IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 August 2023, accepted 3 September 2023, date of publication 11 September 2023,
date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314336

RESEARCH ARTICLE

Approximate Inverse Model Explanations (AIME): Unveiling Local and Global Insights in Machine Learning Models

TAKAFUMI NAKANISHI, (Member, IEEE)
Department of Data Science, Musashino University, Tokyo 135-8181, Japan
e-mail: takafumi.nakanishi@ds.musashino-u.ac.jp

ABSTRACT Data-driven decision-making has become pervasive in the fields of interpretive machine learning and Explainable AI (XAI). While both fields aim to improve human comprehension of machine learning models, they differ in focus. Interpretive machine learning centers on deciphering outcomes in transparent, or 'glass-box,' models, whereas XAI focuses on creating tools for explaining complex 'black-box' models in a human-understandable way. Some existing interpretable machine learning and explainable AI methods have utilized a forward problem to derive how the prediction and estimation output results of a black-box model change with respect to the input. However, methods adopting the forward problem lead to non-intuitive explanations. Therefore, hypothesizing that the inverse problem can yield more intuitive explanations, we propose approximate inverse model explanations (AIME), which offer unified global and local feature importance by deriving approximate inverse operators for black-box models. Additionally, we introduce a representative instance similarity distribution plot, aiding comprehension of the predictive behavior of the model and target dataset. In our experiments with LightGBM, AIME proved effective across diverse data types, from tabular and handwritten digit images to text data. Results demonstrate that AIME's explanations are not only simpler but more intuitive than those generated by well-established methods like LIME and SHAP. It also visualizes similarity distribution with the target dataset, illustrating the relation between different predictions. Furthermore, AIME estimates local and global feature importance and provides fresh insights by visualizing the similarity distribution between representative estimation instances and the target dataset.

INDEX TERMS Approximate inverse models, explainable artificial intelligence (XAI), feature importance, generalized inverse matrices, interpretability, model explanation techniques.

I. INTRODUCTION

In today's digital age, machine learning and artificial intelligence (AI) play a pivotal role in driving the advancement of fields as diverse as automated driving, medical diagnostics, and financial transactions, with their ubiquitous presence shaping the essence of modern decision-making. However, given their growing influence on important decisions, it is crucial to understand how the predictions and estimates are derived and which data features affect the outcomes the most. These insights include identifying unfair biases, confirming the reliability of the model, and assessing accountability from both legal and ethical perspectives. In this context, complex models such as deep learning and black-box models are problematic because it is difficult to intuitively understand their internal behavior.

Many studies have analyzed interpretable machine learning and explainable AI (XAI) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], which have gained significant attention in recent years as data-driven decision-making has become more prevalent.

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

In the case of complex models, local interpretation and explanation methods help obtain a surrogate model close to the target model to further observe its behavior with respect to a data instance of interest.

Existing interpretable machine learning and XAI methods, such as LIME [1] and SHAP [3], extract the significance of local explanations by constructing and observing approximate black-box models. These methods determine the extent to which a feature in a dataset contributes to the predictions and estimations. To understand the behavior of black-box models, it is important to show the relationship between the input data and the predictions and estimates derived from it. Thus, there is a need to develop a new method to evaluate the global and local feature importance and visualize the relevance of the predicted and estimated values. Some existing interpretable machine learning and XAI methods have utilized forward analysis or a forward problem to derive how the prediction and estimation output of a black-box model changes with respect to the input. Alternatively, when seeking explanations for black-box models, it is often useful to address the inverse problem of understanding why the predictions and estimations are derived for a given input. Because some existing methods that adopt the forward problem lead to non-intuitive explanations, we hypothesize that solving the inverse problem of the black-box model would yield more intuitive explanations.

In this study, we introduce approximate inverse model explanations (AIME), a novel approach that explains the behavior of black-box models and the properties of data by deriving approximate inverse operators for the black-box model and using them in conjunction with the target dataset to estimate the significance of local and global features. AIME has proven versatile enough for its effective application across diverse data types. This unique approach manages model complexities by using representative instance similarity distribution plots to visualize relationships among the approximate inverse operators. Thus, through multiple explanation and visualization techniques, AIME offers a comprehensive understanding of the model behavior, marking a significant advancement in the field of XAI. The distinctive characteristic of this method is that it estimates and constructs an approximate inverse operator for the black-box model from data and estimates. This further helps evaluate the significance of features on the estimates obtained from the black box, both locally and globally. Furthermore, we propose a representative instance similarity distribution plot that uses representative estimation instances to identify ''ideal'' or ''typical'' instances, wherein a model may belong to a particular value, and understand the predictive behavior of the representative estimation instances and target dataset. In addition, we visualize the similarity distribution with the target dataset to demonstrate how a particular prediction is related to other predictions.

Fig. 1 presents an overview of the proposed AIME method, which provides four explanations: global feature importance, representative estimation instance, local feature importance, and representative instance similarity distribution plot. The global feature importance explains the behavior of the model. A representative estimation instance provides a typical example for deriving a specific prediction and estimate. The local feature importance explains why a particular prediction or estimation is derived from a certain instance. Lastly, the representative instance similarity distribution plot provides insight into the complexity of the target dataset distribution, i.e., it helps visualize the similarity distribution between a representative estimation instance and each instance in a target dataset. AIME allows not only global and local explanations but also explanations based on the generation of typical instances and the distribution of the target dataset that would derive specific predictions and estimations.

Unlike existing interpretable machine learning and XAI methods, AIME provides both global and local feature importance and helps visualize the relationship between these estimations using representative instance similarity distribution plots. Furthermore, AIME can estimate the local and global feature importance and provide a new interpretation by visualizing the similarity distribution between representative estimation instances and the target dataset. This method could be potentially integrated into several applications, including analyzing medical data, processing natural language, and evaluating creditworthiness. It would be particularly valuable in cases where it is necessary to interpret and explain the results of machine learning estimations.

This study demonstrated that AIME can be effectively applied to different types of data (tabular, image, and text). Although deriving approximate inverse operators based on a generalized inverse matrix can pose a few challenges when applied to highly complex black-box models, the AIME representative instance similarity distribution plot provides important insights into these complexities.

The main contributions of this study are as follows:

- It introduces the AIME approach, which can explain the behavior of black-box models and data properties by deriving the approximate inverse operators.
- It discusses the capability of AIME to derive multiple explanations, including the overall (global) and individual (local) explanations of black-box models, and visualize the distribution of predictions and estimates using representative instance similarity distribution plots. These plots provide insights into the complexities of the model estimations. The insights obtained from AIME offer a clearer understanding of the model predictions and estimates and identify potential challenges in deriving them. This is particularly useful when working with complex models because it helps identify areas where the model might face difficulties in making accurate predictions.

The remainder of this paper is organized as follows: Section II discusses previous works related to this study. Section III presents the formulation of AIME. Section IV describes the implementation of AIME and its application to various data formats, along with experiments comparing AIME with

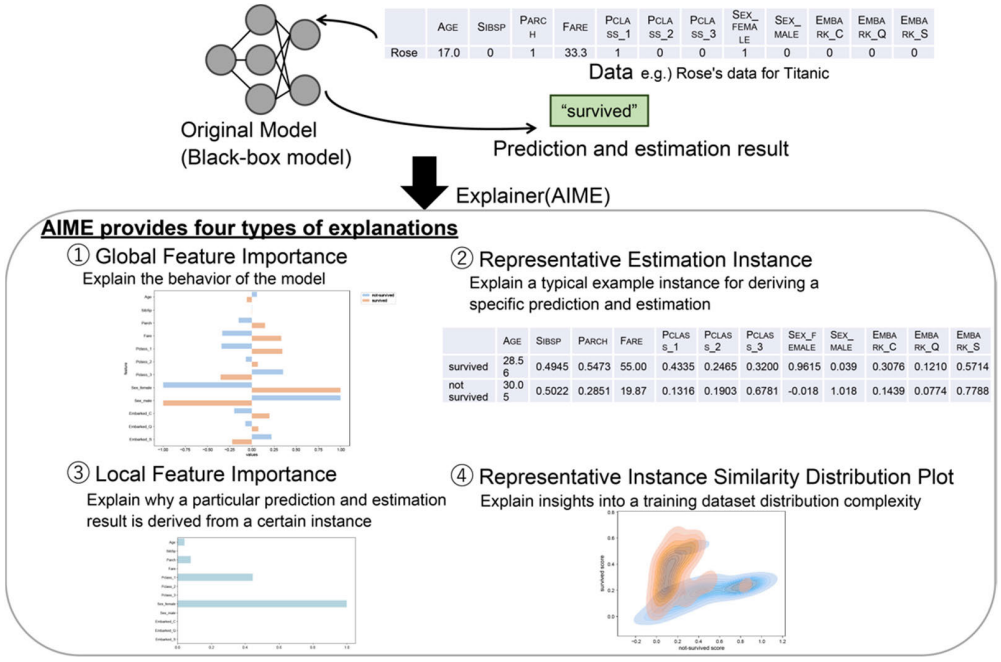

FIGURE 1. Explanation of the original model (black-box model) by the proposed AIME method. The original model (black-box model) outputs "survived" as the prediction from the Rose data (in accordance with the movie "Titanic") in the example of the Titanic dataset [44]. AIME provides four explanations: global feature importance, representative estimation instance, local feature importance, and representative instance similarity distribution plot. The global feature importance explains the behavior of the model. In the case of the Titanic data, we show which features contribute to the "survived" and "not survived" predictions (see Fig. 2). The representative estimation instance explains a typical example instance for deriving a specific prediction and estimate. In the case of the Titanic data, AIME creates "ideal" or "typical" instances representing "survived" and "not survived" as representative estimation instances (see Table 1). The local feature importance explains why a particular prediction estimation is derived from a certain instance. In the case of the Titanic data, AIME extracts features that contribute to why Rose is predicted as "survived" (see Fig. 7). The representative instance similarity distribution plot offers insights into a target dataset distribution complexity by helping visualize the similarity distribution between a representative estimation instance and each instance of the target dataset. It allows for a better understanding of the behavior of the model and the reliability of its estimates. In the case of the Titanic data, the distribution of data sets that are predicted as "survived" (red in the plot) and the distribution of data sets that are predicted as "not survived" (blue in the plot) are visualized (see Fig. 3).

existing methods, such as LIME and SHAP. The paper concludes with a summary in Section V.

II. RELATED WORKS

In recent years, exhaustive research has been conducted on machine learning interpretation and XAI. A recent study presented a review of available XAI methods [32], [33], [34], [35], [36], [37], [38], [39], [40].

The significance of XAI research and implementation was also mentioned in another study [34]. Many users consciously and subconsciously use machine learning and AI in various fields to make daily decisions. Although humans remain personally responsible for decisions and actions, they cannot account for the internal processes of existing machine learning and AI systems. Thus, responsibility for predictions and inferences derived from these tools is ambiguous. Studies on XAI have multiplied considerably because they help understand how machine learning or AI models make predictions and estimates. The interpretability of machine learning models and XAI is particularly important for trust [1], fairness [2], legal and regulatory requirements [41], and model improvement [3]. Another study [1] differentiated two types of trust: (1) trust in a particular prediction, which relates to whether a user is sufficiently confident in an individual prediction to act based on it, and (2) trust in the model as a whole, which refers to the user's confidence in the model's ability to behave reasonably if deployed. In [2], a fairness method that aims to prevent individual discrimination in classification results from machine learning models was presented. The EU General Data Protection Regulation (GDPR) effectively creates a ''right to explanation,'' by which users can ask for an explanation of decisions made about them based on machine learning models and AI [41]. This means that the XAI methodology is important for complying with these regulations. In [3], it was shown that increasing the interpretability of machine learning models is also important for improving accuracy.

A detailed explanation of various classifications of XAI methods is provided in [39]. According to one of the classifications presented, XAI methods are divided into ante-hoc and post-hoc methods, which are further divided into model-specific and model-agnostic methods.

In the ante-hoc method, models were built focusing on interpretability during the design and training stages. Typical ante-hoc methods include those realized using transparent

models, such as linear regression and logistic regression [4], and those that estimate the importance of features derived from the uncertainty of the weights of neural networks as random variables [5].

In contrast, the post-hoc method involves the interpretation of the predictions of a trained model. Model-specific methods include interpretability methods that are only applicable to specific types of models, primarily because they exploit the internal structure and properties of the model. Although they yield a high interpretability, they have a narrow range of applicability. For example, Grad-CAM [6] is an interpretability method for convolutional neural networks. DeepLIFT [7] is an effective method for computing importance scores in neural networks. Other model-specific interpretability methods for deep neural networks have also been proposed in [6], [7], [8], [9], [10], and [11]. Model-agnostic methods are interpretability tools applicable to any machine learning model, i.e., these methods can be integrated regardless of the model type; however, their interpretation is less detailed than that of model-specific methods. This type of approach includes a partial dependence plot [12], [13], which helps visualize the estimated value and impact of each feature, and an individual conditional explanation [14], which evaluates the importance of a feature through random reorder or removal of certain features, permutation feature importance [15], and leave-one-feature-out (LOFO) importance [16], [17]. In [18], a feature importance extraction method based on sensitivity analysis was presented. In this method, feature importance is extracted by varying the input features of all subsets to reflect the interaction and redundancy among the features. However, this method is computationally expensive for high-dimensional data because it evaluates the impact of n features on all n^2 subsets for n number of features. In general, feature importance extraction methods based on sensitivity analysis have proven computationally demanding for high-dimensional data sets. A quantitative input-influence measure was adopted in a method aimed at quantitatively evaluating the impact of individual input variables on algorithm prediction and estimation [19]. For instance, LIME [1] is a model-agnostic method aimed at generating explanations for individual predictions and estimates by evaluating the contribution of specific features. Other methods, such as ALIME [20], DLIME [21], OptiLIME [22], ILIME [23], QLIME-a [24], and S-LIME [25], have been proposed as extensions of LIME [1]. SHAP [3] is a model-agnostic method aimed at interpreting the contribution of each feature to a given prediction. Shapley values from cooperative game theory can be used to represent the contribution of each feature to the predicted outcome. Shapley Flow [26] applied SHAP to determine direct and indirect feature contributions when a causal graph was provided. Shapley Chains [27] incorporated label interdependence into the explanation design process to ensure that explanations reflect the interdependence of multiple output forecasts. Anchors [28] is a model-agnostic interpretability method that provides rule-based explanations.

A method was proposed [29] for extracting rule-like descriptions for classification models based on the contribution matrix derived from non-negative matrix factorization (NMF). A new visualization technique for convolutional networks was presented to understand the function of the intermediate feature layers and the behavior of the classifiers [30]. A method for synthesizing input exemplars that maximize the desired estimation results using generative models, such as VAEs and GANs, was presented in [31].

Notably, the exploration of feature importance in black-box models has garnered attention across various disciplines. Deep-learning frameworks have been developed for ranking feature importance to understand the roles of features within neural architectures. For example, possibly influenced by deep learning paradigms, the field of energy and electrical systems has moved towards utilizing advanced techniques aimed at securing critical infrastructure against threats. Moreover, to improve model interpretability, hybrid approaches to dynamic feature importance analysis are being developed by merging traditional and contemporary methodologies. Another area that has been attracting interest and with potentially broad applicability is the extraction of weak physical cues from noise signals. These diverse studies reinforce the critical nature of our inquiry into improving the interpretability of complex models and understanding the significance of features within them.

This study focuses on post-hoc, model-agnostic methods. In other words, it centers on interpretability methods that can be integrated into any machine learning model. In particular, the study in [3] defines methods such as LIME [1] and SHAP [3] as additive feature-attribution model-agnostic methods. Additive feature-attribution methods are expressed by the following formula:

$$g'\left(z'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

where g is the explanatory model for the original model f, $z' \in \{0, 1\}^M$, M is the number of simplified input features, and $\phi_i \in R$. The output $f(x)$ of the original model was approximated by attributing an effect ϕ_i to each feature and summing the effects of all feature attributions. In [3], it was stated that this formula encompasses several model-agnostic methods. This equation simply adds up the multiplication of ϕ_i (a real number) by $z'_i(\{0, 1\})$ after transforming x_j into z'_i. This equation is valuable not only because it is applicable to most model-agnostic methods, but also because, in terms of interpretability, it expresses the explanatory model in the simplest possible manner.

Existing model-agnostic interpretability methods primarily focus on deriving the significance of the local features. Although it is also possible to determine the importance of global features by computing and visualizing the feature importance for all given data instances, as in SHAP [3], these methods only derive and visualize the feature importance. They do not offer a comprehensive and intuitive

understanding of the complexities associated with the model estimations or the data distribution used for training.

In contrast, the AIME approach proposed in this study provides a visualization of the complexities associated with model estimation, training data distribution, and global and local feature importance. Unlike the additive feature attribution methods presented in [3], AIME derives an approximate inverse operator for the original model. The value of the general inverse matrix that comprises the inverse operator constructed by AIME indicates the contribution of its predicted and estimated values and can express the global feature importance. In addition, the explanatory inverse operator computed in AIME can be used to derive a representative estimation instance that identifies the data points on which the model relies to generate a particular estimate. AIME can compute the importance of local features from these representative estimations and data instances. Furthermore, AIME calculates the similarity between representative estimation instances and each data point in the dataset. This further elucidates the similarity of the distribution, highlighting the complexity and distribution of the dataset when the model outputs a specific prediction or estimated value. In this study, this visualization is referred to as the representative instance similarity distribution plot.

Because AIME is based on the inverse operator of the original model, using a generalized inverse matrix to determine the intended local and global feature importance can be challenging in cases where the original model is overly complex. However, the complexities of the model and the distribution of the data for each estimate can be clarified by using a representative instance similarity distribution plot. Therefore, AIME can provide insights into the behavior of the original model and its feature importance. Furthermore, the representative instance similarity distribution plot in AIME can be leveraged to visualize inherent biases or imbalances in the data. Consequently, by highlighting the similarity of the distribution, AIME can reveal any potential biases in the dataset, making it not only an effective tool for model interpretation but also a diagnostic tool to recognize and understand underlying data biases.

III. FORMULATION OF APPROXIMATE INVERSE MODEL EXPLANATIONS (AIME)

This section describes the formulation of the proposed AIME approach. Section III-A describes the conceptual idea behind the process of deriving explanations for models, predictions, and estimates by creating approximate inverse operators for the original model, which is the key characteristic of AIME. Section III-B presents the formulation of the inverse operator in AIME. Section III-C describes the process for deriving the global feature importance using AIME. Section III-D introduces the concept of representative estimation instances and provides a detailed explanation of the methods used to derive them. Section III-E explains the approach for deriving local feature importance using representative estimation instances. Finally, Section III-F presents the representative instance similarity distribution plot, which is a visualization method that provides an intuitive understanding of the data points on which the model relies for specific predictions or estimates, along with the proportion of these data points in the entire training dataset.

A. CONCEPTUAL FRAMEWORK FOR AIME

Fig. 2 shows a comprehensive schematic of the AIME process. In contrast to many existing model explanation methods that focus on demonstrating the importance of specific features when the model makes individual predictions, providing essentially local feature importance, AIME can generate both global and local explanations for the black-box model. A global explanation interprets how the model behaves and which features contribute the most to the overall prediction, providing a broader view of the model's operation. Conversely, a local explanation focuses on a specific data point or estimate and identifies the characteristics that affect that particular result.

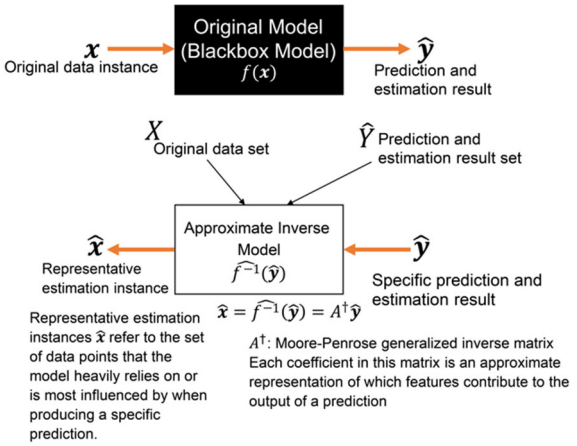

FIGURE 2. Comprehensive schematic of the AIME process. The first approximate inverse operator, denoted as $\widehat{f^{-1}}$, of the original model f is constructed. Then, a generalized inverse matrix $A^{\dagger}$, which composes $\widehat{f^{-1}}$, is defined; each coefficient of $A^{\dagger}$ represents the approximate contribution of each feature when a specific prediction is outputted, which helps derive global feature importance. Local feature importance can be derived from representative estimation instances corresponding to an original data instance, along with prediction and estimation values. Furthermore, based on the original data instances and the representative estimation instances, a similarity distribution plot is generated, which provides a visualization of the distribution of similarities between the data instances and the representative estimation instances.

The basic idea of AIME is to construct an explanation of the original black-box model f by creating an approximate inverse operator $\widehat{f^{-1}}$, which is constructed by deriving the Moore–Penrose generalized inverse matrix $A^{\dagger}$ [42], [43] from the original dataset X and the predictive estimation result set $\hat{Y}$. The approximate inverse operator $\widehat{f^{-1}}$ explains how the model combines the input features to produce a specific output. This is particularly useful for understanding the importance of each feature because it directly indicates the dependence of the output on every input. The approximate inverse operator $\widehat{f^{-1}}$ provides a deeper understanding

of the overall behavior of the model. This further helps to understand the behavior of complex models and investigate the errors and biases in the predictions. Each coefficient of the generalized inverse matrix $A^\dagger$ is an approximation of the contribution of each feature to the output of a specific prediction and estimation because the generalized inverse matrix inverts the operations of the original matrix A. Specifically, the generalized inverse matrix is responsible for the inverse operation of how the original model processes and combines the features to generate predictions. Based on an examination of how the generalized inverse matrix combines features to generate a given prediction, it is possible to understand the contributions of the features to generating such a prediction. Based on these considerations, the AIME outputs each coefficient of the generalized inverse matrix $A^\dagger$ as a global feature importance that approximates the behavior of the entire original model.

Fig. 2 also introduces the concept of representative estimation instances. The representative estimation instance $\widehat{x}$ refers to the specific data points on which the model most relies to output a certain prediction or estimation $\widehat{y}$. The local feature importance can be derived from the representative estimation instances corresponding to the original data instance of interest, along with the predicted and estimated values.

Furthermore, based on the original data instances, we generated a representative estimation instance similarity distribution plot that provides a visualization of the distribution of similarities between the data instances and representative estimation instances. The plot shows the reliance of the model on different data portions to make predictions, as well as how each data instance affects the estimate, which helps to further understand the impact of a particular data instance on the model's output.

AIME visualizes the data distribution across predictions and estimations via representative instance similarity distribution plots, indicating the complexity of deriving such estimates. Thus, the outcomes of the AIME allow for a clearer understanding of the model's results and give insights into the complexity of deriving these predictions and estimates. This is particularly important when dealing with complex models, as it highlights the specific areas where the model may struggle to make accurate predictions.

B. CREATION OF THE APPROXIMATE INVERSE OPERATOR FOR THE ORIGINAL MODEL

This section describes the method for generating the approximate inverse operator $\widehat{f^{-1}}$ of the original model f. AIME can be used to determine whether the original model is a regression or a classification model. In this study, a classification model was used. In such type of model, the objective variable $\widehat{y}$ is assumed to be a vector, wherein the probabilities in each possible classification are stored.

The original model f was trained using datasets X and $\widehat{Y}$. Here, $\widehat{Y}$ represents the results predicted by applying the original model f to dataset X. To determine the approximate inverse operator $\widehat{f^{-1}}$ of the original model f, we used normalized datasets X and $\widehat{Y}$. When dealing with large datasets, the use of appropriately resampled subsets of X and $\widehat{Y}$ is acceptable.

The approximate inverse operator $\widehat{f^{-1}}$ is expressed as follows:

$$\widehat{x} = \widehat{f^{-1}}(\widehat{y}) = A^\dagger \widehat{y},$$

where $\widehat{x}$ is the estimated data instance derived by applying the approximate inverse operator $\widehat{f^{-1}}$ to a particular prediction and estimate $\widehat{y}$. In this study, $\widehat{x}$ is referred to as the representative estimation instance, which is explained in detail in Section III-D. $A^\dagger$ denotes the Moore–Penrose generalized inverse matrix [42], [43] of matrix A, constituting a linear approximation of the original model f. Based on these assumptions, deriving the approximate inverse operator $\widehat{f^{-1}}$ equates to deriving the generalized inverse matrix $A^\dagger$.

Now, we consider deriving the generalized inverse matrix $A^\dagger$ from the above equation using datasets X and $\widehat{Y}$. Substituting X, $\widehat{Y}$ into the previous equation and expanding, yields:

$$\begin{aligned} X &= A^\dagger \widehat{Y}, \\ X\widehat{Y}^T &= A^\dagger \widehat{Y}\widehat{Y}^T, \\ X\widehat{Y}^T \left(\widehat{Y}\widehat{Y}^T\right)^{-1} &= A^\dagger \left(\widehat{Y}\widehat{Y}^T\right)\left(\widehat{Y}\widehat{Y}^T\right)^{-1}, \\ A^\dagger &= X\widehat{Y}^T \left(\widehat{Y}\widehat{Y}^T\right)^{-1} = X\widehat{Y}^\dagger, \end{aligned}$$

where $\widehat{Y}^\dagger$ denotes the Moore–Penrose generalized inverse matrix [42], [43] of matrix $\widehat{Y}$. It should be noted that this formula does not compute the approximate forward operator A, but directly computes the approximate inverse operator $A^\dagger$ from datasets X, $\widehat{Y}$.

An important assumption made when applying AIME is that matrix $\widehat{Y}\widehat{Y}^T$ is nonsingular and invertible because the inverse matrix $(\widehat{Y}\widehat{Y}^T)^{-1}$ must be computed. Thus, if matrix $\widehat{Y}\widehat{Y}^T$ is singular and does not have an inverse, the AIME method is not applicable. Regarding the assumption that matrix $\widehat{Y}\widehat{Y}^T$ is nonsingular, if this condition is not satisfied (i.e., the matrix is singular, this might suggest high correlations among the predictors. In such cases, the model may struggle to discern the individual impact of each predictor, potentially resulting in unstable and, hence, unreliable estimations of the original black box model. Nonetheless, it is uncommon in practice for $\widehat{Y}\widehat{Y}^T$ to be singular, such as when one class is perfectly linearly correlated with other classes. Therefore, when applying AIME to real-world datasets, it is of paramount importance to verify that matrix $\widehat{Y}\widehat{Y}^T$ is nonsingular and invertible. If this is not satisfied, there might be underlying issues with the model or dataset, and thoroughly investigating these concerns before employing AIME would be advisable. In such cases, it can be suspected that the black-box model may be inappropriate for using AIME or other interpretive machine learning and XAI approaches.

The generalized inverse matrix $A^\dagger$ is a matrix wherein the row and column dimensions are the number of

TABLE 1. Representative estimation instances of "survived" and "not survived" in the titanic dataset.

	Age	Sibsp	Parch	Fare	Pclass_1	Pclass_2	Pclass_3	Sex_female	Sex_male	Embark_C	Embark_Q	Embark_S
Survived	28.56	0.4945	0.5473	55.00	0.4335	0.2465	0.3200	0.9615	0.039	0.3076	0.1210	0.5714
Not survived	30.05	0.5022	0.2851	19.87	0.1316	0.1903	0.6781	-0.018	1.018	0.1439	0.0774	0.7788

features in dataset X and the number of classification classes, respectively. These steps contribute to the development of the approximate inverse operator $\widehat{f^{-1}}$ of the original model f.

C. GLOBAL FEATURE IMPORTANCE IN AIME

The generalized inverse matrix $A^{\dagger}$, which constitutes the approximate inverse operator $\widehat{f^{-1}}$ derived in Section III-B, is a matrix in which the number of rows and columns are equal to the number of features in dataset X and the number of considered classes, respectively. This generalized inverse matrix $A^{\dagger}$ indicates global feature importance, as its component reflect the impact of each particular feature on a particular class. Specifically, the generalized inverse matrix $A^{\dagger}$ indicates the contribution of every feature to each class. A particular component of $A^{\dagger}$ (e.g., the component in column j of row i) represents the contribution of the i-th feature to the output of the j-th class. This provides a global understanding of the extent to which each feature affects the overall behavior of the model.

The effect of each feature on the model's predicted results is explained in detail in Section IV. This provides a clear understanding of how the components of the generalized inverse matrix $A^{\dagger}$ contribute to specific prediction results.

D. REPRESENTATIVE ESTIMATION INSTANCE

A representative estimation instance is a theoretical construct that encapsulates the essential characteristics of an instance that can be classified into a particular class according to the model's understanding.

In more formal terms, a representative estimation instance $\boldsymbol{x}^*$ is a solution obtained by applying an approximate inverse operator $A^{\dagger}$ to a vector $\boldsymbol{y}^*$ that represents the perfect prediction of a particular class. The representative estimation instance $\boldsymbol{x}^*$ is derived using the following equation:

$$\boldsymbol{x}^* = A^{\dagger}\boldsymbol{y}^*,$$

where $\boldsymbol{y}^*$ is a vector of predictions that exactly corresponds to a particular class, where only the component corresponding to that class is one and the other components are zero, and $\boldsymbol{x}^*$ is obtained by applying the approximate inverse operator $A^{\dagger}$ to such a vector. Here, $\boldsymbol{y}^*$ is the representative estimated instance of that class and an m-dimensional vector, which is the number of classes. If we assume a particular class k, each element y_i^* of vector $\boldsymbol{y}^*$ for that class can be evaluated using the following equation:

$$y_i^* = \begin{cases} 1 & (if\ i = k) \\ 0 & (otherwise) \end{cases}$$

Representative estimation instances represent "ideal" or "typical" instances that a model expects to belong to a particular class. A closer look at these instances provides a better understanding of the overall significance of features in a particular class.

By examining representative estimation instances, we can gain unique insights into the characteristics that most closely resemble the class membership. Therefore, these instances are of vital importance when interpreting and validating the model predictions.

Table 1 presents the output result examples of the representative estimation instances of "survived" and "not survived" for the Titanic data [44] in AIME. The Titanic dataset [44], which was used as an example in this study, is popular for data analysis and machine learning tasks. The data included various features of passengers, such as age, gender, ticket class, fare, embarkation port, and whether they were traveling alone or with family. Predictive tasks using this dataset typically aim to determine whether a passenger would have survived based on these features. Representative estimation instances listed in Table 1 are estimated with continuous values, including those represented by the dummy variables {0,1}. The representative estimated instance of "survived" has an age of approximately 28 years, a fee of approximately 55, a Pclass of 1, and is female. The representative instance of "not survived" has an age of approximately 30 years, a fee of approximately 20, a Pclass of 3, and is male. Thus, representative instances that accurately capture the characteristics of people who "survived" and "not survived" are extracted.

E. LOCAL FEATURE IMPORTANCE IN AIME

The local feature importance in AIME measures the extent to which each feature contributes to the classification of a particular data instance. This helps to understand why a data instance is classified into a particular class by the model and can reveal the features that directly affect the classification of a particular instance. These local feature importance values provide a clear understanding of how a feature's impact varies in the prediction for a particular data instance and can indicate the extent to which the prediction depends on a particular feature. Therefore, the local importance of a feature is a crucial factor in identifying the features that the model considers significant for a particular instance and integrating that understanding into the interpretation of the model.

For the data instance $\boldsymbol{x}$ of interest, we consider the local feature importance for a particular class k. By incorporating the data instance $\boldsymbol{x}$, the vector $\boldsymbol{y}^*$, which represents the full

prediction for a particular class k described in Section III-D, and the generalized inverse matrix $A^{\dagger}$, the local feature importance vector $\boldsymbol{l}$ can be derived using the following equation:

$$\boldsymbol{l} = A^{\dagger}\boldsymbol{y}^{*} \circ \boldsymbol{x},$$

where the local feature importance vector $\boldsymbol{l}$ consists of the number of dimensions of the total features, where each value represents the significance of that feature, and $\circ$ represents the Hadamard product.

The vector $\boldsymbol{y}^{*}$ represents a ''perfect'' prediction for a particular class, and $A^{\dagger}$ is an approximate inverse operator that maps this prediction to the feature space. This provided an overall representation of the importance of a feature in a particular class. The value of each feature in data instances $\boldsymbol{x}$ and $A^{\dagger}\boldsymbol{y}^{*}$ is multiplied in an element-wise manner to reflect the extent to which each feature's value affects its significance. This is realized by the Hadamard product $\circ$. This operation emphasizes the significance of a particular feature of $\boldsymbol{x}$ if it has a large value. However, if a particular feature of $\boldsymbol{x}$ has a small value or is zero, its importance is reduced. This allows for a specific expression of the contribution of each feature to the classification into a particular class for a particular data instance, $\boldsymbol{x}$. This is the importance of the local features in AIME.

Here, the data instance $\boldsymbol{x}$ should be normalized with the same parameters as dataset X, which is used to obtain the generalized inverse matrix $A^{\dagger}$.

When the local feature importance of AIME is derived using the above formula, the importance of a feature may be high, even if the value of the feature in the data instance $\boldsymbol{x}$ before normalization is zero. This implies that the value of zero for a feature is important. However, for one-hot encoding features, a feature value of zero for a data instance $\boldsymbol{x}$ before normalization indicates that it does not possess that feature. Therefore, in this study, once the value of the feature of the data instance $\boldsymbol{x}$ before normalization is zero, the local feature importance determined by AIME should be zero. However, this may vary, depending on the dataset used.

F. REPRESENTATIVE INSTANCE SIMILARITY DISTRIBUTION PLOT

The AIME visualization method provides a representative instance similarity distribution plot. It visualizes the similarity distribution between a representative estimation instance and each instance in a target dataset. The representative instance similarity distribution plot provides a visual understanding of the behavior of the model when it outputs a particular estimate and the region of the training data to which it corresponds. Specifically, the plot shows the similarity distribution between the representative estimation instance corresponding to the output estimated by the model and each instance in the entire dataset. This helps us understand from which part of the dataset a particular estimate is most likely to be drawn. In addition, the plot provides specific visual information about the model's behavior when outputting a particular estimate and the complexity associated with that estimate. This provides a better understanding of the behavior of the model and estimates its reliability.

First, we derive a representative estimation instance corresponding to each class, as described in Section III-D. Let $\boldsymbol{x}_k^{*}$ be the representative estimation instance of class k. Next, for each class, we compute the similarity between each data instance $\boldsymbol{x}$ in dataset X and the representative estimated instance $\boldsymbol{x}_k^{*}$. In this study, we introduced an RBF kernel K to calculate similarity, as shown in the following equation:

$$K(\boldsymbol{x}_k^{*}, \boldsymbol{x}) = exp(-\gamma \left\| \boldsymbol{x}_k^{*} - \boldsymbol{x} \right\|^2),$$

where γ is a parameter that controls RBF kernel spread. Although we set γ to 0.1 in this study, the optimal value may vary according to the dataset characteristics. Therefore, γ should be chosen carefully, by examining the resulting visualizations or employing other model selection criteria.

The kernel density estimation in the representative instance similarity distribution plot was used to visualize the similarity distribution between a representative instance for a particular estimate and other instances in the entire dataset. A smooth distribution can be obtained by applying kernel density estimation to the similarity scores, which further provides a more intuitive understanding of the parts of the overall dataset to which a particular estimate is most closely related.

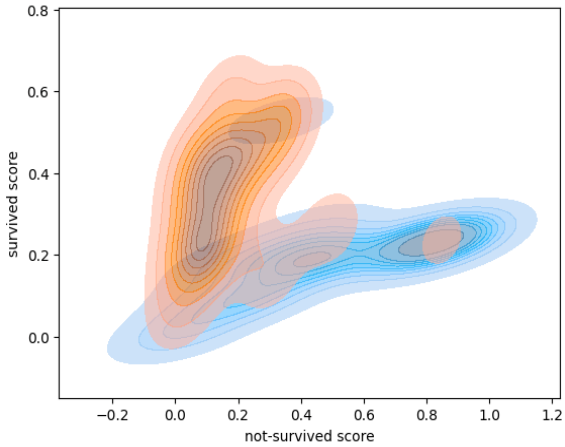

FIGURE 3. Representative instance similarity distribution plot for the case of the Titanic dataset [44]. The horizontal axis, colored blue, represents the "not survived" class; the vertical axis, colored red, represents the "survived" class. A larger value in the horizontal axis indicates a higher similarity to the representative estimated instance of "not survived," while a larger value in the vertical axis indicates a higher similarity to the representative estimation instance of "survived." This distribution suggests that many instances are well represented by their corresponding representative instances in the Titanic dataset. However, there are overlapping red plots on the horizontal axis and overlapping blue plots on the vertical axis. These "overlapping" instances exhibit high similarity with representative instances of both classes, indicating more complex decision boundaries and potential difficulties in prediction.

A representative instance similarity distribution plot for a simple Titanic dataset [44] is shown in Fig. 3. The Titanic dataset, which was used as an example in this study, is popular for data analysis and machine learning tasks. The data included various features of passengers, such as age, gender, ticket class, fare, embarkation port, and whether they were

TABLE 2. Rose and jack data instances in the titanic.

	Age	Sibsp	Parch	Fare	Pclass_1	Pclass_2	Pclass_3	Sex_female	Sex_male	Embark_C	Embark_Q	Embark_S
Rose	17.0	0	1	33.3	1	0	0	1	0	0	0	0
Jack	19.0	0	0	0.0	0	0	1	0	1	0	0	0

traveling alone or with family. A predictive task using this dataset is typically aimed at determining whether a passenger would have survived based on these features. In our study, the labels ''not survive'' and ''survive'' are used to generate each representative estimation instance. Fig. 3 shows the RBF kernel for each instance of the entire dataset and the representative estimation instances corresponding to the two ''not survive'' and ''survive'' cases, and the distribution is visualized using kernel density estimation.

The horizontal axis, colored blue, represents the ''not survived'' class; the vertical axis, colored red, represents the ''survived'' class. A larger horizontal axis indicates a higher similarity to the representative estimated instance of ''not survived,'' while a larger vertical axis indicates a higher similarity to the representative estimation instance of ''survived.'' Because this shows the distribution of each class in dataset X, the data instances with a higher similarity to the ''not survived'' representative instance are spread out along the horizontal axis, while those with a higher similarity to the ''survived'' representative instance are spread out along the vertical axis. This distribution suggests that many instances are well represented by their corresponding representative instances in the Titanic dataset. However, there were overlapping red plots on the horizontal and vertical axes. These ''overlapping'' instances exhibit high similarity with the representative instances of both classes, indicating a more complex decision boundary and potential difficulties in prediction. These complexities and difficulties refer to instances in which a model may be less certain or less accurate in its predictions. Intuitively, in the representative instance similarity distribution plot, if the distribution is dispersed along the horizontal and vertical axes, the data instances are well represented by their respective representative instances, implying that the decisions of the model are relatively straightforward. Conversely, if the distribution is concentrated between the horizontal and vertical axes, the data instances may have overlapping similarities with representative instances from multiple classes. This indicates a more complex model in which decisions are potentially more difficult because of the overlapping feature spaces of the classes.

The representative instance similarity distribution plots intuitively visualize the complexity of the prediction and estimation across the entire dataset.

IV. EXPERIMENTS

We evaluated the system that constitutes the AIME. We focused on the problem of the classification of tabular, image, and text data and studied how AIME can generate explanations.

In Section IV-A, the experimental system environment is presented. In Section IV-B, we present the results of Experiment 1, using tabular data. In Section IV-C, we present the results of Experiment 2 using image data. In Section IV-D, we present the results of Experiment 3, which used text data. Section IV-E discusses the results of these experiments.

A. EXPERIMENT ENVIRONMENT

In this experiment, we used the Titanic data [44] as tabular data, MNIST [45] data as image data, and 20 newsgroups [46] as text data.

For the Titanic data, we used ''Pclass,'' ''Sex,'' ''Age,'' ''SibSp,'' ''Parch,'' ''Fare,'' and ''Embarked'' as explanatory variables, and performed one-hot encoding for ''Pclass,'' ''Sex,'' and ''Embarked.'' For missing values, the average value of ''Age'' is substituted for ''Age,'' ''Embarked'' for ''S,'' and data with missing values are deleted for ''Fare.'' For the 20 newsgroups, we used data from the ''rec.sport.baseball,'' ''rec.sport.hockey,'' and ''sci.electronics'' categories, resulting in a training dataset of 1079 cases. Here, ''rec.sport.baseball'' and ''rec.sport.hockey'' are similar in terms of the sport and their difficulty in classifying categories, while ''sci.electronics'' is a completely different field and is relatively easy to classify.

The experimental system was implemented in Python 3.10.12 and run on a Google Colab Pro. The black-box model chosen for this experiment was LightGBM, based on the lightgbm 4.0.0 package. Although there is no specific reason for choosing LightGBM, it should be noted that theoretically, any model can be used as the original model in LIME, SHAP, and the proposed AIME method because all three are model-agnostic methods. LIME uses the package lime 0.2.0.1, and SHAP uses the package shap 0.42.0. Although visualization methods are implemented in each of them, we used the visualization we implemented using matplotlib 3.7.1 and seaborn 0.12.2 for comparison.

B. EXPERIMENT 1: APPLICATION OF AIME TO TABULAR DATA

In Experiment 1, the Titanic data [44] was used as tabular data to examine the generation of AIME explanations. To facilitate the interpretation of the experimental results, Rose, and Jack data instances are created by us regarding the movie ''Titanic'' as presented in Table 2. According to the movie's storyline, Rose ''survives,'' and Jack does ''not survive.''

First, the results for global feature importance, which describe the overall model with AIME, are presented, and the results for local feature importance are compared. This comparison describes the prediction and estimation results

for the respective data instances of Rose and Jack for LIME, SHAP (Shapley value), and AIME. A representative instance similarity distribution plot obtained using AIME is shown in Fig. 3.

1) GLOBAL FEATURE IMPORTANCE

AIME outputs the global feature importance of the Titanic data, as shown in Fig. 4. For each feature in the Titanic data, ''not survived'' and ''survived'' indicate the degree of contribution of the model to the prediction and estimation, respectively. According to Fig. 4, Sex_female contributes as ''survived'' and Sex_male contributes as ''not survived.'' Pclass_1, or the first class, contributes positively to ''survived.'' ''Fare'' contributes positively to ''survived,'' and Class_3 contributes positively to ''not survived.'' These results show that the global feature importance of AIME provides an intuitive explanation for the trends in the Titanic data and the behavior of the model.

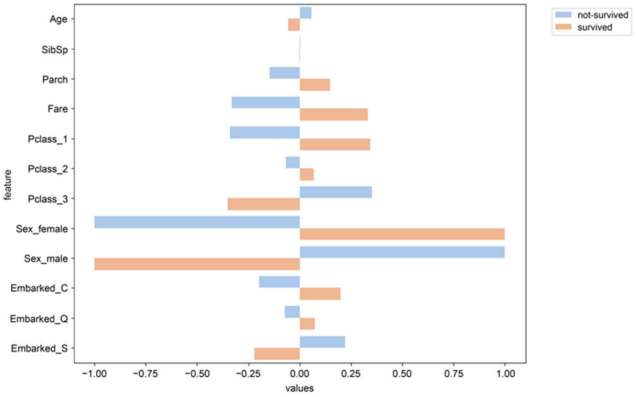

FIGURE 4. Output results for the global feature importance of AIME in the case of the Titanic data. Sex_female contributes as "survived" and Sex_male contributes as "not survived." Pclass_1, or first class, contributes positively to "survived." Fare contributes positively to "survived," and Class_3 contributes positively to "not survived." These results show that the global feature importance of AIME provides an intuitive explanation for trends in the Titanic data and the behavior of the model.

2) LOCAL FEATURE IMPORTANCE

The results for the local feature importance of LIME, SHAP (Shapley value), and AIME are shown in Figs. 5, 6, and 7 as a ''survived'' explanation of the Rose instance presented in Table 2. Rose ''survived'' according to the movie ''Titanic,'' and the LightGBM prediction and estimation results are also judged to be ''survived'' with a probability of 99.7%.

In the case of LIME, as shown in Fig. 5, ''Sex_female'' and ''Embarked_C'' are identified as contributing positively, whereas ''Pclass_3'' and ''Age,'' contribute negatively. The Rose data are listed in Table 2, and the ''Pclass_3'' for Rose is zero. However, the LIME results indicate that ''Pclass_3'' contributes negatively, suggesting that LIME finds ''Pclass_3'' to be of importance. As described in Section III-E, LIME must also include setting the importance of zero features to zero for some datasets.

In the case of SHAP (Shapley values) shown in Fig. 6, ''Sex_female,'' Fare, ''Pclass_3,'' and ''Pclass_1'' contribute

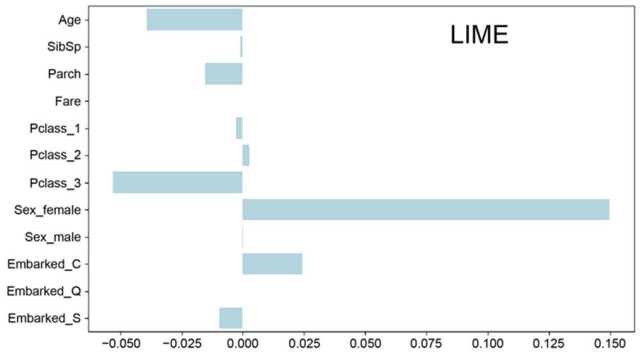

FIGURE 5. Output results of LIME in the case of the Rose data in Titanic. Sex_female and Embarked_C are derived as contributing positively and Pclass_3 and Age as contributing negatively.

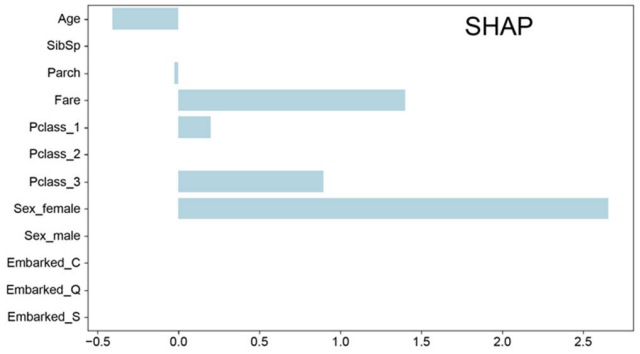

FIGURE 6. Output results for the Shapley values of SHAP in the case of the Rose data in Titanic. Sex_female, Fare, Pclass_3, and Pclass_1 contribute positively, while Age is output as contributing negatively.

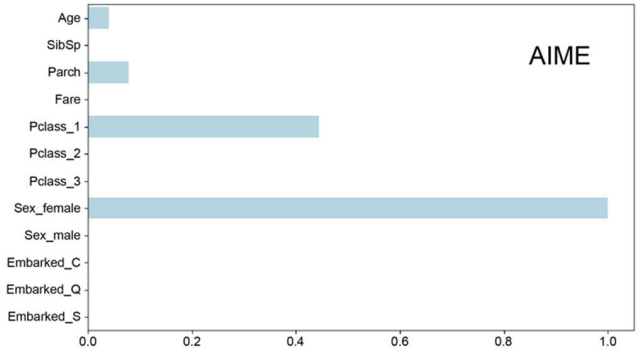

FIGURE 7. Output results of AIME in the case of the Rose data in Titanic. Sex_female, Pclass_1, Parch, and Age contribute positively in that order.

positively, while Age is output as contributing negatively. It is thought that the Shapley value was high because it is important that ''Pclass_3'' is zero, as shown in LIME. However, ''Pclass_3'' should be considered a negative contributor to ''survived,'' but SHAP outputs it as a positive contributor. However, this interpretation is difficult.

In the case of AIME shown in Fig. 7, the output shows that ''Sex_female,'' ''Pclass_1,'' ''Parch,'' and ''Age'' contribute positively in that order. The operation mentioned in Section III-E, which sets the importance of a feature with a value of zero to zero, is in effect, and the output is such that

the importance of the more contributing features among the features with nonzero feature values is higher. This result was simple and easy to interpret.

The results for the local feature importance of LIME, SHAP (Shapley value), and AIME are shown in Figs. 8, 9, and 10 as a ''survived'' explanation of the Jack instance listed in Table 2. Jack is ''not survived'' according to the movie ''Titanic,'' and the LightGBM prediction and estimation results are also judged to be ''not survived'' with a probability of 97.2%.

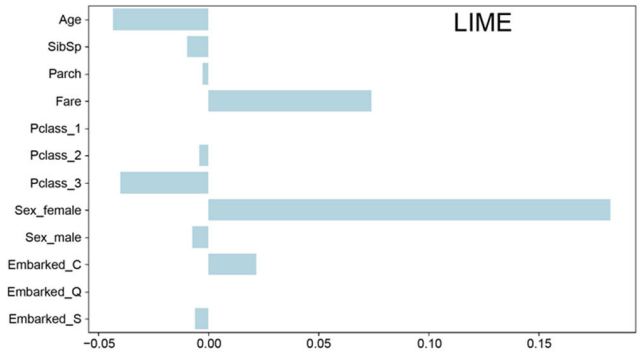

FIGURE 8. Output results of LIME in the case of the Jack data in Titanic. Sex_femle, Fare, and Embarked_C are derived as contributing positively, and Pclass_3 and Age are derived as contributing negatively.

In the case of LIME shown in Fig. 8, ''Sex_female,'' ''Fare,'' and ''Embarked_C'' are derived as contributing positively, while ''Pclass_3'' and ''Age'' contribute negatively. In the Jack data listed in Table 2, the values of ''Sex_female'' and ''Fare'' were zero. However, the LIME results show that Sex_female and Fare contribute positively, which suggests that LIME finds ''Sex_female'' and ''Fare'' of zero importance. However, interpreting the positive and negative values of importance may seem esoteric for users.

In the case of SHAP (Shapley values) shown in Fig. 9, ''Age'' contributes positively, and ''Sex_female,'' ''Fare,'' ''Pclass_3,'' and ''Pclass_1'' contribute negatively in that order. The result that ''Fare'' and ''Pclass_3'' features contribute negatively to ''survive'' is intuitive. However, in the Jack data listed in Table 2, ''Sex_female'' and ''Pclass_1'' are zero. Thus, the fact that the feature is zero is positively significant. The value of ''Pclass_1'' is assumed to contribute positively to ''survived.'' However, in the case of this result, the local negative contribution means that the ''Pclass_1'' value of zero can be considered to contribute negatively to ''survived.''

In the case of AIME shown in Fig. 10, ''Sex_male'' and ''Pclass_3'' contribute negatively in that order, and ''Age'' contributes positively. This result is very simple compared with those of LIME and SHAP. However, it is intuitive how the values of the Jack data listed in Table 2 contribute to ''survived.'' Furthermore, in the case of AIME, it is possible to compare the global feature importance shown in Section IV-B-I and its importance as a model with the importance specific to a particular data instance.

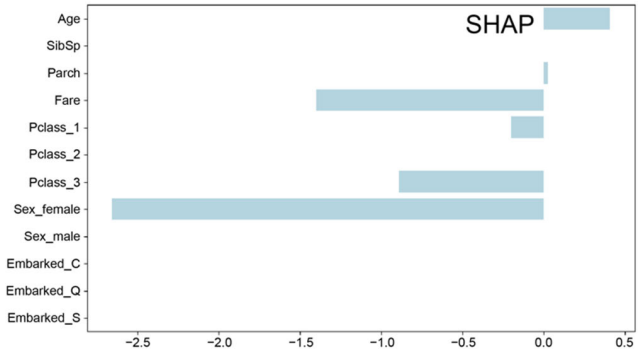

FIGURE 9. Output results for the Shapley values of SHAP in the case of the Jack data in Titanic. Age contributes positively, while Sex_female, Fare, Pclass_3, and Pclass_1 contribute negatively in that order.

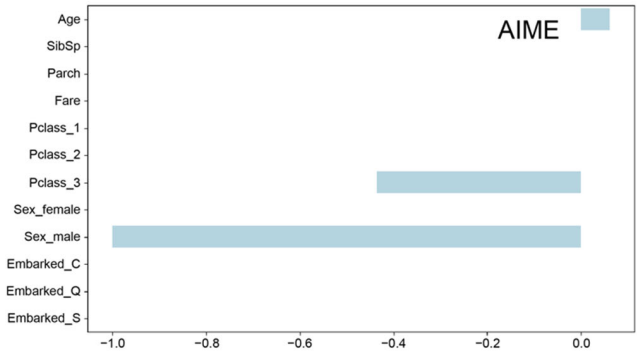

FIGURE 10. Output results of AIME in the case of the Jack data in Titanic. Sex_male and Pclass_3 contribute negatively in that order, while Age contributes positively.

This confirms that different ways of deriving explanations lead to different results for each case. For LIME and SHAP, it is difficult to partially interpret the positive and negative meanings. In this respect, AIME can provide simple local importance results for the Titanic data. However, LIME and SHAP can also be improved by setting the importance of features with a value of zero to zero, as indicated in Section III-E. In addition, depending on the complexity of the dataset, it may be useful to derive the relationships between features, particularly in SHAP. By contrast, AIME can be considered by comparing the global and local feature importance in Section IV-B-I.

3) REPRESENTATIVE INSTANCE SIMILARITY DISTRIBUTION PLOT

A representative instance similarity distribution plot obtained using AIME for the LightGBM for the Titanic data is shown in Fig. 3. This shows the distribution of each class in dataset X, and data instances with higher similarity to the ''not survived'' representative instance are spread out along the horizontal axis, while those with higher similarity to the ''survived'' representative instance are spread out along the vertical axis. This distribution suggests that many instances are well represented by their corresponding representative

instances in the Titanic dataset. In other words, the Titanic data can be easily classified into ''not survived'' and ''survived'' (Fig. 3).

Overlaying the Rose and Jack data on the representative instance similarity distribution plot in Fig. 3 helps observe the complexity of each classification. The results are shown in Figs. 11 and 12.

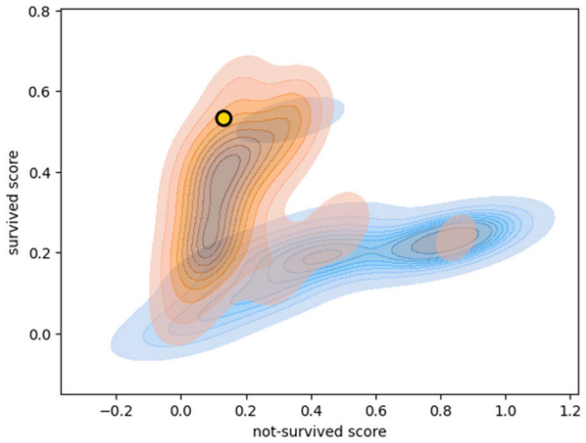

FIGURE 11. Representative instance similarity distribution plot of the Titanic data with the Rose data instances plotted. It is easy to classify the Rose data instance as "survived" because it is close to a representative estimated instance of "survived" and far from a representative estimated instance of "not survived."

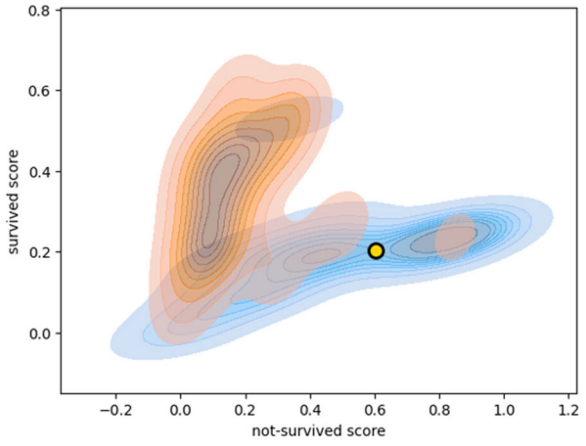

FIGURE 12. Representative instance similarity distribution plot of the Titanic data with the Jack data instances plotted. It is easy to classify the Jack data instance as "not survived" because it is close to the representative estimated instance of "not survived" and far from the representative estimated instance of "survived."

Fig. 11 plots the RBF kernel values for the Rose data instances and the ''not survived'' and ''survived'' representative estimation instances in Fig. 3. and the representative estimation instance of ''survived,'' i.e., the closer it is to the representative data of ''survived.'' Because the plot is located at a smaller position on the horizontal axis and a larger position on the vertical axis, it is closer to the representative estimated instance of ''survived'' and can be objectively estimated as ''survived'' easily.

Fig. 12 shows a plot of the RBF kernel values for the Jack data instances and the ''not survived'' and ''survived'' representative estimation instances in Fig. 3. In contrast to Fig. 10, because the plot is located at a larger position on the horizontal axis and a smaller position on the vertical axis, it is closer to the representative estimated instance of ''not survived'' and can be objectively estimated as ''not survived'' easily.

These representative instance similarity distribution plots allow us to check the complexity of the instance of interest while checking the complexity of the classification of the entire dataset.

C. EXPERIMENT 2: APPLICATION OF AIME TO IMAGE DATA

Experiment 2 used MNIST [45] image data to examine the generation of AIME. The MNIST data [45] consists of grayscale handwritten numeric image data with a size of 28×28 pixels and labels. We converted the 28×28 grayscale handwritten digit image data into a 784-dimensional vector as input data. LightGBM was used to build a model that predicts which numbers from 0 to 9 are written from a 784-dimensional vector converted from image data of handwritten numbers.

First, as in Experiment 1, we show the results for the global feature importance representing the entire model using AIME by reconstructing 28×28 images. Then, for the local feature importance representing the prediction and estimation results for several data instances, the results of LIME, SHAP (Shapley value), and AIME are compared. In addition, we describe the similarity distribution plots for representative instances using AIME.

1) GLOBAL FEATURE IMPORTANCE

The output results for the global feature importance of the MNIST data obtained using the AIME are shown in Fig. 13. Each result shows a 784-dimensional global feature importance reconstructed into a 28×28 image.

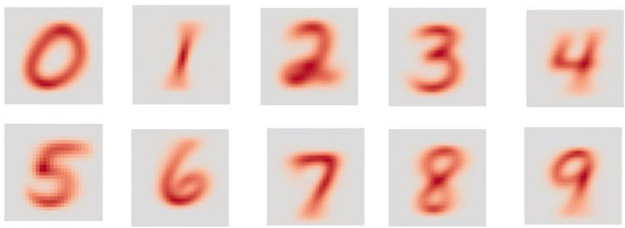

FIGURE 13. Output results for the global feature importance of the AIME in the case of the MNIST data. The redder the pixel color, the higher the positive global feature importance. Gray has global feature importance of zero, and the bluer the pixel (which does not appear in this figure), the higher the negative global feature importance. We can see characteristics of each class of handwriting are captured in all classes.

The generalized inverse matrix $A^{\dagger}$ shows the contribution of each feature to each class. A particular component of $A^{\dagger}$ (e.g., the component in column j of row i) represents the contribution of the i-th feature to the output of the j-th class.

This provides a global understanding of the extent to which each feature affects the overall behavior of the model. In the MNIST data, $A^{\dagger}$ is a 10×784 matrix, meaning 28×28 $(=784)$ global feature importance for 10 classes from 0 to 9. In other words, by taking each row of $A^{\dagger}$ and making it 28×28, the global importance of 0–9 can be visualized as an image.

The redder the pixel color, the higher the positive global feature importance. Gray has global feature importance of zero, and the bluer the pixel (which does not appear in this figure), the higher the negative global feature importance. The characteristics of each handwriting class were captured in all classes.

From this result, we can extract the global key features of the original MNIST model, thereby capturing the key features of the entire model and providing its behavior.

2) LOCAL FEATURE IMPORTANCE

The results for the local feature importance of LIME, SHAP (Shapley value), and AIME are shown in Figs. 14 and 15. Each result shows a 784-dimensional local feature importance reconstructed into a 28×28 image.

Fig. 14 shows the LIME, SHAP (Shapley value), and AIME output for the ''3'' handwritten character. Incidentally, the handwriting character of three in LightGBM is recognized with 99.6% certainty. The redder the pixel color, the higher the positive global feature importance; gray has a global feature importance of zero. Further, the bluer the pixel, the higher the negative global feature importance. For LIME and SHAP, the light red area represents the feature importance of the three, but it is difficult for users to determine how to capture the negative feature importance in blue. In contrast, AIME clearly indicated a feature importance of three.

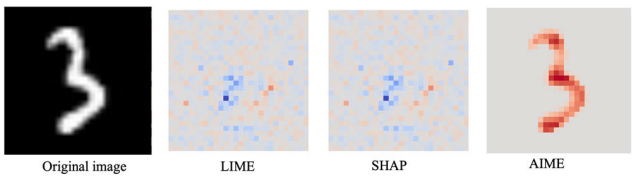

FIGURE 14. LIME, SHAP (Shapley value), and AIME output for "3" handwritten character. The redder the pixel color, the higher the positive local feature importance. Gray has a local feature importance of zero, and the bluer the pixel, the higher the negative local feature importance. For LIME and SHAP, if we consider the light red area, it represents feature importance of three, but it will be difficult for users to figure out how to capture the negative feature importance in blue. In contrast, AIME clearly indicates the feature importance of the three.

Fig. 15 shows the LIME, SHAP, and AIME outputs for the ''8'' handwritten character. As in the case of Fig. 14, LIME and SHAP are characterized, but not as clearly indicated as AIME.

For the derivation of local feature importance in LIME and SHAP, we converted 28×28 image data into a 784-dimensional vector as input, which can pose challenges for LIME and SHAP in representing local feature importance. However, under similar conditions, AIME clearly indicated the local feature importance of the characters.

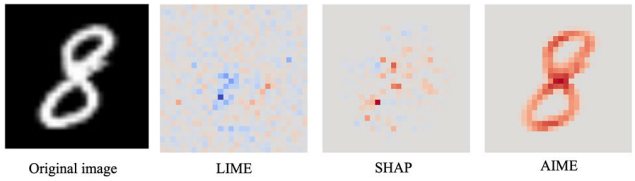

FIGURE 15. LIME, SHAP (Shapley value), and AIME output for "8" handwritten character. The redder the pixel color, the higher the positive local feature importance, gray has a local feature importance of 0, and the bluer the pixel, the higher the negative local feature importance. For LIME and SHAP, if we consider the light red area, it represents feature importance of eight, but it will be difficult for users to figure out how to capture negative feature importance in blue. In contrast, AIME clearly indicates the feature importance of eight.

Further experiments regarding the explanatory derivation of LIME, SHAP, and AIME using a convolutional neural network as a black-box model with 28×28 image data as input are presented in the Appendix.

Fig. 16 shows the local feature importance of the handwritten character ''8'' when it is recognized as ''zero,'' when it is recognized as ''three,'' and when it is recognized as ''eight.'' Incidentally, the handwritten character ''8'' in LightGBM is recognized as eight with 99.3% certainty. In the case of ''zero'' recognition, the rounded areas at the top or bottom are considered to have high local feature importance, and the ''zero'' feature in the letter ''0'' is captured well. In explaining the case of ''three,'' the left side of the letter ''8'' is less important than the right. This is because the left side of the letter ''3'' is cut off.

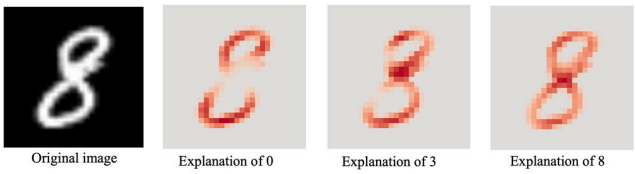

FIGURE 16. Local feature importance of the handwritten character "8" by AIME when it is recognized as "zero," when it is recognized as "three," and when it is recognized as "eight." In the case of "zero" recognition, the rounded areas at the top or bottom are considered to have high local feature importance, and the "zero" feature in the letter "0" is well captured. While explaining the case of "three," the left side of the letter "8" is less important than the right. This is because the left side of the letter "3" is cut off.

Based on the aforementioned factors, LIME, SHAP (Shapley value), and AIME can extract the local feature importance of instances of MNIST data, but AIME can explicitly show it. This indicates that the inverse operator in AIME is well constructed.

3) REPRESENTATIVE INSTANCE SIMILARITY DISTRIBUTION PLOT

A representative instance similarity distribution plot of classes ''three'' and ''eight'' with the ''3'' handwritten character in the MNIST data for AIME is shown in Fig. 17. The representative instance similarity distribution plot in Fig. 17 shows that the distributions of classes ''three'' and ''eight''

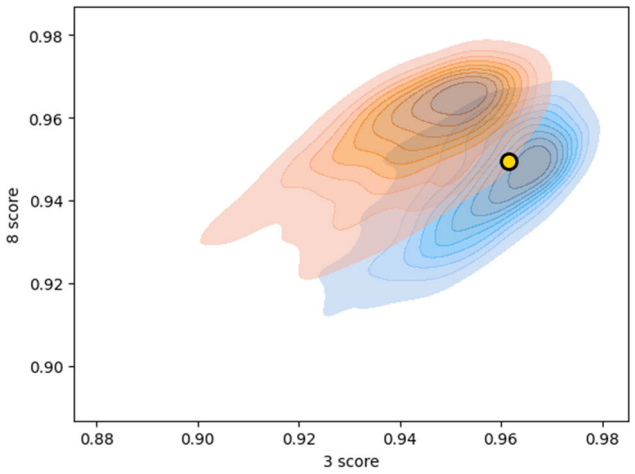

FIGURE 17. Representative instance similarity distribution plot of classes "three" and "eight" with the "3" handwritten character in the MNIST data for AIME. This representative instance similarity distribution plot shows that the class three and eight distributions are located between the horizontal and vertical axes, indicating that the discrimination is relatively complex and well-captured.

are located between the horizontal and vertical axes, indicating that discrimination is relatively complex.

Classes "zero" and "one" in the MNIST data possess relatively simple classifications, as shown in Fig. 18. A distribution with large values on the horizontal axis indicates a class zero dataset, whereas a distribution with large values on the vertical axis indicates a class one dataset. This figure shows that the complexity was relatively low because of the lack of distribution overlap.

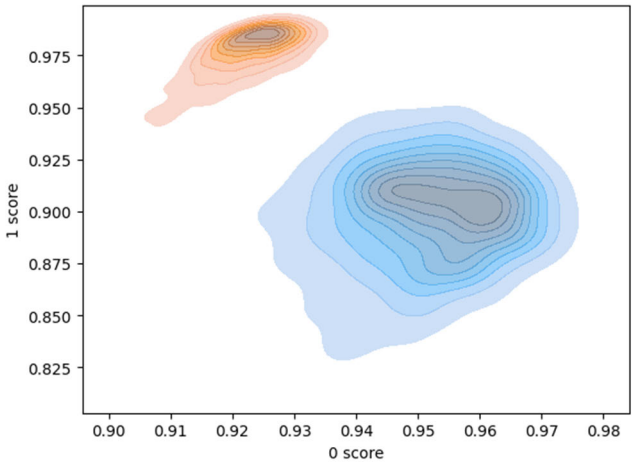

FIGURE 18. Representative instance similarity distribution plot of class "zero" and class "one" in the MNIST data. A distribution with large values on the horizontal axis indicates a class zero data set, while a distribution with large values on the vertical axis indicates a class one data set. This figure shows that complexity is relatively low owing to a lack of overlap in distributions.

Fig. 19 shows the representative instance similarity distribution plots for each class. The diagonal plots of the same class in Fig. 19 are not significant because they are plotted on a $y = x$ straight line. From this figure, it is possible to observe which classes are more complex.

This indicates that the AIME explanation is also valid for the MNIST data.

D. EXPERIMENT 3: APPLICATION OF AIME TO TEXT DATA

Experiment 3 used 20 newsgroup data [46] as an example of text data to study the generation of AIME explanations. For the newsgroup data, we used data from the "rec.sport.baseball," "rec.sport.hockey," and "sci.electronics" categories, resulting in a training dataset of 1079 cases. Here, "rec.sport.baseball" and "rec.sport.hockey" are similar in terms of the sport and their difficulty in classifying categories, whereas "sci.electronics" is a completely different field and is relatively easy to classify.

1) GLOBAL FEATURE IMPORTANCE

The output results for the global feature importance of "rec.sport.baseball" class, "rec.sport.hockey" class, and "sci.electronics" class in the 20-newsgroup data by AIME are shown in Figs. 20, 21, and 22.

Fig. 20 shows the global feature importance of the top 20 in the "rec.sport.baseball" class. The words "year," "wa," "game," "run," "team," "baseball," "hit," "think," "player," "last," "ha," "one," "pitching," "would," "brave," "fan," "good," "win," "pitcher," and "first" appear in the order of their importance.

The following words appear in order of importance: Many of the words that appear here are related to "baseball."

Fig. 21 shows the global feature importance of the top 20 in the "rec.sport.hockey" class. The words "game," "wa," "team," "hockey," "player," "play," "would," "playoff," "ha," "nhl," "year," "go," "season," "goal," "one," "think," "leaf," "win," "time," and "like" appear in this order of importance. The following words appear in order of importance: Many of the words appearing here are related to "hockey."

Fig. 22 shows the global feature importance of the top 20 in the "sci.electronics" class. The words "use," "one," "would," "anyone," "circuit," "know," "doe," "power," "chip," "like," "get," "used," "work," "line," "thanks," "wa," "ground," "current," "output," and "could" appear in this order of importance. The following words appear in order of importance: Many of the words appearing here are related to "electronics."

This indicates that the global feature importance determined by AIME for text data can be weighted to output words that exhibit the model's behavior well.

2) LOCAL FEATURE IMPORTANCE

We used the 0th, 2nd, and 5th data from the test data in the 20-newsgroup data to evaluate the local feature characteristics. These are the instances of "rec.sport.baseball," "rec.sport.hockey," and "sci.electronics" in the test data, respectively. The test data are presented in Table 3. These data were classified into the correct class because of the prediction and inference by LightGBM of the original model (black-box model) that was made beforehand.

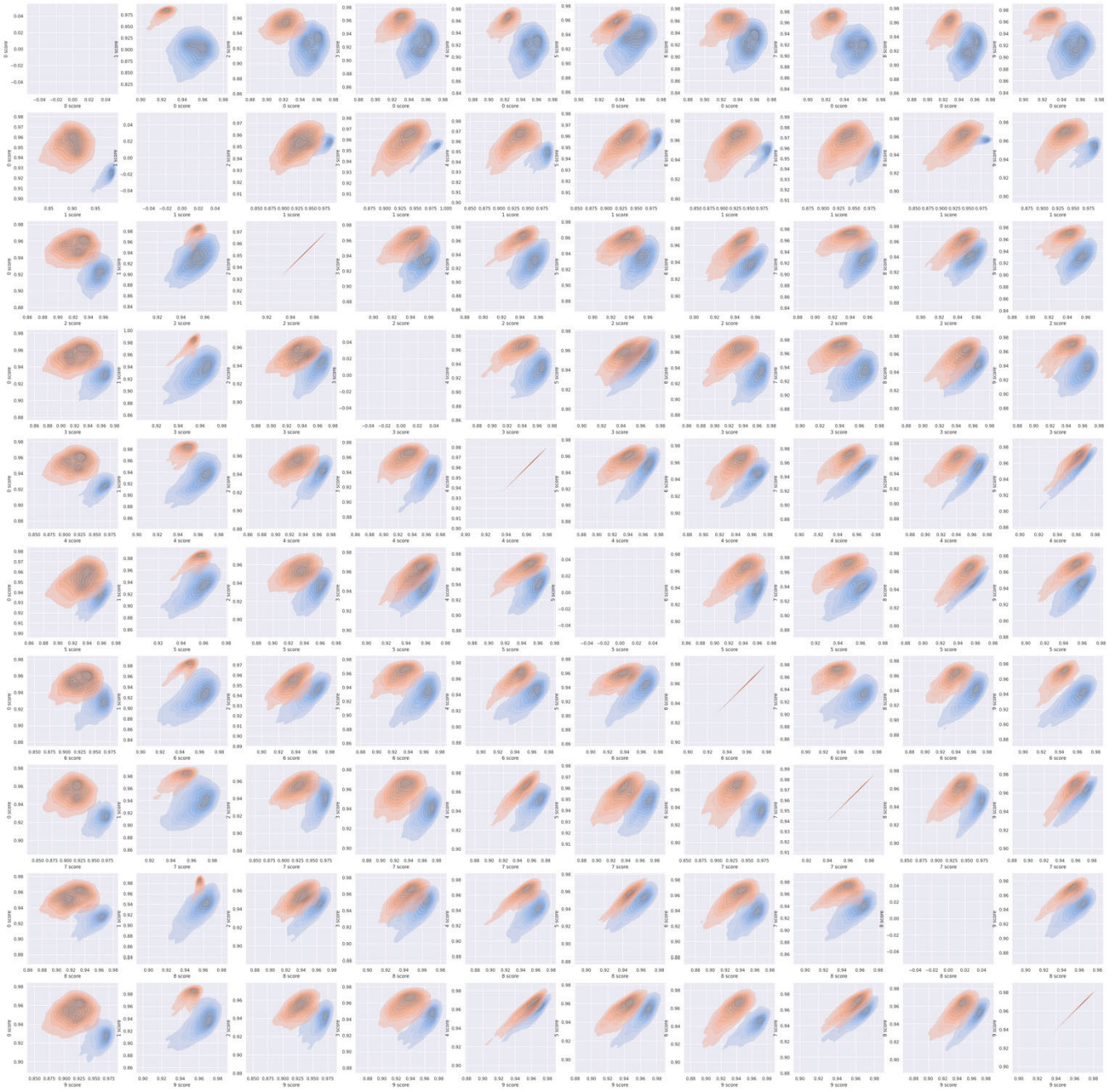

FIGURE 19. Representative instance similarity distribution plots between each class. Diagonal plots of the same class are not significant because they are plotted on a y = x straight line. From this figure, it is possible to observe which classes are more complex to classify.

In Fig. 24, in the case of SHAP (Shapley value), ''baseball,'' ''hockey,'' ''would,'' ''use,'' ''wa,'' ''ha,'' ''hit,'' ''pitcher,'' ''know,'' ''one,'' ''playoff,'' ''day,'' ''need,'' ''year,'' ''power,'' ''much,'' ''good,'' ''make,'' and ''pitching'' are derived as contributing as the 0th data instance. In this result, words that do not appear in the 0th instance, such as ''baseball'' and ''hockey,'' are derived as having local feature importance. This implies that the absence of these words is an important feature. This result may be difficult for users to interpret when used as an explanation. When using this result as an explanation, it is easier to find AIME explanations to understand compared to using LIME or SHAP.

In Fig. 25, in the case of the AIME, ''april,'' ''jose,'' ''sunday,'' ''davis,'' ''oakland,'' ''pm,'' ''05,'' ''california,'' ''storm,'' ''pdt,'' and ''wared'' are derived as contributing as the 0th data instance. These are the words that appear in the 0th data instance. In contrast to the LIME case, ''davis'' and ''jose'' were extracted and derived as contributing positively. In contrast to LIME and SHAP, users may easily interpret this result when using it as an explanation.

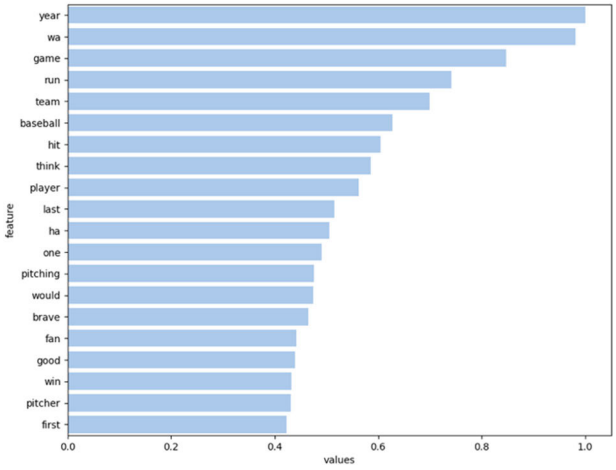

FIGURE 20. Global feature importance of the top 20 in the "rec.sport.baseball" class. The following words appear in this order of importance. Many words appearing here are related to "baseball."

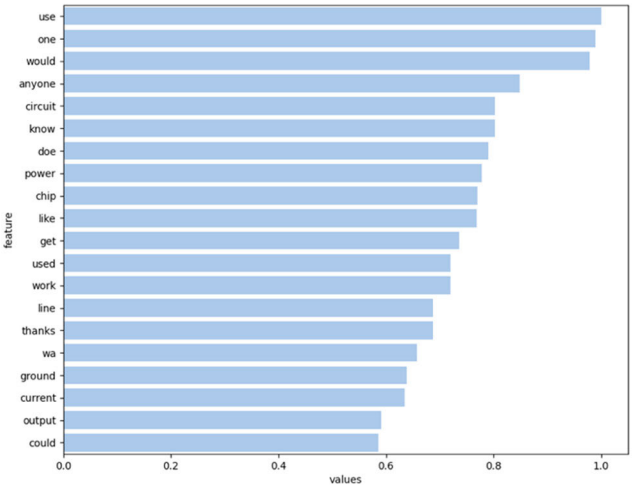

FIGURE 22. Global feature importance of the top 20 in the "sci.electronics" class. The following words appear in this order of importance. Many words appearing here are related to "electronics."

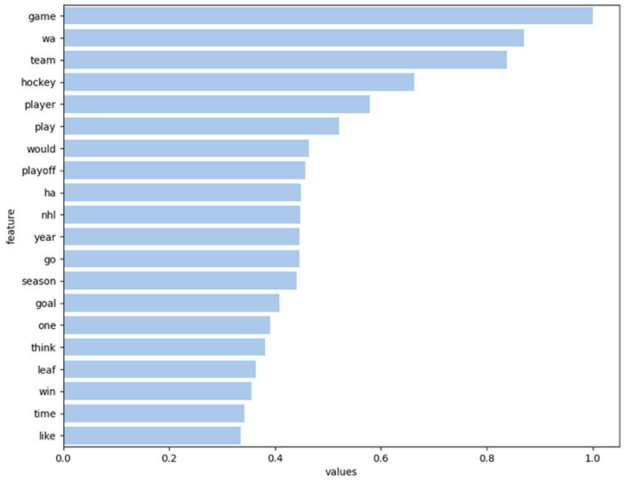

FIGURE 21. Global feature importance of the top 20 in the "rec.sport.hocky" class. The following words appear in this order of importance. Many words appearing here are related to "hockey."

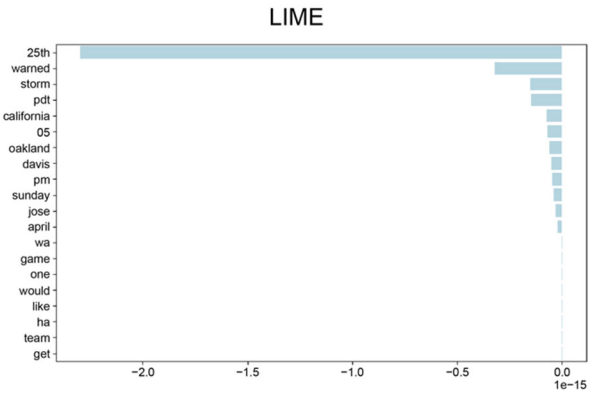

FIGURE 23. Result of the local feature importance of the top 20 for LIME of the 0th instance. "25th," "warned," "storm," "pdt," "california," "05," "oakland," "davis," "pm," "sunday," "jose," "april," and "wa" are derived as contributing negatively as the 0th data instance. Jose Mesa or Storm Davis is the name of a baseball player. In this result, the words "davis" and "jose" are extracted, but they are output as contributing negatively to "rec.sport.baseball."

Figs. 26, 27, and 28 show the results of the local feature importance of the top 20 for LIME, SHAP (Shapley value), and AIME in the 1st instance data.

In Fig. 26, in the case of the LIME, ''delight,'' ''playoff,'' ''hockey,'' ''demise,'' ''game,'' ''nhl,'' ''rag,'' ''streak,'' ''whenever,'' ''losing,'' ''loss,'' ''sure,'' ''hawk,'' ''else,'' ''notice,'' ''wondering,'' ''fan,'' ''basically,'' ''lost,'' and ''championship'' are derived as contributing as the 1st data instance. Words related to hockey tended to have greater importance for local features.

In Fig. 27, in the case of the SHAP, ''playoff,'' ''nhl,'' ''pen,'' ''even,'' ''wa,'' ''pretty,'' ''sure,'' ''hockey,'' ''game,'' ''mike,'' ''least,'' ''need,'' ''one,'' ''think,'' ''blue,'' ''much,'' ''run,'' ''year,'' ''use,'' and ''goal'' are derived as contributing as the 1st data instance. Words related to hockey tended to have greater importance for local features. However, the contribution of words that could be related to ''rec.sport.hockey,'' such as ''hockey'' and ''goal,'' is negative. This can be interpreted as a negative value because ''hockey'' and ''goal'' are related to ''rec.sport.hockey,'' but do not appear in the instance 1st data. However, deriving a feature that does not appear in an actual instance as an explanation may be difficult to interpret, especially for text data. It may be necessary to include the setting of the importance of zero features to zero for some datasets, as described in Section III-E.

In Fig. 28, in the case of AIME, ''game,'' ''hawk,'' ''wa,'' ''last,'' ''nhl,'' ''playoff,'' ''pen,'' ''since,'' ''year,'' ''chicago,'' ''one,'' ''fan,'' ''blue,'' ''even,'' ''record,'' ''need,'' ''series,'' ''sure,'' ''streak,'' and ''mike'' are derived as contributing as the 1st data instance. Words related to hockey tended to have greater importance for local features.

Figs. 29, 30, and 31 show the results of the local feature importance of the top 20 for LIME, SHAP (Shapley value), and AIME in the 5th instance data.

TABLE 3. Contents of data instance in the 20 newsgroup dataset for an experiment of extraction of local feature importance.

	TEXT	LABEL
0	Oakland, California, Sunday, April 25th, 1:05 PM PDT: Jose Mesa vs. Storm Davis. You have been warned. oakland california sunday april 25th 1 05 pm pdt jose mesa v storm davis warned	rec.sport.baseball
1	With the recent demise of the Chicago Blackhawks (much to my delight), I noticed their 8 (?) game playoff losing streak (4 to the Pens last year, and now 4 to the Blues), and I am wondering what the NHL record for consecutive losses are, if there even is one... I'm pretty sure that the Hawks have at least a 9-game losing streak, since they've had to have lost a series since their last championship (whenever that was) Basically, I need something else to rag on my Hawks-fan friend with :) Mike, the insomniaced recent demise chicago blawkhawks much delight noticed 8 game playoff losing streak 4 pen last year 4 blue wondering nhl record consecutive loss even one pretty sure hawk least 9 game losing streak since lost series since last championship whenever wa basically need something else rag hawk fan friend mike insomniaced	rec.sport.hockey
5	[request for WDC65c816 Mac cross-development stuff] Apple themselves sell a 65816 cross-developer for the Macintosh called 'MPW IIgs' (it's intended for use with the Apple IIgs computer, which uses the '816). request wdc65c816 mac cross development stuff apple sells 65816 cross developer macintosh called mpw iigs intended use apple iigs computer us 816	sci.electronics

In Fig. 29, in the case of LIME, "developer," "816," "apple," "macintosh," "cross," "intended," "mac," "development," "request," "us," "called," "stuff," "computer," "sell," "use," "wa," "one," "would," "game," and "like" are derived as contributing as the 5th data instance. Many output words are related to "sci.electronics." However, these contributions are negative. This may be difficult to interpret when users use the results from their explanations.

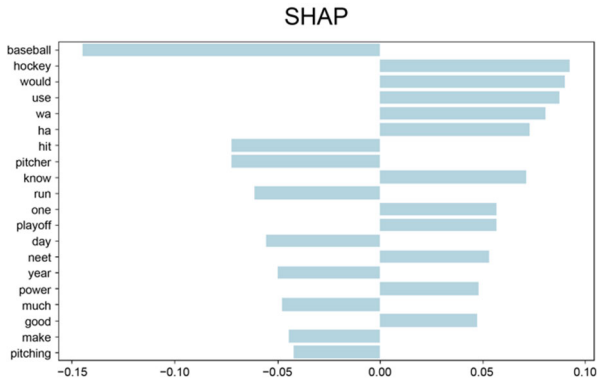

FIGURE 24. Result of Shapley values of the top 20 for SHAP of the 0th instance. "baseball," "hockey," "would," "use," "wa," "ha," "hit," "pitcher," "know," "one," "playoff," "day," "need," "year," "power," "much," "good," "make," and "pitching" are derived as contributing as the 0th data instance. Words that do not appear in the 0th instance, such as "baseball" and "hockey," are derived as having local feature importance. This means that the absence of these words is an important feature.

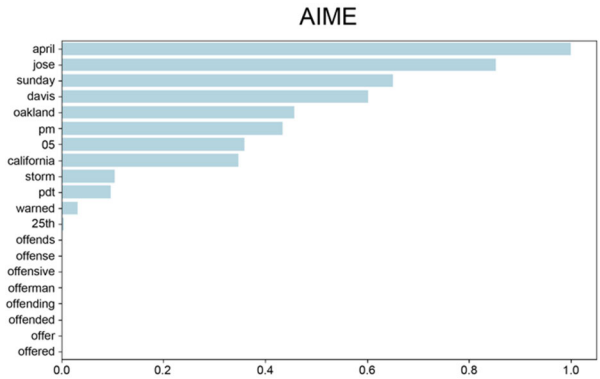

FIGURE 25. Result of the local feature importance of the top 20 for AIME of the 0th instance. "april," "jose," "sunday," "davis," "oakland," "pm," "05," "california," "storm," "pdt," and "wared" are derived as contributing as the 0th data instance. "davis" and "jose" are extracted and derived as contributing positively. In contrast to LIME and SHAP, this result may be easy for users to interpret when using it as an explanation.

In Fig. 30, in the case of SHAP, "use," "team," "game," "computer," "player," "last," "year," "one," "wa," "play," "season," "chip," "power," "get," "circuit," "baseball," "number," "league," "also," and "used" are derived as contributing as the 5th data instance. Words like "team," "game," "baseball," and "league" are related to "rec.sport.baseball" and "rec.sport.hockey." The absence of these words can be considered a contributing factor; however, they are difficult to interpret.

In Fig. 31, in the case of the AIME, "use," "computer," "sell," "stuff," "us," "called," "intended," "apple," "mac," "development," "request," "cross," and "macintosh" are derived as contributing as the 5th data

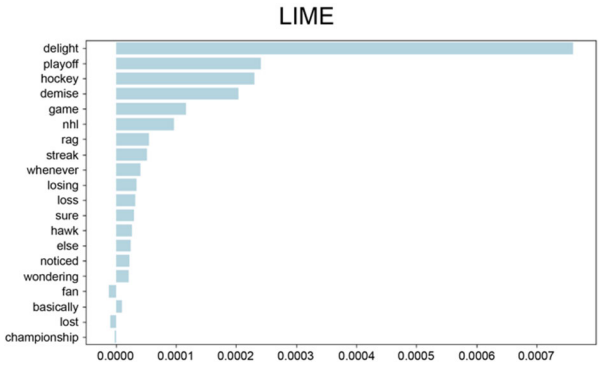

FIGURE 26. Result of the local feature importance of the top 20 for LIME of the 1st instance. "delight," "playoff," "hockey," "demise," "game," "nhl," "rag," "streak," "whenever," "losing," "loss," "sure," "hawk," "else," "notice," "wondering," "fan," "basically," "lost," and "championship" are derived as contributing as the 1st data instance. Jose Mesa or Storm Davis is the name of a baseball player. Words related to hockey tend to have greater local feature importance.

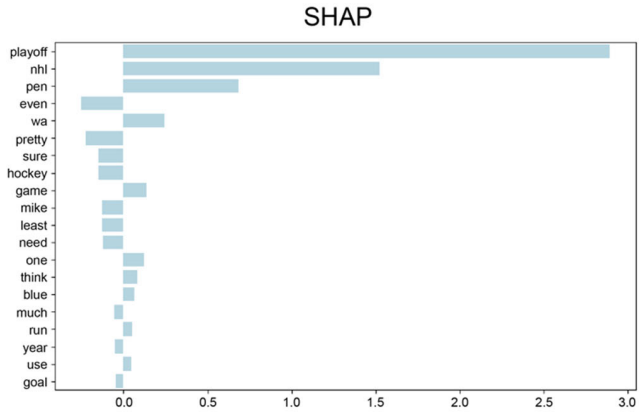

FIGURE 27. Result of the Shapley values of the top 20 for SHAP of the 1st instance. "playoff," "nhl," "pen," "even," "wa," "pretty," "sure," "hockey," "game," "mike," "least," "need," "one," "think," "blue," "much," "run," "year," "use," and "goal" are derived as contributing as 1st data instances. Words related to hockey tend to have greater local feature importance. However, the contribution of words that could be related to "rec.sport.hockey," such as "hockey" and "goal," is negative. This can be interpreted as a negative value because "hockey" and "goal" are related to "rec.sport.hockey," but do not appear in the instance 1st data. However, deriving a feature that does not appear in the actual instance as an explanation, especially for text data, may be difficult to interpret. It may need to include setting the importance of 0 features to 0 for some datasets, as described in Section III-E.

instance. Words related to electronics tended to be of greater importance for local features. In addition, because these words are extracted as locally key features among the words that occur in a sentence, they are more interpretable than LIME and SHAP when used as an explanation.

These results confirm that local feature importance can be extracted from LIME, SHAP (Shapley values), and AIME methods for text data. Compared with LIME and SHAP, the proposed AIME method finds local feature importance in words that appear in the data instances of interest, and the AIME can derive simple and highly interpretable results.

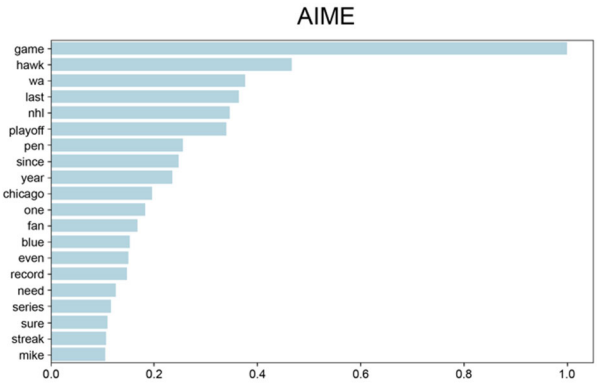

FIGURE 28. Result of the local feature importance of the top 20 for AIME of the 1st instance. "game," "hawk," "wa," "last," "nhl," "playoff," "pen," "since," "year," "chicago," "one," "fan," "blue," "even," "record," "need," "series," "sure," "streak," and "mike" are derived as contributing as the 1st data instance. Words related to hockey tend to have greater local feature importance.

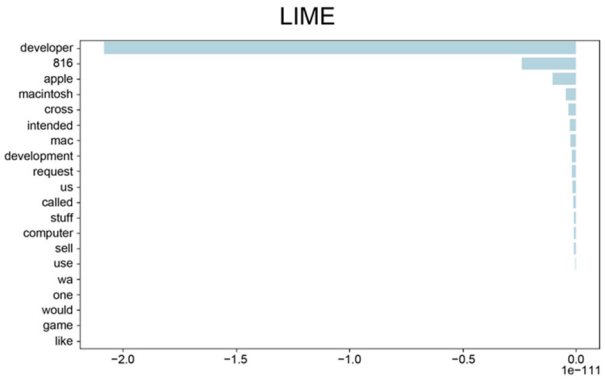

FIGURE 29. Result of the local feature importance of the top 20 for LIME of the 5th instance. "developer," "816," "apple," "macintosh," "cross," "intended," "mac," "development," "request," "us," "called," "stuff," "computer," "sell," "use," "wa," "one," "would," "game," and "like" are derived as contributing as the 5th data instance. Many of the output words are related to "sci.electronics." However, these contributions are derived as negative. This may be difficult to interpret when users use this result in their explanations.

3) REPRESENTATIVE INSTANCE SIMILARITY DISTRIBUTION PLOT

A representative instance similarity distribution plot of classes ''rec.sport.baseball'' and ''rec.sport.hockey'' in the 20-newsgroup data for the AIME is shown in Fig. 32. The distribution of the ''rec.sport.baseball'' data was wide, resulting in an approximately diagonal distribution. This indicates that some data instances are extremely close to the representative estimation instance of ''rec.sport.baseball,'' whereas others are significantly far away. In other words, we can assume that there is high diversity in the words that characterize the ''rec.sport.baseball'' data.

For a detailed relationship between the classes ''rec.sport.baseball'' and ''rec.sport.hockey,'' Fig. 33 shows the graph of Fig. 32 with the display range of both the horizontal and vertical axes changed to 0.88–0.94. According to Fig. 33, the distribution of the class ''rec.sport.hockey''

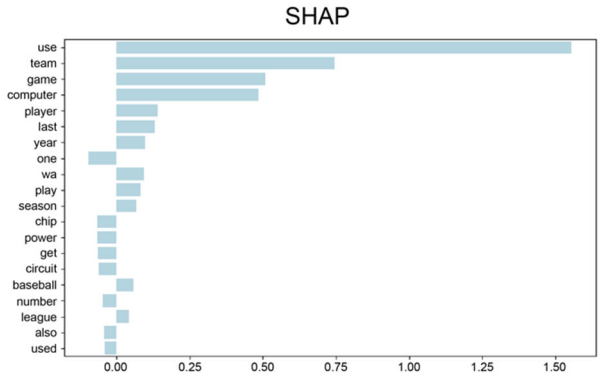

FIGURE 30. Result of the Shapley values of the top 20 for SHAP of the 5th instance. "use," "team," "game," "computer," "player," "last," "year," "one," "wa," "play," "season," "chip," "power," "get," "circuit," "baseball," "number," "league," "also," and "used" are derived as contributing as the 5th data instance. Wors like "team," "game," "baseball," and "league" related to "rec.sport.baseball" and "rec.sport.hockey" are derived. Of course, the absence of this word can be seen as a contribution; however, it is difficult to interpret.

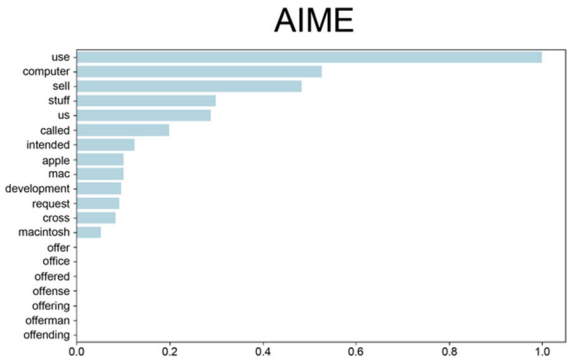

FIGURE 31. Result of the local feature importance of the top 20 for AIME of the 5th instance. The terms "use," "computer," "sell," "stuff," "us," "called," "intended," "apple," "mac," "development," "request," "cross," and "macintosh" are derived as contributing as the 5th data instance. Words related to electronics tend to have greater local feature importance. In addition, these words are extracted as local key features among the words that occur in the sentence.

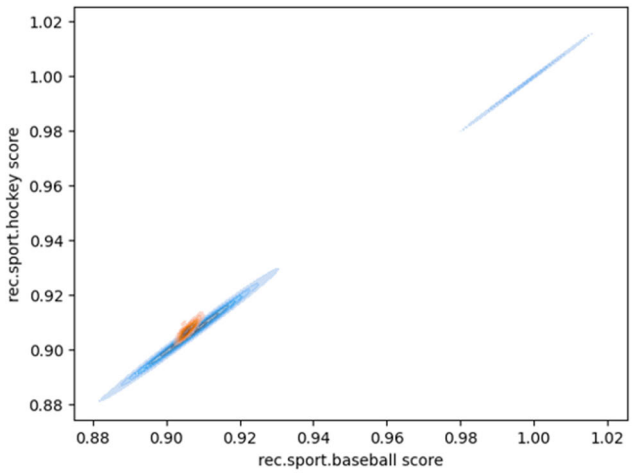

FIGURE 32. Representative instance similarity distribution plot of class "rec.sport.baseball" and class "rec.sport.hockey" in the 20-newsgroup data for AIME. The distribution of the "rec.sport.baseball" data is wide, resulting in an approximately diagonal distribution. This indicates that some data instances are extremely close to the representative estimation instance of "rec.sport.baseball," while others are significantly far away.

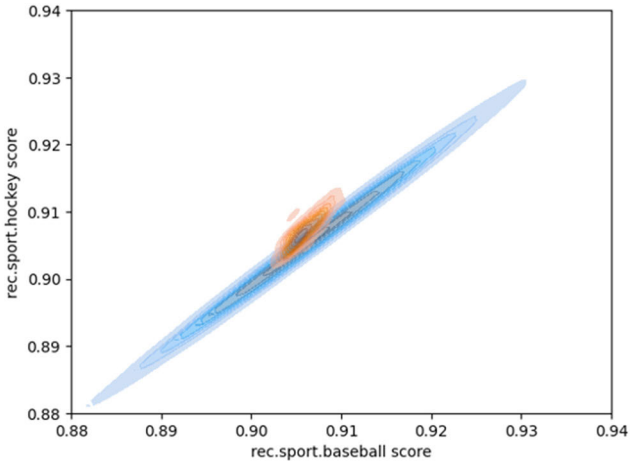

FIGURE 33. Graph of Fig. 30 with the display range of both horizontal and vertical axes changed to 0.88–0.94. The distribution of class "rec.sport.hockey" is smaller than that of class "rec.sport.baseball." Furthermore, there are areas where the two distributions overlap.

is smaller than the class "rec.sport.baseball." Furthermore, there were areas in which the two distributions overlapped. This objectively indicates that the class "rec.sport.hockey" contains some data that are difficult to determine in the classification of the class "rec.sport.baseball." Here, "rec.sport.hockey" cannot be easily judged in the classification of the class "rec.sports.baseball."

Fig. 34 shows a representative instance similarity distribution plot for the classes "rec.sport.baseball" and "sci.electronics," with horizontal and vertical axes in the range 0.88–0.94. The distribution of the class "sci.electronics" has a smaller range than that of "rec.sport.hockey" as shown in Fig. 33. The distribution of "rec.sport.baseball" is wider; thus, we can assume that the complexity of these two classes is also higher.

Fig. 35 shows the representative instance similarity distribution plot for the classes "rec.sport.hockey" and "sci.electronics." Both "rec.sport.hockey" and "sci.electronics" have small distributions. Here, both "rec.sport.hockey" and "sci.electronics" distributions are relatively separated. This suggests that the classification of these classes was relatively less complex.

From these results, the representative instance similarity distribution plot of the AIME proposed in this study can determine the complexity of the classification problem and help to understand the distribution of the data comprising the model. However, the complexity visualization in this method shows the distribution of the representative estimation instance and each instance of the dataset by the RBFkernel and does not visualize the actual contents of the original model (black-box model). This is a limitation of model-agnostic interpretability methods as well as AIME.

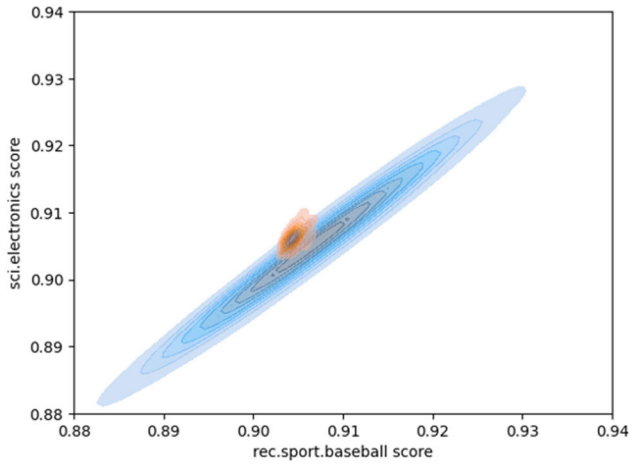

FIGURE 34. Representative instance similarity distribution plot for class "rec.sport.baseball" and the class "sci.electronics" with horizontal and vertical axes in the range 0.88–0.94. The distribution of the class "sci.electronics" has a smaller range than that of "rec.sport.hockey," as shown in Fig. 33.

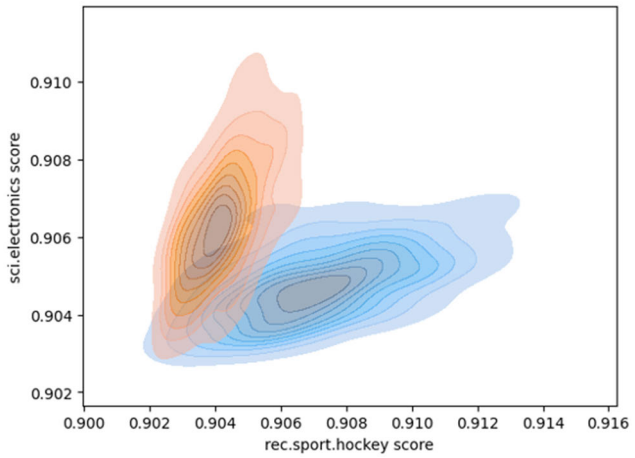

FIGURE 35. Representative instance similarity distribution plot for the classes "rec.sport.hockey" and "sci.electronics." Both classes "rec.sport.hockey" and "sci.electronics" have a small distribution. Both "rec.sport.hockey" and "sci.electronics" distributions are relatively separated.

E. DISCUSSION

The experimental results suggest that AIME could be an efficient method for obtaining explanations from original models (black-box models). In particular, deriving the generalized inverse matrix allowed us to construct an approximate inverse operator and understand the predicted results of the model. AIME can show global and local feature importance and allows the visualization of data distribution and classification complexity in representational instance similarity distribution plots. The LIME, Shapley values of the SHAP, and AIME were compared for local feature importance, and it was found that AIME derives a simpler and more easily interpretable explanation.

The AIME results are a major step toward a better understanding of model behavior. Explaining the predicted results of the black-box models improves model reliability and transparency, which may ultimately lead to better decision-making.

It is also possible that the AIME method proposed in this study may not provide sufficient interpretability, particularly for high-dimensional data and multilayer network models. This is because there are many parameters and interactions in these models, which makes the generalized inverse computation in AIME more difficult. A potential solution to this problem is the introduction of regularization to derive a generalized inverse matrix. Regularization is a common technique for controlling the model complexity and preventing overlearning. Thus, using it to derive the AIME generalizable invertible matrix may improve consistency and reliability by limiting the complexity of the generated generalized inverse matrix and creating generalized inverse elements that prevent overlearning. Investigating formulations using these regularization methods and validating their effectiveness remains a subject for future research.

AIME derives the approximate inverse operator $A^{\dagger}$ from datasets X, $\widehat{Y}$ without deriving the approximate forward operator A. The key point is that $A^{\dagger}$ is primarily derived from $\widehat{Y}$, rather than X. This could potentially render AIME robust to multicollinearity in the explanatory variables X. In situations of high correlation between explanatory variables, traditional regression analysis and some XAI methods, which are centered on explanatory variables, may produce unstable estimates. However, because $\widehat{Y}$ is a product of the black-box model, if the model does not appropriately handle multicollinearity, it may indirectly inherit this issue. Nevertheless, the influence of X's multicollinearity on AIME is smaller than its influence on other methods, owing to AIME's approach of focusing on $\widehat{Y}$ when deriving $A^{\dagger}$. However, note that AIME cannot completely avoid the multicollinearity issue of X. To verify these hypotheses, future work will include experiments and validations using datasets that feature highly correlated explanatory variables.

In addition, deriving the strength of dependencies and causal relationships between features, such as Shapley Flow [26], is an important measure of interpretability. The realization of new application methods for AIME that introduce dependencies and causal relationships among features will be studied in future work.

A key distinction of AIME from previous interpretive machine learning and XAI techniques is that it achieves interpretability by constructing an approximate inverse operator for the original model (black-box model). This helped visualize the distribution and complexity of the dataset using representative estimation instances while capturing the global and local feature importance. Elucidating the original model (black-box model) while using all features of AIME and various interpretive machine learning and XAI techniques is highly significant for the future development of machine learning and AI technologies and for ensuring reliability, transparency, responsibility, and accountability. Exploring methods to integrate and evaluate these numerous

interpretative machine learning approaches and XAI may be addressed in future studies.

V. CONCLUSION

In this study, AIME is a method capable of explaining the behavior of black-box models and the properties of data by deriving approximate inverse operators. AIME derives the local and global feature importance using an approximate inverse operator for the black-box model and target dataset. This method uniquely constructs an approximate inverse operator from the data and estimates and thus provides an assessment of the feature importance at both the local and global levels.

Additionally, we introduced a representative instance similarity distribution plot, which provides visual insights into the predictive behavior of the model and the target dataset. Although deriving approximate inverse operators using a generalized inverse matrix can pose several challenges when applied to highly complex black-box models, a representative instance similarity distribution plot offers insights into these complexities.

Our implementation of AIME was validated based on its application to interpretable machine learning. The AIME was compared with existing methods, LIME and SHAP, for local feature importance. The results demonstrated the ability to derive simple and highly interpretable explanations. Furthermore, our research suggests that the representative instance similarity distribution plot provides an objective visualization of data distribution and classification complexity.

Furthermore, our study suggests that AIME, which estimates an approximate inverse operator $A^{\dagger}$ primarily from output $\widehat{Y}$, may show robustness against multicollinearity in explanatory variables X, an advantage over other interpretive machine learning and XAI methods that use linear regression. However, if the original black-box model does not adequately handle multicollinearity, it can indirectly inherit this issue. Future studies should focus on conducting experiments with datasets containing highly correlated explanatory variables to further explore and validate the potential advantages of AIME.

Future work will explore the application of regularization methods when deriving approximate inverse operators for AIME, the use of AIME in scenarios where causal relationships exist between features and practical applications in real-world problems.

In the broader context of machine learning interpretability, AIME represents a groundbreaking approach that not only demystifies black-box models but also paves the way for more transparent and reliable decision-making processes. Its unique method of deriving feature importance through approximate inverse operators holds the potential to redefine our comprehension and trust in complex algorithms. By emphasizing both local and global feature importance and introducing visual aids, such as the representative instance similarity distribution plot, AIME offers a comprehensive framework poised to significantly advance the field of machine learning interpretability. Given the escalating demand for transparent AI and machine learning models, particularly in critical sectors like healthcare, finance, and public policy, AIME's contributions are poised to play an indispensable role in bridging the divide between intricate algorithms and human understanding.

APPENDIX

Experiment 2, detailed in Section IV-C, utilized the MNIST dataset [45]. In the initial phase of the experiment, 28×28 grayscale handwritten digit image data was converted into a 784-dimensional vector, and a black-box model was constructed using LightGBM. However, when dealing with image data, it is more practical to use 28×28 image data as input and process it using neural networks. Accordingly, we created a black-box model using convolutional neural networks (CNNs), with 28×28 image data serving as the input. We evaluated the global feature importance using AIME, compared the local feature importance using LIME, SHAP, and AIME, and produced a representative instance similarity distribution plot utilizing AIME.

Fig. 36 presents the configuration of the CNNs constructed for this study. To handle image data features, we employed LimeImageExplainer from the lime 0.2.0.1 package. We also used DeepExplainer from the shap 0.42.0 package to calculate the Shapley values for our deep learning models. Additionally, for the generation of explanations using AIME,

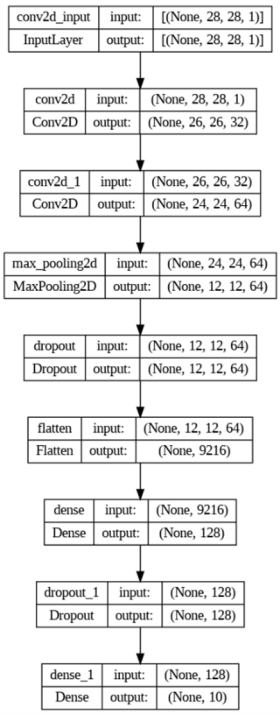

FIGURE 36. Convolutional neural network configuration. The input is 28×28 image data, and the output consists of 10 dimensions representing 0 to 9.

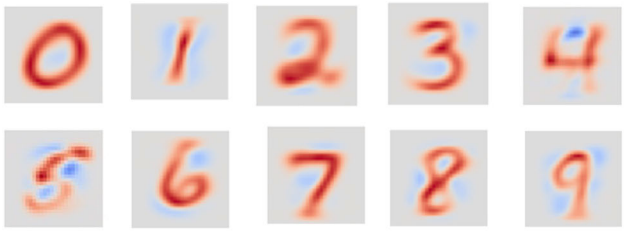

FIGURE 37. Output results for the global feature importance of the AIME in the case of MNIST data. The redder the pixel color, the higher the positive global feature importance. Gray has global feature importance of zero, and the bluer the pixel (not visible in Fig. 13, but apparent in Fig. 37), the higher the negative global feature importance. It is evident that the characteristics of each class of handwriting are captured across all classes.

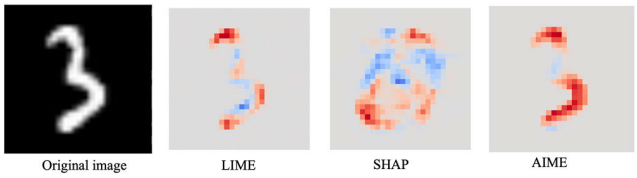

FIGURE 38. LIME, SHAP, and AIME output for "3" handwritten character. The pixel color scheme corresponds to local feature importance: the redder the pixel color, the higher the positive local feature importance. Gray has a local feature importance of zero, and the bluer the pixel, the higher the negative local feature importance. LIME shows relatively moderate feature importance for three. SHAP shows some feature importance for three, though user interpretation is challenging. Conversely, AIME clearly indicates the feature importance of three.

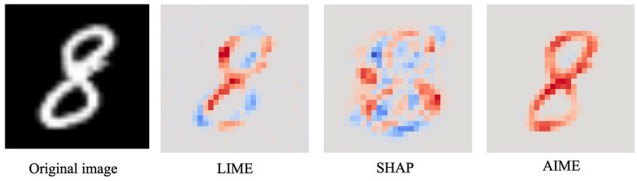

FIGURE 39. LIME, SHAP (Shapley value), and AIME output for "8" handwritten character. The color scheme remains consistent with previous figures. LIME displays some feature importance for the number eight, but SHAP is as difficult to interpret as in the "3" example shown in Figure 38. AIME, however, clearly indicates the feature importance of eight.

28 × 28 grayscale handwritten digit image data was converted into a 784-dimensional vector.

A. GLOBAL FEATURE IMPORTANCE

Fig. 37 shows the output results for the global feature importance of the MNIST data, as obtained using AIME. Each result shows the 784-dimensional global feature importance reconstructed into a 28 × 28 image.

The pixel color's intensity of red indicates the degree of positive global feature importance. Gray signifies global feature importance of zero, while the bluer a pixel (not apparent in Fig. 13 but visible in Fig. 37), the higher the negative global feature importance. Each handwriting class's unique features were captured across all classes.

From these findings, we can conclude that CNNs allow us to extract the global key features of the original MNIST model, similar to the results presented in Section IV-C-I. This understanding enables us to capture the key features of the entire model and infer its behavior.

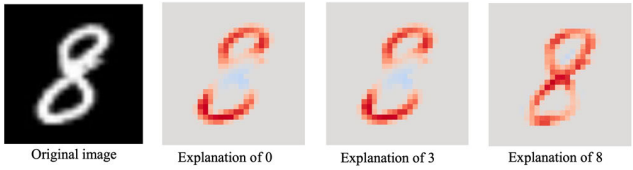

FIGURE 40. Local feature importance of the handwritten character "8" by AIME when it is recognized as "zero," "three," and "eight." As in Fig. 16, in the case of "zero" recognition, the rounded areas at the top or bottom are considered to have high local feature importance, and the "zero" feature in the number "0" is well captured. While explaining the case of "three," the importance of the left side of the number "8" diminishes, reflecting the fact that the left side of the number "3" is cut off.

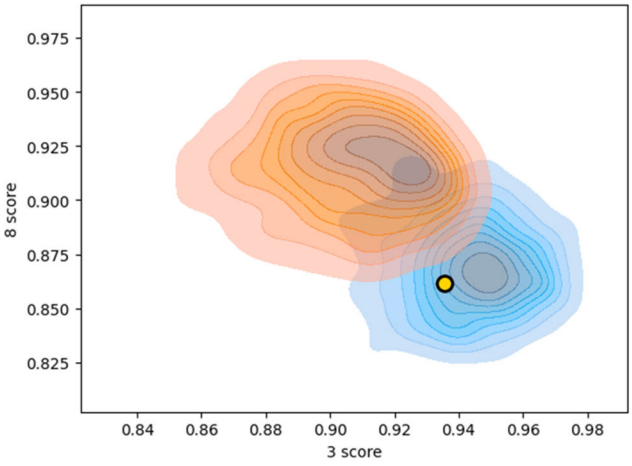

FIGURE 41. Representative instance similarity distribution plot of classes "three" and "eight" in the MNIST data for AIME. This representative instance similarity distribution plot shows that the distributions for classes "three" and "eight" are slightly farther apart than in the case of LightGBM shown in Fig. 17. Additionally, for CNNs, this representative instance similarity distribution plot suggests a clear differentiation.

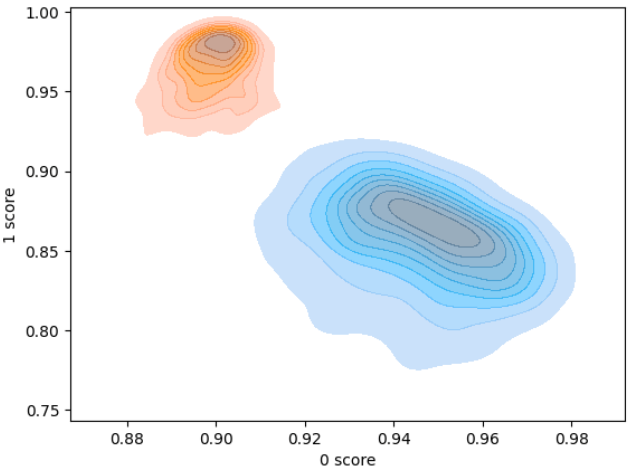

FIGURE 42. Representative instance similarity distribution plot of classes "zero" and "one" in the MNIST data. This figure shows that complexity is relatively low owing to a lack of overlap in distributions than in the case of Fig. 18.

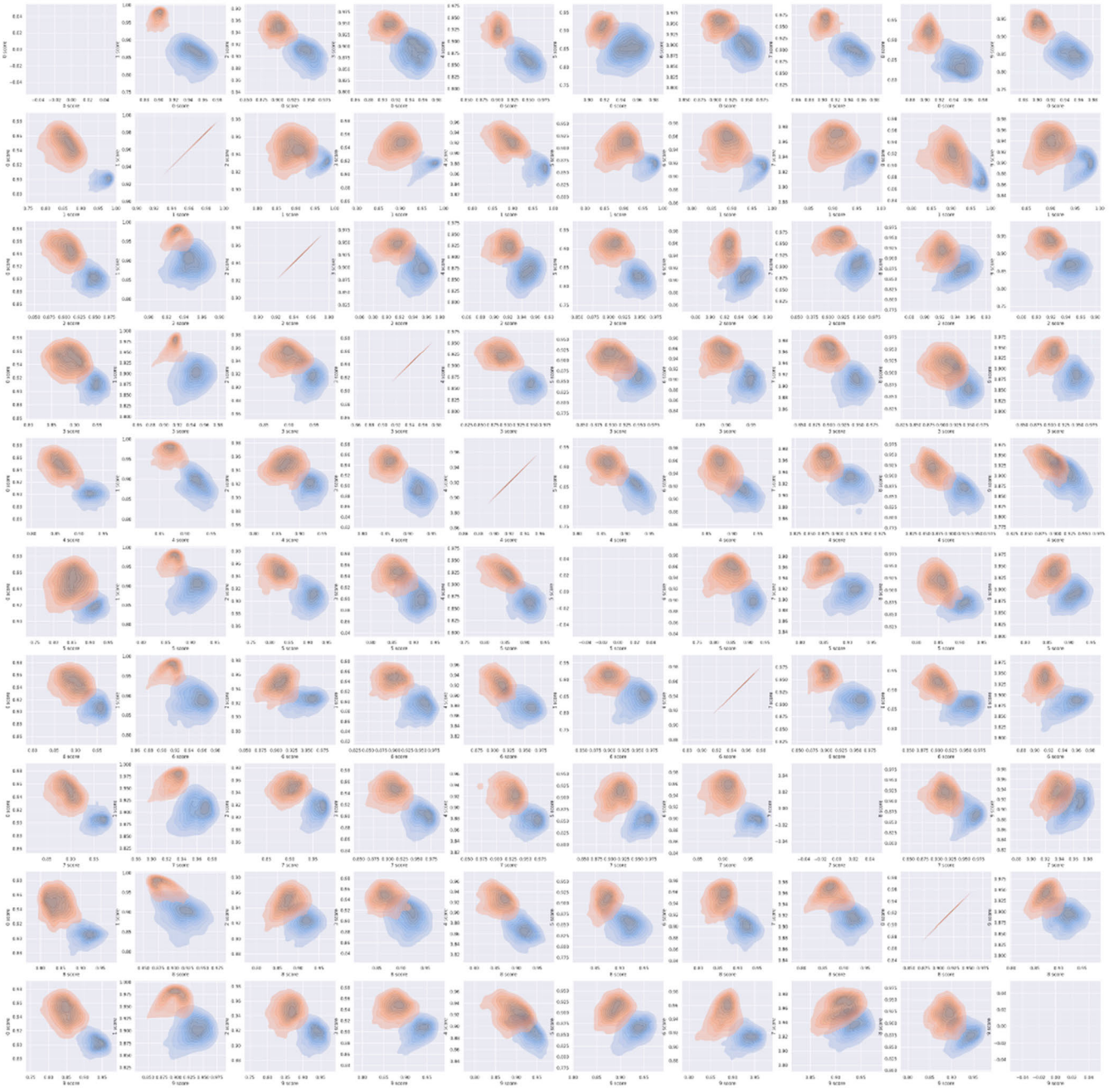

FIGURE 43. Representative instance similarity distribution plots between each class. Overall, there is less overlap in the distributions compared to Fig. 19, indicating that convolutional neural networks have lower classification complexity.

B. LOCAL FEATURE IMPORTANCE

Figs. 38 and 39 show the results for local feature importance using LIME, SHAP (Shapley value), and AIME. Each result shows the local feature importance of a 28×28 image.

Fig. 38 presents the LIME, SHAP (Shapley value), and AIME output for the ''3'' handwritten character. Notably, the handwritten character ''3'' is recognized with 44.6% certainty in the CNNs. The redder the pixel color, the higher the positive global feature importance. Gray has a global feature importance of zero. Further, the bluer the pixel, the higher the negative global feature importance. Both LIME and SHAP demonstrate some feature importance for digit three, but it is difficult for users to interpret, which clearly indicates the feature importance of the number three.

Fig. 39 shows the LIME, SHAP, and AIME outputs for the ''8'' handwritten character. In this case, the handwritten character ''8'' in the CNNs is recognized with 29.8% certainty. LIME shows a relative feature importance of eight. As in Fig. 38, LIME and SHAP are challenging to interpret, while AIME clearly demonstrates a feature importance of eight.

Both Figs. 38 and 39 illustrate that the interpretability of LIME and SHAP improved when compared to Figs. 14 and 15. However, AIME clearly shows the importance of character features in all instances.

Fig. 40 shows the local feature importance of the handwritten character ''8'' when it is identified as ''zero,'' ''three,'' and ''eight.'' Notably, the CNNs recognize the handwritten character ''8'' as ''eight'' with a certainty of 29.8%. As in Fig. 16, in the case of ''zero'' recognition, the rounded areas at the top or bottom are considered to have high local feature importance, and the ''zero'' feature in the number ''0'' is well captured. Like in the case of ''zero,'' when considering the case of ''three,'' the top or bottom rounded areas are considered to have high local feature importance. This is because the left side of the letter ''3'' is cut off.

Based on the aforementioned factors, LIME, Shapley value (SHAP), and AIME can extract the local feature importance of instances of MNIST data, with AIME offering a clearer presentation. This suggests that the inverse operator in AIME is well designed.

1) REPRESENTATIVE INSTANCE SIMILARITY DISTRIBUTION PLOT

A representative instance similarity distribution plot for classes ''three'' and ''eight'' pertaining to the ''3'' handwritten character in the MNIST data for AIME in the case of CNNs is illustrated in Fig. 41. The representative instance similarity distribution plot in Fig. 41 shows that the distributions for classes three and eight are located between the horizontal and vertical axes, indicating a relatively complex discrimination process.

In the MNIST data, classes ''zero'' and ''one'' have simple classifications, as demonstrated in Fig. 42 for CNNs. A distribution with large values on the horizontal axis corresponds to a class zero dataset, whereas a distribution with large values on the vertical axis corresponds to a class one dataset. This figure shows that the complexity was lower due to less overlap in the distribution compared to Fig. 18.

Fig. 43 shows representative instance similarity distribution plots for each class in the case of CNNs. The diagonal plots of the same class in Fig. 43 are not significant as they are plotted on a $y = x$ straight line. Overall, there is less overlap in the distributions compared with Fig. 19, indicating that CNNs have lower classification complexity.

These findings suggest that AIME can also incorporate deep learning techniques such as CNNs for explaining black-box models. Moreover, by comparing the representative instance similarity distribution plot in the case of LightGBM and the one in the case of CNNs (shown in Section IV-C-III), it becomes possible to demonstrate which model reduces classification complexity more effectively. Indeed, CNNs have exhibited lower classification complexity, indicating that CNNs might solve the classification problem while preserving the features of the image data, thereby being more expressive than the LightGBM.

ACKNOWLEDGMENT

The author would like to thank Ayako Minematsu, a Researcher with the Asia AI Institute, Musashino University, for her invaluable comments and insights, which greatly contributed to the enhancement of this article and to Editage (www.editage.com) for English language editing and also would like to thank the role of OpenAI's AI system ChatGPT in facilitating discussions and inspiring innovative ideas during the writing process. Although no direct AI-generated text was used in this study, it was instrumental in conceptualizing and refining the content. The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] M. 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. 97–101, doi: 10.18653/v1/N16-3020.

[2] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, Jan. 2012, pp. 214–226, doi: 10.1145/2090236.2090255.

[3] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, p. 10. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[4] P. Latouche, S. Robin, and S. Ouadah, "Goodness of fit of logistic regression models for random graphs," *J. Comput. Graph. Statist.*, vol. 27, no. 1, pp. 98–109, Jan. 2018, doi: 10.1080/10618600.2017.1349663.

[5] K. Bykov, M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, and M. Kloft, "Explaining Bayesian neural networks," 2021, *arXiv:2108.10346*.

[6] 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. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[7] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.

[8] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[9] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[10] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[11] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.

[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[13] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *J. Bus. Econ. Statist.*, vol. 39, no. 1, pp. 272–281, Jan. 2021.

[14] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, Jan. 2015, doi: 10.1080/10618600.2014.907095.

[15] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," 2018, *arXiv:1801.01489*.

[16] J. Liu, N. Danait, S. Hu, and S. Sengupta, "A leave-one-feature-out wrapper method for feature selection in data classification," in *Proc. 6th Int. Conf. Biomed. Eng. Informat.*, Dec. 2013, pp. 656–660, doi: 10.1109/BMEI.2013.6747021.

[17] (2019). *LOFO Importance*. [Online]. Available: https://github.com/aerdem4/lofo-importance
[18] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014, doi: 10.1007/s10115-013-0679-x.
[19] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 598–617, doi: 10.1109/SP.2016.42.
[20] S. M. Shankaranarayana and D. Runje, "ALIME: Autoencoder based approach for local interpretability," in *Proc. 20th Int. Conf. Intell. Data Eng. Automated Learn. (IDEAL)*, Manchester, U.K., vol. 20, 2019, pp. 454–463.
[21] M. R. Zafar and N. M. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," 2019, *arXiv:1906.10263*.
[22] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME explanations for diagnostic computer algorithms," 2020, *arXiv:2006.05714*.
[23] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "ILIME: Local and global interpretable model-agnostic explainer of black-box decision," in *Proc. 23rd Eur. Conf. Adv. Databases Informat. Syst. (ADBIS)*, vol. 23, 2019, pp. 53–68.
[24] S. Bramhall, H. Horn, M. Tieu, and N. Lohia, "QLIME—A quadratic local interpretable model-agnostic explanation approach," *SMU Data Sci. Rev.*, vol. 3, no. 1, p. 4, 2020.
[25] R. Gaudel, L. Galárraga, J. Delaunay, L. Rozé, and V. Bhargava, "S-LIME: Reconciling locality and fidelity in linear explanations," in *Proc. Int. Symp. Intell. Data Anal.*, 2022, pp. 102–114.
[26] J. Wang, J. Wiens, and S. Lundberg, "Shapley flow: A graph-based approach to interpreting model predictions," in *Proc. 24th Int. Conf. Artif. Intell. Stat.*, 2021, pp. 721–729.
[27] C. W. Ayad, T. Bonnier, B. Bosch, and J. Read, "Shapley chains: Extending Shapley values to classifier chains," in *Proc. Int. Conf. Discovery Sci.*, 2022, pp. 541–555.
[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018. [Online]. Available: https://ojs.aaai.org/index.php/aaai/article/view/11491, doi: 10.1609/aaai.v32i1.11491.
[29] J. Kim and J. Seo, "Human understandable explanation extraction for black-box classification models based on matrix factorization," 2017, *arXiv:1709.06201*.
[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, vol. 13, Sep. 2014, pp. 818–833.
[31] A. Barbalau, A. Cosma, R. T. Ionescu, and M. Popescu, "A generic and model-agnostic exemplar synthetization framework for explainable AI," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, Ghent, Belgium, 2021, pp. 190–205, doi: 10.1007/978-3-030-67661-2_12.
[32] C. Molnar, *Interpretable Machine Learning*. Lulu Com, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 2239–2250, doi: 10.1145/3531146.3534639.
[40] W. Samek and K. R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Lecture Notes in Computer Science), vol. 11700. Cham, Switzerland: Springer, 2019, pp. 5–22. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-28954-6_1, doi: 10.1007/978-3-030-28954-6_1.
[41] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Sep. 2017, doi: 10.1609/aimag.v38i3.2741.
[42] E. H. Moore, "On the reciprocal of the general algebraic matrix," *Bull. Amer. Math. Soc.*, vol. 26, pp. 394–395, 1920.
[43] R. Penrose, "A generalized inverse for matrices," *Math. Proc. Cambridge Phil. Soc.*, vol. 51, no. 3, pp. 406–413, Jul. 1955, doi: 10.1017/S0305004100030401.
[44] T. F. E. Harrell Jr., and T. Cason, "Titanic dataset," Tech. Rep., 2017.
[45] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.
[46] *20 Newsgroups Dataset*. Accessed: Sep. 14, 2023. [Online]. Available: http://qwone.com/~jason/20Newsgroups/

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 2014. In April 2018, he became an Associate Professor with the Global Communication Center. He was appointed as an associate professor. He has been engaged in the research and development of knowledge cluster systems with the National Institute of Information and Communications Technology, International University, where he was engaged in the research and development of text mining and data mining methods, in March 2006. At the Department of Mathematical Engineering, Faculty of Engineering, Musashino University, since April 2019, where he has been an Associate Professor with the Department of Data Science, Faculty of Data Science. His research interests include XAI, data mining, emotional information processing, and media content analyses.