



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/243381/>

Version: Accepted Version

Proceedings Paper:

Liu, L., Bogachev, L., Rezaei, M. et al. (2026) Identifying OM4AnI's Effectiveness in the Context of Explainable AI. In: 2026 IEEE 19th Pacific Visualization Conference (PacificVis). 2026 IEEE 19th Pacific Visualization Conference (PacificVis), 20-23 Apr 2026, Sydney, Australia. Institute of Electrical and Electronics Engineers (IEEE), pp. 147-152. ISSN: 2165-8765. EISSN: 2165-8773.

<https://doi.org/10.1109/pacificvis68791.2026.00022>

This is an author produced version of a conference paper published in the Proceedings of 2026 IEEE 19th Pacific Visualization Conference (PacificVis), made available via the University of Leeds Research Outputs Policy under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Identifying OM4AnI’s Effectiveness in the Context of Explainable AI

Liqun Liu*
University of Leeds

Leonid Bogachev†
University of Leeds

Mahdi Rezaei‡
University of Leeds

Nishant Ravikumar§
University of Leeds

Arjun Khara¶
University of Leeds

Mohsen Azarmi||
University of Leeds

Roy A. Ruddle**
University of Leeds

ABSTRACT

Scatterplots are widely used in Explainable Artificial Intelligence (XAI) to investigate misclassifications and patterns across instances. However, a significant limitation of scatterplots is overplotting, especially when working with large datasets. Although several quality metrics have been proposed to measure the degree of overplotting, none have been demonstrated to be effective in the context of XAI. This paper aims to evaluate the effectiveness of a quality metric, called OM4AnI, in XAI scenarios. We begin by summarizing two visual patterns—cluster-based and regression-based patterns—that support three common XAI tasks: feature importance, feature dependency, and model accuracy. We also introduce how to select the parameters of OM4AnI based on these patterns. We construct two case studies to identify the effectiveness of OM4AnI using public datasets: Census Income dataset and MNIST dataset. OM4AnI is applied to both scenarios under various visual conditions (e.g., marker size and rendering order) to assess its effectiveness. The results demonstrate that OM4AnI serves as an effective quality metric for these two common XAI scenarios, paving the way for adapting other quality metrics to be scalable within XAI contexts.

Index Terms: Visual quality metrics, scatterplots, explainable AI, evaluation of quality metric

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) [35] refers to Artificial Intelligence systems or methods that maintain human oversight, and it has become a popular research area. XAI helps machine learning (ML) experts better understand and improve their models. The XAI process typically involves three primary stages: model understanding, performance evaluation and enhancement, and stakeholder communication [31]. Common XAI tasks within these stages include investigating *feature importance*, *feature dependency*, and *model accuracy*. Scatterplots are widely utilized in these tasks, e.g. to visualize SHAP values [30] representing local feature dependencies or to display misclassifications using t-SNE [33] output.

Scatterplots are a fundamental technique in data visualization, used to represent the relationship between two numerical variables. By plotting individual data points on a two-dimensional grid, scatterplots help in identifying patterns, trends, and potential outliers within the dataset. A key issue with scatterplots is overplotting, especially with large datasets, where the overplotting affects the observation of users from scatterplots because of the covered data

points. Therefore, quality metrics that assess the effectiveness of scatterplots are valuable techniques for visualization designers aiming to improve their designs.

Existing quality metrics can be categorized into three groups [6]. The first category includes fundamental quality metrics, which measure basic properties of scatterplots, such as aspect ratio or screen occupation percentage. The second category consists of metrics that assess the visual effectiveness of scatterplots, such as the correlation between two variables [32] or cluster separation [29, 16], to evaluate how well a scatterplot conveys the underlying data. The third category includes feature preservation metrics, which assess whether the abstracted data or dimensions retain the characteristics of the original data [15, 5].

However, existing quality metrics across all three categories for scatterplots have not demonstrated sufficient effectiveness in XAI applications, particularly in the context of feature importance and model accuracy. This paper aims to address this gap by applying a recently developed quality metric, OM4AnI (Overlap Measure for Anomaly Identification) [23], which belongs to the second category of metrics, those that assess the visual effectiveness of scatterplots, within the context of XAI. We begin by summarizing two visual patterns relevant to three common XAI tasks. We then evaluate the effectiveness of OM4AnI in two typical XAI scenarios corresponding to the two visual patterns. The first scenario involves identifying feature importance using an XGBoost model applied to the Census Income dataset; the second assesses model accuracy using a Convolutional Neural Network (CNN) [22] applied to the MNIST dataset [11]. The results show that OM4AnI accurately reflects changes in scatterplot quality as visual factors, such as marker size and rendering order, cause information to become more or less visible. It demonstrates strong potential for evaluating scatterplot quality in XAI contexts, although some limitations remain.

2 BACKGROUND AND RELATED WORKS

Our paper focuses on evaluating quality metrics for scatterplots in the context of XAI. Our literature review is structured into two parts: (1) the main XAI tasks that ML experts perform and the ways that scatterplots are used in those tasks, and (2) the quality metrics associated with scatterplots.

2.1 XAI Tasks and Scatterplots

In XAI, the process is categorized into three primary stages as outlined by Spinner et al. [31]: (1) understanding the functionality of a specific model; (2) evaluating the model to identify issues and enhance its performance through analysis of its outputs; and (3) engaging with stakeholders to communicate insights. Within these stages, we identify three popular tasks in XAI: investigating feature importance, feature dependency, and identifying model accuracy. Scatterplots are extensively used in XAI to analyze the three tasks through various visual patterns, such as *cluster*, *outlier*, *linear correlation*, and *manifold*. Cluster refers to points that form distinct groups or clusters, indicating subgroups within the data. Outlier identifies points that significantly deviate from the general data pattern. Linear correlation represents points that align in a straight-line

*e-mail: l.liu6@leeds.ac.uk

†e-mail: L.V.Bogachev@leeds.ac.uk

‡e-mail: M.Rezaei@leeds.ac.uk

§e-mail: N.Ravikumar@leeds.ac.uk

¶e-mail: A.Khara@leeds.ac.uk

||e-mail: tsmaz@leeds.ac.uk

**e-mail: R.A.Ruddle@leeds.ac.uk

pattern, indicating a linear relationship between variables. Manifold describes points that form a non-linear yet continuous shape or curve, reflecting more complex relationships.

Both cluster and outlier detection can be employed to assess model accuracy (see, e.g., [20]). ML experts often use scatterplots to explore misclassification instances through projection views, where high-dimensional data is mapped onto a 2D plane [18]. Scatterplots help reveal data distributions and are commonly integrated into interactive systems, enabling instance selection across linked views [14]. Some studies focus on visualizing neural network activations from single or multiple layers [26, 12], while others examine feature distributions [34, 20].

Outlier, linear correlation, and manifold can be used to identify feature dependency and feature importance [13, 8], which is closely related to the analysis of SHAP values. Based on cooperative game theory (see Shapley [30]), SHAP values have become central to XAI [21]. The SHAP (SHapley Additive exPlanations) package (<https://github.com/shap/shap>) is widely used to explain how input features influence model predictions through advanced visualizations. Scatterplots support the interpretation of feature importance and inter-feature relationships by revealing trends, peaks, turning points, and other significant patterns [13].

2.2 Quality Metrics for Scatterplots

The concept of visual quality metrics has been essential in the information visualization community since its inception [3]. These metrics serve two main purposes: (1) to assess how effectively a visualization communicates information to users, and (2) to aid in selecting the most appropriate visualization techniques [7]. Among various visualization methods, scatterplots are particularly popular and widely used.

Scatterplot quality metrics can be categorized into three groups [6]. The first category concerns basic computations, such as the number of data items, density, and screen occupation percentage. The second concerns visualization effectiveness, particularly cluster separation [29, 1, 16], often assessed using point-distance-based statistical methods. Aupetit et al. [2] showed that the Class-Proportion of 0.35-Observable Neighbors (CPON) aligned better with human judgment than Distance Consistency. ClustMe [1] uses a Gaussian Mixture Model to evaluate cluster clarity, while CLAMS [16] applies a regression model to estimate local ambiguity. Quadri et al. [27, 28] examined cluster perception using a topology-based merge tree.

In addition, correlation detection is also a key quality metric for assessing the effectiveness of visualizations. Behrisch et al. [4] introduced a sorting pipeline for scatterplot matrices, extended by Micallef et al. [25] with optimization techniques using Pearson’s coefficient and covariance. Wöhler et al. [36] trained a CNN using perception-based metrics aligned with human annotations. Beyond task-based methods, task-agnostic metrics assess hidden information regardless of task, such as Nearest Neighbor Distance (NND) [9]. OM4AnI is a task-based method that uniquely quantifies how anomalies are occluded by overplotting in multi-class scatterplots.

The third type is feature preservation metrics, which evaluate whether dimensions or abstracted data retain their original characteristics [15, 5]. The first assesses how well high-dimensional data structures are preserved during projection to two dimensions, as in the “Classes Are Not Clusters” metric [15]. The second evaluates sampling methods by measuring how accurately data density is retained, such as the metric by Bertini and Santucci [5], which uses active pixel counts, or the distance-based method by Johansson and Cooper [17].

Quality metrics can generally be evaluated in two ways: (1) by conducting user studies that compare metric outputs with human judgments [2, 29]. However, these approaches are not suitable for evaluating OM4AnI [23], which measures how many anomalies are

hidden in multi-class scatterplots. Since such anomalies are visually occluded, users cannot perceive them directly, making human judgment an unreliable baseline. (2) by analyzing their usage in various contexts [6]. However, existing quality metrics have not been validated in XAI contexts, particularly for tasks such as feature importance and model accuracy. This paper addresses this gap by evaluating OM4AnI within the XAI domain, demonstrating its potential for assessing visualization quality in support of model interpretability.

3 OM4AnI DESCRIPTION

OM4AnI is designed to measure the extent to which anomalies in multi-class scatterplots are occluded, providing users with a score ranging from 0 to 1. We use Q to denote the result of OM4AnI. A score of $Q = 0$ indicates the worst case, while $Q = 1$ represents an ideal scatterplot in which no anomalous data instances are occluded. This score helps users assess whether their scatterplots conceal anomalous instances, and if so, to what extent. In addition, OM4AnI identifies regions within the scatterplot where these anomalous instances are occluded, using heatmaps. OM4AnI was implemented in Python, leveraging the `scikit-learn` and `scipy` libraries to support various machine learning relevant algorithms, including Mahalanobis Distance, Gaussian Kernel Density Estimation (`gaussian_kde`), and t-SNE. We only introduce the key configurations of OM4AnI in this paper. For more detailed information, please refer to [23].

This section first introduces two visual patterns used to address three XAI tasks. Based on two visual patterns, we then describe how to select the parameters in OM4AnI, with particular emphasis on the computation of the anomaly index and the weighting parameters.

3.1 Visual Patterns

Building on the scatterplot usage in XAI tasks introduced in Section 2.1, we summarize two groups of visual patterns for addressing the XAI tasks:

- (1) **Cluster-based patterns.** This pattern involves analyzing scatterplots to identify closely positioned groups of points, or clusters. Such clusters often reveal relationships or similarities among data points within each group, providing insights into the underlying data structure. It includes both cluster and outlier patterns, where most—but not all—data points from the same category form a single cluster. In XAI tasks focused on model accuracy, cluster-based patterns help users assess the separation between different classes and identify misclassified instances.
- (2) **Regression-based patterns.** This pattern focuses on examining and interpreting the relationship between two variables represented on the scatterplot. A positive correlation is typically indicated by an upward trend of points, suggesting that as one variable increases, so does the other. Conversely, a negative correlation is characterized by a downward trend. Tatu [32] noted that ‘good correlations are represented as long, skinny structures in the visualization.’ This pattern includes linear correlation, outliers, and manifold. In XAI feature importance and feature dependency tasks, the regression-based patterns can reveal insights such as the point where SHAP values become zero, the stability of SHAP values over a specific range, and significant increases or decreases.

3.2 Parameters Selection

To use OM4AnI, two key configurations are required. First, users must define the method for **anomaly index computation**, which determines how the anomaly score for each instance is calculated. The core idea is to assess the relative relationship between each

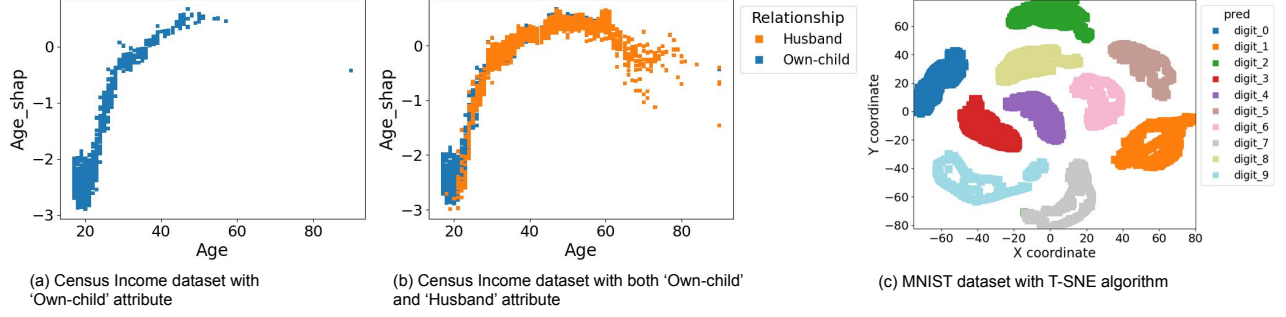


Figure 1: Scatterplot examples in the field of XAI. (a) and (b) display SHAP values, where the x-axis represents the feature ‘Age’ and the y-axis shows the corresponding SHAP values. (a) contains only the category ‘Own-child’, while (b) includes two categories. (c) presents a scatterplot of classification instances, with the x- and y-axes representing the coordinates of each instance as computed by the dimensionality reduction method t-SNE.

data instance and its corresponding category or group. OM4AnI supports several approaches for computing this index, such as the Mahalanobis distance [24] for cluster-based patterns and Cook’s distance [10] for regression-based patterns.

Mahalanobis Distance [24] calculates the distance for each data point to clusters. It first determines the mean vector (μ) and the covariance matrix (S) of the dataset. These elements capture the central tendency and the variability of the data across all dimensions, respectively. For a given data point x , the Mahalanobis Distance is calculated by computing the deviation of x from μ , then transforming this deviation vector using the inverse of the covariance matrix S^{-1} , and finally multiplying the transformed vector by the deviation vector itself. The square root of this scalar product yields the Mahalanobis Distance of the data point from the dataset’s mean, accounting for the correlation between variables. Mathematically, this is expressed as:

$$D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}, \quad (1)$$

where x represents the vector of the data point, μ is the mean vector of the dataset, and S^{-1} is the inverse of the covariance matrix.

Cook’s Distance [10] is one of the common influence functions measuring how each data point affects the estimated regression coefficients, which combines residuals and leverage values to assess influence:

$$D_i = \frac{e_i^2}{s^2 \cdot p} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}, \quad (2)$$

where e_i is the residual, s^2 is the mean squared error, p is the number of model parameters, and h_{ii} is the leverage. Observations with a high Cook’s Distance are considered influential.

Second, users must specify the **weighting parameters** λ and β . The parameter λ reflects the relative importance of data points occluded by others from the same class, where larger values indicate greater importance. In contrast, β controls the importance of data points occluded by markers from different classes. The values of λ and β range from 0 to $+\infty$.

To help users identify where anomalies are located, OM4AnI provides a variable $o(k)$, referred to as the *hidden information degree*, which quantifies how many data instances are occluded by others in each pixel. The value of $o(k)$ ranges from 0 to 1, where 0 indicates minimal hidden information and 1 indicates severe hidden information.

4 XAI CASE STUDIES

This section evaluates OM4AnI using two examples based on public datasets and implemented with basic machine learning models. We first introduce the datasets used in our study: the Census Income

dataset [19] for feature importance and the MNIST dataset [22] for model accuracy. Next, we describe the visual factor settings. Finally, we present two studies illustrating how OM4AnI’s values change in response to different visual factors.

Additionally, we employed the SHAP library to train an XGBoost model on the Census Income dataset, and the TensorFlow library to train a CNN on the MNIST dataset. All computational experiments were conducted on a Windows system equipped with a 13th Gen Intel(R) Core(TM) i7-13700H processor at 2.40 GHz and 32 GB of RAM.

4.1 Datasets

The Census Income dataset [19], derived from the 1994 Census database, is widely used to test the performance of classification models in machine learning. It encompasses various demographic and employment-related attributes of adults, including age, work class, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours worked per week, and native country. The dataset information is shown in Table 1. The primary analytical objective of this dataset is binary classification: to ascertain whether an individual’s annual income surpasses \$50,000. Using an XGBoost model, we predict the income classes and analyze the classification outcomes alongside SHAP values for each attribute.

The MNIST dataset (see Table 1) [22] is a widely used benchmark in machine learning for evaluating image recognition algorithms. It contains 70,000 grayscale images of handwritten digits, with 60,000 used for training and 10,000 for testing. Each image is a 28×28 pixel bitmap representing a digit from 0 to 9. The primary task is to classify each image into its corresponding digit category. In our study, a CNN is trained to perform this classification. The model outputs 10 attributes per instance, each representing the probability of belonging to a specific digit class. To visualize the model’s output, we apply the t-SNE dimensionality reduction (DR) technique, commonly used in XAI, to project the results into a two-dimensional space. The resulting two attributes serve as coordinates for generating scatterplots that cluster instances based on model predictions.

Fig. 1 shows two typical examples of scatterplots used in XAI. Fig. 1(a) and (b) are derived from the XGBoost (<https://xgboost.ai/>) model’s predictions on the Census Income dataset. It presents SHAP values that illustrate how the feature ‘Age’ affects the model’s output across different demographic groups, such as those labeled ‘Husband’ or ‘Own-child’, which are distinguished by color. Based on the trends and values in Fig. 1(a), an ML expert can observe that as ‘Age’ approaches around 30 years old, it begins to have a positive impact on the model output. However, there are fewer data instances when ‘Age’ exceeds 60. When visualizing two demographic groups in Fig. 1(b), many data instances

Table 1: Summary of the datasets in our XAI case studies.

Datasets	# of Data records (k)	Classification model	Data type	Generating scatterplot	Visual patterns	XAI task
Census Income	6.513	XGBoost	Tabular	SHAP	Regression-based	Feature importance
MNIST	10	CNN	Image	t-SNE	Cluster-based	Model accuracy

from the ‘Own-child’ group are occluded by those from the ‘Husband’ group. This occlusion prevents users from clearly observing the trends and specific impact of the ‘Own-child’ group on the model output. Fig. 1(c) is based on the MNIST dataset, where a CNN model is used to classify digit images, followed by DR using t-SNE. In this scatterplot, the x- and y-axes represent the reduced dimensions, and the colors indicate the digit classes predicted by the CNN. The purpose of this scatterplot is to identify misclassifications: data instances are projected into clusters where the majority of neighboring instances belong to different classes. However, in Fig. 1(c), users cannot observe the misclassified data instances, as all clusters appear to be uniformly colored. This visual impression is misleading; some misclassified instances are indeed present but are occluded.

4.2 Visual Factor Selection

A range of visual factors in scatterplots can influence the effectiveness of quality metrics, particularly in the context of specific XAI tasks. In our case studies, we tested three visual factors: rendering order, marker size, and marker shape.

The rendering order of data points can significantly impact scatterplot readability. Since points are drawn sequentially, the rendering order determines which points remain visible and which are obscured. If the order consistently favors certain categories or features, it may introduce visual bias. While users typically consider marker size, shape, and color when designing scatterplots, rendering order is often overlooked. For example, in Fig. 1(c), the default rendering order results in substantial overlap among classes, making it difficult to observe occluded instances. However, by re-ordering the rendering so that misclassified instances are drawn last, these points will become more visible, allowing users to identify more misclassified data points.

In the case studies, we tested with different rendering orders to examine how scatterplots and their quality metrics change with various rendering strategies. In general, the most commonly used rendering order strategy is random rendering, which draws the scatterplot based on the inherent order of the raw data. The second most commonly used strategy for multi-class scatterplots is the category-based method, which renders the data instances class by class, drawing one category of instances after another. OM4AnI also provides a rendering order strategy, which is based on the value of the anomaly index. Specifically, OM4AnI renders the data instances in ascending order of anomaly index values. This approach ensures that instances with higher anomaly index values are rendered later, making them less likely to be occluded by other points and thus reducing the possibility of biased observations. Our evaluation included category-based and OM4AnI rendering orders.

Marker size significantly influences the level of overplotting: larger markers tend to overlap more easily, leading to visual clutter, while smaller markers may be more difficult to distinguish. We used values of 100 and 25 in both examples to examine how scatterplots change with varying marker sizes and how the quality metrics respond accordingly. Additionally, we tested different marker shapes, rectangles and plus symbols, for the Census Income dataset. However, this aspect is not discussed in the paper, as it had minimal effect on quality values of OM4AnI.

4.3 Feature Importance for Census Income Dataset

To assess the positive and negative impacts of individual features on the XGBoost model applied to the Census Income dataset, we computed the SHAP values for each feature and visualized them in the scatterplots shown on the left side of Fig. 2. The x-axis represents age, while the y-axis displays the corresponding SHAP values, illustrating how age influences the model’s output in terms of both direction (positive or negative) and magnitude. Although the datasets used in both scatterplots are identical, the plots differ in rendering order. Fig. 2(a) uses the category-based rendering order, where instances with the label ‘Own-child’ are rendered first, followed by those labeled ‘Husband’. In contrast, Fig. 2(b) uses the OM4AnI rendering order, which sorts and renders the data instances based on their anomaly index values. For both Fig. 2(a) and (b), age has a negative impact (i.e., it contributes to predicting an annual income below \$50,000) before age 30, a positive impact between ages 30 and 60, and a negative impact after age 60.

To compute the quality metrics, we used Cook’s Distance to derive the anomaly index. We set the parameters λ and β to 0 and 100, respectively. Additionally, we calculated the hidden information degree $o(k)$ for each pixel and visualized the results as heatmaps, shown on the right of Fig. 2, corresponding to the two scatterplots.

In the scatterplot shown in Fig. 2(a), we observed that nearly all data points labeled ‘Own-child’ are obscured by those labeled ‘Husband’. The heatmap provided by OM4AnI and shown on the right side of Fig. 2(a), highlights the amount of occluded anomalous data points. For example, although one data instance labeled ‘Own-child’ is completely invisible in the scatterplot, the corresponding heatmap highlights this as an anomaly. To better reveal the ‘Own-child’ data, we reordered the scatterplot using the OM4AnI order, as shown in Fig. 2(b). This adjustment led to a higher quality metric value ($Q = 0.69$) compared to the scatterplot in Fig. 2(a) ($Q = 0.03$). In Fig. 2(b), many ‘Own-child’ data points become visible, including the previously hidden instance on the right (highlighted one) of the scatterplot in Fig. 2(a).

4.4 Model Accuracy for MNIST Dataset

To analyze model accuracy, we plotted scatterplots in which the x- and y-axes represent the coordinates generated by the t-SNE algorithm, and the colors indicate the predicted class labels from the CNN model for each instance (see Fig. 3). The three scatterplots on the left display the same data but differ in rendering order and marker size. The rendering order in Fig. 3(a) follows a category-based ordering strategy: $digit_9 \rightarrow digit_0 \rightarrow digit_2 \rightarrow digit_4 \rightarrow digit_5 \rightarrow digit_6 \rightarrow digit_8 \rightarrow digit_1 \rightarrow digit_3 \rightarrow digit_7$. In contrast, Figs. 3(b) and (c) use the OM4AnI rendering order. As a result, misclassified instances of digits 2, 5, and 8 are hidden in Fig. 3(a), but become visible in Figs. 3(b) and (c).

Using OM4AnI, we applied the Mahalanobis distance to compute the anomaly index. The weighting parameters were set as $\lambda = 0$ and $\beta = 100$. Additionally, we calculated the hidden information degree $o(k)$ for each pixel and visualized the results using heatmaps, shown in the right-hand column of plots in Fig. 3. These visualizations help users identify regions where anomalous data points are occluded.

Fig. 3(a) is low quality ($Q = 0.12$) and the heatmap highlights the three locations where information about the misclassified instances (anomalous data points) is hidden. Fig. 3(b) is higher qual-

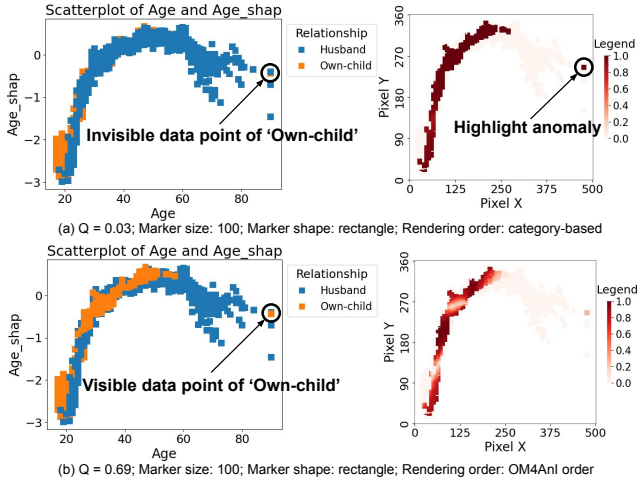


Figure 2: The application of OM4AnI to feature importance analysis for Census Income dataset. The two scatterplots on the left show how age affects the model's output, using different rendering orders: category-based in (a) and OM4AnI order in (b), each with a different quality metric value. The heatmaps on the right indicate the extent of information obscured in the corresponding scatterplots and highlight regions where anomalous data points are hidden.

ity ($Q = 0.44$) and the misclassified digits are clearly visible in the scatterplot. To assess the sensitivity of our quality metrics to marker sizes, we plotted another scatterplot with smaller marker sizes (25 instead of 100) in Fig. 3(c). The resulting value was 0.66, indicating that the quality metric values tend to increase with smaller marker sizes.

5 DISCUSSION AND LIMITATIONS

The two case studies illustrate the effectiveness of OM4AnI when applied to different visual patterns and tasks in the context of XAI. These studies show that OM4AnI can accurately reveal the overplotting of a scatterplot. Particularly when the rendering order of the scatterplots is changed, some anomalous data points may be obscured, which should not happen. If these data are hidden, ML experts might draw incorrect conclusions about their model and make erroneous decisions when adjusting ML parameters or comparing models.

Moreover, the heatmap can help ML experts identify where hidden information is located. For example, in Fig. 3, while ML experts might not notice any hidden information in the scatterplot of Fig. 3(a), OM4AnI indicated that there are indeed misclassified data points hidden. Specifically, instances incorrectly predicted as Digit 3, Digit 7, and Digit 1 are highlighted. This guidance is crucial for accurately assessing model performance and making informed adjustments.

Additionally, OM4AnI is sensitive to marker size, as the degree of overlap within each pixel can vary. When marker sizes are reduced, OM4AnI tends to increase, since overplotting is minimized. However, there is a limitation: if the marker size is reduced to the point where the data points in the scatterplot become barely visible to users, OM4AnI may yield extremely high values due to the absence of overlap. This situation is not practical for real-world applications. We also believe that OM4AnI is sensitive to marker shape, as different marker areas can influence the computed results.

Moreover, OM4AnI has limitations related to scatterplot opacity. Specifically, OM4AnI assumes that all data points are fully opaque; thus, any point obscured by others, regardless of the degree of overlap, remains entirely invisible to users.

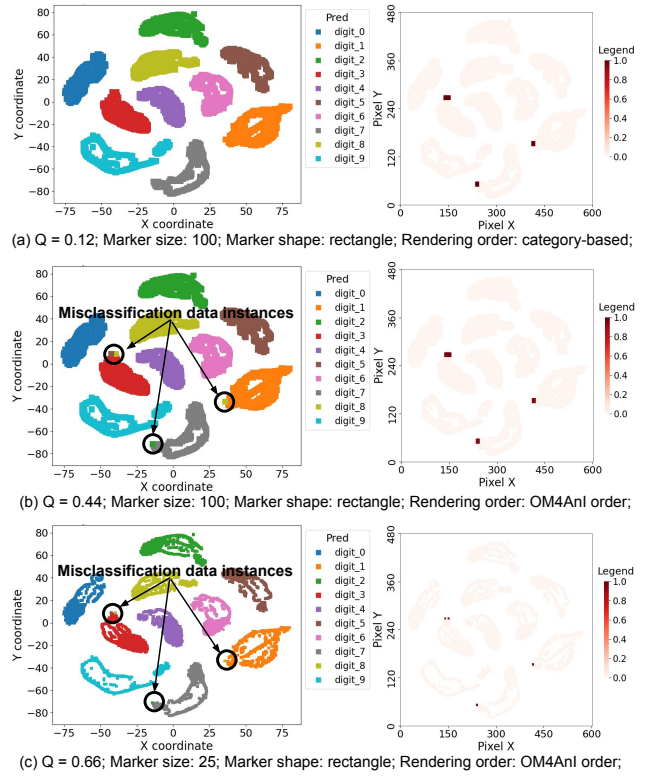


Figure 3: The application of OM4AnI to model accuracy analysis for MNIST dataset. The three scatterplots, derived from the same dataset, are used to identify misclassifications in the CNN model's predictions, with corresponding quality metric values computed using OM4AnI. The three heatmaps on the right show the extent of information obscured in the scatterplots and highlight regions where anomalous data points (i.e., misclassified instances) are hidden.

6 CONCLUSION

This paper evaluated our visual quality measure OM4AnI in the context of XAI to assess its effectiveness and usability. We began by introducing two visual patterns related to three common XAI tasks and explained how to select the parameters of OM4AnI based on these patterns and tasks. We then conducted case studies by applying OM4AnI in two scenarios to examine how its scores respond to various visual factors, including marker size, marker shape and rendering order. The results showed that OM4AnI not only provides accurate scores that reflect the extent to which anomalous data points are occluded under different visual factors, but also identifies the regions where such anomalous points are located, when they exist. Our findings demonstrate that OM4AnI is suitable for use in XAI scenarios and helps pave the way for adapting other quality metrics to similar contexts.

ACKNOWLEDGMENTS

This research was supported by the Engineering and Physical Sciences Research Council (grant EP/X029689/1).

The datasets used in this study are publicly available on [Dropbox](#). The source code is available in the [GitHub repository](#).

REFERENCES

- [1] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. *ClustMe: A visual quality measure for ranking monochrome scatterplots based on cluster patterns*. *Computer Graphics Forum*, 38(3):225–236, 2019. doi: 10.1111/cgf.13684 2
- [2] M. Aupetit and M. Sedlmair. *SepMe: 2002 new visual separation measures*. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*,

- pp. 1–8. IEEE, Taipei, Taiwan, 2016. doi: 10.1109/PACIFICVIS.2016.7465244 2
- [3] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, T. Schreck, D. Weiskopf, and D. A. Keim. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446 2
 - [4] M. Behrisch, L. Shao, J. Buchmüller, and T. Schreck. Quality metrics driven approach to visualize multidimensional data in scatterplot matrix. In *Eurographics Conference on Visualization (EuroVis 2015)*, 2016. <https://bib.dbvis.de/uploadedFiles/QualityMetricsSPLOM.pdf>. 2
 - [5] E. Bertini and G. Santucci. Quality metrics for 2D scatterplot graphics: Automatically reducing visual clutter. In A. Butz, A. Krüger, and P. Olivier, eds., *Smart Graphics*, vol. 3031 of *Lecture Notes in Computer Science*, pp. 77–89. Springer, Berlin, Germany, 2004. doi: 10.1007/978-3-540-24678-7_8 1, 2
 - [6] E. Bertini and G. Santucci. Visual quality metrics. In *BELIV’06: Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for Information Visualization*, pp. 1–5. Association for Computing Machinery, Venice, Italy, 2006. doi: 10.1145/1168149.1168159 1, 2
 - [7] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011. doi: 10.1109/TVCG.2011.229 2
 - [8] D. Bertsimas, G. Lukin, L. Mingardi, O. Nohadani, A. Orfanoudaki, B. Stellato, H. Wiberg, S. Gonzalez-Garcia, C. L. Parra-Calderón, K. Robinson, M. Schneider, B. Stein, A. Estirado, L. A. Beccara, R. Canino, M. D. Bello, F. Pezzetti, A. Pan, K. Akinosoglou, A. Antoniadou, K. Argyraki, G. N. Dalekos, M. Gaga, N. K. Gatselis, C. Gogos, I. Kalomenidis, E. Kranidioti, E. Korompoki, G. Lourida, E. Margellou, G. Ntaios, P. Panagopoulos, A. Pefanis, V. Petrakis, C. Psarrakis, V. Sakka, K. Thomas, and E. Zervas. COVID-19 mortality risk assessment: An international multi-center study. *PLoS ONE*, 15(12):1–13, 2020. doi: 10.1371/journal.pone.0243262 2
 - [9] P. J. Clark and F. C. Evans. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453, 1954. doi: 10.2307/1931034 2
 - [10] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977. doi: 10.1080/00401706.1977.10489493 3
 - [11] L. Deng. The MNIST database of handwritten digit images for machine learning research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477 1
 - [12] H. S. Garcia Caballero, M. A. Westenberg, B. Gebre, and J. J. van Wijk. V-Awake: A visual analytics approach for correcting sleep predictions from deep learning models. *Computer Graphics Forum*, 38(3):1–12, 2019. doi: 10.1111/cgf.13667 2
 - [13] A. D. Haimovich, N. G. Ravindra, S. Stoytchev, H. P. Young, F. P. Wilson, D. v. Dijk, W. L. Schulz, and R. A. Taylor. Development and validation of the quick COVID-19 severity index: A prognostic tool for early clinical decompensation. *Annals of Emergency Medicine*, 76(4):442–453, 2020. doi: 10.1016/j.annemergmed.2020.07.022 2
 - [14] T. Jaunet, R. Vuilleumot, and C. Wolf. DRLViz: Understanding decisions and memory in deep reinforcement learning. *Computer Graphics Forum*, 39(3):49–61, 2020. doi: 10.1111/cgf.13962 2
 - [15] H. Jeon, Y.-H. Kuo, M. Aupetit, K.-L. Ma, and J. Seo. Classes are not clusters: Improving label-based evaluation of dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):781–791, 2024. doi: 10.1109/TVCG.2023.3327187 1, 2
 - [16] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafir, and J. Seo. CLAMS: A cluster ambiguity measure for estimating perceptual variability in visual clustering. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):770–780, 2024. doi: 10.1109/TVCG.2023.3327201 1, 2
 - [17] J. Johansson and M. Cooper. A screen space quality method for data abstraction. *Computer Graphics Forum*, 27(3):1039–1046, 2008. doi: 10.1111/j.1467-8659.2008.01240.x 2
 - [18] S. Kaski and J. Peltonen. Dimensionality reduction for data visualization. *IEEE Signal Processing Magazine*, 28(2):100–104, 2011. doi: 10.1109/MSP.2010.940003 2
 - [19] R. Kohavi. Census Income [dataset], 1996. UC Irvine Machine Learning Repository. Irvine, CA, USA. doi: 10.24432/C5GP7S 3
 - [20] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, 2019. doi: 10.1109/TVCG.2018.2865027 2
 - [21] J. Lange and C. Lange. Applying machine learning and AI explanations to analyze vaccine hesitancy. Preprint, Jan 2022, arXiv:2201.05070 [cs.LG]. doi: 10.48550/arXiv.2201.05070 2
 - [22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541 1, 3
 - [23] L. Liu, L. Bogachev, M. Rezaei, N. Ravikumar, A. Khara, M. Azarmi, and R. A. Ruddle. OM4AnI: A novel overlap measure for anomaly identification in multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, Early Access:1–12, 2025. doi: 10.1109/TVCG.2025.3642219 1, 2
 - [24] G. J. McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999. doi: 10.1007/BF02834632 3
 - [25] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauff. Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1588–1599, 2017. doi: 10.1109/TVCG.2017.2674978 2
 - [26] N. Pezzotti, T. Holtt, J. Van Gemert, B. P. Lelieveldt, E. Eisemann, and A. Vilanova. DeepEyes: Progressive visual analytics for designing deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):98–108, 2018. doi: 10.1109/TVCG.2017.2744358 2
 - [27] G. J. Quadri, J. A. Nieves, B. M. Wiernik, and P. Rosen. Automatic scatterplot design optimization for clustering identification. *IEEE Transactions on Visualization and Computer Graphics*, 29(10):4312–4327, 2022. doi: 10.1109/TVCG.2022.3189883 2
 - [28] G. J. Quadri and P. Rosen. Modeling the influence of visual density on cluster perception in scatterplots using topology. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1829–1839, 2021. doi: 10.1109/TVCG.2020.3030365 2
 - [29] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015. doi: 10.1111/cgf.12632 1, 2
 - [30] L. S. Shapley and M. Shubik. A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, 48(3):787–792, 1954. doi: 10.2307/1951053 1, 2
 - [31] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. ExplAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, 2020. doi: 10.1109/TVCG.2019.2934629 1
 - [32] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnork, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pp. 59–66, 2009. doi: 10.1109/VAST.2009.5332628 1, 2
 - [33] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. <https://jmlr.org/papers/v9/vandermaaten08a.html>. 1
 - [34] J. Wang, L. Gou, H. Yang, and H.-W. Shen. GANViz: A visual analytics approach to understand the adversarial game. *IEEE Transactions on Visualization and Computer Graphics*, 24(6):1905–1917, 2018. doi: 10.1109/TVCG.2018.2816223 2
 - [35] Wikipedia contributors. Explainable artificial intelligence. Wikipedia, The Free Encyclopedia, 2024. Page Version ID: 1242878433. 1
 - [36] L. Wöhler, Y. Zou, M. Mülhausen, G. Albuquerque, and M. Magnor. Learning a perceptual quality metric for correlation in scatterplots. In *Vision, Modeling and Visualization (VMV19)*, pp. 55–62. The Eurographics Association, 2019. doi: 10.2312/VMV.20191318 2