Fast Approximation of Shapley Values with Limited Data

Amr Alkhatib[1] and Henrik Boström[1]

Abstract— **Shapley values have multiple desired and theoretically proven properties for explaining black-box model predictions. However, the exact computation of Shapley values can be computationally very expensive, precluding their use when timely explanations are required. FastSHAP is an approach for fast approximation of Shapley values using a trained neural network (the explainer). A novel approach, called FF-SHAP, is proposed, which incorporates three modifications to FastSHAP: i) the explainer is trained on ground-truth explanations rather than a weighted least squares characterization of the Shapley values, ii) cosine similarity is used as a loss function instead of mean-squared error, and iii) the actual prediction of the underlying model is given as input to the explainer. An empirical investigation is presented showing that FF-SHAP significantly outperforms FastSHAP with respect to fidelity, measured using Spearman's rank-order correlation. The investigation further shows that FF-SHAP even outperforms FastSHAP when using substantially smaller amounts of data to train the explainer, and more importantly, FF-SHAP still maintains the performance level of FastSHAP even when trained with as little as 15% of training data.**

I. Introduction

The application of state-of-the-art machine learning algorithms in solving real-world problems in many domains, e.g., medicine and law, is limited by that the algorithms often produce black-box models [1]. Additionally, comprehending the reasoning behind the predictions is essential for verification and building confidence in such models [2]. Employing algorithms that produce interpretable (white-box) models, such as generalized linear models and decision trees, can provide the needed insights into how the predictions are derived. However, in many cases, using white-box models results in a significant reduction in predictive performance [3]. Therefore, the field of explainable machine learning has become an active research area as a way to achieve interpretability without compromising performance.

Explanation methods fall into two categories: model-agnostic methods that can explain any black-box model and model-specific methods that leverage the characteristics of the underlying black-box model to generate explanations, targeting models such as random forests [4], [5] and deep neural networks [6], [7]. Model-agnostic methods, such as LIME [8] and SHAP [9], focus on explaining a single prediction by feature scores that reflect the relative importance of each feature toward the predicted outcome. Methods that produce Shapley values as explanations are favored since they provide a solution that has been shown by [9] to be unique in the class of additive feature attribution methods, and satisfies the desired properties of local accuracy (the explanation matches the underlying model), missingness (a missing feature is attributed a value of zero), and consistency (when a model changes and a feature's contribution remains the same or increases, the Shapley value does too). However, exact computation of Shapley values requires forming coalitions of features and multiple model evaluations, and the number of the required coalitions grows exponentially with the number of features. Methods that do not produce Shapley values, e.g., LIME and Anchor [10], can also be computationally intensive. For instance, LIME involves creating a local (white-box) surrogate model that can be used to explain a single prediction. Consequently, methods have been proposed to reduce the cost of model-agnostic explainers, e.g., L2X [11], INVASE [12], REAL-X [13], and FastSHAP [14]. Notably, the state-of-the-art technique FastSHAP differentiates itself from the others by approximating Shapley values using a trained neural network (the explainer).

In this work, we propose a novel approach, called FF-SHAP (high **f**idelity **f**ast approximation method of **Shap**ley values), which makes three important modifications to FastSHAP: i) the explainer is trained using ground truth Shapley values, ii) cosine similarity is used as an objective function to maximize the similarity between the approximated and ground truth Shapley values, and iii) the black-box model prediction is given as input to the explainer.

We will argue for why these modifications can be expected to improve fidelity of the approximated explanations, without sacrificing computational performance. This argumentation is supported by presented results from an empirical investigation, in which FF-SHAP is compared to FastSHAP, and fidelity is measured using Spearman's rank-order correlation [15]. We also provide an ablation study where the effect of the two last components is investigated.

The next section provides a brief background on explainable machine learning. In Section III, we briefly discuss related work. In Section IV, the proposed method for approximating Shapley values is described and motivated. In Section V, we present and discuss the results of the empirical investigation. Finally, in Section VI, we summarize the main findings and outline directions for future work.

II. Background

Explainable Machine Learning is a field that focuses on making opaque machine learning models more understandable to users. While state-of-the-art machine learning models often deliver impressive performance, they usually act as black boxes, making it challenging to understand how

[1] KTH Royal Institute of Technology
Electrum 229, 164 40 Kista, Stockholm, Sweden
{alkhat,bostromh}@kth.se

they arrive at their decisions. Explainable Machine Learning methods aim to bridge this gap by providing human-understandable explanations for model predictions, which allow users to trust, validate, and comprehend the reasoning behind the model's outputs. Explainable Machine Learning methods come in various forms, including visualizations, feature importance scores, surrogate models, and rule extraction methods.

Examples of popular approaches for explaining machine learning models by visualizations are Partial Dependence Plots (PDPs) [16] and Individual Conditional Expectation (ICE) plots [17], which visualize the relationship between a feature and the model's predictions while marginalizing the remaining features. Another popular approach is rule-based explanation methods, e.g., Anchors [10], which aim to provide explanations by generating human-readable rules that mimic the decision-making process of the model. Explaining models through additive feature importance scores is one more favored approach. The class of additive feature importance scores involves methods that quantify the contribution of each input feature toward the model's predictions in a straightforward additive form. The importance scores provide a clear understanding of which features greatly impact the model's output, making it a widespread method for interpreting and explaining complex machine learning models. However, it's essential to recognize that these scores may not capture interactions between features accurately.

The concept of Shapley values is borrowed from cooperative game theory and has found significant application in explainable machine learning. Developed by Lloyd Shapley in the early 1950s [18], Shapley values provide a principled way to allocate each player's contribution in a coalition game. In the context of machine learning, the "players" represent the input features, and the "game" represents the predictive model. Explaining machine learning predictions using Shapley values involves calculating the marginal contribution of each feature towards a particular prediction across all possible combinations of features [19]. Shapley values ensure that the contributions of features are additive and sum up to the overall prediction. An example of an explanation based on Shapley values is illustrated in Figure 1.

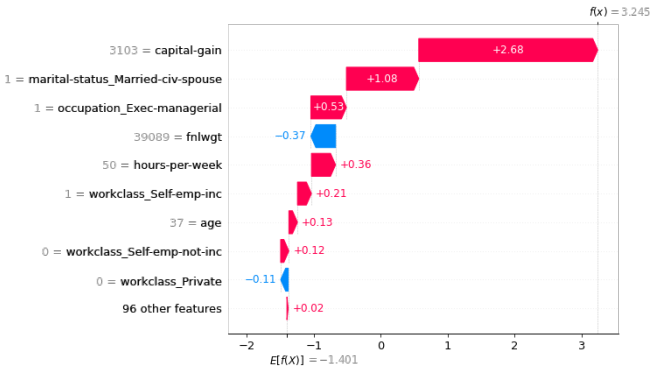


Fig. 1: An example of an explanation generated by KernelSHAP for a positive prediction made by an XGBoost model on the Adult dataset.

III. Related Work

Since the computation of the exact Shapley values can be infeasible due to the number of coalitions that need to be generated, recent research efforts on Shapley value explanations have focused on reducing the computational cost. Lundberg et al. [9] introduced KernelSHAP, a method that approximates Shapley values by randomly sampling feature coalitions and subsequently training a linear model to approximate the Shapley values. Model-specific variants can provide relatively faster approximations since they utilize specific properties of the explained model, e.g., TreeSHAP [20] for tree-based models and DASP [21] for deep neural networks. [22] proposed L-Shapley and C-Shapley for text and image classification, which employ a graphical data representation. H-Shap (Hierarchical Shap) [23] has also been introduced for image classification explanations as a fast and precise implementation to compute Shapley coefficients. [24] proposed the unbiased version of KernelSHAP alongside a convergence detection technique and variance reduction through paired sampling that also helps in faster convergence.

Methods to generate explanations using a pre-trained model have been investigated. [11] proposed to train a feature selection model by maximizing the mutual information between the selected features and the predicted variable by the black-box model. INVASE [12] is also conducting feature selection, however, INVASE is composed of 3 neural networks (a selector, a predictor, and a baseline), which are employed to train the feature selector. CXPlain (causal explanation) [25] trains a model to estimate the extent to which specific inputs influence the outcomes of another machine-learning model. Situ et al. [26] suggested that any off-the-shelf explanation algorithm can be distilled into an explainer neural network, with their approach named L2E (Learning to Explain), primarily concentrating on emulating explanations for text classification tasks. [13] introduced REAL-X, an amortized explanation method designed to generate explanations that align closely with the observed data in a single forward pass. As previously mentioned in Section I, FastSHAP [14] is distinguished by approximating the Shapley values using a trained model, a demanded property as Shapley values provide the sole solution that satisfies local accuracy, missingness, and consistency properties. FastSHAP evades the need for generating training data of ground truth Shapley values in order to train a model to approximate these values, which is achieved by employing a custom loss function with mean squared error (MSE) component that ensures the global optimizer functions as a means that produces the Shapley values. This methodology enables the training of the explainer model in a convenient time.

IV. The Proposed Method

The performance of FastSHAP has yet to be compared to the training based on pre-generated ground truth Shapley

values, as it is not clear if FastSHAP is achieving the same levels of fidelity as explainers trained on ground truth values. Moreover, at the inference time, FastSHAP receives only the features of the data instances without information about the outcome of the underlying black-box model. Hence, it is helpful to assess the impact of providing the explainer not only with the input features but also with the output of the underlying black-box model. Additionally, FastSHAP allows only the use of MSE in the loss function. Consequently, using ground truth Shapley values allows experimenting with other objective functions rather than MSE.

[24] showed that KernelSHAP converges to the true Shapley values when provided with a large number of samples. Consequently, the ground truth training data (Φ) can be obtained by allowing KernelSHAP to sample data and evaluate until it converges to some values, which can be time-consuming for high-dimensional data. However, the ground truth values are generated once at the training time. In contrast to FastSHAP, we propose that the input data instance $\mathbf{x}$ composed of d features $\mathbf{x} = [f_1, f_2, ..., f_d]$ can be supplemented by the predicted outcome of the black-box model $\mathbf{p} = [p_1, p_2, ..., p_c]$ to provide $\mathbf{x}^* = [f_1, f_2, ..., f_d; p_1, p_2, ..., p_c]$, and an explainer $\phi_{ff}(\mathbf{x}^*; \theta)$ can be trained to learn a mapping from $\mathbf{x}^*$ to $\phi = [\delta_1, \delta_2, ..., \delta_d]$. The FF-SHAP model $\phi_{ff}(\mathbf{x}^*; \theta)$ predicts an approximation of Shapley values $\hat{\phi}_i$ for the i-th data instance, and a gradient-based optimization is carried out to minimize the difference between $\hat{\phi}_i$ and the ground truth ϕ_i using a loss function, e.g., MSE. The proposed method is summarized in algorithm 1.

Algorithm 1: FF-SHAP

Data: data instances $\mathbf{X}$, black-box model β, a loss function γ, number of training epochs n and KernelSHAP $\phi_{kernel}(\mathbf{x}, \beta)$
Result: FF-SHAP $\phi_{ff}(\mathbf{x}; \theta)$
Initialize $\phi_{ff}(\mathbf{x}; \theta)$
$\Phi \leftarrow \{\}$
for $\boldsymbol{x}_i \in \boldsymbol{X}$ **do**
  explain $\Phi \overset{+}{\leftarrow} \phi_{kernel}(\mathbf{x}_i, \beta)$
end
for *number of training iterations* n **do**
  for $\boldsymbol{x}_i \in \boldsymbol{X}$ **do**
    $\mathbf{p}_i \leftarrow \beta(\mathbf{x}_i)$
    $\mathbf{x}_i^* \leftarrow (\mathbf{x}_i; \mathbf{p}_i)$
    $\hat{\phi}_i \leftarrow \phi_{ff}(\mathbf{x}_i^*; \theta)$
    $\mathcal{L} \leftarrow \gamma(\hat{\phi}_i, \phi_i \in \Phi)$
    Compute gradients $\nabla_\theta \mathcal{L}$
    Update $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}$
  end
end

Similarity metric. Picking the correct performance metric sets the compass for a machine learning process, as it shapes the optimization process and impacts the model's ability to meet the desired outcomes. [27] showed that Spearman's rank-order correlation is a suitable metric when it comes to similarity measurement between explanations, and the Euclidean distance, for example, can fail to detect similarity.

Since different estimations of Shapley values may bear different scales, metrics affected by the magnitudes of the features, e.g., l_2 distance, can lead to a misleading impression of closeness or similarity between approximated values and the ground truth values. We devise a toy example for illustration, where the ground truth is ϕ = [0.15, 0.2, 0.1] with two estimations $\hat{\phi}_1$ = [0.3, 0.45, 0.2] and $\hat{\phi}_2$ = [0.01, -0.01, 0.0]. According to the results as shown in Table I, l_2 distance indicates that $\hat{\phi}_2$ is a better approximation to the ground truth than $\hat{\phi}_1$ since it is a smaller distance, which is not true if the cosine similarity or Spearman's rank-order correlation are used, where the cosine similarity measures the similarity in the orientation between two vectors of feature scores [28], and the Spearman's rank-order measures the similarity in ranking the feature scores [27].

TABLE I: The similarity between the ground truth ϕ and two different approximations $\hat{\phi}_1$ and $\hat{\phi}_2$ using 3 possible metrics.

	l_2 distance	Cosine	Spearman
$f(\phi, \hat{\phi}_1)$	0.308	0.998	1.0
$f(\phi, \hat{\phi}_2)$	0.27	-0.131	-0.5

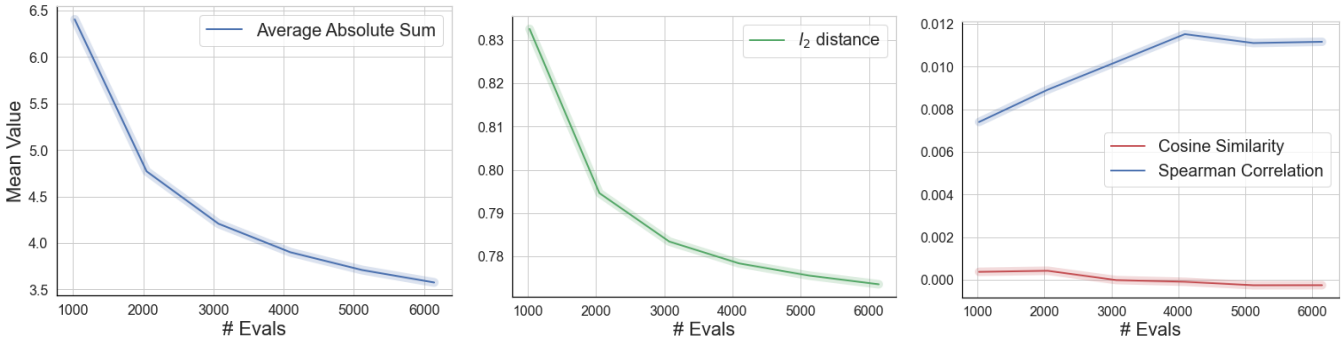
The previous claims are also supported by the following observation from the Scene dataset[1], where the magnitudes of the computed Shapley values tend to get smaller with more data sampling and KernelSHAP evaluation when explaining the predictions of an XGBoost model, as shown in Figure 2a. Consequently, the user can get a false impression of an increase in the accuracy of approximating the true values if the l_2 distance is used as a similarity measure where the distance between smaller magnitudes is smaller. Such effect is displayed in Figure 2b, where the l_2 distance is computed between FastSHAP explanations using a surrogate model and the approximated Shapley values after each iteration of KernelSHAP evaluation. However, such an increase in similarity does not appear if a different metric is used, e.g., cosine similarity or Spearman's rank-order correlation, as illustrated in Figure 2c.

Since the cosine similarity, as well as Spearman's rank-order can provide better metrics to measure the performance of an explainer in terms of how accurate the predicted scores are in approximating the true Shapley values, it can be useful to use such metric as a loss function to be optimized. Therefore, we propose to use the cosine similarity as an objective function to learn an accurate mapping from the input features to the corresponding Shapley values, as shown in Equation 1.

$$\mathcal{L} = 1 - \frac{\hat{\phi}\phi}{\|\hat{\phi}\|\|\phi\|} \tag{1}$$

where $\hat{\phi}$ is the approximated values, and ϕ is the ground truth Shapley values.

[1]The dataset is available on `openml.org` with ID: 312



(a) (b) (c)

Fig. 2: **Comparison of different similarity metrics.** Figure a shows that the summation of the absolute Shapley values tends to get smaller with more evaluations of KernelSHAP. Figure b shows the l_2 distance between kernelSHAP values after each iteration and the values approximated by FastSHAP. In Figure c, we use cosine similarity and Spearman's rank-order correlation to measure the similarity instead of l_2.

V. Empirical Investigation

In this section, we present results from two sets of experiments. In the first experiment, we compare FF-SHAP to the baseline method, FastSHAP. Afterward, we conduct an ablation study where we evaluate the effect of using the cosine similarity as an objective function and also the effect of augmenting the input features with the predicted outcome by the underlying black-box model.

A. Experimental Setup

In the experiments, we used ten publicly available datasets.The black-box models are XGBoost classifiers trained using the default settings. Each dataset is split into training, development, and test sets, where the training set is used to train the black-box model as well as training FF-SHAP and FastSHAP models. The development set is used for early stopping detection during the training phase. Finally, the test set is used to evaluate the trained explainers. The ground truth Shapley values are obtained using an online efficient open-source implementation[2], and the values are determined after KernelSHAP's convergence. The Spearman's rank-order correlation is the similarity metric between explanations.

FastSHAP and FF-SHAP share identical architectures and use the same set of hyperparameters. **Therefore, both have the same computational cost at the inference time, i.e., explanation time.** [3]

B. Experiments

In the following experiments, first, we compare the performance of FF-SHAP to FastSHAP when trained on the full training set. Then, we assess the effect of using different training set sizes.

FastSHAP is trained on the entire training data set, while FF-SHAP is compared when trained on the entire set, 60% of the training data, 30% of the training data, and 15% of the training data, in order to find out if FF-SHAP can achieve the performance level of FastSHAP using substantially smaller-sized datasets, which is particularly important since generating ground truth values can be computationally costly in high-dimensional data.

The trained FF-SHAP explainers generally showed higher fidelity than FastSHAP, even when trained using only 15% of the available training data. To test the null hypothesis that there is no difference in the fidelity, as measured by the Spearman's rank-order test, between FastSHAP and FF-SHAP explainers when compared to the ground truth Shapley values, we carried out statistical significance tests between FastSHAP and each training split size of FF-SHAP using the Wilcoxon signed-rank test [29]. The null hypothesis may be rejected at the 0.05 level for all the pairs compared except for FastSHAP and FF-SHAP trained using 15% of the data, which indicates that FF-SHAP can significantly achieve higher fidelity using substantially smaller size datasets. The detailed results are available in Table II

C. Ablation Study

In the following experiments, first, we assess the effect of using the cosine similarity as an objective function instead of MSE, and then, we evaluate the effect of augmenting the features with the predicted outcome by the black box on the fidelity of the generated explanations.

1) Objective Function: The results of training FF-SHAP using both MSE and cosine similarity as objective functions are available in Table III. The results demonstrate that cosine similarity helps to learn explainers with higher fidelity to the ground truth Shapley values. The results have been proven to be statistically significant when the Wilcoxon signed-rank test is applied, and the null hypothesis that there is no difference can be rejected at the 0.05 level.

2) Features Augmentation: In order to evaluate the effect of augmenting the input features with the predicted outcome by the black-box model, we train the FF-SHAP explainers without any augmentation to the input features and compare

[2] https://github.com/iancovert/shapley-regression/

[3] The source code is available at: https://github.com/amrmalkhatib/ff-shap

TABLE II: The similarity between the ground truth Shapley values and the explanations generated by FastSHAP and FF-SHAP. FastSHAP is trained using all the training data, while FF-SHAP is trained using different training data sizes.

Dataset	FastSHAP	FF-SHAP	FF-SHAP 60%	FF-SHAP 30%	FF-SHAP 15%
Abalone	0.81	0.861	0.851	0.827	0.803
Bank32nh	0.598	0.692	0.67	0.632	0.6
Churn	0.311	0.534	0.511	0.49	0.462
Delta Ailerons	0.867	0.906	0.891	0.868	0.848
Electricity	0.625	0.702	0.699	0.678	0.655
Elevators	0.828	0.855	0.848	0.836	0.829
Higgs	0.678	0.721	0.698	0.638	0.58
JM1	0.781	0.849	0.835	0.808	0.787
MC1	0.198	0.723	0.717	0.71	0.692
PC2	0.299	0.588	0.581	0.572	0.565

TABLE III: The similarity of the generated explanations to the ground truth Shapley values when FF-SHAP is trained using the mean squared error (MSE) vs. when trained using the cosine similarity as an objective function.

Dataset	Cosine	MSE
Abalone	**0.861**	0.857
Bank32nh	**0.692**	0.652
Churn	**0.534**	0.404
Delta Ailerons	**0.906**	0.905
Electricity	0.702	**0.725**
Elevators	**0.855**	0.853
Higgs	**0.721**	0.72
JM1	**0.849**	0.837
MC1	**0.723**	0.208
PC2	**0.588**	0.43

the similarity to the ground truth Shapley values of the test set. The results in Table IV show better performance for the explainers trained using augmented features. Again, these results are subjected to the Wilcoxon signed-rank test, which also allowed us to reject the null hypothesis at the 0.05 level that there is no difference in the fidelity when the explainers are trained with and without input features augmentation with the black box's prediction.

TABLE IV: The similarity between the ground truth Shapley values and the explanations generated by FF-SHAP when trained with and without augmentation of the features by the predicted outcome by the underlying black box.

Dataset	Augmented Input	Original Input
Abalone	**0.861**	0.843
Bank32nh	**0.692**	0.684
Churn	**0.534**	0.525
Delta Ailerons	**0.906**	0.905
Electricity	**0.702**	0.676
Elevators	**0.855**	0.847
Higgs	**0.721**	0.712
JM1	**0.849**	0.844
MC1	0.723	**0.724**
PC2	**0.588**	0.585

VI. Conclusion

We proposed a method to approximate Shapley values of the predictions using a pre-trained neural network with higher similarity to the ground truth values compared to the baseline method, FastSHAP. The proposed method employs cosine similarity as an objective function and augments the input features with the underlying model's prediction when fitting the explainer. We showed through an empirical investigation that the proposed approach outperforms the baseline, even when using a substantially smaller amount of training data and reaches the performance level of the baseline using only 15% of the training data. Moreover, we carried out an ablation study to evaluate the effect of using cosine similarity instead of MSE as a loss function, as well as the effect of augmenting the input features with the predicted outcome by the black-box model. The results indicate that using cosine similarity as an objective function and augmenting the input features significantly improve the learned explainer's performance.

A possible direction for future work is to quantify the uncertainty of the approximated Shapley values using, for instance, Venn prediction [30]. Also, validity guarantees for all the approximated scores using the conformal prediction framework can be investigated using an approach similar to conformal multi-target regression that has been proposed by Messoudi et al. [31]. Finally, the effects of additional loss functions on the fidelity of the trained explainer can be studied.

ACKNOWLEDGMENT

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

APPENDIX

VII. Information about the Used Datasets

This subsection provides a summary of the datasets utilized in the experiments. In Table V, we provide information about the used datasets including the number of features, the size of the dataset, the size of the training, validation, and test splits, and finally the ID of each dataset on OpenML.

TABLE V: The dataset information.

Dataset	Features	Size	Train. Set	Dev. Set	Test Set	OpenML ID
Abalone	8	4,177	2,672	669	836	720
Bank 32 nh	32	8,192	5,242	1,311	1,639	833
Churn	20	5,000	3,200	800	1,000	40701
Delta Ailerons	5	7,129	4,562	1,141	1,426	803
Electricity	8	45,312	28,999	7,250	9,063	151
Elevators	18	16,599	10,623	2,656	3,320	846
Higgs	28	98,050	88,245	4,903	4,902	23512
JM1	21	10,885	6,966	1,742	2,177	1053
MC1	38	9,466	6,057	1,515	1,894	1056
PC2	36	5,589	3,576	895	1,118	1069

REFERENCES

[1] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," *AI Mag.*, vol. 38, no. 3, p. 50–57, sep 2017.

[2] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & explorable approximations of black box models," *CoRR*, vol. abs/1707.01154, 2017.

[3] O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154 096–154 113, 10 2019.

[4] C. Bénard, G. Biau, S. da Veiga, and E. Scornet, "Interpretable random forests via rule extraction," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 937–945.

[5] H. Boström, R. B. Gurung, T. Lindgren, and U. Johansson, "Explaining random forest predictions with association rules," *Archives of Data Science, Series A (Online First)*, vol. 5, no. 1, pp. A05, 20 S. online, 2018.

[6] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. Polo Chau, "Cnn explainer: Learning convolutional neural networks with interactive visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1396–1406, 2021.

[7] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, *GNNExplainer: Generating Explanations for Graph Neural Networks*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[11] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 883–892.

[12] J. Yoon, J. Jordon, and M. van der Schaar, "INVASE: Instance-wise variable selection using neural networks," in *International Conference on Learning Representations*, 2019.

[13] N. Jethani, M. Sudarshan, Y. Aphinyanaphongs, and R. Ranganath, "Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations." in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 1459–1467.

[14] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath, "FastSHAP: Real-time shapley value estimation," in *International Conference on Learning Representations*, 2022.

[15] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. [Online]. Available: http://www.jstor.org/stable/1412159

[16] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 10 2001.

[17] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[18] L. S. Shapley, *A Value for N-Person Games*. Santa Monica, CA: RAND Corporation, 1952.

[19] C. Molnar, *Interpretable Machine Learning*, 2022.

[20] S. M. Lundberg, G. G. Erion, and S. Lee, "Consistent individualized feature attribution for tree ensembles," *CoRR*, 2018.

[21] M. Ancona, C. Oztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 272–281.

[22] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "L-shapley and c-shapley: Efficient model interpretation for structured data," in *International Conference on Learning Representations*, 2019.

[23] J. Teneggi, A. Luster, and J. Sulam, "Fast hierarchical games for image explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] I. Covert and S.-I. Lee, "Improving kernelshap: Practical shapley value estimation using linear regression," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 3457–3465.

[25] P. Schwab and W. Karlen, *CXPlain: Causal Explanations for Model Interpretation under Uncertainty*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[26] X. Situ, I. Zukerman, C. Paris, S. Maruf, and G. Haffari, "Learning to explain: Generating stable explanations fast," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5340–5355.

[27] A. H. A. Rahnama, J. Bütepage, P. Geurts, and H. Boström, "Can local explanation techniques explain linear additive models?" *Data Mining and Knowledge Discovery*, Sep 2023.

[28] J. Han, M. Kamber, and J. Pei, "2 - getting to know your data," in *Data Mining (Third Edition)*, third edition ed., ser. The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 39–82.

[29] F. Wilcoxon, "Individual comparisons by ranking methods. biometrics bulletin 1, 6 (1945), 80–83," 1945.

[30] A. Alkhatib, H. Boström, and U. Johansson, "Assessing explanation quality by venn prediction," in *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, ser. Proceedings of Machine Learning Research, U. Johansson, H. Boström, K. An Nguyen, Z. Luo, and L. Carlsson, Eds., vol. 179. PMLR, 24–26 Aug 2022, pp. 42–54.

[31] S. Messoudi, S. Destercke, and S. Rousseau, "Conformal multi-target regression using neural networks," in *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, ser. Proceedings of Machine Learning Research, A. Gammerman, V. Vovk, Z. Luo, E. Smirnov, and G. Cherubin, Eds., vol. 128. PMLR, 09–11 Sep 2020, pp. 65–83.