Learning Confidence Intervals for Feature Importance: A Fast Shapley-based Approach

Learning Confidence Intervals for Feature Importance: A Fast Shapley-based Approach

Davide Napolitano[1], Lorenzo Vaiani[1] and Luca Cagliero[1]

[1]Politecnico di Torino, Turin, Italy

Abstract

Inferring feature importance is a well-known machine learning problem. Giving importance scores to the input data features is particularly helpful for explaining black-box models. Existing approaches rely on either statistical or Neural Network-based methods. Among them, Shapley Value estimates are among the mostly used scores to explain individual classification models or ensemble methods. As a drawback, state-of-the-art neural network-based approaches neglects the uncertainty of the input predictions while computing the confidence intervals of the feature importance scores. The paper extends a state-of-the-art neural method for Shapley Value estimation to handle uncertain predictions made by ensemble methods and to estimate a confidence interval for the feature importances. The results show that (1) The estimated confidence intervals are coherent with the expectation and more reliable than baseline methods; (2) The efficiency of the Shapley value estimator is comparable to those of traditional models; (3) The level of uncertainty of the Shapley value estimates decreases while producing ensembles of larger numbers of predictors.

1. Introduction

Machine learning and deep learning have achieved remarkable results in various classification tasks. However, due to the inherent complexity end-users often treat them as black-boxes as these models do not provide the necessary insights into the reasons behind the generated predictions [1]. Understanding feature importance is a relevant branch of Explainable AI. The main goal is to estimate the predictive power of a feature for a response variable [2]. To successfully cope with arbitrary predictive models, especially the non-intepretable ones such as neural networks or ensemble methods (e.g., random forests or Gradient Boosting [3]), a particular research interest has been devoted to studying *model-agnostic methods*. They compute the feature importance scores disentangling the approximations from the underlying model characteristics. Within this field, statistics-based approaches to feature importance have two major issues [4]: (1) they make arbitrary distributional assumptions, which are often hard to verify in practice on real data, (2) they neglect, in most cases, the uncertainty of model estimates thus providing end-users with unreliable feature importance scores. This paper addresses the above-mentioned issues as follows: (1) It adopts a state-of-the-art neural network model learning the underlying data distribution at training time; (2) It quantifies the uncertainty of the feature importance scores by learning the corresponding confidence scores.

Shapley Values [5] are known concepts from cooperative game theory that have become established for AI model explanation. Specifically, they quantify the contribution of a given feature to the prediction of a particular instance. Thanks to the additive property, they can be also used to estimate the global contribution of a feature to an AI model [6]. Since the exact Shapley Value estimate is computationally intractable on most real datasets different approximation methods have been proposed. They can be classified as stochastic approaches (e.g., [7, 6, 8]) or model-based ones (e.g., [9, 10]). Among the latter ones, recent approaches based on Neural Network models [10] have shown to be particularly efficient as allow real-time Shapley Value estimate in a single forward pass using a learned explainer model.

The main paper contributions are outlined below.

- **Conceptualization.** We propose to extend existing Shapley Value approximation methods to cope with uncertain predictors by leveraging the concepts of Coalition Interval Game [11] and Interval Shapley Value [12].
- **Design and Implementation.** We introduce *Interval FastSHAP*, a novel and efficient methodology for the approximation of Shapley values, which builds upon the existing state-of-the-art model, FastSHAP [10]. Given an ensemble of predictors associated with the corresponding confidence intervals, it returns the Shapley Values enriched with the corresponding confidence intervals.
- **Comparative study.** To compare Interval FastSHAP with baseline methods, we also extend a statistical approach based on Montecarlo sam-

Published in the Workshop Proceedings of the EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023, Ioannina, Greece)

davide.napolitano@polito.it (D. Napolitano); lorenzo.vaiani@polito.it (L. Vaiani); luca.cagliero@polito.it (L. Cagliero)

0000-0001-9077-4103 (D. Napolitano); 0000-0002-3605-1577 (L. Vaiani); 0000-0002-7185-5247 (L. Cagliero)

pling [13] and a recently proposed regression-based model, namely Biased KernelSHAP [6], to handle confidence intervals.
- **Empirical evaluation.** We report a selection of empirical outcomes achieved on benchmark tabular datasets [14]. The estimated intervals have shown to be more reliable than Biased KernelSHAP; Interval FastSHAP has a complexity that is comparable to FastSHAP (and superior to Biased KernelSHAP); the uncertainty of the Shapely Value approximations decreasing by increasing the number of predictors.

The rest of the paper is organized as follows. Section 2 provides an overview of related works in the field of explainability and feature importance estimation. Section 3 presents some preliminary notions about Shapley Values and FastSHAP architecture. Section 4 introduces Interval FastSHAP, the proposed methodology for estimating Shapley Values with associated confidence intervals. Section 5 presents the empirical evaluation of Interval FastSHAP and compares it with the baseline methods. Finally, Section 6 concludes the paper and outlines possible future works.

2. Related works

In recent years, various model-agnostic feature importance scores have already been proposed in the literature. They can be classified as follows:

a) **Feature exclusion/occlusion methods** (e.g., [15, 16]), which investigate the impact of excluding part of the input features;
b) **Feature permutation** (e.g., [17, 18]), which ensemble predictive models by combining different feature sets;
c) **Shapley value-based feature importance** (e.g., [6, 19]), which quantifies the relevance score of a feature by approximating the per-class global Shapley values [6].

This work belongs to category (c). Few works have focused on quantifying the reliability or uncertainty of the feature importance based on statistical models. For example, [16] performs leave-one-covariate-out inference, [20] adopt MonteCarlo feature sampling, whereas [4] uses minipatch ensembles. However, these approaches are computationally expensive and may not scale well to high-dimensional feature spaces. We believe that our approach provides a robust and scalable solution to the problem of quantifying feature importance and uncertainty in high-dimensional feature spaces. Unlike [16, 4, 20] the approach proposed in the present work relies on neural network learning for efficient Shapley Value approximation. A regression-based approach to estimate Shapley Value residuals has been proposed in [21]. The goal is to warn practitioners against overestimating the extent to which Shapley-value-based explanations give them insights into a model. Unlike [21], the approach presented in the current paper also considers the uncertainty of black-box models consisting of predictor ensembles, focusing on quantifying the uncertainty of feature importance rather than the accuracy of Shapley values.

3. Preliminaries

Shapley Value Introduced in 1951, the Shapley Value assigns a value ϕ_p to each player p in a cooperative game based on the contribution to the total payoff of the group [5].

Formally speaking, the Shapley Value for a player p in a cooperative game with a set of players P and a characteristic function is defined as follows:

$$\phi(p) = \sum_{C \subseteq P \setminus p} \frac{|C|!(|P| - |C| - 1)!}{|P|!} [v(C \cup p) - v(C)]$$

where C is player coalition, $v : 2^N \rightarrow \mathbb{R}$ is a characteristic function, $C \subseteq (P \setminus p)$ is the sum taken over all subsets C of players in P excluding p, $|C|$ is the cardinality of set C, and $|P|$ is the total number of players.

The Shapley Value is computed as the sum over all possible coalitions that do not contain coalition C. The term $v(C \cup p) - v(C)$ is the marginal contribution of player p to the coalition C.

The Shapley value satisfies the axioms of *efficiency*, *symmetry*, *linearity*, and *dummy player* [5]. Efficiency indicates that the sum of the Shapley values for all players is equal to the total payoff of the game; symmetry indicates that if two players have the same marginal contributions to all possible coalitions, their Shapley values are equal; linearity holds because the total payoff of the game can be decomposed into two independent parts and the Shapley value of each player can be obtained by summing their individual Shapley values for each part; dummy player indicates that the Shapley value of a player having no marginal contribution to the total payoff is zero. Notably, linearity allows us to sum the instance-level contributions of a feature for global explanability [6].

Coalition Interval Game For every coalition C in a cooperative game the achieved outcomes have a certain level of uncertainty. We assume that the prediction pr^i of a model M on instance i has a confidence interval $[pr_i^{lower}, pr_i^{upper}]$. The aforesaid range is bounded from below by the pessimistic prediction obtained using the lower value of the associated zero-sum game and it is bounded from above by the optimistic prediction obtained using the upper value of that game. In compliance with [11], we associate with each strategic game

a *coalitional interval game* consisting of a pair (P, v), where P is the set of players, and v is a correspondence that associates with every coalition $C \subseteq P$ an interval $v(C)$ that indicates that the worth of the coalition will be somewhere in this range.

Shapley Value with confidence interval Analogously to coalition interval games, the estimate of the Shapley Value ϕ_f of feature f can be extended to define the corresponding confidence interval on a given instance i [$\phi_i^{f,lower}$, $\phi_i^{f,upper}$] [20]. The confidence interval quantifies the range of uncertainty of the importance of feature f. The traditional Shapley Value ϕ_f is expected to be the mean interval value [12].

FastSHAP FastSHAP [10] is a state-of-the-art model-agnostic approach to real-time Shapley Value approximation. Unlike prior works (e.g., Biased KernelSHAP [6], Unbiased KernelSHAP [22]) it exploits surrogate models to simulate the original, complex, black-box model to be explained. Surrogate models are trained on the same data used to train the original model and aim at predicting the outputs of the black-box model generated by taking into consideration different subsets of features. Based on the surrogate model outcomes, FastSHAP returns the Shapley value approximation in a single-forward pass by minimizing the difference between the output of the surrogate models and the local normalized output. By using surrogate models, FastSHAP can estimate the global Shapley values for the original model more quickly and with a lower computational cost than if we computed the exact Shapley values for the original model directly.

The FastSHAP explainer model is trained by minimizing an objective function inspired by the Shapley Value's weighted least squares characterization [6], thus enabling efficient gradient-based optimization. It also seeks to balance the trade-off between accuracy and fairness in the model's explanations. More specifically, the loss function $L(\cdot)$ is defined as follows:

$$L(\theta) = \mathop{\mathbb{E}}_{p(x)} \mathop{\mathbb{E}}_{U(y)} \mathop{\mathbb{E}}_{p(s)} \left[\left(v_{x,y}(s) - v_{x,y}(0) - s^T \phi_{fast}(x, y; \theta) \right)^2 \right]$$

where x is the feature representation vector corresponding to a sample, y is the response variable for a classification problem, $U(y)$ represents the Uniform distribution over the classes, s represent a subset of feature to be considered to infer the label of a data sample, $v_{x,y}(s)$ is the expected value of the model's prediction when considering only features in s, $v_{x,y}(0)$ is the expected value of the model's prediction when all features are absent and $\phi_{fast}(x, y; \theta)$ is the learned parametric function that should outputs exact Shapley values.

To meet the efficiency constraint, FastSHAP applies a normalization factor to all predictions, namely the *additive efficient normalization*:

$$\phi_{fast}^{eff}(x, y; \theta) = \phi_{fast}(x, y; \theta) + \\ + \frac{1}{d} \Big((v_{x,y}(1) - v_{x,y}(0) - 1^T \phi_{fast}(x, y; \theta) \Big)$$

where $v_{x,y}(1)$ is the expected value of the model's prediction when all features are present in the sample, and d is the total number of features. By incorporating additive efficient normalization into the loss function, the FastSHAP model ensures that the resulting feature attributions are consistent with the Shapley value, providing a theoretically grounded and transparent method for interpreting the model's predictions.

4. Interval FastSHAP

Figure 1 shows a sketch of the Interval FastSHAP workflow. The goal is to explain a black-box prediction model $\mathbf{M}$ consisting of an ensemble of N predictors M_1, M_2, ..., M_N. Without any loss of generality, hereafter we will address the problem of explaining an ensemble of binary classifiers predicting either the positive ($+$) or the negative ($-$) class.

Black-box model Given a dataset D, for each instance i in D let pr_i be the prediction of model M for instance i. Let

$$\begin{aligned} CI^+ &= [pr_i^{+,lower}, pr_i^{+,upper}] \\ CI^- &= [pr_i^{-,lower}, pr_i^{-,upper}] \end{aligned} \tag{1}$$

be the confidence intervals of predictor $\mathbf{M}$ associated to instance i for positive and negative classes, respectively. For instance, if $\mathbf{M}$ is an ensemble of decision trees then the per-class confidence levels can be defined by the range of variation of N trees' predictions. Since the per-class probabilities $\mathbf{P}(i, -)$ and $\mathbf{P}(i, +)$ of a given instance i are linearly dependent, i.e., $\mathbf{P}(i, -)$ = 1 - $\mathbf{P}(i, +)$, we simplify the model output setting as target the crossed pairs

$$\begin{aligned} V_1 &= [pr_i^{-,lower}, pr_i^{+,upper}] \\ V_2 &= [pr_i^{+,lower}, pr_i^{-,upper}] \end{aligned} \tag{2}$$

In details, since the variance on each class is the same, i.e. $Var_i^- = Var_i^+$, rather than considering four vectors, one for each bound

$$\begin{aligned} V1 &= [pr_i^{-,lower}, 1 - pr_i^{-,lower}] \\ V2 &= [pr_i^{+,lower}, 1 - pr_i^{+,lower}] \\ V3 &= [1 - pr_i^{-,upper}, pr_i^{-,upper}] \\ V4 &= [1 - pr_i^{+,upper}, pr_i^{+,upper}] \end{aligned} \tag{3}$$

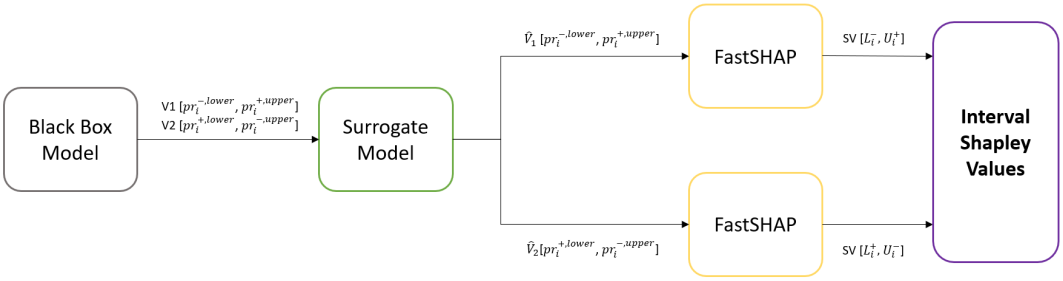


Figure 1: Interval FastSHAP architecture.

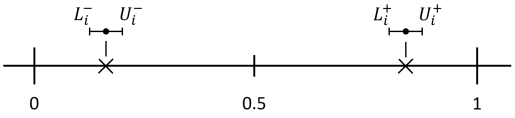


Figure 2: Example of prediction targets.

and recalling the following equivalences, as shown in Figure 2,

$$\begin{aligned} 1 - pr_i^{-,lower} &= pr_i^{+,upper} \\ 1 - pr_i^{+,lower} &= pr_i^{-,upper} \\ 1 - pr_i^{-,upper} &= pr_i^{+,lower} \\ 1 - pr_i^{+,upper} &= pr_i^{-,lower} \end{aligned} \tag{4}$$

we can rewrite the original target vectors as

$$\begin{aligned} V1 &= [pr_i^{-,lower}, pr_i^{+,upper}] \\ V2 &= [pr_i^{+,lower}, pr_i^{-,upper}] \\ V3 &= [pr_i^{+,lower}, pr_i^{-,upper}] \\ V4 &= [pr_i^{-,lower}, pr_i^{+,upper}] \end{aligned} \tag{5}$$

and simply consider two of them:

$$\begin{aligned} V1 = V4 &= [pr_i^{-,lower}, pr_i^{+,upper}] \\ V2 = V3 &= [pr_i^{+,lower}, pr_i^{-,upper}] \end{aligned} \tag{6}$$

Surrogate model The surrogate model **SM** is designed to approximate the behavior of a black-box model **M**. The objective of the surrogate model is to predict the class label $\hat{pr}_i^{+/-}$ of instance i as determined by the black-box model. In order to achieve this, the surrogate model may employ any suitable prediction algorithm that is computationally more efficient and able to accommodate the utilization of varying subsets of features. Regarding the implementation, in this study, a multi-layer perceptron is utilized as the surrogate model, trained to predict the outcome of the random forest approach which is employed as the black-box model. It takes as input the original data points and the vectors V_1 and V_2 as target variables and outputs two vectors

$$\begin{aligned} \hat{V}_1 &= (\hat{pr}_i^{-,lower}, \hat{pr}_i^{+,upper}) \\ \hat{V}_2 &= (\hat{pr}_i^{+,lower}, \hat{pr}_i^{-,upper}) \end{aligned} \tag{7}$$

Combining FastSHAP explainers Two FastSHAP explainers are trained in parallel to infer the interval Shapley values SV_i^+ and SV_i^-. Given an arbitrary instance i, Interval FastSHAP aims at the Interval Shapley Values consisting of the two vector pairs SV_i^+=(L_i^+,U_i^+) and SV_i^-=(L_i^-,U_i^-), where SV_i^+/SV_i^- are the interval Shapley values associated to instance i for the positive and negative classes, respectively. Due to the linear dependency of per-class probabilities, the interval boundaries are predicted by the FastSHAP network in a crossed fashion, accordingly to the previous explanation:

$$\begin{aligned} SV1 &= [L_i^-, U_i^+] \\ SV2 &= [L_i^+, U_i^-] \end{aligned} \tag{8}$$

Specifically, vectors L_i^+ and L_i^- contain the lower bound estimates of the positive/negative confidence intervals of instance i, where the f-th vector dimension corresponds to feature with index f in the input dataset D. The same holds for U_i^+ and U_i^- in the context of upper bounds.

5. Preliminary results

Data We perform experiments on four benchmark tabular datasets belonging to the UCI repository [14], i.e., Monks, Heart, Census, and WBC.

Models and settings To implement the Random Forest classifier, we employed the implementation provided by the widely used scikit-learn library [23]. For all experiments, the number of trees was set to 100, except in the

	FastSHAP				Biased KernelSHAP				MonteCarlo			
	Mean		CI		Mean		CI		Mean		CI	
	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1
Monks	**0.0111**	**0.0224**	**0.0059**	**0.0118**	0.0239	0.0489	0.0194	0.0386	0.1184	0.2241	0.0151	0.0275
WBC	**0.0291**	**0.0721**	0.0186	0.0449	0.0927	0.2272	0.0783	0.1871	0.1102	0.2706	**0.0101**	**0.0240**
Heart	0.0479	0.1373	0.0227	0.0628	**0.0434**	**0.1268**	0.0376	0.1053	0.1334	0.3748	**0.0108**	**0.0282**
Census	**0.0308**	**0.0776**	0.0135	0.0341	0.0445	0.1256	0.0388	0.1044	0.1093	0.2642	**0.0112**	**0.0274**

Table 1
L_1 and L_2 distances computed separately for each dataset (rows) and for each method (columns). The subcolumns Mean and CI indicate the distance between the interval mean point and the distance between interval widths, respectively. All distances are computed against the ground truth and averaged over 100 random samples.

studies exploring the impact of varying the number of trees.

As a surrogate model, we implemented a Multi-Layer Perceptron (MLP), which is a commonly used neural network architecture. The MLP consisted of three linear layers, of hidden size = 512 and interspersed with Rectified Linear Unit (ReLU) activation functions, and two classification heads, one for each target vector. The surrogate model has been trained for a maximum of 200 epochs using the Kullback-Leibler divergence loss, the AdamW optimizer [24], learning rate = 10^{-4}, batch size = 8 and weight decay = 10^{-2}.

For the explainer we use the implementation provided by the FastSHAP authors [10]. It is built as a MLP of 3 linear layers of hidden size = 128 and interspersed with ReLU activation functions. It has been trained for a maximum of 200 epochs using the custom loss described in section 3 together with the additive efficient normalization, the AdamW optimizer [24], learning rate = 10^{-2}, batch size = 8 and weight decay = $5 * 10^{-2}$.

Ground truth To generate the ground truth, we compute the Shapley Values estimates using Unbiased KernelSHAP [22] (with paired sampling) as the model is known to converge to the true Shapley Values given infinite samples. The confidence interval of the true Shapley values is computed using a modified version of Unbiased KernelSHAP estimating both interval boundaries at the same time.

Baselines We extend the following baseline methods to estimate the lower and bounds of the Shapley Value confidence intervals:

- A statistical approach based on Montecarlo sampling [25], hereafter denoted by *Montecarlo*.
- A recently proposed regression-based model [6], i.e., *Biased KernelSHAP*[1].

[1]Biased KernelSHAP is the predecessor of Unbiased KernelSHAP, from which we derive the true Shapley Values.

	True S.V.	FastSHAP	Biased KernelSHAP	MonteCarlo
Monks	0.0054	0.0056	0.0099	0.0059
WBC	0.0040	0.0073	0.0241	0.0034
Heart	0.0029	0.0063	0.0100	0.0026
Census	0.0037	0.0050	0.0112	0.0040

Table 2
Confidence Interval width.

5.1. Accuracy of the explanations

We test how Interval FastSHAP estimates are close to the ground truth Interval Shapley values. To this end, we compute the proximity of the Interval FastSHAP estimates with the ground truth in terms of mean L_1 and L_2 norms. The obtained results are reported in Table 1. The Interval FastSHAP approach demonstrates improved performance in terms of the distance between the interval mean points on three out of four datasets, i.e., Monks, WBC and Census, whereas achieves particularly close results on the Heart dataset, approaching the performance of the best-performing competitor, i.e., Biased KernelSHAP. In terms of interval width prediction, the Montecarlo approach outperforms the other tested methods, providing reasonable interval ranges while centering the interval away from the target mean point. FastSHAP achieves slightly worse results, while, in contrast, Biased KernelSHAP consistently exhibits wider interval predictions.

To quantify the reliability of the mean Shapley Value estimate we also compare the widths of the confidence intervals of the estimated and true Shapley Values. Table 2 reports the confidence interval width for both the ground truth and the tested approaches. Montecarlo produces intervals with minimal width despite that the reliability of the mean Shapley Value is averagely low (see Table 1). KernelSHAP significantly overestimates the width of the confidence interval, showing low reliability of the generated feature importance scores. Conversely, FastSHAP achieves a good trade-off between mean accuracy and

	FastSHAP	Biased KernelSHAP	MonteCarlo
Monks	**0.02s**	2.47s	182.22s
WBC	**0.01s**	207.63s	224.40s
Heart	**0.04s**	6.30s	335.57s
Census	**0.01s**	6.98s	310.81s

Table 3
inference times in seconds computed for each dataset (rows) and for each method (columns).

confidence of the estimate, with a slight overestimation of the actual confidence interval width.

In Figure 3 we plot the Global Shapley Values [26] as an estimate of the global measure of feature importance. They are computed as the mean of the per-instance Shapley Value estimates. The results confirm the bias of Montecarlo sampling-based approaches and the comparable performance of FastSHAP and KernelSHAP estimates on the majority of the input features.

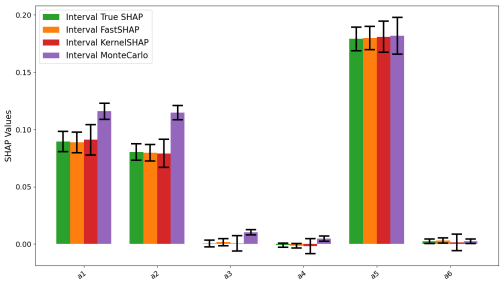
5.2. Execution times

Table 3 compares the inference times per sample spent by all analyzed approaches separately for each tested dataset. The inference step has been performed on a 18 core Intel Xeon Gold 6140. The reported statistics show that Interval FastSHAP is more efficient than MonteCarlo and competitive against Biased KernelSHAP. Notably, Interval FastSHAP also requires a training time overhead. However, similar to FastSHAP [10] (and unlike MonteCarlo and Biased KernelSHAP) it can be used for real-time Interval Shapley Value estimation.

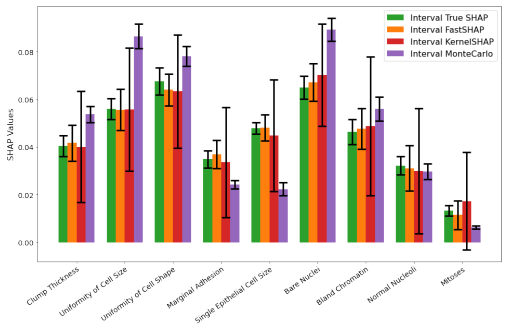
5.3. Effect of the number of predictors

The results of our study indicate that the uncertainty inherent in the predictions of black-box models can have a substantial effect on the reliability of Shapley Value estimates. In particular, the size of the Confidence Interval, which provides an estimate of the degree of uncertainty in the predictions, has shown to be dependent on the number of predictors used in the ensemble method (see Figures 4 and 5 for the Heart and Monks datasets). As the number of predictors increases, the mean Shapley Value estimate converges to a steady state and the Confidence Interval gets smaller, indicating a decrease in uncertainty. This underscores the importance of utilizing a sufficient number of predictors in order to ensure reliable estimates of the Shapley Values.

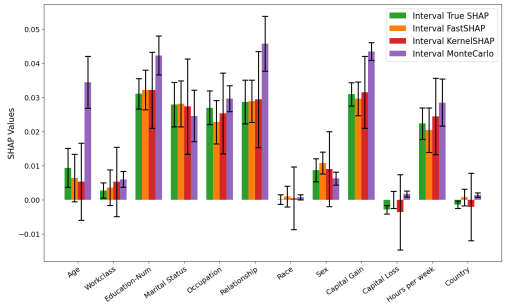
However, it is important to consider that increasing the number of predictors may also result in overfitting, which could lead to a decrease in the overall performance of the model. As such, it is necessary to balance the need



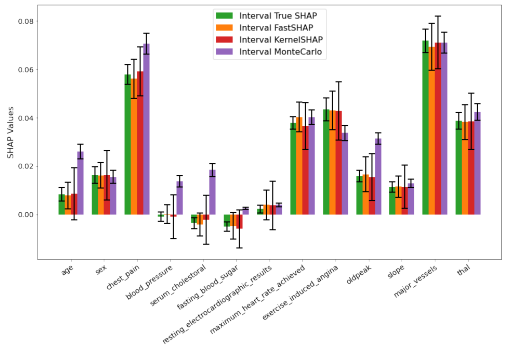
(a) *Monks* dataset



(b) *WBC* dataset



(c) *Census* dataset



(d) *Heart* dataset

Figure 3: Visual representation of Global Shapley values

for a sufficient number of predictors with the need to avoid overfitting.

6. Conclusions and future work

The paper presented a Shapley-based approach to learn confidence intervals for feature importance. It is suited to explain ensemble methods, whose predictors return uncertain outcomes. We leverage the concept of Coalition Interval Game and Interval Shapley Value to adapt the real-time neural network-based approach to handle uncertain input and produce as output confidence intervals in conjunction with the Shapley Value estimates.

The main takeaways can be summarized as follows:

- **Explanation accuracy**: Interval FastSHAP turns out to be significantly more reliable than MonteCarlo in estimating the mean Shapley Value and less susceptible to uncertainty than Biased KernelSHAP.
- **Real-time confidence interval estimation**: Interval FastSHAP is comparable to existing approaches in terms of inference time. Although requiring a computational time overhead for model training, Interval FastSHAP leverages the capability of the original FastSHAP to perform real-time Shapley Value estimates. Notably, the estimation of the confidence interval does not invalidate the efficiency of the original model.
- **Number of predictors**: The results of this study highlight the need for careful consideration of the number of predictors used when estimating the Shapley Values of ensemble black-box models, as the uncertainty inherent in the predictions can have a significant impact on the reliability of the estimates.

As future work, we plan to explore the applicability of the proposed approach to real application scenarios related to predictive maintenance, finance, and user profiling. We also aim to explore the use of different black-box and surrogate models.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

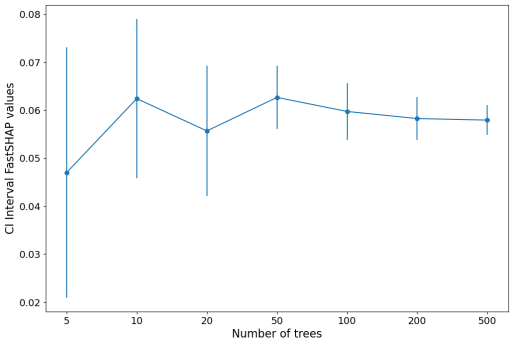


Figure 4: Effect of the number of model predictors. Hearth dataset.

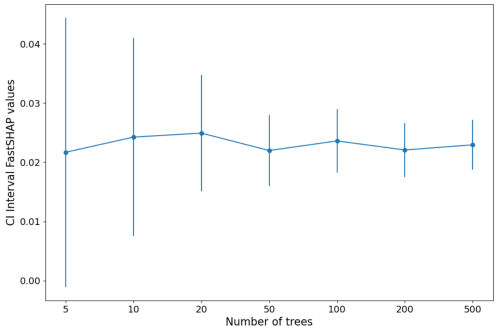


Figure 5: Effect of the number of model predictors. Monks dataset.

References

[1] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Trans. Interact. Intell. Syst. 11 (2021).

[2] I. Batal, M. Hauskrecht, Constructing classification features using minimal predictive patterns, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, ACM, New York, NY, USA, 2010, p. 869–878.

[3] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011.

[4] L. Gan, L. Zheng, G. I. Allen, Model-agnostic confidence intervals for feature importance: A fast and powerful approach using minipatch ensembles, 2022.

[5] L. S. Shapley, Notes on the N-Person Game II: The Value of an N-Person Game, RAND Corporation, Santa Monica, CA, 1951.

[6] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in

Neural Information Processing Systems, Curran Associates, Inc., 2017, pp. 4765–4774.
[7] J. Castro, D. Gómez, J. Tejada, Polynomial calculation of the shapley value based on sampling, Computers & Operations Research 36 (2009) 1726–1730.
[8] I. Covert, S. M. Lundberg, S.-I. Lee, Understanding global feature contributions with additive importance measures, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 17212–17223.
[9] S. M. Lundberg, G. G. Erion, H. Chen, A. J. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S. Lee, Explainable AI for trees: From local explanations to global understanding, CoRR abs/1905.04610 (2019).
[10] N. Jethani, M. Sudarshan, I. C. Covert, S. Lee, R. Ranganath, Fastshap: Real-time shapley value estimation, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022.
[11] L. Carpente, B. Casas-Méndez, I. García-Jurado, A. van den Nouweland, Coalitional interval games for strategic games in which players cooperate, Theory and Decision (2008) 253–269.
[12] S. Z. A. Gök, R. Branzei, S. Tijs, The interval shapley value: an axiomatization, Central Eur. J. Oper. Res. 18 (2010) 131–140.
[13] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2019.
[14] D. Dua, C. Graff, UCI machine learning repository, 2017.
[15] B. D. Williamson, P. B. Gilbert, M. Carone, N. Simon, Nonparametric variable importance assessment using machine learning techniques, Biometrics 77 (2021) 9–22.
[16] J. Lei, M. GaSell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association 113 (2018) 1094–1111.
[17] G. Smith, R. Mansilla, J. Goulding, Model class reliance for random forests, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.
[18] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. (2019).
[19] B. Williamson, J. Feng, Efficient nonparametric statistical inference on population feature importance using shapley values, in: H. D. III, A. Singh (Eds.), Proc. of the 37th Int. Conf. on Machine Learning, volume 119, PMLR, 2020, pp. 10282–10291.
[20] D. V. Fryer, I. Strümke, H. D. Nguyen, Shapley value confidence intervals for attributing variance explained, Frontiers Appl. Math. Stat. 6 (2020) 587199.
[21] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: Quantifying the limits of the shapley value for explanations, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 26598–26608.
[22] I. Covert, S. Lee, Improving kernelshap: Practical shapley value estimation via linear regression, CoRR abs/2012.01536 (2020).
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011).
[24] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[25] R. Y. Rubinstein, D. P. Kroese, Simulation and the Monte Carlo method, John Wiley & Sons, 2016.
[26] C. Frye, C. Rowat, I. Feige, Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability, 2019.