

A Trustworthy View on Explainable Artificial Intelligence Method Evaluation

Ding Li ^{ID}, Yan Liu ^{ID}, Jun Huang, and Zerui Wang ^{ID}, Concordia University

In this article, we follow a process of explainable artificial intelligence (XAI) method development and define two metrics in terms of consistency and efficiency in guiding the evaluation of XAI explanations.

Explainable artificial intelligence (XAI) practices perform an interpretive analysis of the input data. They produce an approximation summary as an explanation relating inputs and outputs, without engaging the internal representations, attributes, and structures of the learning models.¹ XAI is emerging as a common goal for data scientists, engineers, and AI practitioners to deal with the problem of AI opacity on the purpose and server and how the models work. Critically, Babic et al.² have highlighted the need for an explainable method to be trusted in the health-care domain. In particular, a trustworthy XAI method should exhibit some robustness, which means the XAI

method should produce similar explanations for similar inputs.²

The discrepancy in the explanation summary leads to ambiguity in understanding the machine learning prediction. This becomes the question of whether the explanations from different XAI methods are trustworthy. Therefore, the consistency of these summaries becomes essential for the trustworthy and accountable assessment of machine learning models.

One source of variation of explanation summaries originates from the XAI operations. The major entities involved in XAI operations are datasets, a trained model, and XAI methods. An XAI method can be applied to the model prediction on each data sample. As a result, a set of explanation summaries are collected for the same prediction model. One XAI method may demonstrate different levels

of similarity among the explanations of each data sample. In addition, multiple XAI methods may yield explanations in variation from one another on the same dataset and model. We describe one example scenario of XAI explanation variation in the case of code vulnerability analysis in the “Observing Explanation Consistency” section.

A conventional approach is to compare several XAI methods’ explanation summaries and make decisions among the available results, based on how consistent the explanation summaries are. If the decision cannot be reached with majority voting, developing a new XAI algorithm becomes necessary. A proposed XAI algorithm needs to define new metrics that measure feature contributions from a perspective not fully addressed by state-of-the-art XAI methods.

In addition to explanation summary consistency, state-of-the-art XAI methods vary in an extensive range of runtime delays inherently due to their intrinsic algorithms of computing features’ contributions. One target of a new XAI algorithm is to reduce the time complexity, with a consistency level comparable to state-of-the-art XAI methods. Coherently, the problem of evaluating existing XAI methods and developing an XAI alternative converges to the question, What are the criteria and core activities in evaluating XAI methods toward consistent and robust explanation?

In this article, we demonstrate XAI engineering in evaluating different XAI methods with well-defined metrics, namely, consistency and runtime efficiency. We further define the measure of consistency by using the distances among explanation summaries. The runtime efficiency is based on asymptotic analysis in terms of

the number of features and size of the data samples. We present two evaluation techniques: 1) comparing with a baseline model and 2) performing cross validation among multiple models. Based on the consistency evaluation of state-of-the-art XAI methods, we develop a new model-agnostic method for the XAI taxonomy, called the *mean centroid prediction difference (PredDiff)*. Together with the other nine XAI methods, we evaluate the consistency and time efficiency of the mean centroid PredDiff on three domain examples, including image classification, code vulnerability detection, and search-based ranking. We demonstrate a working path of systematically evaluating and making decisions about an XAI method from the trustworthiness point of view.

RELATED WORK

XAI has been an emerging research topic in recent years, aiming to explain AI models’ logic and decision-making processes for users in the interest of safety and fairness. Conventionally, post hoc XAI methods are categorized as model agnostic and model specific.¹ Model-specific methods probe and extract model gradients and neuron activation states from neural network models. Examples include the family of XAI methods based on class activation mapping (CAM),³ including EigenCAM,⁴ GradCAMElementWise,⁵ Grad-CAM++,⁶ XGrad-CAM,⁷ and Hi-ResCAM.⁸ They have been applied to explain feature contributions to image classification algorithms and tasks. Existing XAI work has applications in various domains. Related to the case studies in this article, the work of Singh et al.⁹ introduces model-agnostic Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP)

to compare ranking models. The study defines completeness and validity measurements of ranking models.

Model-agnostic methods are black box based and nonintrusive to specific machine learning algorithms.⁹ The PredDiff¹⁰ and causal explanations¹¹ analyze the PredDiff by masking an individual data feature or a group of input data features. LIME¹² leverages a weighted regularized linear model to interpret the input representation. SHAP¹³ utilizes the Shapley value¹⁴ to calculate feature contribution values.

The robustness of XAI methods in terms of the consistency of their outputs is essential for adopting algorithmic explanations to enhance the trust and accountability of AI.¹⁵ The trust can be based on an individual prediction by users. The XAI explanation summary related to this level of trust indicates the robustness of the explanation for each data sample prediction. Based on the prediction trust, a user may accept the trust at the overall model level. Accordingly, the consistency across multiple XAI methods on the same model relates to the accountability of AI models. At both levels, existing work¹⁶ highlights quantified metrics that have been defined to measure XAI explanation results toward the progress of developing trustworthy AI.

OBSERVING EXPLANATION CONSISTENCY

We introduce an observation of explanation consistency across multiple XAI methods. In software code vulnerability detection, understanding how code features affect the accuracy of vulnerable code classification helps reduce vulnerability risks¹⁷ and enables automated corrections on vulnerable code.¹⁸ We have trained an XLNet model¹⁹ as a classifier to classify vulnerable

software code at the method level to different common weakness enumeration (CWE) types. The feature masking is configured in three types, namely, 1) code only, involving program code without comments and import statements; 2) comment only; and 3) import only, involving only import statements. An XAI method explains the feature importance of *code*, *comment*, and *import statement* in terms of their contributions to code vulnerability classification. We eliminate the CWE label tokens in the code comments to avoid training the machine learning model so that it will not “remember” the CWE labels.

Three model-agnostic XAI methods are applied to the preceding feature masking scenarios, namely, the PredDiff,¹⁰ Shapley value,¹⁴ and KernelSHAP.¹³ Our results show that both the Shapley value and KernelSHAP rank

the feature importance in descending order as *comment*, *code*, and *import statement* on public datasets of Juliet (<https://samate.nist.gov/SARD/test-suites/111>) and the Open Web Application Security Project (OWASP) (<https://owasp.org/www-project-benchmark/>). However, the PredDiff ranks in the order of *code*, *import*, and *comment* on the OWASP dataset. We observe the difference in the explanation summary among the XAI methods and datasets. Furthermore, we measure the time consumption of running three XAI methods against the same XLNet model on two datasets. The Shapley value and KernelSHAP consume approximately three times and 20 times as much time as the PredDiff, respectively. The complete results are available at GitHub (<https://github.com/DataCentricClassificationofSmartCity/Mean-Centroid-PredDiff>).

In this case, one can select one of the three XAI methods as the baseline model to draw further conclusions about the XAI explanation consistency. One option is selecting the Shapley value as the baseline. Shapley values are initially created to assign attributions to specific participants in coalition games. Shapley values have been adopted for explaining machine learning models since they have the properties of efficiency, symmetry, dummy variables, and linearity.¹³ Alternatively, a new XAI method can be developed to cross validate the existing explanation summary.

EVALUATE XAI EXPLANATIONS

The preceding discussion motivates a unified process of evaluating the existing XAI methods and guiding the

EXPLANATION OF FEATURES

Explainable artificial intelligence (XAI) methods that derive explanations via features include masking-based methods and mutation-based methods. Feature masking-based methods remove certain features or set the features with default values.²⁰ Then, the output prediction is evaluated. On the other hand, mutation-based methods assign possible input values to the model and then obtain prediction.⁵¹ The feature influence is measured by inputting the black box models with feature masking and mutation. These methods then measure prediction changes compared to the original model and inputs. These methods vary from each other in terms of

- 1) feature masking and mutation techniques and

- 2) summary techniques to compute the feature importance.²⁰ Other XAI methods, such as explain by visualization, apply digital patterns, plots, and heatmaps to explain the feature classification and localization.⁵²

REFERENCES

- S1. C. Molnar, “Interpretable machine learning,” Lulu, 2020. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- S2. K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” 2014, *arXiv:1312.6034*.

development of a new XAI method. Both activities share core activities, as presented in Figure 1.

The process starts with the task of setting XAI goals and criteria. Survey works^{1,20} provide taxonomies and classifications to select candidate XAI methods that target the same goals. The follow-up steps carry out the measurements on defined metrics directly related to decision-making factors in XAI. In this article, we focus on the algorithmic complexity and consistency of the explanation summary in feature masking and feature removal.

Define consistency for feature-based XAI explanation

Consider $\hat{f}(x^{[l]})$, the model prediction on instance $x^{[l]} = \langle x_1^{[l]}, x_2^{[l]}, \dots, x_p^{[l]} \rangle$, where p is the number of features. Suppose S is the subset of all the features by masking or removing a feature j ; that is, $S \subseteq \{1, 2, 3, \dots, p\} \setminus \{j\}$ and P contains the whole features: $P = S \cup \{j\}$. Under feature masking, the prediction on the masked feature set S and whole feature set P for each instance x has the difference $\delta_j^{x^{[l]}} = \hat{f}_S(x^{[l]}) - \hat{f}_P(x^{[l]})$. Hence, the feature contribution to the payout by masking feature j on the prediction of instance $x^{[l]}$ is defined as a function, $\phi_j(\delta_j^{x^{[l]}})$. An XAI method develops the aggregation of $\phi_j(\delta_j^{x^{[l]}})$ on all the data samples differently. Finally, by masking the features one by one, the feature importance order is derived by ranking the feature contribution values.

After the transformation from feature contribution values to the feature importance order, the Kendall tau ranking distance²¹ is applied to measure the distance of any two pairs of the XAI method's explanation results.

Analyze time complexity

Asymptotic analysis for $\phi_j(\delta_j^x)$ depends on the size of the data instances number N and number of features P . The Shapley value computes the feature value difference under feature masking δ_j^x for the whole dataset for each masked feature. The Shapley value considers the permutation when selecting one feature to mask and makes the reverse value of the permutation the weight to sum the feature contribution value ϕ_j . Overall, we derive that the Shapley value has the complexity $\Theta(N \times P \times 2^P)$. KernelSHAP¹³ uses the linear LIME explanation model and classical Shapley value. According to the definition, KernelSHAP depends on the LIME loss function,¹² weighting kernel, and regularization term. Therefore, KernelSHAP has the complexity $\Theta(N \times (2^P + P^3))$. The PredDiff removes

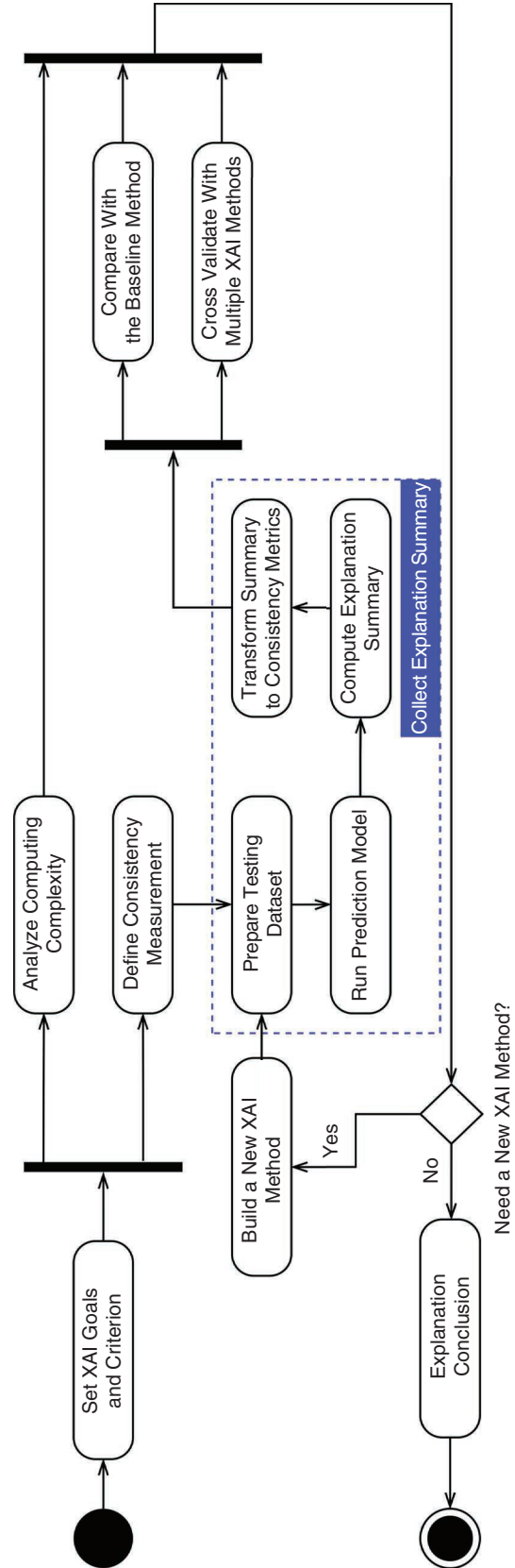


FIGURE 1. The core development activities of the trustworthy evaluation of XAI methods.

each feature individually and measures the difference between each instance's prediction and the feature removal prediction. The time complexity of the PredDiff is $\Theta(N \times P)$. In the "Develop a New XAI Method" section, we present a newly proposed XAI method with the time complexity of $\Theta(N \times P^2)$.

Collect explanation summary

The evaluation of the XAI explanation summary begins with the configuration of the test dataset into several versions, including the whole features and each subset of masked features. The machine learning model runs on the configured test dataset to output $\delta_j^{x^{[i]}}$ the prediction. Next, each XAI method computes the feature contribution $\phi_j(\delta_j^x)$ by aggregating all the data samples.

Measure explanation consistency

The evaluation process in Figure 1 provides a guideline concentrating on the

consistency of the explanation summary as a primary attribute to decide the selection of the XAI method and development of a new XAI method. The computing complexity of an XAI method is another additional attribute of decision making. At the condition checking point, XAI practitioners decide to use the existing XAI method or build a new one. In both cases, the consistency definition enables a unified understanding of XAI methods in three phases, including 1) computing the difference of prediction under the feature configuration, 2) computing the feature contribution value based on the PredDiff, and 3) converting the contribution values to the explanation. A distance metric, such as the Kendall tau ranking distance, is applied to measure the distance between two explanation summaries. The larger the distance value, the less consistent the two explanations are.

DEVELOP A NEW XAI METHOD

We aim to explain the effects of feature masking by the relative difference in the ratio to the prediction without feature masking. The state-of-the-art methods consider the absolute PredDiff. The objective of this new XAI method is to achieve consistency of explanation summaries comparable to the state-of-the-art methods and reduce computing time consumption. Figure 2 describes the core tasks of computing the PredDiff under feature masking and the feature contribution value for each masked feature in three phases.

Phase 1: Compute the PredDiff under feature masking

Algorithm 1 shows that the PredDiff $\delta_j^{x^{[i]}}$ (as the x coordinate) and its corresponding prediction $\hat{f}_P(x^{[i]})$ (as the y coordinate) form a data point in the 2D Euclidian plane. Hence, N numbers of 2D points are created for each masking feature j .

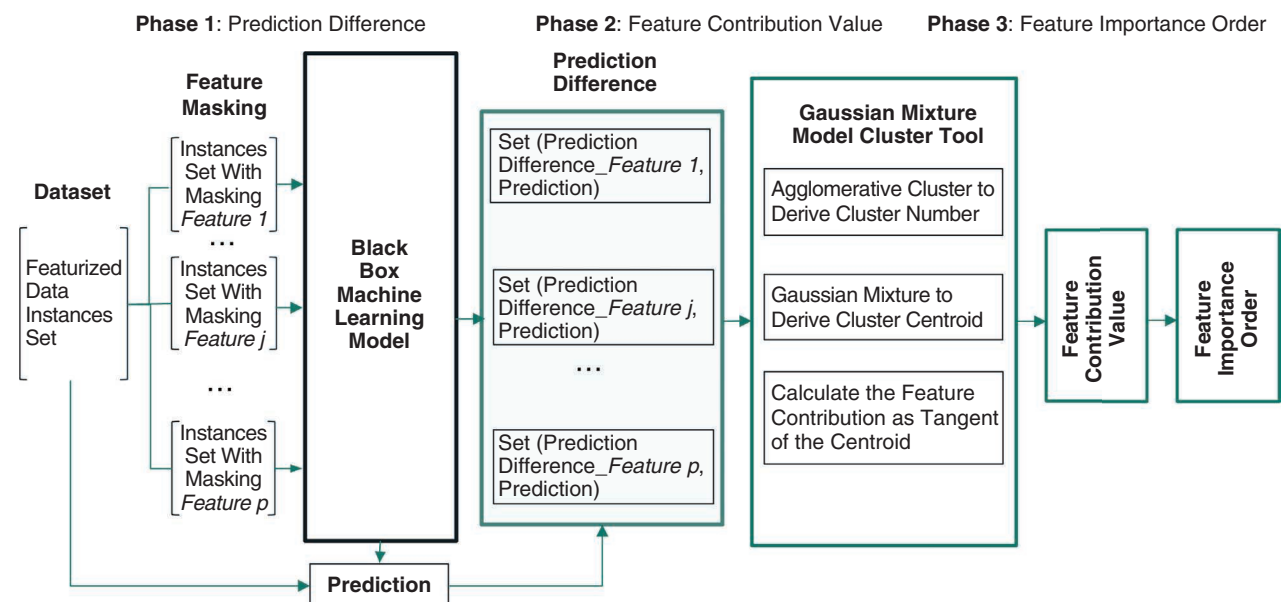


FIGURE 2. The dataflow of the mean centroid PredDiff.

Phase 2: Compute feature contribution values

We observe from the phase 1 output that the data points form clusters. We further group the data points into k_j numbers of clusters by the agglomerative clustering algorithm.²² We then estimate the centroid data point of these clusters, using a Gaussian mixture model.²³ For each masked feature j , we define its feature contribution value ϕ_j aggregated for all the input data samples as the slope, or tangent, of the centroid data point to the origin point in a 2D plane. An example in Figure 3 depicts how the Gaussian mixture clusters aggregate the contribution values of two feature markings. Data points are grouped into two clusters for each feature. The centroid data point is derived as the weighted average by the clusters' density points generated by the Gaussian mixture model. This algorithm has considered the distribution density of the prediction changes of all the data samples.

Phase 3: Convert to feature importance order

The conversion is simply ranking the features in descending order according to their feature contribution value. The consistency of the two explanations is then measured as the distance between two orders.

Asymptotic analysis of time complexity

Given the number of features P and number of instances N , computing the PredDiff is based on the time complexity $\Theta(N \times P)$ in phase 1. In phase 2, computing the clusters takes $\Theta(N \times P^2)$. Overall, the time complexity is $\Theta(N \times P^2)$.

CASE STUDY ON IMAGE CLASSIFICATION

The first case study evaluates the explanation summaries of image classification

in face mask detection. The mean centroid PredDiff method is cross validated with six state-of-the-art model-specific XAI methods. The open source pretrained ResNet50²⁴ is applied to detect the face mask categories from images. The dataset

(https://github.com/youyinnn/ai_face_mask_detection_project.git) contains 2,630 images with five different labels: wearing an N95 mask, wearing a cloth mask, wearing a surgical mask, mask worn incorrectly, and no mask.

ALGORITHM 1: MEAN CENTROID PREDICTION DIFFERENCE EXPLANATION.

- ▷ \hat{f}_{agg} , the agglomerative clustering algorithm²²
- ▷ \hat{f}_{gmm} , the Gaussian mixture model²³
- ▷ k_j , the number of clusters under feature masking j
- ▷ *Centroid* as the cluster centroid point
- ▷ $S \subseteq \{1, 2, 3, \dots, P\} \setminus \{j\}$, the subset of all the features by masking or removing a feature j
- ▷ P , the whole features, $P = S \cup \{j\}$

Input: Input dataset X , full feature set P , masking feature set S , and model prediction $\hat{f}(x^{[i]})$

/ Phase 1: Compute the difference of prediction under feature configuration*/*

for all $j \in P$ **do**

for all $x_i \in X$ **do**

$$\delta_j^{x^{[i]}} \leftarrow |\hat{f}_S(x^{[i]}) - \hat{f}_P(x^{[i]})|$$

$$\nu_j^{x^{[i]}} \leftarrow \langle \delta_j^{x^{[i]}} \rangle, -\hat{f}_P(x^{[i]})|$$

end for

end for

$$V_j \leftarrow \{\nu_j^{[1]}, \nu_j^{[2]}, \dots, \nu_j^{[N]}\}$$

/ Phase 2: Compute the feature contribution value*/*

/ Group V_j to k_j clusters */*

$$k_j \leftarrow \hat{f}_{agg}(V_j)$$

/ Derive the centroid of k_j clusters */*

$$centroid_j \leftarrow \hat{f}_{gmm}(k_j, V_j)$$

/ Compute the contribution value as the tangent of the centroid data point in 2D coordinates */*

$$\phi_j(\delta_j^X) = \tanh(centroid_j)$$

/ Phase 3: Convert the contribution values to the feature importance orders */*

$$order = \text{sort}(\text{abs}(\phi_j(\delta_j^X)))$$

Output: $\phi_j(\delta_j^X)$, *order*

Applying mean centroid PredDiff to image explanation

As illustrated in Figure 4, we generate a kernel masking matrix to mask the pixels iteratively by filling in zeros. We then obtain $(l \times l) / (n \times n)$ masked images for the model prediction, where the image size is $(l \times l)$ and the kernel masking matrix has size $(n \times n)$. The mean centroid PredDiff summarizes the pixel feature contributions from the PredDiff between the original image and the masked ones. In the experiment, l is 256, and we take kernel masking matrix size $n = 8$.

Explanation evaluation analysis

Six XAI methods are selected for explaining the saliency map of the input images, including Grad-CAM²⁵, EigenCAM,⁴ GradCAMElementWise,⁵ Grad-CAM++,⁶ XGrad-CAM,⁷ and HiResCAM.⁸ The saliency map explanation shows the active area of the image that contributes to the model's prediction.

Consistency observation. CAM-based methods compare the prediction change between the original image and masked image by the saliency map. The mean centroid PredDiff summarizes the prediction change from kernel-based image masking.

Figure 5 displays the prediction change distance distribution of 2,630 images. EigenCAM has the longest distributed range compared to other methods. This indicates that EigenCAM varies the most in explaining the feature contributions of the 2,630 images. In contrast, the mean centroid PredDiff plot has the lowest range. This shows that the mean centroid PredDiff method is more consistent across all the images.

Time complexity analysis. The mean centroid PredDiff has the time complexity of $\Theta(N \times P^2)$, given the number of images N and number of features P . In this case, an image with a masking kernel matrix is counted as one feature. Hence, $P = (l \times l) / (n \times n)$ is the number of features.

CASE STUDY ON CODE VULNERABILITY DETECTION

Referring back to the discussion in the "Observing Explanation Consistency" section, we observe that the explanation produces different feature importance orders from three state-of-the-art methods. We reevaluate the XAI methods by adding the mean centroid PredDiff method. Table 1 shows that the PredDiff and mean centroid PredDiff have the same feature importance order. In both values, *code* is the most important feature. The importance order of *comment* and *import statement* varies from the Juliet and OWASP datasets. The Shapley value and KernelSHAP share consistent results but value *comment* more than *code* and *import statement*. Security experts can make further decisions on XAI methods based on the preceding explanation.

CASE STUDY ON SCHOLAR SEARCHING RANKING SYSTEM

This case study attempts to explain the feature influence of an open source semantic scholar search (S2Search) ranking model.²⁶ S2Search provides a prediction tool to output a ranking score for each scholarly article, given a query keyword and list of features. The arXiv dataset, selected from Kaggle (<https://www.kaggle.com/datasets/Cornell-University/arxiv>), has 40 categories and a total of 542,877 articles. An article contains six relevant features: title, authors, abstracts, citation numbers, venue, and publication year. Each category is performed as a single dataset.

Cross validation of consistency

Across-datasets comparison. The median contribution values sort

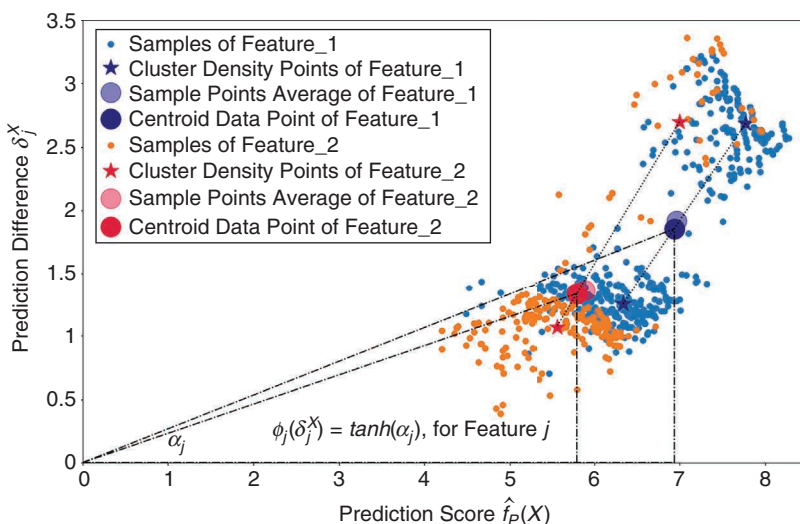


FIGURE 3. The derivation of two features' contribution values via Gaussian mixture clusters.

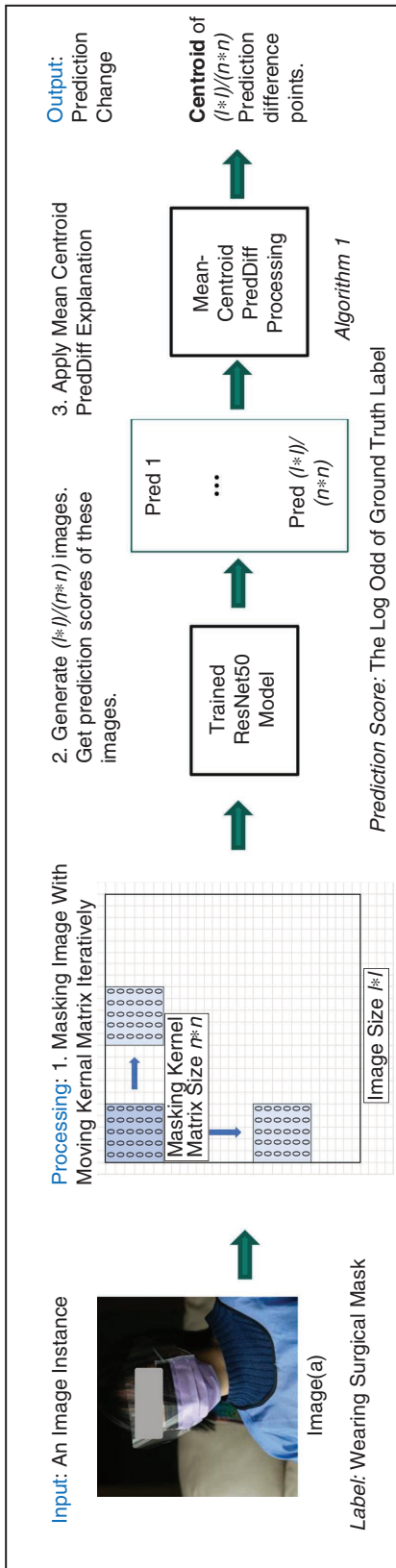


FIGURE 4. The mean centroid PredDiff process in image explanation.

the feature importance order of an XAI method from 40 datasets. We measure the Kendall Tau Ranking Distance (KTRD) distance between this aggregated feature importance order and the orders of 40 datasets as the across-datasets comparison. Figure 6(a) shows the median value of KTRD distances across datasets.

It demonstrates that the mean centroid PredDiff, Shapley value, and KernelSHAP are more consistent than the PredDiff.

Across-XAI-methods comparison. The baseline method is selected in rotation out of the four methods. Figure 6(b) plots the 50th percentile of KTRD

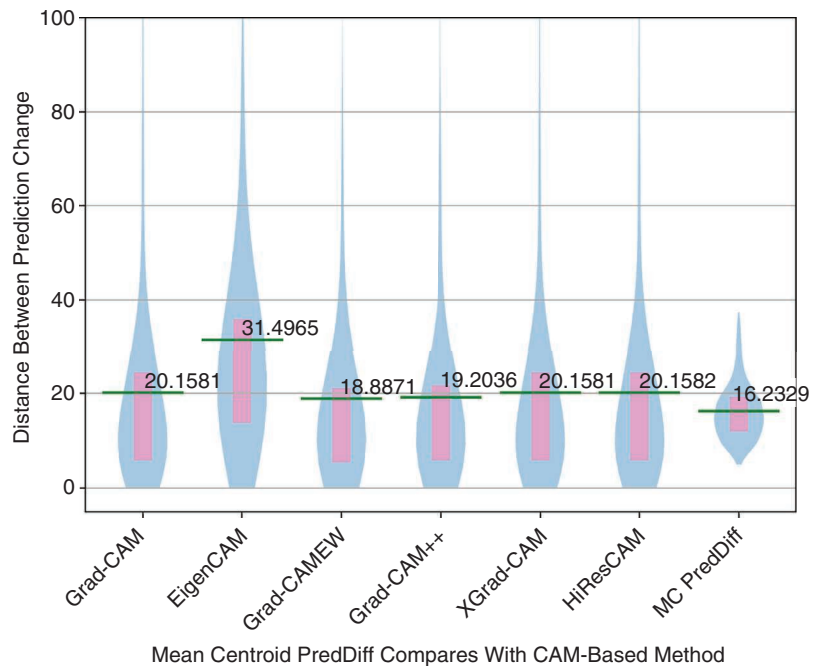


FIGURE 5. The explanation consistency between the mean centroid PredDiff and other CAM-based methods. The green line indicates the mean value.

TABLE 1. The feature importance order summary of the code vulnerability detection case study.

XAI method	Juliet test case	OWASP test case
PredDiff	Comment > code > import	Code > import > comment
Mean centroid PredDiff	Comment > code > import	Code > import > comment
Shapley value	Comment > code > import	Comment > code > import
KernelSHAP	Comment > code > import	Comment > code > import

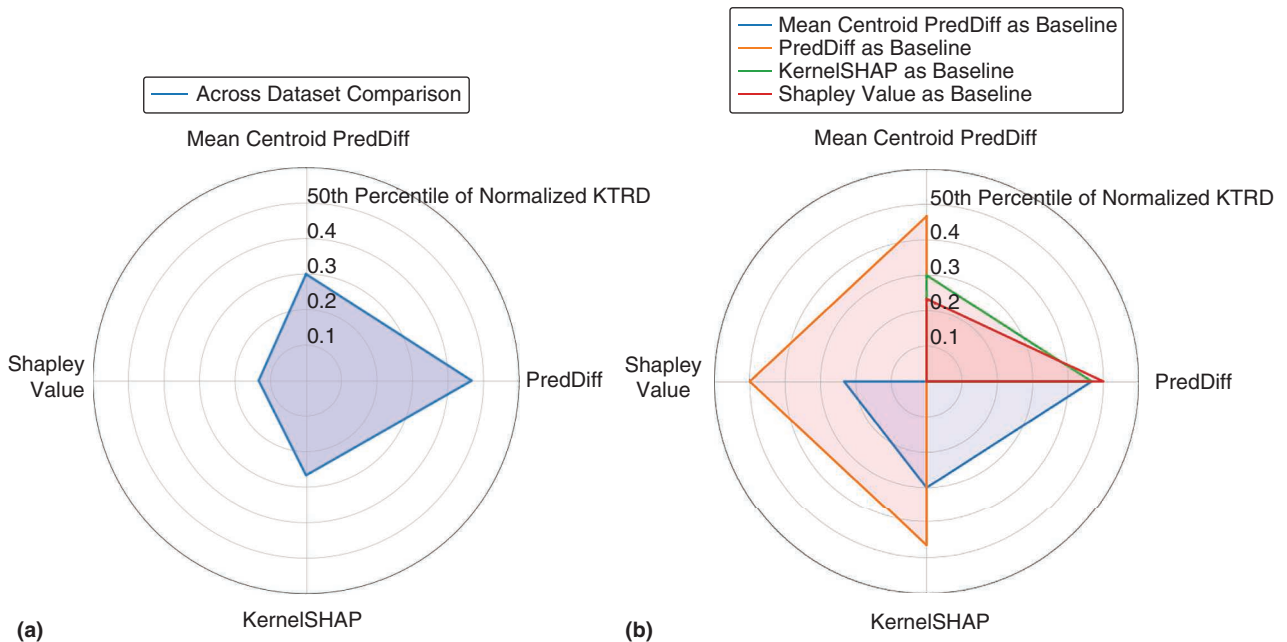


FIGURE 6. A consistency comparison across datasets and XAI methods. A shorter link edge indicates a more consistent XAI method. The results (a) across datasets and (b) across XAI methods.

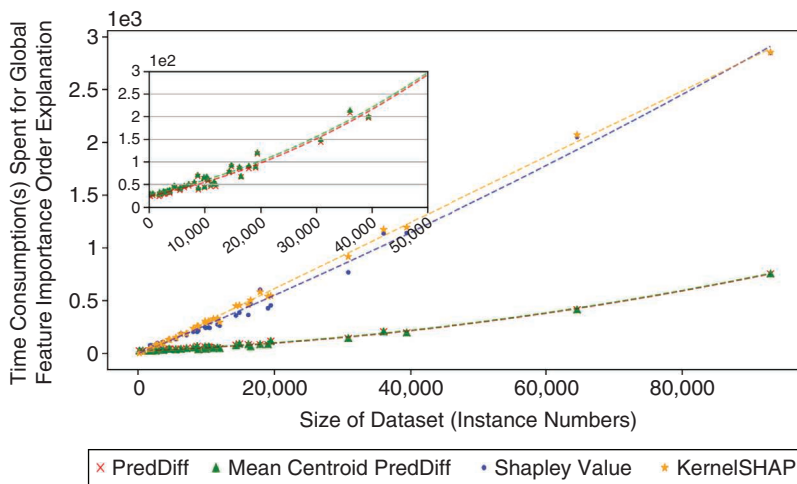


FIGURE 7. The time consumption among XAI methods as the dataset size increases in the S2Search case study.

distances. In summary, the mean centroid PredDiff is more consistent than the PredDiff but less than the other two methods.

Computing time consumption

Figure 7 indicates that the time consumption curve increases as the number of data samples grows. The

PredDiff and mean centroid PredDiff are more time efficient than KernelSHAP and the Shapley value. The mean centroid PredDiff spends approximately 10% more time than the PredDiff due to the clustering computation.

This article discussed the trustworthy view of XAI methods by defining consistency and efficiency metrics. Two metrics, consistency and time efficiency, provide a tradeoff view to evaluate XAI methods. In the case that higher consistency and faster time efficiency cannot be achieved simultaneously, users are left to prioritize the metrics for decision making. Through case studies, we observe that state-of-the-art XAI methods may produce explanation summaries that vary at the dataset level and

ABOUT THE AUTHORS

DING LI is an M.Sc. student in the Department of Electrical and Computer Engineering, Concordia University, Montréal, QC H3G 1M8, Canada. His research interests include explainable artificial intelligence in natural language processing and software vulnerability analysis. Li received an M.Eng. in electronic and communication engineering from Shanghai University, Shanghai, China. He is a Graduate Student Member of IEEE. Contact him at ding.li@mail.concordia.ca.

YAN LIU is a tenured associated professor and Gina Cody Research and Innovation Fellow in the Department of Electrical and Computer Engineering, Concordia University, Montréal, QC H3G 1M8, Canada. Her research interests include large-scale software systems to support government, enterprise, and scientific applications. Liu received a Ph.D. in computer science from the University of Sydney, Sydney, Australia. She is a Member of IEEE. Contact her at yan.liu@concordia.ca.

JUN HUANG is an M.Sc. student in the Department of Electrical and Computer Engineering, Concordia University, Montréal, QC H3G 1M8, Canada. His research interests include explainable artificial intelligence (AI) and explainable AI in convolution neural networks. Huang received a B.Eng in software engineering from South-Central Minzu University, Wuhan, China. He is a Graduate Student Member of IEEE. Contact him at jun.huang@concordia.ca.


ZERUI WANG is a Ph.D. student in the Department of Electrical and Computer Engineering, Concordia University, Montréal, QC H3G 1M8, Canada. His research interests include machine learning, explainable artificial intelligence, graph neural networks, and simulation. Wang received a master of science in process modelling and simulation from Technical University Dortmund, Dortmund, Germany. Contact him at zerui.wang@concordia.ca.

across methods. Hence, this motivates work to develop a unified evaluation method that helps to assess the explanation consistency of existing XAI methods as well as guide the development of a new XAI method. This evaluation method is the base for constructing the service pipeline of XAI operations. ■

REFERENCES

1. A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
2. B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen, "Beware explanations from AI in health care," *Science*, vol. 373, no. 6552, pp. 284–286, Jul. 2021, doi: 10.1126/science.abg1834.
3. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.
4. M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–7, doi: 10.1109/IJCNN48605.2020.9206626.
5. "Jacobgil/pytorch-grad-cam: Advanced AI explainability for computer vision. support for CNNs, vision transformers, classification, object detection, segmentation, image similarity and more." GitHub. Accessed: 2021. [Online]. Available: <https://github.com/jacobgil/pytorch-grad-cam>
6. A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.
7. R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-cam: Towards accurate visualization and explanation of CNNs," 2020, *arXiv:2008.02312*.
8. R. L. Draelos and L. Carin, "HiRes-CAM: Faithful location representation in visual attention for explainable 3D medical image classification," 2020, *arXiv:2011.08891*.
9. J. Singh, M. Khosla, W. Zhenye, and A. Anand, "Extracting per query valid explanations for blackbox learning-to-rank models," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retrieval*, 2021, pp. 203–210, doi: 10.1145/3471158.3472241.
10. L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," 2017, *arXiv:1702.04595*.
11. P. Schwab and W. Karlen, "CXPlain: Causal explanations for model interpretation under uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 1–11.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
13. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4768–4777.
14. H. W. Kuhn and A. W. Tucker, *Contributions to the Theory of Games*, vol. 2.

- Princeton, NJ, USA: Princeton Univ. Press, 1953.
15. K. de Bie, A. Lucic, and H. Haned, "To trust or not to trust a regressor: Estimating and explaining trustworthiness of regression predictions," 2021, *arXiv:2104.06982*.
 16. J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, Mar. 2021, Art. no. 593, doi: 10.3390/electronics10050593.
 17. M. Parmar, 2021, "XAIsec - Explainable AI security: An early discussion paper on new multidisciplinary subfield in pursuit of building trust in security of AI systems," doi: 10.31219/osf.io/rc92f.
 18. A. Wikekoon and N. Wiratunga, "Reasoning with counterfactual explanations for code vulnerability detection and correction," in *Proc. CEUR Workshop*, 2021, pp. 1-3.
 19. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 5753-5763.
 20. I. Covert, S. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," *J. Mach. Learn. Res.*, vol. 22, no. 209, pp. 1-90, 2021.
 21. R. Kumar and S. Vassilvitskii, "Generalized distances between rankings," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 571-580, doi: 10.1145/1772690.1772749.
 22. D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165-193, Jun. 2015, doi: 10.1007/s40745-015-0040-1.
 23. D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, vol. 741, S. Z. Li and A. Jain, Eds. Boston, MA, USA: Springer Science & Business Media, 2009, pp. 659-663, doi: 10.1007/978-0-387-73003-5_196.
 24. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
 25. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
 26. "Allennai/s2search: The semantic scholar search reranker." GitHub. Accessed: 2020. [Online]. Available: <https://github.com/allenai/s2search>




**SUBMIT
TODAY**

IEEE TRANSACTIONS ON

BIG DATA


▶ SUBSCRIBE AND SUBMIT


For more information on paper submission, featured articles, calls for papers, and subscription links visit: www.computer.org/tbd



TBD is financially cosponsored by IEEE Computer Society, IEEE Communications Society, IEEE Computational Intelligence Society, IEEE Sensors Council, IEEE Consumer Electronics Society, IEEE Signal Processing Society, IEEE Systems, Man & Cybernetics Society, IEEE Systems Council, and IEEE Vehicular Technology Society

TBD is technically cosponsored by IEEE Control Systems Society, IEEE Photonics Society, IEEE Engineering in Medicine & Biology Society, IEEE Power & Energy Society, and IEEE Biometrics Council





Digital Object Identifier 10.1109/MC.2023.3254040