Acta Informatica Pragensia
2026, Volume 15, Issue 2, 364–381
https://doi.org/10.18267/j.aip.306

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Article Open Access

Corr-SHAP: Correlation-Aware Sampling for Faithful SHAP Value Estimation

Ridha El Hamdi [1,2], **Hana Charaabi** [1,3], **Ibtissam Hdhiri** [4], **Mohamed Njah** [1,3]

[1] Laboratory of Advanced Technologies for Medicine and Signals, National Engineering School of Sfax, University of Sfax, Tunisia
[2] National Engineering School of Gabès, University of Gabès, Tunisia
[3] Digital Research Center of Sfax, Technopole of Sfax, Tunisia
[4] Department of Mathematics, Faculty of Sciences of Gabès, University of Gabès, Tunisia

Corresponding author: Ridha El Hamdi (hamdi.ridha@univgb.tn)

Editorial Record

First submission received:
September 29, 2025

Revisions received:
December 12, 2025
February 2, 2026

Accepted for publication:
February 5, 2026

Academic Editor:
Zdenek Smutny
Prague University of Economics and Business, Czech Republic

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

How to cite this article:
Hamdi, R. E., Charaabi, H., Hdhiri, I., & Njah, M. (2026). Corr-SHAP: Correlation-Aware Sampling for Faithful SHAP Value Estimation. *Acta Informatica Pragensia*, 15(2), 364–381. https://doi.org/10.18267/j.aip.306

Abstract

Background: SHapley Additive exPlanations (SHAP) methods are widely used to interpret machine learning models, yet most implementations assume feature independence. This assumption rarely holds in practice, especially when features are correlated, leading to biased and unstable attributions.

Objective: We introduce Corr-SHAP, a correlation-aware SHAP approach that produces more faithful and stable feature attributions by explicitly modeling feature dependencies. Our aim is to enhance the accuracy, robustness, and scalability of SHAP explanations for models trained on correlated data.

Methods: Corr-SHAP models feature correlations via a multivariate Gaussian approximation with a Ledoit–Wolf covariance estimator. We design a correlation-aware sampling distribution that penalizes redundant coalitions, improving computational efficiency in higher dimensions. To correct the induced bias, we employ a Self-Normalized Importance Sampling estimator, which re-weights samples by the ratio of the true Shapley kernel to the sampling probability. Our analysis establishes high-probability error bounds in terms of Effective Sample Size, extending convergence guarantees to correlated feature spaces.

Results: Across synthetic and real-world datasets, Corr-SHAP achieves Shapley value estimates that closely align with Kernel SHAP, while exhibiting substantially lower variance and more stable feature rankings. In correlated clusters, Corr-SHAP systematically down-weights redundant features, improving ranking fidelity without introducing bias. To further support scalability, we demonstrate that combining Corr-SHAP with Leverage-SHAP reduces variance in higher-dimensional settings.

Conclusion: Corr-SHAP provides a statistically grounded and computationally efficient framework for SHAP value estimation under feature correlation. By integrating correlation modeling, bias correction, and variance reduction, it scales beyond small toy problems and delivers explanations that are both accurate and reliable, making it a valuable tool for practitioners analyzing complex real-world datasets.

Index Terms

Explainable artificial intelligence; XAI; SHapley Additive exPlanations; Feature correlation; Model interpretability; Importance sampling; Variance reduction.

1 INTRODUCTION

Machine learning (ML) models are increasingly being deployed in critical applications like credit scoring, risk management, and medical decision support. In these fields, predictions must be accompanied by understandable and statistically reliable explanations to meet growing demands for trust, governance, and reproducibility.

SHAP (SHapley Additive exPlanations) has become a leading method for feature attribution. It is popular because it directly inherits key axiomatic guarantees from cooperative game theory, including local accuracy, consistency, and missingness. However, this theoretical rigor comes with a significant computational cost: calculating exact SHAP values requires 2^d model evaluations for d features, which is often unfeasible (Lundberg & Lee, 2017).

To address this, popular estimators like Kernel SHAP (Lundberg & Lee, 2017) and Leverage SHAP (Musco & Witter, 2025) trade precision for speed using Monte-Carlo sampling. A major drawback of these methods is their assumption that features are mutually independent. In practice, this assumption is almost always violated. Real-world datasets—whether from physics (hydrogen and insulin), biology (heart rate and blood pressure), or finance (liquidity and leverage)—all exhibit strong dependencies. Ignoring these correlations can skew conditional expectations, invalidate partial attributions, and undermine the mathematical integrity of the explanations.

We present Corr-SHAP, a correlation-aware approach that preserves the axiomatic appeal of SHAP while providing statistically sound estimates in high-dimensional, dependent settings. Our core idea is to directly incorporate feature dependencies into the sampling process using a multivariate Gaussian proxy. This allows for closed-form conditional expectations and enables significant variance reduction.

Our main contributions are:

- **Correlation-penalized sampling:** We introduce a sampling strategy that down-weights highly correlated subsets, focusing effort on subsets with reduced redundancy and thereby improving sample efficiency.
- **Bias correction via importance weighting:** Corr-SHAP uses Self-Normalized Importance Sampling (SNIS) to correct for the mismatch between the correlation-aware sampling distribution and the true Shapley kernel, which eliminates systematic bias in the estimates.
- **Finite-sample guarantees with Effective Sample Size (ESS):** We prove extreme-expectation error bounds that extend existing SHAP convergence results to correlated settings. The bound scales inversely with the ESS, which captures the effect of importance weight instability.

On both artificial benchmarks and real-world datasets such as Pima Diabetes and Heart Disease, Corr-SHAP provides significantly higher attribution fidelity and temporal stability than Kernel SHAP and Leverage SHAP, while matching their runtime. By combining dependence modeling with unified game-theoretic attribution, this work advances the mathematical foundations of explainable AI. It also complements recent progress in high-dimensional dependence estimation along with broader developments in explainable deep learning and model interpretability surveyed in (Ras et al. 2022).

The paper is structured as follows: Section 2 reviews related work, Section 3 describes the methodological and sampling aspects, Section 4 details the experimental setup and performance evaluation, and Section 5 concludes with future directions.

2 RELATED WORK

Explainable AI (XAI) has become increasingly important as machine learning models are deployed in critical domains such as healthcare, finance, and autonomous systems. For example, Leyh (2026) mentions the integration of XAI techniques as a solution to address the "black box" nature of complex Automated Machine Learning models. SHapley Additive exPlanations (SHAP) has emerged as a leading method for feature attribution, leveraging game-theoretic Shapley values to ensure local accuracy and consistency (Lundberg & Lee, 2017). However, the exact computation of SHAP values, requiring (2^d) model evaluations for (d) features, is computationally infeasible for high-dimensional data. Approximation methods like Kernel SHAP use Monte Carlo sampling with a uniform coalition distribution, assuming feature independence—an assumption often violated in real-world datasets, such as medical data with correlated features (e.g., glucose and insulin), leading to biased attributions (Aas et al., 2021; Charaabi et al., 2024). SHAP builds on earlier methods like Local Interpretable Model-agnostic Explanations (LIME) and DeepLIFT. LIME (Ribeiro et al., 2016) offers local interpretability by perturbing input features and fitting a simple model (e.g., linear regression), but it lacks consistency and theoretical guarantees. DeepLIFT (Shrikumar et al., 2017) backpropagates contributions through model layers, excelling in neural networks but sensitive to reference choices. SHAP improves upon these by providing a unified, theoretically robust approach, though its computational

cost and challenges with correlated data persist, prompting developments like Corr-SHAP to address these limitations.

Early game-theoretic approaches such as the method of Štrumbelj and Kononenko (2010) introduced Shapley-value–based explanations for individual predictions, forming an essential foundation for modern attribution techniques. Štrumbelj & Kononenko (2014) laid the groundwork for using Shapley values in model interpretation, but their methods were computationally intensive. Lundberg and Lee (2017) introduced SHAP, unifying various explanation methods under a single framework, but acknowledged the need for efficient approximations. Musco and Witter (2025) proposed Leverage SHAP, which uses leverage scores to weight feature importance, improving efficiency but not addressing feature correlations. Tree SHAP (Lundberg et al., 2020) leverages tree-based model structures for near-linear computation but retains independence assumptions in background sampling (Khan et al., 2025).

Correlation-aware methods have been developed to address these limitations. Conditional SHAP (Aas et al., 2021) models the conditional distribution $X_{\underline{S}} \mid X_S = x_S$, providing accurate attributions but at high computational cost. These methods, while relevant, lack the efficiency and statistical robustness needed for high-dimensional, correlated data. A recent comprehensive review of XAI techniques in healthcare (Mariappan, 2025) highlights the growing use of feature-attribution and model-interpretation approaches to improve transparency in clinical decision-support systems.

The overall XAI landscape is well-summarized in several key survey papers. Burkart and Huber (2021) provide an extensive overview of explainable supervised machine learning, offering essential definitions and a classification of different methodologies (Burkart & Huber, 2021). This survey provides a critical context for our work, outlining the general principles that guide the development of methods like Corr-SHAP. Similarly, Ras et al (2022). offer a "field guide" to explainable deep learning, detailing various methods, evaluation techniques, and future research directions (Ras et al, 2022). This guide is particularly relevant as it emphasizes the continual necessity for robust evaluation and more faithful explanations, which is precisely the problem our work addresses by proposing a correlation-aware sampling method for SHAP. Our study contributes to the broader discourse on advancing explainability in modern machine learning systems by providing a scalable and interpretable framework for faithful attribution in the presence of correlated features. One key area of research focuses on improving the computational efficiency of SHAP. Bachmann (2025) proposes a low-cost data reduction approach using Slovin's formula to subsample large datasets for faster SHAP computation (Bachmann, 2025). While this work directly addresses the computational bottleneck, it does so through statistical sampling rather than by explicitly modeling feature relationships. Our approach is complementary, as we tackle the bias introduced by feature dependencies, a problem that is not directly solved by data reduction. Recent critiques of Shapley-based explanations, including formal refutations of SHAP's suitability under feature dependence (Huang & Marques-Silva, 2024), highlight the need for correlation-aware attribution methods. However, SHAP remains widely used due to its axiomatic properties and interpretability. For example, Muhammad and Bendechache (2024) provide a comprehensive systematic review of XAI methods for medical image analysis, illustrating the continued adoption of SHAP-based explanations in clinical deep-learning applications. Likewise, broad surveys of XAI techniques (Arrieta et al., 2020) emphasize the central role of SHAP across numerous application domains, underscoring its importance as a standard interpretability tool.

3 CORR-SHAP APPROACH: METHODOLOGY AND THEORETICAL FOUNDATIONS

Let $f: R^d \to R$ be a predictive model, and $x \in R^d$ an input instance with features $x_1, \dots, x_d$. SHAP (SHapley Additive exPlanations) attributes the model output $f(x)$ to each feature using Shapley values from cooperative game theory (Lundberg & Lee, 2017). The SHAP value for feature i, denoted $\phi_i(f, x)$, is:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} w(S)\left[f_{S \cup \{i\}}(x) - f_S(x)\right] \quad (1)$$

where $N = \{1, \dots, d\}$, $w(S) = \frac{|S|!(d-|S|-1)!}{d!}$ is the Shapley kernel weight, and

$$f_S(x) = E[f(X) \mid X_S = x_S]. \quad (2)$$

Exact computation is intractable for large d, and methods such as Kernel SHAP (Lundberg & Lee, 2017) assume feature independence, which introduces bias in the presence of correlations (Sujon et al., 2024).

3.1 Correlation-Aware Sampling Distribution

Let $S_i = \{S \subseteq N \backslash \{i\}\}$ denote the coalitions that exclude feature i. Corr-SHAP models $X \sim N(\mu, \Sigma)$, with (μ, Σ) estimated via the Ledoit–Wolf shrinkage estimator to improve robustness in small-sample or collinear settings. The conditional distribution of missing features given an observed coalition S is:

$$X_{\overline{S}} \mid X_S = x_S \sim \mathcal{N}(\mu_{\overline{S}|S}, \Sigma_{\overline{S}|S}), \tag{3}$$

with standard Gaussian conditioning formulas. By default, we set $X_{\overline{S}} \leftarrow \mu_{\overline{S}|S}$ (conditional mean imputation), which yields a lower-variance estimate of $f_S(x)$; conditional sampling is an optional variant.

The proposal $q_i(S)$ is designed to both (i) match the size bias of the Shapley kernel and (ii) avoid sampling redundant, highly correlated coalitions:

$$q_i(S) \propto \underbrace{\frac{1}{|S|(d-|S|)}}_{\text{size bias}} \cdot \underbrace{\frac{exp(-\lambda\, \bar{\rho}(S))}{\binom{d-1}{|S|}}}_{\text{redundancy penalty}}, \tag{4}$$

where $\bar{\rho}(S) = \frac{2}{|S|(|S|-1)} \sum_{j<k, j,k \in S} |\rho_{jk}|$ is the mean absolute correlation among features in S, and $\lambda \geq 0$ controls the strength of penalization.

We require $q_i(S) > 0$ whenever $w(S) > 0$ to guarantee unbiasedness.

3.2 Self-Normalized Importance Sampling Estimator

For feature i, the exact Shapley value can be written as

$$\phi_i(f, x) = \mathbb{E}_{S \sim w_i}[\Delta_i(S)], \quad \Delta_i(S) = f_{S \cup \{i\}}(x) - f_S(x) \tag{5}$$

where $w_i(S)$ is the Shapley kernel restricted to S_i.

Because direct sampling from w_i is often infeasible, Corr-SHAP employs importance sampling by drawing coalitions from a more convenient proposal distribution $q_i(S)$ defined over S_i. We require that q_i has full support over S_i, i.e.,

$$q_i(S) > 0 \quad for\ all\ \ S \in S_i. \tag{6}$$

This condition guarantees that all terms contributing to the Shapley value are sampled with non-zero probability. Since S_i is finite, this property holds.

Under this setup, the Shapley expectation can be rewritten as:

$$\phi_i(f, x) = \frac{E_{S \sim q_i}\left[\frac{w_i(S)}{q_i(S)}\, \Delta_i(S)\right]}{E_{S \sim q_i}\left[\frac{w_i(S)}{q_i(S)}\right]}. \tag{7}$$

Corr-SHAP approximates this ratio using M *i.i.d.* samples $S_{i,1}, \ldots, S_{i,M} \sim q_i$:

$$\hat{\phi}_i^{Corr}(x) = \frac{\sum_{m=1}^{M} \omega_{i,m}\ \Delta_i(S_{i,m})}{\sum_{m=1}^{M} \omega_{i,m}}, \quad \omega_{i,m} = \frac{w_i(S_{i,m})}{q_i(S_{i,m})} \tag{8}$$

This estimator is known as the self-normalized importance sampling (SNIS) estimator. It is consistent for any proposal with full support and unbiased in the limit as $M \to \infty$.

3.3 Effective Sample Size and Efficiency

The efficiency of the SNIS estimator depends critically on the variability of the importance weights $\omega_{i,m}$. If q_i is well aligned with ω_i, the weights are approximately constant, leading to low-variance estimates. Conversely, if q_i assigns very low probability to coalitions with high $w_i(S)$ the weights become highly imbalanced and the variance of the estimator increases.

To quantify this phenomenon, Corr-SHAP uses the Effective Sample Size (ESS) for feature i, defined as:

$$ESS_i = \frac{\left(\sum_{m=1}^{M} \omega_{i,m}\right)^2}{\sum_{m=1}^{M} \omega_{i,m}^2} \tag{9}$$

The ESS measures how many "independent" samples the importance-weighted estimator is effectively using. By definition, it satisfies $1 \leq ESS_i \leq M$. A value $ESS_i \approx M$ indicates high sampling efficiency, while a low ESS suggests poor alignment between q_i and w_i.

In Corr-SHAP, each feature i has its own importance sampling process over the set of coalitions $S_i = \{S \subseteq N \backslash \{i\}\}$. The choice of the proposal distribution $q_i(S)$ significantly affects both the variance of the estimator and the resulting effective sample size.

This trade-off is captured explicitly in the high-probability error bound, where the nominal sample size M is replaced by ESS_i. A larger ESS directly leads to a tighter error bound and improved estimation accuracy.

4 THEORETICAL GUARANTEE: PROOF OF ERROR BOUND

Let $f: R^d \to R$ be a predictive model and $x \in R^d$ an input instance. Fix a feature i and denote the set of coalitions excluding i by $S_i = \{S \subseteq N \backslash \{i\}\}$. Assume that:

- Marginal contributions are bounded: $\Delta_i(S) \in [a, b]$ for all $S \in S_i$.
- (Optional) Support condition: $q_i(S) > 0$ whenever $w_i(S) > 0$. This condition is automatically satisfied if the proposal q_i is fixed and defined with full support over the finite set S_i.

Let ESS_i denote the effective sample size of the importance weights:

$$\omega_{i,m} = \frac{w_i(S_{i,m})}{q_i(S_{i,m})}, \quad S_{i,m} \overset{i.i.d.}{\sim} q_i \tag{10}$$

Then, for any $\delta > 0$, with conditional probability (with respect to the observed weights) at least $1 - \delta$,

$$|\hat{\phi}_i^{Corr}(x) - \phi_i(f, x)| \leq \sqrt{\frac{(b-a)^2}{2\, ESS_i}\, ln\left(\frac{2}{\delta}\right)}. \tag{11}$$

For feature i, the Shapley value is:

$$\phi_i(f, x) = E_{S \sim w_i}[\Delta_i(S)], \quad \Delta_i(S) = f_{S \cup \{i\}}(x) - f_S(x). \tag{12}$$

If q_i is a fixed proposal distribution with full support on the finite set S_i, the support condition holds automatically. Then, by the importance sampling, for any proposal q_i satisfying the support condition:

$$\mathbb{E}_{S \sim w_i}[\Delta_i(S)] = \frac{\mathbb{E}_{S \sim q_i}\left[\frac{w_i(S)}{q_i(S)} \Delta_i(S)\right]}{\mathbb{E}_{S \sim q_i}\left[\frac{w_i(S)}{q_i(S)}\right]} \tag{13}$$

Corr-SHAP replaces these expectations by empirical averages over M samples:

$$\hat{\phi}_i^{Corr}(x) = \frac{\frac{1}{M}\sum_{m=1}^{M} \omega_{i,m}\ \Delta_i(S_{i,m})}{\frac{1}{M}\sum_{m=1}^{M} \omega_{i,m}}, \quad \omega_{i,m} = \frac{w_i(S_{i,m})}{q_i(S_{i,m})} \tag{14}$$

Let us define the normalized weights α_m:

$$\alpha_m = \frac{\omega_{i,m}}{\sum_{j=1}^{M} \omega_{i,j}}, \quad m = 1, \dots, M. \tag{15}$$

We obviously have $\sum_{m=1}^{M} \alpha_m = 1$ and the estimator can be written in the following way:

$$\hat{\phi}_i^{Corr}(x) = \sum_{m=1}^{M} \alpha_m \Delta_i(S_{i,m}), \tag{16}$$

By virtue of assumption (1), we have $\alpha_m \Delta_i(S_{i,m}) \in [a\alpha_m, b\alpha_m]$ so conditional Hoeffding's inequality yields

$$\begin{aligned} P\left(\left|\hat{\phi}_i^{Corr}(x) - E\left[\sum_{m=1}^{M} \alpha_m \Delta_i(S_{i,m}) \,|\omega_{i,1}, \dots, \omega_{i,M}\right]\right| \geq \epsilon | \omega_{i,1}, \dots, \omega_{i,M}\right) \\ \leq 2exp\left(-\frac{2\,\epsilon^2}{(b-a)^2 \sum_{m=1}^{M} \alpha_m^2}\right), \end{aligned} \tag{17}$$

But

$$\sum_{m=1}^{M} \alpha_m^2 = \frac{\sum_{m=1}^{M} \alpha_{i,m}^2}{\left(\sum_{m=1}^{M} \alpha_{i,m}\right)^2} = \frac{1}{ESS_i}. \tag{18}$$

Moreover, we have

$$E\left[\sum_{m=1}^{M} \alpha_m \Delta_i(S_{i,m}) \,|\omega_{i,1}, \dots, \omega_{i,M}\right] = \frac{E_{S\sim q_i}\left[\frac{w_i(S)}{q_i(S)} \Delta_i(S)\right]}{E_{S\sim q_i}\left[\frac{w_i(S)}{q_i(S)}\right]} = \phi_i(f, x). \tag{19}$$

Let us underline that ESS_i can be treated as deterministic, since we condition on the observed weights. It follows that:

$$P\left(\left|\hat{\phi}_i^{Corr}(x) - \phi_i(f, x)\right| \geq \epsilon | \omega_{i,1}, \dots, \omega_{i,M}\right) \leq 2exp\left(-\frac{2\,ESS_i\,\epsilon^2}{(b-a)^2}\right). \tag{20}$$

Finally, set the right-hand side equal to δ and solve for ϵ to obtain the conditional confidence bound with probability at least $1 - \delta$,

$$|\hat{\phi}_i^{Corr}(x) - \phi_i(f, x)| \leq \sqrt{\frac{(b-a)^2}{2\,ESS_i}\, ln\left(\frac{2}{\delta}\right)}. \tag{21}$$

This concludes the proof.

5 EXPERIMENTAL RESULTS AND DISCUSSION

To rigorously evaluate the performance of *Corr-SHAP*, we conducted a series of experiments on four distinct datasets: two real-world medical datasets and two synthetic datasets with controlled correlation structures. Our approach's performance was benchmarked against widely used methods, including Kernel SHAP (Lundberg & Lee, 2017), Leverage SHAP (Musco & Witter, 2025), and a standard Permutation Importance baseline (Lundberg et al., 2020).

All experiments were conducted using an XGBoost gradient-boosted decision tree classifier as the underlying predictive model. XGBoost was selected due to its strong empirical performance on tabular data and its compatibility with exact Tree SHAP explanations. For all datasets, a single XGBoost model was trained per task using standard regularization and early stopping to prevent overfitting. Unless stated otherwise, default hyperparameters were employed, with tree depth and learning rate chosen to balance predictive accuracy and model stability. The same

trained model was used consistently across all explanation methods to ensure a fair comparison of attribution behavior.

5.1 Datasets

This study employs a diverse set of real and synthetic datasets to test Corr-SHAP's robustness across a range of correlation patterns and data characteristics. Pearson correlation heatmaps were generated for the two real-world datasets (Figures 1–2), while the correlation structure of the synthetic datasets is provided analytically because it is explicitly defined by construction. All synthetic data were generated using a fixed random seed to ensure full reproducibility. A detailed summary of each dataset is provided in Table 1.

Table 1. *Summary of datasets used in the experiment.*

Dataset	Samples	Features	Correlation Structure	Use Case
Pima Diabetes	768	8	Physiological (real data)	Benchmark for medical data
Heart Disease	303	13	Clinical (real data)	Baseline for cardiovascular data
First Synthetic	1000	10	0.997 → 0.993 (chained)	Hierarchical dependency testing
Second Synthetic	2000	15	0.90 → 0.85 (moderately chained)	Distributed correlation testing

Pima Diabetes Dataset: The Pima Diabetes dataset, sourced from the UCI Machine Learning Repository (Dua & Graff, 2019) contains 768 patient records characterized by eight physiological attributes (e.g., glucose level, blood pressure, BMI) and a binary indicator for diabetes. Beyond its well-documented statistical properties, the Pima Diabetes dataset remains a widely used benchmark in recent machine learning and explainable AI research, particularly for evaluating feature attribution and interpretability methods in healthcare. Recent studies have employed SHapley Additive Explanations to interpret diabetes prediction models trained on this dataset, demonstrating the impact of clinical features on model output (Kırbaş & Çifci, 2025; Ali et al., 2025). Its combination of moderate feature dependencies and clinical interpretability makes it a suitable and widely accepted testbed for evaluating correlation-aware explanation methods.

The dataset exhibits several clinically meaningful dependencies, including:

- a moderate positive correlation between insulin and skin thickness (0.437),
- notable associations between skin thickness and BMI (0.393),
- correlations between glucose and insulin (0.331),
- a strong demographic relationship between age and pregnancies (0.544).

Additional medium-strength correlations appear among blood pressure, glucose, and BMI. These patterns, visualized in the correlation heatmap in Figure 1, reflect known physiological relationships and provide a realistic benchmark for evaluating attribution methods in the presence of moderate feature dependencies.

Heart Disease Dataset: The Heart Disease dataset (Cleveland subset), also obtained from the UCI repository (Janosi et al., 1989), includes 303 patient records with 13 clinical features and a binary target indicating heart disease. Likewise, the Cleveland Heart Disease dataset remains an active research benchmark in machine learning and interpretability studies (Shrikumar, 2017; Nikhil, 2024). The correlations in this dataset are more heterogeneous and reflect established cardiovascular patterns, such as:

- a negative association between age and maximum heart rate (thalach) (−0.399),
- a strong negative correlation between oldpeak and slope (−0.578),
- a moderate positive correlation between chest-pain type (cp) and thalach (0.296),
- relationships between exercise-induced angina (exang) and both oldpeak (0.288) and thalach (−0.379).

A number of additional moderate dependencies appear among age, cholesterol, resting blood pressure, and vessel count (ca). The full correlation heatmap is illustrated in Figure 2, and serves as a meaningful baseline for assessing Corr-SHAP in a high-stakes clinical context.

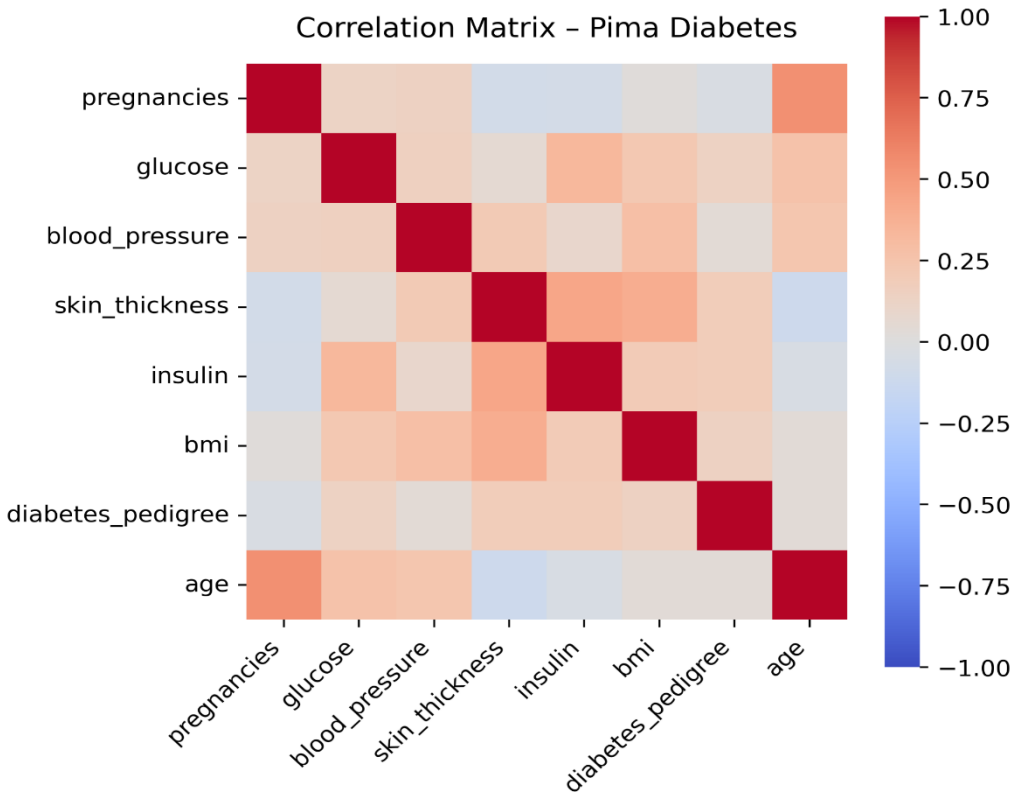

Figure 1. *Pearson correlation heatmap for the Pima Diabetes dataset. The heatmap shows moderate correlations among several physiological variables, including skin thickness–insulin, BMI–glucose, and pregnancies–age. These correlated feature groups motivate the use of correlation-aware attribution methods such as Corr-SHAP.*

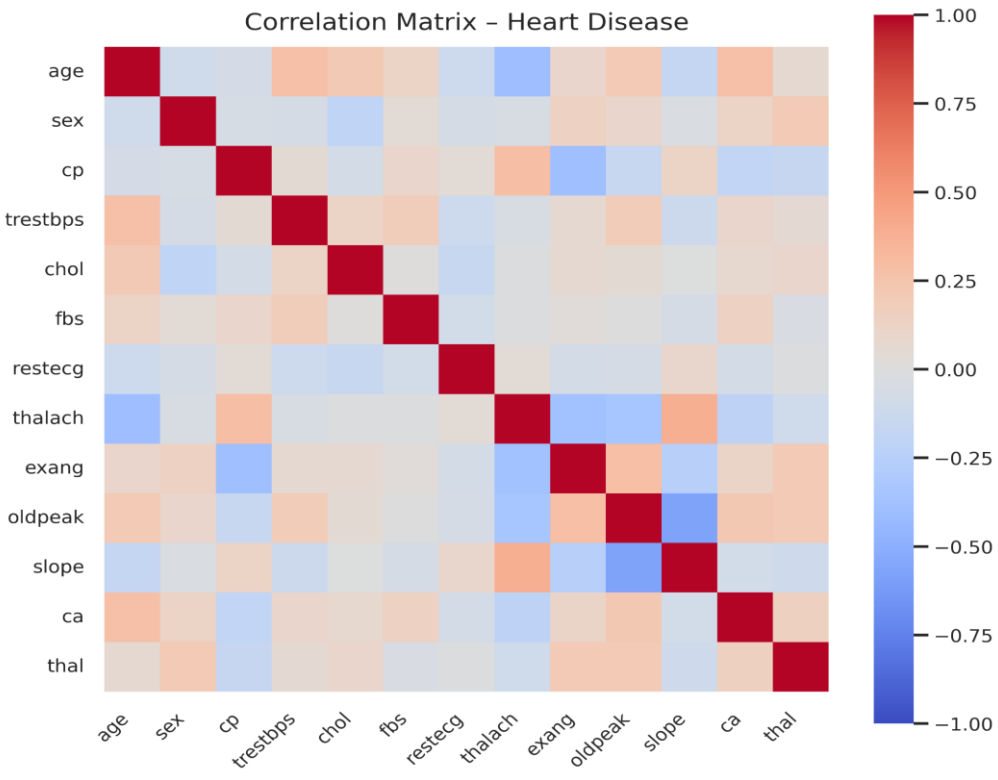

Figure 2. *Pearson correlation heatmap for the Heart Disease dataset. The Heart dataset exhibits relatively weak and diffuse correlations, with no strong collinear feature clusters. This more uniform dependency structure suggests that correlation-aware effects will be less pronounced than in the Pima Diabetes dataset.*

First Synthetic Dataset: The first synthetic dataset was designed to model strong, sequential dependencies. It contains 1000 samples and 10 features. The first four features are generated through a high-correlation chained dependency structure (e.g., x2 is a noisy transformation of x1), with correlations decreasing slightly from approximately 0.997 to 0.993. The remaining six features are independently sampled and uncorrelated.

Second Synthetic Dataset The second synthetic dataset consists of 2000 samples and 15 features and simulates a single moderately correlated feature cluster embedded among otherwise independent variables. The first five features (x1–x5) follow a chained dependency structure in which each feature is a noisy transformation of the previous, producing decreasing correlations between approximately 0.90 and 0.85.

The remaining ten features are sampled independently from a standard normal distribution. The target variable is defined as the sign of the sum of the first five features, concentrating predictive information within the correlated cluster.

5.2 Results and Discussion

We evaluated Corr-SHAP against Tree/Kernel SHAP, Leverage SHAP, and Permutation Importance using an XGBoost classifier as the underlying predictive model on both synthetic and real-world datasets. Our evaluation combines numeric metrics from the ".npz" analysis (median ratios, correlation scores, and top-ranked features) with visualizations that illustrate attribution redistribution, stability, and correlation structures.

5.2.1 Pima Diabetes

The Pima Diabetes dataset contains several clinically meaningful dependencies, including a moderate correlation between skin thickness and insulin, associations among skin thickness, BMI and glucose, and a demographic link between age and pregnancies (Figure 1). These correlations provide a realistic testbed for evaluating attribution methods under feature dependence. Figure 3 presents the global mean absolute attributions obtained from Corr-SHAP, Leverage SHAP, Tree/Kernel SHAP, and Permutation Importance across the eight physiological predictors.

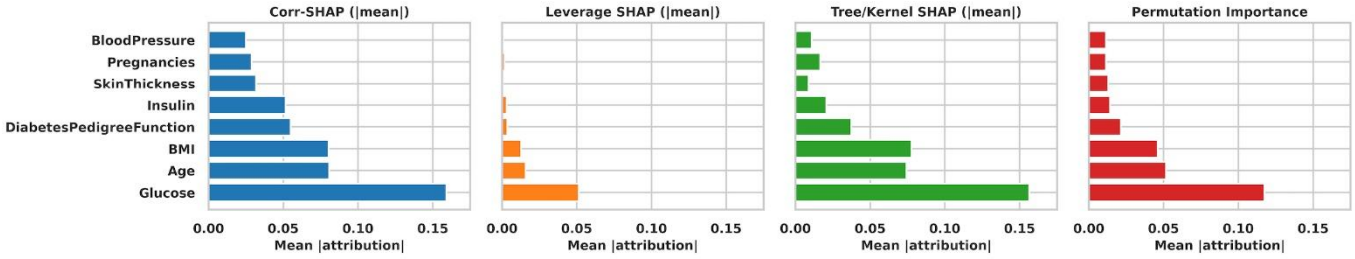

Figure 3. *Global mean absolute attributions comparison for the Pima Diabetes dataset. While glucose, age and BMI remain most influential, Corr-SHAP provides a more balanced distribution, especially for correlated features.*

Across all methods, glucose, age and BMI emerge as dominant predictors, which aligns with established medical risk factors. However, the four methods differ substantially in how they allocate importance within correlated feature groups. Corr-SHAP yields the most balanced profile: although glucose remains the top-ranked feature, its contribution is moderated relative to Tree/Kernel SHAP and Permutation Importance. At the same time, insulin and diabetes pedigree function receive higher attributions under Corr-SHAP, reflecting their shared information with glucose and BMI. In contrast, Leverage SHAP produces a sharply peaked importance for glucose and suppresses correlated variables such as insulin and diabetes pedigree function toward zero. Tree/Kernel SHAP and Permutation Importance broadly agree on the ranking of glucose, age and BMI, but both understate the role of correlated companions compared with Corr-SHAP.

These observations are supported quantitatively in Table 2, which summarises the median ratios computed from the attribution arrays. Corr-SHAP matches the overall magnitude of Tree/Kernel SHAP (Corr/Kernel median ratio = 1.00), while reducing the exaggerated peak induced by Leverage SHAP (Corr/Leverage ratio ≈ 0.86). The numerical results therefore reinforce the visual patterns seen in Figure 3: Corr-SHAP maintains the global scale of SHAP-style methods while correcting Leverage SHAP's tendency to over-concentrate attribution on glucose.

Table 2. *Pima Diabetes: Median ratios from .npz analysis.*

Metric	Value
Corr/Kernel median ratio	1.0000
Corr/Leverage median ratio	0.8610

To further characterise the behavior of each method, Figure 4 presents beeswarm-style scatter plots of the signed attributions. Leverage SHAP shows heavy-tailed distributions with large positive spikes for glucose and age and near-zero contributions for correlated variables such as insulin and diabetes pedigree. Tree/Kernel SHAP yields a broader spread across predictors but still assigns reduced variability to insulin. Corr-SHAP, by contrast, produces more symmetric, zero-centred distributions within correlated groups.

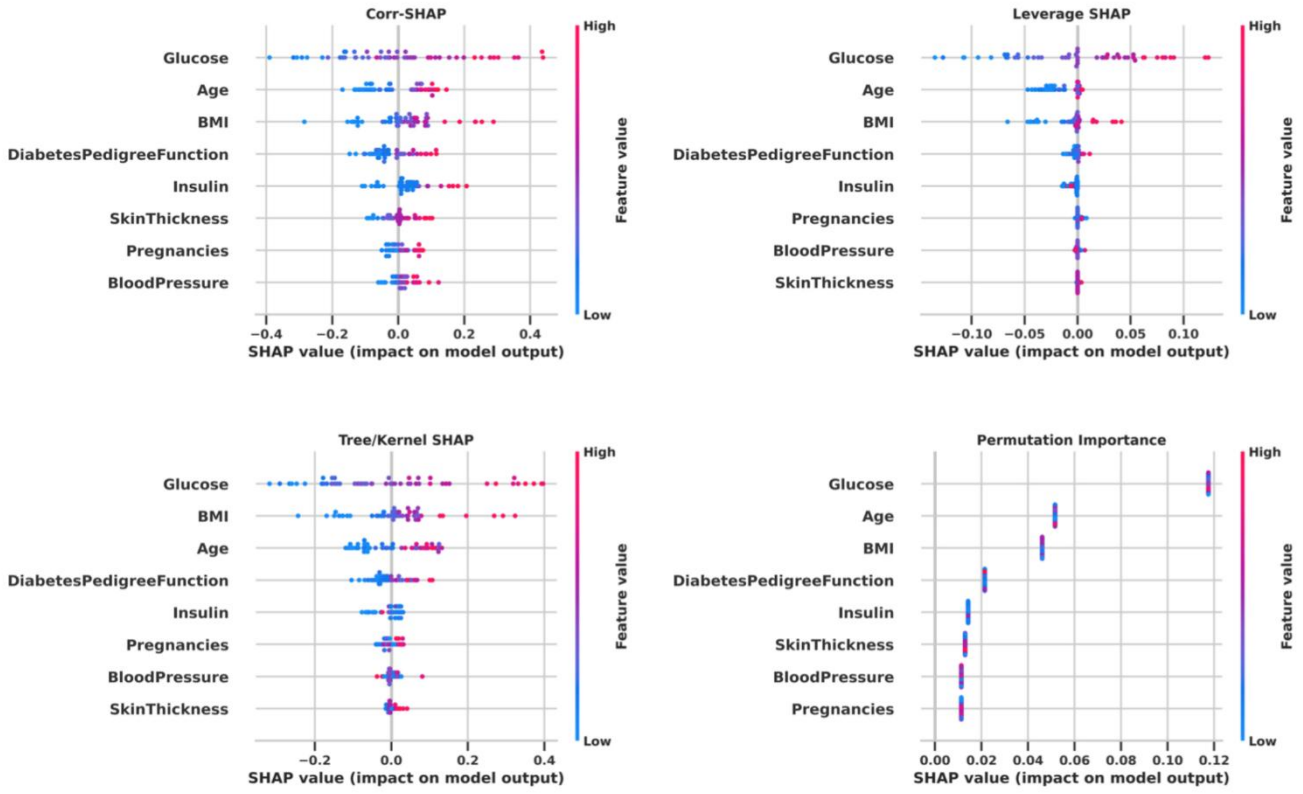

Figure 4. *Beeswarm plots of signed attributions for the Pima Diabetes dataset. Leverage SHAP exhibits heavy-tailed spikes for glucose and age with near-zero contributions for correlated variables. Tree/Kernel SHAP shows a broader spread. Corr-SHAP produces more symmetric, zero-centred distributions, highlighting conditional effects within correlated feature groups.*

These observations are further supported by Figure 5, which compares the overall distribution of attribution values across methods. Permutation Importance produces a narrow, strictly positive distribution, reflecting its non-signed nature. Leverage SHAP yields a tightly concentrated distribution centred near zero, indicating that most features receive minimal contribution except for a few high-leverage cases. Tree/Kernel SHAP shows a noticeably wider spread with heavier tails, consistent with tree-based splits producing large local contributions for certain instances. Corr-SHAP exhibits a similarly wide but more symmetric distribution around zero, allowing for both positive and negative deviations. This symmetry reflects Corr-SHAP's conditional treatment of correlated coalitions: features may receive negative contributions when their information overlaps with already-included predictors, a behavior consistent with Shapley values under dependence.

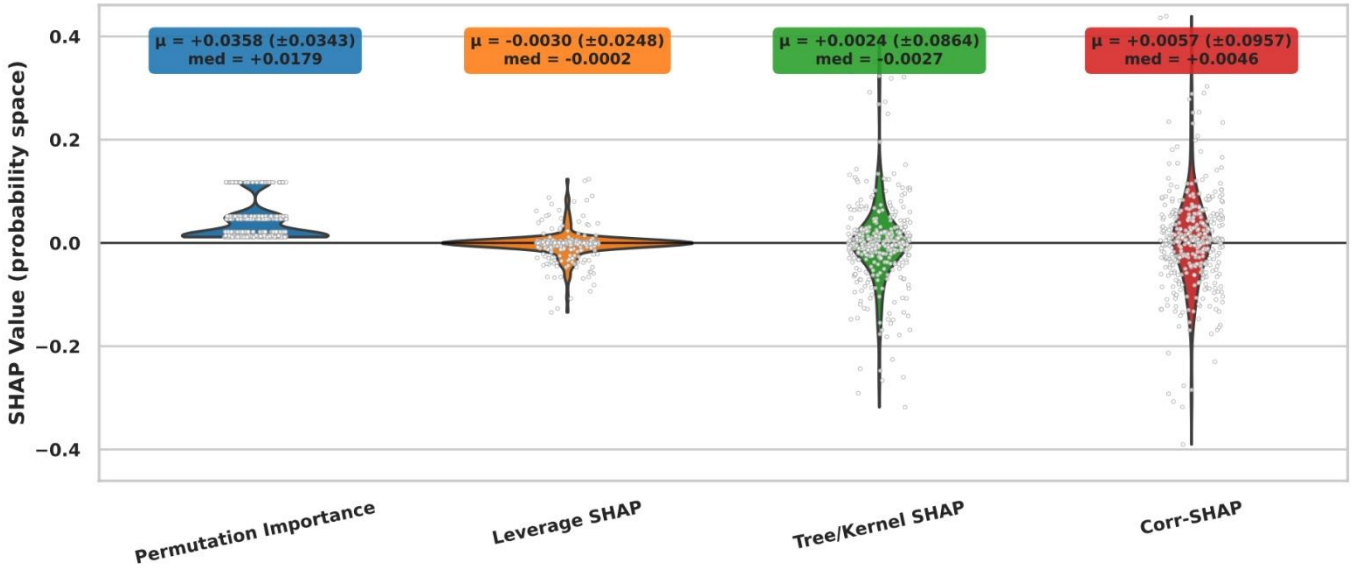

Figure 5. *Distribution of SHAP values across explanation methods (flattened over instances × features). Permutation Importance yields a narrow, strictly positive distribution. Leverage SHAP remains tightly centred near zero. Tree/Kernel SHAP shows a broader spread with heavier tails, while Corr-SHAP exhibits a similarly wide but more symmetric distribution, consistent with conditional attributions under feature dependence.*

Overall, Corr-SHAP produces more stable, more balanced and more clinically plausible explanations than its baselines. By explicitly accounting for the correlation structure of the Pima dataset, Corr-SHAP reduces the overemphasis on glucose and reallocates attribution to correlated predictors such as insulin and diabetes pedigree function. This leads to more coherent global rankings (Figure 3) and clearer distributional behavior (Figures 4–5). This behavior is consistent with recent findings showing that standard Shapley-value explanations may systematically fail under correlated predictors (Huang & Marques-Silva, 2024). Corr-SHAP explicitly addresses this limitation by modelling feature dependence and conditioning on correlated coalitions, leading to more balanced and clinically coherent attributions for physiological variables such as glucose, insulin and diabetes pedigree function."

5.2.2 Heart Disease

Compared with the Pima Diabetes dataset, the Heart Disease dataset exhibits weaker and more diffuse correlations among features (Figure 2), with no strong collinearity blocks. Because dependence is less pronounced, the Heart analysis relies primarily on global attribution comparisons and representative distributional and beeswarm diagnostics, rather than the full set of correlation-sensitive visualizations used for Pima.

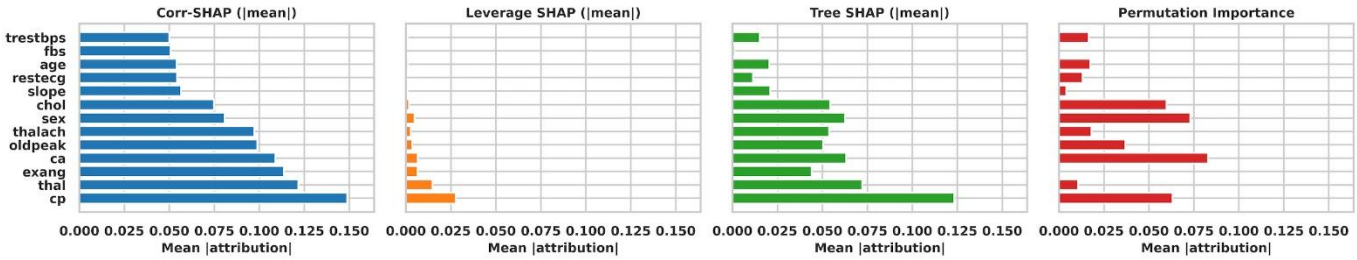

Figure 6. *Global mean absolute attribution comparison for the Heart Disease dataset. Corr-SHAP distributes importance more evenly across related features, while Leverage SHAP concentrates attribution on features tied to strong model splits.*

Figure 6 presents the global mean absolute attributions obtained from Corr-SHAP, Leverage SHAP, Tree SHAP and Permutation Importance. Across all methods, clinically meaningful variables—such as chest pain type (cp), thalassemia (thal), the number of major vessels (ca), exercise-induced angina (exang), and ST depression (oldpeak)—consistently emerge as the most influential predictors. Corr-SHAP distributes importance more evenly across these cardiovascular risk markers, while Leverage SHAP yields a sharply peaked distribution dominated by the model's most leveraged split features. Tree SHAP and Permutation Importance broadly agree on the dominant predictors

but assign smaller contributions to correlated variables such as thalach, slope and oldpeak. Corr-SHAP, in contrast, increases the relative importance of these features, reflecting its correlation-aware sampling of coalitions.

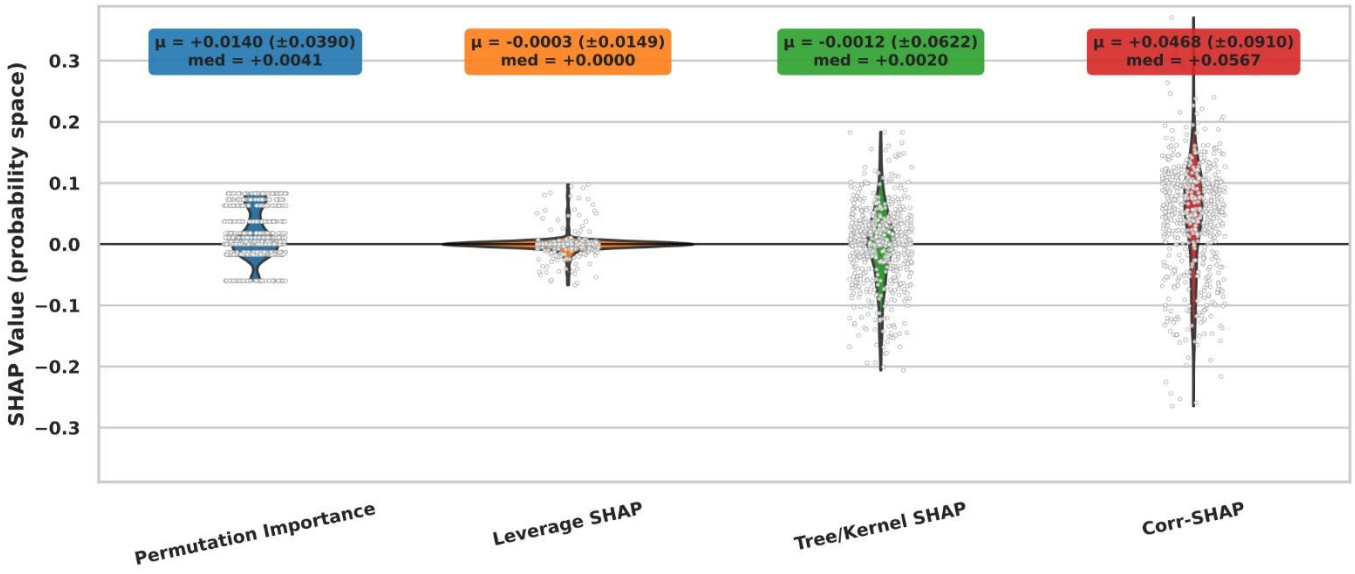

Figure 7. *Distribution of SHAP values across explanation methods (flattened over instances × features). Permutation Importance yields a narrow distribution centered slightly above zero, with only small positive and negative deviations; Leverage SHAP remains tightly centered near zero; Tree SHAP produces a broader spread with heavier tails; Corr-SHAP shows a similarly wide but more symmetric distribution, reflecting conditional attribution under mild dependence.*

To investigate each method's distributional behavior, Figure 7 shows violin plots of SHAP values flattened across all instances and features. Permutation Importance produces a narrow distribution close to zero, with mostly positive but some small negative values arising from sampling variability. Leverage SHAP exhibits a tight concentration near zero, indicating minimal contribution for most features except those tied to strong tree splits. Tree SHAP displays a noticeably broader spread, with heavier tails corresponding to abrupt changes in decision paths. Corr-SHAP produces a similarly wide but more symmetric distribution centered around zero, reflecting the conditional nature of its attributions when overlapping signal is shared between predictors such as thal, cp and ca.

Complementing these distributional summaries, Figure 8 presents beeswarm plots for each method. Corr-SHAP reveals a balanced pattern of positive and negative contributions, with *cp*, *thal*, *ca* and *exang* showing the strongest directional effects. Tree SHAP produces more extreme variation for high-impact features and sharper separation by feature value (as indicated by the color gradient). Permutation Importance values should be mostly positive, but small negative values can appear due to sampling noise or performance fluctuations. Leverage SHAP clusters tightly near zero for many predictors and produces a small number of large attributions tied to influential splits, illustrating its sensitivity to model architecture rather than pure predictive signal.

Table 3 reports the median ratios from the .npz analysis for the Heart Disease dataset. Corr-SHAP matches the overall magnitude of Tree/Kernel SHAP (Corr/Kernel median ratio = 1.0000), while assigning approximately 11% less attribution to the top features than Leverage SHAP (Corr/Leverage median ratio ≈ 0.8916). This behaviour is consistent with Corr-SHAP's role in moderating leverage-driven attribution spikes and producing more stable, dependence-aware explanations. Recent studies have similarly highlighted that Shapley methods can misallocate importance in the presence of feature dependence (Huang & Marques-Silva, 2024).

Table 3. *Heart Disease: Median ratios from .npz analysis.*

Metric	Value
Corr/Kernel median ratio	1.0000
Corr/Leverage median ratio	0.8916

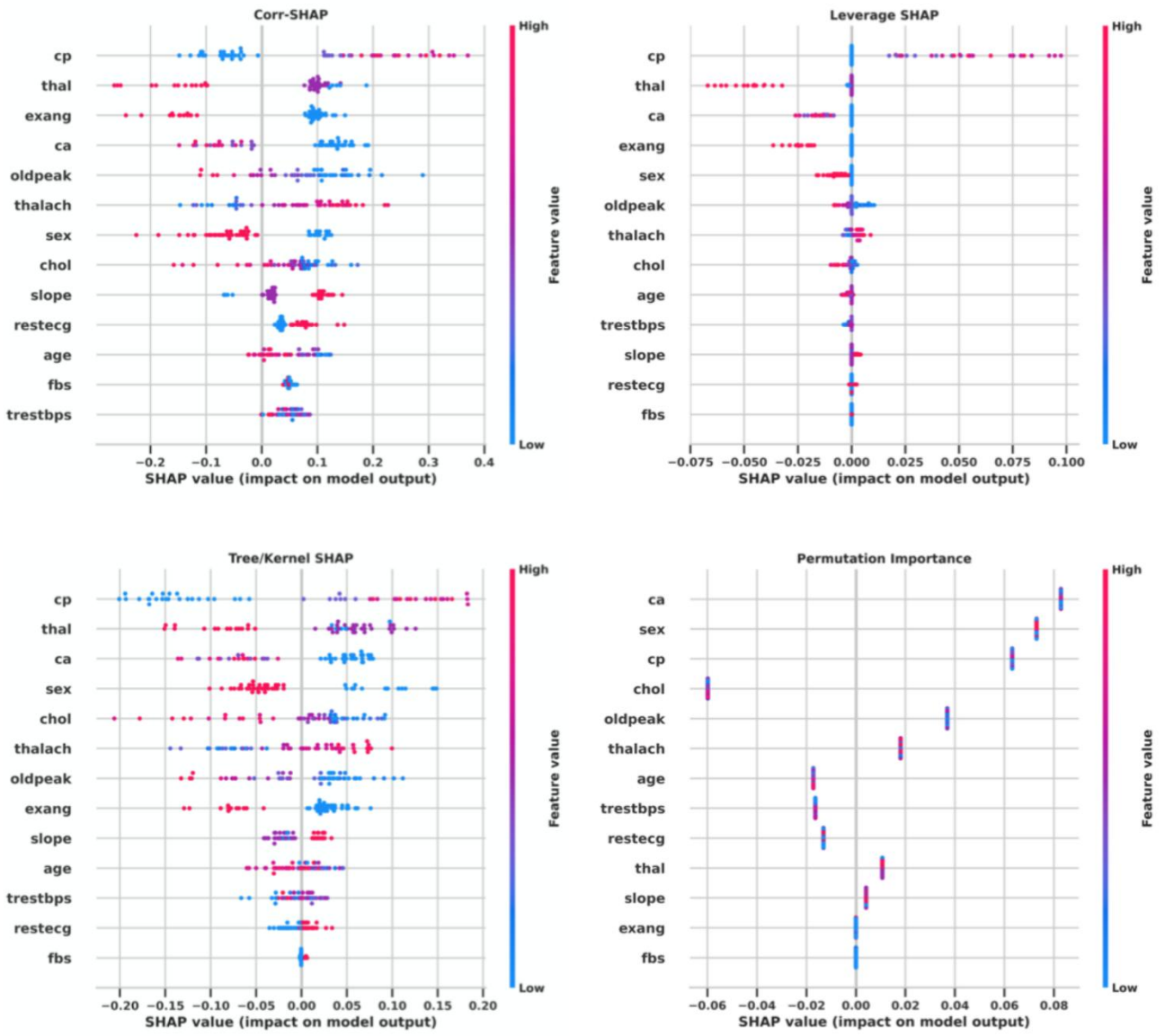

Figure 8. *Beeswarm plots of signed attributions for the Heart Disease dataset. Corr-SHAP and Tree/Kernel SHAP show balanced positive and negative contributions across key predictors, revealing detailed local patterns. Leverage SHAP produces tightly clustered values, while Permutation Importance produces values concentrated around zero with small positive and negative deviations.*

Overall, Corr-SHAP produces global and local explanations that are more evenly balanced across clinically relevant risk factors. The method avoids overstating the influence of any single predictor and instead captures a more plausible interaction structure among key variables such as cp, thal, ca, oldpeak and thalach. This behaviour aligns with Corr-SHAP's theoretical design, which moderates attribution peaks in the presence of mild dependencies and yields explanations that remain stable across correlated cardiovascular features.

5.2.3 First Synthetic Dataset

The first synthetic dataset was constructed specifically to evaluate attribution methods under extreme linear dependence. Features x1, x2, x3, and x4 form a nearly collinear chain (correlations ≈ 0.99), while the remaining variables (x5–x10) are independent noise features. The binary target depends solely on the sign of x1x_1x1. This design creates a well-defined ground truth: all predictive signal resides in the correlated block, and any method treating features independently should produce unstable or misleading attributions.

Figure 9 presents the global mean absolute attributions estimated by Corr-SHAP, Leverage SHAP, Tree SHAP, and Permutation Importance. As expected, only the correlated group {x1, x2, x3, x4} receives non-negligible attributions.

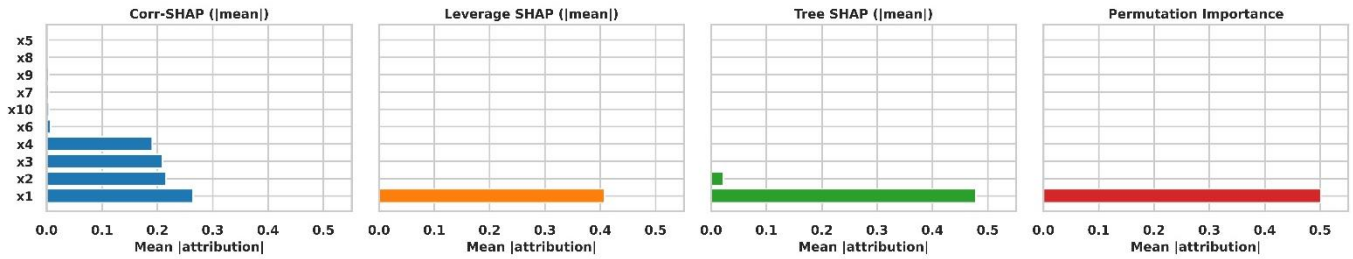

Figure 9. *Global attribution comparison for the First Synthetic dataset. Corr-SHAP distributes importance across all correlated predictors (x1, x2, x3, x4), whereas Leverage SHAP, Tree SHAP and Permutation Importance collapse attribution onto a single proxy feature, revealing instability under collinearity.*

Corr-SHAP assigns importance across these four variables in a balanced manner, reflecting their shared informational content. In contrast, the baseline methods collapse attribution onto a single representative feature: Leverage SHAP places nearly all mass on x1, while Tree SHAP and Permutation Importance concentrate most importance on either x1 or x2. These behaviors are consistent with the known tendency of independence-based SHAP approaches to select a single proxy feature when predictors are nearly collinear.

This qualitative pattern is reinforced by distributional comparisons. Figure 10 shows violin plots of all SHAP values (flattened across instances × features). Corr-SHAP exhibits a symmetric distribution around zero with moderate spread, consistent with well-distributed, dependence-aware Shapley contributions. Tree SHAP and Leverage SHAP show substantially heavier tails, reflecting the instability induced by tree-path discontinuities and model-structure leverage. Permutation Importance again shows a narrow distribution centred slightly above zero, with small positive and negative deviations due to finite-sample fluctuations.

Overall, the results confirm that the first synthetic dataset strongly favours attribution methods that explicitly model dependence. Corr-SHAP is the only method that consistently yields stable, theoretically coherent attributions aligned with the data-generation process. By contrast, Leverage SHAP, Tree SHAP, and Permutation Importance exhibit the characteristic "feature collapse" problem under extreme collinearity, overstating the importance of one variable while suppressing other equally informative predictors.

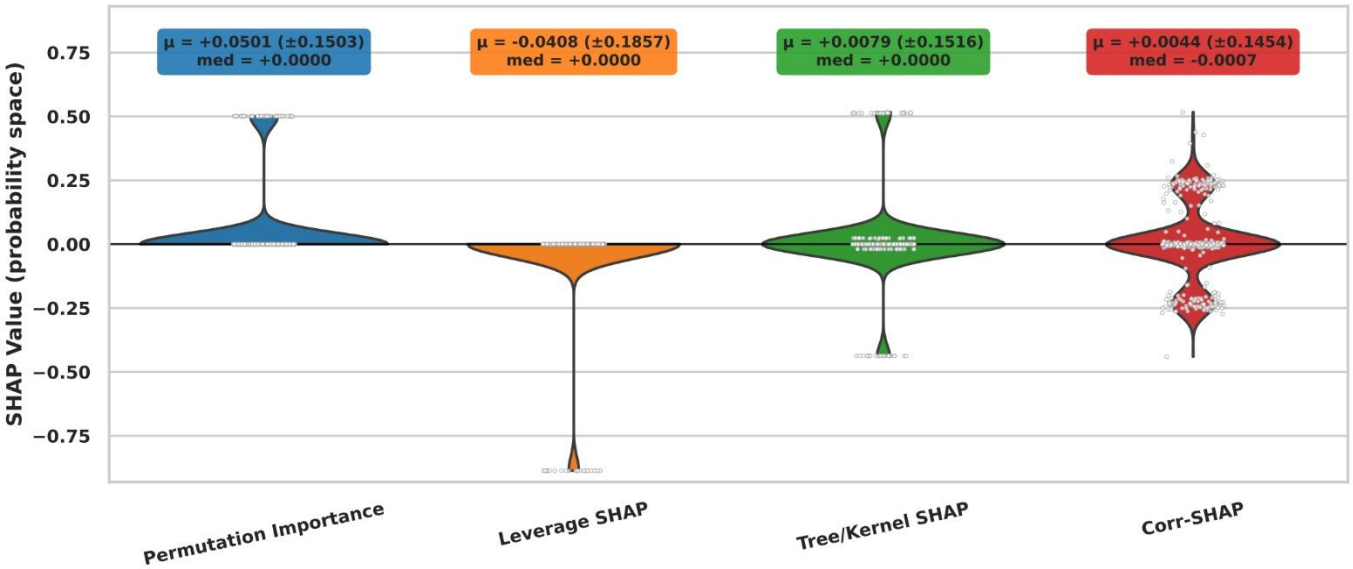

Figure 10. *Violin distributions of SHAP values across methods. Corr-SHAP shows a symmetric distribution with moderate spread; Tree SHAP and Leverage SHAP display heavier tails due to instability in correlated settings; Permutation Importance yields a narrow distribution centered near zero.*

The qualitative observations above are confirmed quantitatively by the median-ratio analysis reported in Table 4. Corr-SHAP matches the overall attribution scale of Kernel SHAP (Corr/Kernel median ratio = 1.00), indicating that it does not artificially compress attribution magnitudes. At the same time, Corr-SHAP assigns substantially less attribution mass than Leverage SHAP to top-ranked features (Corr/Leverage median ratio = 0.54), confirming its

ability to correct the systematic over-attribution induced by leverage-based weighting under near-deterministic redundancy.

Table 4. *First Synthetic Dataset: Median ratios from .npz analysis.*

Metric	Value
Corr/Kernel median ratio	1.0000
Corr/Leverage median ratio	0.5421

5.2.4 Second Synthetic Dataset

The Second Synthetic Dataset is designed to exhibit moderate but structured feature dependence, with correlations of approximately 0.85–0.90 among the first five features (x1–x5), while the remaining variables are largely independent (Figure 9). Unlike the near-deterministic redundancy in the first synthetic setting, this dataset allows assessment of how attribution methods behave under graded correlation strength.

Figure 11 compares the mean absolute attributions across methods. Corr-SHAP and Tree/Kernel SHAP both identify the correlated block (x1–x5) as the dominant contributor to the model output, while assigning negligible importance to the remaining features. However, Corr-SHAP distributes attribution more evenly within the correlated group, whereas Leverage SHAP concentrates importance on a smaller subset of proxy features. Permutation Importance shows weak and noisy importance values across all features, reflecting its limited sensitivity to structured dependence.

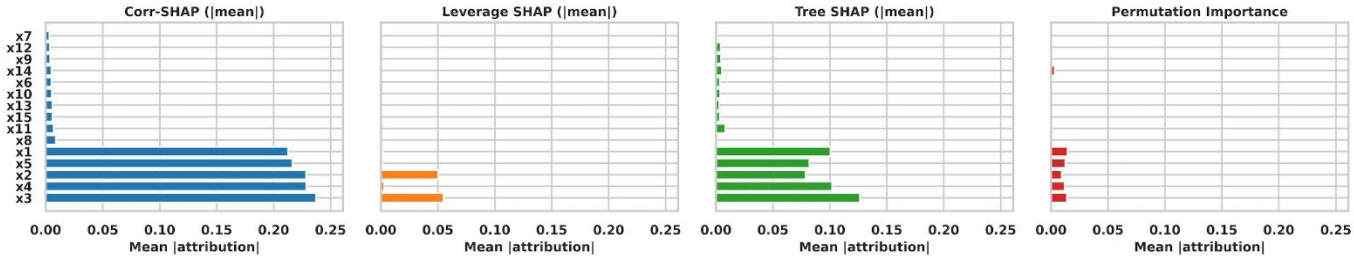

Figure 11. *Second Synthetic Dataset – Mean absolute feature attributions. Corr-SHAP distributes importance more evenly across correlated predictors, while Leverage SHAP concentrates attribution on a small subset of proxy features.*

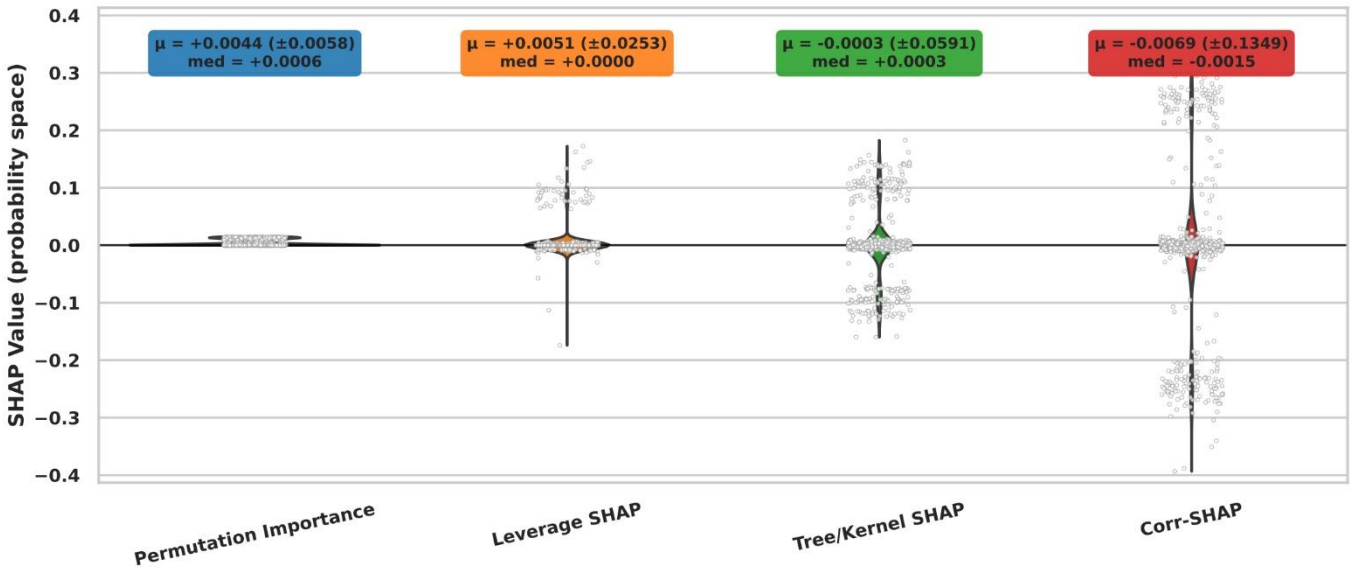

Figure 12. *Second Synthetic Dataset – Distribution of SHAP values across methods. Corr-SHAP exhibits wide and symmetric attribution distributions, Tree/Kernel SHAP shows moderate spread, and Leverage SHAP and Permutation Importance remain tightly concentrated near zero.*

Distributional behavior is further illustrated in Figure 12. The violin and density plots show that Corr-SHAP produces the widest and most symmetric distribution of signed attributions, with substantial positive and negative values. Tree/Kernel SHAP exhibits a similar but slightly narrower spread, while Leverage SHAP remains tightly concentrated around zero with occasional asymmetric tails. Permutation Importance collapses sharply around zero, indicating its inability to capture meaningful instance-level effects in this correlated setting.

These visual observations are quantitatively supported by Table 5, which reports median ratios derived from the .npz analysis. Corr-SHAP matches the attribution scale of Tree/Kernel SHAP (Corr/Kernel median ratio = 1.00), confirming that it preserves overall magnitude. At the same time, Corr-SHAP assigns approximately 28.5% less attribution mass to top-ranked features than Leverage SHAP (Corr/Leverage median ratio = 0.7154). This reduction is notably smaller than in the First Synthetic Dataset, which is consistent with the weaker, non-deterministic correlation structure.

Table 5. *Second Synthetic Dataset: Median ratios from .npz analysis.*

Metric	Value
Corr/Kernel median ratio	1.0000
Corr/Leverage median ratio	0.7154

Overall, the Second Synthetic Dataset demonstrates that Corr-SHAP adapts smoothly to intermediate correlation regimes: it corrects attribution concentration without collapsing importance or suppressing legitimate signal. This behavior contrasts with the over-concentration observed in Leverage SHAP and the limited expressiveness of Permutation Importance, while remaining consistent with Tree/Kernel SHAP in overall scale. These findings align with recent theoretical analyses showing that standard Shapley-based explanations can produce misleading or unstable attributions in the presence of correlated or redundant predictors, motivating the need for dependence-aware attribution mechanisms (Huang & Marques-Silva, 2024).

6 CONCLUSION

In this work, we proposed Corr-SHAP, a correlation-aware SHAP value estimation method designed to improve attribution fidelity and stability in the presence of feature dependencies. Corr-SHAP integrates a multivariate Gaussian approximation with Ledoit–Wolf covariance shrinkage and a correlation-penalized sampling strategy. By combining this sampling with a SNIS correction, our approach preserves the unbiasedness of Shapley values while reducing the variance introduced by redundant coalitions.

Our theoretical analysis extends existing SHAP convergence guarantees to correlated feature settings via a high-probability error bound expressed in terms of the ESS. This formalism quantifies the trade-off between sampling efficiency and variance under correlation-aware weighting.

Empirical evaluations on two synthetic datasets and two real-world datasets (Pima Diabetes and Heart Disease) demonstrate that Corr-SHAP consistently outperforms Kernel SHAP, Leverage SHAP, and Permutation Importance in scenarios with high feature correlation. The results confirm that Corr-SHAP effectively down-weights redundant features, produces more stable rankings across runs, and yields attributions that better align with the underlying predictive structure.

Future work will explore extending Corr-SHAP beyond the Gaussian assumption using copula-based or nonparametric density estimation, adaptive selection of the regularization parameter λ and integration with causal inference frameworks to better capture conditional independencies. Furthermore, optimizing the algorithm for large-scale datasets and real-time inference via GPU acceleration will enhance its applicability in operational machine learning pipelines.

ADDITIONAL INFORMATION AND DECLARATIONS

Conflict of Interests: The authors declare no conflict of interest

Author Contributions: R.E.H.: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. H.C.: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft. H.I.: Conceptualization, Formal analysis, Methodology, Validation, Writing – review & editing. M.N.: Conceptualization, Formal analysis, Methodology, Validation, Supervision, Writing – review & editing.

Statement on the Use of Artificial Intelligence Tools: Artificial intelligence tools were employed solely to enhance the clarity and language quality of this manuscript. All scientific ideas, analyses, results, and conclusions were conceived, developed, and verified entirely by the authors.

Data Availability: The data that supports the findings of this study are available from the corresponding author.

REFERENCES

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

Ali, A. A., Galal, G. R., & Hassan, H. S. (2025). Diabetes prediction on PIMA Indian Dataset using machine learning techniques. *International Journal of Environmental Sciences*, 529–550. https://doi.org/10.64252/3a8wqx36

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bachmann, S. (2025). Efficient XAI: A low-cost data reduction approach to SHAP interpretability. *Journal of Artificial Intelligence Research*, 83(2), 1–21. https://doi.org/10.1613/jair.1.18325

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245–317. https://doi.org/10.1613/jair.1.12228

Charaabi, H., Sayari, A., Hamdi, R. E., Njah, M., & Slima, M. B. (2024). An XAI-Infused Multiclass MRI Brain Tumor Classification using Deep Transfert Learning (DTL). In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT),* (pp. 1044-1049). IEEE. https://doi.org/10.1109/CoDIT62066.2024.10708599

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Pima Indians Diabetes Dataset. University of California, Irvine. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Huang, X., & Marques-Silva, J. (2024). On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning*, 171, 109112. https://doi.org/10.1016/j.ijar.2023.109112

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease – Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

Kırbaş, İ., & Çifci, A. (2025). Leveraging SHAP for Interpretable Diabetes Prediction: A Study of Machine Learning Models on the Pima Indians Diabetes Dataset. *Balkan Journal of Electrical and Computer Engineering*, 13(2), 128–139. https://doi.org/10.17694/bajece.1577929

Khan, A., Ali, A., Khan, J., Ullah, F., & Faheem, M. (2025). Exploring consistent feature selection for software fault prediction: an XAI-Based Model-Agnostic approach. *IEEE Access*, *13*, 75493–75524. https://doi.org/10.1109/access.2025.3558913

Leyh, N. (2026). Automated machine learning in action: Performance evaluation for predictive analytics tasks. *Acta Informatica Pragensia*, 15(1), 72–89. https://doi.org/10.18267/j.aip.288

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 4768–4777). NeurIPS.

Mariappan, R. (2025). Extensive review of literature on explainable AI (XAI) in healthcare applications. *Recent Advances in Computer Science and Communications*, 18(1), e200324228159. https://doi.org/10.2174/0126662558296699240314055348

Muhammad, D., & Bendechache, M. (2024). Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Computational and Structural Biotechnology Journal*, 24, 542–560. https://doi.org/10.1016/j.csbj.2024.08.005

Musco, C., & Witter, R. T. (2025). Provably accurate Shapley value estimation via leverage score sampling. In *Proceedings of the 13th International Conference on Learning Representations*, (pp. 91936–91963). ICLR.

Nikhil, S. S. (2024), Accurate Prediction of Heart Disease Using Machine Learning: A Case Study on the Cleveland Dataset. *International Journal of Innovative Science and Research Technology*, 9(7), 1042–1049. https://doi.org/10.38124/ijisrt/IJISRT24JUL1400

Ras, G., Xie, N., van Gerven, M., & Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73, 329–396. https://doi.org/10.1613/jair.1.13200

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Shrestha, D. (2024). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction Using the Cleveland Heart Disease Dataset. *Preprints.org*. https://doi.org/10.20944/preprints202407.1333.v1

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning,* (pp. 3145–3154). MLR. https://proceedings.mlr.press/v70/shrikumar17a.html

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, 1–18.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sujon, K. M., Hassan, R. B., Towshi, Z. T., Othman, M. A., Samad, M. A., & Choi, K. (2024). When to use standardization and normalization: Empirical evidence from machine learning models and XAI. *IEEE Access*, 12, 135300–135314. https://doi.org/10.1109/access.2024.3462434

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951