

A Data-Centric Literature Review of SHAP and LIME for Tabular Decision Support Systems

Raheleh Yoosefzadeh

Independent Researcher

Email: r.yoosefzadeh@gmail.com

Abstract

Decision Support Systems (DSS) in healthcare and finance increasingly rely on SHAP and LIME to mitigate the "black box" problem of machine learning models. While traditional research focuses on algorithmic mechanics, this paper provides a data-centric literature review. It shifts the focus from the internal logic of the explainer to how intrinsic data characteristics and preprocessing decisions shape the resulting output.

The review identifies critical vulnerabilities linked to feature dependence, where collinearity leads to marginalization failures and "attribution splitting". It further examines data quality gaps, specifically how imputation artifacts act as confounding factors that can lead explainers to highlight reconstructed values rather than observed evidence. Additionally, the paper addresses how class imbalance and synthetic resampling distort decision boundaries, while sampling stochasticity introduces instability through out-of-distribution perturbations. Security and privacy concerns are also explored, including adversarial manipulation and membership inference attacks facilitated by specific attribution patterns.

*Understanding these data-driven limitations is particularly essential for **data scientists and AI practitioners to interpret models outputs reliably** before deployment. Ultimately, this work argues that the fidelity of an explanation is inextricably linked to the underlying **data** distribution. Achieving robust and trustworthy explainability is a fundamentally data-driven challenge that requires the development of **data-aware XAI methods** and standardized evaluation protocols.*

To understand this shift, consider that a machine learning explanation is like a photograph of a landscape; its clarity depends less on the camera's brand (the algorithm) and more on the weather conditions and terrain (the data) being captured.

Keywords: Explainable AI (XAI), SHAP and LIME, Data-Centric Review, Tabular Data, Decision Support Systems (DSS)

1.0 Introduction

Decision Support Systems (DSS) rely heavily on structured and semi-structured data stored in tables, spreadsheets, databases, and data warehouses. Across domains such as healthcare, finance, and business, tabular data serve as the backbone of DSS operations, enabling informed decision-making and predictive analytics (Guan et al. 2025). The adoption of machine learning (ML) models in high-stakes DSS has been rapidly increasing, particularly in data-rich sectors like healthcare and finance, where predictive insights can have critical consequences (Batko and Ślęzak 2022; Černevičienė and Kabašinskas 2024; Weber et al. 2024).

However, the growing complexity of AI models often leads to the "black box" problem, where the internal logic behind predictions is opaque to human users. This lack of transparency undermines trust, accountability, and regulatory compliance, posing a major barrier to broader adoption in clinical and financial contexts (Kabir et al. 2025).

To address this challenge, Explainable AI (XAI) techniques have been developed to illuminate the decision-making processes of complex models. Among these, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are the most widely used. As model-agnostic, post-hoc explanation methods, they are valued for their flexibility and are commonly applied to models trained on tabular data (Salih et al. 2025; Martens et al. 2025).

Despite the prevalence of tabular data in DSS, most XAI methods—including SHAP and LIME—are not specifically designed for this data type (O'Brien Quinn et al. 2024). Unlike reviews that focus primarily on algorithmic innovations, this work adopts a data-centric perspective to address a central yet often overlooked question: How do the inherent characteristics of tabular data—and the preprocessing decisions made by practitioners—affect the reliability, stability, and interpretability of SHAP and LIME explanations? By examining features such as correlations, sparsity, and preprocessing choices, this review highlights data-related limitations that purely algorithm-focused studies often overlook.

This paper proceeds by defining the data-centric approach and outlining its scope, followed by a detailed examination of the methodological foundations of SHAP and LIME. It then systematically analyzes key data-related challenges, including feature dependence, data quality, and class imbalance, before presenting a comparative evaluation and summarizing common evaluation frameworks. Finally, the review concludes with key findings and directions for future research, emphasizing that the fidelity of an explanation depends not only on the algorithm but also on the characteristics of the data it interprets.

1.1 What We Mean by a Data-Centric Review

This review adopts a data-centric perspective, diverging from a method-centric approach that prioritizes algorithmic variations, computational performance, or visualization techniques. The

focus is shifted from “How does the explainer work?” to “How does the data shape the explainer’s output?”

The analysis is structured along three axes:

- **Data Properties:** How do intrinsic data characteristics such as feature dependence, class imbalance, and overall data quality influence the stability and faithfulness of SHAP and LIME explanations?
- **Preprocessing Choices:** What is the impact of common data handling strategies, including missing value imputation, on the resulting explanations?
- **Sampling Strategies:** Impact of perturbation and neighborhood generation—the core techniques in LIME and KernelSHAP—on explanation fidelity and robustness.

The following table provides a clear contrast between a traditional method-centric review and the data-centric focus of this paper.

Table 1. Method-Centric vs. Data-Centric Perspectives in XAI Literature Reviews

Axis	Method-Centric Review	Data-Centric Review (This Paper)
Unit of Analysis	Variants of SHAP/LIME algorithms and their mathematics	Properties of input data and interactions
Central Question	How does the explainer algorithm work?	How does the data shape the explainer’s output?
Evaluation Focus	Computational performance, algorithmic efficiency	Explanation stability, reliability, robustness to data shifts
Key Artifacts	Algorithm pseudocode, performance benchmarks	Preprocessing pipelines, data sampling strategies

1.2 Review Scope and Methodology

This paper presents a narrative, targeted literature review aimed to synthesize key themes, challenges, and insights from existing research on SHAP and LIME applied to tabular data within Decision Support Systems (DSS), with a focus on healthcare and finance, where explainability is a critical requirement. The scope is deliberately narrow to ensure depth, emphasizing post-hoc explanation methods rather than algorithmic variants or performing an exhaustive systematic search.

Relevant literature was selected through a structured methodology applied to peer-reviewed articles, conference proceedings, and influential pre-prints. Sources were drawn from Google

Scholar, IEEE Xplore, ACM Digital Library, PubMed, Scopus, and ScienceDirect, using keywords such as “Explainable Artificial Intelligence,” “XAI,” “survey,” and “tabular data”. Inclusion criteria prioritized studies that applied SHAP or LIME to real-world or realistic synthetic tabular datasets, particularly in healthcare and finance, or critically evaluated these methods with respect to input data characteristics.

2.0 Methodological Foundations of SHAP and LIME

To fully comprehend the data-centric challenges that SHAP and LIME face in real-world Decision Support Systems (DSS), it is strategically important to first understand their core mechanics. Both methods aim to explain black-box model predictions but rely on distinct theoretical frameworks that introduce specific vulnerabilities when applied to complex tabular data.

2.1 LIME: Local Interpretable Model-agnostic Explanations

LIME (Local Interpretable Model-agnostic Explanations) explains individual predictions by approximating the black-box model locally with an interpretable surrogate, typically a linear model (Arunraju Chinnaraju 2025; Amith Kumar Reddy 2024). It generates a neighborhood of perturbed samples around the instance, predicts outcomes using the original model, weights samples by proximity, fits a simple model to derive feature contributions, and the explanation for the resulting prediction is derived from the coefficients or rules of this local surrogate model (Wehner et al. 2025).

LIME’s model-agnostic nature has valued it as being broadly applicable to any underlying ML model (Amith Kumar Reddy 2024), but its reliance on random perturbations can lead to unstable explanations, posing challenges for reliability in high-stakes DSS (Arunraju Chinnaraju 2025).

2.2 SHAP: SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) uses cooperative game theory to assign each feature a Shapley value representing its contribution to a prediction relative to a baseline (Amith Kumar Reddy 2024). Its key properties—local accuracy, missingness, and consistency—ensure theoretically sound and reliable explanations (Suffian et al. 2022).

Exact Shapley value computation is NP-hard; practical approximations, such as KernelSHAP, apply a weighted linear regression based on the Shapley kernel. This approach preserves theoretical guarantees while remaining model-agnostic, but computational complexity and sensitivity to feature correlations can challenge its application in DSS with tabular data (Kolpaczki et al. 2025; Janzing et al. 2019).

2.3 Summary

Both LIME and SHAP offer interpretable insights for complex models, yet their operational assumptions—local linearity for LIME and additive feature contributions for SHAP—can

exacerbate challenges inherent in real-world tabular DSS (Amith Kumar Reddy 2024; Salih 2024b).

3.0 Data-Centric Vulnerabilities in SHAP and LIME Explanations

While the methodological foundations of SHAP and LIME are theoretically grounded in game theory and local surrogacy respectively, their practical reliability is heavily mediated by the characteristics of the input data (Arunraju Chinnaraju 2025; Ahmed et al. 2025). An explanation method cannot be evaluated in a vacuum; its performance is inextricably linked to the data distribution it seeks to interpret. This section systematically categorizes how common properties of data—ranging from structural dependencies to quality degradation—introduce instability, unreliability, and potentially misleading interpretations (Arunraju Chinnaraju 2025; Slack et al. 2020; Mesinovic et al. 2023; Carmichael and Scheirer 2023; Roberts et al. 2022).

3.1 Feature Dependence and the Collinearity Trap

A fundamental vulnerability of both SHAP and LIME is the assumption of feature independence, which is rarely met in real-world observational data (Ortigossa et al. 2024). **Marginalization Failure** occurs because KernelSHAP simulates a feature's "absence" by sampling from its marginal distribution, which is valid only when features are uncorrelated. In the presence of collinearity—such as between systolic and diastolic blood pressure—this process generates unrealistic, off-manifold instances within the coalitions (Salih et al. 2025; Mesinovic et al. 2023; Ortigossa et al. 2024; Kumar et al., n.d.). **Attribution Splitting** arises when collinearity causes SHAP to divide importance scores among correlated variables, obscuring the true influence of a single driver. In contrast, LIME's local linear model interprets feature weights as isolated effects, assuming other features remain constant—an assumption violated in highly dependent feature spaces where variables co-vary (Salih 2024b; 2024a; Ning et al. 2022). **Instability Metrics** such as the Normalized Movement Rate (NMR) and Modified Index Position (MIP) have been proposed to measure how collinearity induces fluctuations or inconsistent reordering of feature rankings across models (Ahmed et al. 2025; Kumar, n.d.; Doumard et al., n.d.).

3.2 Imputation Artifacts and Missing Data Mechanisms

The interpretability process is often a "black box" that includes the model-imputer pipeline, yet explanations are frequently misinterpreted as being of the model alone (Rudin 2019; Cappiello et al., n.d.). From a data centric perspective, missing data handling alters the statistical identity of the dataset by reshaping variance, dependencies, and latent structure. These data-level transformations act as confounding factors for explainability, as imputation methods attribute importance to patterns that may originate from reconstructed or inferred values rather than observed evidence. **The Confounding Variable** arises from the choice of imputation (Mean, MICE, or Deep Learning), which can influence the decision boundaries perceived by the explainer (Golchian and Wright 2025). Mean imputation may reduce variance and make informative features appear less important, MICE generally preserves inter-feature relationships

but introduces some uncertainty, and deep learning – based imputers can create synthetic patterns that might be highlighted by XAI (Cappiello et al., n.d.). If a dataset does not account for complex missingness mechanisms like Missing Not At Random (MNAR), the resulting explanations may reflect proxies for missingness rather than the underlying predictive relationships (Thomas et al. 2022). **Quality Gaps** are evident in surveys showing that many studies across clinical research, finance, and AI fail to report or justify missing data and rarely document their handling strategies. This lack of detailed dataset descriptions and explanation protocols compromise the generalizability and faithfulness of the subsequent explanations (Xin et al. 2025).

3.3 Class Imbalance and Decision Boundary Distortion

When the data distribution is imbalanced, models often underperform on the minority class, a common issue in fraud detection and rare disease diagnosis. This prompts the use of resampling techniques (e.g., SMOTE), which fundamentally reshape the data landscape and affect XAI interpretations (Ahmed et al. 2025; 2025; Keerthana et al. 2025; Vivek et al. 2024). These **synthetic boundary shifts** introduce artificial examples that modify the model's decision boundaries, causing feature attributions or explanations to reflect patterns learned from the synthetic data rather than the original imbalanced distribution, which can mislead interpretation and downstream decision-making (Li et al. 2022; Shuvo et al. 2025).

3.4 Sampling Stochasticity and OOD Perturbations

The reliability of explanations is highly sensitive to the sampling strategy used to define the "local neighborhood" (Amparore et al. 2021; Jesus et al. 2021). LIME, in particular, is affected by **randomized sampling variance**, where minor changes in the random seed or the number of perturbed samples can produce substantially different feature importance rankings for the same instance (Ortigossa et al. 2024; Slack et al. 2021). Moreover, **off-manifold perturbations**, such as graying out pixels or mean-imputing tabular cells, generate out-of-distribution (OOD) input. As a result, the model's responses to these unrealistic points can be unpredictable, causing the explainer to highlight artifacts that do not reflect the true reasoning process (Beechey et al. 2023; Ademi et al., n.d.).

3.5 Modality-Specific Gaps

Across data modalities, structural characteristics routinely undermine the assumptions of traditional XAI methods, leading to explanations that are misleading, non-credible, or contextually invalid (Abbas et al. 2025; Nascita et al. 2025).

Temporal data such as ECG signals or financial time series are often ignored when features are treated independently, thereby breaking temporal logic (Mesinovic et al. 2023; Abbas et al. 2025); **textual** data for example in NLP suffers from the loss of syntactic and contextual meaning due to bag-of-words-based or overly fragmented attribution methods (Lyu et al. 2024; Sakai and Lam 2025); **spatial** data such as in Medical Images explanations ignore spatial autocorrelation and fail to capture coherent physical structures (Aysel et al. 2025); **graph** data

(e.g., Social Networks, Molecules) is inadequately explained when relational edges and topological dependencies are overlooked (Doumard et al., n.d.); **multicollinear** tabular data leads perturbation-based methods to generate implausible instances; **multimodal** (e.g., images + text) systems fail to isolate the specific contribution of each modality, often completely obscuring non-visual signals (Pahud De Mortanges et al. 2024; Multimodal Intelligence as the Dominant Paradigm in 2026 AI Systems, n.d.; Silva and Keller 2024); **longitudinal** data (e.g., Health Records) is reduced to isolated snapshots that neglect trend-dependent semantics which cause failing to capture the diagnostic importance of shifting values over time (Lyu et al. 2024; Sakai and Lam 2025; Pahud De Mortanges et al. 2024); **omics** or **high-dimensional** data (e.g., Genomics) limit the capacity of standard XAI methods to represent complex biological pathways, leading to feature attributions that are formally correct but biologically uninformative or implausible (Pahud De Mortanges et al. 2024); **audio** data is misrepresented when harmonic superposition is flattened into static representations (Das and Rad 2020; Frommholz et al. 2023); and **hierarchical** or **nested** data - where data is usually grouped - is explained at the instance level without accounting for higher-level contextual dependencies (Sakai and Lam 2025; Dubey et al. 2024). Collectively, these limitations demonstrate a fundamental misalignment between generic XAI assumptions and modality-specific data structures.

3.6 Security and Privacy Considerations in Data-Centric Explanations

From a data-centric view, explainability methods, such as SHAP and LIME, introduce notable security and privacy vulnerabilities beyond conventional statistical concerns (Mia and Pritom 2025). Perturbation-based sampling exposes a surface for adversarial manipulation, enabling “fairwashing” attacks where models behave benignly on out-of-distribution (OOD) perturbed points while maintaining biased logic on real data, thereby concealing discriminatory patterns from auditors (Hegde et al. 2024). High-fidelity explanations can also leak sensitive information: stable feature attribution patterns may reveal individual behavioral or physiological routines (property inference) (Ezzeddine 2024), and low-entropy attributions can facilitate membership inference attacks, revealing whether specific records were used in training (Sharma et al. 2025; Shokri et al. 2017). Furthermore, the stochastic nature of sampling introduces integrity risks, as attackers can perform explanation poisoning by subtly altering the background dataset used in KernelSHAP, thereby shifting feature importance and misleading the explainer regarding the model’s true reliance on features (Hwang et al. 2025). These vulnerabilities highlight the necessity of integrating privacy-preserving mechanisms, such as differential privacy or entropy regularization, within the explanation process to safeguard both individual data contributions and the integrity of model interpretations.

Table 2. Summary Taxonomy of Data-Centric Failures

Vulnerability Category	Primary Data-Centric Cause	Resulting XAI Failure
Dependency	Multicollinearity / Interaction neglect	Misleading importance; split attributions.
Quality	MNAR Missingness / Imputation bias	Confounded model-imputer pipeline.
Distribution	Class Imbalance / SMOTE artifacts	Explanation of synthetic boundaries.
Integrity	Stochasticity / Off-manifold sampling	Unstable, non-deterministic explanations.
Modality	Temporal / Sequential links	Violation of data semantics (ECG/NLP).
Security	OOD Scaffolding / Fairwashing	Concealment of model bias from auditors.
Privacy	Memorized attribution patterns	Membership/Property inference leakage.

In practice, SHAP tends to provide more stable and theoretically consistent explanations, especially when aggregating local attributions for global insights, making it preferable in high-stakes tabular DSS like healthcare or finance (Arunraju Chinnaraju 2025). LIME, while faster and conceptually simpler, exhibits higher sensitivity to sampling and feature correlations, which necessitates careful interpretation when used in decision-critical contexts (Mesinovic et al. 2023).

4.0 Evaluation Frameworks and Practices

Evaluating the credibility of SHAP and LIME explanations requires both quantitative and qualitative perspectives. Key quantitative metrics include fidelity, stability, robustness, completeness, and simplicity, which measure how well the explanations reflect the model's decision process (Amparore et al. 2021; 2021; Qureshi et al., n.d.; Nazat et al. 2024; Nakanishi 2026). Qualitative and human-centered assessments focus on trust, usefulness, and

comprehensibility, ensuring that explanations are actionable for domain experts in DSS contexts (Kabir et al. 2025; Arunraju Chinnaraju 2025; Amith Kumar Reddy 2024).

Importantly, a data-centric lens emphasizes that these evaluation metrics must consider data properties—such as feature correlation, missingness, class imbalance, and preprocessing choices—as these factors directly influence explanation reliability. Standardized protocols combining computational metrics with user-centered assessment remain an open challenge for achieving reproducible and practical explainability.

5.0 Conclusion and Future Directions

This data-centric literature review has examined SHAP and LIME, two of the most influential post-hoc explanation methods for tabular data in Decision Support Systems. The analysis reveals that while these methods are widely used for interpreting complex machine learning models, their reliability is highly dependent on the characteristics of the underlying data. Challenges such as feature correlation, data quality, class imbalance, and sampling instability are central determinants of explanation fidelity. In practice, SHAP generally provides more stable and theoretically consistent explanations, particularly when aggregating local attributions to obtain global insights, whereas LIME offers computational efficiency and intuitive local explanations but is more sensitive to sampling variability and feature dependencies. Collectively, these findings underscore that achieving robust and trustworthy explainability in tabular DSS requires careful attention to data preprocessing and management in addition to the selection of the explanation method itself.

To advance practically reliable explainers, several directions emerge from the literature:

- **Development of Data-Aware Explanation Methods:** Future research should focus on designing XAI techniques that explicitly account for feature dependencies, class imbalance, and preprocessing-induced uncertainties. Data-aware explainers would contextualize model predictions relative to the properties and limitations of the dataset, thereby improving the fidelity and interpretability of explanations in applied DSS.
- **Establishment of Standardized, Data-Centric Evaluation Protocols:** Evaluation frameworks should integrate quantitative metrics—such as fidelity, stability, robustness, and completeness—with human-centered assessments that consider the influence of data characteristics on explanation reliability. Standardized, data-aware protocols are essential for reproducible and actionable evaluation of XAI methods across domains.
- **Addressing Security and Robustness:** As XAI becomes embedded in decision-critical pipelines, it is important to develop methods resilient to adversarial manipulation, model inversion, or fairwashing. Ensuring that explanations accurately reflect model behavior without concealing biases or vulnerabilities is crucial for maintaining trust in high-stakes DSS.

- **Enhancing Scalability for Real-World Deployment:** Computational demands, particularly for methods like KernelSHAP, remain a barrier to real-time deployment in large-scale DSS environments. Research into efficient, scalable, and accurate approximations is necessary to enable timely and transparent explanations in applications requiring low-latency decisions, such as dynamic clinical monitoring or fraud detection.

By highlighting the interplay between data characteristics and explanation methods, this review highlights that effective, reliable, and actionable explainability in tabular DSS is fundamentally a data-driven challenge. Future work that synthesizes algorithmic innovation with rigorous data-centric considerations will be critical to advancing trustworthy AI in real-world decision-making contexts.

6.0 References

- Abbas, Qaiser, Woonyoung Jeong, and Seung Won Lee. 2025. "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges." *Healthcare* 13 (17): 2154. <https://doi.org/10.3390/healthcare13172154>.
- Ademi, Luan, Maximilian Noppel, and Christian Wressnegger. n.d. **POMELO: Black-Box Feature Attribution with Full-Input, In-Distribution Perturbations.**
- Ahmed, Shamim, M. Shamim Kaiser, Mohammad Shahadat Hossain, and Karl Andersson. 2025. "A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions." *IEEE Access* 13: 37370–88. <https://doi.org/10.1109/ACCESS.2024.3422319>.
- Amith Kumar Reddy. 2024. "Bridging AI and Human Understanding: Interpretable Deep Learning in Practice." *Journal of Informatics Education and Research* 4 (3). <https://doi.org/10.52783/jier.v4i3.2200>.
- Amparore, Elvio G., Alan Perotti, and Paolo Bajardi. 2021. "To Trust or Not to Trust an Explanation: Using LEAF to Evaluate Local Linear XAI Methods." *PeerJ Computer Science* 7 (April): e479. <https://doi.org/10.7717/peerj-cs.479>.
- Arunraju Chinnaraju. 2025. "Explainable AI (XAI) for Trustworthy and Transparent Decision-Making: A Theoretical Framework for AI Interpretability." *World Journal of Advanced Engineering Technology and Sciences* 14 (3): 170–207. <https://doi.org/10.30574/wjaets.2025.14.3.0106>.
- Aysel, Halil Ibrahim, Xiaohao Cai, and Adam Prugel-Bennett. 2025. "Explainable Artificial Intelligence: Advancements and Limitations." *Applied Sciences* 15 (13): 7261. <https://doi.org/10.3390/app15137261>.

- Batko, Kornelia, and Andrzej Ślęzak. 2022. "The Use of Big Data Analytics in Healthcare." *Journal of Big Data* 9 (1): 3. <https://doi.org/10.1186/s40537-021-00553-4>.
- Beechey, Daniel, Thomas M. S. Smith, and Özgür Şimşek. 2023. "Explaining Reinforcement Learning with Shapley Values." arXiv:2306.05810. Preprint, arXiv, June 9. <https://doi.org/10.48550/arXiv.2306.05810>.
- Cappiello, Cinzia, Federico Cerutti, Camilla Sancricca, and Riccardo Zanelli. n.d. **About the Effects of Data Imputation Techniques on ML Uncertainty.**
- Carmichael, Zachariah, and Walter J. Scheirer. 2023. "How Well Do Feature-Additive Explainers Explain Feature-Additive Predictors?" arXiv:2310.18496. Preprint, arXiv, October 27. <https://doi.org/10.48550/arXiv.2310.18496>.
- Černevičienė, Jurgita, and Audrius Kabašinskas. 2024. "Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review." *Artificial Intelligence Review* 57 (8): 216. <https://doi.org/10.1007/s10462-024-10854-8>.
- Das, Arun, and Paul Rad. 2020. "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey." arXiv:2006.11371. Preprint, arXiv, June 23. <https://doi.org/10.48550/arXiv.2006.11371>.
- Doumard, Emmanuel, Julien Aligon, Elodie Escriva, Jean-Baptiste Excoffier, Paul Monsarrat, and Chantal Soulé-Dupuy. n.d. **A Comparative Study of Additive Local Explanation Methods Based on Feature Influences.**
- Dubey, Akshat, Zewen Yang, and Georges Hattab. 2024. "A Nested Model for AI Design and Validation." *iScience* 27 (9): 110603. <https://doi.org/10.1016/j.isci.2024.110603>.
- Ezzeddine, Fatima. 2024. "Privacy Implications of Explainable AI in Data-Driven Systems." arXiv:2406.15789. Preprint, arXiv, June 22. <https://doi.org/10.48550/arXiv.2406.15789>.
- Frommholz, Annika, Fabian Seipel, Sebastian Lapuschkin, Wojciech Samek, and Johanna Vielhaben. 2023. "XAI-Based Comparison of Input Representations for Audio Event Classification." arXiv:2304.14019. Preprint, arXiv, April 27. <https://doi.org/10.48550/arXiv.2304.14019>.
- Golchian, Pegah, and Marvin N. Wright. 2025. "Imputation Uncertainty in Interpretable Machine Learning Methods." arXiv:2512.17689. Preprint, arXiv, December 19. <https://doi.org/10.48550/arXiv.2512.17689>.
- Guan, Li, José M. Merigó, and Ghassan Beydoun. 2025. "40 Years of Decision Support Systems: A Bibliometric Analysis." *Decision Support Systems* 194 (July): 114469. <https://doi.org/10.1016/j.dss.2025.114469>.
- Hegde, Achyut, Maximilian Noppel, and Christian Wressnegger. 2024. "Model-Manipulation Attacks Against Black-Box Explanations." 2024 Annual Computer Security Applications

- Conference (ACSAC), December 9, 974–87.
<https://doi.org/10.1109/ACSAC63791.2024.00081>.
- Hwang, Hyunseung, Andrew Bell, Joao Fonseca, Venetia Pliatsika, Julia Stoyanovich, and Steven Euijong Whang. 2025. “**SHAP-Based Explanations Are Sensitive to Feature Representation.**” arXiv:2505.08345. Preprint, arXiv, May 13.
<https://doi.org/10.48550/arXiv.2505.08345>.
- Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. 2019. “**Feature Relevance Quantification in Explainable AI: A Causal Problem.**” arXiv:1910.13413. Preprint, arXiv, November 27. <https://doi.org/10.48550/arXiv.1910.13413>.
- Jesus, Sérgio, Catarina Belém, Vladimir Balayan, et al. 2021. “**How Can I Choose an Explainer? An Application-Grounded Evaluation of Post-Hoc Explanations.**” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 3, 805–15. <https://doi.org/10.1145/3442188.3445941>.
- Kabir, Sami, Mohammad Shahadat Hossain, and Karl Andersson. 2025. “**A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities.**” Algorithms 18 (9): 556. <https://doi.org/10.3390/a18090556>.
- Keerthana, Chirumamilla Satya, Siri Chandana Nalluri, Simrah Muskaan, and Poorvie Sadagopan. 2025. “**Explainable AI in Credit Card Fraud Detection: SHAP and LIME for Machine Learning Models.**” 2025 10th International Conference on Signal Processing and Communication (ICSC), February 20, 387–92.
<https://doi.org/10.1109/ICSC64553.2025.10968935>.
- Kolpaczki, Patrick, Tim Nielen, and Eyke Hüllermeier. 2025. “**Antithetic Sampling for Top-k Shapley Identification.**” arXiv:2504.02019. Preprint, arXiv, July 22.
<https://doi.org/10.48550/arXiv.2504.02019>.
- Kumar, I Elizabeth, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A Friedler. n.d. **Problems with Shapley-Value-Based Explanations as Feature Importance Measures.**
- Kumar, Indra Elizabeth. n.d. **Explainability, Fairness, and Evaluation in Machine Learning: From Theory to Policy and Back.**
- Li, Yiming, Wei-Wen Hsu, and for the Alzheimer’s Disease Neuroimaging Initiative. 2022. “**A Classification for Complex Imbalanced Data in Disease Screening and Early Diagnosis.**” Statistics in Medicine 41 (19): 3679–95. <https://doi.org/10.1002/sim.9442>.
- Lyu, Daoming, Xingbo Wang, Yong Chen, and Fei Wang. 2024. “**Language Model and Its Interpretability in Biomedicine: A Scoping Review.**” iScience 27 (4): 109334.
<https://doi.org/10.1016/j.isci.2024.109334>.

- Martens, David, Galit Shmueli, Theodoros Evgeniou, et al. 2025. **"Beware of 'Explanations' of AI."** arXiv:2504.06791. Preprint, arXiv, April 9. <https://doi.org/10.48550/arXiv.2504.06791>.
- Mesinovic, Munib, Peter Watkinson, and Tingting Zhu. 2023. **"Explainable AI for Clinical Risk Prediction: A Survey of Concepts, Methods, and Modalities."** arXiv:2308.08407. Preprint, arXiv, August 16. <https://doi.org/10.48550/arXiv.2308.08407>.
- Mia, Maraz, and Mir Mehedi A. Pritom. 2025. **"Explainable but Vulnerable: Adversarial Attacks on XAI Explanation in Cybersecurity Applications."** arXiv:2510.03623. Preprint, arXiv, October 4. <https://doi.org/10.48550/arXiv.2510.03623>.
- Multimodal Intelligence as the Dominant Paradigm in 2026 AI Systems. n.d.
- Nakanishi, Takafumi. 2026. **"CausalAIME: Leveraging Peter-Clark Algorithms and Inverse Modeling for Unified Global Feature Explanation in Healthcare."** In Explainable Artificial Intelligence, edited by Riccardo Guidotti, Ute Schmid, and Luca Longo, vol. 2577. Communications in Computer and Information Science. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-08324-1_15.
- Nascita, Alfredo, Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio Persico, and Antonio Pescapé. 2025. **"A Survey on Explainable Artificial Intelligence for Internet Traffic Classification and Prediction, and Intrusion Detection."** IEEE Communications Surveys & Tutorials 27 (5): 3165–98. <https://doi.org/10.1109/COMST.2024.3504955>.
- Nazat, Sazid, Osvaldo Arreche, and Mustafa Abdallah. 2024. **"On Evaluating Black-Box Explainable AI Methods for Enhancing Anomaly Detection in Autonomous Driving Systems."** Sensors 24 (11): 3515. <https://doi.org/10.3390/s24113515>.
- Ning, Yilin, Marcus Eng Hock Ong, Bibhas Chakraborty, et al. 2022. **"Shapley Variable Importance Cloud for Interpretable Machine Learning."** Patterns 3 (4): 100452. <https://doi.org/10.1016/j.patter.2022.100452>.
- O'Brien Quinn, Helen, Mohamed Sedky, Janet Francis, and Michael Streeton. 2024. **"Literature Review of Explainable Tabular Data Analysis."** Electronics 13 (19): 3806. <https://doi.org/10.3390/electronics13193806>.
- Ortigossa, Evandro S., Thales Gonçalves, and Luis Gustavo Nonato. 2024. **"EXplainable Artificial Intelligence (XAI)—From Theory to Methods and Applications."** IEEE Access 12: 80799–846. <https://doi.org/10.1109/ACCESS.2024.3409843>.
- Pahud De Mortanges, Aurélie, Haozhe Luo, Shelley Zixin Shu, et al. 2024. **"Orchestrating Explainable Artificial Intelligence for Multimodal and Longitudinal Data in Medical Imaging."** Npj Digital Medicine 7 (1): 195. <https://doi.org/10.1038/s41746-024-01190-w>.

- Qureshi, M Atif, Abdul Aziz Noor, Awais Manzoor, Deedahwar Mazhar Qureshi, Arjumand Younus, and Wael Rashwan. n.d. **Explainability in Action: A Metric-Driven Assessment of Five XAI Methods for Healthcare Tabular Models.**
- Roberts, Claudia V., Ehtsham Elahi, and Ashok Chandrashekar. 2022. **"On the Bias-Variance Characteristics of LIME and SHAP in High Sparsity Movie Recommendation Explanation Tasks."** arXiv:2206.04784. Preprint, arXiv, June 9. <https://doi.org/10.48550/arXiv.2206.04784>.
- Rudin, Cynthia. 2019. **"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead."** arXiv:1811.10154. Preprint, arXiv, September 22. <https://doi.org/10.48550/arXiv.1811.10154>.
- Sakai, Hajar, and Sarah S. Lam. 2025. **"Large Language Models for Healthcare Text Classification: A Systematic Review."** arXiv:2503.01159. Preprint, arXiv, March 3. <https://doi.org/10.48550/arXiv.2503.01159>.
- Salih, Ahmed M. 2024a. **"Explainable Artificial Intelligence and Multicollinearity: A Mini Review of Current Approaches."** arXiv:2406.11524. Preprint, arXiv, June 17. <https://doi.org/10.48550/arXiv.2406.11524>.
- Salih, Ahmed M. 2024b. **"Explainable Artificial Intelligence for Dependent Features: Additive Effects of Collinearity."** arXiv:2411.00846. Preprint, arXiv, October 30. <https://doi.org/10.48550/arXiv.2411.00846>.
- Salih, Ahmed, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, et al. 2025. **"A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME."** Advanced Intelligent Systems 7 (1): 2400304. <https://doi.org/10.1002/aisy.202400304>.
- Sharma, Dilli Prasad, Xiaowei Sun, Liang Xue, Xiaodong Lin, and Pulei Xiong. 2025. **"Privacy-Preserving Explainable AIoT Application via SHAP Entropy Regularization."** arXiv:2511.09775. Preprint, arXiv, November 12. <https://doi.org/10.48550/arXiv.2511.09775>.
- Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. **"Membership Inference Attacks against Machine Learning Models."** arXiv:1610.05820. Preprint, arXiv, March 31. <https://doi.org/10.48550/arXiv.1610.05820>.
- Shuvo, Mrinal Basak, Arjon Talukder, Sadia Islam Neela, Pankaj Bhowmik, and Md. Delowar Hossain. 2025. **"An Adaptive AI Approach for Cervical Cancer Prediction with Explainability."** 2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), June 27, 1–6. <https://doi.org/10.1109/NCIM65934.2025.11160255>.
- Silva, Sam J., and Christoph A. Keller. 2024. **"Limitations of XAI Methods for Process-Level Understanding in the Atmospheric Sciences."** Artificial Intelligence for the Earth Systems 3 (1): e230045. <https://doi.org/10.1175/AIES-D-23-0045.1>.

- Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. **"Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods."** arXiv:1911.02508. Preprint, arXiv, February 3. <https://doi.org/10.48550/arXiv.1911.02508>.
- Slack, Dylan, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. **"Reliable Post Hoc Explanations: Modeling Uncertainty in Explainability."** arXiv:2008.05030. Preprint, arXiv, November 6. <https://doi.org/10.48550/arXiv.2008.05030>.
- Suffian, Muhammad, Pierluigi Graziani, Jose M. Alonso, and Alessandro Bogliolo. 2022. **"FCE: Feedback Based Counterfactual Explanations for Explainable AI."** IEEE Access 10: 72363–72. <https://doi.org/10.1109/ACCESS.2022.3189432>.
- Thomas, Diana M., Samantha Kleinberg, Andrew W. Brown, et al. 2022. **"Machine Learning Modeling Practices to Support the Principles of AI and Ethics in Nutrition Research."** Nutrition & Diabetes 12 (1): 48. <https://doi.org/10.1038/s41387-022-00226-y>.
- Vivek, Yelleti, Vadlamani Ravi, Abhay Mane, and Laveti Ramesh Naidu. 2024. **"Explainable Artificial Intelligence and Causal Inference Based ATM Fraud Detection."** 2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFER), October 22, 1–7. <https://doi.org/10.1109/CIFER62890.2024.10772906>.
- Weber, Patrick, K. Valerie Carl, and Oliver Hinz. 2024. **"Applications of Explainable Artificial Intelligence in Finance—a Systematic Review of Finance, Information Systems, and Computer Science Literature."** Management Review Quarterly 74 (2): 867–907. <https://doi.org/10.1007/s11301-023-00320-0>.
- Wehner, Nikolas, Anika Seufert, Tobias Hoßfeld, and Michael Seufert. 2025. **"A Tutorial on Data-Driven Quality of Experience Modeling With Explainable Artificial Intelligence."** IEEE Communications Surveys & Tutorials, 1–1. <https://doi.org/10.1109/COMST.2025.3583227>.
- Xin, Yu, Ruomeng Song, Jun Hao, et al. 2025. **"Poor Reporting Quality and High Proportion of Missing Data in Economic Evaluations alongside Pragmatic Trials: A Cross-Sectional Survey."** BMC Medical Research Methodology 25 (1): 61. <https://doi.org/10.1186/s12874-025-02519-z>.