IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 June 2024, accepted 19 August 2024, date of publication 26 August 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450299

RESEARCH ARTICLE

PCAIME: Principal Component Analysis-Enhanced Approximate Inverse Model Explanations Through Dimensional Decomposition and Expansion

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

This work was supported by the Department of Data Science, Musashino University, Tokyo, Japan.

ABSTRACT Complex "black-box" artificial intelligence (AI) models are interpreted using interpretive machine learning and explainable AI (XAI); therefore, assessing the importance of global and local features is crucial. The previously proposed approximate inverse model explanation (AIME) offers unified explanations of global and local feature importance. This study builds on that foundation by focusing on assessing feature contributions while also examining the multicollinearity and correlation among features in XAI-derived explanations. Given that advanced AI and machine learning models inherently manage multicollinearity and correlations among features, XAI methods must be employed to clearly explain these dynamics and fully understand the estimation results and behaviors of the models. This study proposes a new technique called principal component analysis-enhanced approximate inverse model explanation (PCAIME) that extends AIME and implements dimensionality decomposition and expansion capabilities, such as PCA. PCAIME derives contributing features, demonstrates the multicollinearity and correlation between features and their contributions through a two-dimensional heat map of principal components, and reveals selected features after dimensionality reduction. Experiments using wine quality and automobile mile-per-gallon datasets were conducted to compare the effectiveness of local interpretable model-agnostic explanations, AIME, and PCAIME, particularly in analyzing local feature importance. PCAIME outperformed its counterparts by effectively revealing feature correlations and providing a more comprehensive perspective of feature interactions. Significantly, PCAIME estimated the global and local feature importance and offered novel insights by simultaneously visualizing feature correlations through heat maps. PCAIME could improve the understanding of complex algorithms and datasets, promoting transparent AI and machine learning in healthcare, finance, and public policy.

INDEX TERMS Approximate inverse model explanation, explainable artificial intelligence, feature correlation, feature importance, model explanation, principal component analysis, principal component analysis-enhanced approximate inverse model explanation.

I. INTRODUCTION

Machine learning and artificial intelligence (AI) are becoming increasingly central to decision-making in various sectors, such as automated driving, medical diagnostics, and financial transactions. The expanding impact of these technologies on critical decisions highlights the need for a clear understanding of how predictions and estimates are derived and which data features most significantly influence outcomes. However, grasping the internal workings of complex models, including deep learning, remains challenging. Nevertheless, the field of explainable AI (XAI) has seen

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

significant research and development in recent years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20].

A previous study derived approximate inverse operators for black-box models and used them in combination with the dataset of interest to explain black-box model behavior and data properties. Further, the local and global feature contributions were estimated using a new and versatile approach, called approximate inverse model explanation (AIME) [21]. AIME can be effectively applied to various data types. In contrast to the existing interpretable machine learning and XAI methods, AIME provides both global and local feature importance and facilitates the visualization of the relationship between these estimates using similarity distribution maps for representative estimation instances. The representative estimation instances are typical data points for estimation results derived from a black-box model. The previous study [21] also showed that AIME can derive clearer and more interpretable explanations than local interpretable model-agnostic explanations (LIME) [22] and Shapley additive explanations (SHAP) [23]. Furthermore, the explanatory features derived by AIME are robust against multicollinearity.

This study highlights the importance of displaying both multicollinearity and correlations among features in the explanations provided by XAI, which is crucial not only for understanding the behavior of a model but also for interpreting the contributions of specific features to the estimated results for individual data points. In simpler, transparent models such as linear and logistic regression, multicollinearity and correlations can decrease the accuracy and complicate the interpretation of explanatory coefficients. Consequently, feature selection and dimensionality reduction are typically employed before model construction.

However, recent advancements in complex AI and machine learning models enable these systems to process multicollinearity and correlations internally, without the need for preliminary feature selection or dimensionality reduction. This capability enables the construction of models that do not require adjustments for multicollinearity and correlations among features. Nevertheless, when analyzing such models to understand their behaviors or the factors contributing to their predictions, demonstrating the presence of multicollinearity and the correlations among features is essential. Although AIME has been noted for addressing multicollinearity, it lacks the capability to explicitly articulate the relationships among features.

This study proposes principal component analysis-enhanced approximate inverse model explanation (PCAIME), which is an extension of AIME with dimensionality reduction and expansion features such as principal component analysis (PCA) [24], [25]. PCAIME derives features with large contributions, simultaneously reveals the multicollinearity and correlations among the features and their contributions using a two-dimensional (2D) heat map of principal components, and highlights features with dimensionality reduction. Thus, PCAIME can simultaneously show the correlations among features and their contributions.

Fig. 1 presents an overview of the proposed method, which introduces decomposition and expansion functions into the conventional XAI approach. Therefore, despite the use of PCA in this study, the developed approach is not limited to PCA because it is a reversible method that can realize decomposition and expansion. In this method, deriving the relationships among the features via decomposition, along with the derivation and visualization of the explanation via expansion, while maintaining the relationships among the features is a meaningful endeavor.

Furthermore, the developed approach introduces PCA for decomposition and expansion and incorporates AIME into the XAI method, which is a combination of PCA and AIME. In particular, this method of decomposition eliminates the ‘‘curse of dimensionality’’ caused by data with numerous dimensions for surrogate XAI methods such as AIME. This method facilitates the extraction of global and local features.

The effects of PCA have been clearly demonstrated in previous studies [24], [25]. However, the objectives of the present study were to combine PCA and AIME and to include the functions of decomposition and expansion in the XAI mechanism. PCAIME reveals the contributing features by examining the relationship among the features and reduces the dimensionality of the features. It aims to show the relationship among the features and to derive a more accurate explanation from the effect of dimensionality reduction.

Experiments on two datasets with multiple linearity were performed to compare LIME, SHAP, AIME, and PCAIME using heat maps, particularly for local feature importance. The corresponding results indicate the effectiveness of PCAIME in exhibiting feature correlation and contributing features in an overhead view. This proposed combined method estimates the global and local feature importance and provides new insights via heat map-based feature correlation visualization.

The primary contributions of this study are as follows:

- By introducing the functions of decomposition and expansion into AIME and combining it with a previously proposed XAI method, this study proposes a new method, PCAIME, which introduces PCA into decomposition and expansion functions. This method derives features with high contributions and simultaneously shows the correlations among the features and their contributions through a 2D heat map of the principal components and features with dimensionality reduction.
- The validity of the proposed method is verified using two datasets that exhibit multicollinearity. The 2D heat map of the principal components for global features for one dataset, including multicollinearity, are examined. This heat map expresses the behavior of the model using PCAIME. For another dataset,

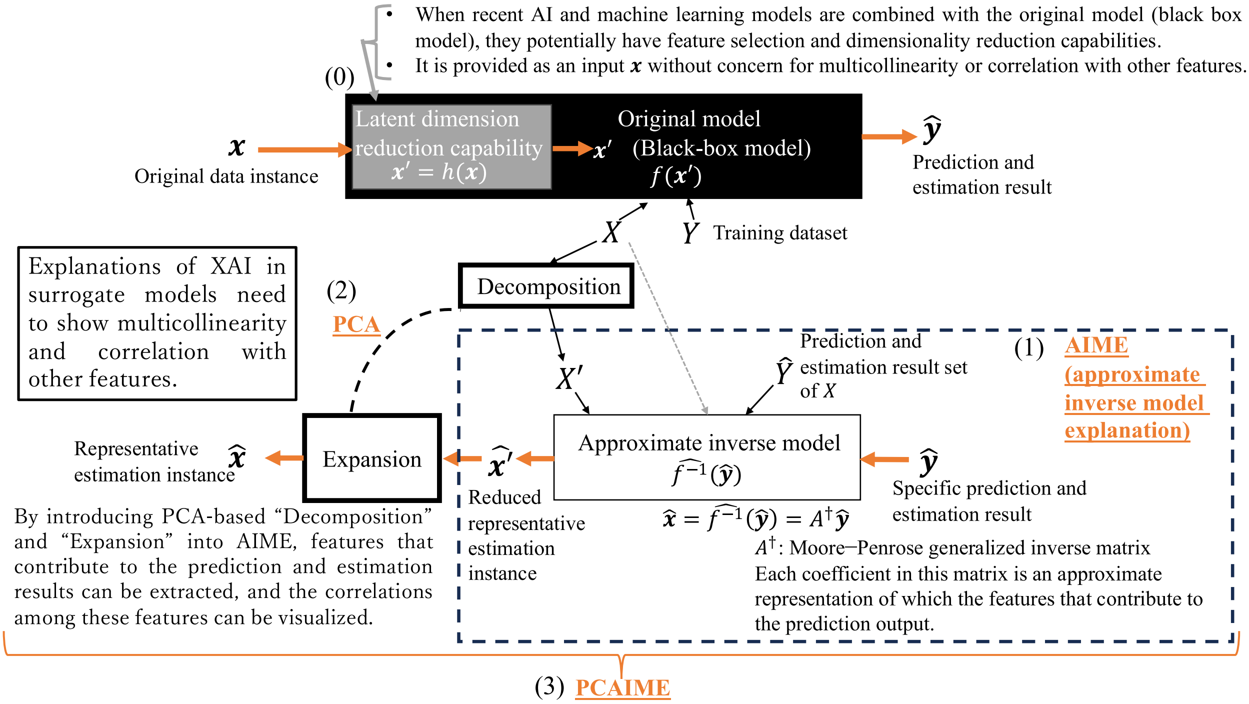
When recent AI and machine learning models are combined with the original model (black box model), they potentially have feature selection and dimensionality reduction capabilities.
It is provided as an input $\boldsymbol{x}$ without concern for multicollinearity or correlation with other features.
(0)
$\boldsymbol{x}$
Original data instance
Latent dimension reduction capability
$\boldsymbol{x}' = h(\boldsymbol{x})$
$\boldsymbol{x}'$
Original model (Black-box model)
$f(\boldsymbol{x}')$
$\hat{\boldsymbol{y}}$
Prediction and estimation result
X
Y
Training dataset
Explanations of XAI in surrogate models need to show multicollinearity and correlation with other features.
(2)
PCA
Decomposition
X'
Prediction and estimation result set of X
$\hat{Y}$
(1)
AIME (approximate inverse model explanation)
Approximate inverse model
$\widehat{f^{-1}}(\hat{\boldsymbol{y}})$
Representative estimation instance
$\hat{\boldsymbol{x}}$
Expansion
$\hat{\boldsymbol{x}}'$
Reduced representative estimation instance
$\hat{\boldsymbol{y}}$
Specific prediction and estimation result
$\hat{\boldsymbol{x}} = \widehat{f^{-1}}(\hat{\boldsymbol{y}}) = A^{\dagger}\hat{\boldsymbol{y}}$
$A^{\dagger}$: Moore–Penrose generalized inverse matrix
Each coefficient in this matrix is an approximate representation of which the features that contribute to the prediction output.
By introducing PCA-based "Decomposition" and "Expansion" into AIME, features that contribute to the prediction and estimation results can be extracted, and the correlations among these features can be visualized.
(3) PCAIME

FIGURE 1. Overview of PCAIME. The proposed method focuses on highlighting multicollinearity and feature correlations alongside explaining model behavior and feature contributions. (0) Recent complex AI and machine learning models inherently manage multicollinearity and feature correlations, eliminating the need for explicit preprocessing steps such as feature selection and dimensionality reduction. (1) However, when interpreting model behavior and feature contributions, multicollinearity and feature correlations must be addressed. Although AIME is robust against multicollinearity, it lacks an explicit representation of feature relationships. (2) Introducing decomposition to delineate feature relationships and expansion to elucidate explanations while preserving these relationships is imperative. (3) PCAIME is proposed, incorporating PCA-based decomposition and expansion functions. The focus of this study is not merely combining PCA and AIME, but rather, devising a new XAI approach that explicitly showcases multicollinearity and feature relationships through decomposition and expansion functions.

PCAIME is used to derive the global feature importance, which describes the behavior of the model and estimates derived from specific data. The local feature importance elucidates the derivation of estimates, which are compared with the results obtained using LIME, SHAP, and AIME. The results indicate that 2D heat maps with PCAIME provide valid insights by simultaneously showing the features contributing to the estimates and the relationships among the features.

The remainder of this paper is organized as follows. Section II presents prior work relevant to this study. Section III provides an overview of the recently demonstrated method, namely, AIME. Section IV details the PCAIME formulation. Section V describes comparative experiments conducted using LIME, SHAP, AIME, and PCAIME to validate the effectiveness of PCAIME. Finally, Section VI highlights the major conclusions drawn from the study findings.

II. RELATED WORKS

As indicated in the previous section, research on XAI has surged in recent years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. Speith [17] broadly divided XAI methods into ante-hoc and post-hoc methods. The ante-hoc method introduces a transparent model that can be explained in advance. The model is constructed using the linear regression, decision tree, and k-nearest neighbor methods, and explanations are derived based on the model values. In contrast, post-hoc methods are employed for complex or less transparent models.

Post-hoc methods are broadly classified into model-specific and model-agnostic methods. Model-specific methods are highly interpretable; however, the range of models is limited. For example, Grad-CAM [26] is an interpretability method for convolutional neural networks and is effective for computing the importance scores of neural networks. In addition, several other model-specific interpretability methods [27], [28], [29], [30] for deep neural networks have been reported to date.

Model-agnostic methods can be divided into three main categories. The methods in the first category are used to understand the behavior of black-box models by varying the input and training data using the target model. These methods include partial dependency plots [31], [32], which aid in the visualization of the estimated value and impact of each feature, and individual conditional explanations [33], which evaluate the importance of a feature by randomly reordering or removing specific features. In addition, certain methods evaluate the feature importance by randomly reordering or deleting specific features and via permutation feature importance [34] and leave-one-feature-out importance [35], [36]. The methods in the second category extract the important features in a forward direction using a different technique. LIME [22] is a model-agnostic method that generates explanations for individual predictions and estimates by evaluating the contributions of specific features. Other methods have been proposed as extensions of LIME [22], including ALIME [37], DLIME [38], OptiLIME [39], ILIME [40], QLIME-A [41], and S-LIME [42]. Anchors [43], an extension of LIME, is a model-agnostic interpretability method that provides rule-based explanations. SHAP [23] is a model-agnostic method that interprets the contribution of each feature to a given prediction. Shapley values based on cooperative game theory can be used to represent the contribution of each feature to a predicted outcome; however, this approach incurs a high computational cost [44]. The methods in the third category extract the important features in an inverse direction using a different technique; AIME belongs to this category [21].

XAI is crucial for explaining the behavior of black-box models, estimating the results derived from such models, and revealing the relationships and multicollinearity among the contributing features. The relationships and multicollinearity among the contributing features are revealed by clarifying the manner in which the input features interact and contribute to the predictions obtained from the decision-making process of the model. As an extension of SHAP [23], Shapley Flow [45] determines direct and indirect feature contributions from a given causal graph. In this method, a user requiring an explanation must provide this relationship. Shapley Chains [46] extends SHAP [23] by incorporating label interdependence into the explanation design process to ensure that the explanations reflect the interdependence of multiple output predictions. This method includes multiple classifiers to obtain the Shapley value, derives the relationships among the features, and derives explanations that consider the interdependence among the features.

GLIME [47] combines LIME with the graphical least absolute reduction and selection operator, which generates undirected graph models to capture both direct and indirect feature effects and reveal the conditional relationships between the features and model decisions, providing novel graphical explainability tools.

Lo and Yin [48] introduced an influence score (namely, I-score) to screen for non-informative variables in images and interactively realize an environment with explainable features that are directly related to predictability. This method has been demonstrated for a pneumonia chest X-ray image dataset.

However, previous methods, such as LIME [23] and SHAP [24], do not reveal the multicollinearity or relationships among features. Thus, several authors [44], [45], [46], [47], [48] have attempted to derive relationships among features. However, these methods are ad-hoc. These methods exhibit two phases: defining or automatically extracting the relationships among features and deriving explanations for model behavior and estimation results; both have been implemented separately. Furthermore, these methods are associated with additional computational costs, even if the relationships among the features are provided in advance or are automatically extracted.

Chaudhury et al. [49] introduced a novel and user-friendly definition of XAI based on the Wasserstein distance as the backbone for i) model explainability, ii) feature explainability, and iii) explainability of decisions made by the model.

Certain researchers have utilized XAI in some applications. Gaspar et al. [50] conducted a perturbation analysis on intrusion detection systems (IDSs) using machine learning to explore their interpretability introducing LIME and SHAP as XAI. Arreche et al. [51] focused on the development of AI techniques for IDSs and analyzed six different metrics of two popular black-box XAI methods, SHAP and LIME. Palkar et al. [52] compared various XAI methods with machine learning-based algorithms and integrated patient data, including medical records and genetic profiles. They found that XAI can foster trust between patients and healthcare professionals, who must rely on AI diagnosis and treatment recommendations.

For PCA, Dorabiala et al. [53] proposed a scalable method called ensemble PCA (EPCA) that simultaneously addressed these issues in the case of data with a low-rank structure.

The incorporation of the functions of decomposition and expansion into the XAI (including AIME) method enables the derivation of explanations for model behaviors and estimation results based on the multicollinearity and relationships among the features. In addition, the decomposition function derives the relationship among the features and determines the contribution values of the features that reflect the relationship. Although these steps increase the computational complexity of dimensionality reduction, they reduce the computational complexity of the XAI (including AIME). Thus, this function is anticipated to be used in the future for processing data with numerous features.

III. AIME

Fig. 2 shows an overview of AIME, where X is a matrix of the number of features × the number of datapoints and Y is a matrix of number of the number of classes × the number of data points.

By learning these data, a black-box model function $f(\boldsymbol{x})$ is created, which outputs an estimate $\hat{\boldsymbol{y}}$ for an input data point $\boldsymbol{x}$. This process forms the basis for machine learning as a black-box model represented by $f(\boldsymbol{x})$. The data X used for training are fed as inputs into the black-box model, and the matrix $\hat{Y}$ of data points × number of classes is obtained. This pair of X and $\hat{Y}$ matrices represents the behavior of the black-box model. In the implementation of AIME, X is a matrix, and each $\boldsymbol{x}$ must be a vector of the number of feature dimensions.

In the above description, X is the same as the training dataset. A dataset with a similar distribution can be used or resampling can be employed to reduce the amount of data. Matrices X and $\hat{Y}$ are used to generate the approximate inverse operator $A^{\dagger}$ of the black-box model as follows [21]:

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

where $\hat{Y}^{T}$ represents the transpose matrix of $\hat{Y}$ and $\hat{Y}^{\dagger}$ represents the Moore–Penrose generalized inverse [54], [55] of matrix $\hat{Y}$. The size of matrix $A^{\dagger}$ is the number of features × number of classes. $A^{\dagger}$ can be used to obtain an approximation $\hat{\boldsymbol{x}}$ of the original data $\boldsymbol{x}$ using the operation $A^{\dagger}\hat{\boldsymbol{y}}$. In other words, $A^{\dagger}$ is defined as the approximate generalized inverse matrix of the black box model. $A^{\dagger}$ is also the approximate inverse operator that serves as a linear map from $\hat{\boldsymbol{y}}$ to $\hat{\boldsymbol{x}}$.

Herein, $A^{\dagger}$ is a matrix of features × numbers of classes (or number of objective variables in the case of regression). The first column, whose values are computed and sorted in the order of increasing absolute values, represents an important feature for recognizing the first class. Similarly, the second column, with absolute values sorted in ascending order, serves as an important feature for the second class. The same trend is observed in the overall behavior of the model.

The approximate inverse operator $A^{\dagger}$ stores the contribution of each feature by computing $A^{\dagger}\hat{\boldsymbol{y}}$ to obtain the estimate $\hat{\boldsymbol{x}}$:

$$
\boldsymbol{l} = A^{\dagger}\hat{\boldsymbol{y}} \circ \boldsymbol{x}, \qquad (2)
$$

where the local feature importance vector $\boldsymbol{l}$ comprises the number of feature dimensions. Each value represents the feature importance, and $\circ$ represents the Hadamard product [56]. The local feature importance vector $\boldsymbol{l}$ indicates the mechanism associated with the estimation of $\hat{\boldsymbol{y}}$ for a given $\boldsymbol{x}$. These absolute values indicate the importance coefficients. Therefore, these absolute values are sorted in descending order. For more information, see [21].

IV. FORMULATION OF PCAIME

A. OVERVIEW OF PCAIME

The overall process flow of PCAIME is illustrated in Fig. 3. This method comprises an original model (assuming a black-box model, complex or nontransparent AI, or machine learning model), an approximate inverse operator construction function, a global feature importance computation function, a local feature importance computation function, a decomposition function, and an expansion function. Notably, in the derivation of global/local feature importance, the explanatory variables X and $\boldsymbol{x}$ must be decomposed, and the expansion function is used to represent the relationship between the features and contributing features in a 2D heat map of the principal components and features. This heat map facilitates the observation of the relationships between features and the importance of features simultaneously. The approximate inverse operator construction function, shown in Fig. 3, creates $A^{\dagger}$ (Section III). The global feature importance computation function extracts the important features that contribute to the behavior of the model (Section III). The local feature importance computation function extracts the features that contribute to the derivation of estimate $\hat{\boldsymbol{y}}$ when a data point $\boldsymbol{x}$ is input into a black-box model (Section III). Here, "features" represents the features obtained after decomposition, or in the case of PCA, principal components. Thus, when the decomposition function is constructed in the PCA, the outputs of the global and local feature importance computation functions are the contributions of each principal component.

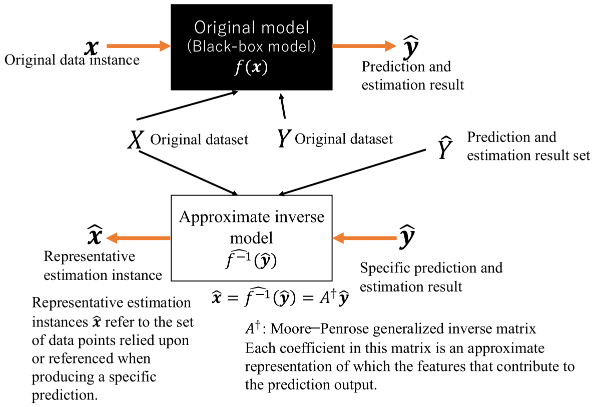
Original model
(Black-box model)
$f(\boldsymbol{x})$
$\boldsymbol{x}$
Original data instance
$\hat{\boldsymbol{y}}$
Prediction and
estimation result
X Original dataset
Y Original dataset
$\hat{Y}$
Prediction and
estimation result set
Approximate inverse
model
$\widehat{f^{-1}}(\hat{\boldsymbol{y}})$
$\hat{\boldsymbol{x}}$
Representative
estimation instance
$\hat{\boldsymbol{y}}$
Specific prediction and
estimation result
$\hat{\boldsymbol{x}} = \widehat{f^{-1}}(\hat{\boldsymbol{y}}) = A^{\dagger}\hat{\boldsymbol{y}}$
Representative estimation instances $\hat{\boldsymbol{x}}$ refer to the set of data points relied upon or referenced when producing a specific prediction.
$A^{\dagger}$: Moore–Penrose generalized inverse matrix
Each coefficient in this matrix is an approximate representation of which the features that contribute to the prediction output.

FIGURE 2. AIME for constructing approximate inverse operators.

The decomposition function processes the multiple linearity and relationships among the features and performs dimensionality reduction to address the curse of dimensionality that impedes the derivation of explanations. Consequently, for high-dimensional explanatory variables, complex or less transparent AI and machine learning models can perform feature selection and dimensionality reduction, whereas regular XAI does not contain these operations. However, the decomposition function solves this problem.

PCAIME is a combination of PCA and AIME, and the decomposition and expansion functions are assumed to be implemented in PCA.

PCAIME offers multiple advantages over conventional LIME, SHAP, AIME, and other similar methods, including better treatment of multicollinearity and relationships among features and improved computational efficiency and interpretability through dimensionality reduction. Specifically, PCAIME effectively solves the problem of multicollinearity at the preprocessing stage and provides more insightful model descriptions by considering the interactions among the features. This step is particularly important for high-dimensional datasets, wherein PCAIME can overcome the curse of dimensionality and significantly improve the model interpretability.

B. DECOMPOSITION FUNCTION

The decomposition function addresses the curse of dimensionality encountered while deriving model explanations and revealing correlations among features within complex datasets [57]. Dimensionality reduction lowers the computational cost by eliminating redundancy while preserving the essential structure of the data. However, when performing dimensionality reduction, maintaining or interpreting the correlations among the features in the original feature space is essential.

Among the diverse approaches for dimensionality reduction, PCA specifically captures the linear correlations among features, identifies the direction of the maximum variance in the dataset, and represents the original feature space using a small number of principal components. This process facilitates both the compression and interpretation of the dataset. In addition, it enables expansion using PCA loadings.

The PCAIME proposed in this study uses PCA-based decomposition to reveal the correlations among features and derive feature contributions that reflect these relationships.

C. EXPANSION FUNCTION

1) BASIC FORMULATION OF PCA LOADINGS AND HEAT MAP DERIVATION

PCA loading serves as a measure of how much the original features contribute to the principal components of PCA. These loadings are crucial for depicting the correlation between the principal components and the original variables, indicating the significance of each variable in each principal component. The loadings p_{ij} are calculated using the elements of the eigenvectors of the principal components (indicating the direction of the principal components) as follows:

$$p_{ij} = v_{ij}\sqrt{\lambda}_i, \tag{3}$$

where v_{ij} is the jth component of the eigenvector of the ith principal component and λ_i is the eigenvalue of the ith principal component. The maximum value of i is the number of principal components, and the maximum value of j is the number of features in the original data.

Let matrix P be a matrix with elements p_{ij}. The matrix P is the number of principal components × the number of features of matrix X. Then, the PCA loading P is used to construct the expansion function. This expansion function uses the PCA loadings to derive the contribution of each feature while extracting the relationship between the original feature space and principal components. Essentially, $A^{\dagger\prime}$ is derived from X' and $\hat{Y}$, which are subjected to dimensionality reduction using decomposition, and the global and local feature importance is derived. Thus, the output is attributed to the principal components. Consequently, it is a vector representing the number of dimensions of the principal components defined in Section IV-B. Let $\boldsymbol{l}$ denote a vector obtained using the following equation:

$$H = P \cdot diag\left(\boldsymbol{l}\right), \tag{4}$$

where $diag$ is a square matrix with $\boldsymbol{l}$ components along the diagonal. The matrix H is the number of principal components × the number of original features. For each value of matrix H, negative and positive values are assumed to be represented by blue and red gradients in the heatmap, respectively. Rows exhibiting similar values indicate the existence of a correlation among the features. In addition, a feature with a large value in a column indicates an important feature with a large contribution.

2) HEAT MAP FOR GLOBAL FEATURE IMPORTANCE

By substituting X', which is the number of datapoints × the number of principal components, that is, X reduced in dimension to X', for X in the formula for $A^{\dagger}$ presented in Section III-A, we obtain

$$A^{\dagger\prime} = X'\hat{Y}^{T}\left(\hat{Y}\hat{Y}^{T}\right)^{-1} = X'\hat{Y}^{\dagger}. \tag{5}$$

By setting X', $A^{\dagger\prime}$ can be represented as the number of principal components × number of classes (or number of objective variables). Thus, the first and second columns represent the principal components that contribute to the first and second classes in the model, respectively. The heat map H_k for the kth class is expressed as follows:

$$H_k = P \cdot diag\left(\boldsymbol{l}_k\right), \tag{6}$$

where $\boldsymbol{l}_k$ is the vector from which the kth column of $A^{\dagger'}$ is extracted and H_k is a matrix of principal components × original features, whose rows represent the principal components and columns represent the features of class k.

3) HEAT MAP FOR LOCAL FEATURE IMPORTANCE

By substituting the dimension-reduced x' for $\boldsymbol{x}$ in the equation shown in Section III-C, we obtain

$$\boldsymbol{l}_{PCA} = A^{\dagger\prime}\hat{\boldsymbol{y}} \circ \boldsymbol{x}'. \tag{7}$$

As discussed in Section IV-B-II, $A^{\dagger\prime}$ is the number of principal components × the number of classes (or the number of objective variables). Therefore, we must substitute for the dimension-reduced $\boldsymbol{x}'$. Consequently, the vector $\boldsymbol{l}_{PCA}$, which highlights the local feature importance, becomes the dimension of the number of principal components and indicates the principal components that contribute to the estimation result of $\hat{\boldsymbol{y}}$ for $\boldsymbol{x}'$ reduced in dimension from $\boldsymbol{x}$. From vector $\boldsymbol{l}$, the 2D heat map H_{local} of the principal components × features can be expressed as

$$H_{local} = P \cdot diag\left(\boldsymbol{l}_{PCA}\right), \tag{8}$$

where H_{local} is a matrix, whose rows represent the number of principal components and columns represent the number of features. Features with similar values in the same row are correlated, and the columns indicate the features that contribute to $\hat{\boldsymbol{y}}$ derived from the black-box model by inputting $\boldsymbol{x}$.

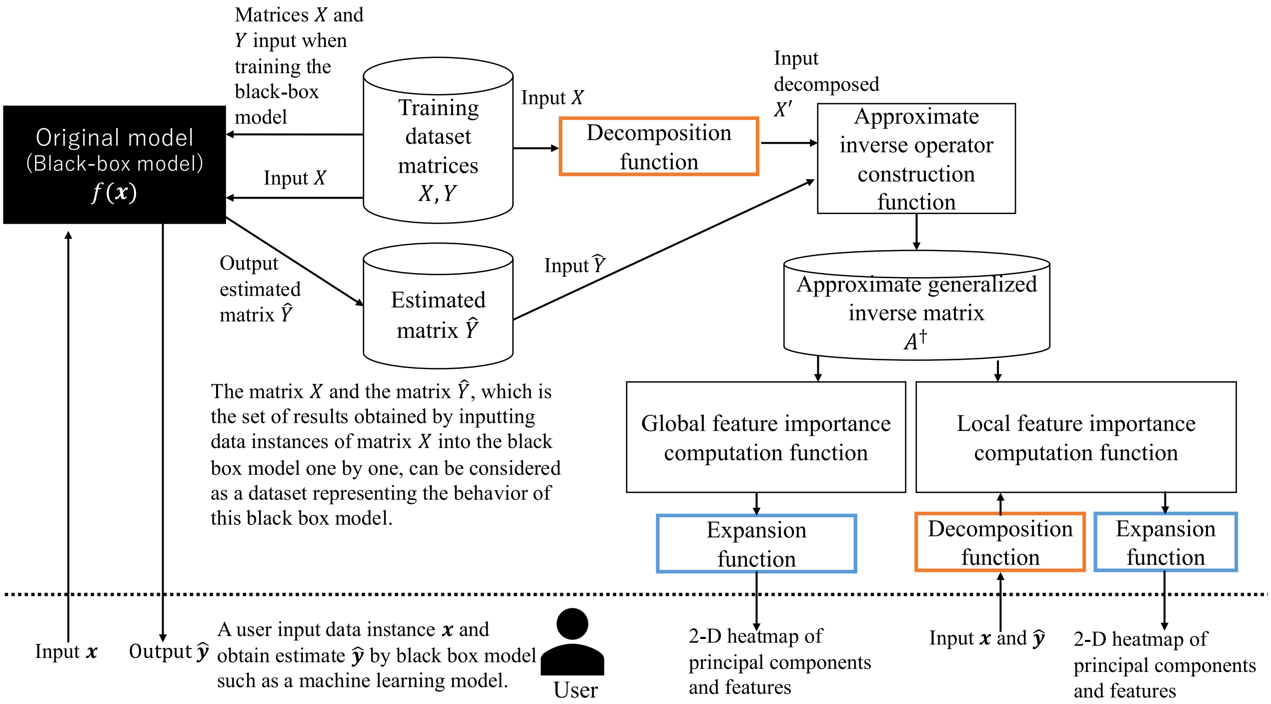
Matrices X and Y input when training the black-box model
Original model (Black-box model) $f(\boldsymbol{x})$
Training dataset matrices X, Y
Input X
Input X
Decomposition function
Input decomposed X'
Approximate inverse operator construction function
Output estimated matrix $\hat{Y}$
Estimated matrix $\hat{Y}$
Input $\hat{Y}$
Approximate generalized inverse matrix $A^{\dagger}$
The matrix X and the matrix $\hat{Y}$, which is the set of results obtained by inputting data instances of matrix X into the black box model one by one, can be considered as a dataset representing the behavior of this black box model.
Global feature importance computation function
Local feature importance computation function
Expansion function
Decomposition function
Expansion function
Input $\boldsymbol{x}$
Output $\hat{\boldsymbol{y}}$
A user input data instance $\boldsymbol{x}$ and obtain estimate $\hat{\boldsymbol{y}}$ by black box model such as a machine learning model.
User
2-D heatmap of principal components and features
Input $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$
2-D heatmap of principal components and features

FIGURE 3. Overview of the proposed method, PCAIME. This method comprises an original model (assuming a black-box model, a complex or nontransparent AI, or a machine learning model), an approximate inverse operator construction function, a global feature importance computation function, a local feature importance computation function, a decomposition function, and an expansion function. The unique feature of this method is that a decomposition function is applied before the mechanism for deriving explanations and an expansion function is used before visualizing explanations. These functions visualize the relationship between the features while explaining them through dimensionality expansion.

V. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT

In this study, experiments were conducted to assess the issue of multicollinearity and its correlation with other features, as well as to explore the mechanisms through which PCAIME generates explanations and elucidates feature relationships. This section outlines the experimental setup and provides an in-depth discussion of the results. The experiment employed the automobile mile-per-gallon (MPG; the fuel consumption rate when driving in a straight line at a constant speed on a flat paved road, expressed as the number of miles that can be traveled on one gallon of fuel) dataset (Auto MPG) [58].

The Auto MPG data consisted of 397 data points, 8 features, and 1 MPG score. These data were divided into training and test datasets. The size of X in the Auto MPG data was 298×8, and the size of Y was 298×1, resulting in dimensions of 8×1 for $A^{\dagger}$.

The black-box models were developed using various algorithms and the most accurate parameters using PyCaret 3.3.0, which is a library of AutoML. An extra tree classifier, random forest classifier, and extra tree regressor were adopted in this study because they yielded the highest prediction accuracies using PyCaret as LIME, SHAP, AIME, and PCAIME are model-agnostic methods. Then, for each dataset, the data were divided into 75% training data and 25% test data, followed by an examination of their accuracies.

In the experiment, the global feature importance was derived and compared for both the red and white wine datasets using AIME and PCAIME. Additionally, the local feature importance for specific data points was analyzed using LIME, SHAP, AIME, and PCAIME. To facilitate comparison, all outputs were visualized using heatmaps, created with Matplotlib 3.7.1, Seaborn 0.13.1, and Pandas 1.5.3. For implementation, LIME was utilized in the form of LIME 0.2.0.1, SHAP was applied in the form of SHAP 0.44.1, and AIME and PCAIME were independently implemented using scikit-learn 1.4.1 post1 and NumPy 1.25.2, respectively.

B. APPLICATION OF PCAIME TO AUTO MPG DATASET

In the first experiment, the Auto MPG dataset was utilized to derive and compare explanations of global feature importance for both AIME and PCAIME. The second experiment employed the Auto MPG dataset to derive and compare explanations based on local feature importance using LIME, SHAP, AIME, and PCAIME.

The dataset includes the following parameters: Cylinders, Displacement, Horsepower, Weight, Acceleration, Model Year, and Origin. Origin is a categorical variable and accepts “2” or “3.” A machine learning model that derives MPG was constructed as a black-box model using these features employing an extra tree regressor.

A previous study on AIME [21] focused on explanations pertinent to classification. In contrast, the current study addressed a numerical dataset—specifically, the MPG dataset—and was focused on deriving explanations for regression problems. $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$ (these are scalers), Y, and $\hat{Y}$ (these are matrices have dimensions of $1\times$the number of datapoints, $(1\ \times 294)$) can be implemented for AIME and PCAIME without modifications, except that they now represent the objective variables. In this case, matrix X has dimensions of $8\times$ 294, because it has 8 features and 294 data points.

1) GLOBAL FEATURE IMPORTANCE FOR AUTO MPG DATASET

This section explores the global feature importance as derived by applying the AIME and PCAIME methodologies to the Auto MPG dataset.

Fig. 4 presents a heat-map correlation matrix for each feature in the dataset. The results highlight that features such as Cylinders, Displacement, Horsepower, and Weight exhibit strong positive correlations.

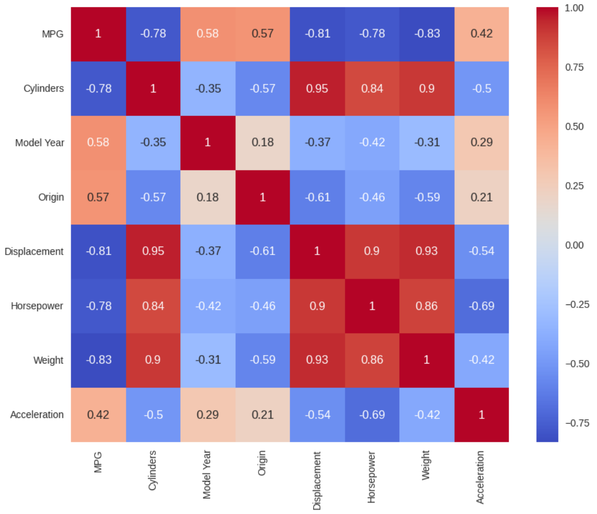

FIGURE 4. **Heat map representation of the correlation matrix of the Auto MPG dataset.**

Table 1 displays the results obtained for the variance inflation factor (VIF), an indicator utilized to assess multicollinearity in the auto MPG dataset. VIF values exceeding 10 are observed for Cylinders, Displacement, and Weight, indicating multicollinearity in the data. Additionally, the high VIF of Horsepower (9.9573) suggests a strong relationship with other characteristics. This finding indicates that the Auto MPG dataset exhibits multicollinearity, with other features showing correlations with each other.

For this analysis, the extra tree classifier was employed as the black-box model for regression. The MAE(Mean Absolute Error), MSE(Mean Squared Erro), RMSE(Root Mean Squared Error), and R2(Coefficient of Determination, R squared) values were 6.364267, 1.762663, 2.52275, and 0.873808, respectively.

TABLE 1. **VIF for the auto MPG dataset.**

Feature	VIF
Const	678.543083
Cylinders	10.737771
Model Year	1.301373
Displacement	22.937950
Horsepower	9.957265
Weight	11.074349
Acceleration	2.625906
Origin_2	1.649271
Origin_3	1.762692

For a comparative analysis, the results are presented in Fig. 5, where the Extra Trees Regressor was utilized to derive feature importance.

The contributions of Cylinders and Displacement to the regression model for the Auto MPG dataset are notable, with respective values of 0.31 and 0.26, suggesting a significant influence on the model predictions. Meanwhile, the Model Year contribution is 0.13, which, although smaller than that of Cylinders and Displacement, is still larger than those of Horsepower and Weight, each contributing 0.12. These characteristics are important predictors within the regression model. However, the nature of these contributions is to fully understand their impact on the model output.

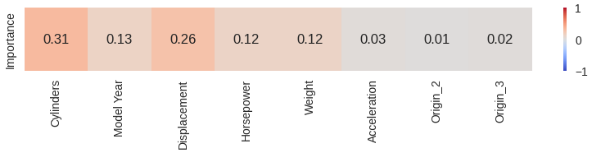

FIGURE 5. **Heat map representation of the feature importance derived using the extra tree classifier for the Auto MPG dataset.**

A heat map of the global feature importance results from the AIME for the Auto MPG dataset is shown in Fig. 6.

For Cylinders, Displacement, Horsepower, and Weight, the model shows a negative contribution, whereas Model Year, Acceleration, Origin_2, and Origin_3 contribute positively. Notably, the contribution of Origin_3 is more positive than that of Origin_2. In contrast to the functional importance depicted in Fig. 5, the global functional importance revealed by AIME clearly indicates whether the model contributes positively or negatively.

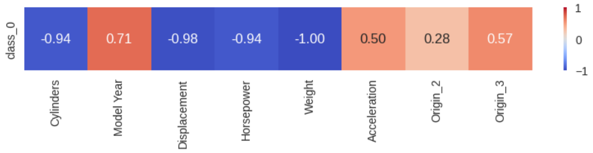

FIGURE 6. **Heat map representation of the global feature importance, derived using AIME, for the Auto MPG dataset.**

The cumulative explanation variance and scree plot for the MPG dataset quality dataset are presented in Figs. 7 and 8, respectively.

The cumulative explanation variance shown in Fig. 7 is approximately 0.85 when the number of principal components is 3. This result implies that it contains approximately 85% of the information content of the original data. The scree plot consists of a gradual slope, referred to as the scree, which indicates the number of principal components. In Fig. 8, the scree plot displays the screes for principal components 2 and 3. Consequently, PCAIME was derived for principal component 3 based on this result. In this case, matrix X' has dimensions of 3 $\times$294, matrix $A^{\dagger\prime}$ is 1$\times$ 3, matrix P is 3 $\times$8, and matrix H_k is 3$\times$8.

Fig. 9 illustrates the corresponding results of the global feature importance derived using PCAIME. Because this regression model contains only one dependent variable, the global feature importance obtained through PCAIME is efficiently captured in a single heatmap. Consistent with the previous figures, the vertical axis in these visualizations represents the principal components, and the horizontal axis denotes the features. PC1 contributes almost exclusively of the global feature importance and negatively to the model, with features such as Cylinders, Displacement, Horsepower, and Weight also contributing negatively. Conversely, Model Year, Acceleration, Origin_2, and Origin_3 contribute positively and are strongly correlated. These correlations arise because these variables move in the same direction along the PC1 axis. The second and third principal components also contribute negatively to the model. Components with positive and negative values in opposite directions within the same principal component indicate a negative correlation between these features. For example, focusing on the PC1 axis, Cylinders, Displacement, Horsepower, and Weight are negatively correlated with Model Year, Acceleration, Origin_2, and Origin _3.

This analysis focuses on the same feature as presented in Fig. 6; however, here it is represented solely by PC1. The ability to represent this feature on a single axis suggests a correlation among all features associated with PC1.

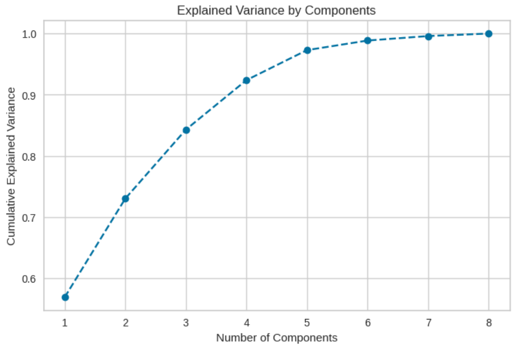

FIGURE 7. Cumulative explanation variance for the Auto MPG dataset.

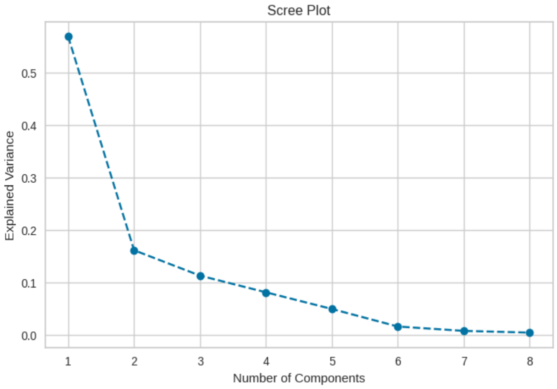

FIGURE 8. Scree plot for the Auto MGP dataset.

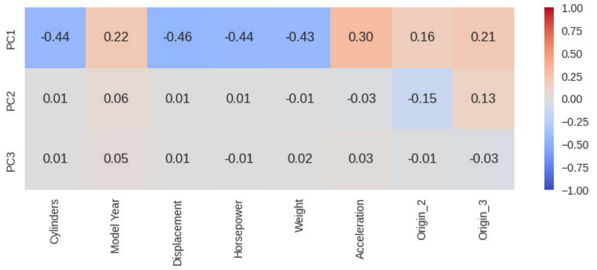

FIGURE 9. Heat map representation of the global feature importance derived using PCAIME for the Auto MPG data.

2) LOCAL FEATURE IMPORTANCE FOR AUTO MPG DATASET

This section discusses the local feature importance derived using LIME, SHAP, AIME, and PCAIME. Despite the strong multicollinearity among the features in the dataset, predicting MPG seems straightforward for extra tree regressors, and the data do not show significant disparities in results. For this analysis, the local features of a relatively fuel-efficient car, with an actual MPG of 26.0 and an extra tree regressor prediction of 25.99, were examined. Similarly, for a fuel-inefficient car, an actual MPG of 15.0 and an extra tree regressor prediction of 14.25 were analyzed. The importance of these features was then derived and compared.

Fig. 10 shows the results derived using LIME, SHAP, and AIME to explain the MPG value of 26.0 and extra tree regressor prediction of 25.99.

The specific values for this data point after normalization are as follows: Cylinder: 4; Model Year: 72; Displacement: −0.94164742; Horsepower: −0.92267201; Weight −0.92958509; Acceleration 0.89232939; Origin_2: 1; Origin_3: 0. In LIME, the positive values of Cylinders, Displacement, Horsepower, and Weight are significant. In SHAP, a high positive correlation between Cylinders and Displacement is observed, and the negative correlation of Model Year is also pronounced. In AIME, Cylinders, Displacement, Horsepower, Weight, and Origin_2 show high positive correlations, and the negative correlation of Model Year is substantial. The actual data point features positive values for Cylinders and negative values for Displacement, Horsepower, and Weight, aligning with the characteristics of a fuel-efficient vehicle. Typically, a positive value for Cylinders indicates a smaller engine size, whereas negative values for Displacement, Horsepower, and Weight suggest lower engine displacement, reduced power, and decreased vehicle mass, respectively—traits commonly associated with fuel efficiency. The accurately reflects the performance of the vehicle. However, the AIME shows a value of 0 for Origin_3, ranging from −1 to 1, which is straightforward to interpret. Yet, these values do not account for multicollinearity or depict relationships among features that are not visualized.

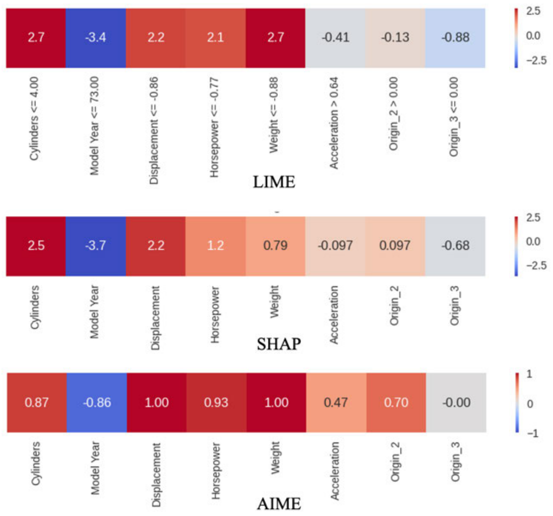

FIGURE 10. Heat map representation of the local feature importance of fuel-efficient vehicle data points derived using LIME, SHAP, and AIME (Auto MPG data).

Fig. 11 shows the results derived using PCAIME to explain the MPG value of 26.0 and the extra tree regressor prediction of 25.99.

Notably, the values of Origin_2 and Origin_3 are reversed in Fig. 11 compared to SHAP and AIME in Fig. 23. This discrepancy can be attributed to the ability of PCAIME to manage relationships between features. Indeed, one can infer that the newer model (Origin_3) is more fuel-efficient than the older model (Origin_2).

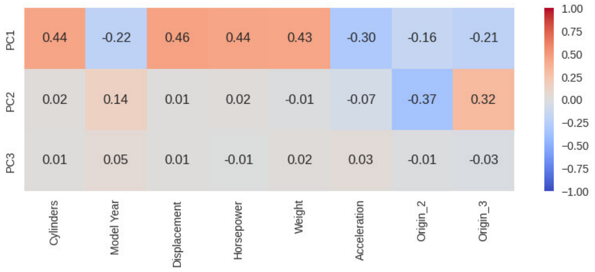

FIGURE 11. Heat map representation of the local feature importance of fuel-efficient vehicle data points derived using PCAIME (Auto MPG data).

The vertical axis represents the principal components, whereas the horizontal axis represents the features. Features that exhibit values along the same direction for the same principal component (row) can be considered correlated with each other. Fig. 11 shows that the results can be explained simply using PC1 and PC2, which is because the value of PC3 is usually low. The columns show that Cylinders, Displacement, Horsepower, and Weight are positively correlated, whereas Acceleration and Origin_2 are negatively correlated. The positive correlations of Cylinders, Displacement, Horsepower, and Weight and the negative correlations of Model Year and Weight are expressed only in PC1, confirming that these features are positively or negatively correlated with each other. In addition, because the original value of Origin_3 is 0, PC1 and PC2 cancel each other out. These results are similar to those of LIME, SHAP, and AIME. PCAIME shows the relationship between features, in addition to the contributions of the features. Thus, the principal component axis illustrates the relationships between features and the columns reveal the features that contribute to the estimation results of the data points of interest. Components with positive and negative values in opposite directions within the same principal component indicate a negative correlation between these features. For example, focusing on the PC1 axis, Model Year, Acceleration Origin_2, and Origin _3 were found to be negatively correlated with Cylinders, Displacement, Horsepower, and Weight. When focusing on the PC2 axis, Origin _2 was found to be negatively correlated with Origin _3.

Fig. 12 shows the results derived using LIME, SHAP, and AIME to explain the MPG value of 15.0 and the extra tree regressor prediction of 14.25.

The specific values for these data after normalization are as follows—Cylinders: 8; Model Year: 70; Displacement: 1.87146856; Horsepower: 2.22492429; Weight: 1.02840604; Acceleration: −2.55551732; Origin_2: 0; Origin_3: 0. For LIME, SHAP, and AIME, Cylinders, Model Year, Displacement, Horsepower, and Weight are negative capitals. However, Acceleration is smaller than −0.74 for LIME. AIME also shows a negative acceleration contribution. In contrast, SHAP exhibits a positive acceleration contribution. However, these results do not reveal any relationship among the features.

Fig. 13 presents the results derived using PCAIME to explain the MPG value of 15.0 and the extra tree regressor prediction of 14.25. The vertical axis represents the principal components, and the horizontal axis denotes the features. Features exhibiting values in the same direction for the same principal component (row) are considered correlated. Fig. 13 demonstrates that the problem can be effectively explained using only PC1. Consequently, Model Year, Acceleration, Origin_2, and Origin_3 are positively correlated with each other, whereas Cylinders, Displacement, Horsepower, and Weight are negatively correlated with each other. Components with positive and negative values in opposite directions within the same principal component indicate a negative correlation between these features. For example, focusing on the PC1 axis, Cylinders, Displacement, Horsepower, and Weight were found to be negatively correlated with Model Year, Acceleration, Origin_2, and Origin _3.

In addition, the results in Fig. 11 are almost identical to those in Fig. 12, particularly in terms of the columnar alignment; however, Acceleration also shows a positive correlation with SHAP. This difference can be attributed to the methodologies employed: PCAIME calculates results by examining the interrelationships among features, whereas SHAP, derived from cooperative game theory, may yield divergent outcomes compared to LIME, which approximates a linear model in a local vicinity, and AIME, which employs inverse operators to establish a linear model. Further refinement in AIME, through the addition of decomposition and expansion functions, enables more nuanced analyses that consider the interplay between features. Whether acceleration contributes positively or negatively to MPG can vary depending on specific cases. Theoretically, a car with efficient acceleration usually features an efficient engine and lightweight design, both of which can enhance fuel economy. Urban driving scenarios, which frequently involve stop-and-go conditions, can benefit from smooth acceleration that minimizes fuel waste. Conversely, vehicles engineered for high acceleration might consume more fuel to deliver such performance, particularly during continuous high-speed conditions such as highway driving. Additionally, vehicle designs focused on enhancing acceleration can sometimes compromise fuel efficiency. Thus, integrating domain knowledge is crucial for informed analysis and interpretation of these dynamics.

Unlike in Fig. 12, Model Year, Acceleration, Origin_2, and Origin_3 show positive correlations in Fig. 13. These differences can be attributed to the influence of other features in the case of Fig. 12, suggesting that factors such as a newer model year and better acceleration typically contribute positively to the fuel efficiency of a car.

These findings demonstrate that by deriving a PCA-like function that summarizes the relationships among features as principal components, feature importance can be assessed independently of other feature influences.

VI. DISCUSSION

The experimental findings indicate that PCAIME is an effective method of interpreting the results from the original model (often referred to as a black-box model), while also uncovering the multicollinearity and interrelationships among features. Specifically, the introduction of a decomposition function within PCAIME enables the extraction and analysis of these relationships, providing a deeper understanding of how each feature influences the outcomes of the model. Additionally, the expansion function plays a crucial role in elucidating the behavior of the model and the estimation results for individual data points, further clarifying the complex dynamics among the features.

In this experiment, the global feature importance derived using PCAIME was compared with that determined using an extra tree regressor, and AIME. The contributions of the feature importance, whether negative or positive, could not be clearly determined for the random forest classifier, extra tree classifier, and extra tree regressor. In contrast, AIME explicitly indicates the positive or negative contribution of the derived feature importance to each class. Notably, PCAIME facilitates clear visualization of the correlations among the features and their contributions—either positive or negative—to the model. Additionally, the local feature importance derived using LIME, SHAP, AIME, and PCAIME was compared. The findings demonstrate that PCAIME not only considers, but also clearly visualizes the interrelationships among features. Consequently, contributions to local feature importance may sometimes appear in opposite directions, likely due to the PCAIME calculations that account for the complex relationships among the features. The highlights the utility of PCAIME in providing a nuanced understanding of feature interactions and their impact on model predictions.

Notably, in this study, PCA was incorporated into AIME, which is a type of XAI, to implement a mechanism for deriving and visualizing the multicollinearity and relationships among features and determine feature importances. The modified approach PCAIME, which incorporates reversible decomposition and expansion functions, is a novel method of realizing such a mechanism. Note that this proposed technique is feasible only if reversible decomposition and

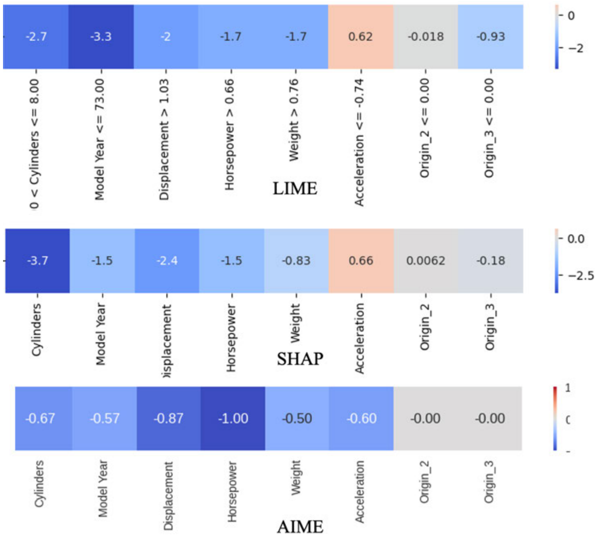

FIGURE 12. Heat map representation of the local feature importance of non-fuel-efficient vehicle data points derived using LIME, SHAP, and AIME (Auto MPG data).

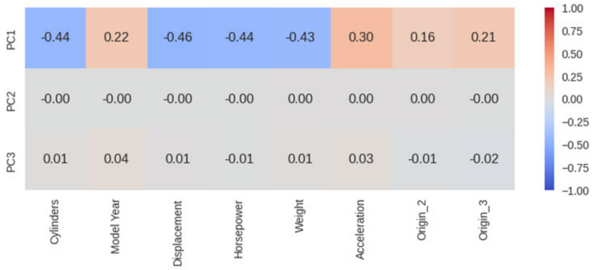

FIGURE 13. Heat map representation of the local feature importance of non-fuel-efficient vehicle data points derived using PCAIME (Auto MPG data).

expansion functions are present. PCAIME can be easily realized via dimensionality reduction performed using PCA and via dimensionality expansion performed based on the relationship between the features and principal components using PCA loadings. In other words, previous studies have demonstrated that PCA is suitable for performing these operations. Complex AI and machine learning models potentially contain functions of dimensionality reduction and feature selection that process multicollinearity and relationships among features while explaining their estimation results. However, explaining the multicollinearity and relationships among features is essential and was investigated in the present study. The present study was primarily focused on the development of an XAI method that implements both reversible decomposition and expansion functions.

However, the proposed PCAIME has certain limitations. Although the reversible decomposition and expansion functions enable visualization of the relationships among features to some extent, using conventional correlation matrices and VIFs is suitable for deriving strict multicollinearity. In other words, PCAIME is mainly aimed at detecting both the existence of a relationship among the features and important features. The correlation matrix and VIF can only determine the relationships between features and multicollinearity. PCAIME provides a higher level of visualization than these systems because the correlation matrix and VIF cannot indicate the importance of a feature.

This method automatically derives the correlations among the features and does not consider causal relationships. Introducing some type of domain knowledge is necessary to verify the causal relationships based on the derived correlations. The construction of such a method will be attempted in a future study.

PCAIME raises questions concerning the use of results obtained via previously developed feature importance methods to recreate AI and machine learning models by simply using only the most important features. This is because PCAIME shows the relationships among features in the form of principal components; therefore, new important features can be created that sometimes serve as components while considering the relationships among features. In other words, considering data with correlations or multicollinearity among features, PCAIME must be used to observe the feature importance while considering the relationships among features. In this study, we examined the relationships between the features and multicollinearity in AIME. However, LIME [22] and SHAP [23] also need to address this issue.

PCAIME can determine the feature contributions by considering the relationships among the features, behavior of the model, and explanation of the estimation results for each data point. In addition, the relationships among the features and their contributions can be visualized using a simple heat map. This process enables the derivation of highly accurate and more interpretable explanations of the results obtained using complex AI and machine learning models for users. Elucidating the original (black-box) model using the various interpretive machine learning and XAI techniques of PCAIME is crucial for the development of next-generation machine learning and AI technology and for ensuring reliability, transparency, responsibility, and accountability. The exploration of methods to integrate and evaluate these numerous interpretive machine learning and XAI approaches may be addressed in future research.

VII. CONCLUSION

In this study, reversible decomposition and expansion functions were applied to previously reported AIME to formulate a visualization method that simultaneously shows the relationship between features and their contributions in a heat map. The relationships were obtained by deriving the behavior of the black-box model and features that contributed to the estimated results when each data point was fed as an input, while considering the multicollinearity and correlation among the features. This approach leverages the decomposition function and approximate inverse operator of AIME to uncover relationships among features, providing an explanation that accounts for these interconnections. Additionally, the expansion function elucidates the relationship between features and their contributions. These insights offer deeper understanding of black-box models, highlighting the superiority of PCAIME over previously reported XAI techniques. These findings suggest that deriving decomposition and expansion functions, which encapsulate feature relationships as principal components, enables the determination of feature importance without impacting other features.

PCAIME implements the decomposition and expansion functions using PCA, where the PCA loadings reflect the relationships between the features and principal components. This study underscores the significance of leveraging these features to derive explanations. PCAIME facilitates the derivation of feature contributions while considering correlations among features, notably in terms of global and local feature importance. Comparative analysis was conducted between the global feature importance derived using PCAIME and that obtained from an extra tree regressor. The results indicate that the extra tree regressor cannot reveal the positive or negative contribution of the feature importance to a class. In contrast, AIME indicates the positive or negative contribution of the features to each class. Notably, PCAIME outperforms all these methods because a simple heat map distinctly reveals the one-to-one correspondence among the features as well as their positive or negative contribution to the model. Similarly, the local feature importance derived using PCAIME was compared with that derived using LIME, SHAP, and AIME, and the results indicate that PCAIME considers as well as reveals the relationships among features.

The proposed method, in which reversible decomposition and expansion functions are applied to AIME to derive the global and local feature importance by considering the multicollinearity and correlation among the features, is one method of solving the problem of multicollinearity and correlation among features indicated in previous studies.

Future studies should focus on conducting experiments using datasets with several explanatory variables to explore and verify the potential advantages of PCAIME. In particular, the decomposition feature should be implemented to realize dimensionality reduction while balancing the computational cost. In addition, PCAIME should be applied in scenarios wherein causal relationships exist between features, as well as in real-world problems.

PCAIME is an innovative approach that unravels complex AI and machine learning-based black-box models and paves the way for the development of transparent and reliable decision-making processes for feature contributions, including multicollinearity and relationships among features. Its unique approach to deriving feature relationships and importance through reversible decomposition and expansion functions has the potential to redefine our understanding of and trust in complex algorithms. PCAIME provides a comprehensive framework that significantly advances the field of machine learning interpretability. Given the growing demand for transparent AI and machine learning models in various critical areas such as healthcare, finance, and public policy, PCAIME will play a vital role in bridging the divide between complex algorithms and datasets and human understanding.

APPENDIX A

This Appendix presents the results of the experiments conducted on the wine data to determine whether PCAIME is also valid for other datasets.

A. ADDITIONAL EXPERIMENTAL ENVIRONMENT

The additional experiment utilized the wine quality dataset [59] to evaluate the framework. The red wine data consisted of 1599 data points, 11 features, and quality score labels ranging from 3 to 8, whereas the white wine data consisted of 4898 data points, 11 features, and quality scores ranging from 3 to 9. These data were divided into training and test datasets. The size of X of the red wine data was 1199×12 and the size of Y was 1199×6, resulting in $A^{\dagger}$ with dimensions of 12×6. The size of X of the white wine data was 3673×12 and the size of Y was 3673×7, resulting in $A^{\dagger}$ with dimensions of 12×7.

B. EXPERIMENT 1: APPLICATION OF PCAIME TO WINE QUALITY DATASET

In an additional experiment, the red and white wine quality datasets were employed to derive and compare explanations using the AIME and PCAIME methods. Both datasets encompassed several parameters including Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulfur Dioxide, Sulfates, and Alcohol. A machine learning model was constructed to predict the quality from these features, operating as a black-box model. The quality of red wine was represented by integral values in the range of 3–8 and 3–9, respectively.

1) FEATURE IMPORTANCE FOR RED WINE QUALITY DATASET

Fig. 14 presents the heat-mapped correlation matrix for each feature, illustrating the degrees of correlation, where red indicates stronger positive correlations and blue signifies stronger negative correlations. The matrix highlights several key relationships: Fixed Acidity and Citric Acid, as well as Fixed Acidity and Density, are highly positively correlated. Free Sulfur Dioxide and Total Sulfur Dioxide also display a strong positive correlation. Conversely, a strong negative correlation is observed between Fixed Acidity and pH. The correlation between Fixed Acidity and pH is negative and significant. Table 2 presents the results of the VIF.

TABLE 2. VIFs for red wine quality data.

Feature	VIF
Const.	1.000000
Fixed Acidity	7.767512
Volatile Acidity	1.789390
Citric Acid	3.128022
Residual Sugar	1.702588
Chlorides	1.481932
Free Sulfur Dioxide	1.963019
Total Sulfur Dioxide	2.186813
Density	6.343760
pH	3.329732
Sulfates	1.429434
Alcohol	3.031160

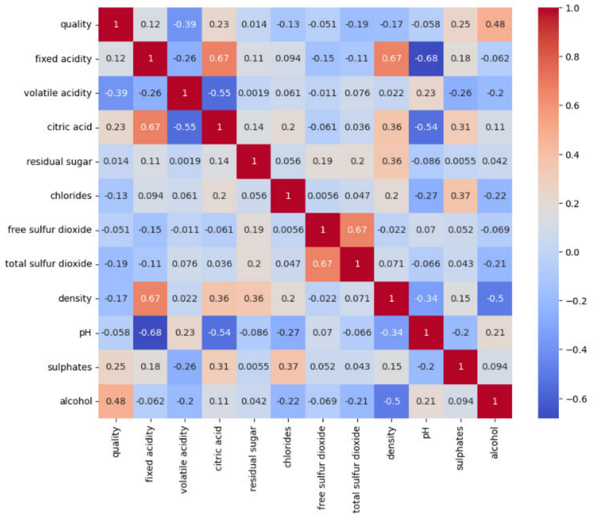

FIGURE 14. Heat map representation of the correlation matrix of the red wine quality data.

Features with VIF values exceeding 10 do not exist; however, the VIF values for Fixed Acidity and Density, that is, 7.768 and 6.3438, respectively, are high. These results indicate that no sufficiently correlated feature group in the red wine quality dataset can be considered multicollinear. However, correlated features are present.

As indicated in Section IV-A, an extra tree classifier was used as the black-box model for these data. The corresponding results were highly accurate, with area under the curve (AUC), recall, precision, and F1 score of 0.6925, 0.857602, 0.661672, and 0.67519, respectively.

The feature importance results obtained using the extra tree classifier are shown in Fig. 15. As the values are not very large (approximately 0.1), the important features cannot be identified.

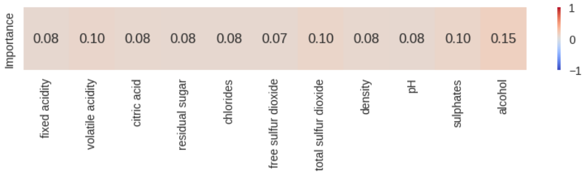

FIGURE 15. Heat map representation of feature importance derived using the extra tree classifier, for the red-wine quality dataset.

The AIME-derived heat map of the global feature importance is presented in Fig. 16. Quality 3 and 4 exhibit positive contributions of Volatile Acidity, whereas Quality 5 shows a positive contribution of Volatile Acidity and Total Sulfur Dioxide but a negative contribution of Alcohol. For Quality 6, 7, and 8, Sulfates and Alcohols are positively correlated. Citric Acid is particularly high for Quality 7 and 8. Thus, the global feature importance derived using AIME revealed the contributions of the features in each class of the black-box model.

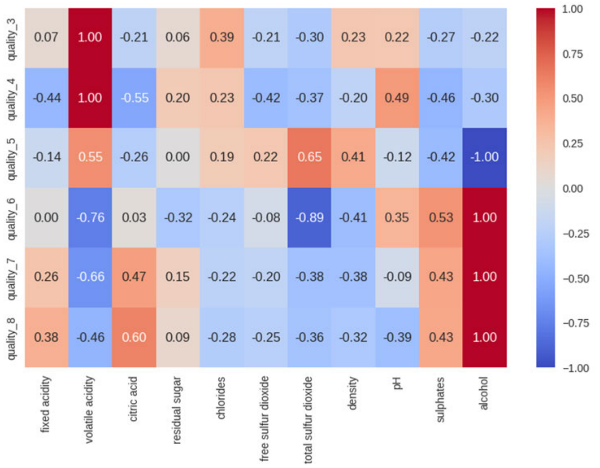

FIGURE 16. Heat map representation of the global feature importancederived using AIME, for the red-wine quality dataset.

Because the proposed PCAIME adopts PCA, determining the number of principal components to be used is crucial. The cumulative explanation variance and scree plots for the red wine quality dataset are shown in Figs. 17 and 18, respectively. The cumulative explanation variance is approximately 0.85 when the number of principal components is 6; this result implies that the contributions of the components contain 85% of the original data. The scree plot in Fig. 8 shows the screes for PC2 and PC6 (herein, PC followed by a number represents principal components). Based on this result, PCAIME is obtained for PC6.

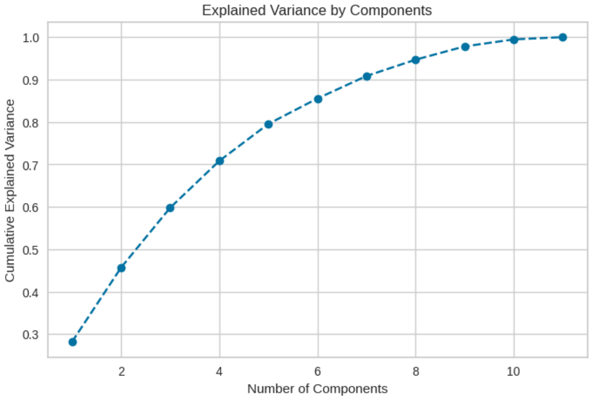

FIGURE 17. Cumulative explanation variance for the red wine quality dataset.

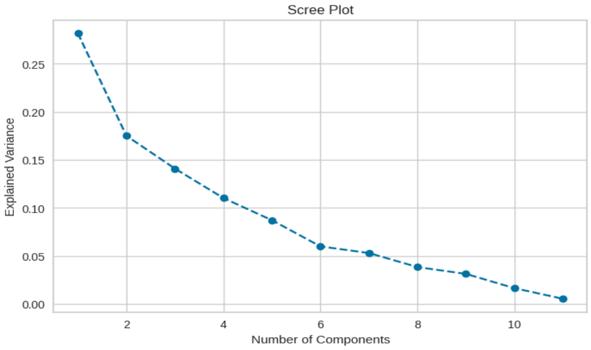

FIGURE 18. Scree plot for the red wine quality dataset.

Negative correlations between quality values and specific features are also observed. Quality 3 is negatively correlated with Citric Acid (−0.21), Free Sulfur Dioxide, Total Sulfur Dioxide (−0.30), Sulfates (−0.27), and Alcohol (−0.22). Thus, these features are negatively associated with Quality 3 wines. For Quality 4, negative correlations are observed with Fixed Acidity (−0.44), Citric Acid (−0.55), Free Sulfur Dioxide (−0.42), Total Sulfur Dioxide (−0.37), Sulfates (−0.46), and Alcohol (−0.30), suggesting that higher levels of these features are not conducive to Quality 4. Quality 5 is negatively correlated with Fixed Acidity (−0.14), Citric Acid (−0.24), pH (−0.12), Sulfates (−0.13), and Alcohol (−1.00), indicating that these features negatively affect this quality level. For Quality 6, negative correlations are observed with Fixed Acidity (−0.14), Volatile Acidity (−0.26), pH (−0.12), Sulfates (−0.42), and Alcohol (−1.00), highlighting its significant negative impact. Quality 7 is negatively correlated with Volatile Acidity (−0.66), Chloride (−0.22), Free Sulfur Dioxide (−0.20), Total Sulfur Dioxide (−0.89), and Density (−0.41), implying that these features reduce wine quality. Quality 8 is negatively correlated with Volatile Acidity

(−0.66), Chlorides (−0.22), Free Sulfur Dioxide (−0.22), Total Sulfur Dioxide (−0.38), Density (−0.64), and pH (−0.09), indicating that these features negatively affect the quality. Finally, Quality 9 is negatively correlated with Volatile Acidity (−0.46), Chlorides (−0.28), Free Sulfur Dioxide (−0.25), Total Sulfur Dioxide (−0.36), Density (−0.32), and pH (−0.39), suggesting that these elements negatively contribute to the quality.

The results for PCAIME Quality 3, 4, 5, 6, 7, and 8 are shown in Figs. 19, 20, 21, 22, 23, and 24, respectively. The PCAIME is represented using a heat map, wherein the principal components and features are displayed on the vertical and horizontal axes, respectively. The rows with the same principal components and similar values are highly correlated. The columns indicate features with strong contributions in the corresponding class.

Fig. 19 shows the PCAIME results for Quality 3. The rows indicate that PC3 and PC5 are effective for representation of feature importance.

In particular, PC3 shows a positive correlation between Volatile Acidity and Density and negative correlations among Free Sulfur Dioxide, Total Sulfur Dioxide, and Alcohol. The correlations among these features are attributed to their same principal components and values in the same direction. In the columns, Volatile Acidity, Residual Sugar, and Density contribute positively, whereas Free Sulfur Dioxide, Total Sulfur Dioxide, and Alcohol contribute negatively. The contribution of Chlorides is lower than that derived using AIME, whereas that of Residual Sugars is higher than that derived using AIME. This result can be ascribed to the ability of PCAIME to organize correlations with other features using principal components. Components with positive and negative values in opposite directions within the same principal component indicate a negative correlation between these features. For example, Free Sulfur Dioxide and Total Sulfur Dioxide have negative correlations with Alcohol and PC2.

Fig. 20 displays the PCAIME results for Quality 4. These results are consistent with those observed for Quality 3, shown in Fig. 19, particularly in the effectiveness of PC3 and PC5 when analyzed row by row. Specifically, PC3 reveals a positive correlation between Volatile Acidity and Density and a negative correlation among Free Sulfur Dioxide, Total Sulfur Dioxide, and Alcohol. The consistency of these principal components and the direction of their values in both quality levels underline their correlated nature. In the column-wise analysis, Volatile Acidity, Residual Sugar, and Density show positive contributions, whereas Free Sulfur Dioxide, Total Sulfur Dioxide, and Alcohol demonstrate negative contributions. Notably, the contribution of residual sugar is more pronounced in PCAIME than in AIME. This difference is attributed to the ability of PCAIME to organize correlations with other features through its use of principal components, providing a clearer delineation of how each feature influences wine quality. Citric Acid, Residual Sugar, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, and Alcohol are negatively correlated with Volatile Acidity and Density in PC3.

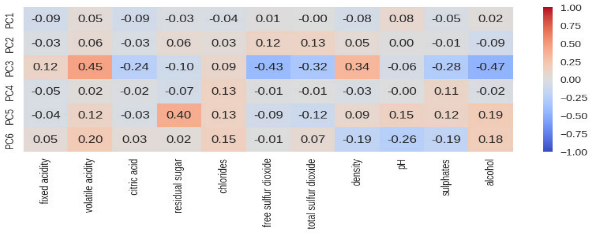

FIGURE 19. Heat map representation of the global feature importance of Quality 3, derived using PCAIME, for the red-wine quality data.

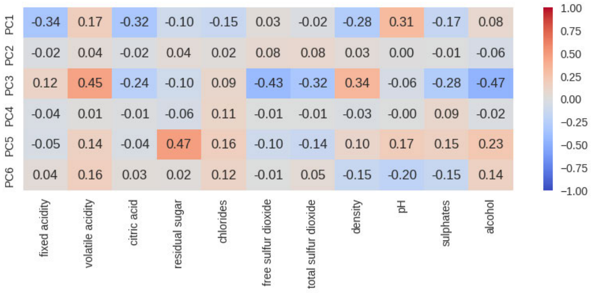

FIGURE 20. Heat map representation of the global feature importance of Quality 4 derived using PCAIME for the red wine quality dataset.

Fig. 21 presents the PCAIME results for Quality 5. In contrast to the Quality 3 and 4 results shown in Figs. 8 and 9, respectively, the results for Quality 5 show a strong influence of PC2.

In particular, Free Sulfur Dioxide and Total Sulfur Dioxide are strongly affected in the same direction, and a strong correlation is observed between these two features. In the columns, Volatile Acidity, Free Sulfur Dioxide, Total Sulfur Dioxide, and Density show positive contributions, whereas Alcohol shows a negative contribution. This trend is similar to that obtained for the Quality 5 row in Fig. 16. However, the negative contribution of Sulfate for Quality 6 is not the same as that obtained for Quality 5 (Fig. 16). This result may be attributed to the PCA processing of the relationships among the features. Thus, Fig. 21 derives feature importance that reflects relationships and multicollinearity among features. Alcohol is negatively correlated with Free Sulfur Dioxide and Total Sulfur Dioxide in PC3.

Fig. 22 shows the PCAIME results for Quality 6, exhibiting strong contributions from PC2, PC3, and PC6, in contrast to those observed in Figs. 19–21.

In particular, the negative contributions of Free Sulfur Dioxide and Total Sulfur Dioxide in PC2 indicate a strong relationship between these two features. In the column analysis, the negative contributions of Free Sulfur Dioxide and Total Sulfur Dioxide outweigh the positive contributions of pH and Alcohol. Additionally, Density also exhibits a negative contribution when included in the column assessments. This trend mirrors the observations in the Quality 6 analysis

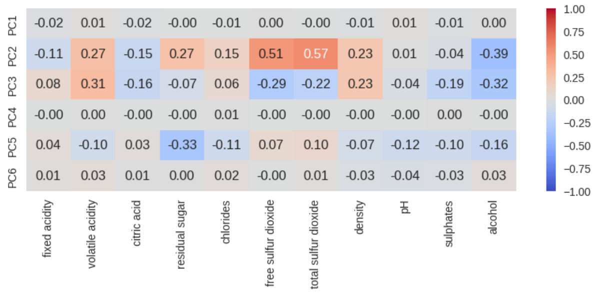

FIGURE 21. Heat map representation of the global feature importance of Quality 5, derived using PCAIME, for the red-wine quality data.

as presented in Fig. 6, underscoring consistent patterns across different quality levels in the dataset. Alcohol is negatively correlated with Free Sulfur Dioxide and Total Sulfur Dioxide in PC2.

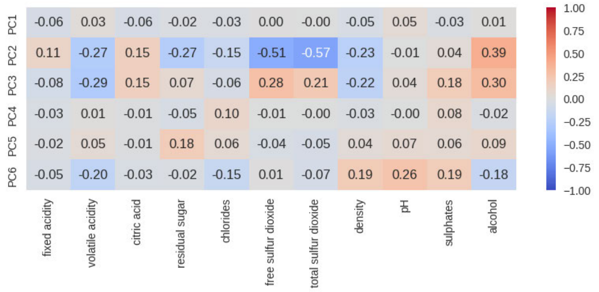

FIGURE 22. Heat map representation of the global feature importance of Quality 6 derived using PCAIME for the red wine quality data.

Fig. 23 depicts the PCAIME results for Quality 7. Notably, PC2 and PC3 of Free Sulfur Dioxide and Total Sulfur Dioxide show positive and negative values that offset each other. This result indicates a strong correlation between these two features. Further, the values of Fixed Acidity and Sulfates are lower than those obtained for Quality 7. Alcohol is strongly negatively correlated with Free Sulfur Dioxide and Total Sulfur Dioxide in PC2.

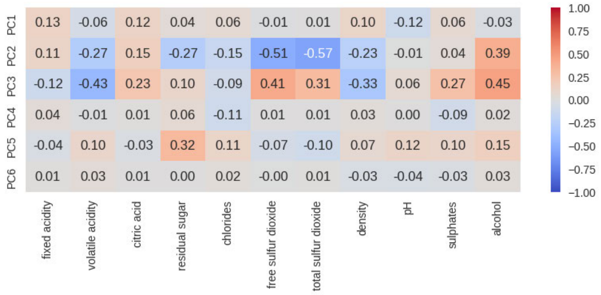

FIGURE 23. Heat map representation of the global feature importance of Quality 7, derived using PCAIME, for the red-wine quality data.

Fig. 24 presents the PCAIME results for Quality 8, which are similar to those observed for Quality 7 in Fig. 13. The analysis highlights that the positive and negative values in PC2 and PC3 for Free Sulfur Dioxide and Total Sulfur Dioxide appear to offset each other, suggesting a strong correlation between these two features. Moreover, when compared to the results for Quality 8 depicted in Fig. 6, the values for Fixed Acidity and Sulfates are relatively low, indicating a distinct pattern in how these components impact wine with higher quality ratings. Free Sulfur Dioxide and Total Sulfur Dioxide consumption are strongly negatively correlated with Alcohol and Volatile Acidity in PC2. Furthermore, Volatile Acidity consumption is strongly negatively correlated with Free Sulfur Dioxide and Total Sulfur Dioxide in PC3.

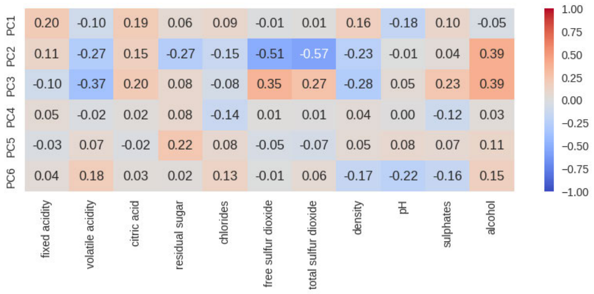

FIGURE 24. Heat map representation of the global feature importance of Quality 8, derived using PCAIME, for the red-wine quality data.

These results suggest that PCAIME reveals the correlations among the features and derives their contributions by considering the relationships among the features.

2) FEATURE IMPORTANCE FOR WHITE WINE QUALITY DATASET

Fig. 25 shows the heat-mapped correlation matrix for each feature. Residual Sugar and Density, Free Sulfur Dioxide, and Total Sulfur Dioxide are highly positively correlated.

Table 3 presents the VIF results. The VIF values for Residual Sugar and Density exceed 10, indicating multicollinearity of the data. The large VIF value of Alcohol (7.7070) indicates its strong relationships with the other characteristics.

As indicated in Section IV-A, a random forest classifier was used as the black-box model for these data; the corresponding accuracy, AUC, recall, precision, and F1 score are 0.659592, 0.842657, 0.659592, 0.663018, and 0.649181, respectively.

The results obtained using the random forest classifier (shown in Fig. 25), which can derive feature importance, were compared. Because the values are low (approximately 0.1), the important features cannot be identified.

Fig. 26 shows a feature-importance heat map, developed using a random forest classifier, for the white-wine quality dataset. In addition, a heat map of the global feature importance results derived using AIME is presented in Fig. 27.

In the case of AIME, Quality 3 shows positive contributions from Fixed Acidity and Free Sulfur Dioxide. In particular, in the case of Quality 3, Free Sulfur Dioxide shows a very large contribution of 1.0. Moreover, Free

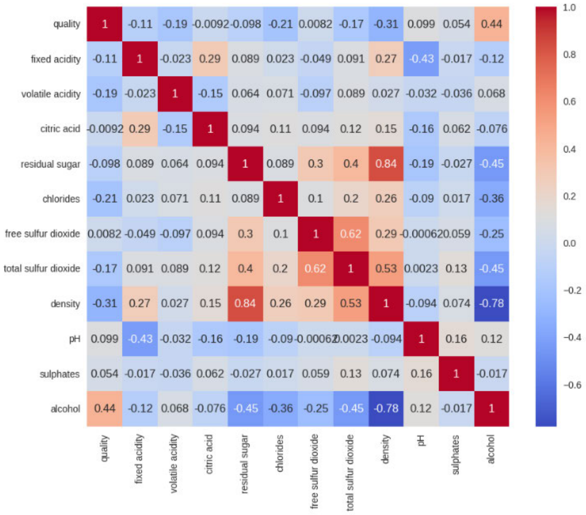

FIGURE 25. Heat map representation of the correlation matrix of the white-wine quality data.

TABLE 3. VIF for white-wine quality Data.

Feature	VIF
Const	1.000000
Fixed Acidity	2.691435
Volatile Acidity	1.141156
Citric Acid	1.165215
Residual Sugar	12.644064
Chlorides	1.236822
Free Sulfur Dioxide	1.787880
Total Sulfur Dioxide	2.239233
Density	28.232546
pH	2.196362
Sulphates	1.138540
Alcohol	7.706957

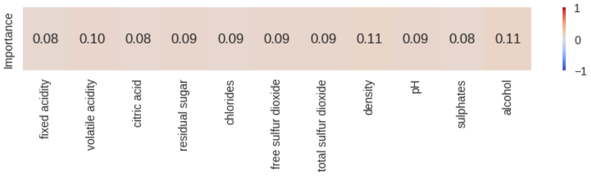

FIGURE 26. Heat map representation of the feature importance, derived using a random forest classifier, for the white-wine quality dataset.

Sulfur Dioxide is often high in low quality wines. Quality 4 exhibits a positive contribution from Volatile Acidity and negative contributions from Residual Sugar, Free Sulfur Dioxide, and Total Sulfur Dioxide. Quality 5 indicates positive contributions from Volatile Acidity, Residual Sugar, Chlorides, Total Sulfur Dioxide, and Density; in contrast, a negative contribution is obtained from Alcohol. Quality 6 indicates a negative contribution from Volatile Acidity. Quality 7 indicates positive contributions from pH and Alcohol, along with negative contributions from Residual Sugar, Chlorides, Total Sulfur Dioxide, and Density. Quality 8 indicates positive contributions from pH and Alcohol and negative contributions from Fixed Acidity, Chlorides, Total Sulfur Dioxide, and Density. Quality 9 indicates positive contributions from Fixed Acidity, Citric Acid, pH, and Alcohol, with negative contributions from Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, and Density.

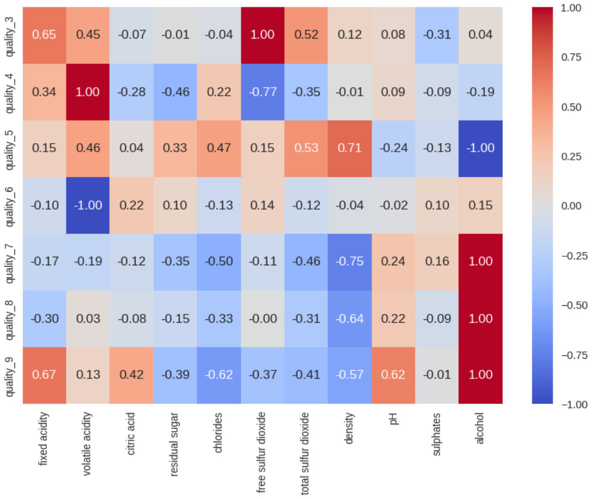

FIGURE 27. Heat map representation of the global feature importance derived using AIME for the white wine quality data.

Negative correlations between quality values and specific features are also observed. For Quality 3, a negative correlation exists with Sulfates (−0.31), which indicates that lower alcohol content is associated with lower-quality wines. For Quality 4, negative correlations are observed with Citric Acid (−0.28), Residual Sugar (−0.46), Free Sulfur Dioxide (−0.77), and Total Sulfur Dioxide (−0.35), suggesting that higher levels of these features are not conducive to higher quality. Quality 5 shows negative correlations with pH (−0.24) and Sulfates (−0.13), indicating that alcohol content negatively affects this quality level. For Quality 6, negative correlations are observed with Volatile Acidity (−1.00), highlighting its significant negative impact. Quality 7 is negatively correlated with Fixed Acidity (−0.17), Volatile Acidity (−0.19), Citric Acid (−0.12), Residual Sugar (−0.35), Chloride (−0.50), Free Sulfur Dioxide (−0.11), Total Sulfur Dioxide (−0.46), and Density (−0.75), implying that these features reduce the quality of wine. Quality 8 is negatively correlated with Fixed Acidity (−0.33), Citric Acid (–0.07), Residual Sugar (−0.15), Chlorides (−0.33), Total Sulfur Dioxide (−0.31), and Density (−0.64), indicating that these features negatively affect quality. Finally, Quality 9 is negatively correlated with Residual Sugar (−0.39), Chlorides (−0.62), Free Sulfur Dioxide (−0.37), Total Sulfur Dioxide (−0.41), and Density (−0.57), suggesting that these elements negatively contribute to each quality.

The AIME results indicate the presence of complex relationships among the features, except that high alcohol consumption results in high quality. The cumulative explanation variance and scree plot for the white-wine quality dataset are shown in Figs. 28 and 29, respectively. The cumulative explanation variance shown in Fig. 18 is greater than 0.80 when the number of principal components is 6; this result indicates that the contribution of PCs contains over 80% of the original data information. The scree plot in Fig. 19 shows the screes for PC2 and PC6, and these results were used to derive the PCAIME for PC6.

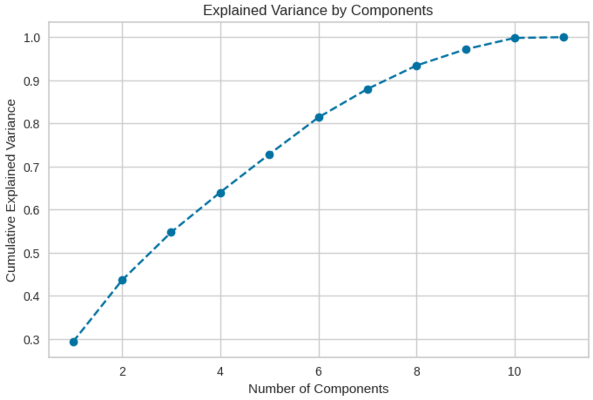

FIGURE 28. Cumulative explanation variance for the white wine quality dataset.

The PCAIME results for Quality 3, 4, 5, 6, 7, 8, and 9 are shown in Figs. 30, 31, 32, 33, 34, 35, and 36, respectively.

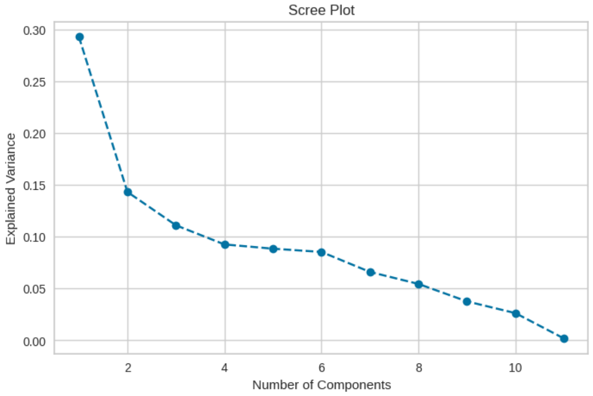

FIGURE 29. Cumulative explanation variance for the white wine quality dataset.

Fig. 30 presents the PCAIME results for Quality 3, indicating that PC5 and PC6 are valid.

Specifically, the PC6 values of Chlorides and Free Sulfur Dioxide exhibit a positive correlation, indicating that these features share the same principal component number and values along the same direction. Conversely, these factors and Sulfates demonstrate negative correlations. Within the column, Volatile Acidity, Free Sulfur Dioxide, and Total Sulfur Dioxide display strong positive correlations. Furthermore, the PC4 and PC6 values of Sulfur Dioxide exhibit a strong positive correlation, whereas the PC5 values demonstrate a negative correlation, suggesting a cancellation effect. Chlorides showcase large positive PC6 values and substantial negative PC4 and PC5 values, indicating a cancellation effect as well. Additionally, compared to the Quality 3 row in Fig. 27, the Fixed Acidity is low due to PCAIME being the main component. This result may be attributed to the use of principal components by PCAIME to determine the correlations with other features. Volatile Acidity has a strongly negative correlation with Chlorides in PC5, and Chlorides, Free Sulfur Dioxide, and Total Sulfur Dioxide have negative correlations with Sulfates, Density, and Residual Sugar in PC6.

Fig. 31 illustrates the PCAIME results for Quality 4, revealing the validity of PC3 and PC4. Additionally, Citric Acid and Sulfates are correlated with the same principal component number, PC3, with values along the same negative direction. Meanwhile, Chlorides and Sulfates share the same principal component numbers and are correlated along the same positive direction. Moreover, Sulfates demonstrate a strong positive correlation in PC4; however, they exhibit a negative correlation in PC3, indicating a cancellation effect. In terms of the columns, Volatile Acidity and Chlorides display positive contributions, whereas Citric Acid and Free Sulfur Dioxide show negative contributions. In addition, compared with the Quality 4 row in Fig. 27, the value of Fixed Acidity is small, and Residual Sugar and Total Sulfur Dioxide do not exhibit large negative contributions, possibly because PCAIME uses principal components to determine the correlations with other features. Citric Acid, Free Sulfur Dioxide, and Sulfates have strong negative correlations with Volatile Acidity in PC3, and Residual Sugar and Free Sulfur Dioxide have negative correlations with Chlorides and Sulphates in PC4.

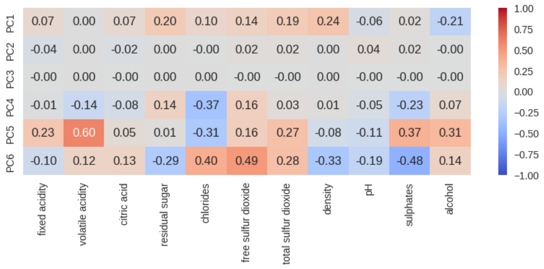

FIGURE 30. Heat map representation of the global feature importance of Quality 3, derived using PCAIME, for the white-wine quality data.

Fig. 32 displays the PCAIME results for Quality 5, highlighting the validity of PC1, PC3, and PC4. Total Sulfur Dioxide and Density are correlated with the same principal components, with their values along the same positive direction. In PC4, Chlorides and Sulfates are correlated due to sharing the same principal components and their values along the same positive direction. Within the columns, Volatile Acidity, Residual Sugar, Chlorides, Total Sulfur Dioxide, and Density demonstrate positive contributions, while alcohol exhibits a negative contribution. Comparing these contributions with those observed in the row for Quality 5 in Fig. 27, similar trends are evident. Alcohol and pH

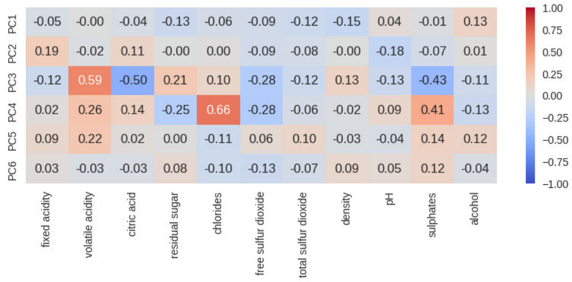

FIGURE 31. Heat map representation of the global feature importance of Quality 4, derived using PCAIME, for the white-wine quality data.

have negative correlations with Residual Sugar, Chorides, Free Sulfur Dioxide, Total Sulfur Dioxide, and Density in PC1, and Citric Acid, Free Sulfur Dioxide, and Sulfates have negative correlations with Volatile Acidity in PC3.

Fig. 33 shows the PCAIME results for Quality 6, confirming the significant contributions of PC3, PC4, and PC5. In PC4, Chlorides and Sulfates are correlated with the same principal component and the values are in the same negative direction. In PC5, Volatile Acidity, Sulfates, Alcohol, and Total Sulfur Dioxide are correlated in the same negative direction. Sulfates are a positive contributor of PC3, whereas PC4 and PC5 cancel each other out. Volatile Acidity exhibits a negative contribution in the columns. Moreover, the trend is similar to that of the Quality 6 row in Fig. 27. Volatile Acidity, Residual Sugar, Chlorides, and Density have negative correlations with Citric Acid in PC3, and Fixed Acidity, Volatile Acidity, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulfates, and Alcohol have negative correlations with Chlorides in PC5.

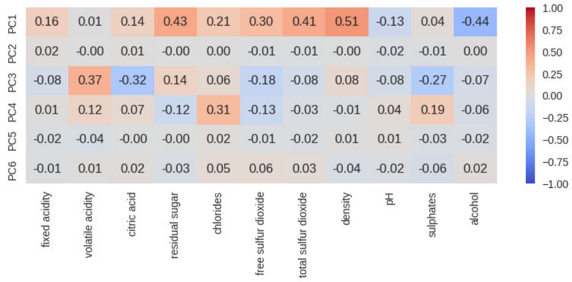

FIGURE 32. Heat map representation of the global feature importance of Quality 5, derived using PCAIME, for the white-wine quality data.

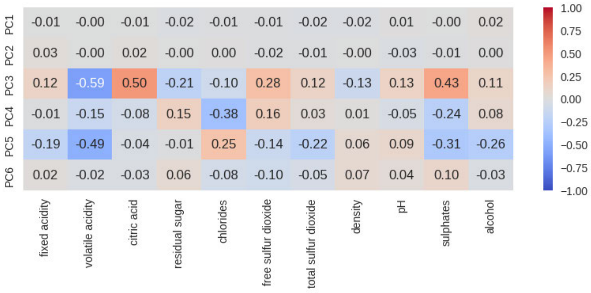

FIGURE 33. Heat map representation of the global feature importance of Quality 6, derived using PCAIME, for the white-wine quality data.

Fig. 34 shows the PCAIME results for Quality 7, indicating that PC1 is valid. For PC1, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, and density are correlated with the same principal components and exhibit the same negative values. Conversely, these variables are negatively correlated with alcohol consumption. In the columns, Residual Sugar, Chlorides, Total Sulfur Dioxide, and Density exhibit negative contributions, whereas Alcohol exhibits a positive contribution. Furthermore, compared to the Quality 7 row in Figs. 27 and 34, the trends are similar in this case. Fixed Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, and Density have negative correlations with Alcohol in PC1.

Fig. 35 presents the PCAIME results for Quality 8, confirming the validity of PC1. Additionally, for Quality 7, Total Sulfur Dioxide is correlated with Density. Conversely, these variables demonstrate a negative correlation with Alcohol. Within the columns, Residual Sugar, Chlorides, Total Sulfur Dioxide, and Density display negative contributions, whereas Alcohol exhibits a positive contribution. Moreover, comparing these trends with those observed in the Quality 8 row in Figs. 27 and 35, similarities are apparent. Fixed Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, and Density have negative correlations with Alcohol in PC1.

Fig. 36 shows the results of PCAIME for Quality 9, showing that PC1, PC3, and PC5 are valid. In PC5, Volatile Acidity, Sulfates, and Alcohol exhibit the same principal components, and their values are in the same positive direction. In the columns, the volatile acidities cancel each other out in PC3 and PC5. Furthermore, compared with the row of Quality 9 in Fig. 27, the contributions of Fixed Acidity and pH are low, whereas that of Sulfates is high. This result may be attributed to the use of principal components by PCAIME to determine correlations with other features. Fixed Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, and Density are negatively correlated with Alcohol in PC1.

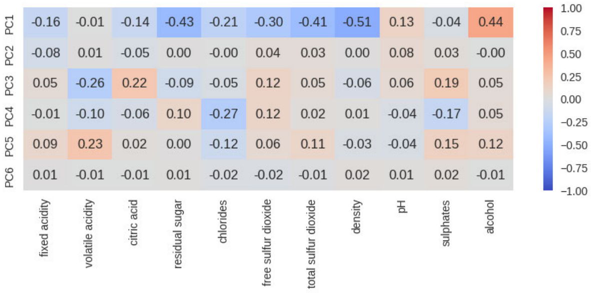

FIGURE 34. Heat map representation of the global feature importance of Quality 7, derived using PCAIME, for the white-wine quality data.

These results indicate that PCAIME can represent the relationship between features in the line direction and identify

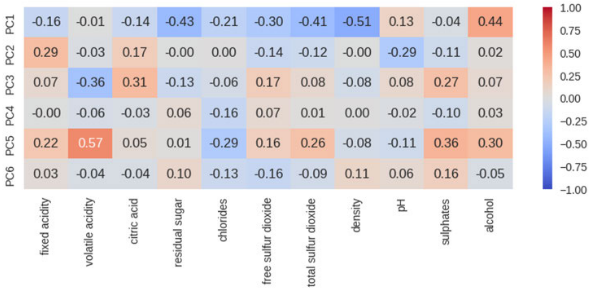

FIGURE 35. Heat map representation of the global feature importance of Quality 8, derived using PCAIME, for the white-wine quality data.

the features contributing to that class along the column direction. Certain results differ from those of AIME; however, these differences may exist because the PCAIME decomposition resulted in a better reflection of the multicollinearity and relationships among features, because PCA reduces the dimensionality and resolves the multicollinearity and relationships between features in the first place.

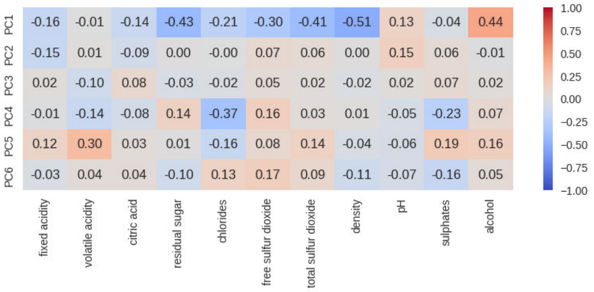

FIGURE 36. Heat map representation of the global feature importance of Quality 9, derived using PCAIME, for the white-wine quality data.

ACKNOWLEDGMENT

The author would like to thank Ayako Minematsu, a Researcher with the Musashino University Asia AI Institute, for her invaluable comments and insights, and Editage (www.editage.com) for English language editing and also would like to thank the role of OpenAI's ChatGPT AI system in facilitating the discussion and inspiring innovative ideas during the writing process. Although no AI-generated text is directly used in this article, ChatGPT was utilized to conceptualize and refine the content.

REFERENCES

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

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

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

[4] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts," *Data Mining Knowl. Discovery*, pp. 1–59, Jan. 2023, doi: 10.1007/s10618-022-00867-8. [Online]. Available: https://link.springer.com/article/10.1007/S10618-022-00867-8

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

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

[7] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805, doi: 10.1016/j.inffus.2023.101805.

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

[9] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.

[10] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu Com, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[11] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.

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

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

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

[15] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/electronics8080832.

[16] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

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

[18] W. Samek and K. R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Berlin, Germany: Springer, Sep. 2019, pp. 5–22, doi: 10.1007/978-3-030-28954-6_1.

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

[20] D. Gunningand and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1145/3301275.3308446.

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

[22] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, San Francisco, CA, USA, Jun. 2016, pp. 97–101, doi: 10.18653/v1/n16-3020.

[23] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[24] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometric Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.

[25] S. M. Holland, "Principal components analysis (PCA)," Dept. Geol., Univ. Georgia, Athens, GA, USA, Tech. Rep., 2008, p. 2501, vol. 30602. [Online]. Available: http://stratigrafia.org/8370/handouts/pcaTutorial.pdf

[26] 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)*, Venice, Italy, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[27] 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*.

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

[29] 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.

[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.

[31] 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.

[32] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *J. Bus. Econ. Statist.*, vol. 39, no. 1, pp. 272–281, Jul. 2019, doi: 10.1080/07350015.2019.1624293.

[33] 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, Mar. 2015, doi: 10.1080/10618600.2014.907095.

[34] 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*.

[35] 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.*, Hangzhou, China, Dec. 2013, pp. 656–660, doi: 10.1109/BMEI.2013.6747021.

[36] (2019). *LOFO Importance*. [Online]. Available: https://github.com/aerdem4/lofo-importance

[37] S. M. Shankaranarayana and D. Runje, "ALIME: Autoencoder based approach for local interpretability," in *Proc. 20th Int. Conf. Intell. Data Eng. Automated Learn. (IDEAL)*, vol. 20, Manchester, U.K., 2019, pp. 454–463.

[38] 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*.

[39] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME explanations for diagnostic computer algorithms," 2020, *arXiv:2006.05714*.

[40] 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)*, Bled, Slovenia, 2019, pp. 53–68.

[41] 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.

[42] 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.*, Rennes, France, 2022, pp. 102–114.

[43] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, vol. 32, no. 1, 2018, pp. 1–9, doi: 10.1609/aaai.v32i1.11491.

[44] L. Utkin and A. Konstantinov, "Ensembles of random SHAPs," *Algorithms*, vol. 15, no. 11, p. 431, Nov. 2022, doi: 10.3390/a15110431.

[45] 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.

[46] C. W. Ayad, T. Bonnier, B. Bosch, and J. Read, "Shapley chains: Extending Shapley values to classifier chains," in *Proc. Int. Conf. Discovery Sci.*, Sofia, Bulgaria, 2022, pp. 541–555.

[47] Z. Dikopoulou, S. Moustakidis, and P. Karlsson, "GLIME: A new graphical methodology for interpretable model-agnostic explanations," 2021, *arXiv:2107.09927*.

[48] S. Lo and Y. Yin, "A novel interaction-based methodology towards explainable AI with better understanding of pneumonia chest X-ray images," *Discov. Artif. Intell.*, vol. 1, no. 1, p. 16, 2021, doi: 10.1007/s44163-021-00015-z.

[49] S. S. Chaudhury, P. Sadhukhan, and K. Sengupta, "Explainable AI using the Wasserstein distance," *IEEE Access*, vol. 12, pp. 18087–18102, 2024, doi: 10.1109/ACCESS.2024.3360484.

[50] D. Gaspar, P. Silva, and C. Silva, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024, doi: 10.1109/ACCESS.2024.3368377.

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

[52] A. Palkar, C. C. Dias, K. Chadaga, and N. Sampathila, "Empowering glioma prognosis with transparent machine learning and interpretative insights using explainable AI," *IEEE Access*, vol. 12, pp. 31697–31718, 2024, doi: 10.1109/ACCESS.2024.3370238.

[53] O. Dorabiala, A. Y. Aravkin, and J. N. Kutz, "Ensemble principal component analysis," *IEEE Access*, vol. 12, pp. 6663–6671, 2024, doi: 10.1109/ACCESS.2024.3350984.

[54] E. H. Moore, "On the reciprocal of the general algebraic matrix," *Bull. Amer. Math. Soc.*, vol. 26, pp. 294–300, Jan. 1920.

[55] 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.

[56] M. Marcus and N. A. Khan, "A note on the Hadamard product," *Can. Math. Bull.*, vol. 2, no. 2, pp. 81–83, May 1959, doi: 10.4153/cmb-1959-012-2.

[57] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, "Digital medicine and the curse of dimensionality," *Npj Digit. Med.*, vol. 4, no. 1, p. 153, Oct. 2021, doi: 10.1038/s41746-021-00521-5.

[58] R. Quinlan, "Auto MPG," UCI Mach. Learn. Repository, 1993, doi: 10.24432/C5859H. [Online]. Available: https://archive.ics.uci.edu/dataset/9/auto+mpg

[59] P. Cortez et al., "Wine quality," UCI Mach. Learn. Repository, 2009, doi: 10.24432/C56S3T. [Online]. Available: https://archive.ics.uci.edu/dataset/186/wine+quality

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