The Model Explanation Journey: An Empirical Evaluation of Shapley Value Approximation Methods [E, A, & B]

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

ABSTRACT

Understanding the choices made by machine learning models holds significant importance in establishing trust in models' predictions, ultimately facilitating their practical application. The Shapley values have gained popularity as a reliable and theoretically robust approach to foster model interpretability. Shapley values quantify each feature's contribution to model predictions by considering all feature subsets, offering comprehensive insights into their impact. The inherent complexity of computing Shapley values as an NP-hard problem has spurred the development of numerous approximation techniques, leading to an increase in the number of choices in the literature. The abundance of options has created a substantial gap in determining the most appropriate approach for practical applications. Through this study, we seek to bridge this gap by comprehensively evaluating various Shapley value approximation methods. With a fusion of quantitative and qualitative analyses, we rigorously assess the performance and reliability of 17 distinct approximation algorithms across 100 datasets spanning different domains and six different model architectures. Our investigation unveils nuanced insights into the strengths and limitations of each technique. Our evaluation highlights that capturing all the feature interactions is paramount for ensuring accurate and granular model explanations. This study explores different dimensions of Shapley value estimations and ultimately lays the groundwork for developing more reliable and efficient techniques. By leveraging the strengths we identified in existing methods, we aim to motivate further research in model explanations using the Shapley values, fostering continued progress in explainable Artificial Intelligence.

PVLDB Reference Format:
Suchit Gupte and John Paparrizos. The Model Explanation Journey: An Empirical Evaluation of Shapley Value Approximation Methods [E, A, & B]. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/TheDatumOrg/ShapleyValuesEval.

1 INTRODUCTION

Over the recent decades, machine learning (ML) and artificial intelligence (AI) have witnessed significant advancements. The deployment of ML models to solve real-world problems has surged due to their ability to outperform humans in terms of efficiency. The application of ML models also extends to various life-critical domains, including healthcare [29] and criminal justice [20], where the decisions made by these models must be accurate and fair. One potential approach to instilling trust in the decisions made by these models is through model interpretability. Model interpretability involves comprehensively understanding a model's underlying decision-making process. However, models with complex architectures [5, 14, 34] pose a challenge to this approach. Unfortunately, such complex models are becoming more popular because they offer better accuracy than simpler ones, thereby hindering model interpretability. Given the increasing prevalence of complex models, focusing on solutions that enhance model interpretability is crucial.

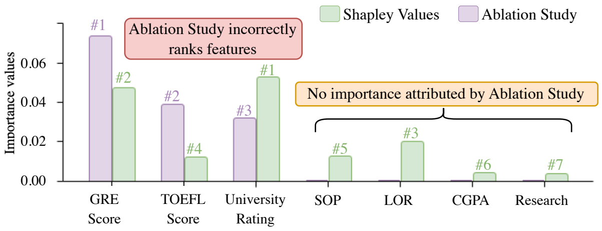


Figure 1: Comparison between Ablation study and Shapley values on graduate admissions dataset [41]. The Ablation study produces a distorted feature ranking due to some features lacking importance values, contrasted with the fine-grained explanations provided by Shapley values.

Model interpretability encompasses various facets, such as feature comprehension, model component analysis, and system understanding. However, fundamentally, each facet aims to explain the rationale behind each choice made within the model. For example, the Ablation study [27] is a popular, widely used solution that systematically removes individual aspects of the model to understand their importance in the model's decision-making process. As straightforward as the Ablation study may appear, its findings can occasionally be misleading. Thus, there is a pressing need within the research community for a more reliable alternative to ensure the accuracy of model interpretation. Shapley values [35, 36, 58], a Nobel Prize-winning concept that surfaced to address this demand, gained substantial recognition for its intricate explanations. 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.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

Table 1: Table depicting the intersection of replacement and estimation strategies in the literature, with checkmarks denoting the existence of methodologies combining each respective pair. Rows represent replacement strategies, and columns represent estimation strategies.

Strategies		**Estimation**				
		Exact	Semi Value	Random Order	Weighted Least Squares	Model specific
Replacement	Occlusion (All-zeros)	✓	-	-	-	✓
	Default (Mean)	✓	-	✓	-	-
	Marginal	✓	-	✓	✓	✓
	Uniform	✓	-	✓	✓	-
	Conditional	✓	✓	✓	✓	✓
	Generative models	-	-	✓	✓	-
	Surrogate models	-	✓	-	✓	-
	Gaussian	✓	-	✓	✓	✓
	Copula	✓	-	✓	✓	-
	Separate models	✓	-	-	-	-

The concept of Shapley values [58] originated in cooperative game theory. It 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. The process of estimating feature contributions using Shapley values is analogous to the Ablation study, 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. Thus, Shapley values enhance interpretability by capturing the nuanced interactions and dependencies between features, providing a detailed perspective on the contribution of each feature.

Utilizing Shapley values for model 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 [24, 31, 35, 36, 53, 61, 67] have been proposed to address this issue, such as imputing missing values or using a surrogate model to capture the behavior of the 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. Various estimation strategies [6, 32, 36, 45, 60] have emerged to efficiently approximate the Shapley values in polynomial time, effectively addressing this drawback. Table 1 highlights the existence of numerous Shapley value approximations resulting from the fusion of replacement and estimation strategies.

The abundance of such approximations has motivated the development of a standardized framework called SHapley Additive exPlanations (SHAP) [36]. While some of these approximations [10, 16, 35, 36, 60] are a part of the SHAP framework, others [1, 4, 15, 26, 32, 39, 45, 61, 63] continue to exist independently. Additionally, these approximations are divided into two categories: model-agnostic and model-specific solutions. The model-agnostic solutions [1, 6, 16, 24, 26, 36, 39, 45, 60, 61] are straightforward but stochastic. In contrast, the model-specific solutions [4, 10, 35, 36, 63] offer a significantly faster estimation of the Shapley values by leveraging model properties to mitigate the exponential complexity.

The abundance of approximations highlights the credibility of Shapley values as a reliable technique in model explanations. However, despite the considerable progress made over decades, a comprehensive evaluation of these approaches is notably absent in the existing literature. Our study represents the pioneering effort to fill this critical gap. We undertake an extensive evaluation encompassing diverse datasets and subject our findings to rigorous statistical analysis. We build up on the surveys [9, 17, 54] and extensively evaluate the approximations based on the Shapley values.

We break down the approximation of 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 estimation 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 100 datasets. This comprehensive evaluation enables us to thoroughly assess the performance and efficacy of different strategies and approximations in estimating Shapley values across varied data scenarios.

Our findings highlight the importance of capturing feature interactions to provide accurate model explanations. Our analysis supports the hypothesis that a replacement approach conditioned on the instance of interest is superior to other methods. Algorithms employing the Weighted Least Squares (WLS) method provide highly accurate approximations of Shapley values in a timely manner. It is worth noting that model-specific algorithms, attain similar levels of accuracy while being approximately ten times more efficient.

We first discuss the necessary background for the Shapley values (Sections 2.1 and 2.2). Then, we present our contributions as follows:

- We identify the computational drawbacks involved in the estimation of Shapley values, followed by reviewing the existing literature on overcoming these drawbacks (Section 2.3).
- We provide an overview of 17 distinct approximation algorithms to estimate the Shapley values (Section 3).
- We present evaluation measures tailored to assess various dimensions of the Shapley value approximations (Section 4).
- We conduct a comprehensive study on 100 datasets, examining the effectiveness of the replacement strategies in conjunction 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 potential future directions in developing more reliable Shapley value approximation techniques (Section 6).

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

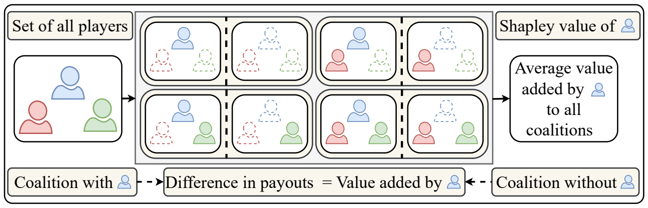


Figure 2: A simple illustration to demonstrate the estimation of the Shapley value for a player playing a three-player game. The Shapley value of the blue player is computed by averaging its marginal contribution across all possible coalitions.

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). Subsequently, we delve into the drawbacks of estimating Shapley values and solutions to overcome them (Section 2.3).

2.1 Shapley values in game theory

Shapley values [58, 65] 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 all the participating players. 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:

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

Figure 2 illustrates a straightforward estimation of the Shapley value for a player playing a three-player game. Despite the simplicity of the Shapley values as a solution, they are supported by robust theoretical properties [58, 65]. 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 $f(x) = w_1 x_1 + \cdots + w_d x_d$, ($w_i$: 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 [53], 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, 51]. 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 ML model f, an explicand, or the instance to be explained x^e, feature set D, and a coalition of the feature set $S \subseteq D$, then 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. In a manner 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.

2.3 Computational Complexity

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 knowledge of the payouts attained by each coalition 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

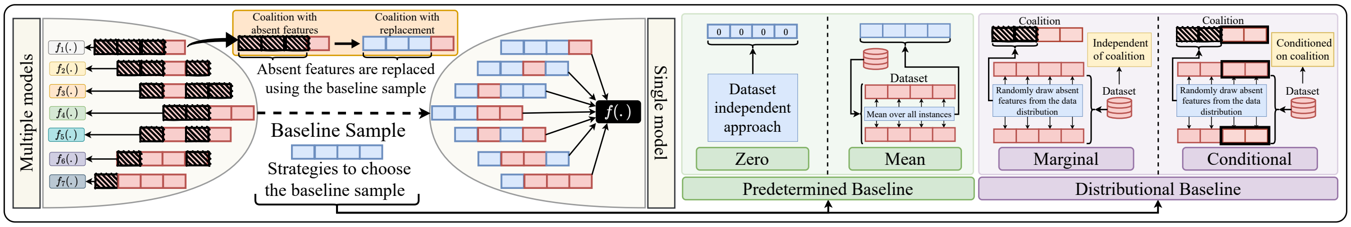


Figure 3: Replacement strategies such as Predetermined Baseline and Distributional Baseline address the absence of features, eliminating the necessity to train an exponential number of models and mitigating computational complexity. The Predetermined baseline imputes missing data with zeros or the mean, 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).

predictions, one must know the model prediction for every possible feature coalition. 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 $f(x_S^e) = f_S(x_S^e)$, where f_S is an extension of the original model trained only on features $\in S$.

Consequently, the estimation of Shapley values of all features demands training a separate model [70] 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.

2.3.1 ***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 3 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:

$$x_i = \begin{cases} x_i^e & \text{... if } i \in S \\ x_i^b & \text{... otherwise} \end{cases}$$

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 $f_S(x_S^e)$ as follows $f(x_S^e, x_{\bar{S}}^b)$. The most predominant choices for the predetermined baseline are the **all-zeros** [50, 57, 67] and the **default** [19, 53, 61] 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 [21, 22, 69] 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** [30, 36, 39, 53, 61]. 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. Consequently, this can result in inaccurate estimates of the Shapley values.

To address this concern, there exist several strategies: **the Parametric Assumption** [24] assumes that the data follows either a *Gaussian* or a *Gaussian-copula* distribution; **the Generative model** [7, 24, 31, 66] trains a deep learning model to predict missing feature values by comprehensively learning all the conditional data distributions; **the Surrogate model** [13, 24, 31] trains a deep learning model to predict the absent features using the target label of the explicand. All these strategies help accurately utilize 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.

2.3.2 ***Tractable estimation strategies:*** Initially, the random sampling approach emerged as the most intuitive estimation technique to address the inherent exponential complexity of computing Shapley values [6, 60]. Instead of exhaustively analyzing every possible combination of features, which becomes impractical as the dimensionality increases, sampling techniques randomly select a subset of combinations for approximation. This approach significantly reduces computational burden while offering a quality of model explanations comparable to the exhaustive estimation. However, random sampling introduces some variability in the estimates of Shapley values. Employing effective sampling techniques such as uniform [55], adaptive [40], and stratified [11, 68], help reduce the variance. Based on the stratified sampling, a Dynamic estimation [37] of Shapley values exists that approximates Shapley values accurately and efficiently. However, variance reduction techniques represent an additional dimension of estimating Shapley values, explicitly for random sampling of combinations. It is a complementary problem to ours; hence, we plan to explore this in the future.

Apart from the random sampling approach, various tractable estimation strategies offer a pragmatic solution by approximating Shapley values in polynomial time. These estimation strategies include multilinear sampling [45], modeling the Shapley value estimation as an optimization problem [16, 36], and a few model-specific solutions [4, 10, 35]. These tractable estimation strategies along with the replacement strategies (Section 2.3.1) form a foundation for the various approximations of estimating the Shapley values. The subsequent section offers a summary of these approximations.

3 SHAPLEY VALUE APPROXIMATIONS

There are several approximations 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 category, followed by a comprehensive list of the approaches falling under each category in Table 2.

Table 2: A detailed list of model-agnostic and model-specific approximations, classified based on estimation and replacement strategies. The approximations stated above form an essential component of our evaluation. "M" denotes replacement via Marginal distribution, while "C" represents Conditional distribution. The "Language" column signifies the implementation language employed for each approximation.

Approaches		Estimation	Replacement	Language
Model-agnostic	Exhaustive sampling	Exact	Separate models	*Python*
	IME [60]	RO	Empirical (M)	*Python*
	CES [61]	RO	Empirical (C)	*Python*
	Cohort [39]	RO	Empirical (C)	*Python*
	MLE [45]	MLE	Empirical (M)	*Python*
	Kernel [16]	WLS	Empirical (M)	*Python*
	SGD-Shapley [26]	WLS	Mean	*Python*
	Parametric [24]	WLS	Gaussian/Copula	*Python/R*
	Non-Parametric [24]	WLS	Empirical (C)	*Python/R*
	FastSHAP [32]	WLS	Surrogate model	*Python*
Model-specific	Linear [10]	Linear	Empirical (M)	*Python*
	Correlated Linear [10]	Linear	Gaussian	*Python*
	Tree interventional [35]	Tree	Empirical (M)	*Python*
	Tree path-dependent [35]	Tree	Empirical (C)	*Python/C++*
	DeepLIFT [52]	Deep	All-zeros	*Python*
	DeepSHAP [11]	Deep	Empirical (M)	*Python*
	DASP [4]	Deep	Mean	*Python*

3.1 Model-agnostic approximations

3.1.1 ***Semi Value (SV):*** The original coalition game, defined using Shapley values, is known as the *SemiValue* [44] 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.* [6] 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.

Local Shapley (L- Shapley) and **Connected Shapley (C- Shapley)** [12] 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

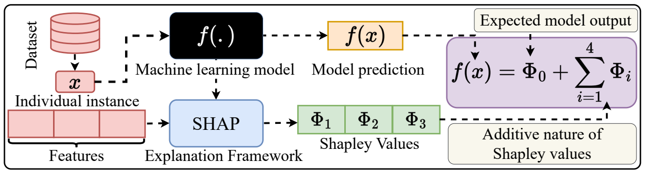


Figure 4: Shapley values, an additive local feature attribution technique central to Weighted Least Squares (WLS) estimation strategy. In this illustration, the instance is composed of 3 features. An individual model prediction is expressed as a sum of the average model output and the Shapley values.

of features instead of feature subsets. Consequently, the modified solution can be formulated in the following manner:

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

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 [42, 58]. Various approaches, including **IME (Interactions-based Method for Explanation) [60]**, **CES (Conditional Expectations Shapley) [61]**, **Shapley Cohort refinement [39]**, and **Generative/Surrogate models [24]** utilize this technique in combination with one of the replacement strategies mentioned in Section 2.3.1.

3.1.3 *Multilinear Extension (MLE):*

Owen introduced a *multilinear extension* [47] 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 [47] 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) [45]** for estimating the Shapley values.

3.1.4 *Weighted Least Squares (WLS):*

In Figure 4, it has been demonstrated that the Shapley values possess an additive property. Consequently, we can represent the model's prediction for a particular instance 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)* [8, 56] 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 [16, 36]** 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 [26]** is an alternative method that extends the principles of *KernelSHAP*. However, it employs projected gradient descent to solve the least squares problem approximately. **FastSHAP [15, 32]** 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 have a significantly high interpretability. As discussed in Section 2.2, there is a linear relationship between the input features and the model prediction, which 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 [36, 62]** and **Correlated LinearSHAP [10]** 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 2.3.1). 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 [25], ensemble learning like random forests [5], and gradient boosting models like XGBoost [14]. These non-linear models are affected by the interdependencies among the input features. **Interventional TreeSHAP [35]** and **Path-dependent TreeSHAP [35]** 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 2.3.1). In contrast, the *Path-dependent TreeSHAP* method adopts a conditional feature removal approach derived from the *Shapley Cohort refinement* approximation [39].

*3.2.3 **Deep learning models:***

Deep neural networks [34] 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 [11, 59]. 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 [36] introduced an extension of *DeepLIFT* [52] called the **DeepSHAP** to produce biased estimates of marginal Shapley values.

Despite its bias, *DeepSHAP* is valuable 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 and the evaluation measures that are a part of our evaluation.

Platform: We run our experiments on a server with the following configuration: AMD EPYC 7713 64-Core. 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) [48], TensorFlow(2.6.0) [2], scikit-learn(0.22.1) [49]. We execute every Shapley value estimation technique on a single core to guarantee accurate assessments of runtime. For experiment reproducibility, we open source the datasets and the code[1].

Datasets: For the scope of the study, we focus on regression-based tabular datasets. We utilize a total of 100 publicly available datasets from the UCI Machine Learning Repository [38]. Within the datasets, there are as many as 60 input features, and the number of instances ranges from 100 to 1 million. Figure 5 illustrates the statistical characteristics of the 100 datasets, explicitly focusing on their dimensions and scale. We split each dataset into training and testing sets to facilitate supervised machine learning. We employ the training split to train the ML models and the test split to generate explanations using the Shapley values. Since the Shapley values are a local feature attribution technique, the number of instances in the dataset has a very insignificant impact on generating explanations. However, the data's dimensionality significantly influences the estimation of Shapley values, as discussed in Section 2.3.

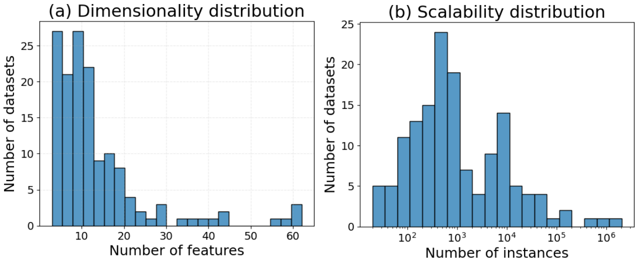


Figure 5: (a) represents the dimensionality distribution and (b) represents the scalability distribution across 100 regression datasets from the UCI ML repository[38].

ML Models: We broadly classify the supervised machine learning models used to tackle regression-based problems into 5 categories - Linear models [25, 33], Ensemble Learning [5], Gradient Boosting [14], Neural Networks [28], and Support Vector Machines [18]. To conduct a thorough evaluation, we integrate models representing each category. Shapley values intend to explain a black box model by leveraging the model itself, thereby negating the significance of the model's fit quality. Consequently, this allows us to use vanilla versions of each model with default hyperparameters.

4.1 Evaluation measures

We bifurcate our analysis into two sub-analyses. The initial segment focuses on evaluating various replacement strategies alongside an exhaustive estimation of the Shapley values. The subsequent segment is dedicated to evaluating different approximations of the Shapley values. We thoroughly perform a quantitative and qualitative evaluation of the strategies. To thoroughly analyze algorithm performance across different dimensions, we employ statistical methods. Specifically, we use the Wilcoxon test [64] for pairwise comparisons across datasets and the Friedman test [23], followed by the posthoc Nemenyi test [43] to rank Shapley value approximations and replacement strategies across multiple datasets. With the above-mentioned structure of the analysis in mind, we elaborate on the evaluation measures employed in conducting the experiments.

*4.1.1 **Explanation Error:***

As we implement replacement strategies to address missing features in Shapley value estimation, the absence of ground truth Shapley values presents a clear obstacle in the evaluation. Consequently, we must employ an alternative evaluation metric to assess the accuracy of the approaches, such as the explanation error [36]. The motivation for explanation error stems from the additive nature of the Shapley values, as illustrated in Figure 4. Shapley values indicate the individual contributions of input features towards shifting the model output from the average model prediction to the actual prediction value given a specific instance. When dealing with a black-box model f and an explicand x^e, the prediction for the explicand can be articulated as follows:

[1] https://github.com/TheDatumOrg/ShapleyValuesEval

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

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 these Φ_is. We can determine the quality of any approximation by measuring the discrepancy between the actual model prediction and the sum of the average model prediction(Φ_0) and the Shapley value approximations($\Phi_i s$). A smaller disparity signifies a higher level of accuracy in the approximation.

We use the R^2 test [46] to analyze the explanation error. The R^2 test [46], or the coefficient of determination, 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 discussed earlier, we divide our experimental analysis into two sections. Section 5.1 deals with the evaluation of different replacement strategies. Section 5.2 focuses on assessing the different approximation strategies, which are a combination of the replacement strategies and the estimation techniques as mentioned in Table 2.

5.1 Analysis of Replacement Strategies

We begin with conducting a comprehensive evaluation of the various replacement strategies (refer to Section 2.3.1). We deploy the replacement strategies against an exhaustive estimation of the Shapley values that encompasses all the potential feature coalitions. Through this evaluation, we aim to understand which replacement strategy efficiently gives accurate Shapley value estimates. Thus, we break down the quality of a replacement strategy into accuracy and efficiency. We evaluate the accuracy using the Explanation Error metric (refer Section 4.1.1) and the efficiency using the Computation Time metric (refer 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.

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 [50, 57, 67]
	Default	Mean [19, 53]
Distributional	Marginal	Empirical[30]
		Uniform distribution [30]
	Conditional	Empirical [39, 61]
		Separate models
		Parametric: Gaussian [24]
		Parametric: Copula [24]

5.1.1 ***Accuracy:*** The performance analysis reveals that the *All-zeros* baseline exhibits the poorest accuracy, while replacement through *Separate models* emerges as the most effective strategy, confirming our initial hypothesis. The *All-zeros* baseline simply substitutes absent features with a neutral value, completely disregarding the present features, and the underlying data distribution, leading to inaccurate Shapley value estimates. Conversely, the *Default (Mean)* replacement strategy demonstrates superior performance as it adheres to the inherent data distribution when replacing values with the mean. The *Separate models* replacement strategy, which employs distinct models for each coalition, aligns with the assumption underlying the Shapley value definition as mentioned in Section 2.2. This approach yields the most precise Shapley value estimates by considering feature dependencies within each coalition. Other distributional replacement strategies show minimal differences in accuracy, with *conditional distribution* strategies slightly outperforming the *marginal distribution* strategies, which is consistent with our expectations. *Conditional distributions* play a crucial role in capturing feature dependencies by conditioning the absent features on the present ones, leading to improved accuracy in estimating Shapley values compared to using the *Marginal distribution.* The critical diagram in Figure 6 assesses significant results by taking the average of the ranks of each method on all 100 datasets. Figure 6 depicts model-agnostic accuracy rankings of the various replacement strategies, whereas the boxplots show a breakdown of accuracies specific to each model.

5.1.2 ***Computational Efficiency:*** The findings depicted in Figure 7 provide valuable insights into the computational time required for estimating Shapley values per instance. Our initial hypothesis suggested that the Separate models replacement strategy would incur the longest computational time. At the same time, the Predetermined baselines (All-zeros and mean) would offer the quickest estimation, given their predefined nature prior to implementation. The results presented in Figure 7 support the hypothesis, indicating that Separate models consistently demand the highest computational overhead, while the Predetermined baseline method consistently yields the fastest estimates. Notably, the Marginal distribution approach consistently outperforms the Conditional distribution method regarding computational efficiency, highlighting its potential for faster estimation. Evaluating the trade-off between computational efficiency and accuracy reveals that replacement strategies employed using conditional distributions yield more robust Shapley value estimates. This analysis highlights the pivotal role of feature interaction effects in comprehending model explanations. Hence, when aiming for precise estimations within tight time constraints, there's a strong rationale supporting the broader implementation of conditional distributions.

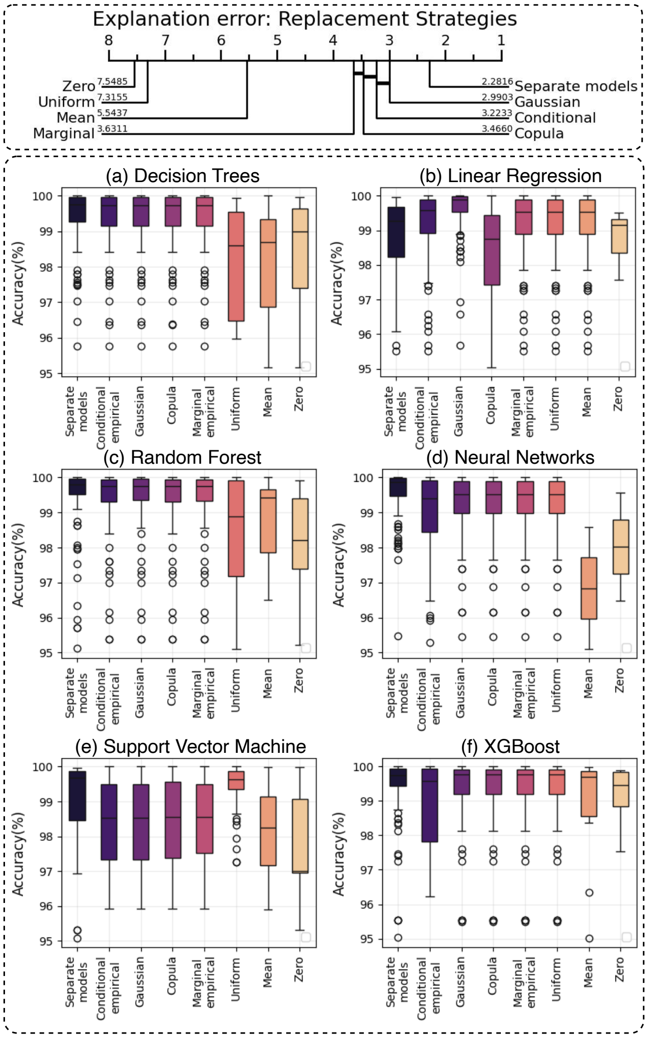


Figure 6: Explanation error of different replacement strategies. The accuracy of each strategy is computed using the R-squared test. The critical diagrams give model-agnostic and model-specific performance rankings of each replacement strategy. The boxplot visualizes these rankings across 100 datasets, offering insights into the variance of accuracies.

5.2 Analysis of approximations

In this section, we conduct an evaluation of both model-agnostic and model-specific approximations. Each approximation comprises a combination of a replacement strategy and a particular estimation strategy. For a comprehensive breakdown of all the approximations and their corresponding replacement and estimation strategies, please refer to Table 2. We divide the evaluation into two subsections. Section 5.2.1 focuses on the quantitative evaluation, whereas Section 5.2.2 addresses qualitative evaluation. *FastSHAP* is excluded from the quantitative evaluation due to its dependency on the quality of the surrogate model. Explicitly curating 100 surrogate models for each dataset is impractical, rendering the inclusion of *FastSHAP* infeasible for this evaluation. Since *FastSHAP*'s accuracy is intricately linked to the quality of the surrogate model, its performance cannot be fairly assessed within the scope of this study.

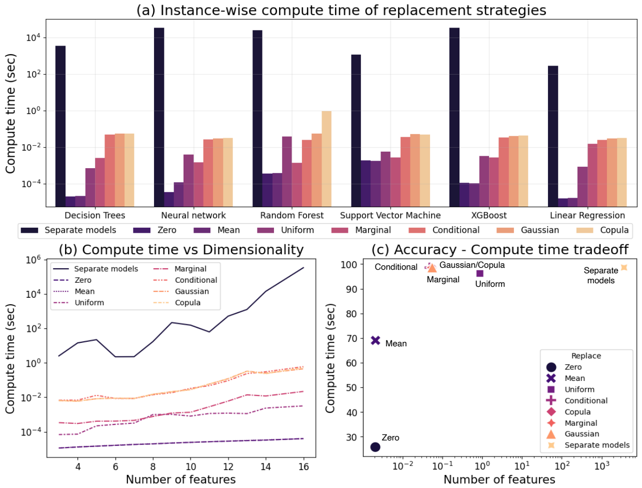


Figure 7: (a) compares the instance-wise compute time of different replacement strategies across 100 datasets. Every colored bar corresponds to a unique replacement strategy. (b) demonstrates the impact of increasing dimensionality on the estimation time of Shapley values. (c) highlights the tradeoff between accuracy and computation time.

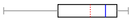
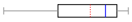
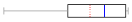
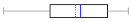
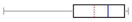
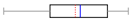
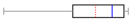
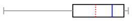
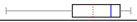
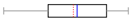
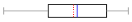
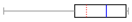
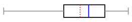
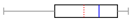
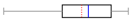
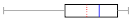
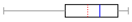
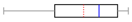
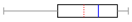
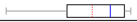
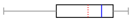
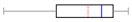
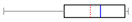
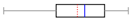
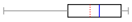
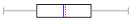
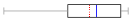
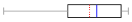
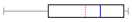
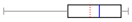
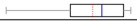
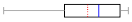
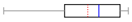
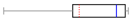
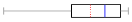
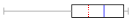
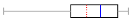
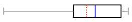
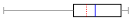
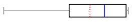
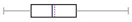
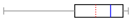
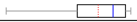
5.2.1 *Quantitative evaluation:*

We conduct a quantitative evaluation of all the approximation techniques across the 100 datasets outlined in Section 4. This evaluation includes the *Exhaustive Sampling* approach, which considers all potential feature coalitions. Despite its significant time inefficiency, particularly when combined with Separate Models, the Exhaustive Sampling approach consistently delivers the most accurate Shapley value estimates, as highlighted in Section 5.1. Table 4 illustrates the accuracy of each approximation technique. The table gives the accuracy rankings for each of the following cases: *Model-agnostic*, *Linear models*, *Tree-based models*, and *Neural networks*.

The *Exhaustive Sampling* technique consistently outperformed other approaches in every scenario. Its comprehensive consideration of all possible feature combinations inherently ensures the most accurate Shapley value estimates. This finding highlights the effectiveness of *Exhaustive Sampling* despite its time-intensive nature. In the model-agnostic context, the poor performance of the *IME* method highlights the inadequacy of random sampling coupled with a marginal distribution. Interestingly, despite the *marginal* replacement's inferiority to *conditional* replacement, *CES* and *Shapley Cohort refinement* exhibit similarly unsatisfactory results as *IME*. On the contrary, *KernelSHAP* and its *Parametric* and *Non-parametric* variations demonstrate superior performance, suggesting that employing a *weighted least squares* solution is more effective than relying solely on random coalition sampling.

Regarding model-specific cases, *IME* emerges as the poorest-performing method, while the approximations explicitly tailored for specific model types consistently rank among the best-performing approximations after the *Exhaustive Sampling* approach. The *Linear*

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 100 datasets).

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

(independent) approach demonstrates robust performance, providing accurate estimates across various datasets. However, the *Linear (correlated)* approach performs comparatively weakly. The poor performance of the *Linear (correlated)* approach might stem from an incorrect underlying assumption that every dataset follows a multivariate Gaussian distribution. On the other hand, both *Tree (interventional)* and *Tree (path-dependent)* approximations exhibit exceptional performance, nearly on par with the Exhaustive Sampling approach. Additionally, *DeepSHAP* and *DASP* approximations also stand out for their ability to produce accurate Shapley value estimates, further emphasizing the effectiveness of these approaches in capturing nuanced relationships within the model.

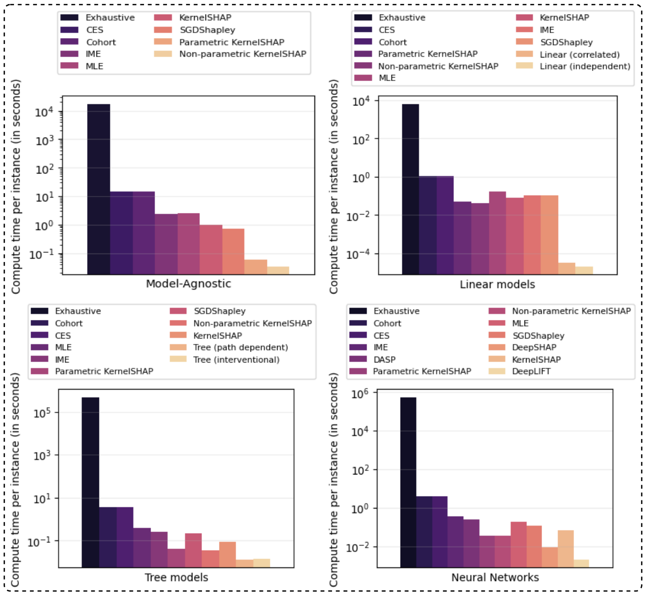


Figure 8: Comparison of per instance computation time of different approximation strategies. The comparison is divided into 1 model-agnostic setting and 3 model-specific settings.

Following the model-specific approximations, the weighted least square approaches, including the *Parametric*, *Non-parametric*, and vanilla versions of *KernelSHAP*, consistently yield superior accuracies compared to random sampling approaches. Hence, employing advanced techniques beyond simple random sampling methods is crucial. Moreover, the choice of replacement strategies significantly influences the robustness of the Shapley value estimates. For instance, *DeepLIFT*, which utilizes the *All-zeros* baseline, notably performs poorly in obtaining accurate estimates. Conversely, employing *Parametric* or *Non-parametric* versions of *KernelSHAP* enhances the accuracy of the vanilla *KernelSHAP* approximation, further emphasizing the significance of thoughtful selection of replacement strategies in Shapley value estimation.

Methods providing accurate estimates of Shapley values also exhibit decent efficiency. As depicted in Figure 8, all model-specific approximations demonstrate significantly faster computation times, leveraging the inherent structure of the model to their advantage. When comparing model-agnostic approximations, the *parametric* and *non-parametric* versions of *KernelSHAP* emerge as the fastest options. However, it is worth noting that *approximations* utilizing a conditional distribution with an empirical variant, such as *CES* and *Shapley Cohort refinement*, are less computationally efficient compared to those employing a *marginal distribution.* Figure 9 further illustrates the tradeoff between accuracy and compute time, providing valuable insights into the performance dynamics of various approximation methods.

5.2.2 *Qualitative evaluation.*

We use the admission dataset [41] to establish a qualitative comparison of different approximation strategies. The admission dataset provides application details of

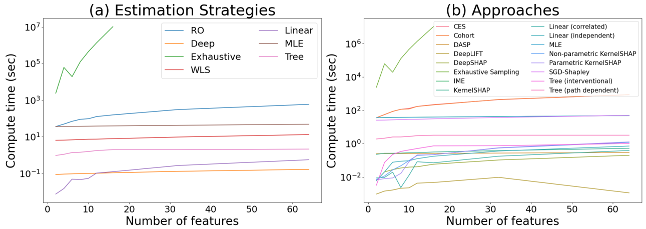


Figure 9: Impact of increasing dimensionality on the per-instance computation time. (a) compares tractable estimation strategies, whereas (b) compares specific approaches to estimating Shapley values.

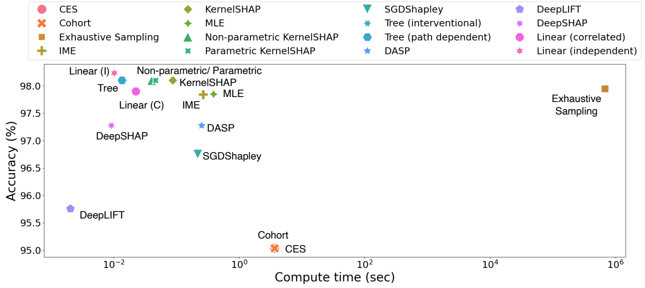


Figure 10: Time-accuracy tradeoff comparison between distinct Shapley value estimation approaches.

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 11 demonstrates the Spearman rank correlation between different approximations across each model type. Figure 12 illustrates the Shapley value distributions of each feature across the entire dataset.

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.

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

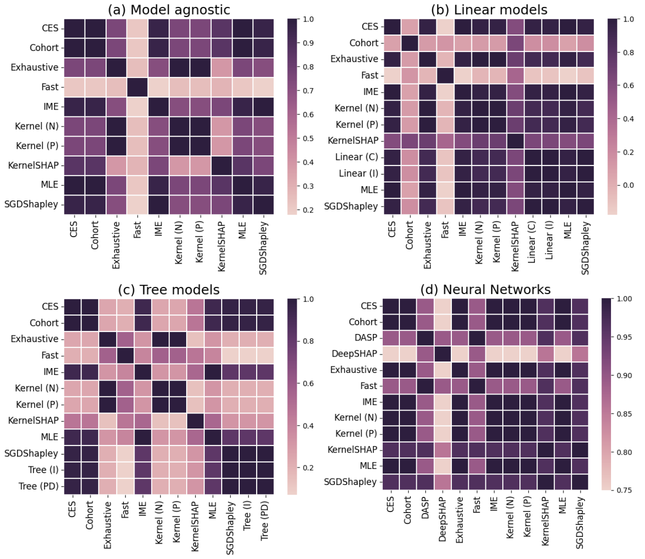


Figure 11: Spearman rank correlation heatmap of a singular instance, comparing the quality of different approximations.

6 DISCUSSION AND FUTURE RESEARCH

The Shapley values are a powerful tool for explaining machine learning models supported by robust properties. However, it is an NP-hard solution. Various sampling approaches attempt to approximate the Shapley values in polynomial time (see Section 3). Despite accounting for every possible feature coalition in the exhaustive estimation of the Shapley values, incorporating replacement strategies introduces a certain level of variability. This variability inherent in the process prevents us from obtaining ground truth Shapley values. However, a lack of ground truth Shapley values makes choosing the most appropriate approximation algorithm difficult.

Like the Shapley values, the Ablation study is another predominant model explanation approach. The core concept of the Ablation Study involves systematically perturbing different aspects of the model or its training process to understand the importance of each aspect in the model's decision-making process. There are various techniques in the Ablation Study, such as Feature Ablation, Architecture Ablation, Data Ablation, Loss Function Ablation, and Evaluation Metric Ablation. Specifically, feature ablation using a leave-one-out technique is a meticulous method to evaluate the influence of individual features in a machine-learning model. This process entails systematically eliminating one feature at a time from the dataset, training the model on the modified dataset, and assessing its performance. Through iterating this procedure for each feature and contrasting the model's performance with and without each feature, researchers can determine the relative significance of each feature in the model's decision-making mechanism.

Intuitively, the above-mentioned feature ablation technique aligns with the notion of the Shapley values. Estimating the Shapley values involves assigning a weight to each feature coalition. A weight function is formulated based on the size of the feature coalition, which signifies the total number of present features as illustrated in Equation 2. This weight function exhibits an inverted bell curve when graphed, as depicted in Figure 13. This inverted bell curve implies that imbalanced coalitions carry more weight than evenly

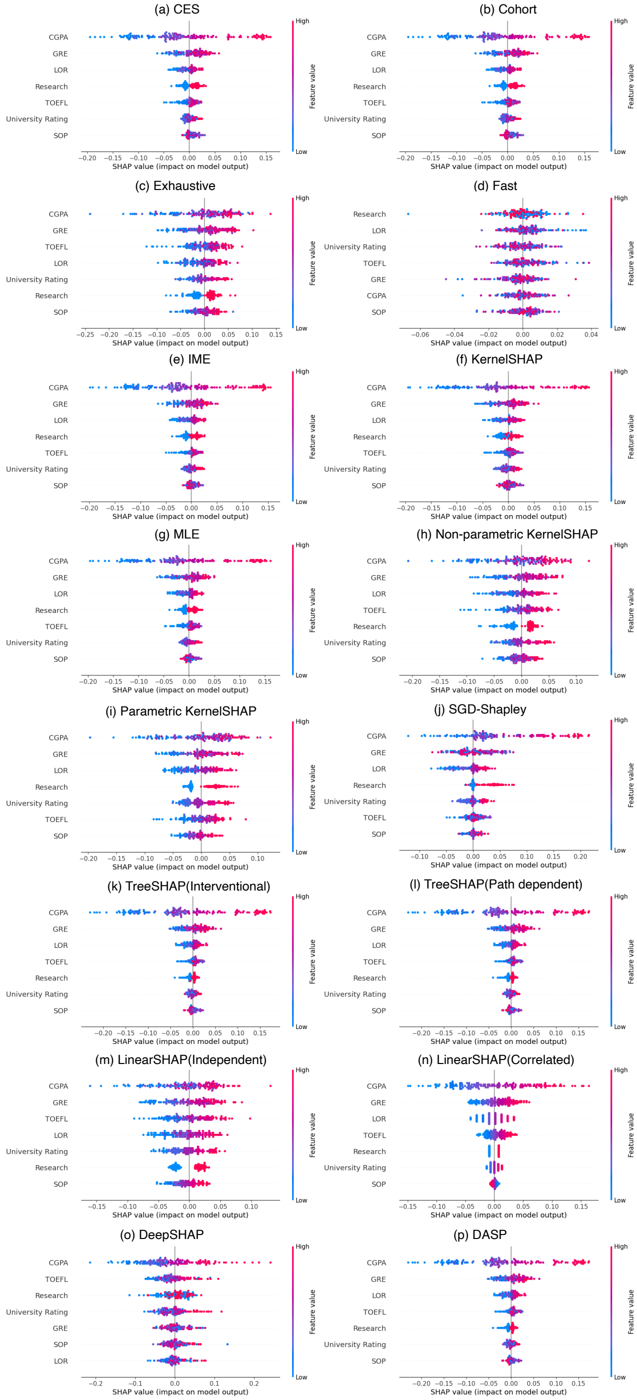

(a) CES (b) Cohort (c) Exhaustive (d) Fast (e) IME (f) KernelSHAP (g) MLE (h) Non-parametric KernelSHAP (i) Parametric KernelSHAP (j) SGD-Shapley (k) TreeSHAP(Interventional) (l) TreeSHAP(Path dependent) (m) LinearSHAP(Independent) (n) LinearSHAP(Correlated) (o) DeepSHAP (p) DASP

Figure 12: Comparison of Shapley value distribution specific to each feature in the "admission" dataset.

distributed coalitions. Consequently, imbalanced coalitions have a more significant impact on the Shapley value of a feature than a coalition with an equal distribution of features.

We can view feature ablation as an extension of the Shapley values framework that explicitly examines the coalitions located at the tail of the bell curve. The coalitions with the highest weight are the primary contributors to the Shapley value. Consequently, the remaining coalitions can be deemed insignificant. Therefore, potential research trajectories for the Shapley values revolve around sampling the skewed coalitions that include 0, 1, $|D| - 1$, $|D| - 2$ features. Given the inherent challenges in obtaining precise ground truth Shapley values, refining the sampling methodologies is a promising avenue for enhancing both efficiency and accuracy. Additionally, the leave-one-out technique of feature ablation requires training a new model by removing one feature. This process demands training models equivalent to the total number of features. Consequently, an alternative route for future exploration entails the integration of replacement strategies into feature ablation. This future direction has the potential to substantially reduce computational complexity and offer improved insights into a model's decision-making process.

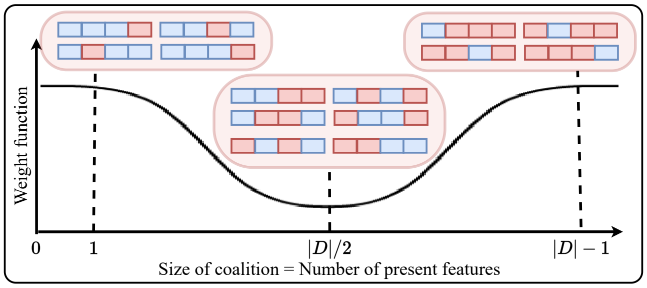


Figure 13: This plot illustrates the correlation between the weight function (refer to Equation 1) and the size of the coalition, which is determined by the number of features present. This graph exhibits an inverted bell curve. Imbalanced coalitions are located at the extremes of this curve. The degree of asymmetry in a coalition is directly linked to the weight assigned to that particular coalition.

7 CONCLUSION

Through this paper, we present a comprehensive study of various Shapley value approximation techniques, shedding light on their strengths, limitations, and implications for interpretability in machine learning models. We provide valuable insights regarding the effectiveness and applicability of different approximation techniques across diverse datasets and model structures. We emphasize that underlying assumptions and replacement strategies play a vital role in the reliability of Shapley value estimations. Moreover, the observed correlation among specific approximation techniques stresses the potential for leveraging conditional distributions as a robust replacement strategy in Shapley value approximation. Moving forward, further research is warranted to explore novel approaches and enhance the interpretability of machine learning models through improved Shapley value estimation techniques. Ultimately, this work contributes to the ongoing dialogue surrounding model interpretability and feature importance analysis, providing valuable insights to guide future developments in this field.

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

[6] 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. https://doi.org/10.1016/j.cor.2008.04.004 Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[7] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2019. Explaining Image Classifiers by Counterfactual Generation. arXiv:1807.08024 [cs.CV]

[8] 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

[9] 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]

[10] 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]

[11] 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). https://doi.org/10.1038/s41467-022-31384-3

[12] 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]

[13] 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]

[14] 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. https://doi.org/10.1145/2939672.2939785

[15] Ian Covert, Chanwoo Kim, and Su-In Lee. 2023. Learning to Estimate Shapley Values with Vision Transformers. arXiv:2206.05282 [cs.CV]

[16] Ian Covert and Su-In Lee. 2021. Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. arXiv:2012.01536 [cs.LG]

[17] Ian Covert, Scott Lundberg, and Su-In Lee. 2022. Explaining by Removing: A Unified Framework for Model Explanation. arXiv:2011.14878 [cs.LG]

[18] Nello Cristianini and Elisa Ricci. 2008. *Support Vector Machines*. Springer US, Boston, MA, 928–932. https://doi.org/10.1007/978-0-387-30162-4_415

[19] Piotr Dabkowski and Yarin Gal. 2017. Real Time Image Saliency for Black Box Classifiers. arXiv:1705.07857 [stat.ML]

[20] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580

[21] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. arXiv:1910.08485 [cs.CV]

[22] 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. https://doi.org/10.1109/iccv.2017.371

[23] 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.

[24] 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]

[25] Johannes Fürnkranz. 2010. *Decision Tree*. Springer US, Boston, MA, 263–267. https://doi.org/10.1007/978-0-387-30164-8_204

[26] Simon Grah and Vincent Thouvenot. 2020. *A Projected Stochastic Gradient Algorithm for Estimating Shapley Value Applied in Attribute Importance*. 97–115. https://doi.org/10.1007/978-3-030-57321-8_6

[27] 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]

[28] Simon Haykin. 1994. *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

[29] 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. https://doi.org/10.1002/widm.1312

[30] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. 2019. Feature relevance quantification in explainable AI: A causal problem. arXiv:1910.13413 [stat.ML]

[31] 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]

[32] Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. 2022. FastSHAP: Real-Time Shapley Value Estimation. arXiv:2107.07436 [stat.ML]

[33] Kamal Kasmaoui. 2019. *Linear Regression*. Springer International Publishing, Cham, 1–11. https://doi.org/10.1007/978-3-319-31816-5_478-1

[34] Bengio Y. Hinton G. LeCun, Y. 2015. Deep learning. *Nature* 521, 436–444 (2015). https://doi.org/10.1038/nature14539

[35] 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). https://doi.org/10.1038/s42256-019-0138-9

[36] 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.

[37] 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. https://doi.org/10.14778/3551793.3551829

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

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

[40] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. 2022. Sampling Permutations for Shapley Value Estimation. arXiv:2104.12199 [stat.ML]

[41] 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).

[42] 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

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

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

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

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

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

[48] 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]

[49] 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]

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

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

[52] 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

[53] 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. https://doi.org/10.1145/2939672.2939778

[54] 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]

[55] Reuven Y. Rubinstein. 1981. Simulation and the Monte Carlo method. In *Wiley series in probability and mathematical statistics*. https://api.semanticscholar.org/CorpusID:39230485

[56] 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

[57] Patrick Schwab and Walter Karlen. 2019. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. arXiv:1910.12336 [cs.LG]

[58] Lloyd S. Shapley. 1988. A Value for n-person Games. https://api.semanticscholar.org/CorpusID:153629957

[59] 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]

[60] Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. *J. Mach. Learn. Res.* 11 (mar 2010), 1–18.

[61] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. arXiv:1908.08474 [cs.AI]

[62] Erik trumbelj 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

[63] Rui Wang, Xiaoqian Wang, and David I. Inouye. 2021. Shapley Explanation Networks. arXiv:2104.02297 [cs.LG]

[64] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* (1945), 80–83.

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

[66] 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]

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

[68] 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. https://doi.org/10.1109/ICDE55515.2023.00055

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

[70] 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. https://doi.org/10.1016/j.datak.2009.01.004