Proceedings of CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF34929

A Review of the Application and Challenges of Shapley Additive Explanations in Data Analysis

Jiayi Ji

Computer Science and Technology, Tongji University, Shanghai, China
2352976@tongji.edu.cn

Abstract. The rapid growth of big data and complex machine learning models—gradient-boosted trees and deep neural networks—has produced highly accurate but opaque "black-box" predictors across medicine, finance, and industry, making interpretability a central concern in data analysis. SHapley Additive exPlanations (SHAP), grounded in cooperative game theory, has become one of the most influential interpretability methods because it provides theoretically consistent feature attributions at both the local and global levels. This paper presents a systematic literature review of SHAP and its role in data analysis. It synthesizes SHAP's theoretical foundations, its main implementations (TreeSHAP, KernelSHAP, and DeepSHAP), its visualization toolkit, and its practical applications, and it reports a compact empirical study comparing the three explainers on a clinical dataset. This study finds that, although SHAP markedly improves transparency and decision support, open challenges remain in computational cost, the reliability of explanations under feature correlation, and consistency across methods. The significance of this work is twofold: theoretically, it organizes SHAP's variants and properties within a single coherent framework; practically, it offers data analysts a structured, evidence-based reference for selecting and applying SHAP appropriately, thereby supporting more transparent, reliable, and accountable model-driven decisions.

Keywords: SHAP, Explainable Machine Learning, Data Analysis, Shapley Value, Feature Attribution

1. Introduction

In recent years, the convergence of big data and machine learning has reshaped decision-making in medicine, finance, urban planning, and industry. High-capacity models such as XGBoost [1], LightGBM [2], and deep neural networks deliver state-of-the-art accuracy, yet their internal reasoning is largely inscrutable. Such "black-box" behavior raises concerns about trust, fairness, regulatory compliance, and error diagnosis, making interpretable machine learning a research priority [3]. Within this field, SHapley Additive exPlanations (SHAP), proposed by Lundberg and Lee [4], unifies several earlier attribution methods under a single game-theoretic framework and yields a unique solution with desirable consistency properties, making it one of the most widely used explanation tools today.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GF34929

Research on SHAP has advanced along three lines: theoretical work refining and extending Shapley-value estimation, including the exact polynomial-time TreeSHAP algorithm [5]; applied work using SHAP for feature selection, model debugging, and decision support; and visualization work developing summary, dependence, and force plots. However, existing studies remain fragmented, and issues of computational cost, correlated features, and inconsistency across variants are still unresolved.

To address this gap, this paper reviews SHAP's theoretical basis, implementations, visualization methods, and applications, and then validates the resulting picture through a compact empirical study before discussing the open challenges. Theoretically, it organizes SHAP's variants within a coherent framework; practically, it offers analysts a structured reference for applying SHAP appropriately, supporting more transparent and accountable model-driven decisions.

2. Research status at home and abroad

Research on SHAP spans three streams—theory, application, and visualization—alongside a growing critical literature on its limitations.

2.1. Theoretical research on SHAP

SHAP derives from the Shapley value in cooperative game theory, which fairly distributes a coalition's payoff among players by their marginal contributions [6]. Lundberg and Lee adapted this idea to machine learning, treating features as players and showing that methods such as LIME [7] and DeepLIFT form a single class of additive attributions whose unique solution satisfies local accuracy, missingness, and consistency [4]. Later work improved estimation: TreeSHAP computes exact Shapley values for tree ensembles in polynomial time [5], while studies of the independence assumption showed that correlated features can yield misleading attributions, motivating conditional approaches [8].

2.2. Applied research on SHAP

SHAP has been applied wherever accuracy and interpretability are both required. In medicine, it explains clinical risk models in real time—for instance, a system predicting intraoperative hypoxaemia and exposing its contributing factors to support anaesthesiologists [9]. In finance, it aids credit scoring and fraud detection; in industry, fault diagnosis and quality prediction; and in urban and environmental analysis, models of traffic, energy demand, and pollution.

2.3. Visualization research on SHAP

Research has also enriched SHAP's visualizations: summary and bar plots give a global importance view, dependence plots reveal how a feature's effect varies and interacts, and force or waterfall plots decompose individual predictions [5]. Recent work integrates these into interactive dashboards that make SHAP accessible to non-specialists.

2.4. Limitations of existing research

Despite this progress, limitations persist. The literature is fragmented, with few integrative reviews oriented toward general data-analysis practice; the reliability of explanations under feature correlation is contested because common approximations assume independence [8]; and conceptual

DOI: 10.54254/2755-2721/2026.GF34929

critiques argue that Shapley-value attributions may not match human-centric explanation goals and that fixing their shortcomings adds complexity, such as the need for causal reasoning [10].

3. Theoretical foundations of SHAP

3.1. Shapley value and cooperative game theory

The Shapley value allocates the total value created by a coalition among its players [6]. Given M players and a characteristic function $v(S)$ assigning a worth to each subset S, player i's Shapley value is the weighted average of its marginal contribution over all coalitions S that exclude i:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M-|S|-1)!}{M!} \bigl[v\left(S \cup \{i\}\right) - v\left(S\right) \bigr] \tag{1}$$

It is the unique allocation satisfying efficiency, symmetry, the null-player property, and additivity. In SHAP, the features are the players, $v(S)$ is the model's expected output given only the features in S, and the Shapley value φ_i is feature i's contribution to a prediction, distributing the gap between the prediction and the average prediction exactly among the features [4].

3.2. SHAP explanation mechanism

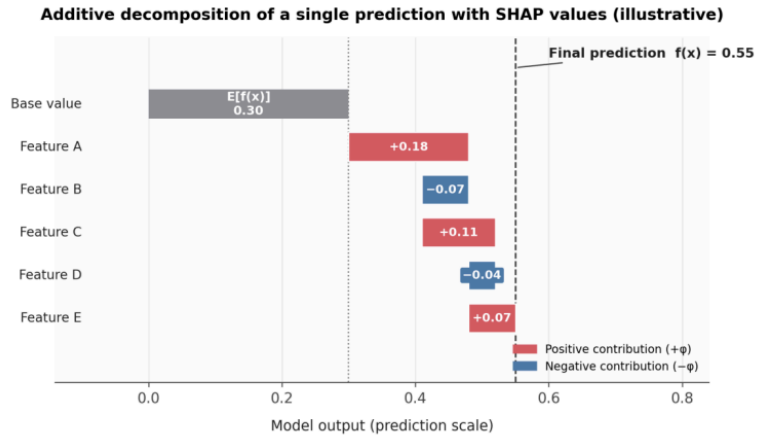

Figure 1. Conceptual illustration of how SHAP additively decomposes a single prediction into a base value E[f(x)] and signed feature contributions that sum to the model output f(x). Values are illustrative

SHAP frames an explanation as an additive feature-attribution model:

$$g\left(z'\right) = \varphi_0 + \sum_{i=1}^{M} \varphi_i z'_i \tag{2}$$

where $\varphi_0 = E[f(x)]$ is the base value and φ_i is the SHAP value of feature i. Lundberg and Lee proved that requiring local accuracy, missingness, and consistency singles out the Shapley values as

DOI: 10.54254/2755-2721/2026.GF34929

the unique solution [4]. This signed, additive decomposition lets a single prediction be read as a base value plus positive and negative feature contributions, as illustrated in Figure 1.

3.3. Main implementations of SHAP

Since exact computation is exponential in the number of features, SHAP relies on several estimators. KernelSHAP is model-agnostic, approximating Shapley values via a weighted linear model fitted over feature subsets, but it is computationally costly and assumes feature independence [4]. TreeSHAP computes exact Shapley values for tree ensembles in polynomial time, making it both fast and accurate for these widely used models [5]. DeepSHAP adapts the approach to deep neural networks by combining SHAP with DeepLIFT-style propagation [4]. The choice among them depends on the model type, the required accuracy, and the available computational budget.

4. Visualization and practical applications of SHAP

Having established the theory, this section turns to how SHAP is used in practice. It is organized around the logic of tool and value: first, the visualization toolkit that turns raw attributions into human-readable evidence, then the generic roles SHAP plays in the data-analysis workflow, and finally its impact in two high-stakes domains.

4.1. Visualization toolkit for multi-level interpretation

SHAP values become actionable only when communicated visually. Four complementary plot families cover the interpretive questions that arise in practice, ranging from global model understanding to the explanation of a single decision; Table 1 summarizes what each plot solves and where it fits best.

Table 1. The four core SHAP visualizations and the interpretive question each one answers

Visualization Type	Core Function	Adaptable Scenario
Feature importance(bar / beeswarm)	Aggregates absolute SHAP values to rank features and show each feature's value distribution and effect direction	Global model understanding; feature screening and reporting
Local explanation(force / waterfall)	Decomposes one prediction from the base value into signed feature contributions	Instance-level accountability in high-stakes decisions
Feature relationship(dependence)	Plots a feature's value against its SHAP value, optionally colored by a second feature, to expose nonlinear, threshold, and interaction effects	Diagnosing how and why a feature moves the prediction
Interactive explanation(dashboard)	Embeds the plots above in filterable, drill-down interfaces linking global summaries to individual cases	Exploratory analysis and communication with non-specialists

4.2. Generic applications in the data-analysis workflow

Beyond explaining predictions, SHAP serves as an active instrument throughout the analysis pipeline. In feature engineering, ranking variables by their aggregated contribution guides feature selection and dimensionality reduction, and redundant variables tend to reveal themselves through

DOI: 10.54254/2755-2721/2026.GF34929

shared or unstable contributions; this reflects actual model behavior more faithfully than correlation-based filters, as demonstrated when SHAP-based ranking is used to prune inputs of tree ensembles without loss of accuracy [5]. In model diagnosis and optimization, inspecting which features drive predictions exposes bias—such as undue reliance on a sensitive attribute—and anomalous contributions that signal data leakage or spurious correlations, while the SHAP decomposition of mispredicted instances helps trace the sources of error [4]. In anomaly detection, SHAP not only flags a suspicious instance but attributes its deviation to specific features, turning an opaque anomaly score into an auditable explanation suited to investigation and regulatory review.

4.3. Domain-specific high-stakes scenarios

The value of SHAP is clearest in domains where an unexplained decision is unacceptable. Two representative high-stakes fields illustrate its practical effect.

Healthcare. Because clinical decisions demand justification, SHAP lets clinicians see the patient-specific factors behind a risk prediction and weigh them against medical knowledge. In the real-time hypoxaemia-prevention system of Lundberg et al. [9], SHAP explanations attached to the risk model raised anaesthesiologists' ability to anticipate intraoperative hypoxaemia and exposed the contributing factors at the moment of decision, demonstrating that interpretability can improve, rather than merely document, expert performance.

Finance. Credit scoring, fraud detection, and risk modeling operate under strict transparency and fairness regulations. Here, SHAP provides per-decision attributions that explain why an individual application was approved or flagged, supporting adverse-action reasoning and audit while helping institutions detect when a model leans on proxy or sensitive variables. In both domains, the same property is decisive: SHAP yields explanations that domain experts can validate against prior knowledge, which is what enables real-world adoption.

5. Empirical study: comparative analysis of SHAP variants

5.1. Experimental setup

To ground the preceding review in concrete evidence, this section reports a compact, reproducible case study comparing the three principal SHAP explainers on a representative tabular classification task. The Breast Cancer Wisconsin (Diagnostic) dataset [11] was used: 569 samples, 30 continuous morphological features, and a binary (malignant vs. benign) label. The data were split into training and test subsets with a stratified split and standardized before training. Two models of comparable quality were trained—a gradient boosting classifier (scikit-learn [12]) and a multilayer perceptron (PyTorch [13])—so that differences between explainers would not be confounded by differences in model skill. The tree model was explained with TreeSHAP and the network with DeepSHAP, each benchmarked against KernelSHAP as a common, model-agnostic reference. A fixed seed of 42, fifty background instances, and twenty explained test instances were used throughout. The explainers were assessed on three criteria: runtime (s), local accuracy (mean absolute reconstruction error of the additive decomposition), and consistency with KernelSHAP (Spearman rank correlation and Top-5 feature overlap).

DOI: 10.54254/2755-2721/2026.GF34929

5.2. Results and analysis

Table 2 reports model performance and Table 3 compares the explainers. The two models were essentially equivalent in predictive quality, so the comparison isolates the behavior of the explainers themselves. Figure 2 and Figure 3 show, respectively, the global feature-importance ranking and a representative local explanation for the tree model.

Table 2. Predictive performance of the two models on the test set

Model	Accuracy	ROC-AUC
GradientBoostingClassifier	0.9580	0.9925
PyTorch MLP	0.9580	0.9884

Table 3. Comparison of the SHAP explainers in runtime, local accuracy, and consistency with KernelSHAP

Model	Explainer	Runtime (s)	Local-acc. MAE	Spearman	Top-5 overlap
Tree model	TreeSHAP	0.0075	0.0000	—	—
Tree model	KernelSHAP (on tree)	1.2869	0.0000	—	—
Neural network	DeepSHAP	0.0405	0.0000	—	—
Neural network	KernelSHAP (on MLP)	1.4419	0.0000	—	—
Tree model	TreeSHAP vs. KernelSHAP	—	—	0.8861	0.8000
Neural network	DeepSHAP vs. KernelSHAP	—	—	0.9021	0.4000

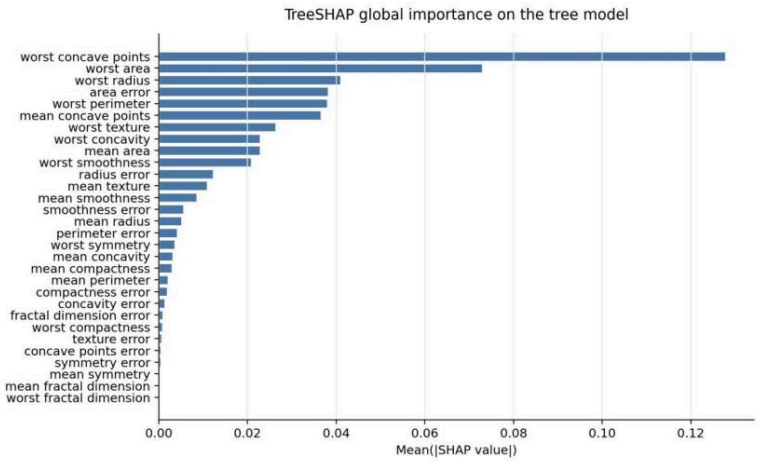

Figure 2. TreeSHAP global feature-importance ranking on the gradient boosting model

Three findings stand out. First, the model-specific explainers were dramatically faster: TreeSHAP ran about 171 times faster than KernelSHAP on the tree model (0.0075 s vs. 1.2869 s) and DeepSHAP about 36 times faster on the network (0.0405 s vs. 1.4419 s), confirming that exploiting model structure yields large computational savings [5]. Second, all four explainers achieved a local-accuracy error of essentially zero, so the additive decomposition faithfully reproduced the model outputs. Third, the explainers agreed strongly but not perfectly: the tree-model comparison gave a Spearman correlation of 0.8861 and a top-5 overlap of 0.8000, whereas the network comparison

DOI: 10.54254/2755-2721/2026.GF34929

gave a higher rank correlation (0.9021) but a lower top-5 overlap (0.4000)—the methods identified the same principal drivers (worst concave points, worst area) while differing on the ordering of secondary features. The global and local plots reinforce this reading: both models concentrated importance on a small set of morphology-related variables consistent with the clinical basis of the task.

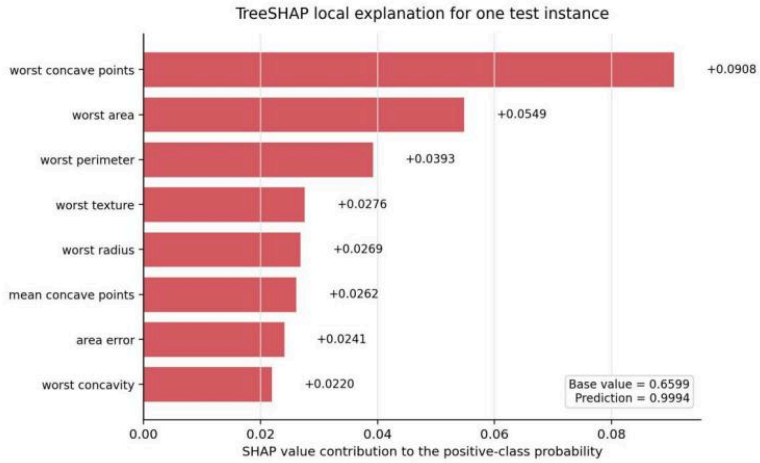

Figure 3. TreeSHAP local explanation for a representative test instance, decomposing the predicted positive-class probability into individual feature contributions

5.3. Theoretical and practical implications

Theoretically, the experiment confirms the additive-attribution guarantee in practice: zero reconstruction error across all explainers shows that local accuracy holds empirically, and the strong cross-method rank agreement supports the claim that the Shapley solution is a stable target that different estimators approximate rather than redefine. Practically, the results sharpen the central trade-off of the review. Model-specific explainers (TreeSHAP, and DeepSHAP) should be preferred whenever the model family supports them because their speed makes repeated or large-scale interpretation feasible; KernelSHAP remains valuable as a universal baseline but is costly in high-dimensional or high-throughput settings. The divergence in secondary-feature ranking also warns that explanation quality cannot be judged by a single global plot: runtime, reconstruction fidelity, and rank consistency should all inform the choice of an explainer for a given analytical task.

6. Conclusion

This review set out to clarify how SHAP improves the transparency of modern machine-learning models and where its use remains limited. Drawing the threads together, three core findings emerge. First, SHAP rests on an unusually firm theoretical footing: by transferring the Shapley value from cooperative game theory to model interpretation, it produces additive feature attributions that uniquely satisfy local accuracy, missingness, and consistency, which is what distinguishes it from earlier, heuristic explanation methods. Second, this single foundation supports a practical ecosystem —model-agnostic and model-specific estimators paired with a coherent family of visualizations— that lets the same conceptual explanation serve feature engineering, model diagnosis, anomaly detection, and high-stakes decision-making in fields such as medicine and finance. Third, our empirical study showed that these promises hold in practice while exposing meaningful differences

DOI: 10.54254/2755-2721/2026.GF34929

among variants: the explainers reconstructed model outputs exactly and agreed on the dominant predictive signals, yet the model-specific methods were one to two orders of magnitude faster than the model-agnostic baseline and the variants diverged on secondary-feature rankings.

Several challenges nonetheless remain open. Computing exact attributions is expensive for model-agnostic estimation on high-dimensional data; the common assumption of feature independence can distort attributions when features are correlated; different variants and background choices can yield inconsistent results that hinder reproducibility; and Shapley attributions describe statistical association rather than causal effect, so they may not always match a human user's intuitive notion of an explanation.

These open problems point to clear directions for future work. Priority should go to more efficient and dependence-aware estimators that relax the independence assumption without prohibitive cost, to standardized protocols for evaluating explanation quality so that methods can be compared on more than visual appeal, and to a closer integration of causal reasoning so that attributions can speak to intervention rather than mere correlation. Pursued together, these advances would allow SHAP to deliver explanations that are not only consistent and faithful but also reproducible and genuinely useful for human decision-making.

References

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
[2] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
[3] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
[4] Lundberg, S. M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
[5] Lundberg, S. M., Erion, G., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.
[6] Shapley, L. S. (1953). A value for n-person games. H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the Theory of Games. Princeton University Press, Vol. II, pp. 307–317.
[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.
[8] Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502.
[9] Lundberg, S. M., Nair, B., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2(10), 749–760.
[10] Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. Proceedings of the 37th International Conference on Machine Learning, PMLR 119, pp. 5491–5500.
[11] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. Biomedical Image Processing and Biomedical Visualization, Proc. SPIE, Vol. 1905, pp. 861–870.
[12] Pedregosa, F., Varoquaux, G., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[13] Paszke, A., Gross, S., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32, 8024–8035.