IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 April 2025, accepted 24 April 2025, date of publication 29 April 2025, date of current version 7 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3565279

RESEARCH ARTICLE

HuberAIME: A Robust Approach to Explainable AI in the Presence of Outliers

TAKAFUMI NAKANISHI, (Member, IEEE)
School of Computer Science, Tokyo University of Technology, Hachioji-shi, Tokyo 192-0982, Japan
e-mail: nakanishitf@stf.teu.ac.jp

This work was supported by the School of Computer Science, Tokyo University of Technology, Tokyo, Japan.

ABSTRACT With the increasing accuracy of machine-learning models in recent years, explainable artificial intelligence (XAI), which allows for an understanding of the internal decisions made by these models, has become essential. However, many explanation methods are vulnerable to outliers and noise, and the results may be distorted by extreme values. This study devised a new method named HuberAIME, which is a variant of approximate inverse model explanations (AIME) and is robust to the Huber loss. HuberAIME limits the impact of outliers by weighting with iterative reweighted least squares and prevents the feature importance estimation of AIME from being degraded by extreme data points. Comparative experiments were conducted using the Wine dataset, which has almost no outliers, the Adult dataset, which contains extreme values, and the Statlog (German Credit) dataset, which has moderate outliers, to demonstrate the effectiveness of the proposed method. SHapley Additive exPlanations, AIME, and HuberAIME were evaluated using six metrics (explanatory accuracy, sparsity, stability, computational efficiency, robustness, and completeness). HuberAIME was equivalent to AIME on the Wine dataset. However, it outperformed AIME on the Adult dataset, exhibiting high fidelity and stability. On the Germain Credit dataset, AIME itself showed a certain degree of robustness, and there was no significant difference between AIME and HuberAIME. Overall, HuberAIME is useful for data that include serious outliers and maintains the same explanatory performance as AIME in cases of few outliers. Thus, HuberAIME is expected to improve the reliability of actual operations as a robust XAI method.

INDEX TERMS Approximate inverse model explanations, explainable AI, global feature importance, Huber loss, iterative reweighted least squares, model-agnostic explanations, outliers, robustness.

I. INTRODUCTION

With machine-learning models achieving higher accuracy and wider adoption in society, the need for humans to understand the reasons behind the outputs of these models has rapidly increased. In recent years, sophisticated methods such as deep learning and ensemble learning have achieved excellent performance across various fields; however, because their internal structures are not intuitively graspable, these methods are often regarded as closed-box models. In high-risk domains such as medicine and finance, if the estimation and decision-making processes of the model cannot be explained, the model may not be accepted in practice or could fail to meet regulatory requirements, regardless of its predictive accuracy. Within this context, numerous studies have been conducted under the collective umbrella of explainable AI (XAI) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20].

XAI approaches can be roughly divided into two types: the design of "transparent models" that provide easier visualization of the model itself, and "model-agnostic" (post-hoc) methods that approximate the behavior of the learned model locally or globally. Well-known examples of the latter include Local Interpretable Model-Agnostic Explanations LIME [21] and SHapley Additive exPlanations (SHAP) [22], which reveal the extent to which a closed-box model depends on each feature by calculating its contribution to the prediction. However, although these methods each have certain strengths, they have also been noted to incur high computational costs and exhibit vulnerability when handling data containing out-

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

liers. In particular, when these models are applied to real datasets with many extreme values, the resulting explanations may be substantially distorted by these outliers.

This study focused on approximate inverse model explanations (AIME) [23], previously proposed by the author, and extended the method to provide greater robustness. AIME separately learns the input–output relationship of a closed-box model and then derives an approximate inverse operator that reconstructs the original input space from the predicted model outputs. For example, if the outputs are k-dimensional class probabilities, AIME learns a data-driven inverse mapping that predicts the input features from the probability vector and subsequently computes the feature importance from this operator. AIME retains versatility for a wide range of models and makes it easy to obtain both local and global explanations because the construction does not require reference to the internal structure of the original model. However, the regression approach used by AIME is based on least-squares estimation (the L2 loss), which may be extremely sensitive to a small number of extreme values (or outliers).

Outliers occur frequently in real-world data. Physiologically impossible or extremely large values often emerge owing to sensor malfunctions or input errors. In addition, when dealing with phenomena that naturally exhibit heavy-tailed distributions, such as market prices or human behavior, removing all outliers can degrade the accuracy. Because the conventional least-squares estimation minimizes squared errors, it is highly sensitive to samples with large residuals, and even a single outlier can distort the overall estimation substantially. Therefore, applying AIME without modification could make it susceptible to such outliers, potentially undermining the reliability of its explanations.

On this basis, HuberAIME, which incorporates insights from robust statistics into AIME by adopting the Huber loss [24], is proposed in this study. The Huber loss is a hybrid function that behaves similar to the squared error for small residuals but changes to a linear penalty when the residual exceeds a certain threshold. This reduces the influence of large-error samples and prevents outliers from having an outsized effect on the estimation compared with least squares. In addition, an iterative solution that automatically downweights the outliers can be implemented by using iterative reweighted least squares (IRLS) [25], [26]. Consequently, HuberAIME is expected to preserve the general explanatory capabilities of AIME while achieving robustness that does not compromise the explanatory accuracy or stability, even in the presence of outliers.

To demonstrate the usefulness of the proposed HuberAIME, this study first compared AIME and HuberAIME using the Wine dataset [27], which contains few outliers. No discernible difference was found between the two methods in the absence of outliers. Subsequently, the performance difference between AIME and HuberAIME was evaluated on the Adult dataset [28], which includes extreme values, as well as the Statlog (German Credit) dataset [29], which contains moderate outliers. A comparative evaluation with SHAP, which is a well-known model-agnostic method, was also performed, and the behaviors of the three methods were comprehensively analyzed based on six indicators: explanatory accuracy, sparsity, stability, computational efficiency, robustness, and completeness. The results revealed that HuberAIME shows higher explanatory accuracy and stability than AIME on data with prominent outliers, whereas the results approximately match those of AIME in cases of few outliers. This behavior aligns with theoretical insights in robust statistics, indicating that the advantage of HuberAIME lies in mitigating the risk of outliers but acting similarly to conventional methods when outliers are not severe. Furthermore, robust regression with IRLS is far less computationally intensive than SHAP, requiring approximately the same runtime as AIME, which suggests that the proposed method is well suited to large-scale data and real-time applications.

As real-world data commonly contain outliers and noise, it is highly desirable for XAI to function reliably in such ‘‘messy’’ environments. From this perspective, the proposed HuberAIME retains the broad applicability of AIME while adding robustness, making it a promising method for providing safe and trustworthy explanations for complex closed-box models. In future work, it may be possible to establish even more robust explanation methods by applying HuberAIME to larger-scale tasks (e.g., images or text) and dynamically adjusting the hyperparameters (such as the Huber threshold).

II. RELATED WORK

In recent years, local model-agnostic explanation methods such as LIME [21] and SHAP [22] have been used extensively to interpret the outputs of machine-learning models. LIME explains predictions by training a simple local surrogate model (e.g., a linear model) around the instance in question, and it is characterized by its straightforward approach and relatively high computational efficiency. SHAP computes the feature contributions based on Shapley values from game theory, enabling a theoretically fair evaluation of the effect of each feature on the output. However, as SHAP requires multiple model evaluations, its computational cost is generally higher than that of LIME. For example, Salih et al. [30] reported that ‘‘LIME is faster than SHAP,’’ particularly when explaining decision tree models, and compared the theoretical foundations and performance of SHAP and LIME in detail, discussing their respective advantages (e.g., the theoretical rigor of SHAP vs. computational efficiency of LIME). Furthermore, a systematic literature review by Saarela and Podgorelec [31] revealed that local feature importance methods, particularly LIME and SHAP, are the most widely applied approaches for interpreting machine-learning models across various domains, underscoring their practical significance.

The presence of outliers can undermine both the accuracy and stability of local surrogate models when explaining closed-box models. Thus, several recent studies have incorporated robust regression techniques into XAI to achieve outlier-resistant explanations. For example, Sim et al. [32]

used the Huber loss to optimize a model that predicts energy consumption, handling small errors via the squared-error (L2) approach and large errors via a linear (L1) penalty. Their method reportedly enhanced the robustness to outliers compared with standard least-squares approaches, while maintaining high computational efficiency (rapid convergence). Huber-loss-based regression is a form of M-estimation that downweights outliers, and it can be solved efficiently using an IRLS algorithm. Consequently, even if the Huber loss is introduced into a local surrogate model, its computational cost remains at approximately the same order as that of regular LIME, potentially surpassing SHAP in terms of the computational speed. Eik et al. [33] implemented a Huber-based regression model for local linear explanations in building air-conditioning prediction, demonstrating the usefulness of a robust explanation for outliers. Their method also integrates feature-importance computations using Shapley values, which suggests that it can achieve high reliability in explaining model predictions. Thus, robust regression-based XAI offers an advantage over traditional methods in terms of outlier tolerance and is particularly promising for noisy data settings, such as industrial fields.

IRLS, which is a well-known solution for M-estimation and robust regression, has also been applied in XAI. Adopting the Huber loss within a local surrogate model typically entails fitting while iteratively updating weights via IRLS. Hence, the approaches of Sim et al. [32] and Eik et al. [33] can be viewed as practical instances of applying IRLS to XAI. Notably, Sim et al. employed IRLS to select input variables in an energy-demand-forecasting model, thereby improving the stability of the explanations. Moreover, researchers have attempted to enhance the reproducibility and stability by adjusting the weighting or selection of neighboring samples when extending the standard LIME. Bora et al. [34] proposed Stabilized LIME for image classification, successfully mitigating the explanation variability caused by randomness and achieving consistent rankings of important features. Although this method does not strictly extend IRLS itself, it relies on refined weighting strategies to improve the local surrogate reliability. It broadly aligns with IRLS-like concepts (assigning an appropriate weight to each data point), thereby contributing to more robust model explanations. Overall, methods that embed IRLS or robust estimation concepts help to balance the explanatory fidelity (i.e., how accurately the surrogate approximates the target model) and stability, and their use in explaining various models is expected to expand.

The industrial applications of XAI have also been actively researched in recent years. The ability to explain AI model decisions is critical for domain experts to deploy these models confidently in many industrial settings. According to Saarela & Podgorelec [31], the range of XAI applications has expanded to include medicine and finance, as well as emerging domains such as law, education, social welfare, and environmental/agricultural management, along with industrial optimization, cybersecurity, finance, and transportation. In particular, XAI is anticipated to support regulatory compliance and enhance safety in the manufacturing and process industries by increasing model transparency [35]. For instance, Eik et al. [32] applied XAI to an industrial use case for building air-conditioning systems and proposed a method for clarifying the model outputs via SHAP. They demonstrated that operators can obtain interpretable feedback for system control and extract insights to improve the energy efficiency. Feilmayr et al. [36] compared and evaluated multiple explanation methods (including SHAP, LIME, and Averaged Local Effects) to model the control of an electric arc furnace and analyzed the limitations of XAI robustness in industrial process modeling. The authors quantified the explanatory accuracy and sensitivity of each method using simulation-based ground truths, and they identified challenges in providing reliable explanations for industrial applications. Such studies indicate that XAI is not merely a technique for model visualization but is directly connected to decision support and process optimization in actual operations. Integrating domain knowledge with XAI techniques for specific industries and embedding explanations into day-to-day workflows (human–AI collaboration) will become increasingly necessary in the future.

AIME [23] can derive explanations such as the global feature importance, representative estimation instances, representative instance similarity distribution plots, and local feature importance. SHAP can also obtain the global feature importance by aggregating local feature attributions. Thus, this study compared SHAP, AIME, and the proposed HuberAIME, which incorporates the Huber loss into AIME to increase the robustness, with a particular focus on extracting the global feature importance. This study aimed to verify the efficacy of HuberAIME in achieving outlier resistance without sacrificing computational efficiency or explanatory fidelity. According to the AIME formulation [23], the local feature importance can be derived with high robustness by calculating the global feature importance with high robustness. Therefore, this study sought to derive the global feature importance to narrow the target and demonstrate its effect. If HuberAIME is verified as effective, the local feature importance, representative estimation instance, and representative instance similarity distribution plot can be derived with high robustness.

As opposed to SHAP, AIME, and HuberAIME, which can produce class-specific global feature importance and distinguish between positive and negative contributions, built-in methods in algorithms such as Random Forest can only generate a single overall feature importance score (which is always non-negative) and lack the ability to indicate whether a given feature contributes positively or negatively. Therefore, they are not directly comparable to our proposed explanation methods.

III. METHODS

A. OVERVIEW OF AIME

AIME constructs an "approximate inverse operator" for a closed-box model by directly learning the correspondence between the inputs and outputs of the model and subsequently uses this operator to evaluate the input feature importance. Specifically, a large number of pairs $\boldsymbol{x}_i \in \mathbb{R}^m$ are sampled in the input space and the corresponding outputs (predictions) $\boldsymbol{y}_i = f(\boldsymbol{x}_i) \in \mathbb{R}^m$ of the closed-box model f are provided. These samples are arranged in matrix form, as follows:

$$X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_N] \in \mathbb{R}^{m\times N} \tag{1}$$

$$Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots \boldsymbol{y}_N] \in \mathbb{R}^{n\times N}, \tag{2}$$

where each $\boldsymbol{x}_i$ is m-dimensional, each $\boldsymbol{y}_i$ is n-dimensional, and X and Y are formed by stacking these column vectors. In AIME, the goal is to learn an "inverse mapping" $A^{\dagger} \in \mathbb{R}^{m\times n}$ such that $\boldsymbol{x}_i \approx A^{\dagger}\boldsymbol{y}_i$, or in matrix notation, $X \approx A^{\dagger}Y$. If this approximation holds, the corresponding input in $\mathbb{R}^m$ can be estimated by computing $A^{\dagger}\boldsymbol{y}$ for a given output $\boldsymbol{y}$ of the closed-box model f. AIME leverages this approximate inverse operator (mapping outputs back to inputs) to assess the feature importance globally and locally. Specifically, $A^{\dagger}$ is obtained by solving the following least-squares problem:

$$\min_{A^{\dagger}\in\mathbb{R}^{m\times n}} \sum\nolimits_{i=1}^{N} \left\| \boldsymbol{x}_i - A^{\dagger}\boldsymbol{y}_i \right\|_2^2, \tag{3}$$

which, in matrix form, is

$$\min_{A^{\dagger}\in\mathbb{R}^{m\times n}} \left\| X - A^{\dagger}Y \right\|_F^2. \tag{4}$$

That is, the square of the Frobenius norm is minimized. As this is a straightforward least-squares solution, if YY^T is invertible (i.e., $\mathrm{Rank}\,(Y) = n$), the closed-form solution is expressed as

$$A^{\dagger} = XY^T(YY^T)^{-1}. \tag{5}$$

If the data are poorly sampled or clustered, YY^T may become singular. However, from both theoretical and practical perspectives, AIME typically assumes that $\mathrm{Rank}\,(Y) = n$ can be maintained by sampling y appropriately.

The resulting $A^{\dagger}$ serves as a linear transform that approximately reconstructs the original input vector $\boldsymbol{x}$ from the output $\boldsymbol{y}$. AIME can estimate which components of $\boldsymbol{y}$ (i.e., which output dimensions) contribute strongly to each feature in the input space by interpreting the elements of this inverse operator appropriately, thereby providing a form of global feature importance. In addition, various visualization and explanation methods, such as representative estimation instances and instance similarity distributions, are reportedly derivable by computing the representative input vectors for specific $\boldsymbol{y}$ values via $A^{\dagger}$ directly, or by performing local analyses on arbitrary samples.

However, the least-squares approximation is highly susceptible to outliers. Specifically, $\left\| \boldsymbol{x}_i - A^{\dagger}\boldsymbol{y}_i \right\|_2$ may become very large if any $\boldsymbol{y}_i$ value is an abnormal output or an extreme point of the model. Owing to the nature of the squared error, even a single extreme sample can skew the entire estimation disproportionately. In effect, if $\left\| \boldsymbol{x}_i - A^{\dagger}\boldsymbol{y}_i \right\|_2^2$ is extremely large, the optimization will attempt to accommodate this sample, potentially sacrificing the approximation accuracy for many other "normal" samples. As a result, the final $A^{\dagger}$ may be excessively influenced by outliers, increasing the risk of the model failing to capture the broader, typical behavior of the original model.

The instability caused by such outliers must be overcome to employ the explanatory ability of AIME in real-world settings. In this study, an extended version of AIME is proposed using the Huber loss (HuberAIME) to address this issue. The Huber loss behaves similarly to the squared error when the residual $\left\| \boldsymbol{x}_i - A^{\dagger}\boldsymbol{y}_i \right\|_2$ is small, but switches to a linear penalty beyond a certain threshold. This curbs the penalty for samples with extremely large residuals (i.e., outliers) and prevents the entire estimation from being overly swayed by these values. Moreover, regression with the Huber loss can be approximated via IRLS, thereby maintaining a computational complexity that is close to that of standard least squares, which is a major practical benefit.

Thus, whereas AIME provides a theoretically simple framework for approximating an inverse operator via least squares and can support both global and local explanations by reconstructing inputs from the outputs of a closed-box model, its reliance on the squared error leads to vulnerability to outliers. This study addresses this vulnerability through HuberAIME, with the aim of establishing a stable explanation method that can robustly capture key factors in the input space, even when dealing with "dirty" data containing outliers.

B. HUBERAIME

HuberAIME introduces the Huber loss instead of least squares to overcome the vulnerability of AIME to outliers. This switching is controlled by the parameter defined in (6), which is greater than zero. When the residual is defined as $r = \left\| \boldsymbol{x}_i - A^{\dagger}\boldsymbol{y}_i \right\|_2$,

$$L\delta\,(r) = \begin{cases} \frac{1}{2}\mathbb{R}^2 & |r| \le \delta \\ \delta(|r| - \frac{\delta}{2}) & |r| > \delta \end{cases} \tag{6}$$

That is, up to a certain threshold value of δ, the squared error is equivalent to $\frac{1}{2}r^2$; however, for large residuals exceeding this threshold, the penalty switches to a linear penalty so that the overall estimation is not significantly affected by samples with outliers. The optimization problem in (7) is obtained by replacing the squared error part of the least-squares problem used in AIME with this Huber loss.

$$\min_{A^{\dagger}\in\mathbb{R}^{m\times n}} \sum\nolimits_{i=1}^{N} L\delta\left(\left\| \boldsymbol{x}_i - A^{\dagger}\boldsymbol{y}_i \right\|_2 \right) \tag{7}$$

When converted into matrix notation, the Huber loss expresses the total loss including all samples of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$; however, it generally does not have a closed-form solution.

Therefore, this study uses IRLS to obtain an approximate solution. In IRLS, an initial estimate $A^{\dagger(0)}$ is provided using a certain method (e.g., the least-squares solution), the residuals for each sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are calculated as $r_i^{(k)} = \left\| \boldsymbol{x}_i - A^{\dagger(k)} \boldsymbol{y}_i \right\|_2$, and the weights $w_i^{(k)}$ are updated according to their magnitude. In the case of the Huber loss,

$$w_i^{(k)} = \begin{cases} 1 & r_i^{(k)} \leq \delta \\ \frac{\delta}{\left| r_i^{(k)} \right|} & r_i^{(k)} > \delta \end{cases} \tag{8}$$

That is, if the residual is within the threshold, the weight is set to 1, and if it exceeds the threshold significantly, it is set to the value of the threshold divided by the absolute value of the residual to suppress the influence of outlying samples.

Subsequently, the following weighted least-squares problem is solved using these weights. Letting the residual matrix be $\left\| X - A^{\dagger} Y \right\|_2$ and the i-th diagonal element of the diagonal weight matrix $W^{(k)}$ be $w_i^{(k)}$,

$$A^{\dagger(k+1)} = \underset{A^{\dagger} \in \mathbb{R}^{m \times n}}{argmin} \left\| W^{(k)} (X - A^{\dagger} Y) \right\|_F^2 . \tag{9}$$

Equation (9) is equivalent to the least-squares solution with small weights for outliers at each iteration step. In matrix notation, it can be updated as follows:

$$A^{\dagger(k+1)} = X W^{(k)} Y^T \left(Y W^{(k)} Y^T \right)^{-1} . \tag{10}$$

If each sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ has a large residual, the corresponding matrix element $w_i^{(k)}$ is reduced so that the estimation effectively ignores the outliers. $A^{\dagger}$, which minimizes the Huber loss approximately, can be obtained if this iteration is repeated sufficiently. The computational cost remains at the same order as that of the least-squares method because only a few weighted loops are added to the least-squares method.

The inverse operator $A^{\dagger}$ of HuberAIME, which is calculated in this manner, treats small residuals similarly to AIME (that is, learning based on the squared error) while applying a linear penalty to outliers with large residuals (abnormal outputs or extreme values of the model), thereby achieving an approximate inverse mapping that suppresses distortion owing to outliers. As a result, HuberAIME can stably capture the typical behavior of closed-box models while maintaining the computational cost at a level similar to that of AIME, even in datasets with substantial noise and outliers.

IV. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT

The global feature importance results of SHAP, AIME, and HubereAIME were compared using the Wine dataset [27], which contains few outliers; Adult dataset [28], which includes extreme values; and Statlog (German Credit) dataset [29], which includes moderate outliers.

The types of objective variables and amounts of data for the respective datasets [27], [28], [29] used are listed in Tables 1, 2, and 3, respectively.

TABLE 1. Values of the objective variables in the wine dataset and number of data points.

Objective variables	Number of data points
1	59
2	71
3	48

TABLE 2. Values of the objective variables in the adult dataset and number of data points.

Objective variables	Number of data points
<=50K	37155
>50K	11687

TABLE 3. Values of the objective variables in the Statlog (German Credit) dataset and number of data points.

Objective variables	Number of data points
1	700
2	300

This system was implemented in Python 3.11.11 on Google Colab Pro+. In the experiments, LightGBM [37] (version 4.5.0) was used as the target machine-learning model. However, because AIME/HuberAIME is model-agnostic, in principle, any closed-box model can be used if desired. Categorical variables were converted into one-hot vectors using scikit-learn 1.6.1. AIME and HuberAIME were implemented using NumPy 1.26.4, pandas 2.2.2, seaborn 0.13.2, scikit-learn 1.6.1, and matplotlib 3.10.0, which are publicly available on GitHub [38] and PyPI [39].

When ensemble learners such as Random Forest and XGBoost are used as closed-box models, they can also be compared using the global feature importance metrics calculated internally by these models. However, AIME and SHAP can calculate the global feature importance for each target variable, whereas Random Forest and XGBoost cannot. In addition, in Random Forest and XGBoost, the importance only indicates the size of the contribution and cannot reveal a positive or negative impact on prediction. Thus, these models were not included in the comparison as it was known in advance that their explanatory power would be lower than that of SHAP, AIME, and HuberAIME.

In the HuberAIME implementation of this study, the IRLS hyperparameters were configured as follows. The Huber loss threshold δ was set to 1.0, up to 50 iterations were permitted ($max\_iter = 50$), and a convergence tolerance $tol = 10^{-5}$ was used. All of these parameters were employed with their default settings within the code used. Any sample with a residual exceeding δ was assigned a weight of $\frac{\delta}{r_i}$ at each IRLS step, whereas samples with smaller residuals were assigned a weight of 1. This approach automatically downweights any present outliers, thereby mitigating the impact on AIME, which is based purely on the squared error. The behavior of the method is similar to that of ordinary least squares for the majority of samples with residuals that remain below δ, maintaining comparable estimation accuracy. They are determined by preliminary experiments and empirically validated to balance robustness and efficiency would be sufficient.

In this study, any missing values in the numerical features were imputed with the median value of the corresponding feature, whereas missing values in the categorical features were assigned to a special "unknown" category. All categorical variables were then transformed through one-hot encoding, ensuring that each distinct category (including "unknown") was represented as a separate binary indicator. Following this imputation and encoding process, the numerical features were standardized through z-score normalization to have zero mean and unit variance. This consistent approach to handling missing data, categorical variables, and feature scaling was applied across all experiments.

B. QUALITATIVE EVALUATION

This section presents a comparison and qualitative evaluation of the global feature importance calculated by the SHAP, AIME, and HuberAIME methods on the Wine, Adult, and Statlog (German Credit) datasets. The global feature importance is an important metric for comparing methods in an integrated manner. In principle, SHAP calculates the local contributions as Shapley values, and the global feature importance can be derived by aggregating these contributions in the form of summation, averaging, and maximization across all samples. AIME and HuberAIME calculate the importance of each feature by directly interpreting the coefficients of the approximate inverse operator.

Figs. 1–3 show the global feature importance values derived by SHAP, AIME, and HuberAIME for the Wine dataset, respectively. Fig. 1 shows the result of integrating the local contribution values for the entire sample, where positive values were assigned to all features for classes 1, 2, and 3. Figs. 2 and 3 show the global importance derived from the coefficients of AIME and HuberAIME, with the contribution in both the positive and negative directions for each feature. For example, Flavanoids had a positive contribution to classes 1 and 2 and a negative contribution to class 3 in AIME and HuberAIME, whereas Hue classes 1 and 3 and Proline had a large positive contribution to class 1, and Hue class 2 and Proline showed negative contributions for classes 2 and 3. Thus, various degrees of importance, including positive and negative, were output.

Notably, almost no difference was observed between the results of AIME and HuberAIME because the Wine dataset is considered clean data with almost no outliers. In Figs. 2 and 3, the size and direction of the bars for both methods are almost the same, and only a slight difference in the error could be confirmed. This is consistent with the theoretical properties of robust statistics, which show that the Huber loss has a mechanism for suppressing samples with large residuals; however, when outliers do not exist, it behaves equivalently to standard least squares (AIME). In contrast, the global feature importance in SHAP is calculated as a positive contribution because SHAP aggregates the local Shapley values in a non-negative direction. However, SHAP is not an inherently "positive-only" method, and it is possible that the overall positive contribution was simply calculated as large in this dataset and class label configuration. In reality, SHAP may also show a mixture of positive and negative Shapley values at the local sample level; however, when class boundaries are relatively simple, such as in the Wine dataset, the positive contribution may predominate.

The following conclusions are suggested based on the comparison on the Wine dataset. First, there was almost no difference between AIME and HuberAIME in the case of clean data with no outliers. This is because the area in which the robust method (Huber loss) behaves in the same manner as the squared error is wide, and there is no disadvantage owing to robustness. Second, all values calculated by SHAP as the global feature importance being positive suggests that the method may integrate the sum and absolute value average of the local contributions as indicators. In reality, local explanation methods, including SHAP, change their positive and negative contributions depending on the sample. However, the class boundaries are relatively simple in the Wine dataset, and it is believed that only the positive direction was omitted through global averaging.

According to these results, focusing only on the Wine dataset, AIME and HuberAIME are almost equivalent in fields without outliers, and there is no need to be particularly conscious of robustness. In contrast, SHAP obtains global importance that is biased towards positive contributions as a set of local explanations, but this is not necessarily a mistake and can be considered a phenomenon caused by the SHAP calculation algorithm, class composition, and data distribution. In terms of outlier tolerance, as no prominent outlier examples exist in the Wine dataset, there is no clear difference in the superiority of AIME and HuberAIME.

Figs. 4–6 show the global feature importance (top 20 features only) of SHAP, AIME, and HuberAIME on the Adult dataset, respectively. As Adult has a large feature dimension, only the top 20 features were extracted, and the results for cases in which the income category was <=50K and >50K were compared. Fig. 4 shows that only positive values were output by SHAP, as with the Wine dataset. However, this is considered to be the result of SHAP aggregating local Shapley values to calculate the global feature importance, with the contribution of each feature combined in a manner that contributes positively overall. Figs. 5 and 6 depict the global feature importance derived from the coefficients calculated by AIME and HuberAIME, respectively. Both methods showed contributions including positive and negative values.

The comparison between AIME and HuberAIME in Figs. 5 and 6 shows that the results on the Adult dataset were clearly different from those on the Wine dataset, with

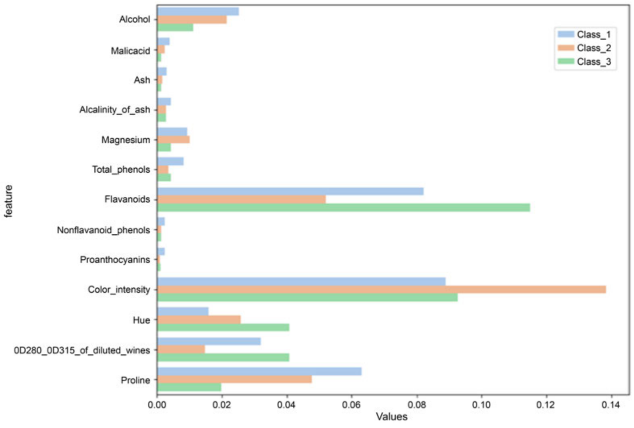

FIGURE 1. Results of the SHAP global feature importance on the Wine dataset.

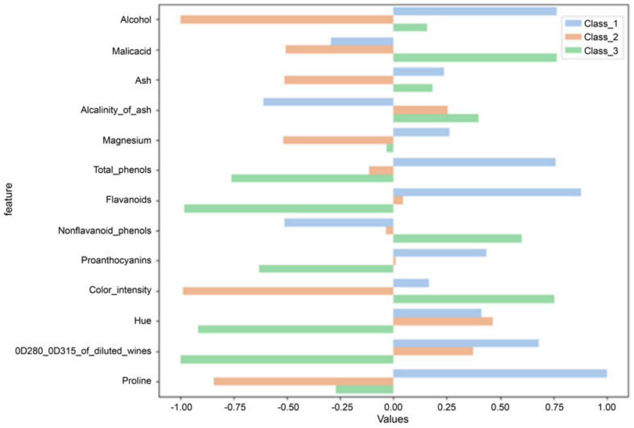

FIGURE 2. Results of the AIME global feature importance on the Wine dataset.

almost no difference between the two methods. For example, ‘‘native-country_United-States’’ appeared in the top 20 in HuberAIME but was outside the top 20 in AIME. Because the Adult dataset contains various outliers (or bias in the sample size) related to country of birth, the contribution of AIME was buried by the outliers, but HuberAIME suppressed this effect and could accurately capture the true importance of the data. In addition, whereas ‘‘workclass_Private’’ was in the top 20 in HuberAIME, it was outside the top 20 in AIME. A certain ratio of missing values and outliers exists in workclass, but HuberAIME could mitigate the effects thereof through robust estimation and by increasing the variable importance. As AIME, which is based on standard least squares, is strongly affected by outliers, it is believed that the overall contribution was distorted or ignored in certain cases. In this respect, HuberAIME suppressed the weight of large residual samples through the Huber loss and estimated the approximate inverse operator to reduce the distortion caused by outliers; thus, it was easier to calculate both positive and negative contributions clearly and significantly.

The results from the Adult dataset suggest a significant difference between AIME and HuberAIME in data environments that include outliers. Thus, when AIME is degraded by outliers and the overall approximation inverse operator yields low precision, HuberAIME can correct this distortion and extract features that are of practical value. In addition, the global importance value output by SHAP is represented as a bar graph that is biased towards positive values because, as mentioned previously, SHAP aggregates the absolute values of the contributions when it sums and averages the local explanation results (Shapley values for the individual samples). Although SHAP can also produce positive and negative values on a local basis, the final aggregation method may result in a different scale or sign direction from that of AIME or HuberAIME.

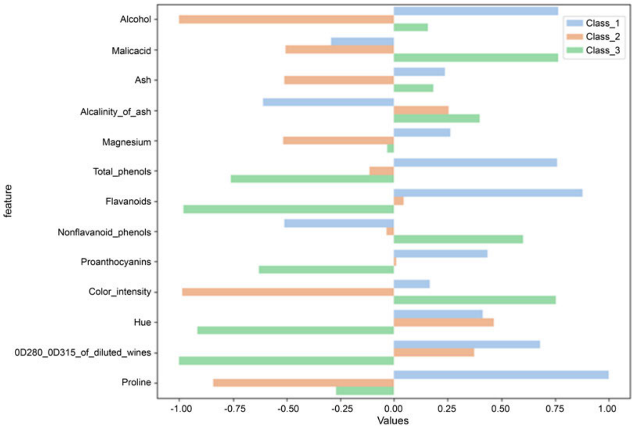

FIGURE 3. Results of the HuberAIME global feature importance on the Wine dataset.

In summary, the significant difference between the results of AIME and HuberAIME on the Adult dataset supports the idea that correction calculations are performed for outliers. That is, by applying IRLS based on the Huber loss, the approximation inverse operator is suppressed from being excessively distorted, even in the presence of outlier samples, and more features are reflected in the original contribution. HuberAIME is more likely to extract features appropriately than AIME in practical applications when explaining data with many outliers (or data of inconsistent quality). In addition, SHAP integrates and averages local explanations across all samples, and only positive contributions are revealed as the result of this aggregation. However, this does not necessarily indicate an error, and the model can be considered to be strongly dependent on specific aggregation methods and task structures. Among the three compared methods, the fact that HuberAIME shows a clearly different contribution to that of AIME in complex data environments that include outliers is a very interesting result from an outlier correction perspective.

Figs. 7–9 show the global feature importance (top 20 features only) of SHAP, AIME, and HuberAIME calculated for the Statlog (German Credit) dataset, respectively. As with the previous dataset, the SHAP results (Fig. 7) only showed positive contributions. This is because SHAP uses a method that integrates local Shapley values across all samples, and in the current data distribution and class composition, they were ultimately aggregated as positive contributions. In contrast,

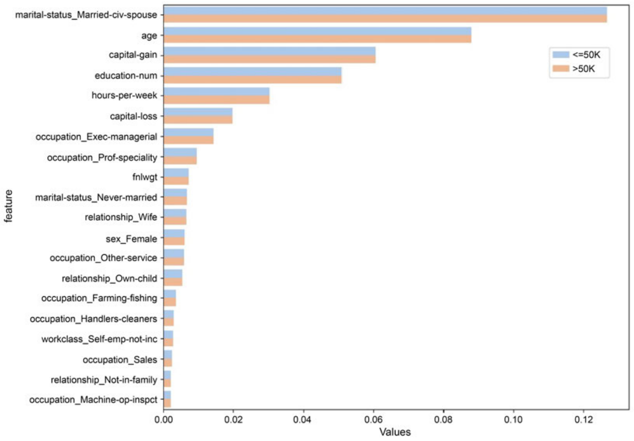

FIGURE 4. Results of the SHAP global feature importance on the Adult dataset.

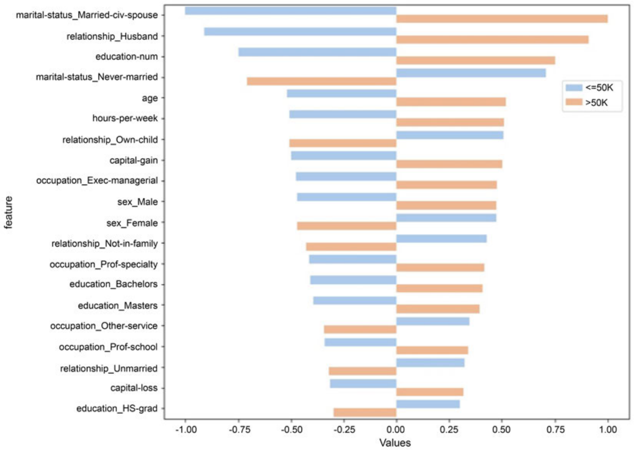

FIGURE 5. Results of the AIME global feature importance on the Adult dataset.

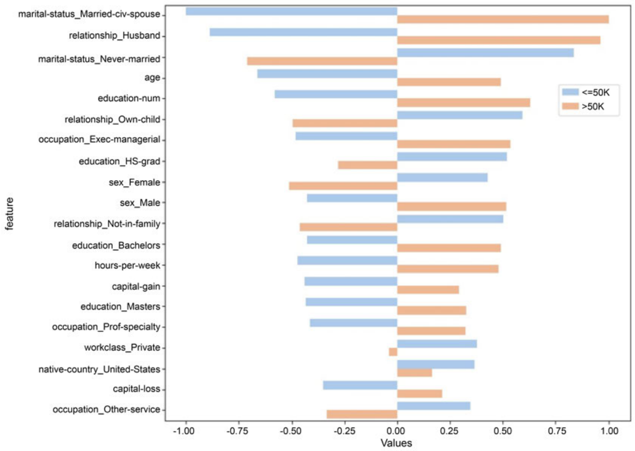

FIGURE 6. Results of the HuberAIME global feature importance on the Adult dataset.

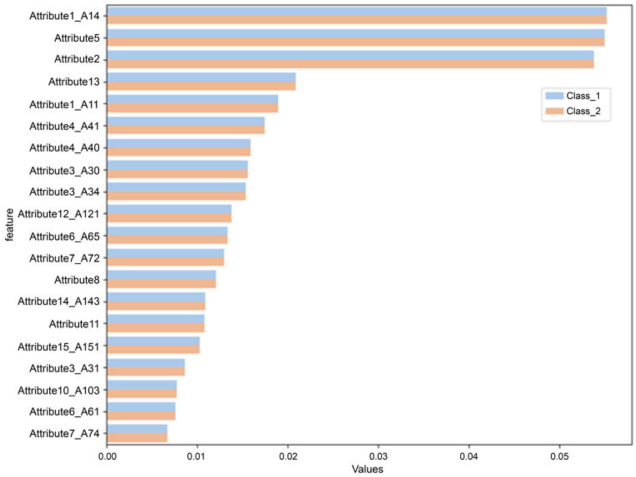

FIGURE 7. Results of the SHAP global feature importance on the Statlog (German Credit) dataset.

AIME (Fig. 8) and HuberAIME (Fig. 9) explicitly showed the contributions of both the positive and negative directions, so that whether each feature had a positive or negative effect on the model output (Class 1/Class 2) could be determined simultaneously.

Figs. 8 and 9 show that the top four features of AIME and HuberAIME were the same (for example, Attribute1_A14, Attribute5, Attribute2, and Attribute13 all showed a large contribution). However, notable differences between the lower rankings were evident. Specifically, in some cases, Attribute14_A141 and other features appeared in the top 20 in HuberAIME, but not in AIME. This is believed to be the result of the robust estimation effect in the presence of outliers.

The Statlog (German Credit) dataset contains a certain percentage of samples with missing and extreme values, and when AIME learned the inverse action matrix based on least squares, the contribution of some features may have been overestimated or underestimated because of certain outliers. In contrast, it is thought that the final contribution ranking of HuberAIME tended to differ from that of AIME because HuberAIME could suppress the weights of samples that produce large residuals through the Huber loss. Thus, the difference in results between AIME and HuberAIME as a result of robustness to outliers could also be confirmed in this dataset. In the case of SHAP, only the positive contributions are visualized globally because the local contributions are integrated and averaged over the entire sample. In particular, in cases with a wide variety of categories, such as Statlog (German Credit), SHAP shows that the contributions of each feature are locally negative in some cases but integrated in a positive manner overall.

In summary, the impact of features on the output class of the model can be understood more intuitively using AIME and HuberAIME, which clearly show positive and negative values. As SHAP also has a strong theoretical basis and stable interpretation, it is possible to use the models complementarily to obtain a multi-faceted understanding of the explanation. Nevertheless, as HuberAIME shows a different importance

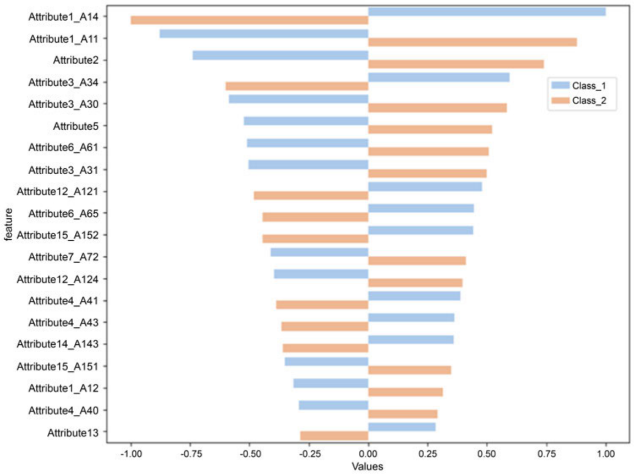

FIGURE 8. Results of the AIME global feature importance on the Statlog (German Credit) dataset.

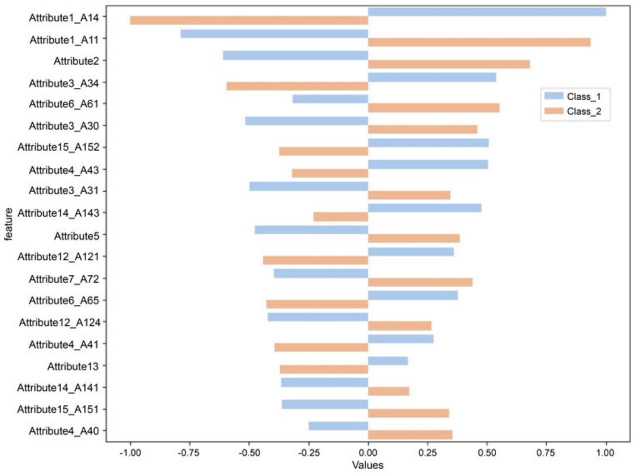

FIGURE 9. Results of the HuberAIME global feature importance on the Statlog (German Credit) dataset.

ranking from AIME in the importance evaluation of global features in datasets with many outliers, it is suggested that the robustness affects the estimation results.

C. QUANTITATIVE EVALUATION

This section presents an evaluation of the results of the Wine, Adult, and Statlog (German Credit) datasets and a discussion of the effectiveness of HuberAIME. SHAP, AIME, and HuberAIME were compared using the six metrics descriptive accuracy, sparsity, stability, efficiency, robustness, and completeness proposed in [40]. Subsequently, the statistical significance of the differences was verified using the Friedman test.

First, because the Wine dataset contains ''clean'' data with almost no outliers, the values for AIME and HuberAIME were very close, and a difference in robustness was not clearly apparent. Table 4 presents the evaluation details. Whereas SHAP showed a slightly different trend in terms of the average values of descriptive accuracy (min–max, avg) and sparsity (min–max, avg), AIME and HuberAIME exhibited very similar values. All methods showed high reproducibility, with a stability of 1.0000. In addition, the efficiency was 10.92 s for SHAP and 0.00 s for AIME and HuberAIME, indicating a very short processing time. The robustness (bias/adv) did not differ significantly owing to few outliers in the Wine dataset. The completeness was also 0.0000 for all three methods, indicating no difference between the surfaces. According to the results of the Friedman test (Table 5), there was no significant difference among the three methods at a significance level of 5% with a p-value of approximately 0.5101. That is, there was no significant difference in the explanatory results when using SHAP, AIME, or HuberAIME on the Wine dataset, and the benefits of the robustness were not apparent. This supports the nature of robust statistics, based on which HuberAIME behaves almost identically to AIME owing to least squares in clean data.

Second, because the Adult dataset contains numerous outliers, the differences between AIME and HuberAIME on this dataset were more pronounced. According to the results in Table 6, the average descriptive accuracy values were 0.80 for SHAP and 0.81 for AIME and HuberAIME, which were almost the same, whereas the sparsity was extremely high (0.97) for SHAP, but only 0.83 for AIME and HuberAIME. The stability differed significantly between SHAP (0.6667) and AIME and HuberAIME (1.0000), with SHAP being susceptible to variations owing to the sampling and approximation processes, whereas AIME and HuberAIME obtained an almost perfectly stable explanation. In terms of the efficiency, SHAP had a very high computational cost of 159.33 s, whereas AIME and HuberAIME completed the process in 0.27 s. AIME and HuberAIME achieved the same values for robustness and completeness. Despite the significant difference between AIME and HuberAIME in the Wilcoxon test, the results were interesting in that they agreed with the average index. The results of the Friedman test shown in Table 7 indicate a significant difference ($p = 1.097792e\text{-}15$), and the results of the post-hoc comparison (Wilcoxon signed-rank test) shown in Table 8 indicate a statistically significant difference in both the SHAP vs. HuberAIME and AIME vs. HuberAIME pairs, which supports the possibility that HuberAIME operates differently in an environment with many outliers.

The results for the Statlog (German Credit) dataset are summarized in Table 9. SHAP showed high sparsity (0.95) and a long execution time (150.69 s), whereas AIME/HuberAIME maintained an excellent computational efficiency of 0.01 s. The descriptive accuracy (min–max, avg) values were 0.72–1.00, 0.81 for SHAP and 0.79–1.00, 0.86 for AIME and HuberAIME, with AIME and HuberAIME achieving slightly higher average values. However, the results of the Friedman test shown in Table 10 ($p = 0.159551$) indicate no significant difference, and there was no significant statistical difference in performance among the three methods. This indicates that, despite the presence of a certain number of outliers in the German Credit data, AIME

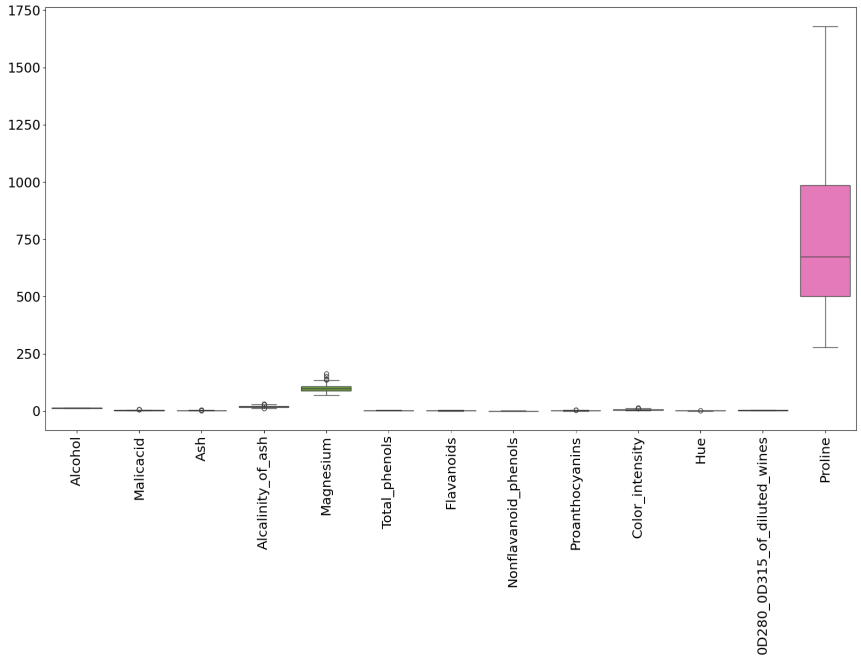

FIGURE 10. Box plot of the Wine dataset [27].

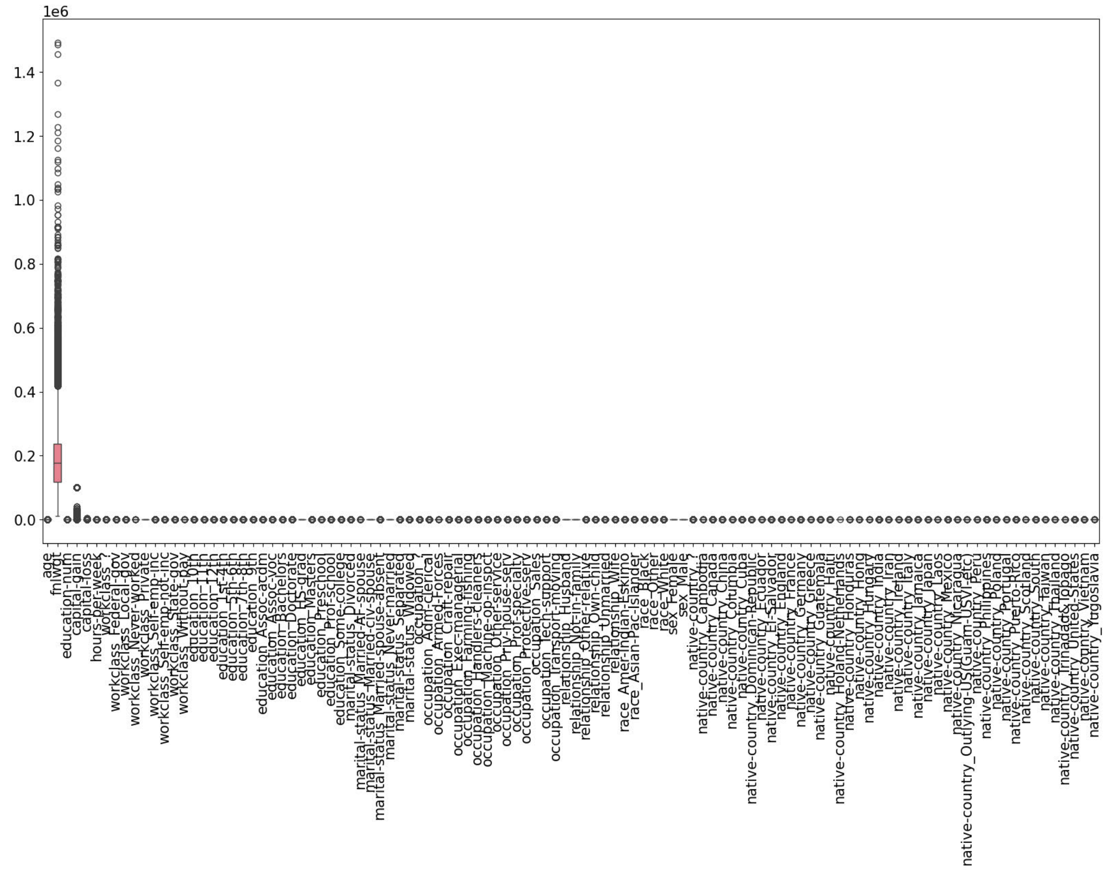

FIGURE 11. Box plot of the Adult dataset [28].

alone could absorb some of the outliers, or the outliers were not sufficiently extreme to have a significant impact; thus, the difference between AIME and HuberAIME was not apparent. Nevertheless, as AIME and HuberAIME showed the same

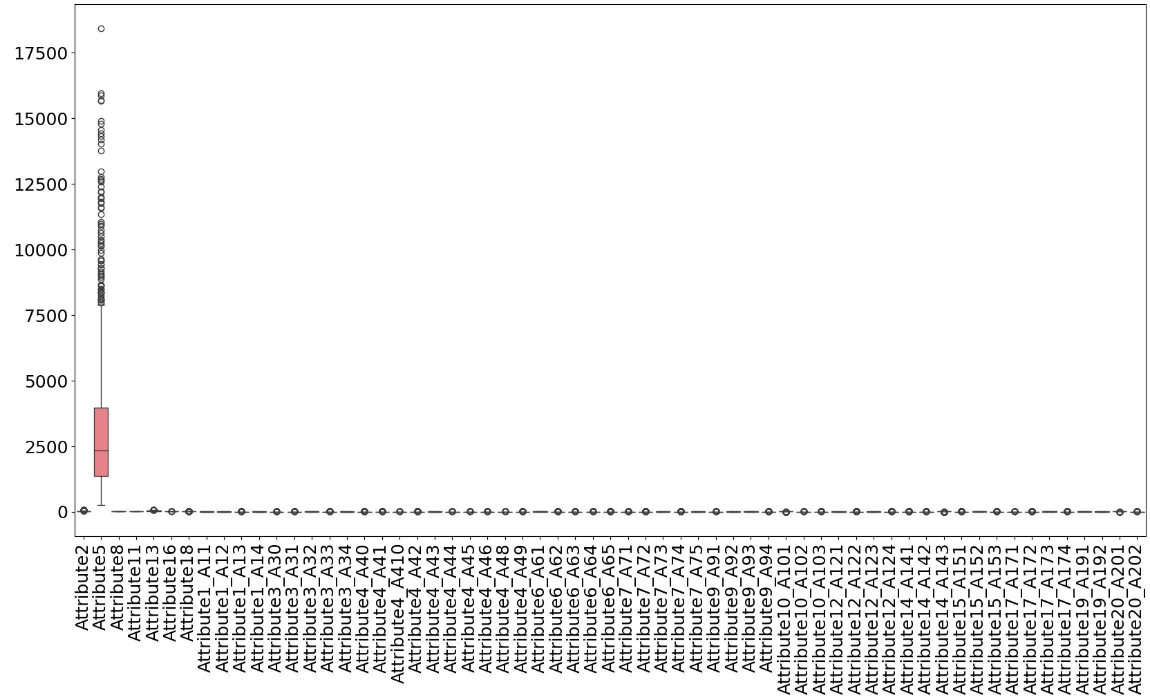

FIGURE 12. Box plot of the Statlog (German Credit) dataset [29].

TABLE 4. Results for the six indicators on the wine dataset.

	Descriptive accuracy (min–max, avg)	Sparsity (min–max, avg)	Stability	Efficiency (s)	Robustness (bias/adv)	Completeness
SHAP	0.42–1.00, 0.61	0.94–0.95, 0.94	1.0000	10.92	1.00 / 0.00	0.0000
AIME	0.42–1.00, 0.70	0.43–0.60, 0.50	1.0000	0.00	0.00 / 0.00	0.0000
HuberAIME	0.42–1.00, 0.70	0.43–0.60, 0.50	1.0000	0.00	0.00 / 0.00	0.0000

TABLE 5. Friedman test for the wine dataset.

Chi-square	p-value	Significance (5%)
1.346405	0.510072	Not significant

results in terms of the average index and had equivalent calculation costs, it can again be confirmed that the introduction of the robust loss does not have a negative impact.

Considering the overall results, in an environment with a few outliers, i.e., the Wine dataset, there was no significant difference between AIME and HuberAIME, and no statistically significant difference among the three methods, including SHAP. However, the Friedman test and Wilcoxon post-hoc comparison showed a significant difference between AIME and HuberAIME on the Adult dataset, which has a relatively high number of outliers, confirming that HuberAIME, which uses a robust loss function, is more resistant to outliers. On the German Credit dataset, the effect of outliers may have been moderate, and AIME could deal with it sufficiently. Thus, the results did not show a statistically significant difference. Nevertheless, the fact that HuberAIME can operate at the same computational cost as AIME and is potentially resistant to outliers is a significant advantage for industrial applications. In addition, although SHAP theoretically provides the most rigorous local explanation, it has a high computational cost, and when many outliers are present, the results are easily biased in only the positive direction, depending on the aggregation method. Overall, although HuberAIME is as efficient as AIME, it exhibits substantial superiority in datasets with apparent outliers. This quantitative evaluation confirmed that the stability and accuracy of the explanation were improved.

V. DISCUSSION

According to the experimental results, the performances of AIME and HuberAIME were almost identical on the dataset with very few outliers (Wine), and the Friedman test revealed

TABLE 6. Results for the six indicators on the adult dataset.

	Descriptive accuracy (min–max, avg)	Sparsity (min–max, avg)	Stability	Efficiency (s)	Robustness (bias/adv)	Completeness
SHAP	0.76–0.88, 0.80	0.97–0.97, 0.97	0.6667	159.33	1.00 / 0.00	0.2696
AIME	0.76–0.88, 0.81	0.83–0.83, 0.83	1.0000	0.27	0.50 / 0.00	0.2279
HuberAIME	0.76–0.88, 0.81	0.83–0.83, 0.83	1.0000	0.27	0.50 / 0.00	0.2279

TABLE 7. Friedman test for the adult dataset.

Chi-square	p-value	Significance (5%)
68.890951	1.097792e-15	Significant

TABLE 8. Wilcoxon signed-rank test for the adult dataset.

	W-stat	p-value	Significance (5%)
SHAP vs. AIME	11583.5	0.883716	Not significant
SHAP vs. HuberAIME	7165.0	7.379135e-07	Significant
AIME vs. HuberAIME	5566.0	5.906642e-11	Significant

no statistically significant difference between the two methods. This finding aligns with the basic properties of robust statistics, whereby the Huber loss behaves the same as the squared error in regions with small residuals. Thus, when outliers are extremely scarce, they produce estimation results that are approximately equivalent to those of AIME. Moreover, there were no statistically significant differences among the three methods, including SHAP, which suggests that all methods could achieve high reproducibility and explanatory accuracy on this "clean" dataset.

In contrast, a clear difference emerged between AIME and HuberAIME, and a statistically significant difference was confirmed in the post-hoc Wilcoxon comparison on the Adult dataset, which is presumed to contain many outliers (or exhibit a skewed distribution). Notably, whereas SHAP aggregates local contributions, AIME and HuberAIME can evaluate features globally in a manner that explicitly shows both the positive and negative contributions. However, AIME may overlook certain features owing to the influence of outliers, whereas HuberAIME is considered to capture the typical model behavior by downweighting outlier samples via a hybrid loss. In terms of the computational cost, SHAP requires a heavy workload owing to numerous model evaluations, whereas AIME and HuberAIME are more lightweight, and it is noteworthy that HuberAIME retains almost the same speed as that of AIME, even with the addition of robustness.

Furthermore, although the average metrics for AIME and HuberAIME were exactly the same on the Statlog (German Credit) dataset, the Friedman test demonstrated no significant difference, indicating that AIME alone may be sufficiently robust when outliers are present, provided that they are not excessively severe. Given that introducing the robust loss does not affect the computational efficiency or accuracy relative to AIME, HuberAIME stands out as a type of "insurance": it imposes no penalty in scenarios with few outliers, yet offers significant benefits when outliers are abundant.

The combined quantitative and qualitative evaluations strongly support the view that HuberAIME, as an extension of AIME, enhances the explanatory accuracy and stability in datasets with evident outliers by leveraging increased robustness, whereas it performs at the same level as the original least-squares-based AIME in cases with few outliers. Thus, HuberAIME can be considered to preserve the performance on "clean" data while providing distinct advantages under "dirty" data conditions, which is a highly desirable attribute in practical applications. In addition, the comparison of HuberAIME and SHAP highlights that, although SHAP can theoretically compute exact Shapley values, it incurs high computational costs, and its robustness is dependent on the aggregation scheme and background data selection. Consequently, AIME or HuberAIME may offer a more efficient and broadly applicable alternative in certain scenarios, depending on the data characteristics and task requirements. These insights underscore the importance of selecting a method that aligns with the presence or absence of outliers and the available computational resources, especially in industrial or large-scale data environments.

In addition to the technical merits of HuberAIME, its robust nature offers clear benefits in real-world scenarios, in which data irregularities and regulatory scrutiny are critical. For example, in finance, loan approval and credit risk models must handle diverse customer profiles and transaction records that often include extreme or erroneous values. The robust explainability of HuberAIME can help financial institutions to justify automated decisions more reliably, especially when demanded by regulations such as the guidelines of the European Banking Authority or local "right to explanation" requirements, by ensuring that high-leverage outliers do not excessively distort the feature attributions. Similarly, in healthcare, patient records frequently contain

TABLE 9. Results for the six indicators on the statlog (German Credit) dataset.

	Descriptive accuracy (min–max, avg)	Sparsity (min–max, avg)	Stability	Efficiency (s)	Robustness (bias/adv)	Completene ss
SHAP	0.72–1.00, 0.81	0.95–0.95, 0.95	1.0000	150.69	1.00 / 1.00	0.0000
AIME	0.79–1.00, 0.86	0.73–0.73, 0.73	1.0000	0.01	0.50 / 0.50	0.0000
HuberAIME	0.79–1.00, 0.86	0.73–0.73, 0.73	1.0000	0.01	0.50 / 0.50	0.0000

TABLE 10. Friedman test for the statlog (German credit) dataset.

Chi-square	p-value	Significance (5%)
3.670782	0.159551	Not significant

anomalies or missing data that can lead to misclassification when standard methods are used. HuberAIME can help maintain accuracy and interpretability even in the presence of such anomalies, supporting more transparent clinical decisions and compliance with regulations such as HIPAA or the EU Medical Devices Regulation. By offering robust and computationally feasible explanations, HuberAIME aligns well with the growing emphasis on trustworthy, high-stakes AI deployment, where both ethical considerations and legal mandates demand transparency, fairness, and reliability in automated decision-making systems.

While HuberAIME offers improved robustness to outliers and retains computational efficiency, it also has certain limitations and requirements. First, the approach is sensitive to hyperparameter selection, most notably the Huber threshold δ. Although a default value ofδ was used throughout the experiments with standardized data, selecting an inappropriate threshold for unscaled features or highly skewed distributions may compromise accuracy or result in failure to mitigate outliers effectively. Similarly, the maximum iteration count and convergence tolerance in IRLS must be tuned to balance the computational cost and solution quality.

Second, HuberAIME assumes that the dataset can be sampled so that YY^T remains invertible (i.e., the model outputs span a sufficient dimension). The pseudo-inverse or IRLS steps may become unstable or degenerate if the dataset is too small or the outputs are highly correlated. Moreover, the method relies on standardization and consistent data preprocessing, particularly with respect to handling missing values and one-hot-encoded categories, to ensure that the residual magnitudes are interpreted consistently.

Third, similar to AIME itself, HuberAIME produces a linear surrogate operator for the inverse mapping, implying that it may not fully capture complex nonlinear interactions that certain closed-box models might learn. Although this linear approximation is advantageous for interpretability, it may be limiting in scenarios where local decision boundaries are extremely nonlinear.

VI. CONCLUSION

Although AIME enables both global and local explanations for closed-box models via an approximate inverse operator based on least squares, it is vulnerable to outliers. To address this issue, this study developed HuberAIME, which introduces the Huber loss into AIME to provide outlier tolerance while retaining the same level of computational efficiency. Experiments were performed on three datasets, and the results were evaluated in comparison with the conventional AIME and SHAP using six metrics along with statistical tests.

First, there was no statistical difference between AIME and HuberAIME on the Wine dataset, which contains very few outliers, and all three methods including SHAP exhibited high stability and explanatory performance. As suggested by robust statistics theory, the Huber loss behaves similar to the squared error in regions with small residuals; hence, when outliers are almost absent, HuberAIME functions in the same manner as AIME and does not incur additional overheads.

In contrast, a clear difference emerged between AIME and HuberAIME on the Adult dataset, which contains many outliers and extreme values. HuberAIME achieved high explanatory accuracy and stability within the same computational budget as AIME, while mitigating the effects of outliers. This outcome indicates that IRLS with the Huber loss prevents the overall estimation from being distorted by downweighting outliers, thereby extracting the contributions of genuinely relevant features more reliably. Furthermore, the average metric values for AIME and HuberAIME were identical, and the Friedman test revealed no significant difference between the models on the Statlog (German Credit) dataset. This suggests that the presence of outliers is moderate enough for AIME alone to exhibit a degree of robustness in certain scenarios. Nevertheless, the ability of HuberAIME to maintain the same cost and accuracy as AIME, even with the introduction of a robust loss function, again demonstrates that it can serve as an effective measure for managing risk in real-world applications.

The comparisons with SHAP indicated that although SHAP can theoretically provide the most rigorous local explanations, it has a high computational cost and its robustness depends on the selection of background data and aggregation methods. AIME or HuberAIME are often more likely to achieve sufficient explanatory performance while reducing the computational load substantially, depending on the data and task characteristics. In particular, HuberAIME shows promise for practical use in industrial settings and other environments with critical demands on computational resources and result interpretation, particularly for stable explanations when dealing with ''dirty'' data that include outliers.

In conclusion, HuberAIME overcomes the outlier vulnerability that is inherent in conventional AIME, preserves its original computational efficiency and explanatory framework, and surpasses previous approaches in terms of explanatory accuracy and stability when outliers occur frequently. Its similar behavior to that of AIME in the absence of outliers and superior performance under heavier outlier conditions owing to its robustness are particularly appealing from reliability and cost-efficiency perspectives in practical applications. In the future, expanding HuberAIME to more diverse fields and automating the hyperparameter (threshold) selection could further promote the adoption of XAI in an even broader range of scenarios.

APPENDIX

A DATASET CHARACTERISTICS (WINE, ADULT, AND STATLOG (GERMAN CREDIT) DATASETS)

In this appendix, the characteristics of the Wine [27], Adult [28], and Statlog (German Credit) [29] datasets used in the experiments are described. These three datasets are open access [27], [28], [29] and can be freely downloaded.

The box plot of the Wine dataset is shown in Fig. 10. It can be observed that most features lie within a relatively small range, whereas the Proline feature extends to a markedly larger scale. This indicates that the amount of proline in wine varies substantially compared with the other components, creating samples that appear to be outliers. However, it is also highly likely that the proline content genuinely differs depending on the wine type and production conditions. Although several observations show slightly elevated Magnesium values, these are not as pronounced as those for Proline; overall, the Wine dataset does not exhibit particularly strong outlier tendencies.

In the box plot of the Adult dataset (Fig. 11), many categorical variables appear compressed near zero—likely because they take on values close to the binary 0/1—while several outliers are visible among the continuous variables, notably ''capital-gain,'' which spans a large range. Indeed, features such as capital-gain and fnlwgt occasionally reach high or extreme values that could strongly influence model estimation. Moreover, elements like education-num and hours-per-week may also display considerable variability, suggesting that the Adult dataset includes relatively ''dirty'' samples.

As shown in the box plot of the Statlog (German Credit) dataset in Fig. 12, some of the continuous values in Attribute2 reach the thousands or even exceed 10,000, implying that certain samples represent very high debts or incomes. Meanwhile, most categorical features are binary and thus cluster near zero on the axis. Owing to this scale discrepancy with the continuous variables, many of the box plots converge around zero. Nevertheless, there are instances of extremely large spikes, suggesting that these samples may behave as outliers when estimating default risk.

Overall, the Wine dataset exhibits a comparatively ''clean'' distribution, apart from Proline; in contrast, the Adult dataset contains broad continuous variations (e.g., capital-gain) and the German Credit dataset spans even higher numerical ranges. These differences in outlier prevalence and distributional characteristics directly affect both the potential distortion of model parameters and the performance of explanation methods (e.g., AIME or HuberAIME). Hence, understanding the data distribution beforehand and carefully planning how to handle outliers are crucial.

ACKNOWLEDGMENT

The author thank Editage [www.editage.com] for English language editing. He gratefully acknowledge the role of OpenAI's ChatGPT, deepL, and Paperpal AI systems in facilitating discussion and inspiring innovative ideas during the writing process. Although AI-generated text was not directly used in this study, it helped to conceptualize and refine the content.

REFERENCES

[1] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani, ''A survey on XAI and natural language explanations,'' *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2023, Art. no. 103111, doi: 10.1016/j.ipm.2022.103111.

[2] W. Saeed and C. Omlin, ''Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities,'' *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110273, doi: 10.1016/j.knosys.2023.110273.

[3] Y.-N. Chuang, G. Wang, F. Yang, Z. Liu, X. Cai, M. Du, and X. Hu, ''Efficient XAI techniques: A taxonomic survey,'' 2023, *arXiv:2302.03225*.

[4] G. Schwalbe and B. Finzel, ''A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts,'' *Data Mining Knowl. Discovery*, vol. 38, no. 5, pp. 3043–3101, Sep. 2024, doi: 10.1007/s10618-022-00867-8.

[5] R. Dazeley, P. Vamplew, and F. Cruz, ''Explainable reinforcement learning for broad-XAI: A conceptual framework and survey,'' *Neural Comput. Appl.*, vol. 35, no. 23, pp. 16893–16916, Mar. 2023, doi: 10.1007/s00521-023-08423-1.

[6] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. B. Amin, and B. Kang, ''Survey on explainable AI: From approaches, limitations and applications aspects,'' *Hum.-Centric Intell. Syst.*, vol. 3, no. 3, pp. 161–188, Aug. 2023, doi: 10.1007/s44230-023-00038-y.

[7] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, ''Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence,'' *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805, doi: 10.1016/j.inffus.2023.101805.

[8] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, ''A review of trustworthy and explainable artificial intelligence (XAI),'' *IEEE Access*, vol. 11, pp. 78994–79015, 2023, doi: 10.1109/ACCESS.2023.3294569.

[9] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan, ''Explainable AI (XAI): Core ideas, techniques, and solutions,'' *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.

[10] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black-Box Models Explainable*. Morrisville, NC, USA: LuluCom, 2020.

[11] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, ''Explainable artificial intelligence: A comprehensive review,'' *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[13] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8827–8836, doi: 10.1109/CVPR.2018.00920.

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[15] 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, Jul. 2019, doi: 10.3390/electronics8080832.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "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.

[17] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*. New York, NY, USA: ACM, Jun. 2022, pp. 2239–2250, doi: 10.1145/3531146.3534639.

[18] W. Samek and K. R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpret*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen and K. R. Müller, Eds., Cham, Switzerland: Springer, Sep. 2019, pp. 5–22, doi: 10.1007/978-3-030-28954-6_1.

[19] M. v. Lent, W. M. Fisher, and M. R. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proc. Nat. Conf. Artif. Intell.* Cambridge, MA, USA: AAAI Press, Jul. 2004, pp. 900–907.

[20] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/aimag.v40i2.2850. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850

[21] M. Ribeiro, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.

[22] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017.

[23] T. Nakanishi, "Approximate inverse model explanations (AIME): Unveiling local and global insights in machine learning models," *IEEE Access*, vol. 11, pp. 101020–101044, 2023, doi: 10.1109/ACCESS.2023.3314336.

[24] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964, doi: 10.1214/aoms/1177703732.

[25] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, May 1974, doi: 10.1080/00401706.1974.10489171.

[26] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Commun. Statist. Theory Methods*, vol. 6, no. 9, pp. 813–827, Jan. 1977, doi: 10.1080/03610927708827533.

[27] S. Aeberhard and M. Forina, "Wine [dataset]," UCI Mach. Learn. Repository, Univ. California, Irvine, CA, USA, Tech. Rep., 1992, doi: 10.24432/C5PC7J.

[28] B. Becker and R. Kohavi, "Adult [dataset]," UCI Mach. Learn. Repository, Univ. California, Irvine, CA, USA, Tech. Rep., 1996, doi: 10.24432/C5XW20.

[29] H. Hofmann, "Statlog (german credit data) [dataset]," UCI Mach. Learn. Repository, Univ. California, Irvine, CA, USA, Tech. Rep., 1994, doi: 10.24432/C5NC77.

[30] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 1, Jan. 2025, Art. no. 2300272, doi: 10.1002/aisy.202400304.

[31] M. Saarela and V. Podgorelec, "Recent applications of explainable AI (XAI): A systematic literature review," *Appl. Sci.*, vol. 14, no. 19, p. 8884, Oct. 2024, doi: 10.3390/app14198884.

[32] T. Sim, S. Choi, Y. Kim, S. H. Youn, D.-J. Jang, S. Lee, and C.-J. Chun, "EXplainable AI (XAI)-based input variable selection methodology for forecasting energy consumption," *Electronics*, vol. 11, no. 18, p. 2947, Sep. 2022, doi: 10.3390/electronics11182947.

[33] M. Eik, A. Kose, H. N. okmabad, and J. Belikov, "Explainable AI based system for supply air temperature forecast," 2025, *arXiv:2501.05163*.

[34] R. P. Bora, P. Terhorst, R. Veldhuis, R. Ramachandra, and K. Raja, "SLICE: Stabilized LIME for consistent explanations for image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2024, pp. 10988–10996, doi: 10.1109/CVPR52733.2024.01045.

[35] D. Patil. (Dec. 3, 2024). *Explainable Artificial Intelligence (XAI) for Industry Applications: Enhancing Transparency, Trust, and Informed Decision-Making in Business Operation*. SSRN. [Online]. Available: https://ssrn.com/abstract=5057402

[36] B. Kantz, C. Staudinger, C. Feilmayr, J. Wachlmayr, A. Haberl, S. Schuster, and F. Pernkopf, "Robustness of explainable artificial intelligence in industrial process modelling," 2024, *arXiv:2407.09127*.

[37] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[38] T. Nakanishi. *AIME: Approximate Inverse Model Explanations*. Accessed: Mar. 6, 2025. [Online]. Available: https://github.com/ntakafumi/aime

[39] T. Nakanishi. *AIME: Approximate Inverse Model Explanations*. Accessed: Mar. 6, 2025. [Online]. Available: https://pypi.org/project/aime-xai/

[40] O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," *IEEE Access*, vol. 12, pp. 23954–23988, 2024, doi: 10.1109/ACCESS.2024.3365140.

TAKAFUMI NAKANISHI (Member, IEEE) was born in Ise, Mie, Japan, in 1978. He received the Ph.D. degree in engineering from the Graduate School of Systems and Information Engineering, University of Tsukuba, in April 2006. Since April 2006, he has been engaged in research and development of knowledge-cluster systems and text/data-mining methods with the National Institute of Information and Communications Technology (NICT). In April 2014, he was appointed as an Associate Professor with the Global Communication Center, International University. He became an Associate Professor with the Department of Mathematical Engineering, Faculty of Engineering, in 2018, and the Department of Data Science, Musashino University, in 2019. He is currently a Professor with the School of Computer Science, Tokyo University of Technology. His research interests include explainable artificial intelligence (XAI), data mining, affective information processing, and media-content analysis.