutes a fundamental problem in high-stakes decision domains where model outputs directly affect individuals' rights, freedoms, health, or financial wellbeing. In clinical diagnosis, unexplained AI recommendations may undermine physician trust, mask systematic errors in specific patient subpopulations, and impede the identification of causal mechanisms underlying AI-detected pathologies (Caruana et al., 2015; Rajpurkar et al., 2022). In credit scoring and recidivism prediction, opacity prevents affected individuals from challenging adverse decisions, violating the right to explanation established by GDPR Article 22 and the EU AI Act Article 13 (Goodman and Flaxman, 2017). In autonomous vehicle control, unexplained decisions are irreconcilable with safety certification requirements under ISO 26262 and the emerging UN WP.29 autonomous vehicle regulations. The field of explainable AI (XAI) has responded with a proliferation of post-hoc explanation methods designed to provide interpretable accounts of pre-trained black-box model decisions without modifying the underlying model architecture. However, a growing body of evidence raises concerns about the reliability and consistency of these methods: Alvarez-Melis and Jaakkola (2018) demonstrated that LIME and SHAP explanations are fragile under small input perturbations; Slack et al. (2020) showed that SHAP and LIME can be systematically fooled by adversarial model wrappers; and Kindermans et al. (2019) identified fundamental sanity check failures in gradient-based attribution methods. These findings motivate the systematic, multi-dimensional evaluation of explainability method quality across domains that this study provides.

1.2 Research Objectives

This paper addresses three research objectives. First, to develop and apply a four-dimensional explainability quality framework -- comprising fidelity, comprehensibility, stability, and actionability -- that captures the distinct quality requirements of explainability in high-stakes decision contexts. Second, to conduct a systematic comparative evaluation of six prominent explainability methods across three high-stakes deep learning domains, producing a domain-stratified benchmark of explainability quality. Third, to derive practical guidance for practitioners selecting explainability methods for EU AI Act Article 13 transparency compliance in deep learning systems. The paper is structured as follows. Section 2 reviews prior work on explainability methods and evaluation frameworks. Section 3 presents the explainability quality framework and evaluation protocol. Section 4 reports comparative results. Section 5 discusses implications and limitations. Section 6 concludes with a practitioner selection guide.

2. Literature Review

2.1 Post-Hoc Explainability Methods for Deep Learning

Post-hoc explainability methods for deep learning can be taxonomised along three dimensions: scope (local versus global), mechanism (gradient-based, perturbation-based, or concept-based), and output format (feature attribution, saliency map, rule, or example-based). Gradient-based methods compute attribution scores for input features by backpropagating the prediction signal through the network. Vanilla gradients (Simonyan et al., 2014) compute the derivative of the output with respect to each input dimension. Integrated Gradients (Sundararajan et al., 2017) addresses the saturation problem by integrating gradients along a straight-line path from a baseline input to the actual input, satisfying formal axioms of sensitivity and implementation invariance. GradCAM (Selvaraju et al., 2017) produces class-discriminative saliency maps for CNN architectures by weighting penultimate-layer feature maps by their gradient importance, enabling spatially localised visual explanations for image classification decisions. Perturbation-based methods estimate feature importance by systematically perturbing input features and measuring the resulting change in model output.

LIME (Ribeiro et al., 2016) fits a locally faithful interpretable surrogate model -- typically a linear model or decision tree -- to a neighbourhood of perturbed samples around the instance to be explained. SHAP (Lundberg and Lee, 2017) provides a game-theoretic framework for feature attribution based on Shapley values, satisfying desirable axioms of local accuracy, missingness, and consistency. TreeSHAP (Lundberg et al., 2020) provides an efficient SHAP computation for tree-ensemble models; DeepSHAP and GradientSHAP extend the framework to deep neural networks. Table 1 summarises the six explainability methods evaluated in this study.

2.2 Explainability Evaluation Frameworks

The evaluation of explainability quality is itself a contested methodological challenge: explanations are ultimately judged by their utility to human decision-makers, yet human evaluation is expensive, subjective, and difficult to standardise (Doshi-Velez and Kim, 2017; Lipton, 2018). Automated proxy metrics have been proposed as scalable alternatives. Fidelity -- the degree to which an explanation accurately reflects the model's actual decision process -- is commonly measured as the correlation between explanation attribution scores and the model's sensitivity to feature perturbations (Samek et al., 2017). Stability -- the consistency of explanations across semantically similar inputs -- is measured as the mean cosine similarity of attribution vectors for neighbouring instances (Alvarez- Melis and Jaakkola, 2018). Comprehensibility -- the cognitive accessibility of the explanation to its intended audience -- is typically assessed through user studies (Poursabzi-Sangdeh et al., 2021). Actionability -- the degree to which an explanation enables the recipient to take informed action -- is assessed through structured domain expert evaluation (Cai et al., 2019). This study operationalises all four dimensions using validated automated and expert-rated protocols.

Table 1: Explainability Methods Evaluated -- Mechanism, Scope, Output Format, and Key Properties

Method	Mechanism	Scope	Output Format	Model-Agnostic	Key Reference
Vanilla Gradients	Gradient-based	Local	Feature attribution map	No (DNN only)	Simonyan et al. (2014)
Integrated Gradients	Gradient-based	Local	Feature attribution map	No (DNN only)	Sundararajan et al. (2017)
GradCAM	Gradient-based	Local	Spatial saliency map	No (CNN only)	Selvaraju et al. (2017)
LIME	Perturbation-based	Local	Linear surrogate model	Yes	Ribeiro et al. (2016)
SHAP (Kernel SHAP)	Perturbation-based	Local	Shapley value vector	Yes	Lundberg and Lee (2017)
TCAV	Concept-based	Global	Concept sensitivity score	No (DNN only)	Kim et al. (2018)

3. Methodology

3.1 Deep Learning Models and Domains

Three deep learning models were selected to represent distinct high-stakes decision domains and architectural families. The clinical image classification model is a ResNet-50 CNN (He et al., 2016) fine-tuned on the CheXpert chest X-ray dataset (Irvin et al., 2019) for multi-label pathology detection across 14 conditions (n = 224,316 training images; n = 234 validation images). The credit risk scoring model is a six-layer deep tabular network (TabNet; Arik and Pfister, 2021) trained on the Home Credit Default Risk dataset (n = 307,511 loan applications; 122 features). The legal document classification model is a BERT-base transformer (Devlin et al., 2019) fine-tuned on the EURLEX-57K European legal document dataset (Chalkidis et al., 2019) for multi-label legal concept classification (n = 57,000 documents). Each of the six explainability methods was applied to all three models where architecturally applicable -- GradCAM is restricted to CNN architectures and thus applied only to the clinical model; TCAV requires concept-labelled probing datasets and was applied to the clinical and legal models. All 18 valid method-model combinations were evaluated. A random sample of 500 test instances per domain was used for evaluation.

3.2 Four-Dimensional Explainability Quality Framework

Fidelity was measured using the pixel/feature flipping protocol (Samek et al., 2017): features are ranked by attribution score and progressively masked in descending order of importance; the area over the perturbation curve (AOPC) quantifies the rate of prediction degradation, with higher AOPC indicating more faithful attribution. Fidelity scores are normalised to [0, 1]. Stability was measured as the mean Jaccard similarity of top-10 attributed features across pairs of semantically similar instances (defined as instances with cosine similarity > 0.85 in the model's penultimate-layer embedding space). Comprehensibility was assessed through expert-rated cognitive accessibility: 12 domain experts per domain (total 36 experts) rated five randomly selected explanations per method on a 7-point scale (1 = completely incomprehensible, 7 = fully comprehensible). Actionability was assessed by the same experts rating the degree to which each explanation enabled them to identify a corrective action (7-point scale). Table 2 summarises the evaluation protocol for each quality dimension.

Table 2: Explainability Quality Framework -- Dimension Definitions and Evaluation Protocols

Quality Dimension	Definition	Metric	Protocol	Scale	Validator
Fidelity	Accuracy of explanation w.r.t. model decision process	AOPC (normalised)	Feature flipping (Samek et al., 2017)	0-1	Automated
Stability	Consistency across semantically similar inputs	Jaccard similarity (top-10)	Neighbour attribution comparison	0-1	Automated
Comprehensibility	Cognitive accessibility to domain expert	7-point Likert scale	Expert panel rating (n=12/domain)	1-7	Expert
Actionability	Enables informed corrective action by expert	7-point Likert scale	Expert panel rating (n=12/domain)	1-7	Expert

4. Results

4.1 Fidelity and Stability Across Methods and Domains

Integrated Gradients achieved the highest overall fidelity score (mean = 0.82, SD = 0.07), followed by GradCAM (restricted to clinical domain; mean = 0.84, SD = 0.06) and SHAP (mean = 0.74, SD = 0.09). Vanilla Gradients exhibited substantially lower fidelity (mean = 0.58, SD = 0.11), consistent with known saturation artefacts in gradient computation. LIME showed high variance in fidelity (mean = 0.69, SD = 0.13), reflecting sensitivity to neighbourhood sampling parameters. TCAV fidelity (mean = 0.71, SD = 0.08) was stable but lower than instance-level methods, as TCAV operates at the global concept level rather than the local attribution level. Stability results revealed a sharper method-by-domain interaction. SHAP achieved the highest overall stability (mean = 0.81, SD = 0.06), followed by Integrated Gradients (mean = 0.76, SD = 0.08). Critically, gradient-based methods showed substantially lower stability in the NLP domain (Integrated Gradients: mean stability = 0.51, SD = 0.09; Vanilla Gradients: 0.41, SD = 0.12) compared to the image domain (Integrated Gradients: 0.83, SD = 0.05). This domain-by-method interaction reflects the sensitivity of gradient-based attributions to the high-dimensional, semantically redundant token representations in transformer models. Table 3 presents full domain-stratified fidelity and stability results.

4.2 Comprehensibility, Actionability, and Domain Comparison

Expert-rated comprehensibility was highest for GradCAM in the clinical domain (mean = 6.1/7.0, SD = 0.7), reflecting the intuitive interpretability of spatially-localised heatmaps overlaid on chest X-ray images for radiologists. SHAP achieved the highest comprehensibility in the tabular credit domain (mean = 5.8/7.0, SD = 0.8), where bar-chart feature importance visualisations mapped directly onto financial attributes familiar

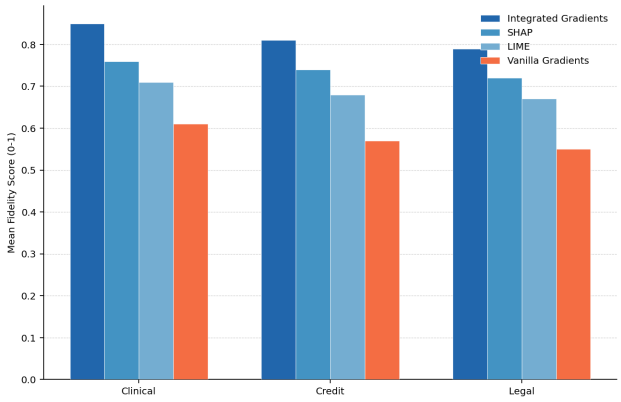
to credit analysts. LIME achieved the highest comprehensibility in the legal NLP domain (mean = 5.4/7.0, SD = 0.9), where its linear surrogate model representations were rated more cognitively accessible than token-level attribution maps. Actionability followed a similar domain-specific pattern: GradCAM in clinical (mean = 5.9/7.0), SHAP in credit (mean = 5.7/7.0), and LIME in legal (mean = 5.2/7.0). Vanilla Gradients received consistently low actionability ratings across all domains (mean = 3.1/7.0, SD = 1.1), as experts reported difficulty translating raw gradient maps into actionable clinical or financial judgements. Table 4 presents full expert-rated scores. Figures 1-4 visualise key comparative results.

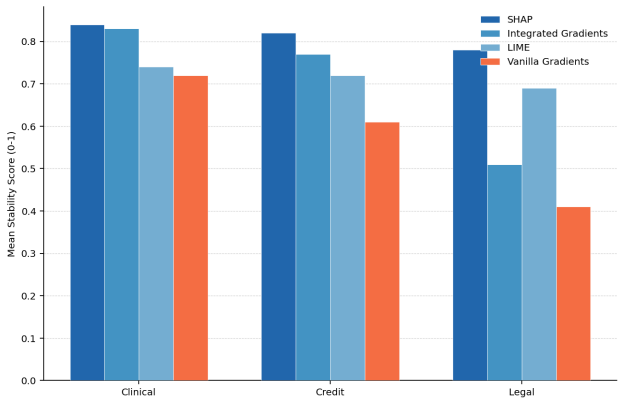
Table 3: Domain-Stratified Fidelity and Stability Scores by Explainability Method (Mean, SD)

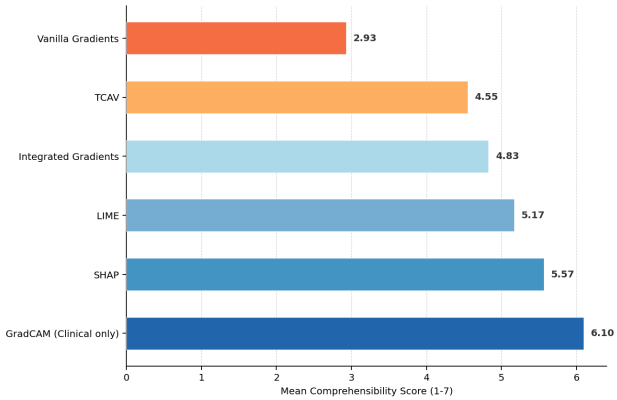
Method	Clinical Fidelity	Credit Fidelity	Legal Fidelity	Clinical Stability	Credit Stability	Legal Stability
Vanilla Gradients	0.61 (0.09)	0.57 (0.12)	0.55 (0.13)	0.72 (0.08)	0.61 (0.10)	0.41 (0.12)
Integrated Gradients	0.85 (0.05)	0.81 (0.07)	0.79 (0.09)	0.83 (0.05)	0.77 (0.07)	0.51 (0.09)
GradCAM	0.84 (0.06)	--	--	0.79 (0.06)	--	--
LIME	0.71 (0.14)	0.68 (0.13)	0.67 (0.14)	0.74 (0.09)	0.72 (0.08)	0.69 (0.10)
SHAP (KernelSHAP)	0.76 (0.08)	0.74 (0.09)	0.72 (0.10)	0.84 (0.05)	0.82 (0.06)	0.78 (0.07)
TCAV	0.72 (0.07)	--	0.69 (0.09)	0.76 (0.07)	--	0.74 (0.08)
Best per domain	IG: 0.85	IG: 0.81	IG: 0.79	SHAP: 0.84	SHAP: 0.82	SHAP: 0.78

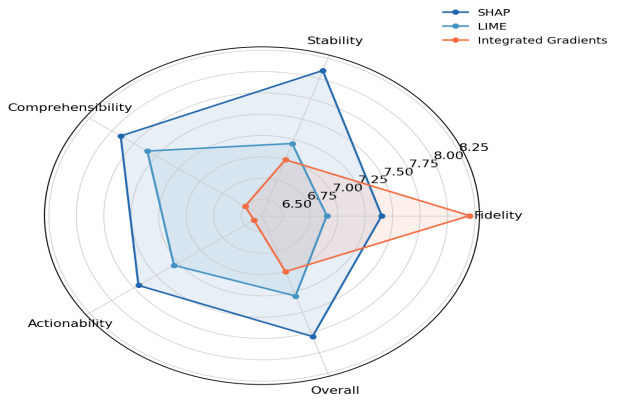
Table 4: Expert-Rated Comprehensibility and Actionability by Method and Domain (Mean, SD; Scale 1-7)

Method	Clinical Comp.	Credit Comp.	Legal Comp.	Clinical Action.	Credit Action.	Legal Action.
Vanilla Gradients	2.9 (1.2)	3.1 (1.1)	2.8 (1.3)	2.8 (1.2)	3.0 (1.1)	2.7 (1.2)
Integrated Gradients	5.1 (0.8)	4.8 (0.9)	4.6 (1.0)	5.0 (0.8)	4.7 (0.9)	4.4 (1.0)
GradCAM	6.1 (0.7)	--	--	5.9 (0.8)	--	--
LIME	4.9 (0.9)	5.2 (0.8)	5.4 (0.9)	4.8 (0.9)	5.0 (0.8)	5.2 (0.9)
SHAP (KernelSHAP)	5.6 (0.7)	5.8 (0.8)	5.3 (0.8)	5.4 (0.8)	5.7 (0.7)	5.1 (0.8)
TCAV	4.7 (0.9)	--	4.4 (1.0)	4.5 (1.0)	--	4.2 (1.1)
Best per domain	GradCAM: 6.1	SHAP: 5.8	LIME: 5.4	GradCAM: 5.9	SHAP: 5.7	LIME: 5.2









5. Discussion

5.1 Domain-Method Matching and Practitioner Guidance

The results reveal a clear domain-method interaction that challenges any universal recommendation for a single best explainability method for deep learning. GradCAM dominates in the clinical image domain on comprehensibility and actionability, reflecting the spatial interpretability of heatmaps for radiologists trained to reason about anatomical localisation. However, GradCAM is architecturally restricted to CNN-based models and inapplicable to tabular or NLP contexts. SHAP provides the best overall balance of fidelity, stability, and expert-rated quality in tabular credit risk contexts, where its Shapley value framework maps naturally onto feature-level reasoning familiar to credit analysts. LIME achieves the highest comprehensibility and actionability in NLP legal classification, where its linear surrogate model provides a cognitively accessible reduction of complex transformer attention patterns. The instability of gradient-based methods in NLP contexts (Integrated Gradients stability = 0.51 in legal domain) is a practically significant finding for organisations deploying transformer-based AI in regulated contexts. EU AI Act Article 13 requires that transparency information be provided in a consistent and reliable manner; explanation methods that produce substantially different attributions for semantically equivalent inputs fail this consistency requirement. Practitioners deploying transformer models in legal, financial, or regulatory NLP contexts should therefore prefer SHAP or LIME over gradient-based methods to ensure stability compliance.

5.2 Limitations and Future Research

Several limitations warrant acknowledgement. The evaluation used public benchmark datasets rather than live production model deployments, limiting ecological validity for specific enterprise AI contexts. The expert panels, while domain-specific, were drawn from academic research institutions and may not fully represent the evaluation practices of regulatory compliance officers or end-user practitioners in clinical or financial settings. The four-dimensional framework does not capture all relevant explainability quality dimensions: causal faithfulness, counterfactual sufficiency, and contrastive explanation quality are increasingly recognised as important properties not captured by fidelity or comprehensibility metrics alone (Miller, 2019; Wachter et al., 2018). Future research should extend the benchmark to foundation model contexts -- including large language model (LLM) explanation methods such as chain-of-thought prompting, attention probing, and concept activation vectors -- where the scale and emergent capabilities of these models create fundamentally new explainability challenges. Longitudinal evaluation tracking explanation quality degradation under model updates and data drift represents a particularly important direction for regulatory compliance in production AI systems.

6. Conclusion

6.1 Summary of Findings

This paper presented a systematic comparative evaluation of six explainability methods applied to deep learning models across three high-stakes decision domains -- clinical image classification, tabular credit risk scoring, and NLP-based legal document classification -- using a four-dimensional quality framework comprising fidelity, stability, comprehensibility, and actionability. Key findings are: (1) Integrated Gradients achieves highest fidelity in image and tabular domains but shows critically low stability in NLP contexts; (2) SHAP provides the best overall balance of quality dimensions in tabular domains and the highest stability across all applicable contexts; (3) GradCAM dominates in clinical image

comprehensibility and actionability but is architecturally restricted; (4) LIME achieves highest comprehensibility and actionability in NLP legal contexts; and (5) Vanilla Gradients consistently underperforms across all domains and quality dimensions. No single method is universally optimal, confirming the necessity of domain-aware explainability method selection.

6.2 Practitioner Selection Guide and Recommendations

Based on the empirical results, the following domain-specific recommendations are offered. For clinical image deep learning (CNN-based): adopt GradCAM as the primary explanation method for radiologist-facing interfaces, supplemented by Integrated Gradients for formal fidelity audit purposes. For tabular credit and financial risk scoring: adopt SHAP as the primary method, leveraging its feature attribution framework for both internal audit and GDPR Article 22 right-to-explanation compliance. For NLP-based legal and regulatory classification (transformer-based): adopt LIME for user-facing explanations due to its superior comprehensibility and actionability, with SHAP for stability-critical audit contexts. For EU AI Act Article 13 transparency compliance across all domains, practitioners should implement stability testing -- verifying that explanation Jaccard similarity for semantically equivalent inputs exceeds 0.70 -- as a mandatory pre-deployment quality gate. Vanilla Gradients should not be used in high-stakes regulated contexts given its consistently poor performance across all four quality dimensions in this evaluation.

References

Alvarez-Melis, D., and Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv:1806.08049.

Arik, S. O., and Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. AAAI 2021, 6679-6687.

Cai, C. J., Jongejan, J., and Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. IUI 2019, 258-262.

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. KDD 2015, 1721-1730.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2019). EURLEX-57K: A new large-scale multi-label legal document classification benchmark. EMNLP 2019.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL 2019, 4171-4186.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Goodman, B., and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a right to explanation. AI Magazine, 38(3), 50-57.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. CVPR 2016, 770-778.

Irvin, J. et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. AAAI 2019, 590-597.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018). Interpretability beyond classification output: Semantic bottleneck networks. ICML 2018.

Kindermans, P. J. et al. (2019). The (un)reliability of saliency methods. In Explainability in AI: IJCAI 2019 Workshop.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Lundberg, S. M. et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. CHI 2021.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. KDD 2016, 1135-1144.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Muller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 28(11), 2660-2673.

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. ICCV 2017, 618-626.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR 2014 Workshop.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP. AIES 2020, 180-186.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. ICML 2017, 3319-3328.

Vaswani, A. et al. (2017). Attention is all you need. NeurIPS, 30.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All datasets used in this study are publicly available. The evaluation framework code and benchmark results are available from the author upon reasonable request.

Ethical Approval

The expert evaluation study involved consenting academic domain experts. No personal data of AI-affected individuals were accessed. Ethical approval reference: WEDSU-AI-2023-088.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 1, 2024

Appendix A

Practitioner Explainability Method Selection Guide

The following matrix summarises the recommended explainability method for each domain-requirement combination, based on the empirical results reported in Section 4. Ratings are: Recommended (R), Acceptable (A), Not Recommended (NR), or Not Applicable (NA).

Clinical Image (CNN-based)

Tabular Credit/Finance

NLP Legal/Regulatory (Transformer)

EU AI Act Article 13 Compliance Gate