Journal of Big Data

METHODOLOGY Open Access

A multi-dimensional hierarchical evaluation system for data quality in trustworthy AI

Hui-Juan Zhang[1,2,3*], Can-Can Chen[1,2], Peng Ran[1,3], Kai Yang[1,3], Quan-Chao Liu[1,2], Zhe-Yuan Sun[1,2], Jia Chen[1,2] and Jia-Ke Chen[1,3]

*Correspondence:
zhanghjyusheng@163.com

[1] Research Institute of China Mobile Communications Corporation, Beijing 100032, China
[2] Future Institute, Beijing 100032, China
[3] Research Institute of Safety Technology, Beijing, China

Abstract

Recently, the widespread adoption of artificial intelligence (AI) has given rise to a significant trust crisis, stemming from the persistent emergence of issues in practical applications. As a crucial component of AI, data has a profound impact on the trustworthiness of AI. Nevertheless, researchers have struggled with the challenge of rationally assessing data quality, primarily due to the scarcity of versatile and effective evaluation methods. To address this trouble, a multi-dimensional hierarchical evaluation system (MDHES) is proposed to estimate the data quality. Initially, multiple key dimensions are devised to evaluate specific data conditions separately by the calculation of individual scores. Then, the strengths and weaknesses among various dimensions can be provided a clearer understanding. Furthermore, a comprehensive evaluation method, incorporating a fuzzy evaluation model, is developed to synthetically evaluate the data quality. Then, this evaluation method can achieve a dynamic balance, and meanwhile achieve a harmonious integration of subjectivity and objectivity criteria to ensure a more precise assessment result. Finally, rigorous experiment verification and comparison in both benchmark problems and real-world applications demonstrate the effectiveness of the proposed MDHES, which can accurately assess data quality to provide a strong data support for the development of trustworthy AI.

Keywords: Data quality, Assessment dimension, Multi-dimensional hierarchical evaluation system, Trustworthy AI

Introduction

With the rapid development of computer software and hardware technology, artificial intelligence (AI) has made significant breakthroughs, which is increasingly applied in multiple fields of human production and life [1–3]. AI has been proven particularly professional to predict stock prices or the stock tendency in the financial field [4]. In the medical field, AI can assist doctors to diagnose diseases and perform surgery [5]. Moreover, AI can identify real-time environmental information for path planning, thus enhancing the likelihood of early arrival of unmanned vehicles in automated driving [6]. However, as the widespread application of AI, the continuous issues, such as the underrepresentation of data or the unfairness of model outputs, have become an obstacle to permeate AI into the actual scenarios [7, 8].

To solve aforementioned issues, trustworthy AI has emerged as a thoroughly important scientific research direction [9]. Over the past decades, numerous investigators have concentrated their efforts on optimizing model structures or enhancing learning algorithms in order to enhance the credibility of AI [10, 11]. For example, Han et al. introduced fuzzy set theory to devise a new building-unit, named fuzzy denoising autoencoder (FDA). This novel unit was used to construct fuzzy deep network [12]. The results displayed that this FDA can extract more robust features compared with the basic unit in traditional DNN, to mitigate the effect of uncertainties. Moreover, Rozsa et al. proposed the batch adjusted network gradients (BANG) for model training, leading to improvements in model accuracy [13]. However, it is undeniable that AI has been criticized in recent years for a major weakness: the lack of interpretability in its decision-making processes [14, 15]. To solve this problem, Dubey et al. devised a scalable polynomial additive model (SPAM) [16]. This SPAM employed tensor rank decompositions of polynomials to create an inherently-interpretable model. In addition, Meng et al. proposed a semantic-enhanced Bayesian personalized explanation ranking (SE-BPER) model, leveraging the interaction information and semantic information [17]. In this SE-BPER model, interaction information was utilized to form a latent factor representation by constructing the interaction matrix. Then, the semantic information was adopted to optimize this factor representation, thereby improving the rationality of decision results. Additional research efforts aimed at the performance improvement of neural network can be found in [18–20]. It imperative to note that the foundation of all these methods, ranging from [12] to [20], is trustworthy of data they adopted. With the maturity of AI technology, including the emergence of automated machine learning (AutoML) platform and industry-standard platforms like PyTorch, it is much easier to develop and improve models when the data are provided than before [21]. Nevertheless, according to surveys conducted in [22, 23], it is staggering that 96% of enterprises encounter challenges related to data quality or labeling in their AI projects, while 40% of them lack the confidence for ensuring the data quality. If a model captures the biases and incorrect correlations of data, it adheres to the principle that "garbage in, garbage out," and will have a significant impact on the credibility of AI [24]. In fact, the breakthrough of AI benefits from the development of high-quality data. which means that the data quality has become absolutely crucial.

In particular, as the emphasis transitions from model optimizing to data improvement, data scientists often spend nearly twice the time on data loading, cleansing and visualization in comparison to model training, selection and deployment [25]. Andrew Ng, a respected figure in the field of AI, has emphasized that 80% data plus 20% model can equal better machine learning [26]. That indicating that more outstanding outcomes of AI can be well achieved by enhancing data quality, especially when the model remains fixed. For instance, ChatGPT, the renowned chat robot, demonstrated greater surprise and credibility in its third iteration compared to its predecessor. This advancement was primarily due to multitudinous efforts in acquiring high-quality data for training, rather than substantial modifications to the model's structure [27]. Furthermore, Professor Songchun Zhu, a globally renowned expert in AI field, once proposed "The key to general AI lies in establishing a "heart" for machines. Data provides learning materials and basis for intelligent agents, and is the foundation for their mental formation" [28].

Good data can guide AI towards goodness and trustworthy. Therefore, more scholars are participating in the evaluation and improvement of data quality. In [29], Liang et al. primarily explored the influence of data pipeline on the credibility of AI, encompassing data design (sourcing and documentation), data sculpting (selection, cleaning and annotation), and data strategies for model testing and monitoring. Their work offers valuable theoretical insights into significance of data quality. However, a specific implementation was not given. Additionally, a data protection strategy, using the backdoor watermarking, was given to authenticate data ownership [30]. This proposed protection method employs poison backdoor attacks for data watermarking. Then, a hypothesis-test-guided method was utilized for verification. The experiments demonstrated that this protection method can effectively prevent data theft to enhance reliability of AI. Caballero proposed a 3Cs model, which is composed of three data quality dimensions for assessing the quality of data: contextual consistency, operational consistency and temporal consistency [31]. Nevertheless, it only focuses on a single dimension of data, potentially overlooking other crucial aspects. A multi-dimensional assessment could provide a more overall understanding of AI credibility. For examples, the significance of data integrity and consistency as pivotal quality assurance dimensions was emphasized together [32]. In addition, a comprehensive assessment methodology, encompassing multiple dimensions, was introduced to estimate the integrity, redundancy, accuracy, timeliness, intelligence and consistency of power data [33]. The result demonstrated that this assessment method can provide a foundation for data analysis and data mining to facilitate the trustworthy AI in power system.

However, the following problems should be future discussed.

1. The evaluation dimensions are currently qualitative, meaning we understand the actions required but lack clarity on the rationale behind them or their potential scope. Practical application guidelines remain somewhat ambiguous. In data-driven scenarios, the performance of AI models is crucial and complex. Due to the qualitative nature of evaluation dimensions, it is often difficult to accurately measure how these variables individually or synergistically affect the model, making it difficult to develop effective optimization strategies for the improvement of AI models.
2. When data is input into an AI model, its performance is not only influenced by a single dimension of the data, but also by multiple factors. The intricate interaction between various dimensions is often overlooked or simplified in the current evaluation methods, resulting in an inability to comprehensively and deeply understand the data and evaluate data. Moreover, a further consideration is needed that the harmonious integration of subjectivity and objectivity criteria to ensure a more precise assessment of data quality.

In order to solve this problem, a multi-dimensional hierarchical evaluation system (MDHES) is proposed to estimate the data quality in this paper. The main contributions of this paper are as follows.

1. Multiple dimensions are designed to evaluate data condition separately. Innovatively, the evaluation dimensions have been not only given, but also the specific quantita-

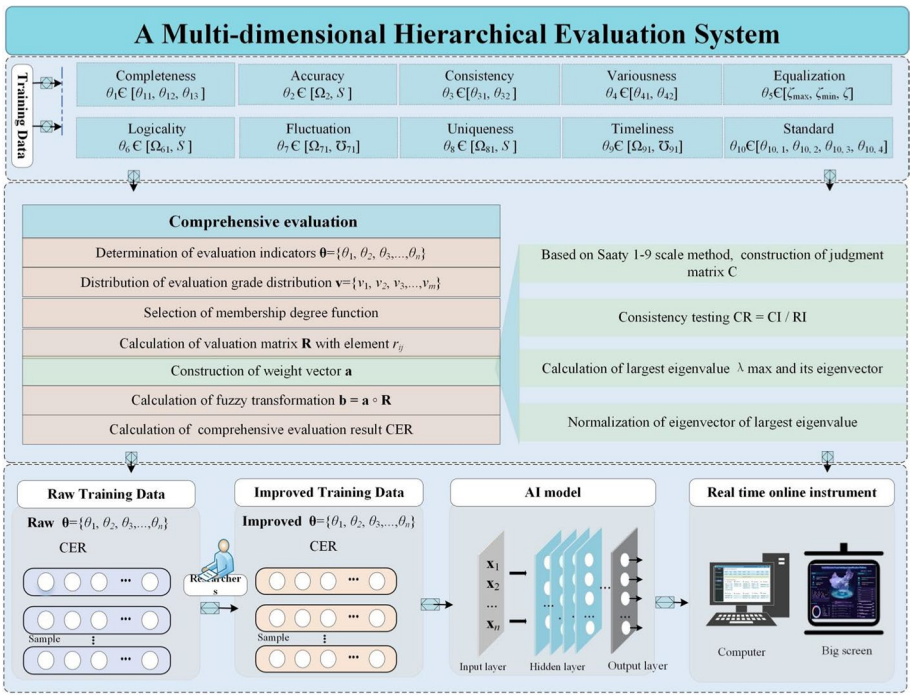

Fig. 1 The framework of the proposed multi-dimensional hierarchical evaluation system

tive formulas, which can provide a clearer understanding of the strengths and weaknesses among these dimensions, by calculating the individual scoring. Moreover, the effects of data improvements can also be explicitly measured, which offers a constructive guidance and feedback on the implementing enhancements for researchers.

2. A comprehensive evaluation method, incorporating fuzzy evaluation model, is developed to synthetically evaluate the data quality, based on the score value of each dimension. This method focuses on interactions of dimensions to achieve the dynamic balance. Furthermore, the adoption of fuzzy evaluation model can achieve a harmonious integration of subjectivity and objectivity criteria to more accurately reflect the data quality.

The outline of this paper is organized as follows. Section "Multi-dimensional hierarchical evaluation system" introduces the proposed MDHES, consisting of quantization of evaluation dimensions and the comprehensive evaluation method in detail. Subsequently, the experimental results and the comparisons on the benchmark problems are discussed to demonstrate the effectiveness of the proposed MDHES in Section "Results and discussion". Section "Practical application of multi-dimension comprehensive evaluation system" highlights real-world applications implemented for cyber-telecoms fraud identification, while the conclusions are given in Section "Conclusion".

Multi-dimensional hierarchical evaluation system

In this section, a MDHES is proposed for data quality evaluation. As shown in Fig. 1, multiple crucial dimensions, consisting of completeness, accuracy, consistency, variousness, equalization, logicality, fluctuation, uniqueness, timeliness, standard, which

encompasses the entire data pipelines (data processing, data usage, data storage and more), are meticulously designed to evaluate data condition separately by calculating individual score of each dimension. Then, a comprehensive evaluation method, integrating individual scores, is developed to provide a synthetically evaluation for the data quality. Next, the proposed MDHES will be introduced in detail.

Design and calculation of dimensions for data quality evaluation

1. Completeness: Data completeness refers to the absence of any gaps or missing values within a training data set. Training data containing miss values will lead to AI models yielding inaccurate assumptions, primarily because the incomplete data only provides a partial information, potentially causing costly mistakes and significant waste of resources [34]. The data completeness encompasses several aspects: the comprehensiveness of features, the fullness of feature values, and the adequacy of data size. For the comprehensiveness of features, θ_{11} can be given as

$$\theta_{11} = \min(1, \Omega_{11}/\mho_{11}) \times 100\%, \tag{1}$$

where $\mho_{11}$ is the number of features in the benchmark data, while Ω_{11} represents the number of features in the training data. Baseline data is meticulously gathered by measurement organizations tailored to specific scenarios, resulting in a comprehensive data set that meets specific requirements. However, acquiring such data demands considerable effort and financial investment.

The presence of null values will affect the availability of data, undermining the decision-making capabilities of AI, especially when the training data contains too many null values. Therefore, it is imperative to consider the fullness of feature values and can be defined as

$$\theta_{12} = (\mathrm{S} - \Omega_{12})/N_s \times 100\%, \tag{2}$$

where Ω_{12} signifies the number of training data samples with null values, and S indicates the total number of training data sample. Having an adequate amount of training data is crucial, which can help the AI model to learn the underlying patterns.

and essential laws, thereby enhancing its generalization ability. It can be described as

$$v_i = \gamma^2 * p_i * (1 - p_i)/\varepsilon^2, \tag{3}$$

where v_i is the ith sample target size in the training data, and γ denotes the z-score associated with the confidence level ε. p_i is the proportion of the ith category in training data. The score of data size can be calculated as

$$\theta_{13} = \min\left(1, \left(\sum_1^{\Omega_{11}} s_i/v_i\right)\Big/\Omega_{11}\right) \times 100\%, \tag{4}$$

where s_i indicates the actual size of ith category. Based on the above analysis, the training data completeness is

$$\theta_1 = \omega_{11}\theta_{11} + \omega_{12}\theta_{12} + \omega_{13}\theta_{13}, \tag{5}$$

where w_{11}, w_{12} and w_{13} are the connect weights for each dimension, respectively.

2. Accuracy: The data accuracy is the degree of closeness between the collected data and the actual true values [35]. The perturbation data, such as the false data by the deep fake algorithm, and adversarial sample in dataset, may hinder AI model to recognize the correct patterns, ultimately affecting the accuracy of AI models. Therefore, for the general perturbation data can be discovered by outlier detection technology, such as the 3-sigma criterion

$$x_{ij} \subseteq \left[\bar{x}_j - 3\kappa_j, \bar{x}_j + 3\kappa_j\right], \tag{6}$$

where x_{ij} is the value of ith row and jth column in training data. $\bar{x}_i$ and κ_i represent the average value and standard deviation of ith column respectively. The accuracy of training data θ_2 can be given as

$$\theta_{21} = \Omega_{21} / S \times 100\%, \tag{7}$$

where Ω_{21} is the number of outliers.

For the carefully designed perturbation data (false data by the deep fake algorithm, the contaminated data), regarded as adversarial sample, the robustness of data can be evaluated by the Adversarial Category Average Confidence (ACAC) [36], which can be defined as the average confidence level of the model classifier in misclassifying

$$\theta_{22} = \begin{cases} \sum_{i=1}^{i=\Omega_{22}} P(F(\mathrm{x}_i^{adv}) \neq y_i) \Big/ \Omega_{22} \text{ non target attack} \\ \sum_{i=1}^{i=\Omega_{22}} P(F(\mathrm{x}_i^{adv}) = y_i) \Big/ \Omega_{22} \text{ target attack} \end{cases} \tag{8}$$

where Ω_{22} represents the number of successful adversarial sample attacks, *xadv* $i$is the ith adversarial sample and y_i is the corresponding label value. Therefore, the accuracy of training data θ_2 can be given as

$$\theta_2 = \Omega_2 / S \times 100\%, \tag{9}$$

where Ω_2 is the number of outliers.

3. Consistency: Contradictory samples in training data can confuse AI models, preventing them from understanding correct relationships among data set. This confusion can reduce the prediction stability of models across various scenarios and even over time [37]. Therefore, it is crucial to measure the consistency of training data. First, the format of features in training data is verified

$$\theta_{31} = (S - \Omega_{31}) / S \times 100\%, \tag{10}$$

where θ_{31} is the score for format consistency, Ω_{31} indicates the number of samples with different formats for a feature. Moreover, by comparing the values of features and labeled values between each sample, the content consistency for the same matter can be given as

$$\theta_{32} = (S - \Omega_{32}) / S \times 100\%, \tag{11}$$

where Ω_{32} denotes the number of samples containing conflicting content. In addition, after content consistency evalution, adversarial samples with the target attacks will also be detected. Therefore, the consistency of training data θ_3 is

$$\theta_3 = \omega_{31}\theta_{31} + \omega_{32}\theta_{32}, \tag{12}$$

where w_{31} and w_{32} are the weight.

4. Variousness: When the training data lacks diversity and is predominantly homogeneous, AI model may become overfitted, limiting its ability to generalize [38]. In fact, the variousness of data can be enhanced by the federated learning. However, too many participants may cause computational waste and increase training costs. Therefore, to ensure the versatility of model across different scenarios, it is essential to evaluate the breadth of data sources θ_{41} and the richness of training data categories θ_{42}

$$\begin{aligned} \theta_{41} &= \min(1, \Omega_{41} / \mho_{41}) \times 100\%, \\ \theta_{42} &= \min(1, \Omega_{42} / \mho_{42}) \times 100\%, \end{aligned} \tag{13}$$

where Ω_{41} is the number of data sources of training data and $\mho_{41}$ is the number of data sources of benchmark data. Additionally, Ω_{42} denotes the number of categories of training data, while $\mho_{42}$ is the number of categories of benchmark data.

Therefore, the variousness of training data can be defined as

$$\theta_4 = \omega_{41}\theta_{41} + \omega_{42}\theta_{42}, \tag{14}$$

where w_{41} and w_{42} signify the connect weights.

5. Equalization: The significant discrepancy in the quantity of samples across different categories can result in discriminatory decision outcomes from the AI models [39]. Thus, to prevent such biases, it is essential to evaluate and ensure the equalization of training data

$$\theta_5 = [1 - (\zeta_{\max} - \bar{\zeta}) / (\zeta_{\max} - \zeta_{\min})] \times 100\% \tag{15}$$

where $\zeta_{\max}$ and $\zeta_{\min}$ are the maximum value of the number of categories respectively.

6. Logicality: The logicality can be utilized to evaluate whether the relationship between features in training data align with factual or commonsense knowledge, which certain data errors may remain undetected during data accuracy assessment [40]. Logical relationships between features encompass comparisons like 'greater than', 'less than', 'equal to', and the like. For example, if feature A and feature B, when multiplied, are expected to be greater than or equal to feature C. If their product falls short of feature C, it indicates a logical inconsistency. Based on the priori knowledge, logicality can be given as

$$\theta_6 = (S - \Omega_{61}) / S \times 100\%, \tag{16}$$

where Ω_{61} represents the number of samples with logical errors.

7. Fluctuation: To investigate the periodic variation or distribution pattern of training data, fluctuation evaluation can be used to assess the difference of historical samples

and latest samples [41]. The evaluation formula for quantifying these fluctuations is expressed as

$$\theta_7 = \min(1, |\mho_{71} - \Omega_{71}| / \mho_{71}) \times 100\%, \tag{17}$$

where $\mho_{71}$ is the sum of historical samples, while Ω_{71} signifies the sum of latest samples, $|.|$ denotes an absolute value operation. It is worth noting that, to ensure computational fairness, the number of historical samples should be equal to the number of latest samples.

8. Uniqueness: Repetitive samples have an influence for the outputs of AI models, often causing the overfitting [42]. Nevertheless, simply eliminating these samples from the training data could compromise the generalization ability of the models. Therefore, data uniqueness assessment is required as a basis for deletion without affecting the performance of models

$$\theta_8 = (S - \Omega_{81}) / S \times 100\%, \tag{18}$$

where Ω_{81} is the number of repetitive samples in training data.

9. Timeliness: AI models need to be trained with the latest data to ensure optimal performance [43]. Failure to update the data in a timely manner may deprive the model of access to the latest information, ultimately diminishing the accuracy of its predictions or decisions. Therefore, timeliness of data is critical to maintain the effectiveness of the AI model. Timeliness of data includes the distributional shifts of data and the update frequency of data. The distributional shifts of data are given as

$$\theta_{9,1} = \min(1, 1 - \Theta) \times 100\%, \tag{19}$$

where

$$\Theta = \frac{1}{100n} \sqrt{\sum_{i=1}^{i=n} \left[\widehat{S}_i(t) - \widehat{S}_i(t-\tau) \right]^2}, \tag{20}$$

$\hat{S}_i$ is the average value of the ith feature of dataset at time t, and τ indicates the time interval. Furthermore, the update frequency of data is

$$\theta_{9,2} = (T_B - T_N) / T_B \times 100\%, \tag{21}$$

where T_B represents the ideal number of update time period and T_N is the time period that has not been updated. Thus, the timeliness of the data can be described as

$$\theta_9 = \omega_{9,1}\theta_{9,1} + \omega_{9,2}\theta_{9,2}, \tag{22}$$

where w_{91} and w_{92} are the connection weights.

10. Standard: By standardizing data formats and naming conventions, the readability and comprehensibility of the data are significantly improved. This process can help researchers rapidly familiarize these data, to reduce the complexity and time required for data processing [44]. Furthermore, data reliability can be enhanced by standardizing data source, collection processes, data storage and data training. The standard of data source means that the data source channel is legitimate. Moreover,

the standard of the collection processes and data storage means that the data is protected by encryption algorithm such as the differential privacy method and federated learning. The standard of data training refers to the standard in the process of using online or offline data to train AI models, such as the homogeneous computation or heterogeneous computing. Based on the manual judgement, θ_{10} score of standard can be given as

$$\theta_{10} = \omega_{10,1}\,\theta_{10,1} + \omega_{10,2}\,\theta_{10,2} + \omega_{10,3}\,\theta_{10,3} + \omega_{10,4}\,\theta_{10,4} + \omega_{10,5}\,\theta_{10,5} + \omega_{10,6}\,\theta_{10,6}, \tag{23}$$

where $\theta_{10,1}$ is the format standard score, $\theta_{10,2}$ is the naming standard score,$\theta_{10,3}$, $\theta_{10,4}$, $\theta_{10,5}$ and $\theta_{10,6}$ indicate scores for data source channel, collection process, data storage channel and data training respectively, $\mathbf{w}_{10}=[w_{10,1},w_{10,2},\ldots,w_{10,6}]$ represents the weight set.

Remark Ten key dimensions have been elaborately designed to evaluate data quality. By calculating the score of each dimension, researchers have a clearer understanding for data condition. This approach not only offers valuable guidance but also provides feedback on the effectiveness of training data enhancement. Moreover, the use of simple naming conventions for dimensions, such as serial numbering, simplifies the expansion of dimensions or the addition of sub-item within each dimension in future studies. It is worth noting that to better align with real-world scenarios, certain dimensions incorporate subjective evaluation criteria. In the following discussion, we will explore strategies to minimize the influence of subjectivity.

Comprehensive evaluation for data quality

Based on the dimensions discussed above, data quality will be comprehensively evaluated in this section. In fact, evaluating multi-dimension problems is difficult, due to the potential fuzziness and subjectivity inherent in certain dimensional values. Moreover, all dimensions have to be considered simultaneously, coupled with varying degrees of importance among them, which makes the problem more complicated. Fuzzy comprehensive evaluation model offers a solution. It can transform qualitative evaluation into quantitative evaluation to make an overall evaluation of things or objects constrained by multiple dimensions, based on fuzzy mathematics. Therefore, this model is introduced into the proposed hierarchical evaluation system, to solve difficult-to-quantify and non-deterministic problems, by better achieving a harmonious integration of subjectivity and objectivity criteria. Next, fuzzy comprehensive evaluation model will be described in detail.

Determination of evaluation dimensions and evaluation grades

According to the discussion in Section "Design and calculation of dimensions for data quality evaluation", the dimension: completeness, accuracy, consistency, variousness, equalization, logicality, fluctuation, repeatability, timelines, and standard, can be applied for data quality evaluation. However, training data in different application scenarios should be evaluated using the appropriate dimensions. Thus, the evaluation

dimensions can be defined as $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \ldots, \theta_n\}$, and $n \in [1, 10]$. The evaluation grades can be given as $\mathbf{v} = \{v_1, v_2, v_3, \ldots, v_m\}$ and m is the number of evaluation grades. For example, when m is four, set $\mathbf{v}$ is {excellent, good, pass, poor}.

Construction of evaluation matrix and weight vector

Firstly, the evaluation matrix $\mathbf{R}$, where the membership degree of dimensions in set $\boldsymbol{\theta}$ for each evaluation grade in set $\mathbf{v}$, needs to be determined

$$\mathbf{R} = \begin{pmatrix} r_{11}, r_{12}, \ldots, r_{1m} \\ r_{21}, r_{22}, \ldots, r_{2m} \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ r_{n1}, r_{n2}, \ldots, r_{nm} \end{pmatrix}, \text{and } \sum_{j=1}^{m} r_{ij} = 1, \tag{24}$$

where r_{ij} is membership degree of ith dimension for the jth evaluation grade and can be given as by adopting the membership function. It is important to highlight that distinct membership functions should be employed in different scenarios. After the evaluation matrix is calculated, weight vector also needs to be determined.

Each dimension has different roles and positions in the evaluation of data quality. Therefore, the weight vector $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, $a_1 \geqq 0$, $\sum a_i = 1$, need to be designed, which can express the importance of each dimension relative to the problem to be evaluated. Since the requirements for each dimension are discrepant in different scenarios. Analytic hierarchy process (AHP) [45] is suitable for the determination of the weights in this paper, which is an effective analytical method that harmoniously combines qualitative and quantitative methods for solving complex problems with multiple objectives. In order to save space, the working principle of AHP will not be presented while the computational steps will be given in detail in the Section Results and discussion.

Make comprehensive evaluation

For the relationship matrix $\mathbf{R}$ and weight vector $\mathbf{a}$, fuzzy transformation result $\mathbf{b}$ can be described as

$$\mathbf{b} = \mathbf{a} \circ \mathbf{R} \tag{25}$$

where vector $\mathbf{b} = [b_1, b_2, \ldots, b_m]$ and symbolic $< \bigcirc >$ represents the fuzzy operator. Moreover, in order to make full use of information of $\mathbf{R}$, the weighted average operator can be adopted

$$b_j = \min \left\{ 1, \sum_{i=1, j=1}^{i=n, j=m} \min(a_i, r_{ij}) \right\}. \tag{26}$$

Thus, the comprehensive evaluation result CER is

$$\mathrm{CER} = \sum_{j=1}^{m} \eta_j b_j, \tag{27}$$

Table 1 The evaluation process of the proposed MDHES for data quality

For training dataset of AI models
%Calculating the sorces of evaluation dimension for data quality
For evaluation object
The selection of appropriate dimension
Calculating the score of evaluation dimension $\boldsymbol{\theta}=\{\theta_1, \theta_2, \theta_3, \ldots, \theta_n\}$ %
The selection form Eqs. (1-23)
%Comprehensively evaluating data quality
Givng as the evaluation grades $v=\{v_1, v_2, v_3, \ldots, v_m\}$
Constructing of evaluation matrix **R** % Eq. (24)
Constructing of judgment matrix **C** % Eq. (28)
Calculating of largest eigenvalue λmax and its eigenvector of matrix **C**
Testing Consistency $CR=CI/RI$ % Eq. (32)
Normalizating of eigenvector of λ_{max}
Obtaing the weight vector a
Making the fuzzy transformation *b* % Eq. (25)
Obtaining the comprehensively evaluation result *CER*
End

where η_j is the source of *j*th evaluation grade. Moreover, the detailed evaluation processes are summarized in Table 1.

Remark The data quality evaluation is complex, because many dimensions to be considered together at the same time, and each dimension has a different level of importance. Therefore, the proposed comprehensive evaluation method, incorporating the fuzzy comprehensive evaluation model, can effectively focus on interactions of dimensions to achieve the dynamic balance. Furthermore, the expert knowledge can be employed in evaluation, and meanwhile the influence of subjectivity in entire evaluation process can be mitigated, by leveraging fuzzy mathematics. Thus, this comprehensive evaluation method will accurately reflect the data quality to increase credibility of AI.

Results and discussion

To further demonstrate the procedures and effectiveness of the proposed MDHES, an example on publicly available data sets: intrusion detection scenario-KDDcup 99 dataset is discussed in this section. The simulations are run on Windows 10.0 operating system with a clock speed 2.6 GHz, 4 GB RAM, and Anaconda 3 development environment with Python 3.6 programming language.

Intrusion detection scenario

Intrusion detection can discover the violations of security policies or signs of attacks in the network and system by collecting and analyzing the operational log information. KDDCup99 dataset was derived from a simulated US Air Force LAN with attacking lasting 7 weeks and can be obtained by https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. The 10_percent_ subset is be adopted in this example. Each sample in subset

Table 2 The distribution details of training data, testing data, benchmark data and validation data

Data	Total number	Normal		Probing		Dos		R2L		U2R	
		Number	Category	Number	Category	Number	Category	Number	Category	Number	Category
Training data	494,021	97,278	1	4107	4	391,458	6	1126	4	52	8
Testing data	311,029	60,593	1	4166	6	229,853	10	228	8	16,189	13
Benchmark data	644,039	218,310	1	9240	6	402,768	10	407	8	13,314	14
Validation Data	322,019	109,537	1	4590	6	201,034	9	204	8	6654	12

Table 3 The Evaluation results of each dimension for training data of intrusion detection

Dimension (%)	Completeness	Accuracy	Consistency	Variousness	Equalization	Logicality	Uniqueness
Evaluation result (Raw)	92.24	96	100	54.5	26	100	30

was labeled as normal and four abnormal categories. Furthermore, anomaly data and another 17 attack types appeared in the test data.

Determination of benchmark data

In order to match the actual scenarios, benchmark data are randomly sampled from the training and testing sets at a ratio of 80%. Additionally, 50% data is randomly sampled from the benchmark data as a validation sample to validate the effectiveness of the AI model (fuzzy neural network, FNN). The distribution details are summarized in Table 2.

Determination of evaluation dimensions

For the training data in intrusion detection scenario, evaluation dimensions are analyzed and used to evaluate, consisting of completeness, accuracy, consistency, variousness, equalization, logicality and uniqueness. Completeness, variousness and equalization can essentially guarantee adequacy and richness of intrusion detection data, thereby increasing the generalization and stability of the AI model. Accuracy, consistency and logicality can ensure that the intrusion data, learned by AI models during training, is real and effective, thereby improving the accuracy of the model in identifying intrusion behaviors. Moreover, the fluctuation, timeliness and standard are not evaluated in this experiment. The reasons are as follows:

Fluctuation: This intrusion detection data does not conform to periodic change. Thus, it is not possible to perform the fluctuation assessment.

Timeliness: This training data, although from 1999, provides a foundation for research on network intrusion detection based on computational intelligence. However, according to the requirements of timeliness, the evaluation scores would be very low. Therefore, the timeliness evaluation is not done in this experiment, to avoid the impact on the final evaluation results.

Standard: The standard of training data cannot be assessed, because this data acquisition process and storage channel is not known.

Evaluation results and analyse for intrusion detection

Dimension evaluation for training data: Based on the completeness, accuracy, consistency, variousness, equalization, logicality and repeatability, KDDCup99 intrusion detection data was evaluated with equal weights for each evaluation dimensions and the details are summarized in Table 3. For completeness of training data, obviously, the scores of features and feature values are 100. The score of data size is the 76.71 with the confidence level 4%, thus the completeness is the 92.24. In addition, the accuracy is 96, where 18,612 anomalous data are filtrated according Eq. (6). No

Table 4 The grade distribution of each dimension in intrusion detection scenario

Dimension (%)	Poor	Pass	Good	Excellent
Completeness	80	85	95	98
Accuracy	85	90	95	99
Consistency	90	95	98	99
Variousness	50	60	70	80
Equalization	35	55	75	85
Logicality	85	90	96	98
Uniqueness	25	35	45	60

contradiction has been discovered between training data and baseline data by comparing each sample, then the score of consistency is 100. Due benchmark data are randomly selected from the training and testing sets, it means that multi-source score of training data is 50. Moreover, there are 23 categories of training data, and 39 categories of benchmark data, the diversity is 56. Therefore, the variousness score is the 53. Intuitively, there is a significant difference in the number of categories of training data and the equalization score is only 1 according to Eq. (13). It can be determined that logicality score is 100 by calibration. Additionally, there are 348,435 duplicate entries in the training data, and the score for uniqueness is calculated to be 30.

Comprehensive evaluation for training data: Based on the above discussion of each dimension, data quality can be comprehensively evaluated in this section. In our experiment, data quality is divided into four evaluation grades: poor, pass, good, excellent. The evaluation matrix **R**, where the membership degree of dimensions for each evaluation grade, needs to be calculated. First, membership function will be determined. For the "poor" grade, the lower the score of dimension, the greater the degree of belonging to it. Thus, "poor" should be a gradually decreasing function. Similarly, "excellent" is a gradually rising function. Both "pass" and "good" are functions that go up and then go down. Based on the above analysis, trapezoidal membership function is adopted with a characteristic of simple calculation. The distribution can be given as

$$r_{i1} = \begin{cases} 1, & x \le \delta_1 \\ (\delta_2 - x)/(\delta_2 - \delta_1), \delta_1 < x < \cdot\delta_2 \\ 0, & x \ge \delta_2 \end{cases}, r_{i2} = \begin{cases} (x - \delta_1)/(\delta_2 - \delta_1), \delta_1 < x < \delta_2 \\ (\delta_3 - x)/(\delta_3 - \delta_2), \delta_2 < x < \delta_3, \\ 0, \quad x \le \delta_1 \text{ or } x \ge \delta_3 \end{cases}$$

$$r_{i3} = \begin{cases} (x - \delta_2)/(\delta_3 - \delta_2), \delta_2 < x < \delta_3 \\ (\delta_4 - x)/(\delta_4 - \delta_3), \delta_3 < x < \delta_4 \\ 0, \quad x \le \delta_2 \text{ or } x \ge \delta_4 \end{cases}, r_{i4} = \begin{cases} 0, & x \le \delta_3 \\ (x - \delta_3)/(\delta_4 - \delta_3), \delta_3 < x \le \delta_4, \\ 1, & x \ge \delta_4 \end{cases} \tag{28}$$

where δ_1, δ_2, δ_3, and δ_4 are the critical values at the four levels of poor, pass, good, excellent, respectively.

For the intrusion detection scenario, the completeness, accuracy, consistency and logicality are critical, which will have a direct impact on the FNN model. However, equalization is hard to guarantee, due to the difficulty and the cost of attacking.

Table 5 1–9 scale method

Scale	Interpretation
1	Equally important compared to two dimensions
3	One dimension is slightly more important than the other
5	One dimension is clearly more important than the other
7	One dimension is more strongly important than the other
9	One dimension is more extremely important than the other
2,4,6,8	Median of two adjacent dimensions
Inverse	$a_{ji} = 1/\, a_{ij}$

Then, a consistency testing is performed for matrix **C**

Uniqueness is also difficult to obtain satisfactory results, because of repeated attack means in this scenario. Therefore, the grade distribution of each dimension is as follows.

According to Table 4, the membership functions of dimensions can be determined. For examples, membership functions of completeness for four grades are

$$r_{i1} = \begin{cases} 1, & x \le 80 \\ (85-x)\big/(85-80), & 80 < x < 85 \\ 0, & x \ge 85 \end{cases}, \; r_{i2} = \begin{cases} (x-80)\big/(85-80), & 80 < x < 85 \\ (95-x)\big/(95-85), & 85 < x < 95, \\ 0, & x \le 80 \text{ or } x \ge 95 \end{cases}$$

$$r_{i3} = \begin{cases} (x-85)\big/(95-85), & 85 < x < 95 \\ (98-x)\big/(98-95), & 95 < x < 98 \\ 0, & x \le 85 \text{ or } x \ge 98 \end{cases}, \; r_{i4} = \begin{cases} 0, & x \le 85 \\ (x-95)\big/(98-95), & 95 < x \le 98, \\ 1, & x \ge 98 \end{cases} \tag{29}$$

thus, the fuzzy set of completeness is [0, 0.276, 0.724, 0]. Sequentially, evaluation matrix **R** can be given

$$\mathbf{R} = \begin{bmatrix} 0, & 0.276, & 0.724, & 1 \\ 0, & 0, & 0.67, & 0.33 \\ 0, & 0, & 0, & 1 \\ 0.55 & 0.45 & 0, & 0 \\ 1, & 0, & 0, & 0 \\ 0, & 0, & 0, & 1 \\ 0.5, & 0.5, & 0, & 0 \end{bmatrix}, \tag{30}$$

Next, the weight vector **a** will be depicted by adopting AHP. In this AHP, judgment matrix **C** is first constructed and the element c_{ij} of **C** are given using 1–9 scale method proposed by Saaty.

Based on Table 5, the judgment matrix **C**, by a two-by-two comparison of completeness, accuracy, consistency, variousness, equalization, logicality and repeatability, can be defined as

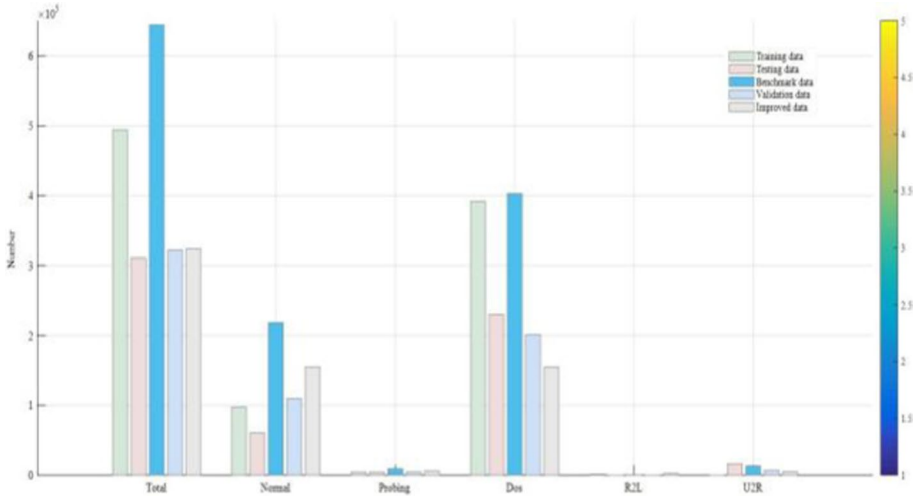

Fig. 2 The number of each data set

$$\mathbf{C} = \begin{bmatrix} 1, & 1/4, & 1/6, & 0.5, & 7, & 1/4, & 6 \\ 4, & 1, & 1/3, & 5, & 9, & 1, & 9 \\ 6, & 3, & 1, & 5, & 9, & 1, & 9 \\ 2, & 1/5, & 1/5, & 1, & 5, & 1/6, & 5 \\ 1/7, & 1/9, & 1/9, & 1/5, & 1 & 1/8, & 1/2 \\ 4, & 1, & 1, & 7, & 8, & 1, & 8 \\ 1/6, & 1/9, & 1/9, & 1/5, & 2, & 1/9, & 1 \end{bmatrix}, \tag{31}$$

$$\begin{aligned} &CR{=}CI\big/RI, \\ &CI{=}\left(\lambda_{\max}{-}\tau\right)\big/\left(\tau{-}1\right), \end{aligned} \tag{32}$$

where CR is the consistency ratio, λ_{max} is the absolute value of the largest eigenvalue of the matrix $\mathbf{C}$ and τ is the number of non-zero eigenvalues order. For matrix $\mathbf{C}$, λ_{max} is 7.588, τ is the 7. Additionally, when the eigenvalue is 7, the corresponding RI is 1.32, which is directly given by Saaty. Therefore, value of CR is 0.074 and passes the consistency testing. The maximum vector of the maximum eigenvalue λmax is [– 0.15, – 0.43, – 0.69, – 0.17, – 0.04, – 0.53, – 0.048]. Then, after normalization, the weight vector is [0.18, 0.25, 0.25, 0.04, 0.04, 0.12, 0.12]. Therefore, the fuzzy transformation result is [0.1, 0.1, 0.47, 0.31]. The scores corresponding to four evaluation grades are [40, 60, 80, 90]. Hence, the comprehensive evaluation result of training data of intrusion detection is 75.5.

In addition, to better compare the advantages of the fuzzy comprehensive evaluation method, it is compared with the weighted average evaluation method, where each dimension is equally weighted. The calculated score for data quality is 71.2. In this scenario, the score is 4 points lower, compared to the score of fuzzy comprehensive evaluation. However, the fuzzy comprehensive evaluation method produces the final evaluation result by setting unequal weights for each evaluation dimension. This weight setting method can reflect the relative importance between the evaluation dimension, making the results more scientific and reasonable. Besides, it can effectively deal with various vague and uncertain information. These advantages make the fuzzy comprehensive

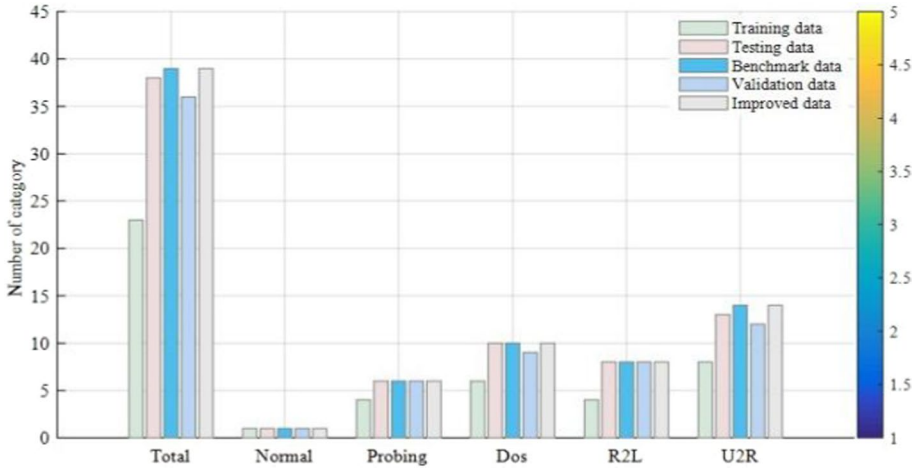

Fig. 3 The number of categories of each data set

evaluation method prominent and become one of the effective means to solve practical evaluation problems.

Based on the above discussion, equalization and uniqueness are the main reasons for causing the unsatisfactory score of data quality. Therefore, some data processing is required for the data quality improvement. After de-duplication and simple random sampling [46], the distributions of all data sets are shown in Figs. 2, 3. It can be seen that, to improve the equalization, the number of category R2L is decreased, while other categories are increased. Moreover, the evaluation results of improved training data are summarized in Table 5 and the comparison with the raw training data is displayed in Fig. 4.

Based on Fig. 4 and Table 6, the accuracy is improved by 2% comparing with the raw training data. Additionally, the scores of variousness, equalization and uniqueness have been improved by at least 50%. It is worth noting that we don't completely remove duplicates, to maintain a balance between the equalization and uniqueness, due to the

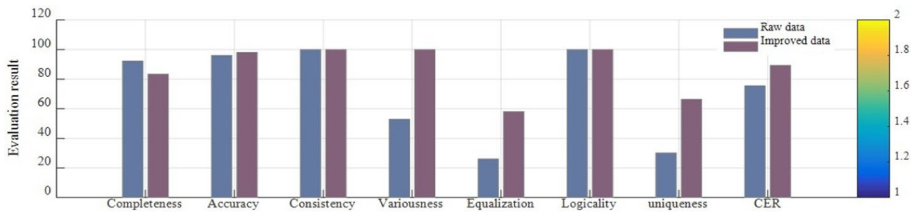

Fig. 4 The evaluation result comparison of raw training data and improved training data for intrusion detection

Table 6 The evaluation results of each dimension for processed training data of intrusion detection

Dimension (%)	Completeness	Accuracy	Consistency	Variousness	Equalization	Logicality	Uniqueness	CER
Evaluation result (improved)	90	98	100	100	58	100	66	89.2

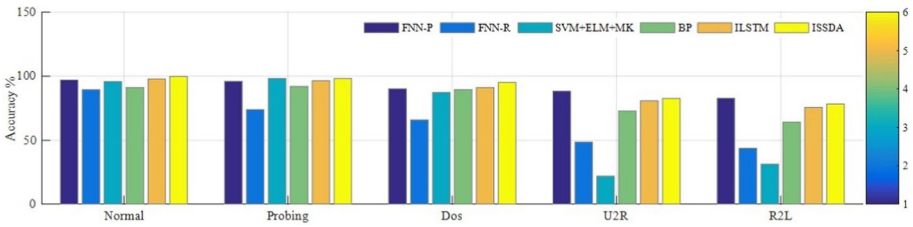

Fig. 5 The accuracy comparison of different AI models

limitations in benchmark database. The comprehensive score of the processed training data is 89.2, whose quality has significantly improved.

In order to further demonstrate the impact of training data improvement on model performance, FNN will be used in the validation data. Moreover, the simulation results of FNN on the processed training data (FNN-P) are compared with FNN on the raw training data (FNN-R), multi-level hybrid support vector machine and extreme learning machine based on modified k-means (SVM + ELM + MK) [47], back propagation (BP) network, improved long-short term memory (ILSTM) [48], and improved sparse denoising autoencoder (ISSDA) [49]. To make a fair comparison, the accuracy T_A, false positive rate T_F, and false negative rate T_M are introduced. The results are displayed in Figs. 5–7 and the details are shown in Table 7.

The accuracy of different AI models is displayed in Fig. 5. The false positive rate, and false negative rate are shown in Figs. 6 and 7. Combining with the Table 7, it can be seen that FNN-P has the significant improvement on all categories, whose performance is shown in bold in Table 7, such as the accuracy of 'Normal' is improved by approximately 8% and the false rate of 'Dos' is reduced to 3.6%, comparing with the FNN-R. It indicates that data quality plays a crucial role in model performance improvement. Moreover, FNN-P performs better than the meticulously designed SVM + ELM + MK, especially in categories 'U2R' and 'R2L'. In addition, the effectiveness of FNN-P can be comparable to that of improved deep networks (ILSTM and ISSDA). Based on the above analysis, the proposed MDHES has a magnitude significance. The strengths and weaknesses of intrusion detection can be clearer understood by quantifying the score of each dimension, which will provide the guidance and feedback on the data quality improvement for the researchers. Furthermore, the proposed comprehensive evaluation method can achieve the dynamic balance of dimensions to avoid the resource waste for excessive pursuing a certain dimension improvement, such that the score equalization is only 58, FNN still achieves good performance.

Table 7 Performance comparison of different AI models for intrusion detection

Types (%)	Normal			Probing			Dos			U2R			R2L		
	T_A	T_F	T_M	T_{IA}	T_F	T_M	T_A	T_F	T_M	T_A	T_F	T_M	T_A	T_F	T_M
FNN-P	**97.01**	**1.66**	**0.84**	**95.93**	**3.62**	**2.51**	**90.05**	**7.50**	**4.31**	**88.10**	**7.10**	**6.20**	**82.60**	**10.80**	**8.31**
FNN-R	89.36	8.51	2.3	73.67	24.6	6.94	65.87	32.87	11.43	48.64	36.65	24.68	43.81	32.41	23.79
SVM + ELM + MK [47]	95.79	4.68	2.05	98.10	0.95	0.46	87.20	11.32	2.40	21.93	57.41	22.69	31.35	48.46	21.92
BP	91.12	5.92	3.38	91.94	5.24	2.86	89.4	6.32	4.28	72.60	23.36	13.31	64.20	28.27	15.54
ILSTM [48]	97.79	1.84	1.27	96.37	3.74	2.83	91.12	8.39	1.27	80.61	15.48	8.45	75.48	21.39	9.24
ISSDA [49]	99.72	0.18	0.11	98.13	1.10	0.77	95.02	5.32	3.61	82.50	23.67	9.82	78.13	16.69	5.19

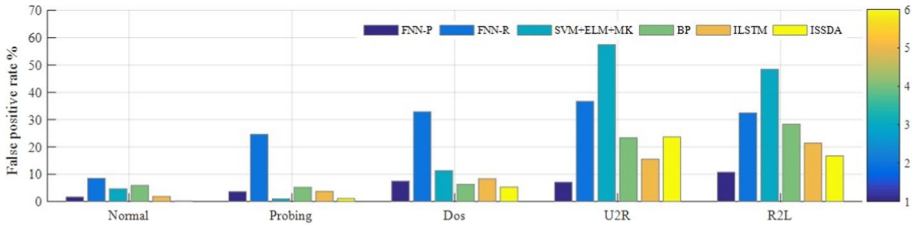

Fig. 6 The false positive rate comparison of different AI models

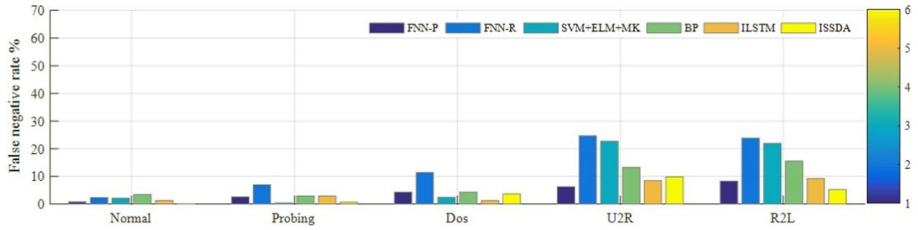

Fig. 7 The false negative rate comparison of different AI models

Practical application of multi-dimension comprehensive evaluation system

In recent years, cyber-telecoms fraud has been a seriously social problem, threatening the property safety of the people, with the characteristics of frequent occurrence, rapid increase, and repeated prohibition. Deep neural networks (DNNs) are the powerful tool to effectively distinguish the fraudulent behaviour. However, this tool is highly questionable, because that the effects are unable to determine, which may cause the unbearable consequences. Therefore, the proposed MDHES are investigated from the data quality aspect for intelligent identification of cyber-telecoms fraud to enhance the credibility of DNNs.

Fraud data from four provincial companies in China has been obtained to form a benchmark database after data processing (deduplication, denoising, and complement of missing values, etc.), which will be encrypted by the differential privacy way and uploaded to the data management library of this proposed evaluation system. The data features can be summarized into three categories: network management data, signaling data, and call ticket data. The network management data contains the IP address, MAC address, port number and so on. Signaling data is some location related information. The features of call ticket data include the number of calls, total call duration, maximum call duration, minimum call duration and last call duration. The important features of this benchmark database are 32 after selection, where the some features are very sensitive and required not to disclose publicly. Moreover, the number of total data is 500 million, contains the approximately 400 million normal samples and 100 million fraud samples. There are mainly types of fraud: traditional fraud (gambling, pornography, pyramid schemes) and new fraud (investment and financial management, pig killing, loans, brushing orders, counterfeit industry and commerce, counterfeit public security, procurator and judicial authorities, counterfeit customer service and so on).

In this scenario, we adopt nine dimensions, except the fluctuation due to the irregular occurrence of fraud activities, to evaluate the training data. Firstly, uploading the

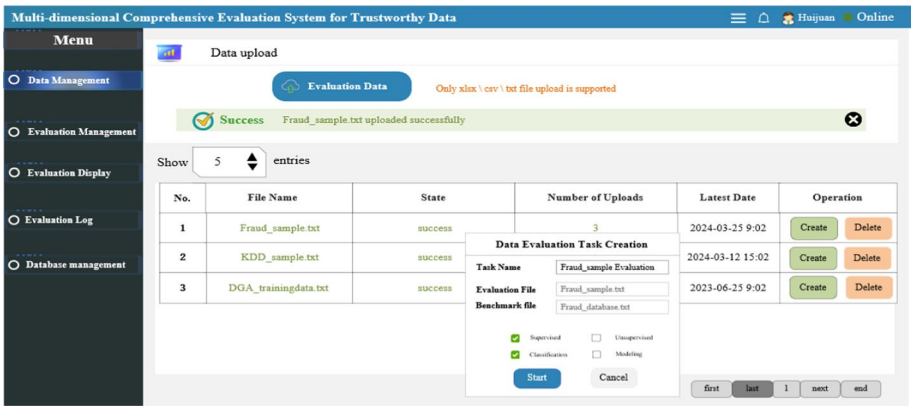

Fig. 8 The evaluation task creation of fraud data

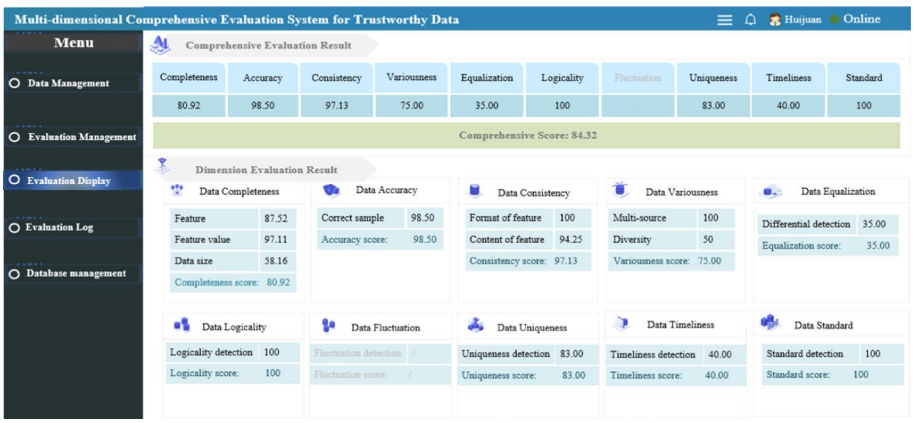

Fig. 9 The evaluation result display of fraud data

training data file with the 40 million samples to the proposed MDHES (seen as the Fig. 8). Then, Clicking the 'Create' button, evaluation task can be created. Next, nine dimensions are selected, and the task starts by clicking the 'Start' button. The evaluation results are shown in Fig. 9. It can be seen that the data size in completeness is 58.16 and the variousness is 75. Therefore, the dimensions: completeness, consistency, uniqueness and timeliness have unsatisfactory scores, which needs to be further improved and repeat the evaluation process. Additionally, in order to clearly demonstrate the changes in the scores of the raw and improved data samples, the comparisons are displayed in Fig. 10, and the details are summarized in Table 7.

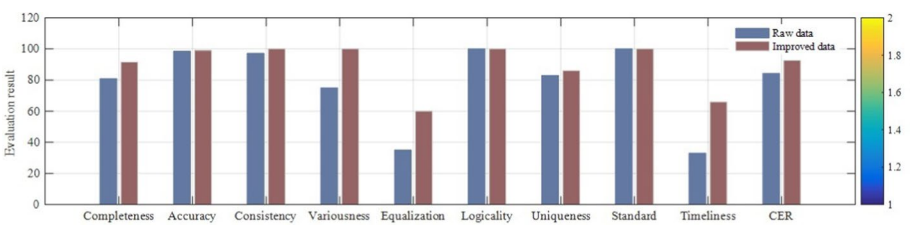

Fig. 10 The evaluation result comparison of raw training data and improved training data for cyber-telecoms fraud identification

Table 8 The evaluation comparisons of the raw and improved data samples for telephone network fraud

Dimension (%)	Completeness	Accuracy	Consistency	Variousness	Equalization	Logicality	Uniqueness	Timeliness	Standard	CER
Score (Raw)	80.92	98.5	97.13	75	35	100	83	40	100	84.32
Score (processed)	91.55	99	100	100	60	100	86	60	100	92.6

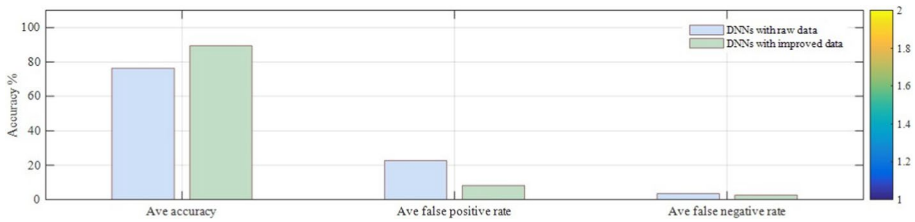

Fig. 11 The average performance comparison of DNNs with the raw training data and improved training data

It can be seen the completeness of training data is improved by 10% in Fig. 10 and Table 8. Consistency and timeliness are two important dimensions for fraud identification DNNs, due to the replacement of objects. For example, the previously normal URL has become a malicious URL, which will lead to a contradiction of label value. Therefore, the consistency is enhanced by correcting contradiction samples based on the latest fraud data. Of course, the timeliness is also increased. Moreover, the comprehensive score for data quality is 92.6, which is enhanced by approximately 8.3%, comparing with the original training data. To reveal the impact of data quality change on DNNs, SDA and LSTM are adopted, where the basic building units of SDA are 32, and the basic units of LSTM are 21. The performances on the validation data are shown in Fig. 11.

As shown in Fig. 11, the average accuracy, average false positive rate, average and false negative rate of DNNs have the positive improvement. The average accuracy of DNNs is 89.46, which has increased by 13% comparing to the original data. Furthermore, the average T_F is reduced to 8.19% and the average T_M is also 2.61%. Based on the evaluation results and performance, this intelligent identification model meets the actual application requirements and is deployed into the business of cyber-telecoms fraud prevention. The actual application is shown in Fig. 12.

Fig. 12 Real-time display screen of cyber-telecoms fraud prevention

It can be seen that the total number of fraudulent websites today, the total number of fraudulent websites within a week, the number of interceptions, and the total number of interceptions within a week. Moreover, the top twelve types of fraudulent websites are given, and at the same time, blocking websites with higher rankings are also displayed. In the first quarter of 2024, the cyber-telecoms fraud prevention system effectively identified 3 million + fraudulent websites and blocked 46 billion + illegal visits, to effectively reduce the incidence of fraud and protect citizens' lives and properties.

Conclusion

In this paper, a MDHES is proposed to estimate the data. In this evaluation system, multiple crucial dimensions are designed to evaluate data condition separately, which can provide a clearer understanding for improvement. Then, a comprehensive evaluation method, incorporating a fuzzy evaluation model, is developed to synthetically evaluate the data quality to achieve the dynamic balance and meanwhile mitigates the impact of subjectivity on the comprehensive evaluation result. Finally, experiment results and comparisons, in intrusion detection benchmark problem and real intelligent application identification of cyber-telecoms fraud, demonstrate the effectiveness of the proposed MDHES, which achieves an accurately and thoroughly data quality assessment to provide strong data support for trustworthy AI. In addition, when there are multiple dimensions (more than 9), the workload of AHP scaling is too large, which can easily cause confusion in judgment. Therefore, in future research, on the hand, we will focus on improving the comprehensive evaluation method to better assess the quality of data. On the other hand, more dimensions and sub-items for the big model will be considered to improve our evaluation system for promoting the application of AI in more practical scenarios in a safe and efficient manner.

Abbreviations

AI	Artificial intelligence
MDHES	Multi-dimensional hierarchical evaluation system
FDA	Fuzzy denoising autoencoder
BANG	Batch adjusted network gradients
SPAM	Scalable polynomial additive model
SE-BPER	Semantic-enhanced bayesian personalized explanation ranking
AutoML	Automated machine learning
FNN	Fuzzy neural network
AHP	Analytic hierarchy process
FNN-P	Processed training data
FNN-R	Raw training data
SVM +ELM+MK	Support vector machine and extreme learning machine based on modified k-means
BP	Back propagation
ILSTM	Improved long-short term memory
ISSDA	Improved sparse denoising autoencoder

Acknowledgements

Zhang Yusheng is acknowledged for his consulting assistance and silent accompany provided to the first author of this manuscript.

Author contributions

H.J. is the main designer of the proposed evaluation system and was also a major contributor in writing the manuscript. C.C., P.R. and K.Y. coordinates with other parties to obtain and analyze the fraud data. Z.Y., J.C., Q.C. and J.K. are mainly responsible for the web development of this multi-dimensional comprehensive evaluation system and real-time display screen of cyber-telecoms fraud prevention. All authors read and approved the final manuscript.

Funding

This work was supported by Research and standardization of key technologies for 6G general computing and intelligent integration R241149BC03, Research on 6G Trusted Endogenous Security Architecture and Key Technologies Grant

R24113V7, and Research and Standardization of Key Technologies for 6G General Computing and Intelligent Integration Grant R241149B.

Data availability

KDDCup99 dataset was derived from a simulated US Air Force LAN with attacking lasting 7 weeks and can be obtained by https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. Fraud dataset is owned by a third party. The data underlying this paper were provided by [third party] under licence/ by permission. data will be shared on request to the corresponding author with permission of [third party].

Declarations

Competing interests

The authors declare no competing interests.

Received: 11 April 2024 Accepted: 14 September 2024
Published online: 27 September 2024

References

1. Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Trans Ind Inform. 2022;18(8):5031–42.
2. Putra MA, Ahmad T, Hostiadi DP. B-CAT: a model for detecting botnet attacks using deep attack behavior analysis on network traffic flows. J Big Data. 2024;11(1):49.
3. Zhang HJ, He S, Chen J. A hierarchical authentication system for access equipment in internet of things. Int J Intell Syst. 2023;1:1–11.
4. Furman J, Seamans R. AI and the economy. Innov Policy Econ. 2019;19(1):1–191.
5. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719–31.
6. Khan A. Role of artificial intelligence in car-following and lane change models for autonomous driving. Adv Hum Asp Transp. 2018;9:307–17.
7. Schetinin V, Li D, Maple C. An evolutionary-based approach to learning multiple decision models from underrepresented data. In: Schetinin V, editor. 2008 Fourth international conference on natural computation, vol. 1. Jinan: IEEE; 2008.
8. Vavra P, Baar JV, Sanfey A. The neural basis of fairness. Interdiscip Perspect Fairness Equity Justice. 2017;5:9–31.
9. Liu HC, Wang YQ, Fan WQ, et al. Trustworthy AI: a computational perspective. ACM Trans Intell Syst Technol. 2022;14(1):1–59.
10. Chatila R, Dignum V, Fisher M, et al. Trustworthy AI. Reflect Artif Intell Hum. 2021;12600:13–39.
11. Malchiodi D, Raimondi D, Fumagalli G, et al. The role of classifers and data complexity in learned bloom flters: insights and recommendations. J Big Data. 2024;11(45):1–26.
12. Han HG, Zhang HJ, Qiao JF. Robust deep neural network using fuzzy denoising autoencoder. Int J Fuzzy Syst. 2020;22(6):1356–75.
13. Rozsa A, Gunther M, Boult TE. Towards robust deep neural networks with BANG. In: IEEE winter conference on applications of computer vision (WACV)2018.
14. Yampolskiy R. Unexplainability and Incomprehensibility of AI. J Artif Intell Conscious. 2020;7(2):1–15.
15. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. ACM Comput Surv. 2019;51(5):1–42.
16. Dubey A, Radenovic F, Mahajan D. Scalabl interpretability via polynomials. Neural Inform Process Syst. 2022;1:1–26.
17. Meng ZL, Wang MH, Bai JJ, et al. Interpreting deep learning-based networking systems. IEEE Commun Surv Tutor. 2019;21(3):2702–33.
18. Nwafor O, Okafor E, Aboushady AA, et al. Explainable artificial intelligence for prediction of non-technical losses in electricity distribution networks. IEEE Access. 2023;11:73104–15.
19. McClure P, Moraczewski D, Lam KC, et al. Improving the interpretability of fMRI decoding using deep neural networks and adversarial robustness. Apert Neuro. 2023;3:1–17.
20. Fernandes FE, Yen GG. Automatic searching and pruning of deep neural networks for medical imaging diagnostic. IEEE Trans Neural Netw Learn Syst. 2021;32(12):5664–74.
21. Barreiro E, Munteanu CR, Monteagudo MC, et al. Net–net auto machine learning (AutoML) prediction of complex ecosystems. Sci Rep. 2018;8(12340):2685–96.
22. Elliott A. What data scientists tell us about AI model training today, Alegion, 2019; 1–10.
23. Forrester Consulting. Overcome obstacles to get to AI at scale. IBM. 2020; 1–12.
24. Kortylewski A. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019; pp. 2261–2268.
25. Jackson A, The state of open data science 2020, Digital Science. 2020: 1–30.
26. Andrew NG A Chat with Andrew on MLOps: From Model-centric to Data-centric AI. 2022. https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning.
27. Zhang D, Lai H. Data-centric artificial intellgence: a survey. 2023; 1–39.
28. Zhu SC. Making mathematical models for the humanities: Chinese thought from the perspective of artificial general intelligence. J Mod Stud. 2024;3(1):42–66.

29. Liang WX, Tadesse GA, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell. 2022;4:669–77.
30. Artamonov I, Deniskina A, Filatov V, et al. Quality management assurance using data integrity model. Matec Web Conf. 2019. https://doi.org/10.1051/matecconf/201926507031.
31. Caballero I, Serrano M, Piattini M. A data quality in use model for big data. Adv Concept Model. 2014;8823:65.
32. Cai L, Zhu YY. The challenges of data quality and data quality assessment in the big data era. Data Sci J. 2015;14(2):78–92.
33. Hongxun T, Honggang W, Kun Z. Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory. In: Hongxun T, editor. 2018 IEEE 3rd international conference on cloud computing and big data analysis. Chengdu: IEEE; 2018. p. 248–52.
34. Cai L, Zhu YY. The challenges of data quality and data quality assessment in the big data era. Data Sci J. 2015;14:69–87.
35. Barchard KA, Verenikina Y. Improving data accuracy: selecting the best data checking technique. Comput Hum Behav. 2013;29(5):1917–22.
36. Li ZT, Sun JB, Yang KW, Xiong DH. A review of adversarial robustness evaluation for image classification. J Comput Res Dev. 2022;59(10):2164–89.
37. Khalfi B, de Runz C, Faiz S, Akdag H. A new methodology for storing consistent fuzzy geospatial data in big data environment. IEEE Trans Big Data. 2021;7(2):468–82.
38. Wang S, Yao X. Relationships between diversity of classification ensembles and single-class performance measures. IEEE Trans Knowl Data Eng. 2013;25(1):206–19.
39. Chae JH, Jeong YU, Kim S. Data-dependent selection of amplitude and phase equalization in a quarter-rate transmitter for memory interfaces. IEEE Trans Circuits Syst. 2020;67(9):2972–83.
40. Yao W. Research on static software defect prediction algorithm based on big data technology. In: Yao W, editor. 2020 International conference on virtual reality and intelligent systems (ICVRIS). Zhangjiajie: IEEE; 2020. p. 610–3.
41. Kim KY, Park BG. Effect of random dopant fluctuation on data retention time distribution in DRAM. IEEE Trans Electron Devices. 2021;68(11):5572–7.
42. Widad E, Saida E, Gahi Y. Quality anomaly detection using predictive techniques: an extensive big data quality framework for reliable data analysis. IEEE Access. 2023;11:103306–18.
43. Xia Q, Xu Z, Liang W, Yu S, et al. Efficient data placement and replication for QoS-aware approximate query evaluation of big data analytics. IEEE Trans Parallel Distrib Syst. 2019;30(12):2677–91.
44. Lee D. Big data quality assurance through data traceability: a case study of the national standard reference data program of Korea. IEEE Access. 2019;7:36294–9.
45. Ge Z, Liu Y. Analytic hierarchy process based fuzzy decision fusion system for model prioritization and process monitoring application. IEEE Trans Industr Inf. 2019;15(1):357–65.
46. Antal E, Tillé Y. Simple random sampling with over-replacement. J Stat Plann Inference. 2011;141(1):597–601.
47. Al-Yaseen WL, Othman ZA, Nazri MZA. Multi-level hybridsupport vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Syst Appl. 2017;67:296–303.
48. Zhang L, Yan H, Zhu Q. An improved LSTM network intrusion detection method. In: Zhang L, editor. 2020 IEEE 6th international conference on computer and communications (ICCC). Chengdu: IEEE; 2020.
49. Guo XD, Li XM, Jing RX, et al. Intrusion detection based on improved sparse denoising autoencoder. J Comput Appl. 2019;39(3):769–73.

Publisher's Note