GAPS: Generality and Precision with Shapley Attribution

Brian Daley
Computer Science Dept
Columbia University
New York City, United States
brian.daley@columbia.edu

Qudrat E Alahy Ratul
Computer Science Dept.
Boise State University
Boise, United States
qudratealahyratu@u.boisestate.edu

Edoardo Serra
Computer Science Dept.
Boise State University
Boise, United States
edoardoserra@boisestate.edu

Alfredo Cuzzocrea
iDEA Lab
University of Calabria,
Rende, Calabria, Italy
alfredo.cuzzocrea@unical.it

***Abstract*—In an age of the growing use of Machine-learning, it has become an imperative task to be able to explain the processes behind the functions of many "black box" models. The explainability feature of artificial intelligence is key to building trust between humans and computers' algorithmic predictions. One of the main ways to generate this interpretability is through attribution methods, which produce importance values of each feature for a single instance in a dataset. There are many different ways of attribution for various Machine-learning models, including ones designed for specific models or "model agnostic" attribution methods—ones that do not require a specific model to achieve importance values. These attribution methods are valued because of their easily understood nature. While evaluation procedures exist such as generality and precision for rule-based explanation methods, these have not been used on attribution methods until recently. A recent experiment by Ratul et al. [1] proved that the two most popular local model-agnostic attribution methods, LIME and SHAP, have poor precision and generality. In this paper, we propose a new attribution method, the Generality and Precision Shapley Attributions (GAPS). To evaluate these models, we use the generality and precision equations used previously to evaluate the other models. We present our findings that GAPS produces higher generality and precision scores than the existing LIME and SHAP models.**

***Index Terms*—Explainable Artificial Intelligence, Interpretable Machine-learning, Attribution Methods, Generality and Precision.**

I. Introduction

Machine-learning has become a tool for businesses, governments, and enterprises alike to guide their decisions and actions in the modern world. Artificial Intelligence can be found in many aspects of people's lives, such as customer recommendations, speech recognition, virtual assistants, and more [2]. Though when it comes to understanding the intricacies of a Machine-learning model's logic and mechanisms, they still remain somewhat of a black box, hidden to the average individual and even experts in the field [3]. When the decision-making process of artificial intelligence is not fully known, there can be dire consequences in fields where many lives are on the line, for instance, aircraft collision detection systems [4]. In addition, Machine-learning models may use undesirable techniques in order to achieve their goals, such as looking at copyright symbols in an image to classify objects instead of classifying the object based on the image itself.

Just as the use of Machine-learning rises in the current era of increasing data, so does the need to be able to explain the predictions made by Machine-learning [5]–[8]. It is essential for there to be a certain level of trust between humans and machines to progress with the expansion of algorithmic predictors.

As a direct response to the black box nature of many models, scientists have started to develop explanation methods [3]. In the field of Explainable Artificial Intelligence (XAI), researchers seek to create interpretable models with predictions that can be interpreted by humans. In this way, transparency can be added to the predictions of ML models without decreasing the performance or accuracy of their results.

Governments are taking great interest in the future of the use of Machine-learning as well. The United States' Defense Advanced Research Projects Agency (DARPA) famously created an XAI program, funded by the U.S. Department of Defense (DoD) [9]. This program created an alternative "glass box" models to be used in various fields: transportation, security, medicine, finance, legal, and military. The White House Office of Science and Technology Policy (OSTP) published reports that state that it is crucial for artificial intelligence to be governable, especially so that AI can work in accordance with social values and human trust [10]. Other governments have also started publishing their plans and roadmaps towards the transparency and interpretability of AI, namely France's Strategy for Artificial Intelligence, The United Kingdom's Academy of Sciences, and AI Portugal 2030 [11]–[13]. Finally, the European Union also issued statements detailing the importance of the understandability of Machine-learning models to humans, which will also reduce bias and error [14].

Companies are also making moves in interpretable Machine-learning. As interpretable A.I.'s nature of easy understandability can be a great resource even to those who are non-experts in the XAI field, some corporations are beginning to commercialize explainable ML. One example is Google's responsible AI practices, wherein they advocate in their product lines of treating interpretability as a main aspect of a user's experience, designing their models to be inherently interpretable in nature, and ultimately communicating those explanations found by

978-1-6654-8045-1/22/ $31.00 ©2022 IEEE

the models to users [15]. Famous credit scoring service company FICO published a paper in 2018 titled "Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach," clearly also using interpretable Machine-learning in their suite [16].

Interoperability can be applied in many different fields which include: Health [17], [18], Criminal Networks [19]–[21], Privacy [22], and Cybersecurity [23]–[26].

This paper focuses on the explanations of outcomes produced by a Machine-learning model, such as a binary classification, of a specific instance in a dataset. Currently, there are two different kinds of explanation methods: attribution methods which assign values of the importance of each feature to an instance's classification, and rule-based methods. To evaluate a rule-based method of explanation, the two main metrics used are precision and generality [27]. Precision scores are based on the ideology that a rule explaining the classification of one instance should not, in turn, explain instances of the opposite classification. Generality scores are based on the ideology that a rule explaining the classification of one instance should explain other instances of the same classes as well. These two metrics are essential to human relationships with Machine-learning, as the explanations behind a Machine-learning model can then be depended upon. Until recently, those precision and generality metrics were exclusively used on rule-based methods and were not tested on attribution methods of explanation. For this experiment, the evaluation methods proposed by Ratul et al. [1] for attribution explanations were used in the experimental design.

In this paper, we propose a new reward function for attribution. The proposed method is called the Generality and Precision Shapley Attributions (GAPS). We will prove that GAPS produces attribution scores that have a higher generality and precision score in the evaluation for single instances, resulting in a more trustworthy explanation for Machine-learning models. This paper first discusses the details of the local model-agnostic attribution methods that will be evaluated as a comparison in Section II. Secondly, it discusses the evaluation techniques of attribution methods of instances in Section III. Thirdly, it discusses the design of the newly proposed attribution method that was used in the experiment in Section IV. Then, it presents the experimental design of the research, as well as the results, in Section V. Lastly, it discusses some conclusions and explorations of future works with this research in Section VII.

II. Local Model-Agnostic Attribution Methods

The appeal of a model-agnostic attribution method is that it does not require a specific Machine-learning model to be used in order to explain the predictions made. Thus, any Machine-learning model, namely Random Forest Classifiers, Support Vector Machines, Logistic Regression, etc. can be used and the local model-agnostic attribution methods can still be utilized. This also extends to local surrogate models, which, as opposed to other Machine-learning classifiers, focus on the classification of a single instance in a dataset.

In this paper, the two main local surrogate models of focus are SHAP and LIME. Both of which are attribution methods, meaning that they provide importance scores for each feature used in the prediction of a single instance.

A. LIME

Local **I**nterpretable **M**odel-agnostic **E**xplanations, otherwise known as **LIME** is a way to explain the results of a large black box model by using a local model around a single data point of interest [28]. With one data instance, LIME modifies—or perturbs—the feature values slightly and observes the slight changes in the prediction of the classifier.

The main mechanism behind LIME is the local sensitivity analysis of the features of an instance. When LIME slightly varies the feature values, it has the ability to tell how much change in predictions is due to the variation of that specific feature. Below in Figure 1 is a graph of the LIME abstraction around a single instance.

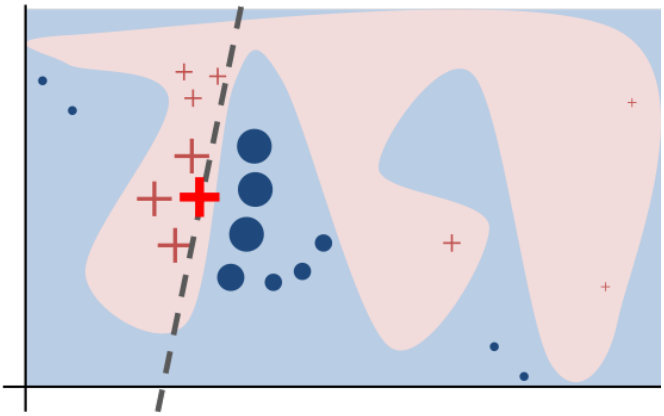

Figure 1: Lime Abstraction [28].

In this figure, the pink and blue colored sections represent the areas in which the instances would be classified as either the cross or circle class. As one can clearly see, the shape of the boundary between the pink and blue classification areas are quite nonlinear and strange, which would be a difficult task for a linear model to classify normal instances. With LIME, it solely examines one instance, in this case, the bold red cross. Then, it generates neighbors close to the point and assigns various weights corresponding to the distance from the instance. The instances close to the instance of interest are larger than the ones far away. Given these perturbations, LIME's approach can create a simple linear abstraction that works locally just to the right of the bold red cross.

The formula for LIME's explanation equation is given as follows:

$$\xi(x) = \underset{g \epsilon G}{argmin}\ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Where x is the instance of interest, f is the function of the actual Machine-learning model used for prediction, g is the surrogate model that estimates the predictions in close proximity to x, and π_x is the locality. g in this equation is

an element of potentially interpretable models, G. $\Omega(g)$ is the complexity of the function, as opposed to interpretability, and $\mathcal{L}(f, g, \pi_x)$ is a measure of local fidelity. The goal of the algorithm is to minimize the unfaithfulness of $\mathcal{L}(f, g, \pi_x)$ while maintaining that the complexity is low enough for humans to understand.

Finally, to determine the importance value of an instance's features, it is passed through a linear model such as Linear Regression. The coefficients of that linear model are then treated as the importances.

B. SHAP

SHaply **A**dditive ex**P**lanations, otherwise known as **SHAP** provides an amalgamated approach for explaining a model's prediction [29]. Like LIME, SHAP gives each feature an importance score for a specific instance. It is a combination of six different interpretation models: LIME [28], Shapley sampling values [30], DeepLIFT [31], QII [32], Layer-wise relevance propagation [33], and Shapley regression values [34]. It does this through the additive feature attribution method. To compute the attribution method, SHAP utilizes game theory with the Shapley value.

Shapley values work by using coalitional game theory to assign payouts to each feature in an instance for the total contributions. Each feature is a "player" in a coalition, and it receives a reward for its contribution to the overall prediction. In an input to a Machine-learning model f, F is the set of all features of a given instance x. $\phi_i^f(x)$ for each feature i in a coalition S is given by the following equation, which is a modification on the equation of the Shapley value:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)\,!}{F!} \left[f_c(x_{S \cup \{i\}}) - f_c(x_S) \right]$$

Where f_c is the confidence of the Machine-learning classifier of that instance, and x_S is defined as an instance where each value of a feature where the input is not an element of the coalition is substituted for the mean of all instances' values for that feature. This score calculates how pertinent the feature is to this instance as opposed to simply using the mean for the calculation of the prediction of a classifier.

One of the problems with this calculation is that the time complexity increases exponentially as the number of features increase. To avoid this complication, approximations of SHAP have been developed. One of those approximations is KernelSHAP [29], which is calculated by fitting a linear model with the following equation:

$$g(S) = \phi_0 + \sum_{j \in S} \phi_j$$

The mechanism behind the fitting technique lies in the minimization of a loss function, defined as follows:

$$L(\hat{f}, g, \pi_x) = \sum_{S \subseteq F} (g(S) - f_c(x_S))^2 \mu(S)$$

Where the kernel of the Kernel SHAP function is given by

$$\mu(S) = \frac{|F|-1}{\binom{|F|}{|S|} |S| (|F|-|S|)}$$

Such that $|F|$ is the maximum number of features and $|S|$ is the total number of features present in a coalition. This makes the SHAP calculation much faster because only a random set of samples $H \subset \{S | S \subseteq F\}$ is used in tandem with the loss function. This Kernel approach can be used with both SHAP and LIME for this experiment [35].

III. Attribution Evaluation Methods

In this section we discuss the workings of the evaluation methods used for the attribution techniques. For this, we use generality and precision, which for rule-based explanations can be found in Ribeiro et al. [27] For attribution-based techniques, we use the methods proposed by Ratul et al. [1] to calculate the precision and generality.

A. Precision

Precision scores are based on the ideology that a rule explaining the classification of one instance should not, in turn, explain instances of the opposite classification. Therefore, if there is a rule whose conditions are satisfied by two different instances, but their classification is different, its faulty nature would not be viewed favorably by humans. Ribeiro et al. [27], articulates that the precision of an explanation rule r with a classification a is given by the inverse of the percentage of instances that are classified as the opposite class by the same rule r.

For the purposes of this experiment, two functions were used to find the attribution precision. The first is $sel(X, x)$, which outputs a vector in $\mathbb{R}^{|S|}$ which gives a binary vector to determine which features are present in a coalition. The formal definition is given as $S = i_1, \ldots, i_k$ the subset of features for each $j \in \{1, \ldots, k\}$, $sel(S, x)[j] = x[i_j]$. The second function is based on the att attribution vector which returns the top-k feature, given by the equation $top_k : \mathbb{R}^n \to 2^{|k|}$.

Given classification outcomes a where I_a is the set of instances that have classification outcome and $I_{\neg a}$ is the set of instances that do not have classification outcome a, the following equation yields the inverse of the attribution precision, where $S_{at}^x = top_k(att_x)$:

$$RP^k(x, att_x) = \frac{|\{\hat{x} | \hat{x} \in I_{\neg a}, sel(S_{at}^x, x) = sel(S_{at}^x, \hat{x})\}|}{|I_{\neg a}|}$$

The reverse precision is a measure of the number of instances that share the same value in top_k with another instance that have the same outcome a. Obviously, the average reverse precision score would be given by the following equation:

$$avgRP^k(I_a) = \frac{\sum_{x \in I_a} RP^k(x, att_x)}{|I_a|}$$

Where the reverse precision is calculated for all instances x and is then divided by the total number of instances with classification a.

B. Generality

Generality scores are based on the ideology that a rule explaining the classification of one instance should explain other instances of the same classes as well. Generality in this sense would take in two instances, x_1 and x_2, and their attribution vectors att_1 and att_2 and find the number of top_k features that are shared between the two instances. The equation for the common features between the two are given below:

$$common_k(att_{x_1}, att_{x_2}) = |top_k(att_1) \cap top_k(att_2)|$$

The equations to find the top_k features are defined in the Precision section. With an instance x that is an element of I_a, the generality of the attribution of x, att_x is calculated by using the top-h neighbor instances from the same classification with the function $generality(x, k, h, agg)$. The equation for this is defined as follows:

$$agg(\{common_k(att_x, att_{\hat{x}}) | \hat{x} \in topNeighbour_h(x, I_a)\})$$

Where $agg \in \{sum, \min, \max\}$ and $topNeighbour_h(x, I_a)$. The $topNeighbor_h$ function defines the top-h neighbor instances as mentioned previously from instances of the same class of x. $common_k$ in this function does not consider the actual feature values from each instance, because by using the top-h neighbors, one can assume that the feature values would be similar. The agg function also gives more insight into the generality because a human can see how the similarities of the attributions are distributed for each instance. Like the average precision function $avgRP^k(I_a)$, the average generality function can be defined similarly as follows:

$$avgGen(I_a, k, h, agg) = \frac{\sum_{x \in I_a} generality(x, k, h, agg)}{|I_a|}$$

IV. GAPS Design

In this paper, we propose a new attribution method, called **GAPS**, which stands for **G**enerality **A**nd **P**recision **S**hapley Attributions. As one can guess from the title, the goal of creating GAPS was to increase the precision and generality values of existing attribution methods. A recent study has shown that the precision and generality values of SHAP and LIME are quite poor, and show results that would be unfavorable.

Given an instance x from a dataset and a coalition s which constituted a binary vector of values depending on which features are present or not in a coalition, the equation for the reward function of an instance's coalition is given as follows:

$$f(s, x) = \left[\begin{array}{c} E_{l \sim m(s,x)}[c(l)] + \\ + \sum_{z \in N(x,s,a)} \frac{\lambda_G c(z)}{|N(x,s,a)|} + \sum_{z \in N(x,s,\neg a)} \frac{\lambda_P (c(z)-1)}{|N(x,s,\neg a)|} \end{array} \right]$$

Where λ_G and λ_P are some constants to scale the generality and precision quantities, respectively, to increase either evaluation score. N is defined as the set of neighbors close to instance x where the features that are in the coalition are not perturbed, and the features that are not in the coalition are perturbed. They are perturbed in accordance with using a normal distribution with the mean as the feature value of x and the standard deviation as the standard deviation of the feature value across all instances in the dataset. $N(x, s, 1)$ is the set of neighbors which are predicted using the same classifier to be in class one, and $N(x, s, 0)$ to be in class zero. $c(z)$ is defined as the confidence of the Machine-learning model of the neighbor z which is an element of N in their respective summations.

Finally, $E_{l \sim m(s,x)}[c(l)]$ is the expected value of the confidence of the Machine-learning classifier of many l which are sampled from $m(x, s)$ to create data points where once again the features that are in a coalition will remain unchanged, but the features that are not will be replaced with a random feature value from an instance across all points in the dataset.

The main idea behind this equation is to create a trade-off between generality and precision scores, where $lambda_g$ and $lambda_p$ can be manipulated to adjust the prevalence of the generality and precision scores in the evaluation. This creates a more trustable model to humans than other attribution equations because it can produce a higher generality and precision score than other models such as LIME or SHAP.

Such as LIME and SHAP, the coalitions and the rewards from the reward function will be passed through a linear Machine-learning model, such as Linear Regression. The sample weight of the Linear Regression is the same kernel used in Kernel SHAP, $\mu(s) = \frac{|F|-1}{\binom{|F|}{|S|}|S|(|F|-|S|)}$. The coefficients from the Linear Regression model for each feature are then treated as the explanation values.

V. Experiment

In this section, we discuss the design of the experiment that we used to carry out the calculations of precision and generality on the local model-agnostic attribution methods for a dataset, as well as the results from the experiment. The design of this experiment was modeled after that of Ratul et al. [1], and the results for the SHAP and LIME procedures are from their research as well.

A. Experimental Design

The dataset used in this experiment was the UNSW-NB15 Dataset [36]. Created by the Cyber Range Lab of UNSW Canberra, Australia, this dataset contains raw network packet data that simulates modern network traffic, including a mix of both real normal network activities and synthetic attack behaviors. There are nine kinds of attacks used in the dataset, with 2,540,044 records and each instance contains a total of 49 features. The collection period of this data was for 16 hours on 2015, January 22, and 15 hours on 2015, February 17.

Of the 49 features, 3 were categorical, and these were transformed to be used by the Machine-learning models with one-hot encoding. The data was split with 70% in the training data and 30% in the testing data. Machine-learning classifiers were then used on the data and the precision, recall, and f1-score was evaluated for each model: Logistic Regression, Random Forest, K-Nearest Neighbors, and Support Vector Classification. The attribution methods, and therefore the precision and generality evaluations were solely performed

Table I: Average Generality for LIME & SHAP with a Varied Number of k Features & h Neighbors from the "UNSW-NB 15" Dataset.

		Mean LIME intersection size			Mean SHAP intersection size		
		Max	**Mean**	**Min**	**Max**	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**						
1	**1**	1.00	0.37	0.00	1.00	0.84	0.00
	5	5.00	1.83	0.00	5.00	4.11	0.00
	10	10.00	5.46	2.00	10.00	8.64	3.00
5	**1**	0.20	0.01	0.00	1.00	0.84	0.00
	5	3.20	1.81	0.80	5.00	3.91	0.80
	10	7.40	5.44	4.00	10.00	8.52	3.00
10	**1**	0.20	0.01	0.00	1.00	0.84	0.00
	5	3.10	1.81	0.50	5.00	3.83	0.90
	10	7.20	5.44	2.50	10.00	8.42	4.20

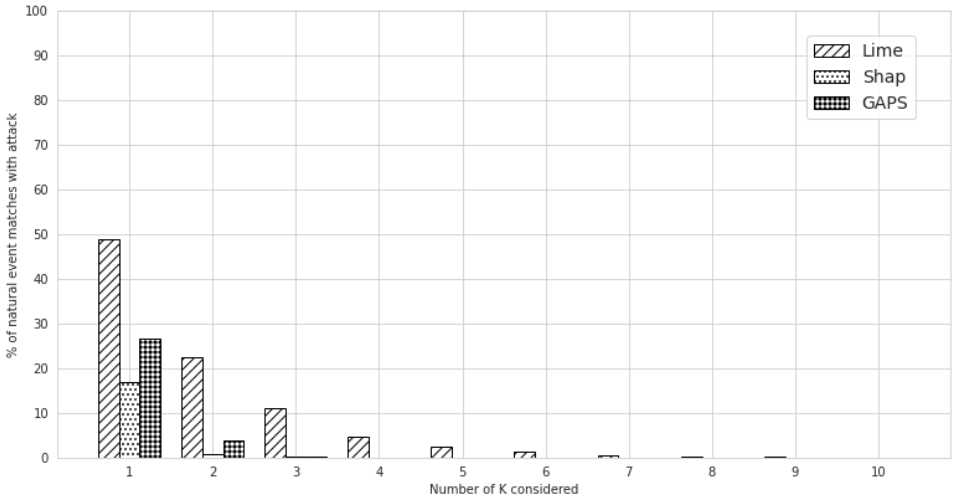

Figure 2: Average Precision Percentage for LIME, SHAP, & GAPS with a Varied Number of k Features from the "ICS: Power System" Dataset.

Table II: Average Generality for LIME & SHAP with a Varied Number of k Features & h Neighbors from the "ICS: Power System" Dataset.

		Mean LIME intersection size			Mean SHAP intersection size		
		Max	**Mean**	**Min**	**Max**	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**						
1	**1**	1.00	0.01	0.00	1.00	0.39	0.00
	5	2.00	0.22	0.00	5.00	2.09	0.00
	10	4.00	0.83	0.00	10.00	4.41	0.00
5	**1**	0.20	0.01	0.00	1.00	0.34	0.00
	5	1.00	0.22	0.00	4.20	1.69	0.00
	10	2.20	0.85	0.00	9.20	3.65	0.40
10	**1**	0.10	0.01	0.00	0.90	0.32	0.00
	5	0.80	0.21	0.00	3.70	1.57	0.00
	10	1.90	0.87	0.20	8.40	3.36	0.40

on the Random Forest Classifier since the computation time would be the most efficient.

B. Experimental Results

The results for this experiment are in two formats: firstly is a bar graph of the average reverse precision ($avgRP^k(I_a)$) when varying the number of k features considered from each dataset for LIME, SHAP, and GAPS; secondly is a chart of the average generality ($avgGen(I_a, k, h, agg)$) when varying the number of k features considered and the number of h close features from each dataset for LIME, SHAP, and GAPS.

Table III: Average Generality for GAPS with a Varied Number of k Features & h Neighbors from the "ICS: Power System" Dataset.

		Mean GAPS intersection size		
		Max	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**			
1	**1**	1.00	0.15	0.00
	5	4.00	0.76	0.00
	10	7.00	1.82	0.00
5	**1**	0.80	0.16	0.00
	5	2.00	0.78	0.00
	10	4.00	1.87	0.20
10	**1**	0.60	0.16	0.00
	5	1.90	0.79	0.00
	10	3.70	1.86	0.30

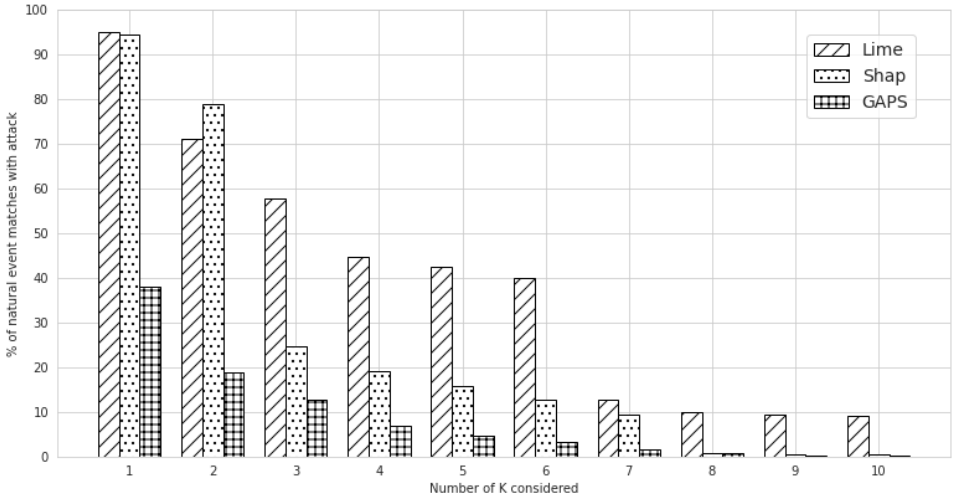

Figure 3: Average Precision Percentage for LIME, SHAP, & GAPS with a Varied Number of k Features from the "UNSW-NB 15" Dataset.

Table IV: Average Generality for GAPS with a Varied Number of k Features & h Neighbors from the "UNSW-NB 15" Dataset.

		Mean GAPS intersection size		
		Max	**Mean**	**Min**
No of Neighbors (h)	**No of Features (k)**			
1	**1**	1.00	0.68	0.00
	5	5.00	3.91	0.00
	10	10.00	8.27	1.00
5	**1**	1.00	0.65	0.00
	5	5.00	3.70	0.00
	10	10.00	7.95	1.00
10	**1**	1.00	0.62	0.00
	5	5.00	3.61	0.30
	10	10.00	7.82	1.30

VI. Acknowledgments

This research was supported by the National Science Foundation awards NSF REU #1820685 and #CCF 1950599, and Idaho Global Entrepreneurial Mission/Higher Education Research Council #IGEM22-001.

VII. Conclusions

When viewing preliminary results, it is clear that GAPS outperformed the existing attribution methods of LIME and SHAP when it comes to the "UNSW-NB 15" dataset, as evidenced by the lower average reverse precision scores in each varied number of k features. The average generality for GAPS additionally outperformed LIME and SHAP in the same dataset, with higher intersections on balance. However, in terms of the "ICS: Power System" dataset results, the graphs and charts show that GAPS outperformed LIME but did not outperform SHAP. The results show that GAPS has a higher average reverse precision score than SHAP and a lower average than LIME, and a lower average generality on balance than SHAP but higher than LIME.

It is clear that the GAPS method of Generality and Precision with Shapley Attribution shows promise in terms of creating better attribution techniques. Future research is necessary into this novel method to better improve the scores and outperform both LIME and SHAP across the board, rather than in one dataset or one technique rather than the other.

References

[1] Q. E. Alahy Ratul, E. Serra, and A. Cuzzocrea, "Evaluating attribution methods in machine learning interpretability," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5239–5245.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[3] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[4] S. Temizer, M. Kochenderfer, L. Kaelbling, T. Lozano-Pérez, and J. Kuchar, "Collision avoidance for unmanned aircraft using markov decision processes," in *AIAA guidance, navigation, and control conference*, 2010, p. 8040.

[5] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[6] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[7] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley, "Google vizier: A service for black-box optimization," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1487–1495.

[8] C. Rudin, "Please stop explaining black box models for high stakes decisions," *arXiv preprint arXiv:1811.10154*, vol. 1, 2018.

[9] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, no. 2, 2017.

[10] P. Press, "Preparing for the future of artificial intelligence," 2016.

[11] C. Villani, "French national strategy for artificial intelligence." 2019. [Online]. Available: https://www.aiforhumanity.fr/en/

[12] "Portuguese national initiative on digital skills. ai portugal 2030. 2019." 2019. [Online]. Available: https://www.incode2030.gov.pt/sites/default/files/draft_ai_portugal_2030\v_18mar2019.pdf

[13] "Machine learning: The power and promise of computers that learn by example." 2019. [Online]. Available: https://royalsociety.org/topics-policy/projects/machine-learning/

[14] "European commission. algorithmic awareness-building. 2018." 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/algorithmic-awareness-building

[15] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[16] G. Fahner, "Developing transparent credit risk scorecards more effectively: An explainable artificial intelligence approach," *Data Anal*, vol. 2018, p. 17, 2018.

[17] A. Shrestha, E. Serra, and F. Spezzano, "Multi-modal social and psycholinguistic embedding via recurrent neural networks to identify depressed users in online forums," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1–11, 2020.

[18] J. Souza, C. K. Leung, and A. Cuzzocrea, "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics," in *International conference on advanced information networking and applications*. Springer, 2020, pp. 669–680.

[19] M. Joaristi, E. Serra, and F. Spezzano, "Inferring bad entities through the panama papers network," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 767–773.

[20] ——, "Detecting suspicious entities in offshore leaks networks," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–15, 2019.

[21] B. Daley, E. Serra, and A. Cuzzocrea, "Identifying malicious users in the offshore leaks networks via structural node representation learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 5095–5101.

[22] E. Serra, A. Shrestha, F. Spezzano, and A. Squicciarini, "Deeptrust: An automatic framework to detect trustworthy users in opinion-based systems," in *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, pp. 29–38.

[23] S. Khamaiseh, E. Serra, Z. Li, and D. Xu, "Detecting saturation attacks in sdn via machine learning," in *2019 4th International Conference on Computing, Communications and Security (ICCCS)*. IEEE, 2019, pp. 1–8.

[24] S. S. Das, E. Serra, M. Halappanavar, A. Pothen, and E. Al-Shaer, "V2w-bert: A framework for effective hierarchical multiclass classification of software vulnerabilities," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021, pp. 1–12.

[25] S. Khamaiseh, E. Serra, and D. Xu, "vswitchguard: Defending openflow switches against saturation attacks," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2020, pp. 851–860.

[26] A. Rullo, E. Serra, E. Bertino, and J. Lobo, "Shortfall-based optimal placement of security resources for mobile iot scenarios," in *European Symposium on Research in Computer Security*. Springer, 2017, pp. 419–436.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[28] ——, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[29] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[30] S. Kaufman, S. Rosset, and C. Perlich, "Leakage in data mining: formulation, detection, and avoidance," in *KDD*, 2011.

[31] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3145–3153.

[32] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 598–617.

[33] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[34] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[35] S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," *arXiv preprint arXiv:1706.06060*, 2017.

[36] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.