Understanding the Black Box: A Deep Empirical Dive into Shapley Value Approximations for Tabular Data

SUCHIT GUPTE, The Ohio State University, USA
JOHN PAPARRIZOS, The Ohio State University, USA

Understanding the decisions made by machine learning models is significant for building trust and enabling the adoption of these models in real-world applications. Shapley values have emerged as a leading method for model interpretability, offering precise insights by quantifying each feature's contribution to predictions. However, computing Shapley values requires exploring all possible combinations of features, which can be computationally expensive, especially for high-dimensional data. This challenge has led to the development of various approximation techniques, often composed of estimation and replacement strategies, to compute the Shapley values efficiently. Our study focuses on the interpretability of machine learning models for tabular datasets, one of the most common and widely used data type. However, the abundance of options has created a substantial gap in determining the most appropriate technique for practical applications. Through this study, we seek to bridge this gap by comprehensively evaluating Shapley value approximations, covering 8 replacement and 17 estimation strategies across diverse regression and classification tasks. The evaluation is conducted exclusively on tabular data, leveraging 200 synthetic and real-world datasets, covering a wide range of model types, from conventional tree-based and linear models to modern neural networks. We focus on computational efficiency and the consistency of Shapley value estimates in handling high-dimensional feature spaces. Our findings reveal that traditional sampling-based approaches significantly reduce computational costs but fail to capture complex feature interactions. On the contrary, model-specific approaches that exploit the structure of the underlying model consistently outperform model-agnostic techniques, delivering higher accuracy and faster computations. Through the study, we aim to encourage further research on Shapley value approximations, advancing data-centric explainable AI.

CCS Concepts: • **General and reference** → **Empirical studies**; Surveys and overviews; • **Computing methodologies** → *Feature selection*; • **Information systems** → Data model extensions.

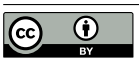
Additional Key Words and Phrases: Shapley Values; Shapley Value Approximations; Data-Centric AI

ACM Reference Format:
Suchit Gupte and John Paparrizos. 2025. Understanding the Black Box: A Deep Empirical Dive into Shapley Value Approximations for Tabular Data. *Proc. ACM Manag. Data* 3, 3 (SIGMOD), Article 232 (June 2025), 31 pages. https://doi.org/10.1145/3725420

1 Introduction

Machine learning (ML) and artificial intelligence (AI) have witnessed significant advances in recent decades. The deployment of ML models to solve real-world problems has increased due to their ability to outperform humans in terms of efficiency [140, 147]. The application of ML models also extends to various domains [51, 77, 89, 105], including healthcare [58, 116, 117] and criminal justice [39], where decisions must be accurate, fair, and transparent. A viable strategy for building confidence in machine learning models is interpretability, which refers to understanding and explaining their decision-making processes. However, as models with increasingly complex

Authors' Contact Information: Suchit Gupte, The Ohio State University, Columbus, Ohio, USA, gupte.31@osu.edu; John Paparrizos, The Ohio State University, Columbus, Ohio, USA, paparrizos.1@osu.edu.



ACM 2836-6573/2025/6-ART232
https://doi.org/10.1145/3725420

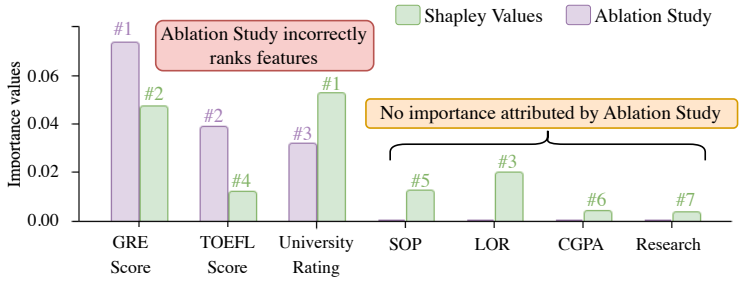


Fig. 1. Comparison between Ablation study and Shapley values on graduate admissions dataset [91]. Shapley values offer more detailed explanations by evaluating all possible feature subsets, unlike the Ablation study

architectures, such as deep networks [40, 69, 71], gain prominence, the challenge of interpretability grows. These black-box architectures make it challenging to trace the specific decisions the models make. Unfortunately, the pursuit of higher accuracy has often driven a shift towards these more sophisticated models, further affecting interpretability [29, 43, 50, 107, 129, 153, 160].

In database systems, interpretability [30, 38, 49, 64, 65, 73, 74, 82, 83, 85, 93–96, 106, 112, 114, 118, 125, 130, 134–136, 139, 145, 148, 152, 161, 162, 164] is especially critical in tasks such as query execution, query optimization, indexing, similarity search, and data cleaning, where understanding the impact of specific tuples or features on aggregate outcomes is essential. However, the growing volume of data complicates the selection of relevant data features and the right data size for training. Redundant data adds noise and makes training expensive, while insufficient data hampers the model performance. Hence, understanding which features contribute most to a model's predictions helps prioritize relevant data, simplifying dataset selection and improving training efficiency and accuracy. To satisfy this need, Shapley values [81, 137] have emerged as a leading feature explanation technique for identifying the impact of individual features in a model's decision-making process.

The concept of Shapley values [137], originally developed in cooperative game theory, was subsequently adopted to explain machine learning models by modeling the prediction task as a cooperative game. In this setting, each feature functions as a player in the game, collectively contributing to the prediction task. Estimating feature contributions using Shapley values is analogous to the conventional Ablation study [55], where a feature is systematically removed to observe its impact on the model output. However, unlike the Ablation study, Shapley values go beyond isolating individual features and instead estimate the contribution of a feature across *all* potential subsets of the feature set. This exhaustive approach allows Shapley values to provide a thorough and detailed feature comprehension. Evidently, as illustrated in Figure 1, performing an Ablation study ignores several critical features and even fails to pinpoint the most influential feature. On the contrary, Shapley values offer a more granular and accurate comprehension of individual feature contributions, showcasing their superior interpretability.

Using Shapley values for feature explanations is a straightforward solution. However, this solution involves two significant drawbacks. The first drawback arises when dealing with absent features. When considering a subset of the feature set, some features are bound to be missing. Handling these missing features without skewing the interpretation of feature contributions is crucial. Various replacement strategies [47, 61, 80, 81, 128, 143, 159] have been proposed to address this problem, such as imputing missing values or using a surrogate model to capture the behavior

Table 1. A detailed comparison between the key aspects covered by works on Shapley value approximations.

	Strategies evaluated		# Datasets	Analysis		
	Replacement	Estimation		Accuracy	Runtime	Statistical
[23]		✓(3)	1-3	✓		
[33]	✓(8)		3-5	✓	✓	
[131]		✓(5)	-			
This work	**✓(8)**	**✓(17)**	**200**	**✓**	**✓**	**✓**

of absent features based on the present features. Another drawback is the exponential complexity of the Shapley values. Due to its exhaustive nature, computing the Shapley values for all features is computationally expensive. Numerous estimation strategies [20, 62, 81, 100, 141] have emerged to efficiently compute the Shapley values in polynomial time, effectively addressing this drawback.

Various Shapley value approximations [1, 4, 20, 24, 32, 47, 52, 80, 81, 88, 100, 141, 143, 150] develop from the effective combination of an appropriate replacement strategy and a robust estimation strategy. The abundance of such approximations has motivated the development of a standardized framework called SHapley Additive exPlanations (SHAP) [81]. Although some of these approximations [24, 32, 80, 81, 141] are part of the SHAP framework, others [1, 4, 18, 31, 52, 62, 86, 88, 90, 100, 143, 150] continue to exist independently. Additionally, these approximations are divided into: model-agnostic and model-specific solutions. The model-agnostic solutions [1, 20, 32, 47, 52, 81, 88, 100, 141, 143] are simple and flexible but rely on random sampling, adding variability. In contrast, the model-specific solutions [4, 24, 80, 150] offer a significantly faster estimation of the Shapley values by leveraging model properties to mitigate the exponential complexity.

The widespread use of Shapley value approximations demonstrates their reliability as an interpretable method. However, despite the considerable progress made over the decades, a comprehensive evaluation of these approximations is notably absent in the existing literature. The existing surveys [23, 33, 131] focus primarily on theoretical discussions, performing very limited empirical evaluations of the various approximation techniques. Our research aims to address this significant gap by systematically evaluating the consistency, reliability, and scalability of the various Shapley value approximations. Table 1 presents a comprehensive analysis of the key aspects addressed in our study relative to existing surveys. In this study, we focus on tabular datasets, specifically for regression and classification tasks, because they offer well-defined features, allowing for precise evaluation of Shapley value approximations. Unlike domains such as image, text, or time series, tabular data minimizes the need for extensive preprocessing or feature engineering, ensuring the evaluation focuses purely on the accuracy and performance of Shapley value approximations.

We break down the approximation of the Shapley values into two principal dimensions. These dimensions also serve as a guide for setting up the evaluation framework. The first dimension involves properly treating missing values with the help of different replacement strategies. We deploy each replacement strategy against an exhaustive computation of Shapley values. This evaluation measure will highlight the strengths and weaknesses of replacement strategies, aiding future research in selecting the most reliable strategy. The second dimension focuses on tractable estimation strategies, which are crucial for efficiently computing Shapley values. We analyze the performance of these tractable estimation strategies using established approximation algorithms. We systematically evaluate 8 distinct replacement strategies and 17 distinct approximation algorithms across a diverse set of 200 datasets. This comprehensive evaluation enables us to thoroughly assess the performance and efficacy of individual strategies and the various approximations in estimating

Shapley values across varied data scenarios. We open-source our code [53, 54] to ensure fairness, reproducibility, and to encourage further research in this field.

Our analysis reveals that traditional sampling-based approaches significantly reduce computational costs but lead to higher variance in the Shapley value estimates. Several other estimation strategies consistently delivered superior performance compared to the conventional sampling approach. Similarly, in terms of replacement strategies, instance-conditioned methods consistently capture high-order interactions more accurately than global strategies, offering a more reliable approach. Furthermore, our study shows that while model-agnostic approximations using suitable replacement and estimation strategies provide reasonable accuracy and efficiency, model-specific approaches leveraging the model's internal structure deliver significantly better results.

We first discuss the related work and the necessary background for the Shapley values (Section 2). Then, we present our contributions as follows:

- We provide an extensive overview of the various model-agnostic and model-specific approximation algorithms designed to efficiently estimate Shapley values. (Section 3).
- We present the different evaluation measures tailored to assess the various dimensions of the Shapley value approximation techniques (Section 4).
- We conduct a comprehensive study on 200 datasets, examining the effectiveness of the replacement strategies with an exhaustive estimation of Shapley values (Section 5.1).
- We perform a quantitative and a qualitative assessment of 17 distinct model-agnostic and model-specific methodologies of approximating the Shapley values (Section 5.2).
- We offer guidelines for selecting appropriate Shapley value computation strategies based on data characteristics, model complexity, and resource availability (Section 6).

Finally, we conclude with the implications of our work (Section 7)

2 PRELIMINARIES AND RELATED WORK

We first introduce the necessary background relevant to Shapley values (Section 2.1), followed by an overview of the application of this solution in explaining ML models (Section 2.2). Further, we introduce the use of Shapley values in databases (Section 2.3), and subsequently, we delve into the drawbacks of estimating Shapley values and solutions to overcome them (Section 2.4).

2.1 Shapley values in game theory

Shapley values [137, 156] have become increasingly popular in game theory due to their ability to ensure fair distribution of credit. In a cooperative game setting, where a group of players work together to receive a payout, a critical concern lies in fairly allocating the payout amongst the participants. The challenge in the fair allocation of the payout is to estimate the exact contribution of each player towards attaining the total payout. To tackle the above challenge, Shapley values are employed as a measure of importance, indicating the significance of each player's contribution.

To facilitate comprehension of the notion of Shapley values, we briefly overview the process of estimating the Shapley value for a player playing a cooperative game. Specifically, given a set of players (D), let us consider a subset ($S \subseteq D$) of the set of all the players. For the remainder of this paper, we will refer to the subset of players as a coalition. Let the payout attained by the coalition S be $v(S)$. Thus, $v(\phi) = 0$, and $v(D)$ is the total attainable payout through the game. Our goal is to allocate $v(D)$ fairly among the members of D with the help of the Shapley values. The difference between payouts attained when player i takes part in the coalition game represents the contribution of player i towards S. We refer to this contribution as player i's marginal contribution towards the coalition S. The total contribution of player i is the average marginal contribution of

player i over all possible coalitions $S \subseteq D$. Assuming that we know the payouts obtained by each coalition $S \subseteq D$, the Shapley value of player i can be defined as follows:

$$\Phi_i = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} [v(S \cup \{i\}) - v(S)] \tag{1}$$

Despite the simplicity of the Shapley values as a solution, they are supported by robust theoretical properties [137, 156]. The theoretical robustness of the Shapley values has led to their widespread recognition. Shapley values are relevant in numerous fields other than just cooperative game theory. Shapley values find significant applications in ML. In the subsequent section, we provide a comprehensive overview of the utilization of the Shapley values to explain complex ML models.

2.2 Shapley values in machine learning

A fundamental supervised machine learning framework involves training a black-box model f on a dataset consisting of features $x_1, \ldots x_d$, where f makes predictions for unknown instances. To establish confidence in the predictions made by f, f must possess a high level of interpretability. When interpreting a simple model, the most efficient strategy is to utilize the model itself. If f is a linear model of the form $f(x) = w_1x_1 + \cdots + w_dx_d$, ($w_i$: weight coefficient of feature x_i in attaining $f(x)$), then the model representation suffices to generalize individual feature contributions. However, using complex models such as ensembles, boosting, or deep networks for self-explanation is not feasible because of their opaque structure.

LIME [128], a widely used approach, leverages the concept of linear models to explain complex models. *LIME* offers an approximate explanation of a complex model by squeezing it into an interpretable version that accurately captures the model's behavior for a specific instance. Specifically, *LIME* trains a local surrogate model to explain individual predictions of the original black-box ML model. However, the prediction capacity of the surrogate model poses a limitation in achieving predictions that would accurately represent the original model [3, 126]. Therefore, an ideal scenario demands the utilization of the original model to provide explanations.

The concept of Shapley values helps to meet the aforementioned demand. The prediction task of the black-box model corresponds to the coalition game. The input features are the players of the game. Consequently, the objective boils down to explaining an individual prediction by allocating a Shapley value to each feature, signifying its contribution towards attaining the prediction. Formally, given a black-box model f, an explicand, or the instance to be explained x^e, feature set D, and a coalition of the feature set $S \subseteq D$, the Shapley value of input feature i can be expressed as follows:

$$\Phi_i = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} [f(x^e_{S \cup \{i\}}) - f(x^e_S)] \tag{2}$$

The model prediction of the explicand, denoted as $f(x^e_S)$, represents the model prediction when only the features $i \in S$ are visible to the model. The total contribution of feature i is the average marginal contribution of feature i over all possible feature coalitions $S \subseteq D$. Therefore, assuming that we know the model prediction for each of the $2^{|D|}$ feature coalitions, we can compute the contribution of individual features towards attaining the model prediction. Shapley values are additive in nature. Analogous to the distribution of the total payout among players in a coalition game, the Shapley values of distinct features sum up to the model's prediction. The additivity attribute enables a thorough scrutiny of the predictions, leading to a deeper insight into the significance of different features in the model's decision-making structure.

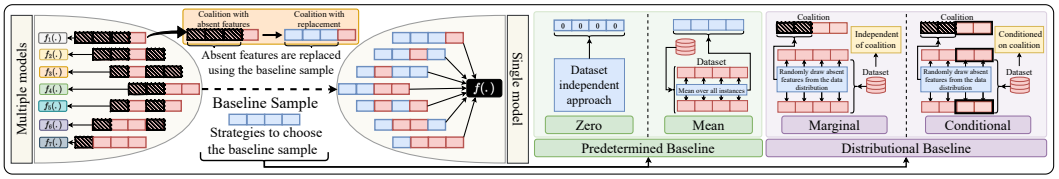


Fig. 2. Replacement strategies such as Predetermined Baseline and Distributional Baseline address the absence of features, eliminating the necessity to train an exponential number of models and mitigating computational complexity. The Predetermined baseline imputes missing data with zeros or the mean, while the Distributional baseline samples missing feature values based on specified distributions, such as the Marginal (independent of explicand) or the Conditional (dependent on explicand).

2.3 Shapley values in databases

The application of Shapley values in the context of databases [7, 8, 37, 49, 59, 63, 66, 67, 79, 82, 84, 85, 125, 161] necessitates a shift in focus from the individual features of data instances to the specific tuples within the database that influence the outcome of a query. Here, the utility function is defined not by the predictions of a model but by the results of an aggregate query [7, 8, 37, 66, 67, 78], which may include operations such as summation, averaging, or counting. The primary challenge is determining each tuple's contribution in obtaining the result. For instance, in a database that holds sales records, a query that calculates the total revenue would allow for determining the Shapley value for each tuple, reflecting its contribution to the overall revenue. In this scenario, the utility function corresponds to the total revenue, while the tuples represent the entities whose contributions are being assessed. Another important application of Shapley values in databases involves treating tables as players [30, 83, 85, 125, 152], particularly in revenue distribution within data markets. These markets require collaboration between data owners to generate datasets, and Shapley values are used to fairly allocate revenue based on each owner's contribution.

While the focus in databases shifts from features to tuples, the computational challenges remain similar to those in machine learning. Similar challenges also arise in fields like genomics or finance, where the interpretation of time series [9, 11, 14, 16, 41, 42, 103, 104, 111, 113, 115, 155] is critical for downstream tasks [5, 10, 12, 13, 15, 75, 76, 108–110, 119, 144]. Explanation techniques like TimeSHAP [6], and WindowSHAP [97] efficiently handle such extremely high-dimensional data with extensive processing. In the above-mentioned domains, efficient sampling and replacement strategies still play a crucial role in managing the complexity of Shapley value computation. Hence, our analysis of these strategies provides valuable insights that can be leveraged in data-centric AI systems and data markets, supporting accurate data valuation and transparent attribution.

Beyond databases, Shapley values are widely used for diverse applications, such as evaluating model performance by quantifying the contributions of individual data points [154]. They facilitate subgroup behavior analysis by uncovering patterns in classifier behavior across various subgroups [121]. In the context of model rankings, Shapley values assist in identifying biased subgroups within ranking and classification tasks, promoting fairness in machine learning models [57, 120]. Furthermore, in federated learning, Shapley values are employed to assess participants' contributions to collaborative learning, ensuring equitable evaluation while preserving data privacy [72]. These applications highlight the adaptability of Shapley values in delivering interpretable insights across multiple areas of data science and machine learning.

2.4 Efficient Computation Strategies

The Shapley values appear to offer a straightforward solution for explaining any black-box ML model. However, this seemingly simple solution comes with a significant drawback. To estimate

the Shapley values of individual players, one must possess the knowledge of the payouts attained by each coalition set of the players (Refer Section 2.1 for the underlying assumption of Equation 1). Similarly, in the context of machine learning, where the objective is to generate predictions, one must know the model prediction for every possible coalition of the feature set. The original model trained on a dataset containing all the input features will not be able to generate a prediction for an arbitrary coalition that may only include a subset of the feature set. Thus, given a feature coalition $S \subseteq D$, an explicand x^e and a black box model f trained on input features $x_1, \ldots, x_d$, the model prediction of the coalition is defined as $f(x^e_S) = f_S(x^e_S)$, where f_S is an extension of the original model trained only on features in the coalition set S.

Consequently, the estimation of Shapley values of all features demands training a separate model [165] for each coalition $S \subseteq D$. However, there are 2^d feature coalitions (d: cardinality of the feature set D), and training a distinct model for every coalition can be pretty cumbersome. Moreover, as the number of features increases, the number of coalitions will grow exponentially, requiring the training of an exponential number of models. Training and maintaining an exponential number of models can be time-consuming, resource-intensive, and impractical. Thus, despite the straightforwardness and the theoretical robustness of the Shapley values, this computational burden poses a significant drawback to its application in explanations. Dealing with the exponentially growing complexity is critical for effectively implementing Shapley values in explaining models. The following section will provide a succinct overview of various strategies to combat this challenge.

2.4.1 ***Strategies for handling the absent features:*** We use the notion of present and absent features to address the aforementioned computational complexity better. When examining a feature coalition S, the features that form S are designated as present features, while the remaining features are regarded as absent. By effectively handling the values of the absent features, we can eliminate the requirement of training an exponential number of models. The strategies for handling absent features can be classified into two categories. Refer to Figure 2 for a description of each category.

- **Predetermined Baseline:** A modified instance is defined by considering a predetermined baseline sample, which serves as a reference point for treating the absent features of the explicand. When given a feature coalition S, a predetermined baseline sample x^b, and an explicand x^e, this approach defines the modified instance as follows:

$$x_i = \begin{cases} x^e_i & \text{... if } i \in S \\ x^b_i & \text{... otherwise} \end{cases}$$

  Thus, the modified instance is comparable to the explicand, except the features not present in the coalition are extracted from the predetermined baseline. Now using the baseline sample, we can approximate the prediction of the coalition-specific extension of the original black box model $f_S(x^e_S)$ as follows $f(x^e_S, x^b_{\bar{S}})$. The most predominant choices for the predetermined baseline are the **all-zeros** [124, 133, 159] and the **default** [36, 128, 143] baseline. As the name suggests, the *all-zeros* baseline involves replacing absent feature values with zeros. This method assumes that the absent feature values have no significant impact on estimating the Shapley values and can be safely replaced with a neutral value. It is a straightforward, easy-to-implement solution, especially with large datasets. On the other hand, the *default* baseline uses a user-defined sample to replace the absent features. Since we are focusing on regression-based models, the mean baseline is of concern. The rest of the approaches [44, 45, 70, 163] are tailored explicitly for computer vision problems and are beyond the scope of this study. The mean baseline calculates the average of the feature column from the training dataset and utilizes it to replace the absent features. The mean value replacement technique

aims to preserve the overall distribution dataset and provides a better approximation of $f_S(x_S^e)$.

- **Distributional Baseline:** Instead of relying on a fixed baseline and imputing absent features with a predetermined value, this approach allows a more flexible and probabilistic treatment of missing data. This replacement strategy treats the absent features as random variables by drawing their values from the data distribution. The data distributions are categorized into two main distributions: **the marginal distribution** and **the conditional distribution**.
  The *marginal distribution* handles absent features independent of the present features by sampling the missing values according to the distribution $p(X_{\bar{S}})$. Consequently, we can modify the definition of the model prediction as follows:

$$f_S(x_S^e) \approx \mathbb{E}[f(x_S^e, X_{\bar{S}})]$$

  Conversely, the *conditional distribution* addresses absent features by leveraging the present features. Unlike the *marginal distribution*, the *conditional* approach does not assume feature independence. The absent feature values are drawn according to the conditional distribution $p(X_{\bar{S}}|X_S = x_S^e)$, thereby altering the definition in the following manner:

$$f_S(x_S^e) \approx \mathbb{E}[f(X)|X_S = x_S^e]$$

  The *marginal distribution* approach is employed through **an empirical strategy** [60, 81, 88, 128, 143]. This strategy entails randomly drawing a set of instances from the training data independent of the explicand and determining the prediction for a particular coalition by averaging over the sampled set of instances. Using an *empirical strategy* to handle *conditional distribution* involves randomly sampling a set of instances from the training data conditioned on the present features of the explicand. However, a caveat associated with this approach is the potential occurrence of an empty set, resulting in inaccurate Shapley value estimates.
  To address this concern, there exist several strategies: **the Parametric Assumption** [47] assumes that the data follows either a *Gaussian* or a *Copula* distribution; **the Generative model** [21, 157] trains a deep learning model to predict missing feature values by comprehensively learning all the conditional data distributions; **the Surrogate model** [27, 61] trains a deep learning model to predict the absent features using the target label of the explicand. All these strategies use the *conditional distribution* approach of treating the absent features.

Thus, implementing one of the aforementioned replacement approaches suffices to eliminate the need to train an exponential number of models. However, handling an exponential number of coalitions still makes the estimation of Shapley values challenging. In the subsequent section, we explore a method for mitigating this drawback by employing random sampling techniques.

2.4.2 ***Tractable estimation strategies:*** The random sampling approach was the first intuitive solution to tackle the exponential complexity of computing Shapley values [20, 141]. Randomly selecting subsets of feature combinations instead of analyzing every possible one significantly reduces the computational burden while maintaining a comparable explanation quality to that of exhaustive methods. However, this random sampling approach can introduce variability in estimates, which can be minimized using more advanced techniques [18, 19, 90, 146]. Beyond random sampling, several estimation strategies offer polynomial-time solutions for Shapley value approximation. These include multilinear sampling [100], optimization-based methods [32, 47, 81], and model-specific solutions [4, 24, 80]. Together with replacement strategies (Section 2.4.1), these methods form the foundation for efficient Shapley value approximations. The various Shapley value approximations are summarized in the subsequent section.

Table 2. A detailed summary of model-agnostic and model-specific Shapley value approximations based on various estimation and replacement strategies. "M" and "C" denote Marginal and Conditional distribution replacements, respectively. The "Complexity" column outlines time and space complexities in big-O notation, while "Approximation Guarantees" indicate theoretical bounds. Notation includes d (features), m (sampled subsets), k (SGD iterations), L (tree leaves), D (tree depth), n (neural network neurons), and r (reference samples). The "Usage" column lists implementation languages and marks methods implemented from scratch with a "✓"; others rely on available GitHub code.

	Approaches	Strategies		Complexity		Approximation Guarantees	Usage
		Estimation	Replacement	Time	Space		
Model-agnostic	Exhaustive sampling	Exact	Separate models	$O(2^d)$	$O(2^d)$	Gold standard for accuracy	*Python* | ✓
	IME [141]	RO	Empirical (M)	$O(md)$	$O(m+d)$	$O(1/\sqrt{m})$ convergence rate	*Python* | ✓
	CES [143]	RO	Empirical (C)	$O(md)$	$O(m+d)$	Reduces variance compared to IME	*Python* | ✓
	Cohort [88]	RO	Empirical (C)	$O(md)$	$O(m+d)$	No formal approximation guarantees	*Python* | ✓
	MLE [100]	MLE	Empirical (M)	$O(md)$	$O(m+d)$	$O(1/m)$ convergence rate	*Python* | ✓
	Kernel [32]	WLS	Empirical (M)	$O(md^2)$	$O(md)$	$O(1/m)$ convergence rate with optimized sampling	*Python*
	SGD-Shapley [52]	WLS	Mean	$O(kd)$	$O(d)$	$O(1/\sqrt{k})$ convergence with SGD	*Python*
	Parametric [47]	WLS	Gaussian/Copula	$O(md^2)$	$O(d)$	Same as KernelSHAP [32]	*Python* | ✓
	Non-Parametric [47]	WLS	Empirical (C)	$O(md^2)$	$O(md)$	Same as KernelSHAP [32]	*Python* | ✓
	FastSHAP [62]	WLS	Surrogate model	$O(md)$	$O(n)$	Amortized approach for real-time Shapley value estimation. $O(1)$: inference time	*Python*
Model-specific	Linear [24]	Linear	Empirical (M)	$O(d)$	$O(d)$	Best if features are independent	*Python*
	Correlated Linear [24]	Linear	Gaussian	$O(d^2)$	$O(d^2)$	Assumes Gaussian data distribution	*Python*
	Tree interventional [80]	Tree	Empirical (M)	$O(LD)$	$O(LD)$	Polynomial time exact computation	*Python*
	Tree path-dependent [80]	Tree	Empirical (C)	$O(LD)$	$O(LD)$	Captures feature interactions	*Python/C++*
	DeepLIFT [127]	Deep	All-zeros	$O(n)$	$O(n)$	No formal approximation guarantees but empirically approximates backprop	*Python*
	DeepSHAP [25]	Deep	Empirical (M)	$O(nr)$	$O(nr)$	Extends DeepLIFT for better accuracy	*Python*
	DASP [4]	Deep	Mean	$O(nd)$	$O(nd)$	Uses uncertainty prop to reduce bias	*Python*

3 Shapley Value Approximations

There are several approximation approaches proposed to make the computation of Shapley values feasible. These approaches can be broadly classified into model-agnostic and model-specific approaches. Model-agnostic approaches can be applied to any model regardless of their type. Model-specific approaches are designed to provide an edge by utilizing that specific model's properties. We will now offer a concise overview of each approach category, followed by a comprehensive list of the approaches falling under each category in Table 2.

3.1 Model-agnostic approximations

3.1.1 ***Semi Value (SV):*** The original coalition game, defined using Shapley values, is known as the *SemiValue* [99] estimation strategy. The Shapley value of a feature can be computed by averaging its marginal contribution across all possible feature coalitions, as shown in Equation 1. However, this strategy still grapples with the issue of handling an exponential number of coalitions. To tackle this challenge, Castro *et al.* [20] introduced an alternative method called **ApproSemiValue**. This approach involves sampling coalitions based on the probability distribution obtained from the weight function. Thus, implementing the *SemiValue* strategy demands sampling of coalitions according to the distribution: $P(S) = \frac{|S|!(|D|-|S|-1)!}{|D|!}$. While *ApproSemiValue* successfully reduces the time complexity, drawing coalitions according to the probability distribution $P(S)$ is quite challenging. Moreover, this method does not offer any solution for handling the absent features.

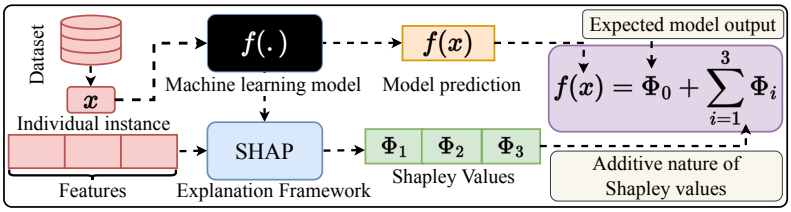


Fig. 3. Shapley values, an additive local feature attribution technique central to Weighted Least Squares (WLS) estimation strategy. In this illustration, the instance is composed of 3 features. An individual model prediction is expressed as a sum of the average model output and the Shapley values.

Local Shapley (L- Shapley) and **Connected Shapley (C- Shapley)** [26] are two approaches based on the *SemiValue* estimation strategy. These approaches are explicitly tailored for structured data like images with significant spatial correlation and, hence, are outside of the scope of our study. Apart from *L-Shapley* and *C-Shapley*, no approximation utilizes the *SemiValue* strategy.

3.1.2 ***Random Order (RO):*** The initial approach to calculating the Shapley values incorporates a weight function assigned to each coalition. The size of the coalition determines the value of this weight function. However, we can eliminate the need for a weight function by modifying the solution to work with permutations of features instead of feature subsets. Consequently, the modified solution can be formulated in the following manner:

$$\Phi_i = \frac{1}{|D|!} \sum_{\pi \in \Pi(D)} (v[Pre_\pi(i) \cup \{i\}] - v[Pre_\pi(i)]) \tag{3}$$

In the above expression, $\Pi(D)$ represents the set of all permutations of the feature set D. $Pre_\pi(i)$ denotes the set of features preceeding feature i in a specific permutation of features $\pi \in \Pi(D)$. The marginal contribution of feature i towards the permutation π is the difference in the model predictions when the feature i is included in $Pre_\pi(i)$. Now, since there are $|D|!$ total permutations, the total contribution of a feature is averaged over all the permutations instead, eliminating the concept of the weight function from Equation 1.

With this modified definition, the focus shifts from randomly sampling subsets of features to randomly sampling permutations from the set of all permutations of the feature set. This estimation technique for the Shapley values is referred to as Random Order [92, 137]. Various approaches, including **IME (Interactions-based Method for Explanation) [141]**, **CES (Conditional Expectations Shapley) [143]**, **Shapley Cohort refinement [88]**, and **Surrogate models [1]** utilize this technique in combination with one of the replacement strategies from Section 2.4.1.

3.1.3 ***Multilinear Extension (MLE):*** Owen introduced a *multilinear extension* [102] of the Shapley values. It involves imposing a probabilistic structure on the feature space, where each feature j is treated as a random variable with a probability $0 \leq q \leq 1$ of including in a coalition. Consequently, each coalition is represented as a random variable E_j. Using the above-mentioned probabilistic structure, the Shapley value can be defined as follows:

$$\Phi_j = \int_0^1 \mathbb{E}[v(E_j \cup \{x_j\}) - v(E_j)]\, dq \tag{4}$$

Owen [102] established that the summation in Equation 2 can be transformed into an integral by treating the coalitions as random variables. Based on Owen's notion of the *multilinear extension*, Okhrati and Lipani later introduced a sampling approximation approach known as the **MLE (MultiLinear Extension sampling) [100]** for estimating the Shapley values.

3.1.4 ***Weighted Least Squares (WLS):*** Figure 3 demonstrates the additive property of the Shapley values. Thus, we can represent the model's prediction as a summation of the average model output and the Shapley values associated with each feature. Hence, in this specific scenario, determining the Shapley values can be perceived as an optimization problem, wherein the objective is to solve the below expression using a *Weighted Least Squares (WLS)* [22, 132] approach.

$$\min_{\Phi_i, \forall 1 \leq i \leq |D|} \sum_{S \subseteq D} W(S) \left[(\Phi_0 + \sum_{i \in S} \Phi_i) - v(S) \right] \tag{5}$$

$$\textbf{Weighting Kernel: } W(S) = \frac{|D| - 1}{\binom{|D|}{|S|} |S| (|D| - |S|)}$$

In the above expression, $W(S)$ is the weighting kernel, and Φ_0 is the average model prediction. When $S = D$, the sum of the average model prediction and the Shapley values is equivalent to the actual model prediction of the instance. Thus, when $S = D$, the inner expression ultimately reduces to zero. **KernelSHAP [32, 81]** aims to approximate this weighted least squares problem by sampling a subset of coalitions. This sampling is done according to weighting kernel $W(S)$. **SGD-Shapley [52]** is an alternative method that extends the principles of *KernelSHAP*. However, it employs projected gradient descent to solve the least squares problem approximately. **FastSHAP [31, 62]** is a novel technique that uses the least squares approximation. Unlike other methods, *FastSHAP* trains a surrogate model to estimate the Shapley values in a single forward pass by amortizing the training process over the training samples.

3.2 Model-specific approximations

Model-specific approximations are curated using a removal strategy and a sampling technique that leverages the model's inherent structure, enabling a significantly faster estimation of Shapley values. The scientific community has proposed approaches for three different model categories: linear, tree, and deep learning. In the following subsections, we briefly discuss each model type and the corresponding approximation techniques suggested for them.

3.2.1 ***Linear models:*** Linear models are self-interpretable. As discussed in Section 2.2, a linear relationship between the input features and the model prediction allows the weight coefficients to effectively explain the impact of individual features on the model's prediction. Thus, Shapley values are integrated to work for linear models by leveraging the concept of weight coefficients. **LinearSHAP [81, 142]** and **Correlated LinearSHAP [24]** are the two approximation techniques designed for linear models. The vanilla version of *LinearSHAP* incorporates a marginal feature removal approach, whereas the correlated version computes conditional Shapley values (refer Section 2.4.1). The *Correlated LinearSHAP* method assumes that the data distribution conforms to a *multivariate Gaussian distribution*, thereby introducing the possibility of producing inaccurate Shapley value estimates when the data does not align with the distribution.

3.2.2 ***Tree-based models:*** Tree-based models include decision trees [48], ensemble learning like random forests [17], and gradient boosting models like XGBoost [28]. These non-linear models are affected by the interdependencies among the input features. **Interventional TreeSHAP [80]** and **Path-dependent TreeSHAP [80]** are able to approximate Shapley values accurately by leveraging the tree structure. We can depict a tree structure by breaking it down into individual outputs for every leaf within the tree. As a result, the impact of each leaf on the Shapley value of a particular feature can be determined at the leaf level, viewing it as a coalitional game where the players are the features found along the path from the root to the current leaf. A dynamic programming approach helps to generate explanations for the Shapley values of all features simultaneously as it traverses

through the nodes in the tree. The *Interventional TreeSHAP* method assumes that the features are independent and employs the empirical marginal feature removal approach (see Section 2.4.1). In contrast, the *Path-dependent TreeSHAP* method adopts a conditional feature removal approach derived from the *Shapley Cohort refinement* approximation [88].

3.2.3 ***Deep learning models:*** Deep neural networks [71] are gaining popularity due to better hardware, more data, and more innovative techniques. They are widely used across industries for their ability to solve complex problems effectively. The structures consist of multiple layers that increase opacity levels, resulting in models that are extremely difficult to interpret. One of the initial approaches to explain deep models, known as **DeepLIFT**, was developed to allocate attributions throughout a deep network for a single explicand and baseline [25, 138]. The method examines the impact of alterations in input data on the network's activations across different layers. Nonetheless, the utilization of certain simplifications and approximations may occasionally produce biased Shapley value estimates. Subsequently, Lundberg & Lee [81] introduced an extension of *DeepLIFT* [127] called the **DeepSHAP** to produce biased estimates of marginal Shapley values.

Despite its bias, *DeepSHAP* is a valuable approach because its computational complexity is proportional to the size of the model and the number of baselines. **Deep Approximate Shapley Propagation (DASP)** [4] is another technique to approximate baseline Shapley values for deep models. It uses uncertainty propagation, modeling input distributions as standard random variables. It has a lower bias than *DeepLIFT* but is computationally costly.

4 Experimental Settings

In the subsequent sections, we discuss the implementation details of our evaluation.

Platform: We run our experiments on a high-performance computing server with the following configuration: 2xAMD EPYC 7713 64-Core processors and 1TB RAM. The server is equipped with two Nvidia A100 GPUs and functions on a 64-bit Ubuntu 22.04.3 LTS Linux Operating System.

Implementation: To ensure fair implementation of all the approximation techniques, we use the official GitHub repositories. In instances of code unavailability, we implement the methods based on our comprehension of the paper. The implementations are in Python(3.10), C++, and R with the following dependencies: Pytorch(1.11) [122], TensorFlow(2.6.0) [2], scikit-learn(0.22.1) [123]. We execute every Shapley value estimation technique on a single core to guarantee accurate assessments of runtime. For experiment reproducibility, we open-source the codebase [53].

Datasets: We focus on tabular datasets curated for regression and classification problems. We utilize 200 publicly available datasets from the UCI Machine Learning Repository [87]. Within the datasets, there are as many as 60 input features, and the number of instances ranges from 100 to 1 million. Figure 4 highlights the dimensions and scale of these datasets. Each dataset is split into training and testing sets for model training and computing Shapley value estimations. Since the Shapley values are a local feature attribution technique, the number of instances in the dataset has a very insignificant impact on the Shapley value estimates; however, data dimensionality significantly affects the estimation, as discussed in Section 2.4. To conduct a qualitative evaluation of the approximations, we also create a synthetic dataset designed to serve as a controlled benchmark. The dataset includes randomly generated features to ensure variability and independence. A predefined target function is used to assign specific weights to each feature, representing their contributions to the outcome. These weights act as the ground truth Shapley values.

ML Models: We utilize the supervised machine learning framework used to tackle regression and classification tasks. We use the following model architectures - Linear models [48, 68], Ensemble Learning [17], Gradient Boosting [28], Neural Networks [56], Nearest neighbors [35], Naive Bayes classifiers [149], and Support Vector Machines [34]. To conduct a thorough evaluation, we integrate

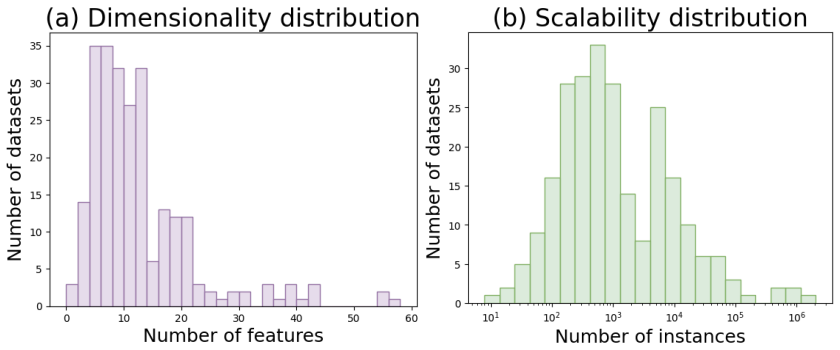


Fig. 4. (a) represents the dimensionality distribution and (b) represents the scalability distribution across 200 regression and binary/mutliclass classification datasets from the UCI ML repository [87].

models representing each category. Shapley values intend to explain a black box model by leveraging the model itself, negating the significance of its fit quality. Consequently, this allows us to use vanilla versions of each model with default hyperparameters.

4.1 Evaluation measures

We divide our analysis into two parts: the first evaluates various replacement strategies while the second assesses different Shapley value approximations. Since Shapley values are highly dependent on the model and dataset, feature importance varies across different contexts, with models assigning different importance to the same features. This variability stems from the unique behavior of each model, preventing the creation of a universal standard. As a result, an exhaustive estimation of Shapley values with models trained on an exponential number of feature subsets cannot be relied upon to produce ground truth values, making it essential to evaluate their accuracy as part of the process. Even in controlled synthetic settings with known ground-truth Shapley values, we hypothesize that Shapley value approximations preserve the relative feature rankings, though their absolute magnitudes might be scaled down. This scaling bias arises from inherent sampling limitations and computational constraints. To address this, our evaluation emphasizes metrics that focus on reconstructing model predictions rather than direct value-to-value comparisons.

We conduct quantitative and qualitative evaluations using statistical methods, including the Wilcoxon test [151] for pairwise dataset comparisons, and the Friedman test [46] with a posthoc Nemenyi test [98] to rank strategies across datasets. Below, we detail the evaluation measures.

4.1.1 ***Explanation Error:*** The absence of ground-truth Shapley values presents a clear obstacle in the evaluation. Consequently, we employ an alternative evaluation metric to assess the accuracy of the approaches, such as the explanation error [81]. The motivation for explanation error stems from the additive nature of the Shapley values, as shown in Figure 3. Shapley values indicate the individual contributions of input features towards shifting the model output from the average model prediction to the actual prediction value. Explanation error is analogous to the fidelity score metric [158], as both evaluate the alignment between the model's predictions and the reconstructed value using the Shapley values. The terms can be used interchangeably, and the choice of terminology in this study is motivated by its historical usage in similar contexts [81]. Precisely, when given a black-box model f and an explicand x^e, the prediction for the x^e can be articulated as follows:

$$f(x^e) = \Phi_0 + \sum_{i=1}^{|D|} \Phi_i \tag{6}$$

Table 3. A consolidated list of replacement strategies that are a part of our extensive evaluation. Approach refers to the primary replacement strategy, and variant refers to the methodology of implementing the primary approach.

Approach	Variant	Strategy
Predetermined	Occlusion	All-zeros [124, 133, 159]
	Default	Mean [36, 128]
Distributional	Marginal	Empirical[60]
		Uniform distribution [60]
	Conditional	Empirical [88, 143]
		Separate models
		Parametric: Gaussian [47]
		Parametric: Copula [47]

In the above equation, Φ_0 symbolizes the average model prediction, while Φ_is refer to the Shapley values assigned to each input feature. The objective of any Shapley value estimation technique is to approximate Φ_is. We can evaluate the accuracy of any approximation by comparing the model's actual prediction to the combined value of the average prediction and the estimated Shapley values. A smaller error between the two values indicates a higher precision in the approximation.

We use the R^2 test [101] to analyze this error. The R^2 test [101] is a statistical test designed for regression analysis to assess the quality of fit. R^2 values, spanning from 0 to 1, are often converted into percentages to represent the accuracy of any regression model. For computing the R^2 value, we treat $f(x^e)$ as the ground truth and $\Phi_0 + \sum_{i=1}^{|D|} \Phi_i$ as the predicted value. A strategy with an R^2 value approaching 1 indicates that it can approximate the Shapley values accurately.

4.1.2 ***Compute time:*** Since Shapley values are a local feature attribution technique, we compare the instance-wise computational efficiency of different approaches. As indicated in Section 4, the evaluation encompasses datasets that contain up to 45 features. Using the per-instance runtime comparison, we anticipate the trend of the runtime results as the dimensionality increases. We determine which methods are most suitable for handling high-dimensional data by analyzing the runtime results of different approaches.

5 Experimental Results

As mentioned earlier, we divide our analysis into two sections: Section 5.1 deals with the evaluation of replacement strategies. Section 5.2 focuses on assessing the approximations, which are a combination of the replacement strategies and the estimation strategies as mentioned in Table 2.

5.1 Analysis of Replacement Strategies

We conduct a comprehensive evaluation of various replacement strategies (Section 2.4.1) by comparing them against exhaustive Shapley value estimations, covering all possible feature coalitions. Through this evaluation, our goal is to understand which replacement strategy provides accurate and efficient Shapley value estimates. We measure accuracy using the Explanation Error metric (Section 4.1.1) and efficiency with the Computation Time metric (Section 4.1.2). By implementing an exhaustive estimation technique, we ascertain that the replacement strategy is the sole factor responsible for impacting the precision of the Shapley value estimates. We explicitly evaluate the performance of the replacement strategies mentioned in Table 3.

5.1.1 ***Accuracy:*** The regression analysis in Figure 5 reveals an exciting trend: the **Separate models** and **Gaussian** replacement strategies consistently outperform the others, delivering

notably higher accuracy across all six evaluated models. Their boxplots are clustered tightly in the upper accuracy range, highlighting their impressive ability to preserve data integrity. Conversely, the **All-Zeros** and **Uniform** replacement strategies struggle the most, particularly with complex models like Neural Networks and XGBoost. Their increased variability and lower accuracy make them stand out for their instability, while the critical diagram vividly highlights their poor performance, making them highly unreliable for providing consistent Shapley value estimates.

In contrast, for classification, the performance of the **Gaussian** and **Conditional** strategies is notably reduced, showing increased variability and less stability. Meanwhile, the **Separate models** technique continues to deliver the highest accuracy in the classification tasks as well. The **Copula** strategy comes next, trailing the **Gaussian** and **Conditional** methods but still offering moderate accuracy. The **Marginal**, **Mean**, **Uniform**, and **Zero** strategies show a clear drop in accuracy, with Uniform and Zero performing the worst. These two strategies have the widest variability and the lowest median values, highlighting their poor performance in maintaining classification accuracy.

5.1.2 ***Computational Efficiency:*** Figure 6 offers an extensive evaluation of the computational requirements for various replacement strategies across regression-based datasets. In panel (a), the compute times for individual instances are evaluated across a range of machine learning models. The findings reveal that the **Separate models** strategy consistently results in the highest computational costs across all models, significantly surpassing the compute times of alternative strategies. In contrast, the **All-Zeros** strategy exhibits the lowest computational requirements, rendering it an attractive option for scenarios where efficiency is paramount. Panel (b) delves into the correlation between feature dimensionality and compute time, indicating that all strategies face increased computation times as the number of features rises; however, the **Separate models** strategy shows a particularly steep increase, highlighting its inefficiency in high-dimensional settings. Panel (c) contrasts accuracy with compute time, showcasing a distinct tradeoff; strategies such as **Conditional**, **Gaussian**, and **Copula** manage to achieve a balance between accuracy and computational efficiency, while the **Separate models** strategy, despite its high accuracy, imposes a substantial computational burden.

Figure 7 extends this analysis to classification-based datasets, revealing similarities and differences in the computational performance of the replacement strategies. In panel (a), the instance-wise compute times are compared across different model architectures. The **Separate models** strategy again emerges as the most computationally expensive, particularly in Decision Trees and Logistic Regression models. Simpler strategies like **Zero** and **Mean** maintain lower compute times, indicating their effectiveness in classification tasks. Panel (b) examines the impact of feature dimensionality on compute time, showing that as the number of features increases, all strategies see a rise in computation time, with the **Separate models** strategy experiencing the most significant increase. However, strategies like **Conditional** and **Gaussian** show a more gradual rise, suggesting better scalability for classification tasks. Finally, panel (c) illustrates the tradeoff between accuracy and compute time, with the **Conditional**, **Copula**, and **Gaussian** strategies offering a favorable balance, whereas the **Separate models** strategy, despite its high accuracy, remains computationally intensive, especially with increasing dimensionality.

5.2 Analysis of approximations

In this section, we analyze the performance of Shapley value approximations, highlighting the trade-offs between model-agnostic and model-specific techniques, including their generalizability, implementation complexity, and performance across different architectures. For a detailed breakdown of these approaches refer to Table 2. We divide the evaluation into two subsections: Section 5.2.1 delves into the quantitative analysis, while Section 5.2.2 explores the qualitative aspects.

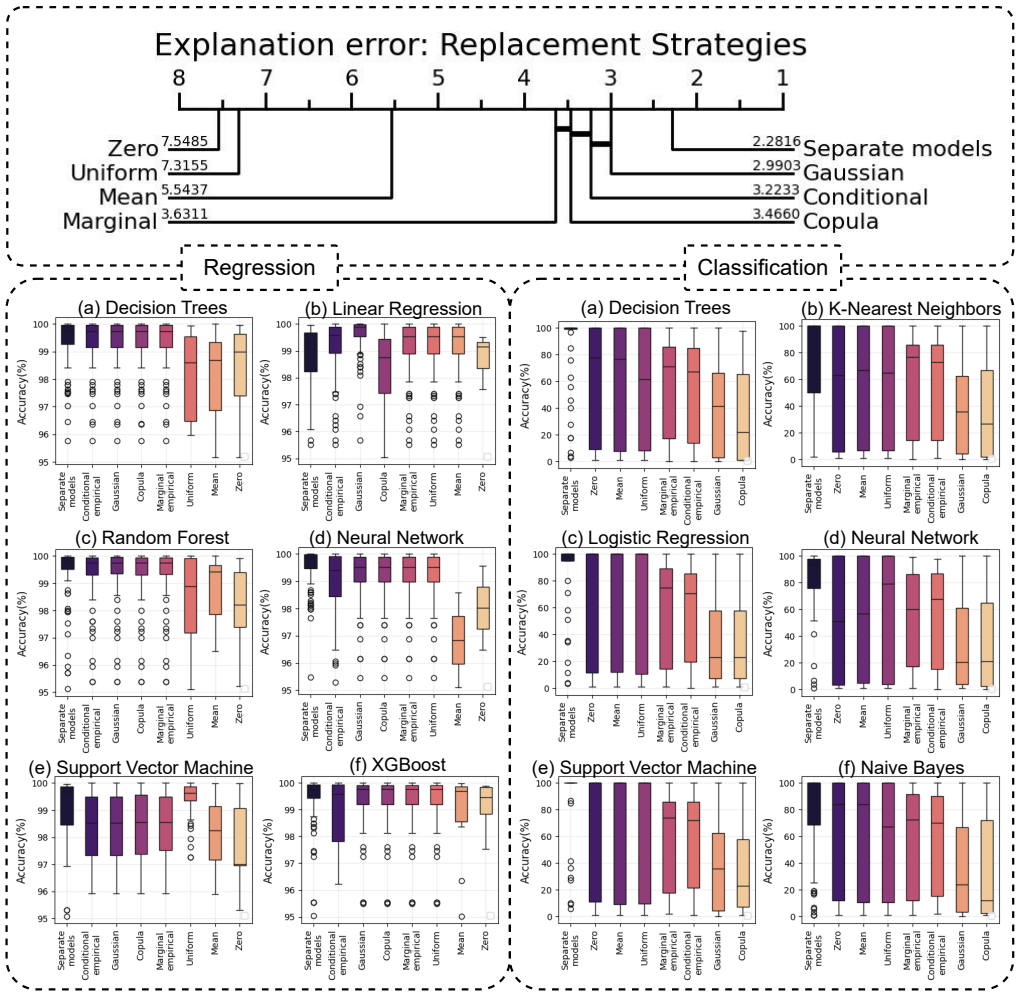


Fig. 5. Explanation error of different replacement strategies. The accuracy of each replacement strategy is computed using the R^2 test. The critical diagram shows model-agnostic and model-specific rankings, while boxplots illustrate ranking variance across 200 regression and classification datasets.

Notably, we exclude *FastSHAP* from the quantitative evaluation due to its reliance on the quality of the surrogate model. Curating 200 surrogate models per dataset is impractical, making its inclusion infeasible. As *FastSHAP*'s performance heavily depends on the accuracy of the surrogate model, a fair assessment of its effectiveness is beyond the scope of this study.

5.2.1 ***Quantitative evaluation:*** We comprehensively evaluate all approximation techniques across the 200 regression and classification tabular datasets as outlined in Section 4. This includes the time-intensive but highly accurate *Exhaustive Sampling* approach, which considers every possible feature coalition. Despite its computational burden, when paired with *Separate Models*, *Exhaustive Sampling* consistently produces Shapley value estimates that are as close to the ground truth as

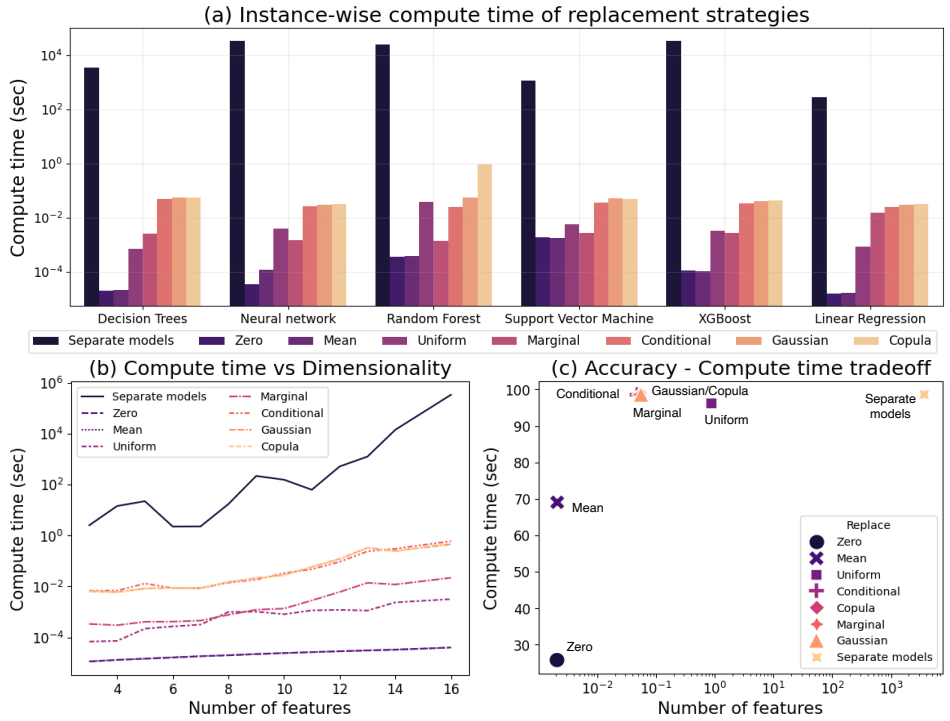


Fig. 6. (a) Compares instance-wise compute time of replacement strategies across Regression datasets. Every bar corresponds to a unique replacement strategy. (b) Demonstrates the impact of dimensionality on the estimation time of Shapley values. (c) Highlights the tradeoff between accuracy and computation time.

possible, as discussed in Section 5.1. Table 4 ranks the accuracy of each technique across four scenarios: *Model-agnostic*, *Linear*, *Tree-based*, and *Neural networks*.

The *Exhaustive Sampling* method, with its $O(2^d)$ time complexity, consistently outperforms other approaches in terms of accuracy across all scenarios. This aligns with its theoretical guarantee of producing exact Shapley values. However, as shown in Figure 8 it is computationally impractical for high-dimensional datasets, highlighting the need for efficient approximations. Among model-agnostic methods, *KernelSHAP* and its variants (*Parametric* and *Non-parametric*) demonstrate superior performance despite their $O(md^2)$ time complexity. This suggests that the weighted least squares approach effectively balances accuracy and efficiency, particularly for datasets with moderate dimensionality. *IME*, with its $O(md)$ time complexity, shows poor performance compared to other methods. This indicates that while random sampling offers computational efficiency, it may not capture complex feature interactions adequately, especially in high-dimensional spaces.

In the model-specific setting, *IME* emerges as the clear underperformer, while the methods tailored for specific models consistently rank among the top performers, just behind the gold-standard *Exhaustive Sampling* approach. The *Linear (independent)* approach demonstrates robust performance, providing accurate estimates across various datasets. In contrast, the *Linear (correlated)* approach struggles, likely due to its flawed assumption that every dataset follows a multivariate Gaussian distribution. On the other hand, both the *Tree (interventional)* and *Tree (path-dependent)* methods with $O(LD)$ complexity (where L is the number of leaves and D is the maximum tree depth) stand out, delivering performance nearly on par with *Exhaustive Sampling*.

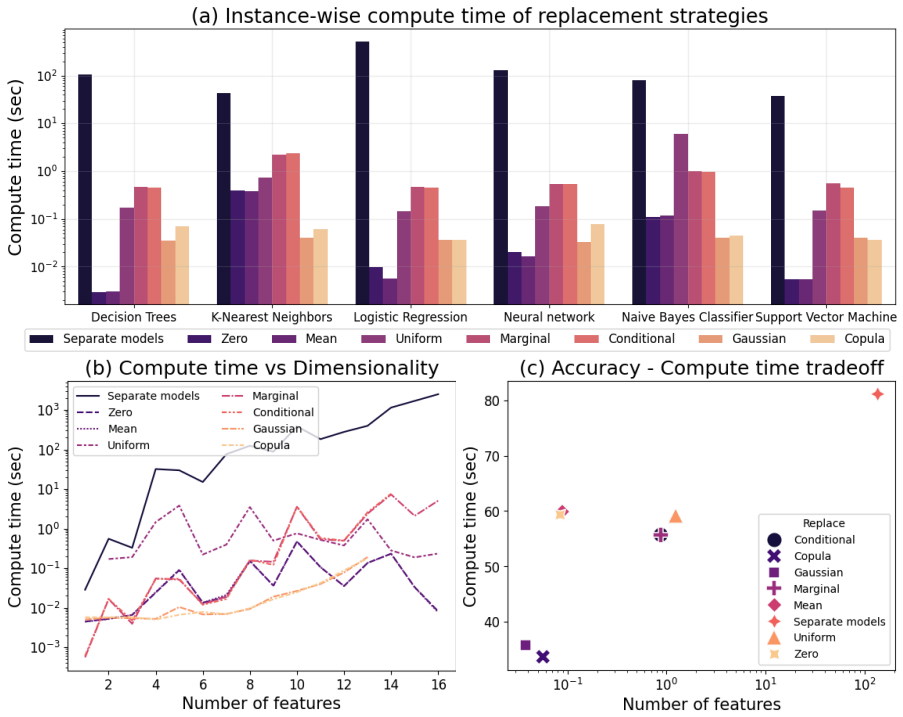


Fig. 7. (a) Compares instance-wise compute time of replacement strategies across Classification datasets. Every bar corresponds to a unique replacement strategy. (b) Demonstrates the impact of dimensionality on the estimation time of Shapley values. (c) Highlights the tradeoff between accuracy and computation time.

This suggests that exploiting the hierarchical structure of tree models allows for highly efficient and accurate Shapley value estimation. For deep learning models, *DeepSHAP* with $O(nr)$ complexity (where n is the number of neurons and r is the number of reference samples) outperforms *DeepLIFT* with $O(n)$ complexity. This indicates that the additional computational cost of *DeepSHAP* translates to meaningful improvements in explanation quality for neural networks. While model-specific methods demonstrate superior performance due to their alignment with the underlying model structure, this specialization limits their applicability across diverse architectures and increases implementation complexity.

As dataset dimensionality increases, methods with lower polynomial complexity in terms of the number of features (d) show better scalability. As depicted in Figure 10 *Exhaustive sampling* with $O(2^d)$ complexity becomes intractable for high-dimensional data. Sampling-based approaches like *IME* with $O(md)$ complexity offer a favorable trade-off between accuracy and scalability. Model-specific methods like *Linear (independent)* with $O(d)$ complexity and *Tree-based* methods with $O(LD)$ complexity (where L is a number of leaves and D is the maximum tree depth) scale well to higher dimensions. As dimensionality grows, model-specific methods leveraging the underlying structure of the models achieve a more favorable balance between accuracy and computational efficiency compared to model-agnostic approaches (Refer Figure 11). In practice, choosing a technique that achieves a balance between accuracy and scalability is crucial. We provide guidelines (Section 6) for careful selection of the approximation method based on usecase.

Table 4. Summary of accuracy evaluation for all Shapley value estimation techniques divided into different categories according to model type. The accuracy is computed using the R^2 test. (red represents mean, and blue represents median accuracy of each method over 200 datasets).

	Approximation	Rank	Accuracy - R^2 value
Model-agnostic	CES	**7**	
	Cohort	**6**	
	Exhaustive Sampling	**1**	
	IME	**9**	
	KernelSHAP	**4**	
	MLE	**8**	
	Non-parametric	**3**	
	Parametric	**2**	
	SGD-Shapley	**5**	
Linear model	CES	**9**	
	Cohort	**8**	
	Exhaustive Sampling	**1**	
	IME	**11**	
	KernelSHAP	**5**	
	MLE	**10**	
	Non-parametric	**4**	
	Parametric	**3**	
	SGD-Shapley	**7**	
	Linear(correlated)	**6**	
	Linear(independent)	**2**	
Tree-based models	CES	**9**	
	Cohort	**8**	
	Exhaustive Sampling	**1**	
	IME	**11**	
	KernelSHAP	**5**	
	MLE	**10**	
	Non-parametric	**7**	
	Parametric	**6**	
	SGD-Shapley	**4**	
	Tree (interventional)	**3**	
	Tree (path dependent)	**2**	
Neural Networks	CES	**10**	
	Cohort	**9**	
	Exhaustive Sampling	**1**	
	IME	**7**	
	KernelSHAP	**4**	
	MLE	**11**	
	Non-parametric	**6**	
	Parametric	**5**	
	SGD-Shapley	**8**	
	DeepLIFT	**12**	
	DeepSHAP	**2**	
	DASP	**3**	

The empirical results suggest that model-specific approaches tend to offer more reliable approximations. This is likely due to their ability to capture model-specific structures and interactions,

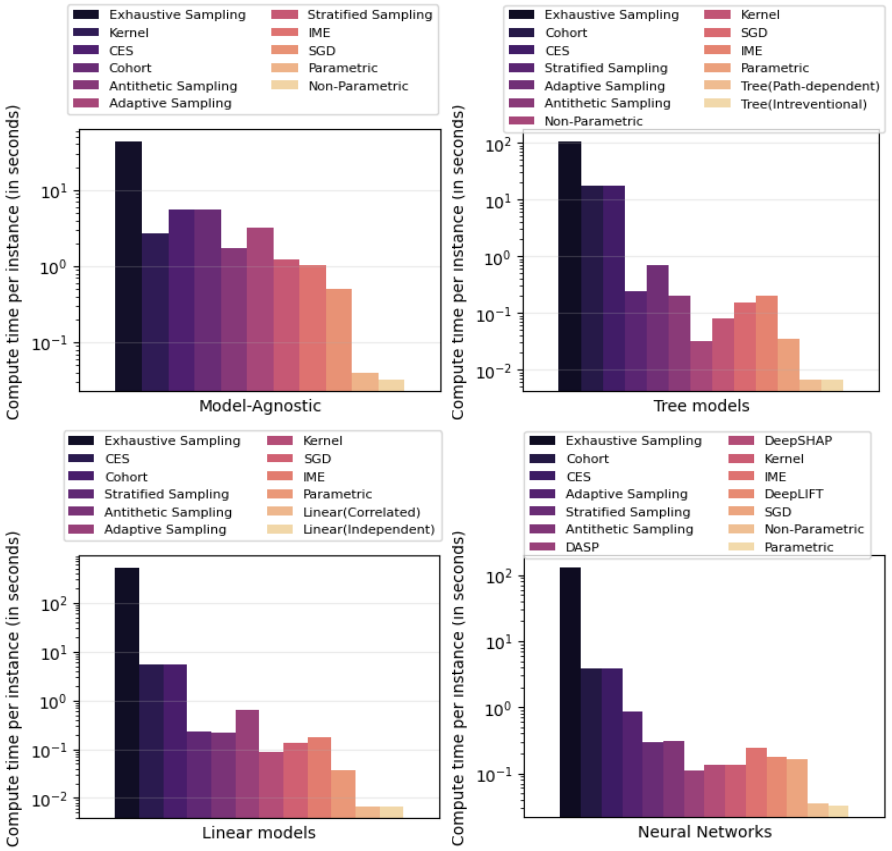


Fig. 8. Comparison of per instance computation time of different approximation strategies. The comparison is divided into 1 model-agnostic setting and 3 model-specific settings.

leading to more consistent and accurate Shapley value estimates. In conclusion, the empirical findings (Table 4) strongly align with the theoretical properties (Table 2), demonstrating a clear connection between the approximation methods' characteristics and their performance in practice. Model-specific approximations demonstrate significantly faster computation times, leveraging the inherent structure of the models. For model-agnostic scenarios, methods like *KernelSHAP* that employ sophisticated estimation strategies outperform more straightforward sampling-based approaches, especially as dataset complexity increases.

Additionally, the experiment in Figure 9 demonstrates the robustness of Shapley value approximation techniques under scenarios involving missing data. While all methods exhibit an expected increase in explanation error as the percentage of sampled subsets decreases, some techniques, such as IME and KernelSHAP, show greater resilience to missing data compared to others. This robustness can be attributed to their systematic sampling strategies, which help maintain the relative feature rankings even with reduced subsets. These findings highlight the importance of selecting techniques that balance accuracy and robustness, particularly in real-world scenarios where missing data is prevalent.

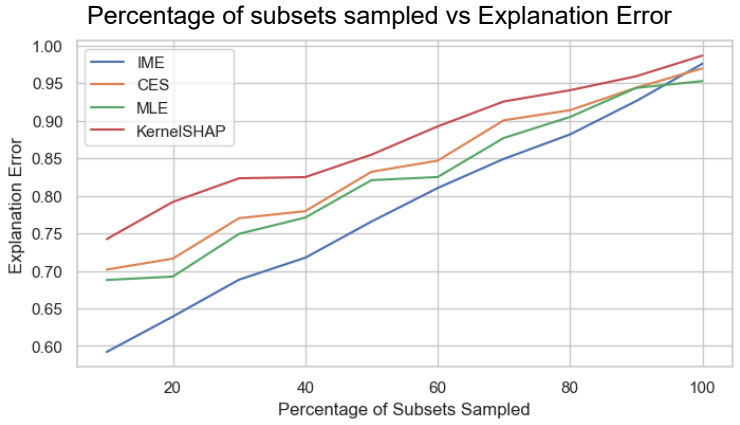


Fig. 9. Comparison of explanation errors for Shapley values approximated using the sampling method across an increasing percentage of subsets sampled.

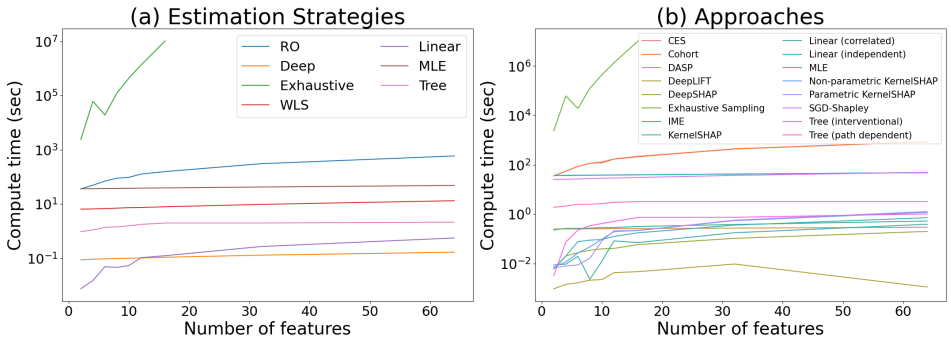


Fig. 10. Impact of increasing dimensionality on the per-instance computation time. (a) compares tractable estimation strategies, whereas (b) compares model-agnostic and model-specific Shapley value approximations.

5.2.2 ***Qualitative evaluation***. We use the admission dataset [91] to compare different approximation strategies qualitatively. The admission dataset provides application details of individual students, and the task is to predict the chances of the student receiving admission. We trained all the mentioned model types using the dataset and then generated explanations using each approximation technique. We have also included *FastSHAP* in the evaluation by training a surrogate model that best suits *FastSHAP*'s explanation process. Figure 12 demonstrates the Spearman rank correlation between different approximations across each model type.

The qualitative evaluation reveals a notable trend: approximation methods that rely on conditional distributions, including *CES*, *Shapley cohort refinement*, and *Parametric/Non-parametric KernelSHAP*, as well as *Exhaustive Sampling*, demonstrate a high degree of correlation in the Shapley values they produce. This finding implies that these methods are consistent in estimating the contributions of features to model predictions. The strong correlation among Shapley values obtained through these techniques suggests that they offer robust and reliable estimations of feature importance. Moreover, it emphasizes the efficacy of utilizing conditional distributions as a replacement strategy in Shapley value approximation.

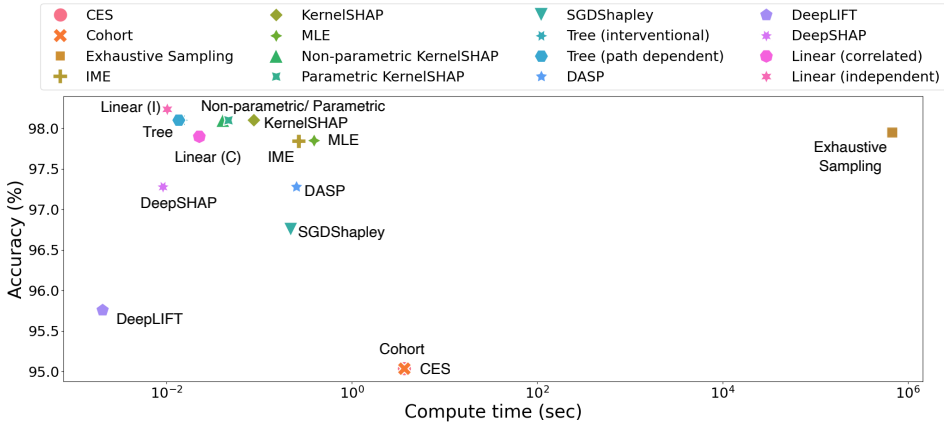


Fig. 11. Time-accuracy tradeoff comparison between distinct Shapley value estimation approaches.

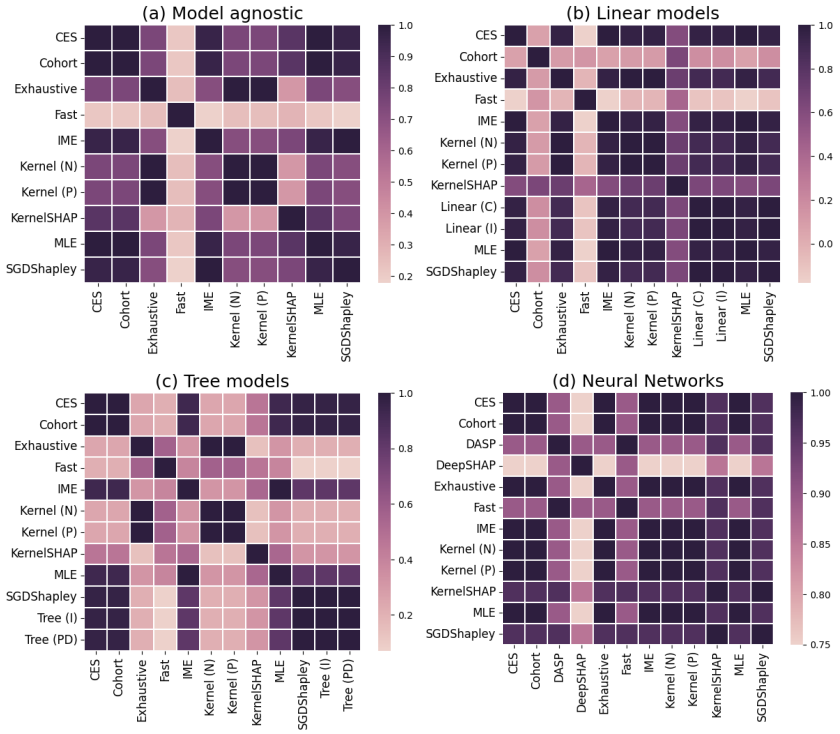


Fig. 12. Spearman rank correlation heatmap of an instance, comparing the quality of different approximations.

The performance of *FastSHAP* appears to be subpar, indicating that it may yield less accurate or reliable results compared to other Shapley value approximation methods. *FastSHAP* may require a highly tuned model to serve effectively instead of a surrogate model. In other words, to achieve satisfactory performance with *FastSHAP*, it may be necessary to meticulously optimize and fine-tune the underlying machine learning model used for approximation. This observation stresses the

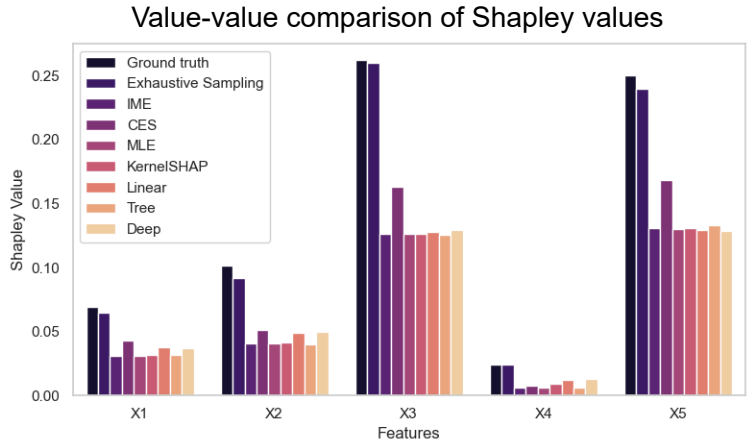


Fig. 13. Magnitude comparison of Shapley values. "X_is" are the features in the synthetic dataset. Features ranked by importance as $X_3 > X_5 > X_2 > X_1 > X_4$

significance of careful model selection and tuning when employing *FastSHAP* for feature importance analysis or interpretability tasks in machine learning applications.

To further evaluate the behavior of Shapley value approximations in a controlled setting, we conducted a synthetic dataset experiment as outlined in Section 4. This experiment validates our hypothesis in Section 4.1, demonstrating that while Shapley value approximations preserve relative feature rankings, their absolute magnitudes are reduced due to biases introduced by limited subset sampling and assumptions of feature independence. Among the methods evaluated, *Exhaustive Sampling* closely aligns with the ground-truth Shapley values, while other approximation techniques maintain consistent feature rankings despite differences in magnitude. As illustrated in Figure 13, features ranked by importance as $X_3 > X_5 > X_2 > X_1 > X_4$ retain their order across all approximations, even though their absolute Shapley values differ from the ground truth. This phenomenon highlights the limitations of direct value-to-value comparisons, as scaling effects obscure the true magnitude of feature contributions. Instead, focusing on relative feature rankings or reconstruction metrics, such as explanation error, offers a more robust and reliable evaluation framework by prioritizing consistent feature rankings and alignment with model predictions, which are essential for interpretability.

6 GUIDELINES

Selecting the appropriate strategy for Shapley value computation depends on the characteristics of the data and the specific use case. These guidelines help align your approach with the nature of your dataset, model complexity, and available computational resources.

- **General Case:** In typical scenarios, the *Marginal Distribution* effectively handles missing features. *WLS* balances accuracy and efficiency, while *KernelSHAP* offers robust model-agnostic explanations but is resource-intensive for high-dimensional datasets.
- **Complex Feature Interactions:** When dealing with complex feature relationships, the *Conditional Distribution* ensures accurate handling of missing features by capturing dependencies. We recommend *TreeSHAP* and *DeepSHAP*, since they offer high accuracy with reduced computational effort.

- **High Dimensional Data:** For high-dimensional datasets, replacing missing features with *Predetermined Baselines* is an efficient strategy to speed up computations. *TreeSHAP* and *DeepSHAP* provide optimized approximations for high-dimensional feature spaces without compromising on quality.
- **Exploratory Analysis and Limited Resources:** Future work should focus on model and dataset-specific techniques, leveraging strategies like stratified and adaptive sampling to enhance accuracy and efficiency. While preliminary findings indicate promising potential, their broader applicability remains unclear, stressing the importance of understanding the influence of dataset characteristics on the effectiveness of these approaches.
- **Evaluating Newer Methodologies:** Approximations should be evaluated using relative feature rankings and reconstruction metrics rather than direct value comparisons. These metrics offer a more reliable evaluation framework, as approximations typically preserve feature importance rankings despite scaling biases.

7 Conclusion

Through this paper, we presented a comprehensive study of various Shapley value approximations in a tabular data setting, shedding light on their strengths, limitations, and implications for interpretability in machine learning models. Our findings highlight that while Shapley value approximations often align with ground-truth values in terms of relative feature rankings, inherent biases from sampling methods and computational constraints can make value-to-value comparisons misleading. Instead, we emphasize the importance of alternate metrics, such as relative feature rankings and reconstruction metrics like explanation error, as robust and reliable evaluation methodologies. By using these metrics, we provide valuable insights into the effectiveness and applicability of different techniques across diverse datasets and model structures. Moreover, the observed correlation among specific approximation techniques highlights the potential of leveraging conditional distributions as a robust replacement strategy. Future research should focus on combining dataset-specific sampling techniques like stratified and adaptive sampling with model-specific approximations to enhance accuracy and computational efficiency. While our exploratory findings suggest potential in such combinations, their generalizability remains unclear, emphasizing the importance of analyzing dataset properties that influence the effectiveness of these approaches. This work lays a foundation for advancing interpretability and feature importance analysis, fostering innovation in developing robust and efficient techniques for machine learning models.

Acknowledgments

The authors thank anonymous reviewers whose comments greatly improved this manuscript. This research was supported in part by Cisco Systems and Meta.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2019. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *ArXiv* abs/1903.10464 (2019). https://api.semanticscholar.org/CorpusID:85497080

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467 [cs.DC]

[3] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. arXiv:1806.08049 [cs.LG]

[4] Marco Ancona, Cengiz Öztireli, and Markus Gross. 2019. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation. arXiv:1903.10992 [cs.LG]
[5] Mohini Bariya, Alexandra von Meier, John Paparrizos, and Michael J Franklin. 2021. k-ShapeStream: Probabilistic Streaming Clustering for Electric Grid Events. In *2021 IEEE Madrid PowerTech*. IEEE, 1–6.
[6] João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, and Pedro Bizarro. 2021. TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. ACM, 2565–2573. doi:10.1145/3447548.3467166
[7] Leopoldo Bertossi, Benny Kimelfeld, Ester Livshits, and Mikaël Monet. 2023. The Shapley Value in Database Management. *SIGMOD Rec.* 52, 2 (aug 2023), 6–17. doi:10.1145/3615952.3615954
[8] Meghyn Bienvenu, Diego Figueira, and Pierre Lafourcade. 2024. When is Shapley Value Computation a Matter of Counting? arXiv:2312.14529 [cs.DB] https://arxiv.org/abs/2312.14529
[9] Paul Boniol, Ashwin K Krishna, Marine Bruel, Qinghua Liu, Mingyi Huang, Themis Palpanas, Ruey S Tsay, Aaron Elmore, Michael J Franklin, and John Paparrizos. 2025. VUS: effective and efficient accuracy measures for time-series anomaly detection. *The VLDB Journal* 34, 3 (2025), 32.
[10] Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparrizos. 2024. Dive into Time-Series Anomaly Detection: A Decade Review. *arXiv preprint arXiv:2412.20512* (2024).
[11] Paul Boniol, John Paparrizos, Yuhao Kang, Themis Palpanas, Ruey S Tsay, Aaron J Elmore, and Michael J Franklin. 2022. Theseus: navigating the labyrinth of time-series anomaly detection. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3702–3705.
[12] Paul Boniol, John Paparrizos, and Themis Palpanas. 2023. New Trends in Time Series Anomaly Detection.. In *EDBT*. 847–850.
[13] Paul Boniol, John Paparrizos, and Themis Palpanas. 2024. An interactive dive into time-series anomaly detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5382–5386.
[14] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. 2021. SAND in action: subsequence anomaly detection for streams. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2867–2870.
[15] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. 2021. SAND: streaming subsequence anomaly detection. *Proceedings of the VLDB Endowment* 14, 10 (2021), 1717–1729.
[16] Paul Boniol, Emmanouil Sylligardos, John Paparrizos, Panos Trahanias, and Themis Palpanas. 2024. Adecimo: Model selection for time series anomaly detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5441–5444.
[17] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 5–32 (2001). doi:10.1023/A:1010933404324
[18] Mark A. Burgess and Archie C. Chapman. 2021. Approximating the Shapley Value Using Stratified Empirical Bernstein Sampling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 73–81. doi:10.24963/ijcai.2021/11 Main Track.
[19] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Comput. Oper. Res.* 36 (2009), 1726–1730. https://api.semanticscholar.org/CorpusID:42828306
[20] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* 36, 5 (2009), 1726–1730. doi:10.1016/j.cor.2008.04.004 Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).
[21] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2019. Explaining Image Classifiers by Counterfactual Generation. arXiv:1807.08024 [cs.CV]
[22] Abraham Charnes, Boaz Golany, Michael S. Keane, and John J. Rousseau. 1988. Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations. https://api.semanticscholar.org/CorpusID:123476789
[23] Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. 2022. Algorithms to estimate Shapley value feature attributions. arXiv:2207.07605 [cs.LG]
[24] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. True to the Model or True to the Data? arXiv:2006.16234 [cs.LG]
[25] Hugh Chen, Scott M. Lundberg, and Su-In Lee. 2022. Explaining a series of models by propagating Shapley values. *Nature Communications* 13, 1 (Aug. 2022). doi:10.1038/s41467-022-31384-3
[26] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. arXiv:1808.02610 [cs.LG]
[27] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. arXiv:1802.07814 [cs.LG]
[28] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM. doi:10.1145/2939672.

2939785
[29] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. 2018. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1244–1253.
[30] Zicun Cong, Xuan Luo, Pei Jian, Feida Zhu, and Yong Zhang. 2021. Data Pricing in Machine Learning Pipelines. arXiv:2108.07915 [cs.LG] https://arxiv.org/abs/2108.07915
[31] Ian Covert, Chanwoo Kim, and Su-In Lee. 2023. Learning to Estimate Shapley Values with Vision Transformers. arXiv:2206.05282 [cs.CV]
[32] Ian Covert and Su-In Lee. 2021. Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. arXiv:2012.01536 [cs.LG]
[33] Ian Covert, Scott Lundberg, and Su-In Lee. 2022. Explaining by Removing: A Unified Framework for Model Explanation. arXiv:2011.14878 [cs.LG]
[34] Nello Cristianini and Elisa Ricci. 2008. *Support Vector Machines*. Springer US, Boston, MA, 928–932. doi:10.1007/978-0-387-30162-4_415
[35] Pádraig Cunningham and Sarah Jane Delany. 2021. k-Nearest Neighbour Classifiers - A Tutorial. *Comput. Surveys* 54, 6 (July 2021), 1–25. doi:10.1145/3459665
[36] Piotr Dabkowski and Yarin Gal. 2017. Real Time Image Saliency for Black Box Classifiers. arXiv:1705.07857 [stat.ML]
[37] Daniel Deutch, Nave Frost, Benny Kimelfeld, and Mikaël Monet. 2022. Computing the Shapley Value of Facts in Query Answering. arXiv:2112.08874 [cs.DB] https://arxiv.org/abs/2112.08874
[38] Guozhu Dong and Jian Pei. 2007. *Sequence data mining*. Vol. 33. Springer Science & Business Media.
[39] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. doi:10.1126/sciadv.aao5580 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580
[40] Adam Dziedzic, John Paparrizos, Sanjay Krishnan, Aaron Elmore, and Michael Franklin. 2019. Band-limited training and inference for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 1745–1754.
[41] Jens E d'Hondt, Haojun Li, Fan Yang, Odysseas Papapetrou, and John Paparrizos. 2025. A Structured Study of Multivariate Time-Series Distance Measures. In *Proceedings of the ACM on Management of Data*, Vol. 3. Article 121. doi:10.1145/3725258
[42] Jens E d'Hondt, Odysseas Papapetrou, and John Paparrizos. 2024. Beyond the Dimensions: A Structured Evaluation of Multivariate Time Series Distance Measures. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 107–112.
[43] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. 2021. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, 6 (2021), 741–760.
[44] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. arXiv:1910.08485 [cs.CV]
[45] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE. doi:10.1109/iccv.2017.371
[46] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* 32 (1937), 675–701.
[47] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. 2021. Shapley explainability on the data manifold. arXiv:2006.01272 [cs.LG]
[48] Johannes Fürnkranz. 2010. *Decision Tree*. Springer US, Boston, MA, 263–267. doi:10.1007/978-0-387-30164-8_204
[49] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. arXiv:2103.11972 [cs.AI]
[50] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Feature attribution and recourse via probabilistic contrastive counterfactuals. In *Proceedings of the ICML Workshop on Algorithmic Recourse*. 1–6.
[51] Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International conference on social informatics*. Springer, 41–57.
[52] Simon Grah and Vincent Thouvenot. 2020. *A Projected Stochastic Gradient Algorithm for Estimating Shapley Value Applied in Attribute Importance*. 97–115. doi:10.1007/978-3-030-57321-8_6
[53] Suchit Gupte. [n. d.]. Shapley Values Eval GitHub. https://github.com/TheDatumOrg/ShapleyValuesEval.
[54] Suchit Gupte and John Paparrizos. 2025. ShapX Engine: A Demonstration of Shapley Value Approximations. In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25)*. ACM, Berlin, Germany, 4. doi:10.1145/3722212.3725135
[55] Isha Hameed, Samuel Sharpe, Daniel Barcklow, Justin Au-Yeung, Sahil Verma, Jocelyn Huang, Brian Barr, and C. Bayan Bruss. 2022. BASED-XAI: Breaking Ablation Studies Down for Explainable Artificial Intelligence. arXiv:2207.05566 [cs.LG]

[56] Simon Haykin. 1994. *Neural networks: a comprehensive foundation.* Prentice Hall PTR.
[57] Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. RankingSHAP – Listwise Feature Attribution Explanations for Ranking Models. arXiv:2403.16085 [cs.IR] https://arxiv.org/abs/2403.16085
[58] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. 2019. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery* 9 (2019). Issue 4. doi:10.1002/widm.1312
[59] Anthony Hunter and Sébastien Konieczny. 2010. On the measure of conflicts: Shapley Inconsistency Values. *Artificial Intelligence* 174, 14 (2010), 1007–1026. doi:10.1016/j.artint.2010.06.001
[60] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. 2019. Feature relevance quantification in explainable AI: A causal problem. arXiv:1910.13413 [stat.ML]
[61] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. 2021. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations. arXiv:2103.01890 [stat.ML]
[62] Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. 2022. FastSHAP: Real-Time Shapley Value Estimation. arXiv:2107.07436 [stat.ML]
[63] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gurel, Bo Li, Ce Zhang, Dawn Song, and Costas Spanos. 2023. Towards Efficient Data Valuation Based on the Shapley Value. arXiv:1902.10275 [cs.LG]
[64] Hao Jiang, Chunwei Liu, Qi Jin, John Paparrizos, and Aaron J Elmore. 2020. Pids: attribute decomposition for improved compression and query performance in columnar storage. *Proceedings of the VLDB Endowment* 13, 6 (2020), 925–938.
[65] Hao Jiang, Chunwei Liu, John Paparrizos, Andrew A Chien, Jihong Ma, and Aaron J Elmore. 2021. Good to the Last Bit: Data-Driven Encoding with CodecDB. In *Proceedings of the 2021 International Conference on Management of Data*. 843–856.
[66] Ahmet Kara, Dan Olteanu, and Dan Suciu. 2023. From Shapley Value to Model Counting and Back. arXiv:2306.14211 [cs.DB] https://arxiv.org/abs/2306.14211
[67] Pratik Karmakar, Mikaël Monet, Pierre Senellart, and Stéphane Bressan. 2024. Expected Shapley-Like Scores of Boolean Functions: Complexity and Applications to Probabilistic Databases. arXiv:2401.06493 [cs.DB] https://arxiv.org/abs/2401.06493
[68] Kamal Kasmaoui. 2019. *Linear Regression.* Springer International Publishing, Cham, 1–11. doi:10.1007/978-3-319-31816-5_478-1
[69] Sanjay Krishnan, Aaron J Elmore, Michael Franklin, John Paparrizos, Zechao Shang, Adam Dziedzic, and Rui Liu. 2019. Artificial intelligence in resource-constrained and shared environments. *ACM SIGOPS Operating Systems Review* 53, 1 (2019), 1–6.
[70] Aditya Lahiri, Kamran Alipour, Ehsan Adeli, and Babak Salimi. 2022. Combining Counterfactuals With Shapley Values To Explain Image Models. arXiv:2206.07087 [cs.LG] https://arxiv.org/abs/2206.07087
[71] Bengio Y. & Hinton G. LeCun, Y. 2015. Deep learning. *Nature* 521, 436–444 (2015). doi:10.1038/nature14539
[72] Jinyang Li, Yuval Moskovitch, and H. V. Jagadish. 2023. Detection of Groups with Biased Representation in Ranking. arXiv:2301.00719 [cs.LG] https://arxiv.org/abs/2301.00719
[73] Chunwei Liu, Hao Jiang, John Paparrizos, and Aaron J Elmore. 2021. Decomposed bounded floats for fast compression and queries. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2586–2598.
[74] Chunwei Liu, John Paparrizos, and Aaron J Elmore. 2024. AdaEdge: A Dynamic Compression Selection Framework for Resource Constrained Devices. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 1506–1519.
[75] Qinghua Liu, Paul Boniol, Themis Palpanas, and John Paparrizos. 2024. Time-Series Anomaly Detection: Overview and New Trends. *Proceedings of the VLDB Endowment (PVLDB)* 17, 12 (2024), 4229–4232.
[76] Qinghua Liu and John Paparrizos. 2024. The Elephant in the Room: Towards A Reliable Time-Series Anomaly Detection Benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems*.
[77] Shinan Liu, Tarun Mangla, Ted Shaowang, Jinjin Zhao, John Paparrizos, Sanjay Krishnan, and Nick Feamster. 2023. AMIR: Active Multimodal Interaction Recognition from Video and Network Traffic in Connected Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–26.
[78] Ester Livshits, Leopoldo Bertossi, Benny Kimelfeld, and Moshe Sebag. 2020. The Shapley Value of Tuples in Query Answering. In *23rd International Conference on Database Theory (ICDT 2020) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 155)*, Carsten Lutz and Jean Christoph Jung (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 20:1–20:19. doi:10.4230/LIPIcs.ICDT.2020.20
[79] Ester Livshits and Benny Kimelfeld. 2021. The Shapley Value of Inconsistency Measures for Functional Dependencies. In *24th International Conference on Database Theory (ICDT 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 186)*, Ke Yi and Zhewei Wei (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 15:1–15:19. doi:10.4230/LIPIcs.ICDT.2021.15

[80] Erion G. Chen H. et al. Lundberg, S.M. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2, 56–67 (2020). doi:10.1038/s42256-019-0138-9

[81] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[82] Xuan Luo, Jian Pei, Zicun Cong, and Cheng Xu. 2022. On shapley value in data assemblage under independent utility. *Proceedings of the VLDB Endowment* 15, 11 (July 2022), 2761–2773. doi:10.14778/3551793.3551829

[83] Xuan Luo, Jian Pei, Zicun Cong, and Cheng Xu. 2022. On shapley value in data assemblage under independent utility. *Proc. VLDB Endow.* 15, 11 (jul 2022), 2761–2773. doi:10.14778/3551793.3551829

[84] Xuan Luo, Jian Pei, Zicun Cong, and Cheng Xu. 2022. On shapley value in data assemblage under independent utility. *Proceedings of the VLDB Endowment* 15, 11 (July 2022), 2761–2773. doi:10.14778/3551793.3551829

[85] Xuan Luo, Jian Pei, Cheng Xu, Wenjie Zhang, and Jianliang Xu. 2024. Fast Shapley Value Computation in Data Assemblage Tasks as Cooperative Simple Games. *Proc. ACM Manag. Data* 2, 1, Article 56 (mar 2024), 28 pages. doi:10.1145/3639311

[86] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. 2014. Bounding the Estimation Error of Sampling-based Shapley Value Approximation. arXiv:1306.4265 [cs.GT] https://arxiv.org/abs/1306.4265

[87] Kolby Nottingham Markelle Kelly, Rachel Longjohn. [n. d.]. The UCI Machine Learning Repository. ([n. d.]).

[88] Masayoshi Mase, Art B. Owen, and Benjamin Seiler. 2020. Explaining black box decisions by Shapley cohort refinement. arXiv:1911.00467 [cs.LG]

[89] Kathy McKeown, Hal Daume III, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R Fleischmann, et al. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2684–2696.

[90] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. 2022. Sampling Permutations for Shapley Value Estimation. arXiv:2104.12199 [stat.ML] https://arxiv.org/abs/2104.12199

[91] Aneeta S Antony Mohan S Acharya, Asfia Armaan. 2019. A Comparison of Regression Models for Prediction of Graduate Admissions. *IEEE International Conference on Computational Intelligence in Data Science* (2019).

[92] Dov Monderer and Dov Samet. 2002. Variations on the shapley value. *Handbook of Game Theory With Economic Applications* 3 (2002), 2055–2076. https://api.semanticscholar.org/CorpusID:150532249

[93] Davide Napolitano and Luca Cagliero. 2024. BONES: a Benchmark fOr Neural Estimation of Shapley values. arXiv:2407.16482 [cs.LG] https://arxiv.org/abs/2407.16482

[94] Davide Napolitano, Luca Cagliero, et al. 2024. Evaluating the Reliability of Shapley Value Estimates: An Interval-Based Approach. In *Proceedings of 1st Human-Interpretable AI Workshop*. CEUR.

[95] Davide Napolitano, Lorenzo Vaiani, and Luca Cagliero. 2024. Efficient Neural Network-based Estimation of Interval Shapley Values. *IEEE Transactions on Knowledge and Data Engineering* (2024), 1–12. doi:10.1109/TKDE.2024.3420180

[96] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish. 2022. Responsible data integration: Next-generation challenges. In *Proceedings of the 2022 International Conference on Management of Data*. 2458–2464.

[97] Amin Nayebi, Sindhu Tipirneni, Chandan K Reddy, Brandon Foreman, and Vignesh Subbian. 2023. WindowSHAP: An Efficient Framework for Explaining Time-series Classifiers based on Shapley Values. arXiv:2211.06507 [cs.LG] https://arxiv.org/abs/2211.06507

[98] Peter Nemenyi. 1963. *Distribution-free Multiple Comparisons*. Ph. D. Dissertation. Princeton University.

[99] Abraham Neyman, Pradeep Dubey, and Roberth J. Weber. 1981. Value Theory without Efficiency. *Mathematics of Operations Research* 6 (1981), 122–128.

[100] Ramin Okhrati and Aldo Lipani. 2020. A Multilinear Sampling Algorithm to Estimate Shapley Values. arXiv:2010.12082 [cs.LG]

[101] C. Onyutha. 2020. From R-squared to coefficient of model accuracy for assessing "goodness-of-fits". *Geoscientific Model Development Discussions* 2020 (2020), 1–25. doi:10.5194/gmd-2020-51

[102] Guillermo Owen. 1972. Multilinear Extensions of Games. *Management Science* 18 (1972), 64–79. https://api.semanticscholar.org/CorpusID:122887906

[103] Ioannis Paparrizos. 2018. *Fast, scalable, and accurate algorithms for time-series analysis*. Ph. D. Dissertation. Columbia University.

[104] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. 2022. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment* 15, 11 (2022), 2774–2787.

[105] John Paparrizos, B Barla Cambazoglu, and Aristides Gionis. 2011. Machine Learned Job Recommendation. In *Proceedings of the fifth ACM Conference on Recommender Systems*. 325–328.

[106] John Paparrizos, Ikraduya Edian, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. 2022. Fast Adaptive Similarity Search through Variance-Aware Quantization. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*.

IEEE, 2969–2983.

[107] John Paparrizos and Michael J Franklin. 2019. Grail: efficient time-series representation learning. *Proceedings of the VLDB Endowment* 12, 11 (2019), 1762–1777.

[108] John Paparrizos and Luis Gravano. 2015. k-Shape: Efficient and Accurate Clustering of Time Series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 1855–1870.

[109] John Paparrizos and Luis Gravano. 2017. Fast and accurate time-series clustering. *ACM Transactions on Database Systems (TODS)* 42, 2 (2017), 1–49.

[110] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S Tsay, Themis Palpanas, and Michael J Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. *Proceedings of the VLDB Endowment* 15, 8 (2022), 1697–1711.

[111] John Paparrizos, Haojun Li, Fan Yang, Kaize Wu, Jens E d'Hondt, and Odysseas Papapetrou. 2024. A survey on time-series distance measures. *arXiv preprint arXiv:2412.20574* (2024).

[112] John Paparrizos, Chunwei Liu, Bruno Barbarioli, Johnny Hwang, Ikraduya Edian, Aaron J Elmore, Michael J Franklin, and Sanjay Krishnan. 2021. VergeDB: A Database for IoT Analytics on Edge Devices.. In *CIDR*.

[113] John Paparrizos, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. 2020. Debunking four long-standing misconceptions of time-series distance measures. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1887–1905.

[114] John Paparrizos, Chunwei Liu, Aaron J Elmore, and Michael J Franklin. 2023. Querying Time-Series Data: A Comprehensive Comparison of Distance Measures. *Data Engineering* (2023), 69.

[115] John Paparrizos and Sai Prasanna Teja Reddy. 2023. Odyssey: An Engine Enabling the Time-Series Clustering Journey. *Proceedings of the VLDB Endowment* 16, 12 (2023), 4066–4069.

[116] John Paparrizos, Ryen W White, and Eric Horvitz. 2016. Detecting devastating diseases in search logs. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 559–568.

[117] John Paparrizos, Ryen W White, and Eric Horvitz. 2016. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice* 12, 8 (2016), 737–744.

[118] John Paparrizos, Kaize Wu, Aaron Elmore, Christos Faloutsos, and Michael J Franklin. 2023. Accelerating Similarity Search for Elastic Measures: A Study and New Generalization of Lower Bounding Distances. *Proceedings of the VLDB Endowment* 16, 8 (2023), 2019–2032.

[119] John Paparrizos, Fan Yang, and Haojun Li. 2024. Bridging the gap: A decade review of time-series clustering methods. *arXiv preprint arXiv:2412.20582* (2024).

[120] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Identifying Biased Subgroups in Ranking and Classification. arXiv:2108.07450 [cs.LG] https://arxiv.org/abs/2108.07450

[121] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. Association for Computing Machinery, New York, NY, USA, 1400–1412. doi:10.1145/3448016.3457284

[122] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG]

[123] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine Learning in Python. arXiv:1201.0490 [cs.LG]

[124] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421 [cs.CV]

[125] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2022. Interpretable Data-Based Explanations for Fairness Debugging. arXiv:2112.09745 [cs.LG]

[126] Amir Hossein Akhavan Rahnama and Henrik Boström. 2019. A study of data and label shift in the LIME framework. arXiv:1910.14421 [stat.ML]

[127] Jacob Reiter. 2020. Developing an Interpretable Schizophrenia Deep Learning Classifier on fMRI and sMRI using a Patient-Centered DeepSHAP. https://api.semanticscholar.org/CorpusID:220050528

[128] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[129] Sudeepa Roy and Babak Salimi. 2023. and Explanations. *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures* 13759

(2023), 105.
[130] Sudeepa Roy and Babak Salimi. 2023. Causal inference in data analysis with applications to fairness and explanations. In *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures*. Springer, 105–131.
[131] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The Shapley Value in Machine Learning. arXiv:2202.05594 [cs.LG]
[132] Luis Ruiz, Federico Valenciano, and José Manuel Zarzuelo. 1998. The Family of Least Square Values for Transferable Utility Games. *Games and Economic Behavior* 24 (1998), 109–130. https://api.semanticscholar.org/CorpusID:120297656
[133] Patrick Schwab and Walter Karlen. 2019. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. arXiv:1910.12336 [cs.LG]
[134] Nima Shahbazi and Abolfazl Asudeh. 2022. Data-centric reliability evaluation of individual predictions. *CoRR, abs/2204.07682* (2022).
[135] Nima Shahbazi and Abolfazl Asudeh. 2024. Reliability evaluation of individual predictions: a data-centric approach. *The VLDB Journal* (2024), 1–28.
[136] Nima Shahbazi, Mahdi Erfanian, and Abolfazl Asudeh. 2024. Coverage-based Data-centric Approaches for Responsible and Trustworthy AI. *IEEE Data Eng. Bull.* 47, 1 (2024), 3–17.
[137] Lloyd S. Shapley. 1988. A Value for n-person Games. https://api.semanticscholar.org/CorpusID:153629957
[138] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. arXiv:1605.01713 [cs.LG]
[139] Michelle Si and Jian Pei. 2024. Counterfactual Explanation of Shapley Value in Data Coalitions. *Proceedings of the VLDB Endowment* 17, 11 (2024), 3332–3345.
[140] Schrittwieser J. Simonyan K. et al. Silver, D. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 354–359 (2017). doi:10.1038/nature24270
[141] Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. *J. Mach. Learn. Res.* 11 (mar 2010), 1–18.
[142] Erik Strumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41 (2014), 647–665. https://api.semanticscholar.org/CorpusID:2449098
[143] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. arXiv:1908.08474 [cs.AI]
[144] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, and Themis Palpanas. 2023. Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. *Proceedings of the VLDB Endowment* 16, 11 (2023), 3418–3432.
[145] Guanting Tang, James Bailey, Jian Pei, and Guozhu Dong. 2013. Mining multidimensional contextual outliers from categorical relational data. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*. 1–4.
[146] Sofiane Touati, Mohammed Said Radjef, and Lakhdar Sais. 2020. A Bayesian Monte Carlo method for computing the Shapley value: Application to weighted voting and bin packing games. *Comput. Oper. Res.* 125 (2020), 105094. https://api.semanticscholar.org/CorpusID:225321234
[147] Wu Y. Le Q.V. et al. Trinh, T.H. 2024. Solving olympiad geometry without human demonstrations. *Nature* 625 476-482 (2024). doi:10.1038/s41586-023-06747-5
[148] Lorenzo Vaiani, Davide Napolitano, and Luca Cagliero. 2023. Learning Confidence Intervals for Feature Importance: A Fast Shapley-based Approach. (03 2023).
[149] Vikramkumar, Vijaykumar B, and Trilochan. 2014. Bayes and Naive Bayes Classifier. arXiv:1404.0933 [cs.LG] https://arxiv.org/abs/1404.0933
[150] Rui Wang, Xiaoqian Wang, and David I. Inouye. 2021. Shapley Explanation Networks. arXiv:2104.02297 [cs.LG]
[151] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* (1945), 80–83.
[152] Haocheng Xia, Jinfei Liu, Jian Lou, Zhan Qin, Kui Ren, Yang Cao, and Li Xiong. 2023. Equitable Data Valuation Meets the Right to Be Forgotten in Model Markets. *Proc. VLDB Endow.* 16, 11 (jul 2023), 3349–3362. doi:10.14778/3611479.3611531
[153] Zhengzheng Xing, Jian Pei, Philip S Yu, and Ke Wang. 2011. Extracting interpretable features for early classification on time series. In *Proceedings of the 2011 SIAM international conference on data mining*. SIAM, 247–258.
[154] Xinyi Xu, Thanh Lam, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. 2024. Model shapley: equitable model valuation with black-box access. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1874, 30 pages.
[155] Fan Yang and John Paparrizos. 2025. SPARTAN: Data-Adaptive Symbolic Time-Series Approximation. In *Proceedings of the ACM on Management of Data*, Vol. 3. Article 220. doi:10.1145/3725357

[156] H. Peyton Young. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14 (1985), 65–72. https://api.semanticscholar.org/CorpusID:122758426

[157] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting with Contextual Attention. arXiv:1801.07892 [cs.CV]

[158] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. 2021. On Explainability of Graph Neural Networks via Subgraph Explorations. arXiv:2102.05152 [cs.LG] https://arxiv.org/abs/2102.05152

[159] Matthew D Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901 [cs.CV]

[160] He Zhang, Bang Wu, Xingliang Yuan, Shirui Pan, Hanghang Tong, and Jian Pei. 2024. Trustworthy graph neural networks: aspects, methods, and trends. *Proc. IEEE* (2024).

[161] Jiayao Zhang, Qiheng Sun, Jinfei Liu, Li Xiong, Jian Pei, and Kui Ren. 2023. Efficient Sampling Approaches to Shapley Value Approximation. *Proc. ACM Manag. Data* 1, 1, Article 48 (may 2023), 24 pages. doi:10.1145/3588728

[162] Jiayao Zhang, Haocheng Xia, Qiheng Sun, Jinfei Liu, Li Xiong, Jian Pei, and Kui Ren. 2023. Dynamic Shapley Value Computation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. 639–652. doi:10.1109/ICDE55515.2023.00055

[163] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2015. Object Detectors Emerge in Deep Scene CNNs. arXiv:1412.6856 [cs.CV]

[164] Jiongli Zhu, Romila Pradhan, Boris Glavic, and Babak Salimi. 2022. Generating interpretable data-based explanations for fairness debugging using gopher. In *Proceedings of the 2022 International Conference on Management of Data*. 2433–2436.

[165] E. Štrumbelj, I. Kononenko, and M. Robnik Šikonja. 2009. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* 68, 10 (2009), 886–904. doi:10.1016/j.datak.2009.01.004

Received October 2024; revised January 2025; accepted February 2025