Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

AISYST: AI-Powered Interactive Visual System to Assist With Fidelity Assessment of Synthetic Tabular Data

Liqun Liu[1] | Leonid Bogachev[2,3] | Netochukwu Onyiaji[4] | Lukas Čironis[3] | János Gyarmati-Szabó[2,3] | Roy A. Ruddle[1,4]

[1]School of Computer Science, University of Leeds, Leeds, UK | [2]Department of Statistics, School of Mathematics, University of Leeds, Leeds, UK | [3]4-Xtra Technologies Limited, Number 3 Acorn Business Park, Airedale Business Centre, Skipton, North Yorkshire, UK | [4]Leeds Institute for Data Analytics (LIDA), University of Leeds, Leeds, UK

Correspondence: Liqun Liu (l.liu6@leeds.ac.uk)

Received: 2 April 2026 | **Revised:** 26 June 2026 | **Accepted:** 10 July 2026

Keywords: data assessment | data quality | large language models (LLM) | synthetic data | visualization design

ABSTRACT

Evaluating synthetic data produced by generative models remains a critical challenge in sensitive domains such as healthcare and finance. Ensuring that such data is 'faithful' to real data is essential for downstream applications and decision-making, including regulatory compliance. This paper introduces an AI-powered interactive visual system—AISYST—designed to assess the fidelity of synthetic tabular datasets. The system supports multilevel comparisons with real datasets, spanning multivariate resemblance analyses based on dimensionality reduction through suitable two-dimensional projections, bivariate correlation and univariate similarity. AISYST also integrates an AI assistant by leveraging state-of-the-art large language models to summarize key findings and generate suggestions for improving synthetic data generation models. We validated the capabilities of AISYST through three case studies, supported by feedback from industrial AI experts who endorsed its broader deployment.

1 | Introduction

Synthetic data are increasingly used in sensitive domains such as medicine, health, business, banking, insurance, and so forth, particularly with the booming adoption of machine learning for training classifiers and making predictions (Yan et al. 2022). In such domains, real data are often inaccessible for model training due to privacy, security, or regulatory constraints. Consequently, synthetic data play an important role in addressing this limitation. However, to be usable and reliable, synthetic data generation models must be systematically assessed within the context of their intended use case before deployment (Yan et al. 2022).

Fidelity, also known as *resemblance* (Yan et al. 2022), refers to the degree to which synthetic data preserve the distributions, correlations and higher-order statistical patterns present in real data. It is an important dimension of synthetic data evaluation (Santangelo et al. 2024). Current methods for evaluating the fidelity of synthetic tabular data fall into two main categories: quality metrics and visualization techniques. Quality metrics assign scores based on how well the synthetic data preserves the statistical properties of the original data, typically focusing on univariate similarity, bivariate correlation and multivariate resemblance (Goncalves et al. 2020; McInnes, Healy, and Melville 2018; McInnes, Healy, Saul, and Großberger 2018). Static visualization techniques also target these three aspects, using histograms and bar charts for univariate similarity analysis (Kaur et al. 2021; Ping et al. 2017; Budu et al. 2024), heatmaps and correlation matrices for bivariate correlation analysis (Che et al. 2017; Livieris et al. 2024) and dimensionality-reduced scatterplots for multivariate resemblance analysis (Budu et al. 2024). However, advanced interactive visualization tools, enabling users to dynamically explore synthetic tabular data, are currently scarce and limited in their functionality (Santangelo et al. 2025; Brito et al. 2025).

Despite recent advancements, the assessment of synthetic data remains challenging, particularly for practitioners who must

determine whether retraining is necessary to generate higher-quality synthetic tabular data and how to adjust the generative models accordingly. Existing quality metrics typically provide only a one-time, holistic assessment and lack detailed diagnostic capabilities, such as identifying specific variables or relationships that deviate from the real data that would support model refinement. Static visualization techniques, meanwhile, offer only a general overview of the differences between real and synthetic data, without enabling deeper, targeted exploration. Although some interactive visualization tools combine various quality metrics to indicate the overall fidelity of synthetic data, they do not allow users to interactively explore the datasets to understand where and how the synthetic data fails.

To address this gap, our paper presents a domain-agnostic interactive visual tool for assessing synthetic data, AISYST (**AI**-powered **SY**nthetic data interactive visual **S**ys**T**em), which compares synthetic data with real data using a range of visualization techniques, including scatterplots, histograms, violin plots and other plot types. The tool not only informs users about the overall fidelity of their synthetic data but also identifies where and why the synthetic data fails to replicate the characteristics of the real data. The system comprises:

i. a multivariate resemblance analysis visualization to incorporate scatterplots to examine the overall distributional resemblance between synthetic and real data and to help users select specific data instances for further exploration;

ii. a bivariate similarity analysis visualization to analyse pairwise correlations among variables in both datasets, enabling the exploration of relationships between variable pairs;

iii. a univariate similarity analysis visualization that compares the distributions of individual variables between the real and synthetic datasets; and

iv. an AI-assistant analysis to generate the summary PDF report to highlight the general issues, insights and recommendations.

The evaluation is based on three case studies demonstrating the effectiveness of AISYST across varying domains and dataset scales. In addition, feedback from industrial AI experts indicates that AISYST contributes significantly to the current field of synthetic data evaluation and highlights the usability of our visual system in helping to fine-tune their generative models by supporting a training process tailored to specific requirements.

Our contribution includes:

- A novel interactive visual system, AISYST, combines multiple visualization techniques to assist AI experts in interactively evaluating synthetic tabular data generated by generative models (Section 4).
- Development of an anomaly detection algorithm to identify anomalous subsets within gridified regions (Section 4.3).
- Integration of a large language model (LLM) agent with carefully crafted prompts to assess the quality of synthetic data, focusing on its fidelity (Section 4.8).
- Validation based on three case studies from different subject domains, demonstrating the tool effectiveness and ease of use, supported by the positive feedback received from industrial AI experts (Section 6).

2 | Related Work

This paper focuses on the fidelity evaluation of synthetic data, analysing it through multivariate resemblance, variable correlation and univariate similarity. We summarize related research along three dimensions: (1) synthetic data generation methods, (2) quality metrics for assessing synthetic data fidelity (accompanied by relevant quantitative scores) and (3) visualization techniques designed to illustrate the fidelity of synthetic tabular data.

2.1 | Synthetic Data Generation

The generation of synthetic tabular data has gained widespread applications across fields such as healthcare, finance and public policy, where data sensitivity is a major concern (Albuquerque et al. 2011). Synthetic datasets enable researchers and developers to conduct exploratory analyses, train machine learning models and test data processing pipelines without compromising the confidentiality of the original data.

Generative models such as *Generative Adversarial Nets (GAN)* (Goodfellow et al. 2014) and *Variational AutoEncoders (VAE)* (Kingma and Welling 2014) are widely used for synthetic tabular data generation. GANs consist of two neural networks—a generator and a discriminator—that are trained in opposition: the generator learns to produce data samples that mimic the real data distribution, while the discriminator attempts to distinguish between real and synthetic samples. Through iterative training, the generator improves its outputs to closely approximate the real data. Tabular-specific adaptations like *CTGAN* (Xu 2020) and *TableGAN* (Park et al. 2018) address the challenges of mixed data types and imbalanced categorical distributions by introducing techniques such as conditional generation and mode-specific normalization. More recently, *CA-GAN* (Nguyen et al. 2025) introduces causal awareness into tabular data generation by extracting a causal graph from real data via causal discovery methods and structuring the generator accordingly, with the aim of preserving underlying causal relationships.

Probabilistic graphical models, such as *Bayesian networks* (Pearl 1988), offer an alternative approach to synthetic data generation by explicitly modelling the conditional dependencies among variables. These models construct a directed acyclic graph where each node represents a variable and edges denote probabilistic dependencies. This method is particularly effective for tabular datasets with interpretable relationships.

2.2 | Fidelity Evaluation Metrics for Synthetic Tabular Data

This paper focuses on the fidelity evaluation of synthetic tabular data, thus we reviewed the quality metrics in fidelity. Fidelity can be classified into three categories (Santangelo et al. 2024):

(i) multivariate resemblance, (ii) variable correlation and (iii) univariate similarity.

Multivariate resemblance analysis focuses on the similarity between synthetic and real data in terms of statistical distributions in a multi-dimensional context. There are two main approaches to performing this analysis. One approach involves *dimensionality reduction (DR)* techniques (which reduce high-dimensional data to lower dimensions), followed by visualization of the results (e.g., via scatterplots) to observe the overall 'shape' of both the synthetic and real datasets. Popular DR techniques include *Uniform Manifold Approximation and Projection (UMAP)* (McInnes, Healy, and Melville 2018; McInnes, Healy, Saul, and Großberger 2018), *Principal Component Analysis (PCA)* (Jolliffe and Cadima 2016) and *t-Distributed Stochastic Neighbor Embedding (t-SNE)* (van der Maaten and Hinton 2008). An alternative approach uses similarity and distance measures, which directly compute the distance between real and synthetic data vectors. For example, *Cosine Similarity (CS)* (Venugopal et al. 2022) calculates the angle between real and synthetic data vectors, while *Jaccard Similarity (JS)* (Venugopal et al. 2022) evaluates the co-occurrence of discrete variables. More recently, advanced methods have been proposed to evaluate the structural fidelity of synthetic tabular data. For example, *TabStruct* (Jiang et al. 2026) assumes that tabular data is generated from an underlying structural causal model (SCM) and assesses whether synthetic data preserve the corresponding causal relationships.

Some methods assess the correlation between synthetic and real data by measuring the correlation between pairs of variables in either the real or synthetic dataset, and then calculating the difference between these correlations. For example, the *Pearson correlation coefficient* (Rodgers and Nicewander 1988) is a commonly used method to compute similarity between two continuous variables. For categorical variables, it is appropriate to use the aforementioned *Cramér's V* (Cramér 1946), a Chi-squared-based statistic that measures the strength of association between two categorical variables on a scale from 0 (no association) to 1 (perfect association). To assess the similarity between categorical and numerical variables, the *correlation ratio* (Fisher 1925) can be applied.

Univariate similarity involves comparing the distribution of a single variable in both synthetic and real data. The metrics used include distribution similarity measures such as *Kullback–Leibler (KL) divergence* (Goncalves et al. 2020; de Benedetti et al. 2020), which calculates the marginal probability mass function (PMF) and measures the similarity between the PMFs of synthetic and real data. Similarly to KL divergence, the *Kolmogorov–Smirnov (KS) test* (Kaur et al. 2021) also assesses the similarity of one-dimensional probability distributions between synthetic and real datasets. Besides, if the variable is categorical, methods such as *categorical Wasserstein distance* (Rubner et al. 2000) or *Cramér's statistic V* (Cramér 1946) can be used.

These methods for assessing the quality of synthetic data focus on assigning a score—or a small set of scores—to indicate how well the synthetic data can serve as a substitute for real data (Hernandez et al. 2025; Kurakova and Homayouni 2025). However, these quality metrics do not help users identify where and how the synthetic data fails, nor do they provide insights into specific issues that may affect downstream usability.

2.3 | Visualization for Synthetic Data Assessment

Visualization techniques are widely used across various domains for data analysis (Conejero et al. 2021; Hernando et al. 2018) and to support human decision-making (Akçay et al. 2012; Shao et al. 2025; Mohedano-Munoz et al. 2023). We summarize the visualization techniques employed in synthetic data evaluation with respect to fidelity, including how these techniques address multivariate resemblance, bivariate correlation and univariate similarity.

For multivariate resemblance analysis, the most commonly used methods involve DR techniques. These techniques reduce high-dimensional data to two dimensions, allowing for the use of scatterplots to visualize and compare the distributions of real and synthetic data (Livieris et al. 2024; Budu et al. 2024). In most cases, separate scatterplots are generated for real and synthetic data to facilitate comparison; in an alternative approach, Livieris et al. (2024) plotted both real and synthetic data within a single scatterplot, using different colours to distinguish between those.

For bivariate correlation analysis in both synthetic and real data, the most widely used visualization techniques are co-occurrence frequency heatmaps for categorical variables (Che et al. 2017) and correlation matrices for numerical variables (Livieris et al. 2024; Ping et al. 2017; Budu et al. 2024; Li et al. 2023; Chandra et al. 2022; Raab et al. 2021). In Noruzman et al. (2021) and Lautrup et al. (2025a), three correlation matrices were plotted: first, for the variables in the real data, then for the synthetic data, and finally a third matrix showing the absolute difference between the two. Lautrup et al. (2025a, 2025b) used the correlation matrix in a different way: they employed a heatmap to visualize the correlations between all variables in the real data and those in the synthetic data. Additionally, scatterplot matrices are also used to visualize pairwise correlations between variables (Budu et al. 2024).

Dimension-wise statistics (DWS) is one method used to evaluate cross-dimensional correlation, serving as a form of correlation assessment. To visualize the results of DWS, researchers typically use scatterplots to compare the DWS values of real and synthetic data (Yan et al. 2022; Kaur et al. 2021; Budu et al. 2024; Li et al. 2023), where the x-axis and y-axis represent the DWS results from the real and synthetic data, respectively. If all data points lie close to the diagonal line, this indicates that the synthetic data exhibits a high degree of correlation with the real data. Otherwise, it suggests that the synthetic data is not a good approximation of the real data.

To analyse univariate similarity, the most commonly used visualization techniques are histograms and bar charts (Kaur et al. 2021; Ping et al. 2017; Budu et al. 2024; Noruzman et al. 2021; Lautrup et al. 2025a). For numerical variables, histograms are typically used to compare the distributions of real

and synthetic data, with both distributions overlaid in a single plot. For categorical variables, bar charts are often employed, using different colours to distinguish between real and synthetic data.

In addition, line charts are also employed to visualize univariate distributions and compare real and synthetic data (Sidorenko et al. 2025). Beyond this, Li et al. (2023) used line and area charts to present the distribution of variables between real and synthetic data at each time point, where the x-axis represents the time series dimension for both datasets.

As was mentioned in the Introduction, most of existing visualization techniques are static, in that they do not allow users to explore synthetic data quality in a progressive or interactive manner. Two interactive visual tools have been developed to assist users in evaluating the quality of synthetic tabular data. The first, *SynthRO* (Santangelo et al. 2025), enables users to progressively select and combine different metrics, resulting in a fine-tuned composite metric for assessing synthetic data quality. The second, *SynthGuard* (Brito et al. 2025), helps users summarize various quality metrics using propensity scoring. However, both methods focus primarily on combining and comparing metrics via basic visualization, rather than on visualization design for analysing the fidelity of synthetic data. Moreover, none of the existing visual tools (either static or interactive) enable users to detect and explore abnormal instances in synthetic data, and thus cannot effectively support AI experts in identifying where and how the synthetic data fails.

3 | Tasks

In close collaboration with AI experts from our partner industrial organization, 4-Xtra Technologies Ltd., we identified key challenges in synthetic data assessment through three steps: (1) summarizing relevant synthetic data generation methods (see Section 2.1), (2) analysing the scope of existing assessment approaches (see Section 2) and (3) extracting five core assessment tasks through a literature review and expert input from our collaborator (discussed in this section).

Focusing on the scope of our paper, which centres on fidelity assessment, we define three tasks corresponding to the three core dimensions of fidelity: multivariate resemblance analysis (task T1), variable correlation analysis (task T2) and univariate similarity analysis (task T3). In addition, we incorporate two more tasks—subgroup fidelity exploration (task T4) and outlier and rare category detection (task T5)—which were derived from identified needs expressed by AI experts. To summarize, the five tasks are as follows:

- **T1: Multivariate resemblance analysis**. Assess the extent to which the global statistical distributions of synthetic data align with those of real data.

  Example: Compare income, age and education distributions across datasets. Identify regions (e.g., high income, age 30–40 and low education) present in the real data but missing or underrepresented in the synthetic data.

- **T2: Bivariate correlation analysis**. Examine whether inter-variable relationships (e.g., correlations, dependencies) are preserved in the synthetic data.

  Example: Analyse the correlation between age and health expenditure. If the real data shows a strong positive correlation, check whether this relationship is similarly reflected in the synthetic data.

- **T3: Univariate similarity analysis**. Identify individual variables that exhibit significant distributional differences between synthetic and real datasets.

  Example: Compare the distribution of the categorical variable 'employment status' or the numerical variable 'monthly rent' across datasets to detect any deviations.

- **T4: Subgroup fidelity exploration**. Evaluate fidelity within specific population subgroups, particularly for fairness and bias analysis.

  Example: Filter the dataset for individuals who satisfy all of the following conditions: gender = female, age > 60, ethnicity = minority group, income level ≤ median and urban residence = true. Then, compare the joint conditional distribution of healthcare access score, number of chronic conditions and annual medical expenditure between the real and synthetic datasets.

- **T5: Outlier and rare category detection**. Determine whether rare patterns or categories in the real data are accurately represented in the synthetic data.

  Example: Identify rare records in the real dataset characterized by uncommon combinations of attributes (e.g., extremely high income, young age, rare job titles, short tenure and rural residence). Then, check whether the synthetic data contains similar multivariate profiles and whether their joint frequencies are consistent with those in the real data.

4 | AISyst

We designed AISyst in collaboration with AI experts from our industrial partner, 4-Xtra Technologies Ltd. We held weekly team meetings in April–September 2025 (as part of a project under the Data Scientist Development Programme at the University of Leeds) to discuss and advance the system design. The AI experts communicated to us the needs and requirements from stakeholders (such as banks and insurance companies). Based on this input, we interactively developed our system, progressively improving its sensitivity and efficiency. The beta-version of AISyst reported in this paper was validated in line with the industry standards and endorsed by the AI experts for further deployment.

In this section, we first provide an overview of AISyst. We then describe how to load the data and select the parameters of the dimensionality reduction (DR) algorithm. Following that, we introduce the four main components for assessing the synthetic data, including multivariate resemblance analysis, bivariate correlation analysis, univariate similarity analysis and AI-assisted structured analysis.

4.1 | Workflow

The workflow of AISYST (see Figure 1) includes a loading data step and four components. AISYST first allows users to load the real and synthetic datasets and then to select the parameters for the chosen DR method to reduce the data multiple dimensions into two dimensions, which are then used to create 2D scatterplots. After that, AISYST enables users to validate the synthetic data through four components: (1) multivariate resemblance analysis, (2) bivariate correlation analysis, (3) univariate similarity analysis and (4) AI-generated structured analysis. Specifically:

1. **Multivariate resemblance analysis** helps users compare the global distribution (task T1) of real and synthetic data through scatterplots generated by DR methods (see Figure 2a2), detect anomalies in specific areas (tasks T4 and T5) and display floating text by linking to other visualizations.
2. **Bivariate correlation analysis** supports the examination of correlations between pairs of variables (task T2). AISYST first visualizes the correlation between each two variables using heatmaps (see Figure 2b1) and the analysis of the relationship between two selected variables using a scatterplot if both are numerical, a beeswarm plot if one is categorical and the other is numerical, or a heatmap if both are categorical or represent a limited number of discrete numerical values (Figure 2b2).
3. **Univariate similarity analysis** allows users to examine differences in individual variables between real and synthetic data. This is done using a table (see Figure 2c1) to identify differences in variable types, and plots to visualize distributional discrepancies (task T3) and outliers (task T5) (see Figure 2c2). For numerical variables, histograms or violin plots are used, while categorical variables are visualized with bar charts.

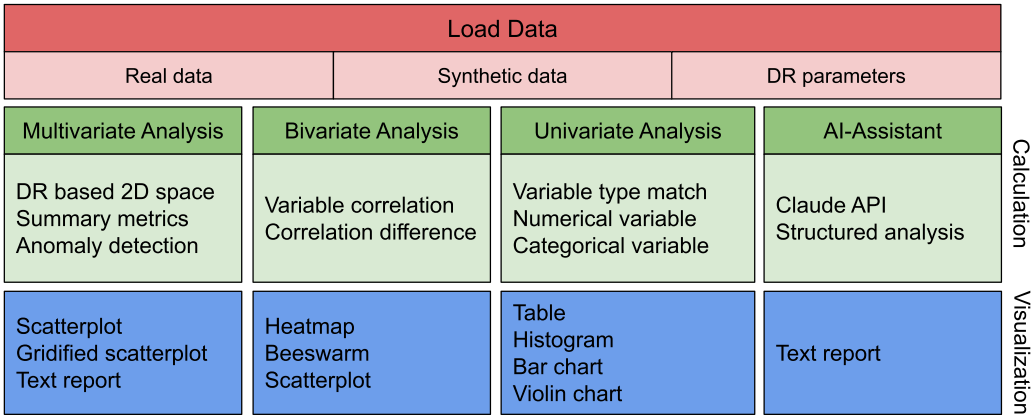

FIGURE 1 | Workflow of AISYST mainly consists of four components sections, based on the uploaded data and selected parameters for the DR algorithm. Each component contains two parts: one for calculating relevant metrics, and the other for visualizing the calculated results based on those parameters.

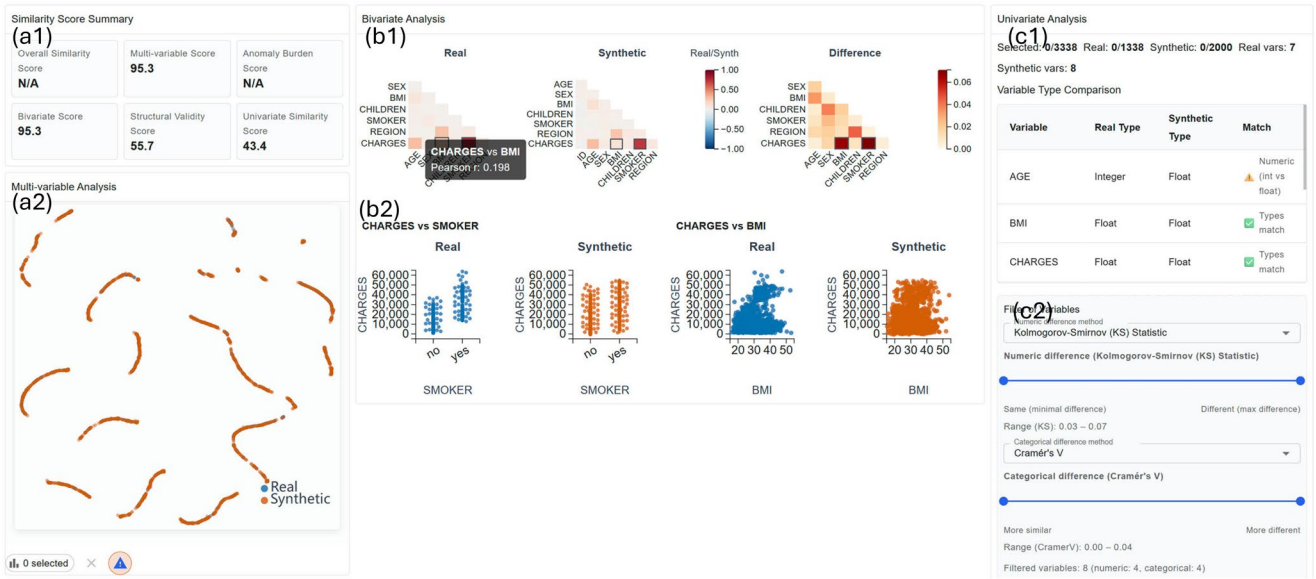

FIGURE 2 | AISYST overview—Visualization. It includes three types of visualizations: A summary score (a1), multivariate resemblance analysis (a2), bivariate correlation analysis (b1) and (b2) and univariate similarity analysis (c1) and (c2).

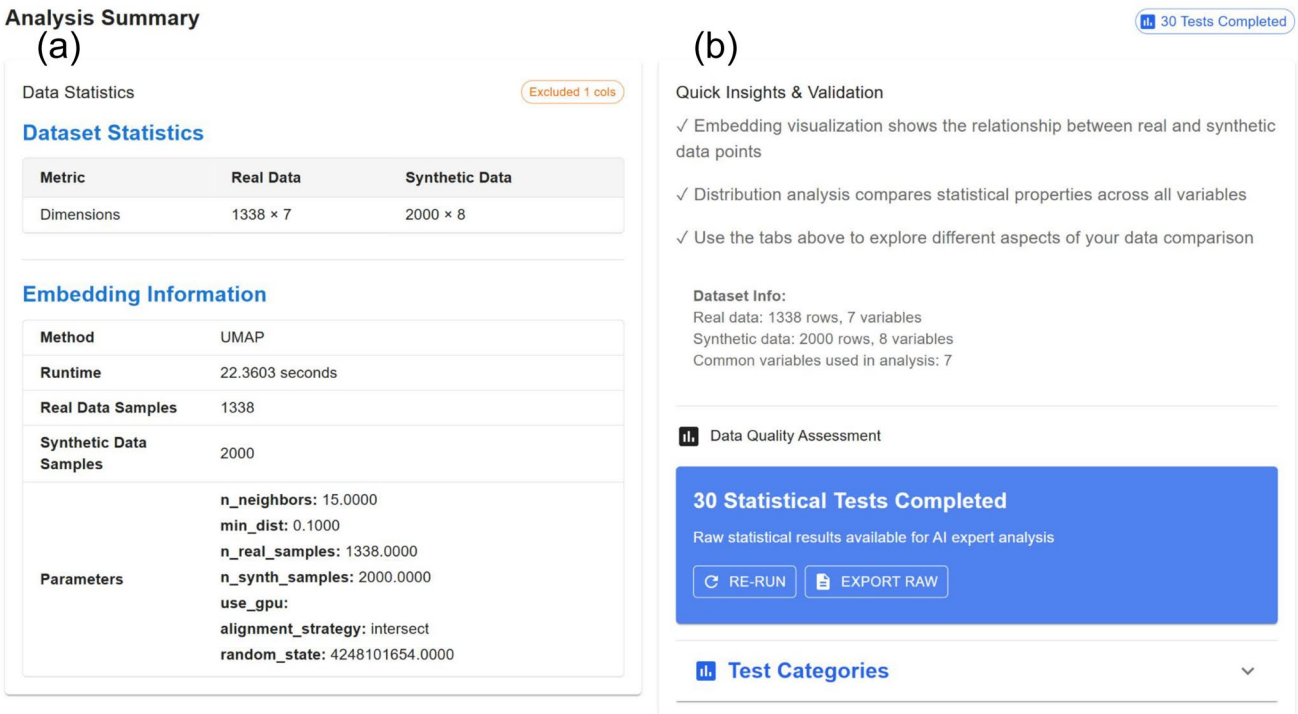

FIGURE 3 | AISYST overview—AI-assistant structured analysis. It includes a summary view that presents key information about the datasets (a) and an AI-generated report (b).

4. **AI assistant** generates a structured PDF report (see Figure 3b) to help users analyse the synthetic data in summary (tasks T3–T5).

The following practical approach may be recommended to the user:

1. Begin by selecting a subset from the multivariate resemblance analysis visualization using either manual selection or automatic anomalous subset detection.
2. Inspect bivariate correlations and refine using univariate similarity analysis.
3. After completing the above loop, capture all relevant issues along with the global comments from the LLM-generated summary report.

This approach, proposed and advocated by our AI experts, enables the user to get a precise answer to where and what the issues are with synthetic data.

4.2 | Data Loading

The data loading consists of two sub-steps. First, both the real and synthetic datasets are loaded, and then a DR algorithm is selected along with its corresponding parameters. This algorithm is used to generate the scatterplot for the subsequent assessment stage.

In this system, we consider two DR methods: *t-SNE* and *UMAP*.

- **t-SNE** (van der Maaten and Hinton 2008) is a nonlinear DR technique particularly well-suited for visualizing high-dimensional data. It operates by converting similarities between data points into joint probabilities and minimizing the Kullback–Leibler divergence between these joint probabilities in the high-dimensional and low-dimensional spaces. While t-SNE is effective at capturing local structures and forming visually distinct clusters, it is sensitive to hyperparameters such as perplexity and can be computationally expensive when applied to large datasets.

- **UMAP** (McInnes, Healy, and Melville 2018; McInnes, Healy, Saul, and Großberger 2018) is another nonlinear DR technique that emphasizes the preservation of both local and global data structures. It constructs a high-dimensional graph representation of the data and then optimizes a low-dimensional projection that maintains this structure. UMAP typically runs faster than t-SNE and scales better to large datasets, making it a practical choice for embedding high-dimensional real and synthetic data for comparative analysis.

4.3 | Multivariate Resemblance Analysis Visualization

A scatterplot in the multivariate resemblance analysis visualization (see Figure 4) provides users with an overview that facilitates the comparison between real and synthetic data. This visualization presents a scatterplot based on two coordinates obtained through the DR process carried out in the loading data step. In this scatterplot, real and synthetic data are represented using two distinct colours, allowing users to visually distinguish between the two datasets and understand how they are distributed within a two-dimensional space.

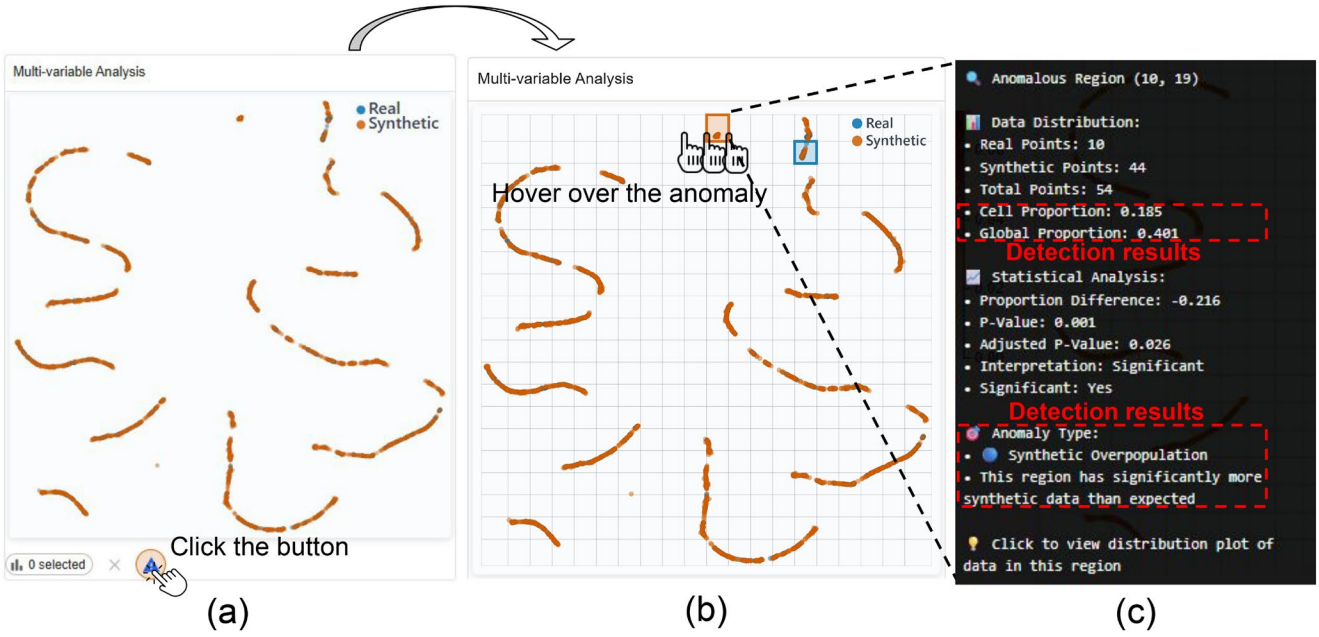

FIGURE 4 | Multivariate resemblance analysis for insurance data: (a) shows a UMAP scatterplot; clicking the triangle transforms it into a gridified view (b) highlighting anomaly regions. Hovering over squares reveals detection results via tooltip (c).

Figure 4a offers valuable insights the relationships of two datasets in a 2D dimensions. For example, it enables users to identify regions that contain exclusively real or synthetic data, as well as regions where both types coexist. Furthermore, it can reveal the presence of outliers (i.e., data points that are positioned far from the primary clusters), thereby helping users assess the structural differences between the real and synthetic data distributions.

To automatically identify anomalous regions and characterize the type of anomaly present in each region, we propose an adaptive histogram-based anomaly detection algorithm operating in the two-dimensional space produced by the dimensionality reduction (DR) method. Let R and S denote the sets of real and synthetic samples, respectively, and let n_R and n_S be their corresponding sizes. The DR projection is partitioned into a set of N grid cells $\{C_1, C_2, \ldots, C_N\}$. The grid resolution is determined adaptively according to the size and spatial distribution of the projected dataset, ensuring that each cell contains a sufficient number of observations for reliable statistical testing while maintaining sensitivity to local distributional differences.

For each cell C_i, we count the number of real and synthetic samples, denoted by r_i and s_i, respectively, and compute the local proportion of real samples as

$$\hat{p}_i = \frac{r_i}{r_i + s_i}. \tag{1}$$

This local proportion is compared against the global baseline proportion

$$p_0 = \frac{n_R}{n_R + n_S}, \tag{2}$$

which represents the expected proportion of real samples under the null hypothesis that real and synthetic data are distributed similarly within the projected space.

To determine whether the observed deviation between $\hat{p}_i$ and p_0 is statistically significant, we perform a binomial proportion test (Howell 2013) for each cell. Specifically, the null hypothesis assumes that the proportion of real samples in the cell equals the global baseline proportion,

$$H_0: \hat{p}_i = p_0, \tag{3}$$

whereas the alternative hypothesis assumes that the two proportions differ,

$$H_1: \hat{p}_i \neq p_0. \tag{4}$$

The test produces a p value for each grid cell, indicating the likelihood of observing the corresponding deviation under the null hypothesis. Because multiple hypothesis tests are conducted simultaneously across all N cells, we apply the *False Discovery Rate (FDR)* correction (Benjamini and Hochberg 1995) to control the expected proportion of false positives. The adjusted p values are ranked and compared against the corresponding FDR thresholds. Cells whose adjusted p values fall below the significance threshold ($\alpha = 0.05$) are designated as anomalous (Figure 4b).

An anomalous cell is further categorized according to the direction of the deviation. If $\hat{p}_i > p_0$, indicating a statistically significant over-representation of real samples, the cell is classified as *real overpopulation* (highlighted with). Conversely, if $\hat{p}_i < p_0$, indicating a statistically significant over-representation of synthetic samples, the cell is classified as *synthetic overpopulation* (highlighted with).

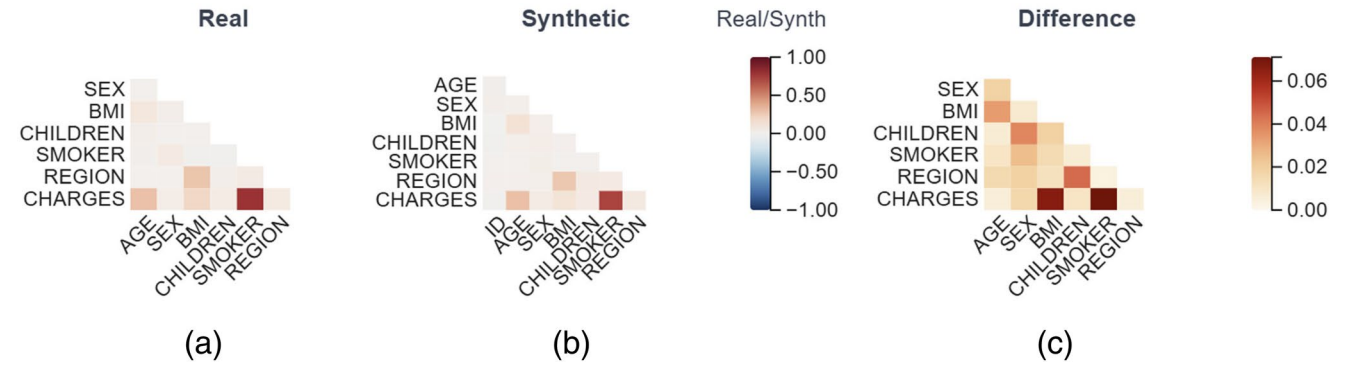

FIGURE 5 | The heatmap view displays the correlations between each pair of variables in the insurance data for both the real and synthetic datasets, as shown in (a, b). Heatmap (c) visualizes the differences in variable correlations between the real and synthetic data.

The computational complexity of the algorithm is $O(n)$ for assigning $n = n_R + n_S$ projected samples to grid cells, followed by $O(N)$ for performing the cell-wise statistical tests and $O(N \log N)$ for the FDR correction. Therefore, the overall complexity is

$$O(n + N \log N), \tag{5}$$

making the method scalable to large synthetic datasets. The adaptive grid construction further improves robustness by balancing spatial resolution and statistical reliability across datasets with different sizes and distributions.

In addition, we provide a tooltip showing the detection results in each region if it is anomalous and present them in Figure 4c, which provides a detailed summary of the region's data distribution, statistical analysis, anomaly type and other relevant metrics. The data distribution section includes the number of real data points, synthetic data points and the total number of points within the region. It also reports the cell proportion and the global proportion: the cell proportion refers to the proportion of real data within the specific region, while the global proportion indicates the proportion of real data across the entire dataset. The statistical analysis section presents the difference between the cell and global proportions, along with additional statistical metrics such as the p value, adjusted p value, interpretation and statistical significance. Lastly, the anomaly type assigned to the region is displayed at the bottom of the summary.

4.4 | Bivariate Correlation Analysis Visualization

In the bivariate correlation analysis section, we first examine the correlations between each pair of variables in the real and synthetic datasets, visualizing them using two heatmaps in Figure 5a,b, corresponding to the real and synthetic data, respectively. We then compute the absolute difference between the two heatmaps by calculating the element-wise differences for each corresponding cell and visualize the result as an additional heatmap, shown in Figure 5c.

The dataset contains both categorical and numerical variables. Therefore, to include all variable pairs in a single heatmap we apply different correlation computation methods depending on the variable types: **num–num**, **cat–cat** and **num–cat**.

For **num–num** pairs, we applied the *Pearson correlation coefficient* (Rodgers and Nicewander 1988), which quantifies the linear relationship between two continuous variables. Alternative methods such as *Spearman's rank correlation coefficient* (Sedgwick 2014) may also be used, particularly when the relationship is monotonic but not necessarily linear. For **cat–cat** pairs, we used *Cramér's V* (Cramér 1946), a Chi-squared-based statistic that measures the strength of association between two categorical variables. An alternative approach is *mutual information* (Kraskov et al. 2004), which captures non-linear dependencies without assuming a specific distribution. For mixed **num–cat** pairs, the *correlation ratio* η (Fisher 1925) was computed to quantify the proportion of variance in the numerical variable explained by the categorical grouping. Alternatives such as the *point-biserial correlation* (Kornbrot 2005) may be considered when the categorical variable is binary. These measures produce a unified correlation matrix that captures relationships among both numerical and categorical variables, allowing a heatmap to provide an integrated view of variable dependencies across the entire dataset.

To provide a deeper understanding of the relationships among variables of different types, a user may click on a specific cell in the heatmap (in our AISYST tool) to view detailed information (see Figure 6a).

For **cat–cat** pairs (we regard discrete variables with fewer than five unique values as categorical variables), heatmaps are constructed based on contingency tables to display the frequency or strength of association between category combinations (Figure 6b shows a heatmap of `SEX` (including female and male) vs. `REGION` (including four values: northeast, northwest, southeast and southwest)). This example illustrates differences in correlations between the real and synthetic data. In the real data, the number of males in the southeast region is higher than the number of females; however, this is not the case in the synthetic data.

For **num–num** pairs, scatterplots are used to depict linear or nonlinear relationships between continuous variables, allowing users to observe distribution patterns, trends and potential outliers; for example, Figure 6c shows a scatterplot of `BMI` versus `CHARGES`. In this example, the scatterplot reveals differences in distribution patterns between real and synthetic data. In the real data, the two variables exhibit a two-cluster distribution pattern; however, this pattern is not present in the synthetic data.

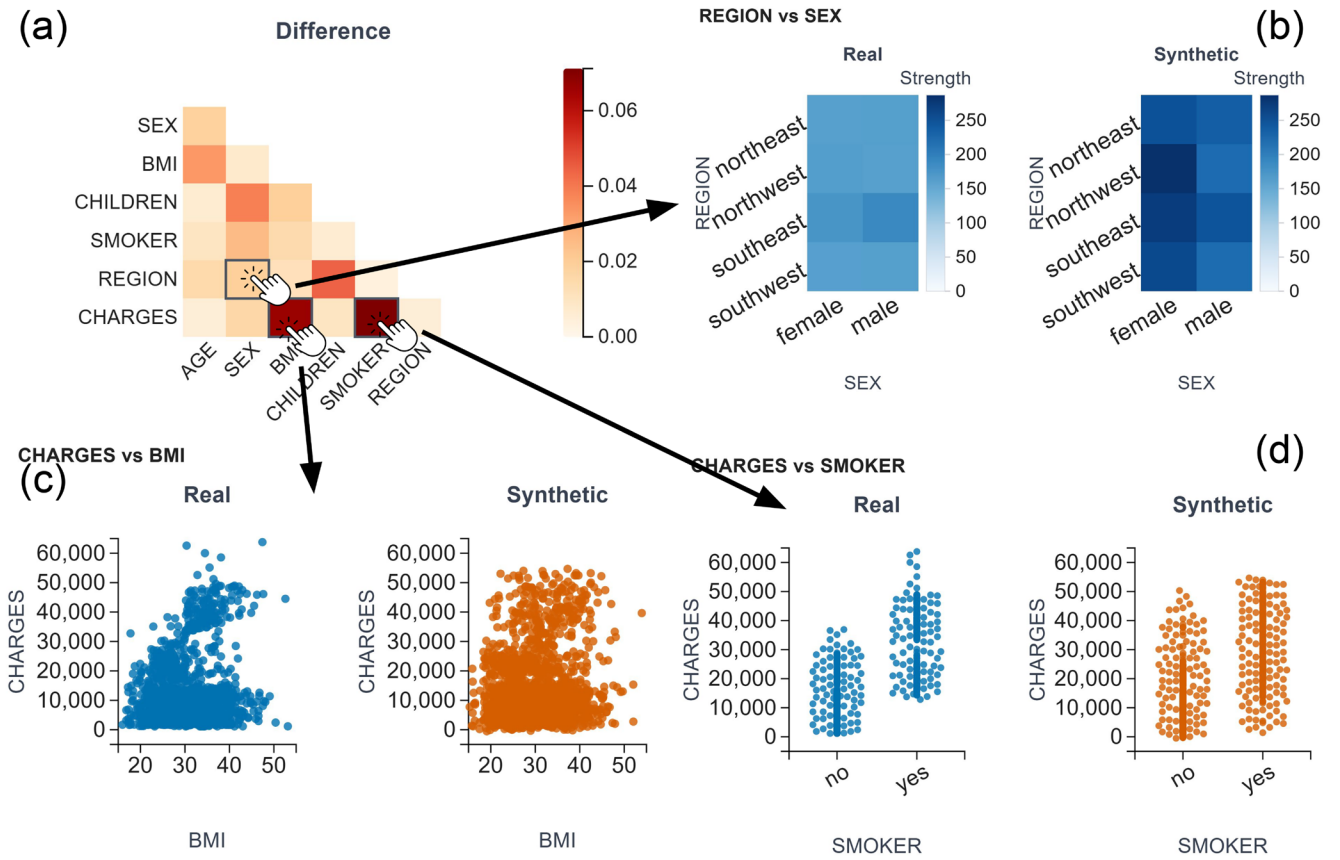

FIGURE 6 | Correlation plots for selected variable pairs in insurance data. By clicking a specific cell in heatmap (a), the correlation is visualized using: Heatmaps (b) for **cat–cat** pairs, scatterplots (c) for **num–num** pairs and beeswarm plots (d) for **num–cat** pairs.

For **num–cat** pairs, beeswarm plots are utilized to illustrate the distribution of numerical values across different categorical groups, revealing how the numerical variable varies within and between categories; for example, Figure 6d shows a beeswarm plot of SMOKER versus CHARGES. This example demonstrates the differing distributions of CHARGES across smokers and non-smokers. In the real data, smokers have much higher CHARGES values, a pattern that does not occur in the synthetic data.

Together, the three visualization techniques enable a unified and flexible framework for exploring correlations across mixed data types, facilitating a more comprehensive interpretation of the relationships present in both real and synthetic datasets.

4.5 | Univariate Similarity Analysis Visualization

The univariate similarity analysis is designed to help users understand the distribution of individual variables and assess whether a subgroup of data instances exhibits a distribution similar to that of the entire dataset. The type of distribution plot used depends on the variable type. If the variable is categorical or a discrete numerical variable with a limited number of unique values, a bar chart is used. For continuous numerical variables, the distribution is displayed using either a histogram, focusing on the comparison of distribution shapes, or a violin plot, which emphasizes differences in the range and scale of values. Although a box plot can serve as an alternative to the violin plot, AISYST does not support box plots in the current version since the box plots hide distributional structure that violin plots reveal.

Figure 7 presents examples of these visualizations. Figure 7a,b shows a histogram and a violin plot, respectively, which illustrate the distribution of a continuous numerical variable. The histogram effectively displays the overall distribution of the data; for example, Figure 7a shows that in the real data, there are more instances around age 18–20 than at other ages. Compared to the histogram, the violin plot provides more detailed information about the distribution's shape and the data range. For instance, Figure 7b shows that the real data contain higher values in the variable CHARGES (medical insurance costs billed to customers). In addition, for categorical variables, the distribution is visualized using a bar chart, as shown in Figure 7c.

In Figure 7a, the histogram reveals substantial differences in total counts between the real and synthetic data. To address this and to enable a fair comparison of distributions, we change the y-axis by replacing absolute counts with densities, which normalizes the area under each distribution curve to 1, regardless of the sample size. This allows users to compare the shape of the distributions rather than their raw frequencies—an essential consideration when sample sizes vary between real and synthetic datasets. The effect of this adjustment is illustrated in Figure 7d. This transformation enables a more meaningful comparison by removing the influence of sample size disparities.

When dealing with a small number of variables, it is feasible to examine the distribution of each one individually. However, when there are 100 or more variables, it becomes impractical for users to manually inspect each of them. To address this, the univariate similarity analysis visualization includes a panel for filtering variables for further analysis, as shown in Figure 8. AISYST provides two metric sections for filtering either numerical or categorical variables (see Figure 8a). For numerical variables, we include six metrics: the Kolmogorov–Smirnov statistic (Kaur et al. 2021), difference in mean, difference in standard deviation, difference in median, difference in minimum and difference in maximum. These differences, based on basic statistical measures, are computed as the absolute differences between the real and synthetic data for each variable. When a metric is

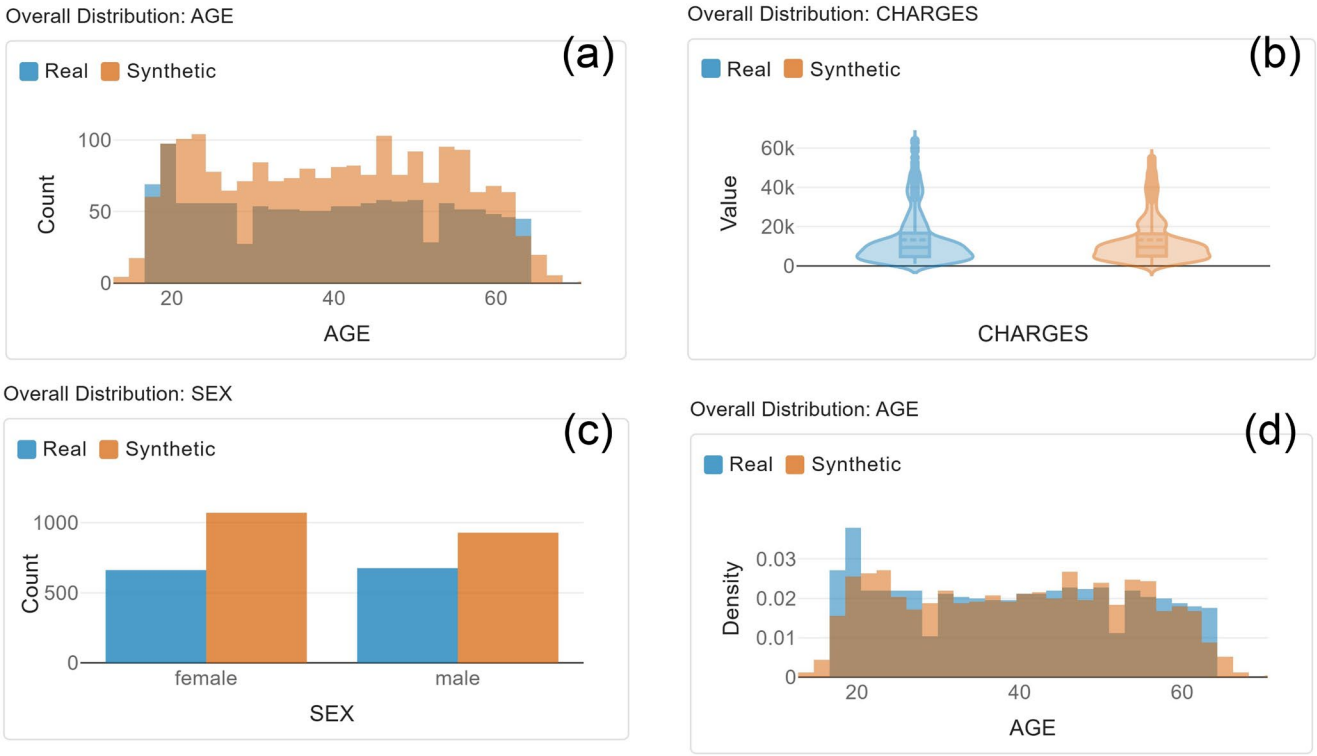

FIGURE 7 | Univariate distribution comparisons between real and synthetic insurance data. Histograms (a) and violin plots (b) illustrate the distribution of continuous numerical variables, while bar charts (c) compare categorical and discrete numerical variables. Normalizing the *y*-axis in histograms (d) enables more meaningful comparisons between real and synthetic data.

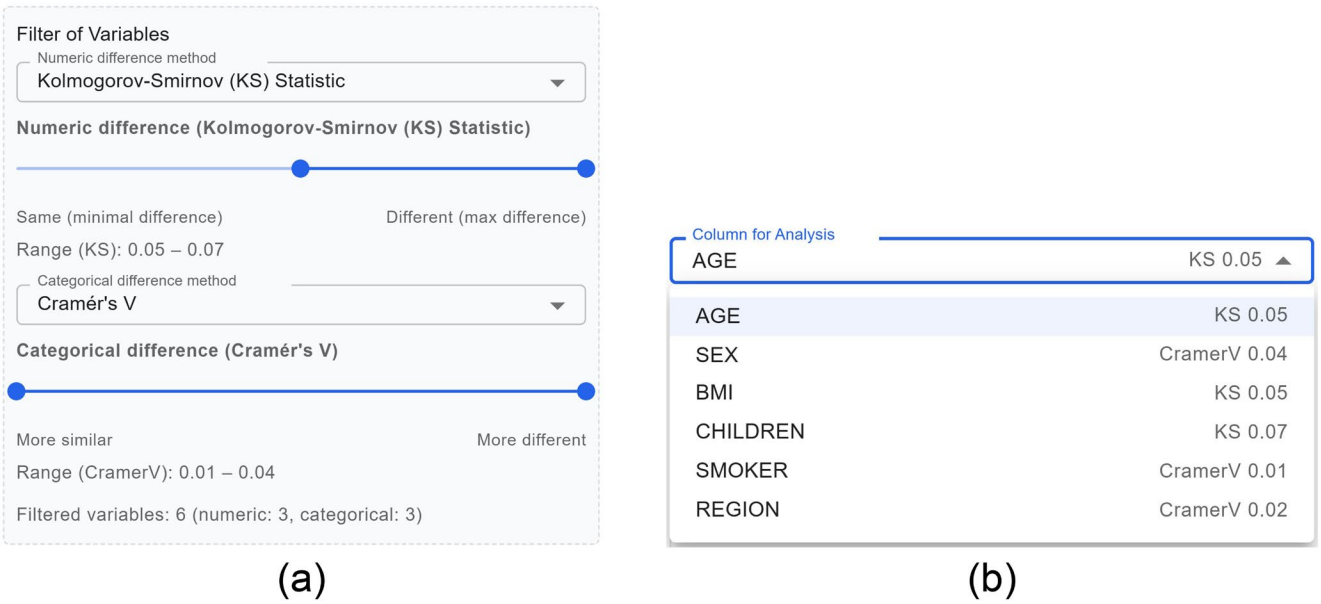

FIGURE 8 | A panel allows filtering variables based on user-specified difference metrics in (a), with results shown in (b). In (a), the KS range for numerical variables is set to 0.05–0.07, filtering out those below 0.05.

selected, a slider allows users to specify a range of values for filtering variables based on the selected metric. Similarly, differences in categorical variables can be assessed using Cramér's V (Cramér 1946) and the categorical Wasserstein distance (Rubner et al. 2000). Based on the selected metric and its specified range, users can identify and select potentially problematic variables, thereby reducing the cognitive load associated with analysing univariate similarity.

AISYST also provides a table that shows whether the variables in the synthetic data match those in the real data. Generally, there are three main types of possible mismatch: (1) a missing column, where either the real or synthetic dataset lacks a variable that is present in the other; (2) a numeric type mismatch, where two different numerical types are used for the same variable (e.g., `integer` vs. `float`); and (3) a data type mismatch, where a variable's type in the synthetic data differs from its type in the real data (e.g., `integer` vs. `categorical`). Figure 9 shows an example of a variable type mismatch table for the real and synthetic data from the insurance case study; here, a detected mismatch is that variable `AGE` is `integer` in the real data but `float` in the synthetic counterpart.

4.6 | Subgroup Analysis

The multivariate resemblance analysis visualization also allows users to select specific groups of data instances from the scatterplot. The selected data instances are then highlighted in the bivariate correlation analysis visualization, and their distribution is displayed in the univariate similarity analysis visualization. There are two ways to select subgroups of data instances in the multivariate scatterplot: either by clicking on the anomaly regions shown in Figure 4b or by using a lasso tool, as shown in Figure 10a.

Figure 10 illustrates the interactions that select the subgroup of data instances from the scatterplot in Figure 10a using a lasso tool. After selecting this specific subgroup, a new plot is triggered to visualize the distribution of the selected data instances, based on the chosen variable and its type. Figure 10b shows the distribution of the `AGE` variable for the selected data instances. It reveals that, for real data instances, most individuals are around 18 years old. In contrast, the synthetic data within this subgroup displays a more diverse distribution of `AGE` values.

Additionally, the selected data instances are highlighted in the bivariate similarity analysis visualization, specifically within the correlation plots of variable pairs. These highlighted instances are displayed in scatterplots or beeswarm plots. In this case, the two beeswarm plots correspond to `SEX` versus `CHILDREN` in Figure 10c and `REGION` versus `CHILDREN` in Figure 10d, where the variable `CHILDREN` refers to the number of dependent children covered by (or living with) the insured individual or policyholder. It can be observed that all selected data instances in the real dataset are male and from the Southeast region; however, this pattern differs notably in the synthetic data.

4.7 | Summary Score

We adopt a composite similarity score to evaluate the fidelity of synthetic data relative to an original dataset (see Figure 2a1). The overall score $S_{\text{overall}} \in [0, 100]$ is defined as a weighted combination of four components:

$$S_{\text{overall}} = w_{\text{struct}} S_{\text{struct}} + w_{\text{uni}} S_{\text{uni}} + w_{\text{bi}} S_{\text{bi}} + w_{\text{anom}} S_{\text{anom}}, \tag{6}$$

where all component scores are scaled to $[0, 100]$ and the weights sum to 1. A reasonable default choice is

$$w_{\text{struct}} = 0.25, \quad w_{\text{uni}} = 0.30, \quad w_{\text{bi}} = 0.30, \quad w_{\text{anom}} = 0.15. \tag{7}$$

Structural validity (S_{struct}): This component evaluates whether the synthetic dataset preserves the schema of the original

Variable Type Comparison

Variable	Real Type	Synthetic Type	Match
AGE	Integer	Float	⚠ Numeric (int vs float)
BMI	Float	Float	✅ Types match
CHARGES	Float	Float	✅ Types match
CHILDREN	Integer	Integer	✅ Types match
REGION	Categorical	Categorical	✅ Types match

FIGURE 9 | A table showing variable type matches for insurance data, with one mismatch: `AGE` is an integer in the real data but a float in the synthetic data. More variables can be viewed by scrolling in AISYST.

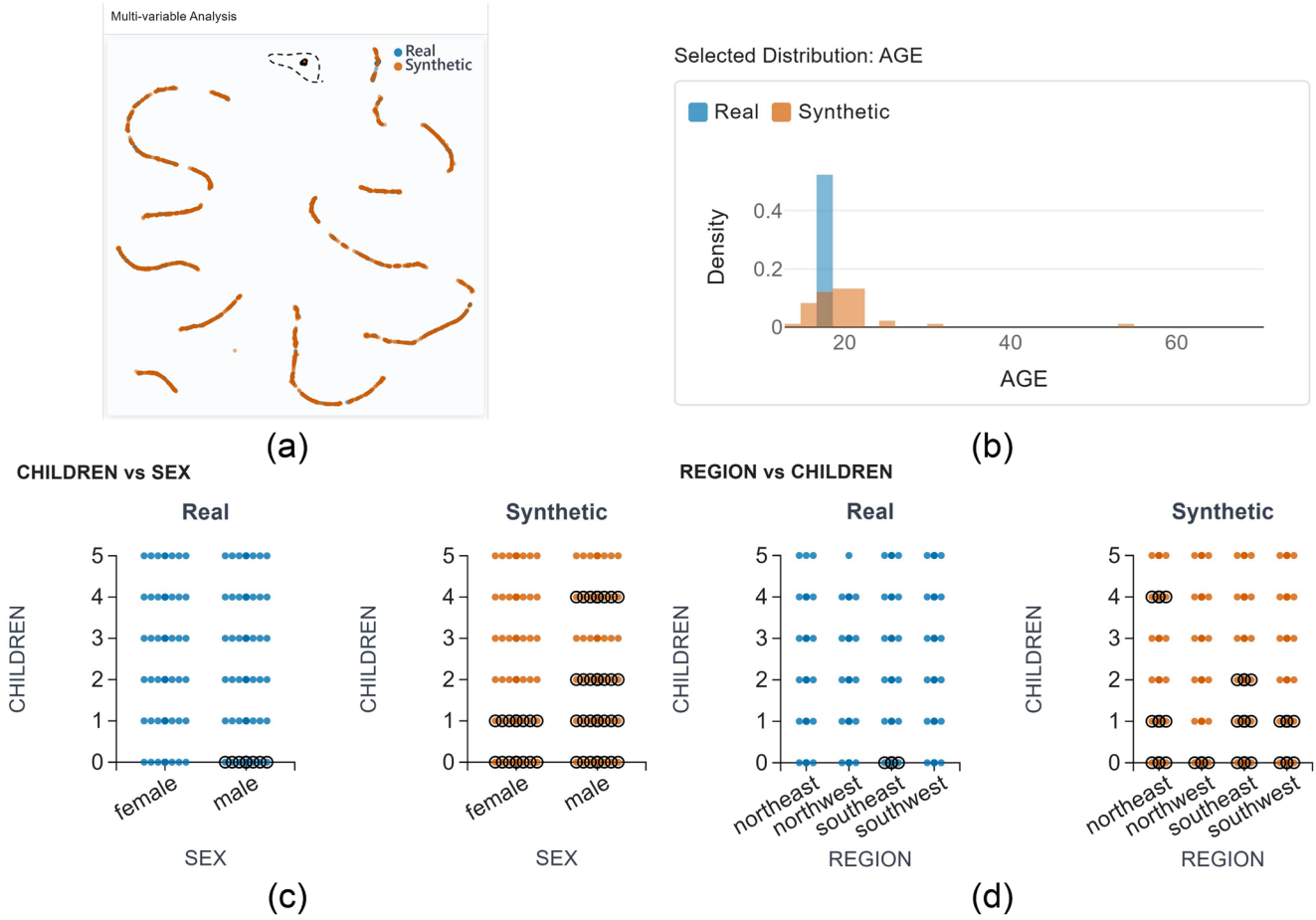

FIGURE 10 | Subgroup analysis for insurance data: (a) shows subgroup selection; (b) displays the `SEX` distribution as a histogram. The selected points are also highlighted in the bivariate views (c) and (d).

data, including consistency in columns, data types, categorical support and valid ranges. It is defined using a penalty-based formulation:

$$S_{\text{struct}} = 100 \cdot \max\left(0, 1 - \sum_{j=1}^{k} \alpha_j e_j\right), \tag{8}$$

where $e_j \in [0,1]$ are normalized structural error terms and α_j are their weights with $\sum_j \alpha_j = 1$.

Univariate Similarity (S_{uni}). This component measures how well marginal distributions of individual variables are preserved. For each variable, a discrepancy $d_i \in [0,1]$ is computed. For numerical variables, this combines the Kolmogorov–Smirnov statistic with normalized differences in mean, standard deviation and median. For categorical variables, total variation distance is used. The overall discrepancy is

$$D_{\text{uni}} = \frac{1}{p} \sum_{i=1}^{p} d_i, \tag{9}$$

and the score is

$$S_{\text{uni}} = 100 \cdot \left(1 - D_{\text{uni}}\right). \tag{10}$$

Bivariate Similarity (S_{bi}). This component evaluates the preservation of pairwise relationships between variables. Let $A_{ij}^{(\text{orig})}$ and $A_{ij}^{(\text{syn})}$ denote the associations between variables i and j in the original and synthetic datasets, respectively. The discrepancy is defined as

$$D_{\text{bi}} = \frac{\sum_{i<j} \beta_{ij} \left| A_{ij}^{(\text{orig})} - A_{ij}^{(\text{syn})} \right|}{\sum_{i<j} \beta_{ij}}, \tag{11}$$

and the corresponding score is

$$S_{\text{bi}} = 100 \cdot \left(1 - D_{\text{bi}}\right). \tag{12}$$

The weights β_{ij} can be uniform or reflect the importance of relationships.

Anomaly Burden (S_{anom}). This component captures localized discrepancies between datasets. Let $\mathcal{A}$ denote the set of anomalous regions. For each region $r \in \mathcal{A}$, define a weight ω_r (proportional to its sample size) and a severity s_r (measuring imbalance between datasets). The anomaly burden is

$$D_{\text{anom}} = \sum_{r \in \mathcal{A}} \omega_r s_r, \tag{13}$$

and the score is

$$S_{\text{anom}} = 100 \cdot \left(1 - D_{\text{anom}}\right). \tag{14}$$

Overall, this framework provides an interpretable and flexible metric: the aggregate score offers a concise summary, while individual components reveal specific sources of discrepancy.

4.8 | AI-Assistant Structured Analysis

AISYST integrates with an AI component to provide structured contextual interpretation by generating a PDF report that summarizes the differences between real and synthetic data; Figure 11 shows an example of such a report. We designed the AI component to act as a senior data quality expert by prompting it to deliver a comprehensive analysis based solely on the validation results and underlying data. The prompt requests detailed sections including an executive summary, key findings, statistical quality, practical usefulness, critical issues and 3–5 actionable recommendations, each with a maximum word requirement. Additionally, the agent was asked to assess the overall risk level (LOW, MEDIUM or HIGH), with a justification of at most 50 words. The prompt explicitly instructed the AI to avoid fabricating information, to justify all findings and recommendations using the provided data, and to include *p* values and references to the statistical tests used when applicable. Importantly, all statistical results are pre-computed using deterministic equations rather than generated by the AI.

The output generated through the Claude API is ultimately summarized as a narrative and exported as a downloadable PDF report (see Figure 11). This approach reduces the cognitive burden on users by transforming raw statistical outputs into human-readable explanations that highlight key differences and potential risks.

FIGURE 11 | An example of an AI-generated report produced by the AI assistant for analysing the insurance data.

5 | Implementation

We have implemented AISYST using Python for the back end and JavaScript for the front end. The back end performs DR methods to project high-dimensional data onto two dimensions for the subsequent generation of scatterplots. The front end provides the user interface that allows users to explore the data in detail. It is built using REACT.JS (Fedosejev 2015), a JavaScript library for building interactive and component-based user interfaces, and D^3 (Bostock et al. 2011), a library for creating dynamic and data-driven visualizations using web standards such as SVG, HTML and CSS. For the AI-assistant functionality, we integrated an LLM CLAUDE 3.5 HAIKU (Claude 2024), enabling our system to generate a readable analysis report.

6 | Evaluation

To evaluate the effectiveness of our AI-powered interactive visual system, we conducted three case studies across different domains: insurance, income and health. Through these use cases, we demonstrated how users can gain insights by employing AISYST and interacting with the various analytical visualizations provided by the system.

In all three case studies, synthetic datasets were supplied by 4-Xtra Technologies Ltd. using their proprietary generation algorithms. The structure and format of variables in the real data input were preserved in the synthetic counterparts but an additional column named `ID` was created to give the entries their unique labels (produced via a random permutation of the numerals from 1 to n, where n is the total number of synthetic rows). Naturally, the `ID` column was excluded from DR algorithms.

As an important part of our evaluation protocol, we collected feedback from three AI experts **E1**, **E2** and **E3** from 4-Xtra Technologies Ltd., who have extensive experience in synthetic data generation. We organized a meeting with the experts, which began with a 10-min introduction to AISYST, followed by a 60-min demonstration illustrating how to use AISYST to assess synthetic data across the three case studies. During the case studies' presentation, the experts were encouraged to ask questions and provide comments. The meeting concluded with a 30-min open discussion on the strengths and weaknesses of AISYST.

6.1 | Insurance Case Study

The datasets used in this case study comprise insurance-related data. The real dataset (Putten 2000) contains genuine policyholder records collected from an insurance company, including attributes such as customer demographics and claim history. The synthetic dataset was generated to mimic statistical characteristics and structural relationships of the real data while preserving confidentiality. The real insurance dataset contains 1338 records; the synthetic dataset mirrors the same structure, with 2000 records.

The first step in assessing the synthetic data is to upload both the real and synthetic datasets and select the corresponding parameters for the DR method. We generated low-dimensional embeddings using the UMAP method. The number of neighbours was set to 15, providing an appropriate trade-off between capturing local detail and preserving global structure. The minimum distance was set to 0.1, allowing moderate clustering of points in the embedded space while avoiding excessive crowding. These parameter settings produced stable and interpretable two-dimensional projections suitable for comparative visualization and analysis. Based on the DR method, we derived 2D coordinates to represent the multiple variables.

After uploading the data and applying the DR method, we initialized AISYST. We proceeded to the multivariate resemblance analysis visualization. The summary score provides an overall quality assessment of the synthetic dataset, as shown in Figure 2a1. The overall similarity score is 63.0, indicating that the quality of the synthetic insurance data is insufficient for reliable use. Additional metrics further describe the quality of the synthetic data. Then, AISYST generated a scatterplot (see Figure 4a) using the 2D coordinates obtained from UMAP to compare the real and synthetic data (task T1). Next, we created gridified regions to further analyse anomalous areas by clicking the anomaly detection button (see Figure 4a). The resulting gridified scatterplot is shown in Figure 4b, which highlights two anomalous regions: one indicating synthetic overpopulation, marked with ☐, and the other indicating real overpopulation, marked with ☐. By hovering over the synthetic overpopulation region, a tooltip appears that summarizes key insights from the anomaly detection results, as shown in Figure 4c. Specifically, it reports that the cell proportion is 0.185, while the global proportion is 0.401 (calculated as the number of real data points divided by synthetic data points). Therefore, this region is identified as a case of synthetic overpopulation, indicating significantly more synthetic data than expected in the hovered region (task T5). Expert **E1** commented that this was especially evident in the automated detection of anomalies in multivariate resemblance visualizations using the gridified method. This approach helped identify specific subsets of the datasets for further exploration by automatically detecting the proportion of real and synthetic data in each region.

Having concluded that the cluster, identified as synthetic overpopulation (see Figure 4b), appeared to be an outlier, we were curious about the features of the data instances within this cluster. We selected the cluster shown in Figure 10a, and the corresponding data instances were highlighted in the bivariate correlation analysis visualization, as shown in Figure 10c,d. Their distribution for the variable `AGE` is presented in Figure 10b: for the selected data instances, all real data points are approximately 18 years old, whereas the synthetic data instances show a more diverse age range (task T4). Figure 10c reveals that all selected real data instances are male and have no children. In contrast, the synthetic data instances in this cluster include both male and female individuals, and their `CHILDREN` values are more diverse. A similar pattern is observed in the `REGION` distribution shown in Figure 10d: the real data instances in the selected cluster are exclusively from the southeast and have no children, while the corresponding synthetic data instances are more widely distributed across regions and child counts.

Next, we examined the bivariate correlation visualization by exploring several cells corresponding to variable pairs, with a particular focus on those involving the AGE and CHARGE variables, as well as cells in the difference heatmap that showed larger discrepancies. This analysis revealed several insights, as shown in Figure 6, where three cells were selected for further examination. As mentioned in Section 4.4, Figure 6b–d illustrates findings from the correlations between individual variable pairs. Expert **E1** commented that it is beneficial to observe the correlation between each pair of variables using different plots. These plots help illustrate the concentration of the data distribution. However, it would also be useful to provide visualizations that show certain proportions in the distribution—for example, how many data instances represent males from the southeast region.

We then examined the table for mismatched variable types, as shown in Figure 9. The table indicated that the AGE variable is not matched between the synthetic and real data: the synthetic data is of type float, whereas the real data is of type integer (task T3). Expert **E1** commented here that it was because there was no post-processing of the data after training. Sometimes, it is acceptable to maintain these data as they are, but in other cases, post-processing is necessary to ensure that the data type strictly matches that of the real data. Next, we reviewed the overall distributions of all variables and found that the CHARGE values in the real data are significantly higher than those in the synthetic data (see Figure 7b). This suggests that the synthetic data did not successfully simulate the extreme values of the CHARGE variable (task T3). Expert **E1** commented that it is good that we can observe the different distribution between real and synthetic data in a simple way, which can help users develop a basic sense of the data quality. Therefore, in the following analysis, special attention should be given to these two variables.

AISYST provided the AI-assistant with a structured analysis of the insurance dataset (see Figure 11). Through the report, we found 7 out of 7 columns failed the distribution similarity assessments. Additionally, the report recommended five actionable suggestions to improve the synthetic data, such as enhancing the balance between privacy and utility by implementing differential privacy techniques. The reader is referred to the file ai-validation-report-Insurance-data.pdf in the Supporting Information (https://github.com/llqsee/SynAccess/tree/main/data) for more information about the AI assistant and a structured analysis report. Expert **E2** commented that the AI assistant relies on third-party models, which do not always generate reliable results. Therefore, it could be used to summarize findings, but should not be regarded as authoritative.

6.2 | Income Case Study

The real dataset used in this case study is a survey-based socioeconomic dataset containing individual-level records with demographic, occupational and income-related attributes, which was provided by 4-Xtra Technologies Ltd. It contains authentic individual records, including information such as sex, age and other demographic variables. The synthetic dataset was generated to replicate the statistical patterns and structural relationships observed in the real data while ensuring data confidentiality. The real dataset consists of 5000 records, whereas the synthetic dataset has 10,000 records.

Similarly to the insurance case, we first loaded the real and synthetic datasets, along with the corresponding DR parameters, to generate 2D coordinates for visualizing the high-dimensional information in a two-dimensional space. In this case, we used the same DR algorithm and parameters. After loading the data, we analysed the bivariate correlation analysis visualization using the difference heatmap (see Figure 12). We examined the areas with the highest differences and selected the variable pair 12CONSCIOUS and 5CREDIT, as this combination exhibited the most significant disparity. Upon selection, we obtained a detailed view of the correlation between these two variables. Since both are discrete numerical variables with less than 5 unique values, their relationship is visualized using heatmap.

From the heatmap, we observed that in the real data whenever 5CREDIT equals 0, 12CONSCIOUS is always 0. In contrast,

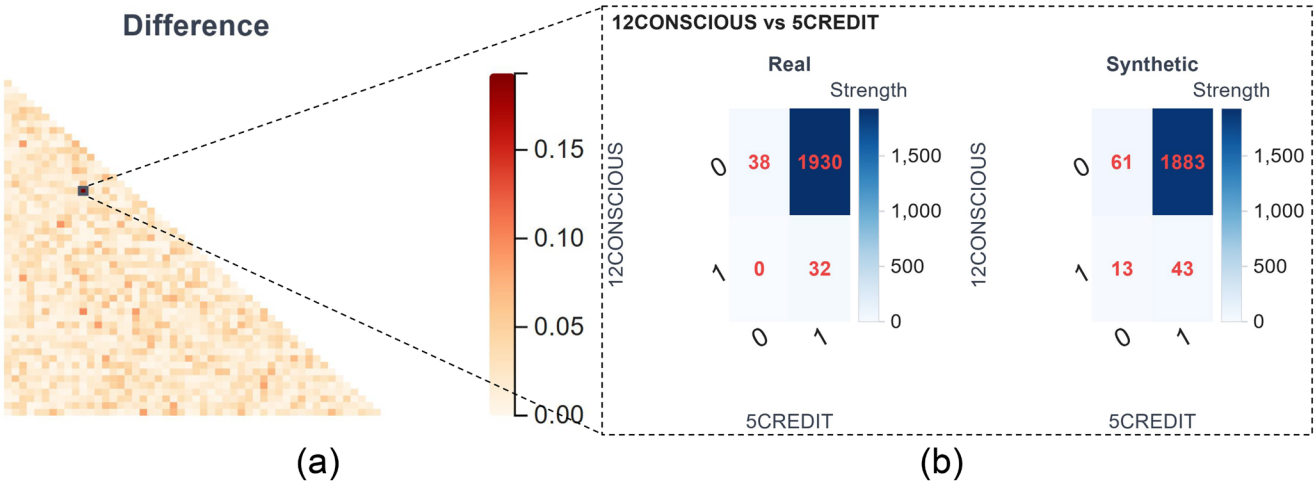

FIGURE 12 | The bivariate correlation analysis of the income data highlights the variable pair 5CREDIT versus 12CONSCIOUS (b), selected based on the difference heatmap (a). In the real data, when 5CREDIT equals 0, 12CONSCIOUS cannot equal 1; however, this combination occurs in the synthetic data (b).

in the synthetic data, there are 13 instances where `5CREDIT` equals 0 and `12CONSCIOUS` equals 1. This indicates a discrepancy in the dependency between the two variables. This discrepancy likely contributes to the observed difference in correlation patterns between the datasets (task T2). In this case, does such a difference between the real and synthetic data matter? Expert **E1** noted that this may depend on domain knowledge to determine whether the discrepancy is meaningful. Nevertheless, our tool can assist by providing relevant information—not only about the co-occurrence of variable values but also about their proportions and density distributions.

There are too many variables to examine individually. Therefore, we applied a metric to filter the variables that exhibited greater differences between the real and synthetic data. Specifically, we selected the KS statistic to filter the numerical variables, focusing on those with values ranging from 0.15 to 0.18, as shown in Figure 13a. This filtering step resulted in only two numerical variables: `10DAILYTRAVEL` (daily travel frequency) and `14SUM` (indicating savings or account balance). Expert **E1** noted that it is a useful approach to first filter certain variables using specific metrics, such as the KS statistic for numerical variables. That avoids users needing to manually examine each variable, which is highly time-consuming.

We selected `10DAILYTRAVEL` for further exploration, as shown in Figure 13b. The variable `10DAILYTRAVEL` is stored as an integer in the real data but as a float in the synthetic data. We observed differences in their distributions using a violin plot. The real data displays a broader range, with values reaching approximately 20, whereas the synthetic data only extend to around 14 (task T3).

6.3 | Diabetes Case Study

The datasets used in this study comprise patient-level medical records related to diabetes diagnostics. The real dataset (NIDDK 2021) contains authentic clinical measurements collected from 768 female patients, with the output value 1 or 0 depending on whether the patient tested positive for diabetes, and with 8 numerical variables (features): `Pregnancies` (number of times pregnant), `Glucose` (glucose concentration in an oral glucose tolerance test, mg/dL), `BloodPressure` (diastolic blood pressure, mmHg), `SkinThickness` (triceps skin fold thickness, mm), `Insulin` (2-h serum insulin, μU/mL), `BMI` (body mass index, kg/m^2), `DPF` (diabetes pedigree function) and `Age` (years). The synthetic dataset (expanding to 10,000 records) was generated to replicate statistical characteristics and structural relationships observed in the real data while ensuring full anonymity.

Similarly to the previous two case studies, we first uploaded the real and synthetic datasets and selected the corresponding parameters for the DR method. We used the same parameters as in the previous two case studies. We then focused on univariate similarity analysis by examining the table comparing variable types and distributions for each variable. We found that only three variables—`BMI`, `DPF` and `Outcome`—had matching data types between the real and synthetic datasets. One issue was a mismatch in variable types for `Pregnancies`, which appeared as a `float` in the synthetic dataset but as an `integer` in the real dataset. Expert **E1** explained that the synthetic dataset was missing a final step to align data types between the real and synthetic datasets. Specifically, the `Pregnancies` variable should have been converted from `float` to `integer`.

We also looked at the variables' distributions. It was observed that the value ranges for `SkinThickness` in the synthetic dataset were narrower than in the real dataset. In contrast, the range of `DPF` in the synthetic dataset was wider than in the real dataset. One of the most critical insights from the univariate similarity analysis was that three variables—`Glucose`, `BMI` and `BloodPressure`—included values below 0 in the synthetic dataset that is impossible in the real context (task T5). Figure 14b–d shows the violin plots visualizing the distributions of these variables. Notably, `BloodPressure` is one of the variables with matched data types, yet it still exhibits unrealistic negative values in the synthetic data (task T3).

We further investigated variables with values below 0—`BloodPressure` and `BMI`—as shown in Figure 15a,b. Figure 15b presents two scatterplots for the real and synthetic data using these two variables. We observed that, in the real data, both `BMI` and `BloodPressure` contain values that are exactly equal to 0. In contrast, the synthetic data includes values approximately near 0, but not exactly 0. Our AI assistant identified the value 0 in the real dataset as a potential anomaly and suggested that it could be due to data collection issues or represent missing values. For more details, please refer to section *Real Data Quality*

FIGURE 13 | Univariate similarity analysis of the income data: (a) numerical variables filtered according to the KS statistic, (b) one filtered variable revealing a difference in scale.

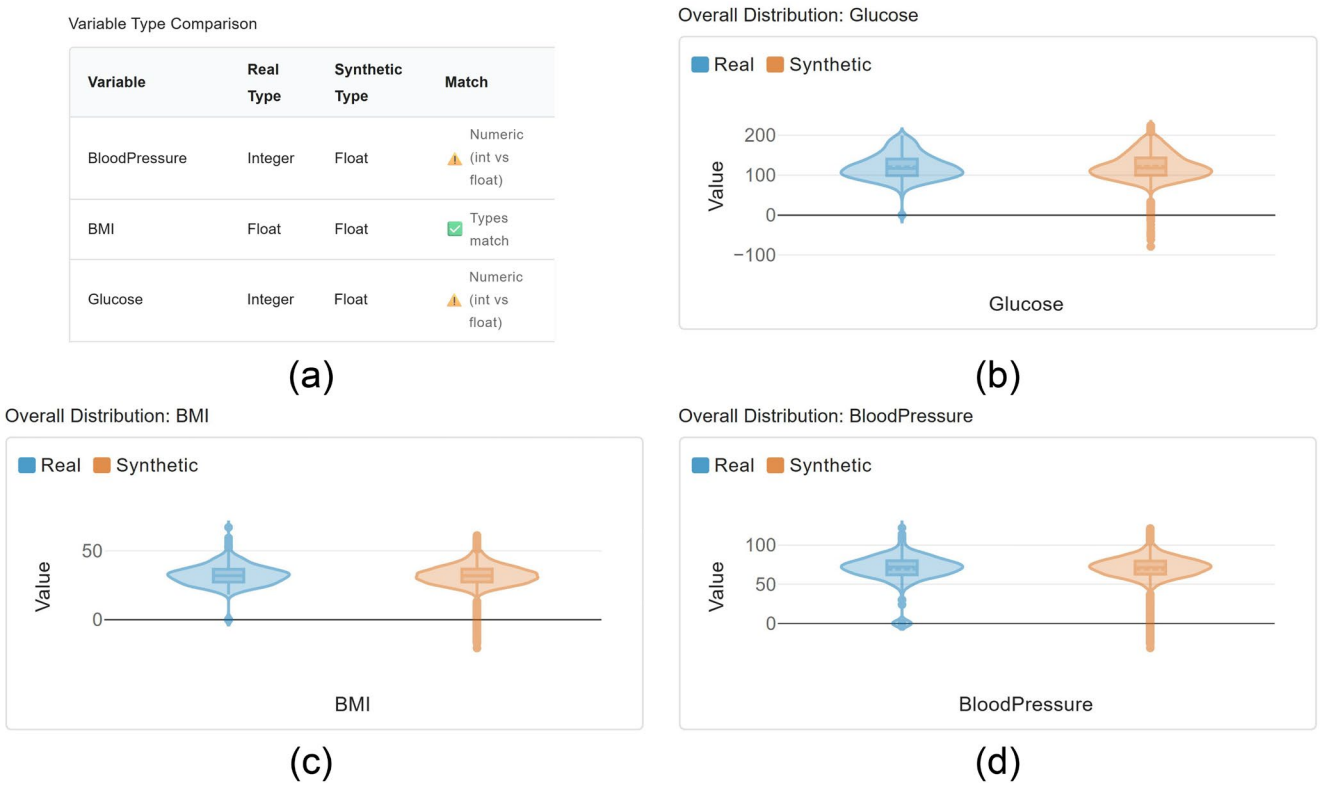

FIGURE 14 | Variables in the diabetes data have some unrealistic values. Specifically, two variables show mismatches between the types of real and synthetic data (a). The corresponding violin plots in (b), (c) and (d) reveal that, for all three variables in (a), the synthetic data include values below 0—an unrealistic outcome in a real-world context.

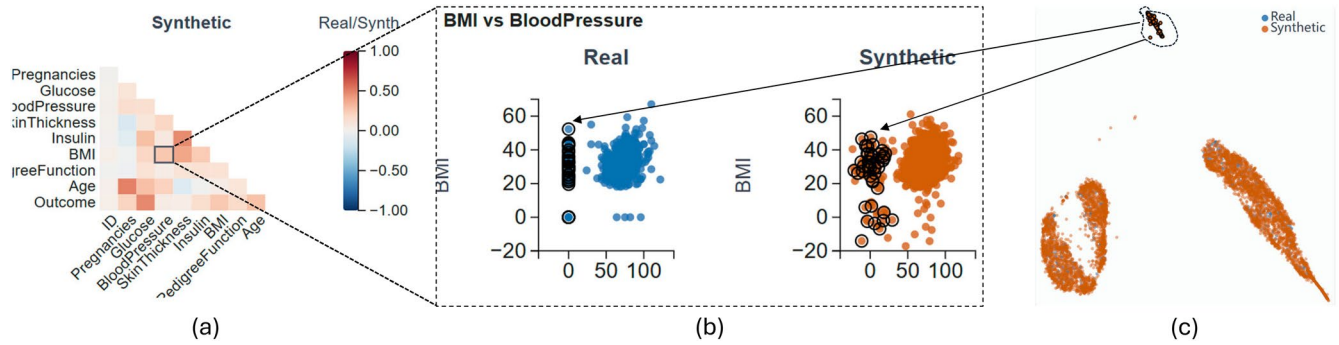

FIGURE 15 | Analysis of a data subgroup from the diabetes dataset using `BMI` and `BloodPressure`: (a) shows their correlation, with detailed plots in (b). Highlighted instances in (b) correspond to selections in (c); real instances have `BloodPressure` = 0, while synthetic ones are near 0.

by Variable in the file `ai-validation-report-diabetes-data.pdf`, available in the Supporting Information (https://github.com/llqsee/SynAccess/tree/main/data). Expert **E2** noted that the univariate similarity visualizations were useful for identifying abnormal values—for example, `BloodPressure` values below 0—which can guide improvements in synthetic data generation based on predefined rules. In the model, such 0 values should either be removed prior to training or handled through rules that enforce realistic, non-zero values.

Where do these unrealistic synthetic data points originate? To explore this, we turned to the multivariate resemblance analysis visualization. Using a lasso tool on the scatterplot, we selected the apparent outlier cluster (see Figure 15c) and found that the corresponding data instances in both the real and synthetic datasets are highlighted in Figure 15b—specifically, at 0 in the real data and around 0 in the synthetic data. This indicates that the selected cluster indeed consists of outlier instances with invalid or unrealistic values (task T5). Expert **E1** commented that this enables users to further investigate the distribution of abnormal subsets identified in the scatterplot, either manually, as shown in Figure 15c, or automatically, using algorithms, as shown in Figure 4b. This helps AI experts quickly identify what is missing in the data and what issues the synthetic data may have, allowing them to consider this information when retraining models.

6.4 | General Expert Feedback

After presenting to them the three case studies, the experts provided overall feedback and suggestions for improvement. All three experts acknowledged the key contribution of

AISYST, emphasizing that training loss alone is insufficient to assess the quality of synthetic data. Instead, post-training validation is essential to uncover issues often hidden in global statistical evaluations. Expert **E1** highlighted that AISYST effectively identifies problematic data subsets by analysing real versus synthetic data similarity in low-dimensional space. This targeted analysis, combined with detailed visualizations and automated reporting, facilitates efficient diagnosis and iteration in the model development process, reducing the reliance on manual debugging.

In addition to that, experts **E1** and **E2** suggested how to improve AISYST for further development. Recent insights from Financial Conduct Authority (FCA) studies (e.g., FCA 2023) emphasize that synthetic data validation should adopt a use-case and risk-based approach, combining quantitative metrics with qualitative assessments and recognizing explicit trade-offs between fidelity, utility and privacy. This contrasts with common regulatory practices (e.g., exercised by the Prudential Regulation Authority, https://www.bankofengland.co.uk/prudential-regulation), which often rely on quantitative thresholds for model approval using aggregated performance metrics such as *Area Under the ROC curve (AUROC)* (Fawcett 2006).

Even in AI-assistant structured analysis, utility and privacy are commonly evaluated. However, the visualization components of AISYST currently focus exclusively on fidelity. Expert **E1** emphasized that AISYST could be improved by incorporating utility and privacy analyses, enabling a more comprehensive assessment of trade-offs among these three key aspects. Experts **E2** and **E3** suggested extending the framework to include regulatory oversight (e.g., for regulated financial institutions) by adding regulators into the loop, thereby supporting more transparent and regulator-aligned validation processes for synthetic data and models, both quantitatively and qualitatively.

Regarding the user interface, expert **E1** noted that the current design is well suited for users with expertise in AI who generate the synthetic data, but may not be accessible to those without sufficient technical background. In particular, the complexity of interactions and the variety of visualizations could be overwhelming for non-expert users (e.g., in regulated financial institutions). Experts **E1** and **E2** suggested that this can be addressed by providing clearer workflow guidance within the interface, helping users navigate the tool more effectively and reducing cognitive overload.

7 | Discussion and Limitations

This section presents the findings from the case studies and expert feedback, highlighting the effectiveness and strengths of AISYST. In addition, we outline several limitations of the current version and propose possible solutions for future work.

7.1 | Discussion

AISYST effectively supports all the tasks mentioned. Multivariate resemblance analysis (task T1) is facilitated through scatterplots generated using DR methods, enabling a two-dimensional visual comparison of the overall distributions of real and synthetic data. Correlation analysis (task T2) is primarily supported by the bivariate correlation visualization, which includes heatmaps for the real data, synthetic data and their differences, as well as a detailed correlation view between two selected variables. Figure 12 presents the distribution of the variables `5CREDIT` and `12CONSCIOUS` in the bivariate correlation analysis visualization. Univariate similarity analysis (task T3) is mainly achieved through both the univariate similarity and bivariate correlation visualization. For example, in the income dataset, Figure 13b shows the distribution of the `10DAILY TRAVEL` variable in the univariate visualization. Besides, in univariate analysis, a table that compares variable types between the real and synthetic datasets to identify mismatches, such as integer versus float types.

Subgroup fidelity exploration (task T4) is primarily supported through a combination of multiple visualizations—particularly the scatterplot in the multivariate analysis visualization—along with other visualizations such as the bivariate correlation visualization, which displays correlations between two variables while highlighting the data records selected in the multivariate scatterplot. Outlier and rare category detection (task T5) is also supported through the integration of several visualizations, in a manner similar to subgroup fidelity exploration. For example, a region in the diabetes dataset highlighted in Figure 15c represents an outlier subgroup in which all data records have blood pressure values close to 0 in Figure 15b.

Compared to existing interactive visual systems for evaluating synthetic data (see, e.g., Noruzman et al. 2021; Santangelo et al. 2025), AISYST reveals where and how synthetic data fail. Most existing approaches provide only a single aggregated score or a set of basic plots, such as scatterplots and heatmaps. They do not support fine-grained exploration to explain why synthetic data may fail to accurately represent certain subgroups, which can ultimately lead to lower utility scores compared to real data. AISYST enables detailed analysis of subgroup-specific data characteristics, helping AI experts not only answer whether the synthetic data has lower fidelity than real data, but also explain why and where the synthetic data fail.

AISYST contributes to assisting AI experts in evaluating whether the current synthetic data needs to be regenerated or whether fine-tuning the generative models could improve the quality of the synthetic data. For example, in the diabetes dataset, we found that the synthetic data includes many blood pressure values less than 0, which are not possible. This insight allows us to fine-tune the generative models to avoid values below 0.

In the case studies, AISYST was applied to three datasets containing up to 10k data records and more than 50 variables, for implementing the DR method and the three visualizations. The DR method, using UMAP, took up to 8 s, which is acceptable in real-world scenarios. This is not a limitation of AISYST, which is capable of handling larger datasets with more records and variables—the capability that can be explored in future work.

Apart from the three case studies, we also conducted sanity checks by randomly selecting some data records from the

three synthetic datasets and treating them as if they were real data to evaluate the generated AI report results. Overall, the AI reports indicated that the synthetic data were near-perfect replications across multiple validation dimensions. For details, please refer to the three AI-generated reports in the `sanity_check_ai_report` folder in the Supporting Information (https://github.com/llqsee/SynAccess/tree/main/data).

7.2 | Limitations

The current version of the synthetic data evaluation focuses exclusively on tabular datasets. However, many other types of generative data exist, such as images and text, which AISyst is not currently designed to evaluate. In future work, we plan to extend AISyst to support not only tabular data but also additional data modalities. Furthermore, the current implementation of AISyst only allows comparison between a single synthetic dataset and the corresponding real dataset. In practice, it is often beneficial to compare multiple synthetic datasets against the same real dataset. We aim to expand AISyst to support such multi-dataset comparisons in future versions. In addition, the present version of AISyst focuses solely on analysing fidelity. Future work will include an exploration of the trade-offs between fidelity, utility and privacy, providing a more comprehensive evaluation framework. Furthermore, AISyst is designed to be generic, which makes it difficult to incorporate domain-specific metrics or background knowledge. For example, in the clinical domain, constraints such as the clinical constraint violation rate or requirements like pregnancy being limited to biologically female patients are not explicitly considered. In future work, we plan to incorporate more concrete domain-specific knowledge. Finally, AISyst currently supports only pairwise correlation analysis between two variables. However, in some cases, three or more variables may exhibit strong mutual dependencies, which cannot be captured by pairwise analysis alone. To address this limitation, we plan to develop a novel visualization technique—based on an extended heatmap—that can effectively represent correlations among three or more variables.

Regarding the scatterplots generated from dimensionality reduction (DR) algorithms, when there are too many data instances, overplotting occurs. Our current visualization strategy first renders the real dataset, followed by the synthetic dataset. As a result, many real data instances are obscured by overlapping synthetic data instances.

To address this issue, numerous enhanced scatterplot designs have been proposed within the visualization community. As introduced by Ellis and Dix (2007), these approaches can be broadly categorized into three groups:

1. Appearance improvement methods, which seek to reduce visual clutter while preserving the original spatial layout, such as sampling techniques that selectively remove points (Bertini and Santucci 2004; Bertini and Santucci 2005) and aggregation techniques that summarize groups of points into higher-level visual representations (Carr et al. 1987).
2. Spatial distortion methods, which intentionally modify the spatial arrangement of data points to reveal hidden structures and reduce overlap, including point or line displacement techniques such as Splatterplots (Mayorga and Gleicher 2013) and topology-preserving distortions that reorganize point locations while maintaining neighbourhood relationships (Liu et al. 2024).
3. Temporal methods, which exploit animation or interactive transitions to reveal information that cannot be displayed simultaneously in a static view (Waldeck and Balfanz 2004).

Each category offers different trade-offs between preserving geometric accuracy, maintaining data readability and supporting analytical tasks. Appearance-improvement methods typically preserve the original data coordinates but may hide individual observations; spatial-distortion methods improve visibility at the cost of positional fidelity; and temporal approaches can reveal additional information but require users to integrate observations over time. Consequently, the choice of technique depends on the characteristics of the dataset and the analytical goals of the visualization system. One or more of these methods could be applied in AISyst in our future work.

It should be noted that the data type of each variable was inferred based on the actual values present in the datasets, without considering their real-world semantics. This posed a challenge for AISyst, which had to 'guess' the type and range of variables based solely on the appearance of the available values, sometimes leading to misconceptions. For example, the values of a variable such as blood pressure may have been measured using a physical or digital instrument, and thus be inherently of `float` type, but were presented in the dataset as rounded values, giving the impression of an `integer` type. Another example is a person's age, which is often reported as a whole number of years (e.g., age at the most recent birthday), despite being effectively a continuous variable that is rounded for practical reporting purposes.

The current evaluation is primarily based on qualitative feedback collected through interviews with domain experts. Although this approach provides valuable insights into the usefulness and practical applicability of AISyst, the number of participants was limited and no controlled user study was conducted to collect quantitative measures such as task completion time, response accuracy, or usability metrics. Consequently, the effectiveness of the system has not yet been assessed through quantitative evaluation. Conducting larger-scale user studies with quantitative performance measures represents an important direction for future work.

We note that the anomaly detection algorithm in AISyst is designed to support interactive visual exploration rather than to function as a standalone anomaly detection method. Therefore, the focus of this work is on the effectiveness of the overall visual analytics system rather than on maximizing anomaly detection performance in isolation. As a result, we did not perform a quantitative benchmarking study comparing the proposed algorithm with existing anomaly detection techniques. While such an evaluation would be valuable for understanding the relative strengths and weaknesses of different detection strategies, it falls outside the scope of the current work. Future research could investigate this aspect through

controlled benchmarking experiments on datasets with known anomaly labels and standard evaluation metrics.

8 | Conclusion

This paper presented an AI-powered interactive visual system, AISyst, designed to evaluate the fidelity of tabular synthetic data. The system comprises four main components: a multivariate resemblance analysis visualization for assessing overall data distribution; a bivariate correlation analysis visualization for examining correlations between variable pairs; a univariate similarity analysis visualization for analysing the distribution of individual variables and detecting variable type mismatches; and an AI assistant for generating general summary reports. AISyst not only assists AI experts in assessing the overall quality of synthetic data, but also enables to explore where and why synthetic data fail. The three case studies demonstrated the effectiveness of AISyst in supporting all evaluation tasks identified through literature review and expert consultation. Expert feedback also affirmed the usability of the system and its potential for a wider deployment. In future work, we plan to extend AISyst beyond tabular data to support other data modalities, such as images and language, expanding its applicability across various domains of synthetic data generation and evaluation.

Author Contributions

L.B., L.Č. and R.A.R. designed the study and provided conceptualization and the overall supervision. N.O. wrote and executed the basic computer code behind AISyst. L.L. implemented additional features in the code, applied AISyst for data analysis in the three case studies and summarized the visualization results. L.B. and N.O. carried out statistical analysis of outputs. R.A.R. and L.L. provided the overall visualization expertise. L.Č., J.G.-S. and L.B. contributed to the expert assessment of the AISyst performance. N.O. and L.L. wrote the first draft of the paper, followed by editing and reviewing by L.L., L.B., R.A.R., L.Č. and J.G.-S. All authors contributed to the interpretation of the results and agreed with the final version of the paper.

Acknowledgements

The research collaboration with the University of Leeds spinout 4-Xtra Technologies Ltd. is gratefully acknowledged.

Funding

This research was supported by an EPSRC grant EP/X029689/1 'Making Visualization Scalable (MAVIS) for explaining machine learning models' (https://gtr.ukri.org/project/4D2D6831-B0A1-40DF-AF11-86376D0A0604). Research of N.O. was carried out as part of the Data Scientist Development Programme at the Leeds Institute for Data Analytics (LIDA).

Conflicts of Interest

This paper expresses personal opinions and views of L.B., L.Č. and J.G.-S., which do not necessarily reflect policies or positions of 4-Xtra Technologies Ltd. The authors declare no conflicts of interest.

Data Availability Statement

The datasets used in this study are available at a (https://github.com/llqsee/SynAccess/tree/main/data). Upon acceptance of this paper, the datasets and the code will be released on GitHub and made publicly available.

References

Akçay, A. E., G. Ertek, and G. Büyüközkan. 2012. "Analyzing the Solutions of DEA Through Information Visualization and Data Mining Techniques: SmartDEA Framework." *Expert Systems with Applications* 39: 7763–7775. https://doi.org/10.1016/j.eswa.2012.01.059.

Albuquerque, G., T. Lowe, and M. Magnor. 2011. "Synthetic Generation of High-Dimensional Datasets." *IEEE Transactions on Visualization and Computer Graphics* 17: 2317–2324. https://doi.org/10.1109/TVCG.2011.237.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57: 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bertini, E., and G. Santucci. 2004. "By Chance Is Not Enough: Preserving Relative Density Through Non Uniform Sampling." In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004*, 622–629. IEEE. https://doi.org/10.1109/IV.2004.1320207.

Bertini, E., and G. Santucci. 2005. "Improving 2D Scatterplots Effectiveness Through Sampling, Displacement, and User Perception." In *Ninth International Conference on Information Visualisation (IV'05)*, 826–834. IEEE. https://doi.org/10.1109/IV.2005.62.

Bostock, M., V. Ogievetsky, and J. Heer. 2011. "D^3 Data-Driven Documents." *IEEE Transactions on Visualization and Computer Graphics* 17: 2301–2309. https://doi.org/10.1109/TVCG.2011.185.

Brito, E., M. Shoush, K. Tamm, P. Etti, and L. Kamm. 2025. "SynthGuard: Redefining Synthetic Data Generation With a Scalable and Privacy-Preserving Workflow Framework." In *Availability, Reliability and Security (ARES 2025)*, edited by B. Coppens, B. Volckaert, V. Naessens, and B. De Sutter, 193–211. Springer. https://doi.org/10.1007/978-3-032-00633-2_12. Lecture Notes in Computer Science, Vol. 15995.

Budu, E., K. Etminani, A. Soliman, and T. Rögnvaldsson. 2024. "Evaluation of Synthetic Electronic Health Records: A Systematic Review and Experimental Assessment." *Neurocomputing* 603: 128253. https://doi.org/10.1016/j.neucom.2024.128253.

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. "Scatterplot Matrix Techniques for Large *N*." *Journal of the American Statistical Association* 82: 424–436. https://doi.org/10.2307/2289444.

Chandra, G., P. Siirtola, S. Tamminen, M. J. Knip, R. Veijola, and J. Röning. 2022. "Impacts of Data Synthesis: A Metric for Quantifiable Data Standards and Performances." *Data* 7: 178. https://doi.org/10.3390/data7120178.

Che, Z., Y. Cheng, S. Zhai, Z. Sun, and Y. Liu. 2017. "Boosting Deep Learning Risk Prediction With Generative Adversarial Networks for Electronic Health Records." In *2017 IEEE International Conference on Data Mining (ICDM)*, 787–792. IEEE. https://doi.org/10.1109/ICDM.2017.93.

Claude. 2024. *Model Claude 3.5 Haiku. Released 22 October 2024*. Anthropic. https://www.anthropic.com. https://claude.com/platform/api.

Conejero, J. M., J. C. Preciado, A. J. Fernández-García, A. E. Prieto, and R. Rodríguez-Echeverría. 2021. "Towards the Use of Data Engineering, Advanced Visualization Techniques and Association Rules to Support Knowledge Discovery for Public Policies." *Expert Systems with Applications* 170: 114509. https://doi.org/10.1016/j.eswa.2020.114509.

Cramér, H. 1946. "Mathematical Methods of Statistics." In *Princeton Landmarks in Mathematics; Princeton Mathematical Series*, vol. 9. Princeton University Press. https://www.jstor.org/stable/j.ctt1bpm9r4.

de Benedetti, J., N. Oues, Z. Wang, P. Myles, and A. Tucker. 2020. "Practical Lessons From Generating Synthetic Healthcare Data With Bayesian Networks." In *ECML PKDD 2020 Workshops*, edited by I. Koprinska, J. Gama, R. Ribeiro, et al., 38–47. Springer. https://doi.org/10.1007/978-3-030-65965-3_3. Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases, Ghent, Belgium, September 14–18, 2020, Proceedings.

Ellis, G., and A. Dix. 2007. "A Taxonomy of Clutter Reduction for Information Visualisation." *IEEE Transactions on Visualization and Computer Graphics* 13: 1216–1223. https://doi.org/10.1109/TVCG.2007.70535.

Fawcett, T. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27: 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

FCA. 2023. "Exploring Synthetic Data Validation—Privacy, Utility and Fidelity." https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity. Insights from a roundtable event hosted by the Financial Conduct Authority, the Information Commissioner's Office, and the Alan Turing Institute. Contributors: Marshall, V., Markham, C., Avramovic, P., Comerford, P., Maple, C., Szpruch, L.

Fedosejev, A. 2015. "React.Js Essentials: A Fast-Paced Guide to Designing and Building Scalable and Maintainable Web Apps With React.Js." In *Community Experience Distilled*. Packt Publishing.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Oliver and Boyd. Scanned version of the 1st edition. Reprint of the 14th edition in: Fisher, R.A. 1990. Statistical Methods, Experimental Design, and Scientific Inference. Oxford University Press, Oxford, UK.

Goncalves, A., P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales. 2020. "Generation and Evaluation of Synthetic Patient Data." *BMC Medical Research Methodology* 20: 108. https://doi.org/10.1186/s12874-020-00977-1.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, et al. 2014. "Generative Adversarial Nets." In *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, vol. 2, 2672–2680. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

Hernandez, M., P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, and N. Aginako. 2025. "Comprehensive Evaluation Framework for Synthetic Tabular Data in Health: Fidelity, Utility and Privacy Analysis of Generative Models With and Without Privacy Guarantees." *Frontiers in Digital Health* 7: 1576290. https://doi.org/10.3389/fdgth.2025.1576290.

Hernando, A., J. Bobadilla, F. Ortega, and A. Gutiérrez. 2018. "Method to Interactively Visualize and Navigate Related Information." *Expert Systems with Applications* 111: 61–75. https://doi.org/10.1016/j.eswa.2018.01.034.

Howell, D. C. 2013. *Statistical Methods for Psychology*. 8th ed. Wadsworth Cengage Learning.

Jiang, X., N. Simidjievski, and M. Jamnik. 2026. "TabStruct: Measuring Structural Fidelity of Tabular Data." Preprint, arXiv:2509.11950 [cs.LG]. https://doi.org/10.48550/arXiv.2509.11950.

Jolliffe, I. T., and J. Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A* 374: 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kaur, D., M. Sobiesk, S. Patil, et al. 2021. "Application of Bayesian Networks to Generate Synthetic Health Data." *Journal of the American Medical Informatics Association* 28: 801–811. https://doi.org/10.1093/jamia/ocaa303.

Kingma, D. P., and M. Welling. 2014. "Auto-Encoding Variational Bayes." In *Conference Proceedings, International Conference on Learning Representations (ICLR)*. ICLR. https://doi.org/10.48550/arXiv.1312.6114.

Kornbrot, D. 2005. "Point Biserial Correlation." In *Encyclopedia of Statistics in Behavioral Science*, edited by B. S. Everitt and D. C. Howell, vol. 3, 1552–1553. John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa485.

Kraskov, A., H. Stögbauer, and P. Grassberger. 2004. "Estimating Mutual Information." *Physical Review E* 69: 066138. https://doi.org/10.1103/PhysRevE.69.066138.

Kurakova, A., and H. Homayouni. 2025. "A Comprehensive Evaluation Framework for Synthetic Medical Tabular Data Generation." *Journal of Biomedical Informatics* 171: 104939. https://doi.org/10.1016/j.jbi.2025.104939.

Lautrup, A. D., T. Hyrup, A. Zimek, and P. Schneider-Kamp. 2025a. "Syntheval: A Framework for Detailed Utility and Privacy Evaluation of Tabular Synthetic Data." *Data Mining and Knowledge Discovery* 39: 6. https://doi.org/10.1007/s10618-024-01081-4.

Lautrup, A. D., T. Hyrup, A. Zimek, and P. Schneider-Kamp. 2025b. "Systematic Review of Generative Modelling Tools and Utility Metrics for Fully Synthetic Tabular Data." *ACM Computing Surveys* 57: 1–38. https://doi.org/10.1145/3704437.

Li, J., B. J. Cairns, J. Li, and T. Zhu. 2023. "Generating Synthetic Mixed-Type Longitudinal Electronic Health Records for Artificial Intelligent Applications." *npj Digital Medicine* 6: 98. https://doi.org/10.1038/s41746-023-00834-7.

Liu, L., R. Vuillemot, P. Rivière, J. Boy, and A. Tabard. 2024. "Generalizing OD Maps to Explore Multi-Dimensional Geospatial Datasets." *Cartographic Journal* 61: 49–68. https://doi.org/10.1080/00087041.2024.2325191.

Livieris, I. E., N. Alimpertis, G. Domalis, and D. Tsakalidis. 2024. "An Evaluation Framework for Synthetic Data Generation Models." In *Artificial Intelligence Applications and Innovations, AIAI 2024. IFIP Advances in Information and Communication Technology*, vol. 713. Springer. https://doi.org/10.1007/978-3-031-63219-8_24.

Mayorga, A., and M. Gleicher. 2013. "Splatterplots: Overcoming Overdraw in Scatter Plots." *IEEE Transactions on Visualization and Computer Graphics* 19: 1526–1538. https://doi.org/10.1109/TVCG.2013.65.

McInnes, L., J. Healy, and J. Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." Preprint, arXiv:1802.03426 [stat.ML]. https://doi.org/10.48550/arXiv.1802.03426.

McInnes, L., J. Healy, N. Saul, and L. Großberger. 2018. "UMAP: Uniform Manifold Approximation and Projection." *Journal of Open Source Software* 3: 861. https://doi.org/10.21105/joss.00861.

Mohedano-Munoz, M. A., C. Soguero-Ruiz, I. Mora-Jiménez, M. Rubio-Sánchez, J. Álvarez Rodríguez, and A. Sanchez. 2023. "A Streaming Data Visualization Framework for Supporting Decision-Making in the Intensive Care Unit." *Expert Systems with Applications* 227: 120252. https://doi.org/10.1016/j.eswa.2023.120252.

Nguyen, T. A. H., D. Nguyen, T. N. Vo, T. D. Le, and S. Gupta. 2025. "Causal-Aware Generative Adversarial Networks With Reinforcement Learning." Preprint, arXiv:2510.24046 [cs.LG]. https://doi.org/10.48550/arXiv.2510.24046.

NIDDK. 2021. *Diabetes Dataset*. Kaggle. https://www.kaggle.com/datasets/mathchi/diabetes-data-set [Dataset originally from the National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov].

Noruzman, A. H., N. A. Ghani, and N. S. A. Zulkifli. 2021. "Gretel.Ai: Open-Source Artificial Intelligence Tool to Generate New Synthetic Data." *Malaysian Journal of Innovation in Engineering and Applied*

Social Sciences (MyJIEAS) 1: 15–22. https://myjieas.psa.edu.my/index.php/myjieas/article/view/27.

Park, N., M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. 2018. "Data Synthesis Based on Generative Adversarial Networks." *Proceedings of the VLDB Endowment* 11: 1071–1083. https://doi.org/10.14778/3231751.3231757.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

Ping, H., J. Stoyanovich, and B. Howe. 2017. "DataSynthesizer: Privacy-Preserving Synthetic Datasets." In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM'17)*. ACM. https://doi.org/10.1145/3085504.3091117.

Putten, P. 2000. *Insurance Company Benchmark (COIL 2000) [Dataset]*. UCI Machine Learning Repository. https://doi.org/10.24432/C5630S.

Raab, G. M., B. Nowok, and C. Dibben. 2021. "Assessing, Visualizing and Improving the Utility of Synthetic Data." Preprint, arXiv:2109.12717 [stat.CO]. https://doi.org/10.48550/arXiv.2109.12717.

Rodgers, J. L., and W. A. Nicewander. 1988. "Thirteen Ways to Look at the Correlation Coefficient." *American Statistician* 42: 59–66. https://doi.org/10.1080/00031305.1988.10475524.

Rubner, Y., C. Tomasi, and L. J. Guibas. 2000. "The Earth Mover's Distance as a Metric for Image Retrieval." *International Journal of Computer Vision* 40: 99–121. https://doi.org/10.1023/A:1026543900054.

Santangelo, G., G. Nicora, R. Bellazzi, and A. Dagliati. 2024. "SynthCheck: A Dashboard for Synthetic Data Quality Assessment." In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies: HEALTHINF*, vol. 2, 246–256. IEEE. https://doi.org/10.5220/0012558700003657.

Santangelo, G., G. Nicora, R. Bellazzi, and A. Dagliati. 2025. "How Good Is Your Synthetic Data? SynthRO, a Dashboard to Evaluate and Benchmark Synthetic Tabular Data." *BMC Medical Informatics and Decision Making* 25: 89. https://doi.org/10.1186/s12911-024-02731-9.

Sedgwick, P. 2014. "Spearman's Rank Correlation Coefficient." *BMJ* 349: g7327. https://doi.org/10.1136/bmj.g7327.

Shao, C., S. Zheng, Y. Xu, H. Gu, X. Qin, and Y. Hu. 2025. "A Visualization System for Dam Safety Monitoring With Application of Digital Twin Platform." *Expert Systems with Applications* 271: 126740. https://doi.org/10.1016/j.eswa.2025.126740.

Sidorenko, A., M. Platzer, M. Scriminaci, and P. Tiwald. 2025. "Benchmarking Synthetic Tabular Data: A Multi-Dimensional Evaluation Framework." Preprint, arXiv:2504.01908 [cs.LG]. https://doi.org/10.48550/arXiv.2504.01908.

van der Maaten, L., and G. Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605. https://jmlr.org/papers/v9/vandermaaten08a.html.

Venugopal, R., N. Shafqat, I. Venugopal, B. M. J. Tillbury, H. D. Stafford, and A. Bourazeri. 2022. "Privacy Preserving Generative Adversarial Networks to Model Electronic Health Records." *Neural Networks* 153: 339–348. https://doi.org/10.1016/j.neunet.2022.06.022.

Waldeck, C., and D. Balfanz. 2004. "Mobile Liquid 2D Scatter Space (ML2DSS)." In *Proceedings, Eighth International Conference on Information Visualisation, 2004. IV 2004*, 494–498. ACM. https://doi.org/10.1109/IV.2004.1320190.

Xu, L. 2020. "Synthesizing Tabular Data Using Conditional GAN." MSc diss., Massachusetts Institute of Technology, Cambridge, MA. https://dspace.mit.edu/handle/1721.1/128349.

Yan, C., Y. Yan, Z. Wan, et al. 2022. "A Multifaceted Benchmarking of Synthetic Electronic Health Record Generation Models." *Nature Communications* 13: 7609. https://doi.org/10.1038/s41467-022-35295-1.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.