Journal of computer science and software applications

https://www.mfacademia.org/index.php/jcssa

ISSN:2377-0430

Vol. 5, No. 5, 2025

Unsupervised Anomaly Detection in Structured Data Using Structure-Aware Diffusion Mechanisms

Honghui Xin[1], Ray Pan[2]

[1]Northeastern University, Seattle, USA

[2]Independent Researcher, Seattle, USA

*Corresponding Author: Ray Pan; raypan.research@gmail.com

Abstract:

To address the challenge of limited accuracy in detecting anomalies within complex structured data, this paper proposes a structure-aware anomaly detection method based on diffusion models. The method builds a generative diffusion framework that models the distribution of normal data through a forward noise perturbation and reverse denoising process. This enables the identification of abnormal data. To enhance the model's understanding of dependencies among fields in structured data, a Structure-Aware Diffusion (SAD) mechanism is introduced. It uses a structural matrix to explicitly model the semantic and logical relationships between fields, allowing the diffusion process to follow structural constraints. In addition, a Dynamic Reconstruction Scoring (DRS) mechanism is proposed. During the anomaly scoring phase, it dynamically adjusts weights based on the reconstruction uncertainty of different fields. This improves detection accuracy for local and sparse anomalies. Experiments on public datasets show that the proposed method outperforms traditional neural network models and baseline generative models in terms of Accuracy, AUC, and F1-score. It effectively identifies multiple types of structured anomalies. Further ablation studies confirm the significant contributions of the two proposed mechanisms. The results demonstrate the effectiveness of structure awareness and dynamic scoring in high-dimensional structured anomaly detection tasks. By integrating generative learning with structural information, this paper provides a high-accuracy, generalizable anomaly detection approach for complex relational data. The method shows strong robustness and practical value.

Keywords:

Structured anomaly detection, diffusion model, structure-aware modeling, reconstruction scoring mechanism

1. Introduction

With the proliferation of data-driven applications across domains such as finance, telecommunications, healthcare, and e-commerce, the volume and structural complexity of digital data have increased substantially [1]. Ensuring the integrity, consistency, and reliability of such data has become a critical prerequisite for stable system operations and informed decision-making. However, the growing complexity of data systems has also led to a rise in anomalous behaviors, including irregular patterns, inconsistent values, and unexpected interactions. These anomalies, if undetected, may degrade system performance, compromise decision accuracy, and pose significant operational and security risks. As such, the development of accurate and efficient anomaly detection methods for complex structured data is of significant practical importance.

Conventional anomaly detection techniques—such as rule-based methods, statistical modeling, and classical machine learning—have been widely used in various scenarios [2]. While these approaches can be effective in narrowly defined contexts, they often suffer from limited generalizability, sensitivity to noise, and reliance on expert knowledge. Their performance typically declines in high-dimensional, heterogeneous, and dynamically evolving environments, where anomalous patterns may be subtle, sparse, or nonstationary. These limitations underscore the need for more adaptive and robust detection frameworks capable of capturing complex data dependencies and rare deviations without extensive manual intervention [3].

Recent advances in generative modeling offer promising avenues for addressing these challenges. In particular, diffusion models have emerged as a powerful class of deep generative models capable of learning rich data distributions through iterative noise perturbation and denoising processes. Their capacity to capture fine-grained structures and to generate realistic reconstructions has led to significant success in domains such as image synthesis, speech modeling, and natural language processing. These capabilities make diffusion models well-suited for detecting anomalies that deviate from learned data distributions, particularly in unsupervised or semi-supervised settings where labeled anomalies are scarce [4].

When applied to structured data, diffusion models can be enhanced with mechanisms that explicitly capture inter-field dependencies and semantic relationships. Such enhancements enable the modeling process to preserve contextual information and improve sensitivity to structurally meaningful anomalies [5]. Moreover, generative approaches inherently support interpretability, as reconstruction deviations can provide insights into the nature and location of anomalies.

In this context, the integration of structure-aware generative modeling and dynamic scoring strategies represents a promising direction for anomaly detection in complex data environments. By leveraging the strengths of diffusion models in distribution learning and reconstruction, such approaches can advance the development of intelligent, self-adaptive monitoring systems. These systems are expected to play a critical role in ensuring the robustness, reliability, and security of next-generation data infrastructures.

2. Related work

2.1 Anomaly Detection

Anomaly detection is a key research area in data mining and machine learning. It aims to identify samples or patterns that deviate significantly from normal behavior. Traditional methods often rely on statistical modeling, rule matching, or clustering techniques based on distance and density. These approaches model data features to determine whether a sample is anomalous[6]. They perform reasonably well in low-dimensional and simple-structured scenarios. However, in high-dimensional spaces, with complex nonlinear relationships or noisy data, these methods often fail to model accurately. This leads to a decline in detection performance. In addition, such methods typically require preset parameters or expert knowledge, making it hard to adapt to changing data characteristics[7].

With the advancement of machine learning, especially deep learning, model-based anomaly detection has gained popularity[8,9]. Neural networks, autoencoders, and generative models have shown strong capabilities in extracting high-dimensional features and capturing complex patterns. Compared to traditional methods, these models offer better nonlinear representation and higher detection accuracy. They are particularly suitable for complex data scenarios such as network traffic analysis, graph-based data monitoring, and industrial sensor anomaly detection. However, their application to relational structured data still faces challenges. These include maintaining logical relationships between data, handling missing values and noise, and interpreting detection results effectively[10].

ISSN:2377-0430

Vol. 5, No. 5, 2025

To improve anomaly detection in complex data, generative models have been increasingly adopted. They show strong potential, especially in unsupervised or semi-supervised tasks. Generative models learn the distribution of normal data and use it to detect samples that do not conform to this distribution[11]. Representative methods such as variational autoencoders and generative adversarial networks have achieved significant results in unstructured data like images and text. In contrast, diffusion models offer a unique approach through gradual perturbation and reconstruction. They provide a new way to model data distributions and detect anomalies. This makes them particularly effective in identifying samples that significantly deviate from normal patterns. As a result, diffusion models offer promising techniques and insights for anomaly detection in structured data.

2.2 Diffusion Model

As an emerging method in deep generative modeling, diffusion models have gained increasing attention in recent years[12,13,14]. The core idea is to gradually add noise to the data, transforming it into a standard Gaussian distribution, and then recover the original data distribution through a learned denoising process. Compared to traditional generative models, diffusion models offer stronger modeling capabilities and more stable training. They can effectively capture complex structures and fine-grained features in data. This progressive generation process improves the quality of generated data and provides a more intuitive framework for understanding and controlling the generation[15]. As a result, diffusion models have been widely applied in image, speech, and text domains.

With the advancement of diffusion models, researchers have expanded their application scope. Initially focused on unstructured data generation, they have since been extended to graph modeling, time series forecasting, and complex distribution learning[16,17]. A key advantage of diffusion models is their high-fidelity fitting of data distributions. This makes them particularly effective in anomaly modeling. By learning the evolution path of normal data, the model can identify abnormal patterns during the reverse reconstruction process[18]. These patterns are often difficult to recover, enabling efficient distinction of anomalous samples. In addition, the diffusion process itself allows for the interpretation and visualization of anomalies. This gives the model both detection capability and a degree of interpretability[19].

In structured data scenarios, especially in relational databases, the use of diffusion models is still in an exploratory phase[20,21]. However, their strong performance in other data types provides a solid foundation for extension. Structured data is characterized by clear patterns and dependencies between fields. When modeling these internal structures, diffusion models can be designed with tailored noise mechanisms and network architectures to preserve semantic consistency and logical relationships. With their powerful generative ability and flexible modeling framework, diffusion models hold promise as a key technique for anomaly detection in relational databases. They can drive the intelligent evolution of database systems and offer new solutions for complex data modeling.

3. Method

This study proposes a diffusion-based anomaly detection algorithm tailored for complex structured data, aiming to enhance detection accuracy by capturing underlying dependencies and distributional deviations. It aims to improve the identification of abnormal patterns in complex structured data using deep generative modeling capabilities. The method first preprocesses and embeds multi-dimensional data from relational databases. It then constructs vector representations suitable for the diffusion process. Through forward noise perturbation and reverse denoising reconstruction, the model learns the latent distribution of normal data. Reconstruction error is used as the basis for anomaly detection. Two innovations are introduced to enhance

ISSN:2377-0430

Vol. 5, No. 5, 2025

model performance and adaptability. The first is Structure-Aware Diffusion (SAD). This mechanism incorporates structural attention based on field relationships within the data. It ensures logical consistency between fields during the diffusion process. The second is Dynamic Reconstruction Scoring (DRS). During the denoising phase, it dynamically adjusts reconstruction weights across dimensions. This improves sensitivity to local anomalies and rare patterns. The proposed algorithm balances modeling accuracy with structural adaptability. It offers a robust and scalable solution for anomaly detection in structured data. The model architecture is shown in Figure 1.

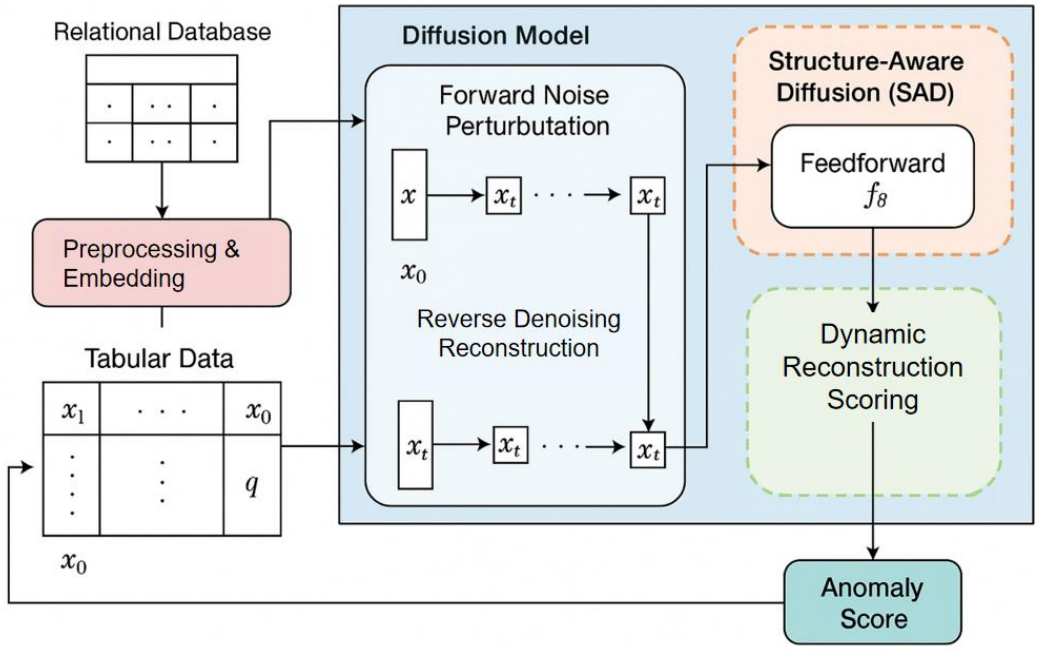

Figure 1. Overall model architecture diagram

3.1 Structure-Aware Diffusion

In relational databases, there are complex logical dependencies and structural relationships between data fields. Simply treating tabular data as an independent set of vectors will ignore this structural information, thus affecting the accuracy of anomaly detection. To this end, we introduce the Structure-Aware Diffusion (SAD) mechanism to enhance the model's ability to express dependency information within the table by fusing the relationship structure between fields during the diffusion modeling process. In each diffusion step, the model not only learns the evolution process of the data itself, but also perceives the correlation between columns through the structural encoding mechanism to improve the modeling ability of structural anomalies. Its module architecture is shown in Figure 2.

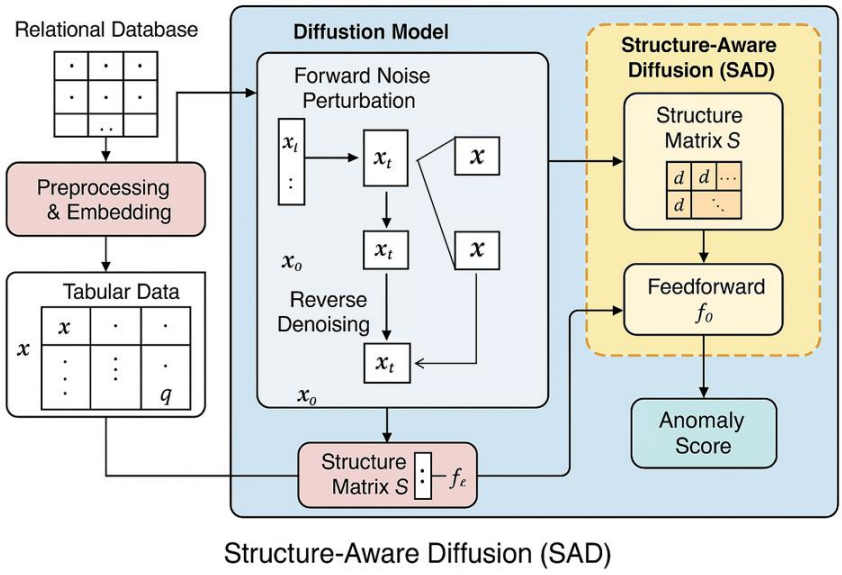

Figure 2. SAD module architecture

Let the original input table data be $X = [x_1, x_2, ..., x_n] \in R^{n \times d}$, where n represents the number of samples and d is the field dimension. In the forward diffusion process, we add noise to each time step t, defined as:

$$q(x_t \mid x_0) = N(x_t; \sqrt{a_t} x_0, (1 - a_t) I)$$

Where $a_t \in (0,1)$ represents the noise attenuation coefficient, x_0 is the initial input, and x_t is the perturbation representation of the t-th step. At the same time, in order to embed structural information, we introduce a structure-aware embedding matrix $S \in R^{d \times d}$, which represents the dependency strength between fields. A structural attention term is added in each diffusion step, so that the network output depends on the weights between fields:

$$x_t = x_t + \lambda S_{x_t}$$

Where λ is the structural adjustment coefficient, which controls the influence of structural perception information on the current representation. This operation forces the model to explicitly learn the interaction between fields during the perturbation process, which helps to more accurately distinguish structural anomalies in the reverse reconstruction stage.

In the reverse process $p_\theta(x_{t-1} \mid x_t)$, we use the conditional denoising network to estimate the noise at each time step and simultaneously fuse the structural information to guide the reconstruction path. The reconstruction process is defined as:

$$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t, S), \sum_\theta(x_t, t))$$

ISSN:2377-0430

Vol. 5, No. 5, 2025

Here μ_θ is the mean estimate output by the neural network, which explicitly depends on the structure matrix S; while $\sum_\theta$ is the prediction variance, which is used to capture the reconstruction uncertainty. After multi-step reconstruction, $\tilde{x}_0$ is finally obtained, and the difference is compared with the original data x_0 as the basis for subsequent abnormality discrimination.

To better adapt to multi-type fields, the structure-aware mechanism further introduces field type embedding representation $E \in R^{d \times h}$, which is obtained through the projection function $\phi : R^d \to R^h$:

$$e_j = \phi(x_j), j = 1,2,...,d$$

The structure matrix S is obtained by calculating the similarity of field embedding, and its form is:

$$S_{i,j} = \frac{e_i^T e_j}{\| e_i \| \cdot \| e_j \|}$$

This design can adaptively capture implicit semantic relationships between fields, making the diffusion process more structurally sensitive and generalizable, and effectively improving the model's ability to identify potential complex anomalies.

3.2 Dynamic Reconstruction Scoring

In relational database anomaly detection, different fields have significant differences in sensitivity to anomalies, and a unified reconstruction error calculation method is often difficult to fully characterize local or rare anomalies. To this end, this study proposes a dynamic reconstruction scoring mechanism (DRS), which introduces dynamic weights and dimensional sensitivity estimates to weighted aggregate reconstruction errors of different fields, thereby enhancing the model's ability to respond to fine-grained anomalies. This mechanism is embedded in the denoising stage of the diffusion model, and automatically adjusts the scoring strategy based on the reconstruction difficulty and uncertainty of each sample in each field dimension.

Assume that the final reconstructed sample of the model is $\tilde{x}_0 \in R^d$, the original input is $x_0 \in R^d$, and its initial reconstruction error vector is defined as:

$$e = | \tilde{x}_0 - x_0 |$$

In order to dynamically perceive the reconstruction sensitivity of different dimensions, the field confidence vector $\gamma \in R^d$ is introduced, each dimension of which is estimated by the cumulative reconstruction variance at each time step during the diffusion process:

$$\gamma_j = \frac{1}{T} \sum_{t=1}^{T} \sigma_\theta^2(x_t)_j$$

Where $\sigma_\theta^2(x_t)_j$ represents the reconstruction uncertainty of the jth dimension at the tth step, and T is the number of diffusion steps. The dynamic weight vector $w \in R^d$ is constructed according to the confidence, which is defined as:

ISSN:2377-0430

Vol. 5, No. 5, 2025

$$w_j = \frac{1}{\gamma_j + \varepsilon}$$

ε is a smoothing factor, which is used to avoid instability caused by a denominator of zero or too small. This weight vector can dynamically focus on reconstructing dimensions with smaller variance and more reliable structure, effectively improving the ability to distinguish abnormal signals.

The final anomaly score is composed of the weighted sum of the reconstruction errors in each dimension:

$$A(x) = \sum_{j=1}^{d} w_j \cdot e_j$$

This scoring function can adjust the importance of different dimensions according to the actual reconstruction performance to avoid the dilution of abnormal signals by the average effect. At the same time, in order to further enhance the robustness of anomaly detection, a standardization operation is introduced to normalize the reconstruction error of each dimension:

$$\tilde{e}_j = \frac{e_j - \mu_j}{\sigma_j}$$

μ_j and σ_j are the mean and standard deviation of the j-th dimension reconstruction error in the training set. This mechanism enables the model to not only detect global structural mutations, but also have the ability to perceive tiny local anomalies, thereby achieving higher-resolution anomaly recognition.

4. Experimental Results

4.1 Dataset

This study employs the publicly available KDD Cup 1999 dataset as the primary experimental benchmark for evaluating the proposed anomaly detection approach. Originally developed for network intrusion detection, this dataset has been extensively utilized in structured anomaly detection research due to its diverse range of abnormal patterns and well-characterized normal behavior. It is organized in a structured tabular format, comprising 41 features that span discrete, continuous, and categorical variables, including attributes such as connection type, service protocol, and data transfer volume.

With nearly five million labeled connection records, the dataset presents a highly imbalanced class distribution, where the majority of instances are normal, and the remainder belong to various attack categories such as Denial-of-Service (DoS), Probe, User-to-Root (U2R), and Remote-to-Local (R2L). This imbalance poses significant challenges for anomaly detection methods, particularly in identifying rare or subtle anomalies. Each record can be interpreted as a high-dimensional structured instance, potentially containing latent dependencies among features, making the dataset suitable for testing models that incorporate structure-aware mechanisms.

To ensure effective model training, redundant or low-contribution features are commonly removed or normalized, and categorical attributes are embedded into continuous vector representations compatible with neural architectures. While not inherently relational in nature, the KDD Cup 1999 dataset provides a well-

ISSN:2377-0430

Vol. 5, No. 5, 2025

structured, labeled, and widely recognized testbed for assessing detection accuracy, generalization capability, and robustness in structured anomaly detection tasks.

4.2 Experimental setup

This study evaluates the proposed anomaly detection method based on the diffusion model in a unified experimental environment to ensure the reproducibility and fairness of the results. All experiments are conducted on a hardware platform with consistent configuration, including a high-performance GPU to accelerate the model training and inference process. In order to enhance the stability and generalization ability of the model, standardized preprocessing and Early Stopping strategies are used in the training process, and the data set is divided into training and test sets according to the conventional division ratio. In the anomaly detection evaluation, the three indicators Acc, Auc and F1-score are mainly referenced to measure the detection performance of the model on different anomaly categories.

In addition, the key parameters of the model are determined by cross-validation, and the number of diffusion steps, noise adjustment factors, structural perception coefficients and dynamic reconstruction weights are all tuned within a reasonable range. The comparison models include traditional statistical methods and advanced deep learning anomaly detection methods to ensure that the evaluation is widely representative. The main experimental hyperparameter settings are shown in Table 1.

Table 1: Hyperparameter setting

Parameter	**Value**
Hardware	NVIDIA RTX 3090 GPU, 128GB RAM, 32 CPU cores
OS & Framework	Ubuntu 20.04, Python 3.10, PyTorch 2.0
Dataset Split	70% Training / 30% Testing
Diffusion Steps (T)	1000
Learning Rate	1e-4
Batch Size	256
Structure Coefficient (λ)	0.5
Weight Smoothing	1e-6
Optimizer	Adam

4.3 Experimental Results

1) *Comparative experimental results*

ISSN:2377-0430

Vol. 5, No. 5, 2025

First, this paper presents a set of comparative experiments conducted with several existing models. These experiments are designed to evaluate the performance of the proposed method under consistent conditions. The comparative results, which reflect the effectiveness of different approaches, are organized and displayed in Table 2.

Table 2: Comparative experimental results

Method	**Acc**	**Auc**	**F1-Score**
MLP[22]	0.872	0.903	0.854
LSTM[23]	0.881	0.915	0.862
CNN[24]	0.889	0.921	0.870
Transformer[25]	0.893	0.930	0.878
CNN+Transformer[26]	0.901	0.937	0.884
Ours	0.918	0.953	0.902

The experimental results show that traditional neural network models, such as MLP and LSTM, have demonstrated certain effectiveness in anomaly detection on structured data. Their F1-scores reached 0.854 and 0.862, respectively. However, these models have limitations in capturing complex dependencies among fields. They are relatively insensitive to data structure and struggle to detect anomalies that arise from coordinated changes across related fields in relational databases. This limitation is more evident in scenarios with sparse anomaly distributions or weak inter-field dependencies.

In contrast, models like CNN and Transformer exhibit better generalization when handling spatial locality and long-range dependencies. The Transformer model, in particular, benefits from the self-attention mechanism, which captures hidden relationships between fields. This results in improved AUC and F1-score, reaching 0.930 and 0.878, respectively. However, despite its global modeling capabilities, the Transformer architecture does not fully incorporate domain-specific structural information. When applied to tasks involving explicit logical relationships among database fields, it lacks the granularity needed for precise modeling. This leads to blurred detection boundaries and reduced sensitivity to subtle anomalies.

The model that combines CNN and Transformer further improves performance. It outperforms each standalone model across all three metrics. This suggests that combining local perception and global attention enhances anomaly detection to some extent. The F1-score achieved 0.884. Nevertheless, without a structure-aware mechanism, this architecture still struggles to fully adapt to the complex organization of fields in relational databases. In cases where anomalies are subtle and affect only local fields, the model's recognition ability remains limited.

The proposed Structure-Aware Diffusion (SAD) model combined with the Dynamic Reconstruction Scoring (DRS) mechanism achieved the best results across all evaluation metrics. The F1-score reached 0.902. By embedding structural information during the diffusion process and dynamically adjusting the importance of each field during scoring, the method significantly improves the modeling and detection of local, rare, and structure-dependent anomalies. These results confirm the advantages of integrating structure-aware

mechanisms with generative modeling. The proposed approach provides a solid technical foundation for applying anomaly detection in real-world relational database systems.

2) Ablation Experiment Results

Secondly, this paper provides the results of ablation experiments to evaluate the individual contributions of key components within the proposed model. These experiments help analyze the impact of each module by systematically modifying the model structure. The detailed results of the ablation study are presented in Table 3.

Table 3: Ablation Experiment Results

Method	**Acc**	**Auc**	**F1-Score**
BaseLine	0.891	0.926	0.872
+SAD	0.903	0.938	0.884
+DRS	0.898	0.934	0.879
Ours	0.918	0.953	0.902

As shown in the ablation results in Table 3, the BaseLine model, without structure-aware and dynamic reconstruction mechanisms, achieves acceptable performance in relational database anomaly detection. However, its overall effectiveness is limited, with an F1-score of only 0.872. This indicates that relying solely on the basic generative capacity of the diffusion model is insufficient to fully capture semantic relations and local structural features among database fields. As a result, the model struggles to distinguish anomalous samples accurately.

When the structure-aware diffusion module (+SAD) is added to the BaseLine model, performance improves significantly. The F1-score rises to 0.884, and AUC increases to 0.938. These results confirm the effectiveness of the SAD module in modeling field dependencies. It allows the model to retain structural information between fields during the denoising process, which enhances sensitivity to structural anomalies. The SAD module enables the diffusion process to account for not only numerical perturbations but also relational constraints. This is especially important for identifying logical anomalies in relational databases.

On the other hand, introducing the dynamic reconstruction scoring mechanism (+DRS) also leads to performance gains. The F1-score improves to 0.879. Unlike SAD, DRS focuses on the anomaly scoring phase. It dynamically adjusts reconstruction weights for each field, improving the model's response to local changes and fine-grained anomalies. In scenarios where anomalies affect only a few fields, DRS helps prevent anomaly signals from being diluted by global error averaging, thereby increasing detection accuracy.

When SAD and DRS are used together, the model achieves the best performance across all three evaluation metrics. The F1-score reaches 0.902. This shows that the two modules complement each other. SAD enhances the expressiveness of anomaly modeling, while DRS improves discriminative power during detection. The final results demonstrate the importance and practical value of the dual-module design proposed in this study for anomaly detection in relational databases.

3) Effect of different diffusion steps on detection performance

ISSN:2377-0430

Vol. 5, No. 5, 2025

This paper also investigates the effect of varying the number of diffusion steps on the overall detection performance of the proposed model. By systematically adjusting the number of steps in the diffusion process, the study aims to understand how the depth of the generative process influences the model's ability to capture data distribution and identify anomalies. This analysis provides insights into the relationship between diffusion step configuration and detection accuracy. The corresponding experimental results that illustrate this impact are comprehensively presented in Figure 3.

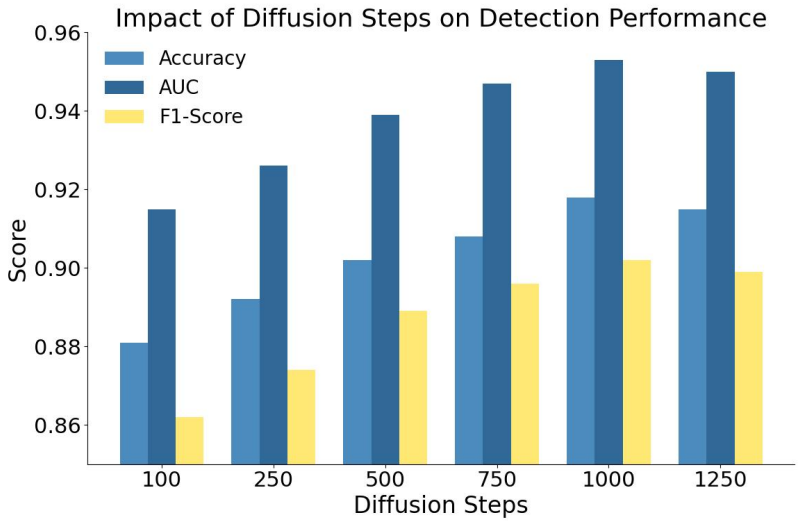

Figure 3. Effect of different diffusion steps on detection performance

The results in the figure show that the number of diffusion steps has a significant impact on anomaly detection performance. As the number of steps increases, the model shows steady improvement in Accuracy, AUC, and F1-score. This suggests that a longer diffusion process helps the model better learn the underlying data distribution, leading to more accurate reconstruction of normal patterns and identification of anomalies.

Between 500 and 1000 steps, the performance improvement is particularly notable. In this range, the model receives sufficient perturbation information without being affected by excessive noise. This demonstrates the robustness of the diffusion model in capturing anomalous patterns. The result indicates that anomalies in relational databases often involve coordinated changes across multiple fields. A moderate number of diffusion steps enhances the model's ability to learn such high-dimensional structural dependencies.

When the number of steps reaches 1000, the model achieves optimal performance. The F1-score and AUC reach 0.902 and 0.953, respectively. This shows that the method improves not only global accuracy but also the detection of sparse, subtle, and structure-sensitive anomalies. The introduction of the Structure-Aware Diffusion (SAD) and Dynamic Reconstruction Scoring (DRS) mechanisms enables dynamic capture of each field's contribution during diffusion. This strengthens the model's ability to detect complex anomaly distributions.

However, when the number of steps increases further to 1250, the performance slightly declines or stabilizes. This suggests that an overly long diffusion chain may lead to reconstruction quality fluctuations or overfitting, which interferes with accurate anomaly detection. Therefore, selecting an appropriate number of diffusion steps is critical. It should balance data characteristics and computational efficiency. These findings further confirm the rationality and practical value of the structure-aware and generative modeling design proposed in this study.

4) Analysis of the impact of embedding dimension selection on model performance

This paper also provides a detailed analysis of how the selection of embedding dimension affects the overall performance of the proposed model. The study systematically explores different embedding dimensions to examine their influence on the model's capacity to represent structured data and capture complex relationships among fields. By evaluating multiple configurations, the analysis aims to identify an appropriate embedding size that balances representational richness and computational efficiency. The experimental setup and outcomes related to this investigation are thoroughly illustrated in Figure 4.

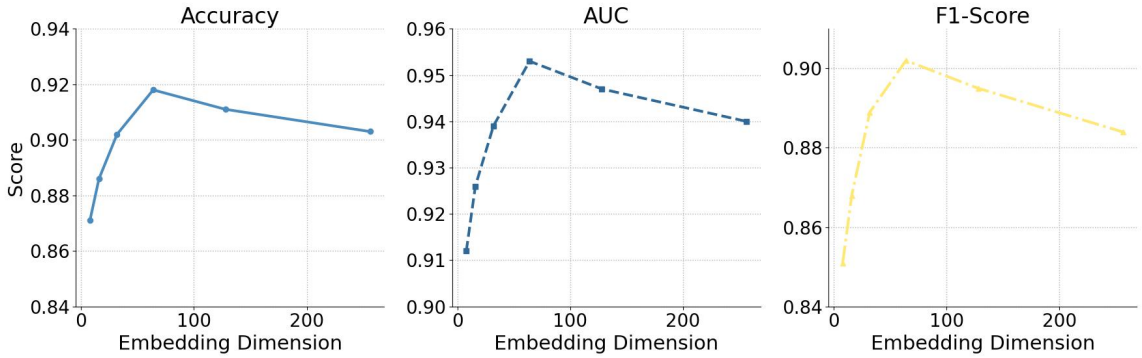

Figure 4. Analysis of the impact of embedding dimension selection on model performance

As shown in Figure 4, the embedding dimension has a clear impact on model performance. With increasing dimension, Accuracy, AUC, and F1-score all show an upward trend in the early stages. This indicates that low-dimensional embeddings are insufficient to capture the complex semantic and structural dependencies among fields in relational databases, which limits the model's expressive power. Increasing the dimension allows the model to better represent potential relationships between fields, which helps in identifying anomalous patterns.

When the embedding dimension reaches 64, all three metrics peak. The model achieves its best performance, with an F1-score of 0.902 and an AUC of 0.953. This shows that at this level, the model retains enough semantic information while maintaining effective parameter control. The structure-aware diffusion model learns field-level structural relations more stably and captures anomaly signals during reverse reconstruction. A 64-dimensional embedding offers a good trade-off between modeling capacity and computational cost, making it suitable for modeling high-dimensional relational data.

When the embedding dimension increases further to 128 or 256, the model performance slightly declines. This is especially evident in the drop of the F1-score. The results suggest that excessively high dimensions may introduce redundant features, add noise during learning, and increase the risk of overfitting in the diffusion process. This reduces the model's generalization ability in detecting anomalies. The problem is more pronounced when training samples are limited or anomalies are sparse, leading to less precise detection boundaries.

These results confirm that the embedding dimension plays a crucial role as a key hyperparameter in the modeling of structured data. The choice of embedding size has a direct and significant impact on both the representational capacity of the model and its ability to accurately detect anomalies. An appropriately selected embedding dimension enables the model to capture complex relationships among fields while maintaining computational efficiency. In this study, the proposed method demonstrates that using a moderate embedding size strikes a balance between sufficient semantic representation and model stability. This finding

highlights the importance of embedding design in structured anomaly detection tasks. Moreover, it suggests that integrating structure-aware mechanisms with embedding optimization forms an effective strategy to enhance the overall performance of diffusion-based models. Such a combination supports not only accurate anomaly identification but also improves the model's generalization across diverse data conditions.

5) The impact of anomaly type distribution on model detection accuracy

This paper also explores the impact of anomaly type distribution on the detection accuracy of the proposed model. Different categories of anomalies can vary significantly in terms of their structural patterns, frequency, and visibility within the data, which may affect how effectively a model can identify them. To better understand this relationship, the study conducts experiments across multiple anomaly types, allowing for a comprehensive evaluation of the model's adaptability and sensitivity to diverse abnormal behaviors. This analysis is crucial for assessing the robustness of the model in real-world scenarios where anomalies may be unevenly distributed and exhibit different levels of complexity. The corresponding experimental results that illustrate the influence of anomaly type distribution on model performance are presented in detail in Figure 5.

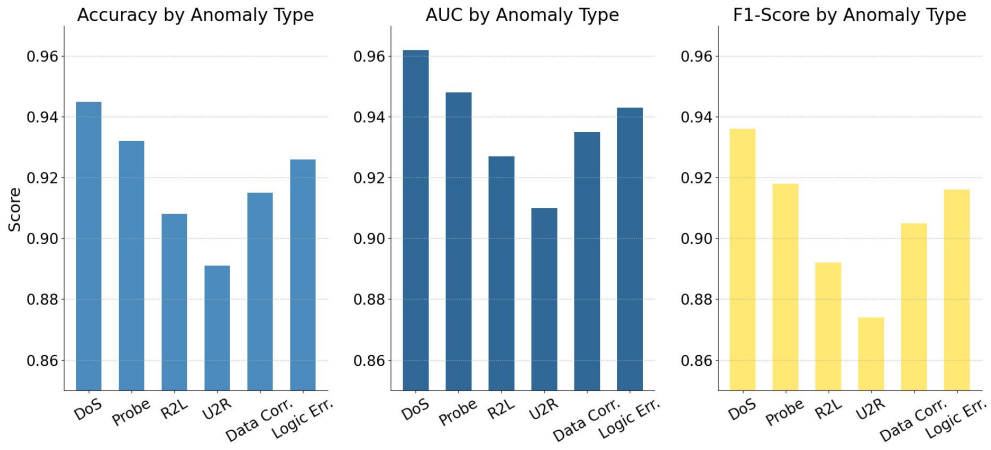

Figure 5. The impact of anomaly type distribution on model detection accuracy

The experimental results in Figure 5 show clear differences in detection accuracy across various types of anomalies. DoS and Probe anomalies perform well across all three metrics: Accuracy, AUC, and F1-score. This indicates that the model can accurately distinguish these types from normal behavior. Such anomalies often involve significant changes across multiple fields. They have strong structural patterns, making them easier to detect through diffusion modeling and structure-aware mechanisms.

In contrast, the detection performance for U2R and R2L anomalies is relatively low, especially in terms of F1-score. These anomalies tend to be more covert. They affect fewer fields and often cause only minor shifts in certain discrete attributes. This presents challenges for the reconstruction process. Although the model integrates structure-aware diffusion and dynamic scoring, sparse and subtle anomalies like these are still prone to false negatives or misclassifications. This highlights the need for further optimization when dealing with low-frequency and hidden anomalies.

In addition, the detection of Logic Error and Data Corruption anomalies achieves moderate performance. These anomalies are usually non-attacking structural issues. They often involve inconsistencies in field values, invalid combinations, or rule violations. After modeling field relationships using the structure matrix

S, the model can partially capture the deviation patterns. However, detection still depends on the precision of structural modeling and the accuracy of embedding representations.

Overall, the results confirm that the type of anomaly directly impacts detection performance. The proposed method shows clear advantages in detecting structural mutations and highly distinctive anomalies. However, in cases of sparse or low-visibility anomalies, further improvement is needed. Enhancing fine-grained structural modeling and increasing sensitivity to local patterns may help boost overall detection capability. These findings provide a clear direction for future optimization.

6) Loss function changes with epoch

This paper also gives a graph of the loss function changing with epoch, as shown in Figure 6.

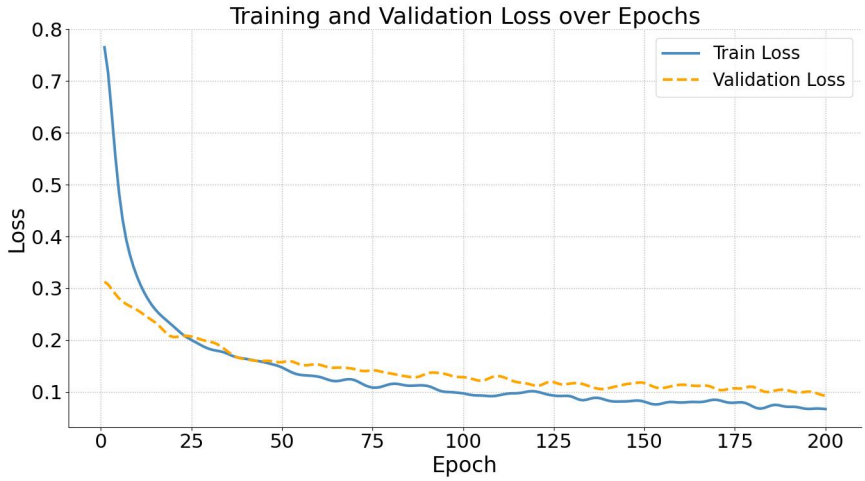

Figure 6. Loss function changes with epoch

As shown in Figure 6, the proposed model exhibits a consistent downward trend in the loss function during training. This indicates that the model is gradually learning the true data distribution in the relational database. The training loss decreases rapidly within the first 50 epochs, reflecting the model's strong ability to quickly fit the distribution of normal samples at an early stage. This aligns with the powerful data modeling capacity of diffusion models, especially for high-dimensional structured data.

Between 50 and 150 epochs, the decrease in training loss slows down and gradually stabilizes. The validation loss also remains relatively steady during this phase. This suggests that the model has effectively learned the pattern features on the training set and generalizes well to unseen data. The combination of Structure-Aware Diffusion (SAD) and Dynamic Reconstruction Scoring (DRS) helps build robust representations across different fields, avoiding performance fluctuations due to overfitting.

The curve shows that training loss is consistently lower than validation loss, but the gap is small and both curves follow a similar trend. This further indicates that the model maintains good structural alignment and anomaly modeling capability throughout training. The absence of a significant rise in the validation loss curve suggests that no clear overfitting occurred during the training process. This is particularly important for anomaly detection tasks in relational databases, where data sparsity and type diversity are common.

Overall, the figure confirms that the diffusion model demonstrates good convergence and training stability in structured data anomaly detection. It also shows that, with appropriate structural guidance and dynamic

scoring, the model can maintain expressive power while ensuring a well-controlled loss convergence process. This leads to higher detection accuracy and improved robustness.

5. Conclusion

This study presents an anomaly detection framework tailored for high-dimensional structured data, grounded in a generative modeling paradigm based on diffusion processes. The proposed method introduces two novel components: a Structure-Aware Diffusion (SAD) mechanism and a Dynamic Reconstruction Scoring (DRS) mechanism. The SAD module integrates inter-feature dependencies directly into the diffusion trajectory, enabling the model to preserve and exploit structural relationships within the data. The DRS module further refines detection sensitivity by dynamically weighting reconstruction errors according to dimension-specific uncertainty. Through a forward noise perturbation and reverse denoising process, the model effectively learns the underlying data distribution, thereby enhancing its capacity to identify anomalous patterns. Experimental evaluations across multiple metrics demonstrate that the proposed approach consistently surpasses conventional neural architectures in anomaly detection accuracy, particularly in the presence of sparse or structure-dependent anomalies. By explicitly encoding feature dependencies and incorporating adaptive scoring strategies, the method addresses key limitations of existing approaches—namely, the inability to generalize across complex data distributions and the lack of interpretability in anomaly attribution. The model's structural awareness facilitates precise modeling of semantic correlations among features, while the dynamic scoring mechanism ensures robustness against noise and local irregularities. This dual-modular design yields a comprehensive framework capable of detecting both global and fine-grained anomalies, offering significant practical utility in real-world data analysis settings.

The proposed method is particularly applicable to domains characterized by complex structured data and stringent accuracy requirements, such as financial fraud detection, network security monitoring, clinical anomaly screening, and large-scale enterprise system diagnostics. In such contexts, the combination of structural sensitivity, generative flexibility, and scoring adaptivity provides a compelling solution to the limitations of traditional rule-based or shallow learning techniques. Furthermore, the model's unsupervised nature reduces dependence on labeled data, making it well-suited for deployment in environments where anomaly labels are scarce or evolving. Future research directions include extending the framework to accommodate graph-structured inputs, integrating temporal modeling for sequential anomaly detection, and exploring contrastive or self-supervised pretraining to enhance generalization in low-data regimes. Additionally, the adaptation of lightweight diffusion models for streaming or online inference scenarios warrants further investigation, particularly in latency-sensitive applications. These extensions have the potential to expand the framework's applicability and further solidify its role in the next generation of intelligent, adaptive anomaly detection systems.

References

[1] Gadde, Hemanth. "AI-Driven Anomaly Detection in NoSQL Databases for Enhanced Security." International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence 14.1 (2023): 497-522.

[2] Singh, Jatin Pal. "Enhancing Database Security: A Machine Learning Approach to Anomaly Detection in NoSQL Systems." International Journal of Information and Cybersecurity 7.1 (2023): 40-57.

[3] Landauer, Max, et al. "Deep learning for anomaly detection in log data: A survey." Machine Learning with Applications 12 (2023): 100470.

[4] Quatrini, Elena, et al. "Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities." Journal of Manufacturing Systems 56 (2020): 117-132.

[5] Gümüşbaş, Dilara, et al. "A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems." IEEE Systems Journal 15.2 (2020): 1717-1731.

[6] Lim, Willone, et al. "Future of generative adversarial networks (GAN) for anomaly detection in network security: A review." Computers & Security 139 (2024): 103733.

[7] Trilles, Sergio, Sahibzada Saadoon Hammad, and Ditsuhi Iskandaryan. "Anomaly detection based on artificial intelligence of things: A systematic literature mapping." Internet of Things 25 (2024): 101063.

[8] Cao, Yunkang, et al. "A survey on visual anomaly detection: Challenge, approach, and prospect." arXiv preprint arXiv:2401.16402 (2024).

[9] Li, Xiaofan, et al. "Promptad: Learning prompts with only normal samples for few-shot anomaly detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[10] Adhikari, Deepak, et al. "Recent advances in anomaly detection in Internet of Things: Status, challenges, and perspectives." Computer Science Review 54 (2024): 100665.

[11] Altulaihan, Esra, Mohammed Amin Almaiah, and Ahmed Aljughaiman. "Anomaly detection IDS for detecting DoS attacks in IoT networks based on machine learning algorithms." Sensors 24.2 (2024): 713.

[12] Huang, Yi, et al. "Diffusion model-based image editing: A survey." arXiv preprint arXiv:2402.17525 (2024).

[13] Fuest, Michael, et al. "Diffusion models and representation learning: A survey." arXiv preprint arXiv:2407.00783 (2024).

[14] Cao, Hanqun, et al. "A survey on generative diffusion models." IEEE Transactions on Knowledge and Data Engineering (2024).

[15] Mousakhan, Arian, Thomas Brox, and Jawad Tayyub. "Anomaly detection with conditioned denoising diffusion models." DAGM German Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2024.

[16] Qi, Pian, et al. "Model aggregation techniques in federated learning: A comprehensive survey." Future Generation Computer Systems 150 (2024): 272-293.

[17] Yao, Hang, et al. "GLAD: towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

[18] Liu, Jing, et al. "A survey on diffusion models for anomaly detection." arXiv preprint arXiv:2501.11430 (2025).

[19] Tebbe, Justin, and Jawad Tayyub. "Dynamic addition of noise in a diffusion model for anomaly detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[20] Fučka, Matic, Vitjan Zavrtanik, and Danijel Skočaj. "TransFusion–a transparency-based diffusion model for anomaly detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

[21] Wei, Changyun, et al. "TDAD: Self-supervised industrial anomaly detection with a two-stage diffusion model." Computers in Industry 164 (2025): 104192.

[22] Zhong, Zhijie, et al. "PatchAD: A lightweight patch-based MLP-mixer for time series anomaly detection." arXiv preprint arXiv:2401.09793 (2024).

[23] Shrestha, Rakesh, et al. "Anomaly detection based on lstm and autoencoders using federated learning in smart electric grid." Journal of Parallel and Distributed Computing 193 (2024): 104951.

[24] Zhang, Jiajia, et al. "A light CNN based on residual learning and background estimation for hyperspectral anomaly detection." International Journal of Applied Earth Observation and Geoinformation 132 (2024): 104069.

ISSN:2377-0430

Vol. 5, No. 5, 2025

[25] Ma, Mingrui, Lansheng Han, and Chunjie Zhou. "Research and application of Transformer based anomaly detection model: A literature review." arXiv preprint arXiv:2402.08975 (2024).

[26] Zhu, Bingke, et al. "ADFormer: Generalizable Few-Shot Anomaly Detection with Dual CNN-Transformer Architecture." IEEE Transactions on Instrumentation and Measurement (2024).