Computers and Artificial Intelligence; Vol.3, No.4; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n4.207219
Published by: Zeus Press

A Survey of Anomaly Detection Methods Based on Clustering Analysis and Deep Learning

Jiawen Peng*

School of Software & Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang 330013, China

**Corresponding author: Jiawen Peng.*

Abstract

As information systems continuously expand, such as networks, the Internet of Things, and cloud computing, security and economic losses, as a consequence of anomalous events, have become increasingly prominent, and anomaly detection has become a fundamental competence for ensuring system reliability and security. High-dimensional, strongly nonlinear, and strongly temporally dependent data restrict the performance of anomaly detection methods based on a single technical route. Within this background, this research provides a formal description of anomaly detection and a classification of anomaly types, and categorizes time-series methods into six paradigms: clustering, prediction, reconstruction, representation learning, statistical, and hybrid. The paper expatiates on the anomaly discrimination mechanisms, parameter sensitivity, and complexity constraints of partitioning-, density-, and hierarchy-based clustering detection methods, as well as reconstruction-based and prediction-based detection frameworks based on autoencoders, long short-term memory networks, convolutional networks, Transformers, and generative adversarial networks, and then presents the objective functions and anomaly scoring forms of each method. A classification framework for fusion methods is constructed on the basis of three strategies: serial, parallel, and embedded. The findings show that combining the structural priors of clustering with the representation learning capability of deep models yields performance superior to that of a single method, however, the benefits are accompanied by increased model complexity and interpretation costs. In the meantime, four unresolved issues are identified: a unified fusion theory, adaptability to small samples and training set contamination, lightweight edge-side inference, and the construction of interpretable criteria.

Keywords

anomaly detection, clustering analysis, deep learning, deep clustering

1. Introduction

The research of anomaly detection is concerned with determining those observations that do not fit into the normal patterns of data and are most likely related to important events, including network intrusions, fraudulent transactions, equipment failures, or rare pathologies. Due to the ever-increasing size of the Internet, the Internet of Things, and cloud computing, the traffic logs, sequences of user actions, and multi-channel sensor inputs produced by information systems have demonstrated a growth in scale, dimensionality, and heterogeneity orders of magnitude greater than before. The anomalous events like an attack, failure, and fraud not only

© 2026 The Author(s). Authors retain copyright in their articles.

disrupt business continuity but also cause significant security and financial damages. Hence, detection of anomalies has now evolved to be one of the basic capabilities of achieving reliability and security of the systems [1].

Conventional methods of detection are primarily based on distributional assumptions or handcrafted features and threshold rules. It provides a good level of interpretability in low-dimensional stationary data; however, when there is high-dimensionality, strong nonlinearity and high temporal dependence, manual features are unable to cover discriminative patterns and distributional assumptions can be easily violated with limited accuracy and cross-scenario generalization capabilities [2]. Simultaneously, the cost of annotation and the imbalance between classes decrease the practical applicability of supervised schemes even further, and both unsupervised and self-supervised paradigms become the dominant ones in this domain [3].

This being said, two supportive technical pathways can be identified: clustering-based approaches describe the intrinsic clustering of data in an unsupervised fashion and offer intuitive principles under the hypothesis that the fact of being distant to cluster structure signifies anomaly [4]; deep learning-based approaches acquire hierarchical representations using multi-layer nonlinear transformations and are good at managing high-dimensional and high-range temporal correlations. Its complementarity has focused on the fusion concept that clustering gives structural priors whereas deep learning gives discriminative representations, and a method like deep clustering has proven to be more effective than a single path on high-dimensional complex data [5].

The current surveys develop in the context of a particular technical lineage: some work on the comparison of algorithms in clustering-based outlier detection [6], others use deep learning architectures (autoencoders, recurrent networks, generative adversarial networks, etc.) as the thread [7], and others are restricted to particular targets, including high-dimensional multivariate data [8], collective anomalies [9], vehicular ad hoc networks [10], and the Internet of Things [11]. All in all, three drawbacks can be noted: firstly, the two routes are not compared horizontally in terms of their mechanism of anomaly discrimination, complexity limitations, and failure factors in the same aspect; secondly, there is no classification at the level of strategy and similar conclusions about how to combine them; thirdly, the foundations of method selection in particular situations are seldom given.

In order to overcome these deficiencies, in this paper, the time-series anomaly detection methods are classified into six paradigms under a single problem description, and the anomaly discrimination mechanisms and objective function forms of the two routes are sorted out separately in order to ensure that the discussion of advantages and disadvantages can be followed back to particular mathematical mechanisms, and three fusion strategies are used to organize hybrid methods, with gaps in research being identified by four dimensions: theoretical framework, cross-dataset generalization, real-time performance, and interpretability.

The present paper addresses the ways to detect anomalies without full labeling. The coverage of the survey is divided into three stages: first, detection methods depending on cluster analysis, namely, K-Means and its variations, DBSCAN, local outlier factor (LOF), and hierarchical clustering; second, detection methods depending on deep learning, i.e. autoencoders and their variations, LSTM and its encoder-decoder framework, convolutional networks, Transformers, and generative adversarial networks; third, fusion approaches of clustering and deep learning, such as deep clustering, clustering-constrained regularization, and multi-detector result fusion. The performance data in this paper are all labeled with their sources and outcomes of a particular dataset are not extrapolated to be universal results.

2. Problem Description and Method Classification for Anomaly Detection

2.1 Problem Definition and Anomaly Types

With a given dataset $D = \{x_i\}_{i=1}^{n} \subset R^{d}$, the process of anomaly detection can be formally defined as learning an anomaly scoring function

$$s(x)=P(x \in ANOM x, D) [0,1] \quad (1)$$

and separating samples into normal and abnormal depending on a specified or adaptive threshold, τ; in the absence of any labels during the training phase, this task is the same as learning the distribution or manifold of normal data and employing the magnitude of deviation as the score. When a limited set of weak labels is

© 2026 The Author(s). Authors retain copyright in their articles.

permitted, the score may be transformed into a decision by using threshold calibration or ranking truncation [4].

Based on the way they manifest themselves, research objects can be classified into three types: point anomalies - these are individual samples with outliers that deviate significantly on a global scale, e.g., a sharp increase in sensor readings; contextual anomalies - these are the samples that are anomalous in a given context (temporal, spatial, or neighborhood conditions), but are normal in other contexts. An example would be the temperature of 30 ℃ in the winter, which is anomalous in the north but normal in the tropics, and it should be introduced explicitly with contextual properties; group anomalies are a set of samples that show an anomaly as a whole, whereas individual samples need not be anomalous, for instance, anomalous combination patterns of several data packets in a traffic segment, and the challenge with detection is developing criteria at the sequence level, not the sample level [9].

2.2 Main Application Scenarios

Network security detectors normally use NetFlow or statistical properties of packets as inputs in order to detect distributed denial-of-service attacks, port scanning, and malware communications [2]. Surveys at the system level in vehicular ad hoc networks and autonomous driving show that the most common ones are LSTM, CNN, and autoencoders, with LSTM taking up a central role due to its ability to characterize continuous dependency relationships of communication sequences [10]. In the context of the IoT, the cost of deploying detection techniques becomes one of the main constraints due to the limitation of resources of devices and heterogeneous protocols [11].

The use of financial risk control depends on some of the characteristics of the transaction volume, the frequency of transactions, and the geographical location of the transaction in order to detect the theft of cards and money laundering, and this poses a great challenge of extreme class imbalance and stringent real-time demands. In the industrial sector, multi-channel sensor data, including vibration, temperature, and current, can be used to monitor equipment condition and predict maintenance, and their time-varying nature makes reconstruction- or prediction-based approaches more beneficial [1]. In the medical sector, ECG signal and medical image detection have stricter requirements regarding interpretability and false positive rate control.

Although there are differences between domains, the given scenarios have four common types of constraints that are: due to the rarity of anomalous samples, they cause class imbalance and lack of labels; normal patterns themselves experience concept drift with the change in business; the finite edge-side computing capability must support a manageable inference latency; and the decision-making process should be explained to operations staff or regulators. These four aspects will serve as the most important dimensions in assessing each of the three method paradigms discussed further along in the text.

2.3 Method Classification Framework

In terms of technical pathways, there are four types of approaches that are currently used in practice: statistical methods, traditional machine learning methods, deep learning methods, and hybrid methods. The first uses distributional assumptions as its point of departure, which have well-defined mechanisms and do not require a lot of computation, but the assumptions are easy to violate. The second consists of clustering methods (K-means), classification methods (support vectors and random forests), and density methods (LOF) whose performance relies heavily on the quality of input features [12]. The third learns representations automatically using neural networks and reconstructs, predicts, or estimates densities. The combination of the above methods is called hybrid methods because it combines both methods and achieves advantages complementary to each other; one such method is the combination of clustering with deep learning [2]. In terms of modelling paradigms, methods of time series can be further divided into six major paradigms: clustering, prediction, reconstruction, representation learning, statistical, and hybrid [1]. The classification relationships between them are depicted in Figure 1, and the anomaly discrimination basis, typical representatives, and important limitations of each paradigm are presented in Table 1.

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 1: Classification framework of time-series methods for anomaly detection

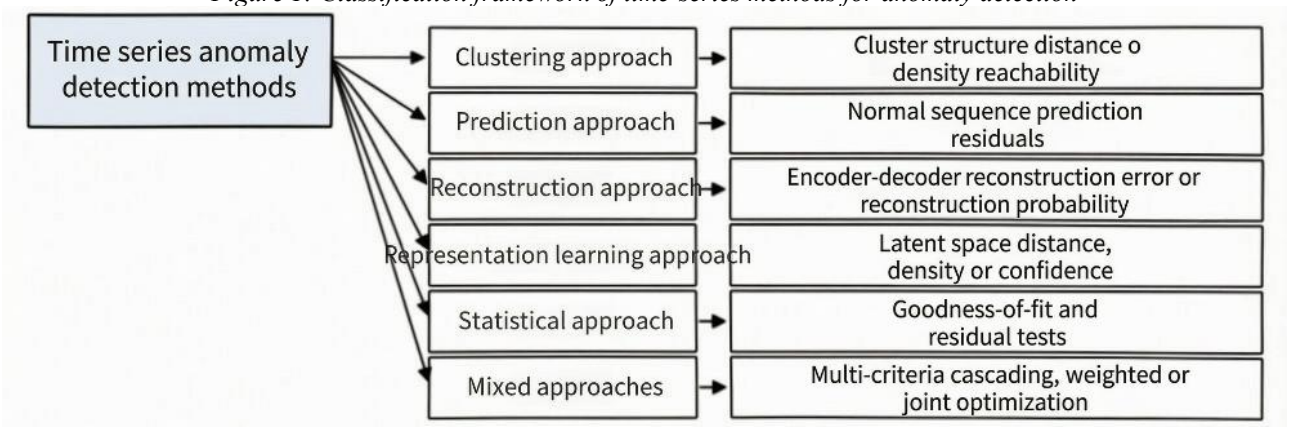


Table 1: Main Paradigms of Time Series Anomaly Detection Methods:Criteria and Limitations

Category	Anomaly Detection Basis	Typical Representatives	Main Limitations
Clustering-based methods	Distance from a sample to the nearest cluster center or density reachability	K-Means, DBSCAN, LOF	High-dimensional distance failure, parameter sensitivity
Prediction-based methods	Residuals between predicted values of normal sequences and true values	LSTM predictor, Transformer predictor	Error diverges as prediction horizon increases
Reconstruction-based methods	Encoding-decoding reconstruction error or reconstruction probability	AE, VAE, EncDec-AD	Sensitive to training set purity
Representation learning-based methods	Distance, density, or classification confidence in latent space	Deep clustering, self-supervised contrastive representation	Mismatch between objective function and clustering objective
Statistical methods	Goodness-of-fit of distribution and residual tests	Gaussian model, extreme value model, ARIMA	Distribution assumptions are easily violated
Hybrid methods	Cascading, weighting, or joint optimization of multiple criteria	CNN cascaded with K-Means, K-Means in parallel with LSTM	Increased complexity and hyperparameter tuning cost

The stated paradigms do not exclude each other: taking deep clustering as an example, it includes both representation learning and clustering criteria, in this paper it is considered as a hybrid approach due to its primary anomaly discrimination step.

3. Clustering Analysis-Based Anomaly Detection Methods

3.1 Basic Principles and Anomaly Discrimination Mechanisms

Cluster analysis is designed to maximize intra-cluster similarity and minimize inter-cluster similarity, and its main components are the distance measures (Euclidean distance, cosine similarity, Mahalanobis distance, etc.) and the computation of representative points, and cluster partitioning techniques [5]. When it comes to anomaly detection, clustering approaches assume that “normal patterns dominate” and treat those samples that are not part of any cluster, or those which are part of a cluster but not close to its representative point, as anomalies. The distance between a sample and the nearest cluster center is the most popular scoring function.

$$s(x) = min_{c_k \in C}\, d(x, c_k) \tag{2}$$

and the threshold is defined using distance quantiles or cross-validation outcomes. Such a belief does not involve any labels with its findings that are simple to interpret for non-experts, and it has become the most popular base in unsupervised anomaly detection [13].

© 2026 The Author(s). Authors retain copyright in their articles.

3.2 Partitioning and Density-Based Methods

3.2.1 K-Means and Its Variants

K-Means reduces the sum of squared errors across clusters by alternating between assigning samples and updating centroids

$$J = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} || x - \mu_i ||^2 \quad (3)$$

and takes the distance in Eq. (2) as the anomaly score [13]. K-Means++ minimizes randomness due to initialization by using probabilistic sampling of initial centroids, and K-Medoids uses actual samples as representative points to enhance resistance to outliers.

Such an algorithm can have a succinct goal function and can also be straightforwardly parallelized, and it has obvious efficiencies with small- and medium-scale data and convex cluster structures, but its drawbacks can be attributed to Eq. (3): the number of clusters k should be predetermined, and the best k is hard to find; it is sensitive to the initial centroids and may converge to local minima; the assumption of convex clusters is implicit in Euclidean distance and does not work well on non-convex forms like rings and crescents; in high-dimensional spaces, the contributions to the distances of the individual dimensions become comparable, and the metric becomes less discriminative. Standard responses include initially reducing the dimensions using principal component analysis followed by clustering, or alternatively switching to other measures that are more resistant to changes in scale and correlation, including: Mahalanobis distance and cosine distance [8].

3.2.2 DBSCAN and LOF

Density-based clusters are based on the level of neighborhood density, with samples located at low densities listed as noise or anomalies. DBSCAN [14] demands that core points must meet

$$| N_{\varepsilon}(p) | \geq MinPts \quad (4)$$

where ε is the radius of the neighborhood and MinPts is the smallest number of samples. Samples that are not able to meet this requirement and are unable to be density-reachable by core points are labeled as anomalies. Such approach does not require any predetermined number of clusters, can find clusters of any shape and is resistant to noise, but the two parameters in Eq. (4) are very susceptible to outcomes, global parameters cannot be used to describe irregular densities, and the neighborhood volume increases sharply in high dimensions, which undermines the idea of density. The current enhancements help in choosing parameter values using adaptive radius or inflection points of k-distance, but it remains difficult to apply them in practice without domain understanding [6]. LOF takes “whether a sample is too sparse compared with its own neighborhood” as the criterion and evaluates the degree of anomaly by comparing the local relative density of a sample with its k-nearest neighbors

$$LOF_k(p) = (1 / | N_k(p) |) \cdot \Sigma_{o \in N_k(p)} lrd_k(o) / lrd_k(p) \quad (5)$$

where lrd represents reachability density. The local density of the sample is significantly smaller than that of its neighborhood, when LOF is much higher than 1, and it is called a local anomaly. LOF can thus identify local anomalies hidden by the global density differences by paying the penalty of huge computational and storage costs due to the nearest neighbor search as well as large scale data sets need approximate nearest neighbors or subsampling methods [15].

3.3 Hierarchical Methods

Hierarchical clustering creates a tree by either bottom-up merging or top-down splitting and can implement pruning algorithms to detect outlier branches that do not join large clusters. The benefits are the absence of a predetermined number of clusters and multigranularity output, which enables the explanation of the conclusions of the detection at various levels; the drawbacks are that the cost of merging increases quickly with the sample size (naive complexity of implementation is between O(n2) to O(n3)) and that it is susceptible to noise - one outlier will alter the entire structure of the merging. Hence, when it comes to anomaly detection, hierarchical clustering has a greater tendency to be applied to exploratory analysis of small- and medium-scale data and to extract cluster structure priors instead of being a direct online detector [6].

© 2026 The Author(s). Authors retain copyright in their articles.

3.4 Research Limitations

The shortcomings of clustering algorithms can be grouped as follows: a) The relative differences between the pairwise distances decrease when the space dimension is large, and the distance and density used by Eq. (2) and Eq. (5) become discriminative to a lesser extent, and the detection performance declines, which has been established in several surveys on dimensionality [8]. b) Both noise and cluster number assumptions are unstable; normal samples can be misjudged due to the shifting of K-Means centroids caused by the presence of outliers, and the choice of parameters for DBSCAN is not stable. c) Conventional clustering techniques work on static feature vectors and are unable to describe the evolution principles and long-range correlations of time series, so that in the temporal situation, it typically needs to perform sliding-window feature engineering before, and the information loss inherent in the process places a direct constraint on the detection upper bound. That lack of quality is exactly where deep learning solutions come into play. The principles, use cases, and complexity comparison of four representative approaches are presented in Table 2.

Table 2: Comparison of Clustering-Based Anomaly Detection Methods

Method	Core Principle	Applicable Scenarios	Advantages	Limitations
K-Means	Partitional clustering, minimizing within-cluster variance	Convex clusters, small-to-medium-scale data	Simple computation, parallelizable	Requires pre-setting k, sensitive to initial values, degradation in high dimensions
K-Means++	Probabilistic initialization of centroid sampling	Same as above	Reduces the impact of randomness	Still requires pre-setting k
K-Medoids	Uses actual samples as representative points	Noisy small-to-medium-scale data	Robust to outliers	Poor scalability
DBSCAN	Neighborhood density reachability	Clusters of arbitrary shape, noisy data	No need to pre-set the number of clusters, noise-resistant	Parameter-sensitive, difficult to define density in high dimensions
LOF	Local relative density comparison	Data with uneven density distribution	Can capture local anomalies	High computational overhead for large-scale data
Hierarchical Clustering	Constructs a tree-like clustering structure	Small-to-medium-scale data	No need to pre-set the number of clusters, multi-granularity	High complexity, sensitive to noise

4. Deep Learning-Based Anomaly Detection Methods

4.1 Technical Advantages and Modeling Paradigms

Deep learning can automatically extract hierarchical features by using multi-layer nonlinear mapping, eliminating the use of manual feature engineering, and is capable of working with high-dimensional, sparse and highly nonlinear data. The benefits are accumulated into three points: automatic representation learning, which means that the features are acquired in a data-driven way based on distance or density criteria; complex distribution fitting, meaning approximating the manifold of normal data by using generative or reconstruction models; and temporal dependency modeling, implying that the long-range connections are captured by recurrent networks and self-attention mechanisms [11].

Depending on different discrimination grounds, deep detection techniques can be generally categorized into two paradigms, namely reconstruction-based and prediction-based.The reconstruction-based paradigm is trained on encoding-decoding structures that operate on normal data and its anomaly score is based on reconstruction error or reconstruction probability which is appropriate with high-dimensional static sample and collective anomaly detection; the prediction-based paradigm employs historical windows of a sequence to estimate the value at the next time step and the prediction residual as the anomaly score, naturally adapting to temporal signals [5].

© 2026 The Author(s). Authors retain copyright in their articles.

4.2 Autoencoder-Based Methods

Autoencoder (AE) is to learn the low-dimensional manifold of normal data by jointly training the encoder E and decoder D, and apply the reconstruction error as the score of anomaly

$$s(x) = || x - D_{\varphi}(E_{\theta}(x)) ||_2^2 \quad (6)$$

The rationality of this approach is based on the fact that the model is trained on normal samples only and its reconstruction error is low, whereas patterns included in anomalous samples remain unlearned and their error rises substantially. The common variants have different focuses: sparse autoencoders apply activation sparsity constraints in the hidden layer to achieve more discriminative representations; denoising autoencoders are used to reconstruct the original input using noisy input, which enhances its resistance to noise and contaminated samples; variational autoencoders treat latent variables as probability distributions and optimize the evidence lower bound

$$L_{VAE} = E_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x) || p(z)) \quad (7)$$

so that reconstruction probability, instead of a single error, can be taken as the score, which is more straightforward to interpret probabilistically and calibrate thresholds by [16]. Autoencoders have a straightforward architecture and may be unsupervisedly trained, however, they are highly sensitive to the quality of the training set: when anomalies are introduced to the training data, the model will overgeneralize the anomalous patterns, resulting in reduced error discriminability. As autoencoders are generative deep networks, their learning object is the distribution of normal data. Since collective anomalies will move away with respect to this distribution altogether, they can be well-detected, and hence autoencoders can be naturally effective at detecting collective anomalies [17].

4.3 LSTM-Based Methods

Recurrent neural networks, along with their long short-term memory (LSTM) versions, are built exclusively to work with sequence data. In LSTM, forget gates, input gates, and output gates control the reading and writing of cell states to solve the vanishing and exploding gradient issues of the standard recurrent network [18]:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f),\ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i),\ g_t = tanh(W_g[h_{t-1}, x_t] + b_g) \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t,\ h_t = o_t \odot tanh(c_t) \quad (9)$$

The central concept of LSTM in detection tasks is sequence prediction: train a prediction model on normal time series and use the residual ŷt - xt between the predicted value $\hat{y}_t$ and the observed value x_t

$$r_t = || x_t - \hat{y}_t ||_2 \quad (10)$$

as the anomaly score at any given time step, and the full threshold determination based on residual statistics in a sliding window. Examples of work are LSTM-NDT (based on multivariate prediction), which builds the scores on the prediction errors of multivariate models, and EncDec-AD (based on an encoder-decoder LSTM architecture), which is able to detect anomalies with reconstruction errors and performs well with NASA turbofan engine data [1]. This kind of method has natural constraints such as the increase in residuals with prediction step length, and inability to parallelize recurrent structures, leading to significant training and inference cost.

4.4 Convolutional (CNN), Attention (Transformer), and Adversarial Network (GAN) Methods

The most popular use of convolutional neural networks is image and video anomaly detection. Two common concepts are: using a pretrained network to extract visual features and then using clustering or distance measures to identify anomalies or directly reconstructing the input using a convolutional autoencoder and finding anomalous areas by error at the pixel level. In a comparative survey of video anomaly detection, it was systematically compared how CNN, autoencoders, VAE, RNN, GAN, and hybrid architectures performed on the same benchmark and it was noted that reconstruction-based and hybrid approaches are more competitive in frame-level localization [7].

© 2026 The Author(s). Authors retain copyright in their articles.

The transformer models the global interactions between sequences of elements using the self-attention mechanism and is able to learn long-range and local relationships at the same time. The way it computes attention is

$$Attention(Q, K, V) = softmax(Q K^T / \sqrt{d_k}) V \quad (11)$$

In comparison to recurrent structures, it is more straightforward to parallelize, although the number of parameters and the cost of inference latency are much larger. Generative adversarial networks are trained by adversarially training a generator and a discriminator to approximate the distribution of normal data

$$min_G max_D V(D, G) = E_{x \sim p_\{data\}}[\log D(x)] + E_{z \sim p_\{z\}}[\log(1 - D(G(z)))] \quad (12)$$

The abnormal samples become apparent as they are hard to produce or as they are poorly accepted by the discriminator, but its training is sensitive to hyperparameters and network sizes, vulnerable to mode collapse, and should be deployed carefully in engineering [19].

4.5 Research Challenges

The deep method is challenged by three issues: a) Lack of interpretability. The model does not offer a direct basis on which one can interpret why it is anomalous. When dealing with a situation that involves clear-cut decisions like security measures and risk control in finance, it is hard to give an explanation of the criterion to the auditors, and it cannot be used as an immediate disposal action. b) Sensitivity to data and contamination. The model demands high data scale and quality. Score discriminability is reduced when anomalies are added to the training sample, and label sparsity constrains the scaling to thresholds and post-processes [17]. c) Performance of large-parameter models and resource consumption. Large-parameter models suffer high inference latency and energy usage at gateways, vehicular terminals, and industrial edge nodes, and this makes them hard to satisfy millisecond-to-second response requirements; the issue is especially significant in the context of the IoT and vehicular ad hoc networks [10]. Comparison of the principles, appropriate data types, benefits, and drawbacks of different methods are discussed in Table 3.

Table 3: Comparison of Deep Learning-Based Anomaly Detection Methods

Method	Core Principle	Applicable Data Types	Advantages	Limitations
AE	Encoding-decoding reconstruction error	Images, tabular data, general	Unsupervised, simple structure	Sensitive to training set purity
SAE	Reconstruction error with sparsity constraint	General	Strong representational discriminability	Constraint strength requires hyperparameter tuning
VAE	Evidence lower bound and reconstruction probability	General	Can output reconstruction probability	Complex training and inference
LSTM	Sequence prediction residuals	Time series	Strong long-range dependency modeling capability	Serial computation, high overhead
CNN	Feature extraction and pixel reconstruction	Images, video	Efficient spatial feature extraction	Limited temporal modeling capability
Transformer	Self-attention dependency modeling	Time series, long sequences	Global dependency modeling	Large number of parameters, slow inference
GAN	Adversarial training to approximate distribution	Images, general	Strong distribution fitting capability	Unstable training

5. Detection Methods Fusing Clustering and Deep Learning

5.1 Fusion Motivation

The fundamental inspiration behind fusion is based on the complementarity between the two methods. On the one hand, clustering offers structural priors to deep models: natural groups in data tend to reflect various

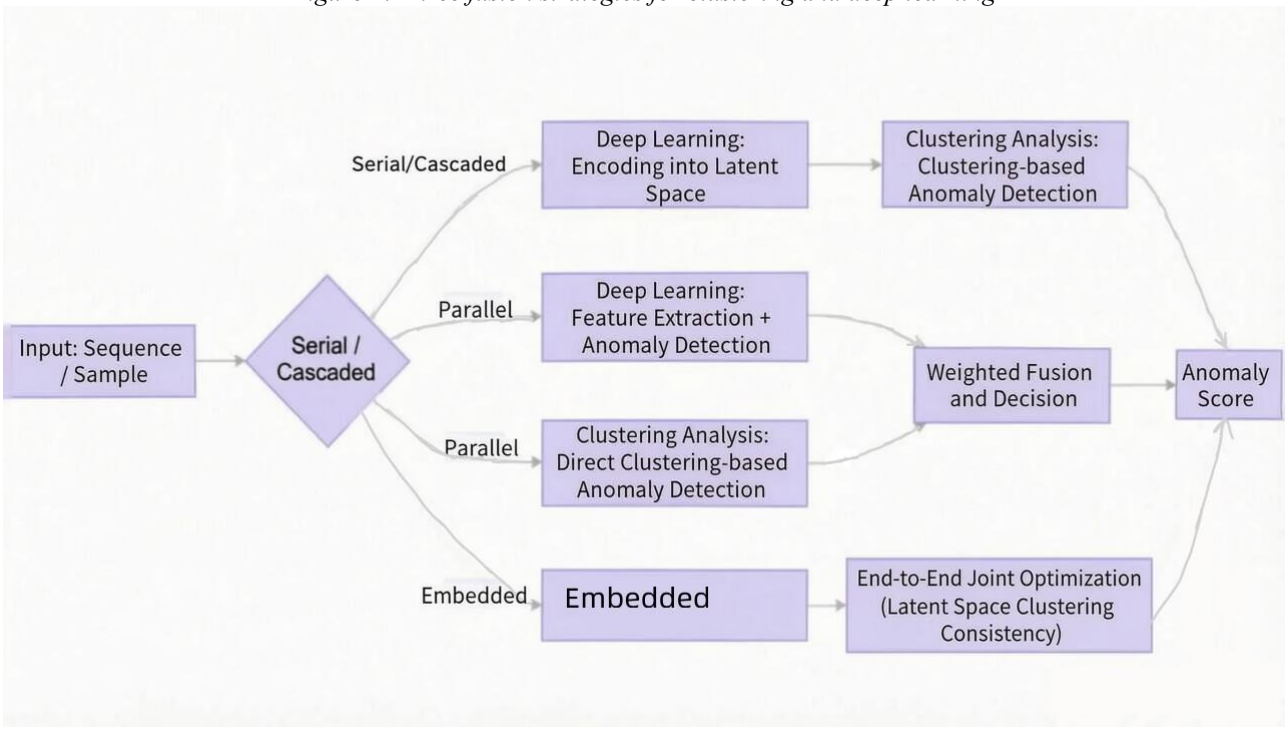
© 2026 The Author(s). Authors retain copyright in their articles.

business behaviors or types of attacks, and previous group data may help the model resist averaging across all data, supporting the strategy of "group-wise modeling and intra-group anomaly discrimination." On the other hand, deep learning has the potential of improving the clustering ability greatly: distances and densities in the original space are destroyed when it is high dimensional, whereas deep representations have the ability to reduce the dimensions yet maintain the discriminative information, and thus, offer an improved input space to cluster. It is exactly where the conception of deep clustering comes into play, simultaneously optimizing the goals of clustering and feature learning [5]. Available survey literature demonstrates that hybrid approaches are typically more effective in generalization capabilities and detection reliability, although they come with higher complexity and tuning costs [17].

5.2 Classification of Fusion Strategies

Depending on the stage of fusion, the current methods may be classified into three approaches. The serial approach appears as a cascade of representation first, clustering second: a deep network learns features, followed by a clustering algorithm that creates distance or density measures. As an instance, a trained convolutional network extracts an image feature set, K-Means creates a standard feature dictionary, and the distance between a given specimen and the closest cluster center is considered to identify the anomaly. It has a well-defined form and it is quite interpretable, however, errors may increase over the cascade. The parallel approach runs two kinds of detectors concurrently and combines the results. One instance would be K-Means and LSTM operating simultaneously to identify network traffic anomalies and combining them using Bayesian weighting, potentially adding value and improving resilience at the expense of more complicated fusion rule formulation and conflict resolution. The embedded approach encodes clustering constraints directly into the deep model, incorporating K-Means-like assignment or clustering consistency loss into the latent space of an autoencoder such that the representation is separable i.e. end-to-end deep clustering. Its peak performance is somewhat high, however, it does not have a single theoretical basis [5].

Figure 2: Three fusion strategies for clustering and deep learning



5.3 Review of Typical Fusion Methods

On the line of network intrusion detection, there has been some research in the area that utilizes traffic pattern clusters data found by K-Means as the context and integrates it with the continuously predicted results of LSTM based on reliability measurements and results in an F1 score higher than any single model on the CICIDS2017 data [20]. In industrial visual defect detection, a typical solution would be to extract image

© 2026 The Author(s). Authors retain copyright in their articles.

features using a pretrained convolutional network, create a normal feature dictionary using K-Means, and use the distance between a feature and the closest cluster center as the anomaly score. The reported average detection accuracy on MVTecAD dataset is more than 95% and its parameter number is much lower than the number of parameters in an end-to-end deep model [7].

Spectral clustering is applied in the direction of multi-sensor systems to extract the spatial correlation structure between sensors and then graph neural networks are used together with message passing and anomaly propagation models to facilitate equipment fault localization in smart buildings; in the direction of multivariate time series, other approaches involve using a Transformer encoder to perform global dependency modeling and applying the DBSCAN algorithm in the latent space with a density threshold to detect collective anomalies, achieving fairly stable results on various publicly available datasets [1]. All the mentioned above demonstrate the fact that the worth of fusion is not found in the stacking of modules, but in allowing clustering to play the roles of "structure and interpretation", and deep models to take the roles of "representation and fitting".

5.4 Strategy Comparison and Research Gaps

Table 4 shows the differences between the three strategies in terms of implementation complexity, interpretability and scenarios where they can be applied.

Table 4: Comparison of Fusion Strategies for Clustering and Deep Learning

Fusion Strategy	Representative Method	Core Idea	Advantages	Limitations	Applicable Scenarios
Serial / Cascaded	CNN cascaded with K-Means	Representation learning followed by clustering-based anomaly detection	Clear structure, interpretable	Error accumulation	Image and defect detection
Parallel	K-Means and LSTM in parallel	Parallel detection followed by result fusion	Complementary strengths, robust	Complex fusion rules	Network traffic detection
Embedded	Deep clustering	Joint optimization of representation and clustering	End-to-end, discriminative representation	Lack of theoretical framework	High-dimensional complex data

The current fusion approaches have four gaps: a) Absence of a common theoretical foundation. Many plans are empirical designs and do not have testable conditions of the situation when fusion can be used and when fusion is most beneficial; b) Poor generalization verification. Many findings only apply to a particular dataset (e.g., MVTec AD, CICIDS 2017, KDD99) and the transferability between domains and types of datasets is not evaluated in a systematic way, which makes it hard to justify the choice of engineering options [20]. c) Performance and efficiency in real time are not considered enough. Deep module parameter counts and computational graph complexities are passed on to fusion methods, and edge device latency and energy consumption are not well measured, and there are no similar measurement outcomes of the trade-off between accuracy and efficiency [11]. d) Fusion does not enhance interpretability. New modules add to the decision chain. When the interpretable source of cluster structure cannot be maintained in the fusion setting, the fusion can concurrently lose the interpretability benefit of clustering and the accuracy benefit of deep models.

6. Research Challenges and Prospects

6.1 Construction of Interpretable Criteria

Interpretability improvement should not be based on post hoc explanation instruments, but should begin with the criteria structure. The provision of intermediate quantities (cluster assignment, density reachability, and gated activation) along with anomaly scores in a standard form could offer operations staff a three-level explanation, i.e., "what kind of normal pattern it is related to, what deviation magnitude is present, and what channel causes it". A fundamental problem that deep clustering must still solve is how to ensure that the cluster semantics are not lost during end-to-end optimization [5].

© 2026 The Author(s). Authors retain copyright in their articles.

6.2 Small Samples and Training Set Contamination

Real systems do not cover all normal patterns, unusual samples are rare, and the training sample is necessarily contaminated. Consistency and contrastive constraints may be incorporated into the representation learning phase to minimize reliance on labels, and semi-parametric limits or robust statistics with tolerance to contamination rates may be considered at the criteria phase, whereas for generative algorithms, the research of the upper bound of manifold estimation bias with unknown contamination ratios is required to prevent overgeneralization of anomalous patterns [17].

6.3 Edge-Side Real-Time Performance and Lightweight Design

Detectors should be able to respond in less than a second on limited computing power for gateways, vehicular terminals, and industrial sites. Some feasible paths are to bring clustering criteria to the edge, deploy deep representations on the central side with a division-of-labor architecture, or use pruning, quantization, and knowledge distillation with deep modules, and assess methods based on dual indicators of accuracy-latency instead of one accuracy indicator [11].

6.4 Unified Theory and Cross-Domain Generalization

It is in the future that there will be a need to have an analytical framework that can explain the origins of fusion gains, e.g., characterize the upper bound of hybrid methodologies with respect to cluster structure separability and representation discriminability consistency. Simultaneously, there must be a uniform evaluation protocol across datasets and data types, and report degradation curves of methods with concept drift, dimensionality change, and scale change so that conclusions are extrapolated [8].

7. Conclusion

The present paper describes the current state of the art in the topic of anomaly detection, both in the form of clustering analysis and deep learning. The major conclusions are the following. Given a common problem formulation, time-series anomaly detection approaches can be divided into six paradigms: clustering, prediction, reconstruction, representation learning, statistical, and hybrid. The bases of anomaly discrimination and applicable boundaries of these paradigms vary: clustering methods use intuitive principles and interpretable solutions and are highly robust to low-dimensional and small/medium-sized data but fail due to high dimensionality distances degradation, parameters sensitivity and inability to model time, whereas the deep learning methods, which use automatic representation learning and the ability to fit complex distributions, excel at detecting images and time series, but are limited by their low interpretability, vulnerability to contamination of the training set and large inference cost. Based on the above, the three strategies of serial, parallel and embedded fusion, created based on the stage of fusion, have clear premises of applicability: serial fusion is appropriate when the representations are rich and criteria require interpretation, parallel fusion is appropriate when the criteria are multimodal and robustness is desired, and embedded fusion is appropriate when data are high-dimensional and complex and end-to-end training cost is affordable. To summarize, current work has significant loopholes in four areas: a common theoretical framework, cross-dataset generalization validation, real-time performance and interpretability.

References

[1] Sayyaf, M. I., Pascacio, P., Zhu, N., et al. (2024). Time-series anomaly detection for sensor data: Models, metrics, and methodologies: A review. IEEE Sensors Journal, 24 (12), 19668–19688.

[2] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19–31.

[3] Pang, G., Shen, C., Cao, L., et al. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys, 54 (2), 1–38.

[4] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41 (3), 1–58.

© 2026 The Author(s). Authors retain copyright in their articles.

[5] Ghasiabadian, M., Moosavi, Z., & Chow, T. W. S. (2023). Deep clustering: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13 (1), e1485.

[6] Campos, G. O., Zimek, A., Sander, J., et al. (2016). On the evaluation of unsupervised outlier detection: Measures, metrics, and empirical analysis. Data Mining and Knowledge Discovery, 30 (6), 1541–1576.

[7] Sabokrou, M., Rozantehder, A., Mohammadi, A., et al. (2022). Deep learning-based video anomaly detection: A survey. The Visual Computer, 38 (1), 337–364.

[8] Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. In Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data (pp. 215–224). ACM.

[9] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings of the 8th IEEE International Conference on Data Mining (pp. 413–422). IEEE.

[10] Olusoyi, K. P., Olugbein, R. A., Folorunso, O. A., et al. (2021). Anomaly detection for vehicular ad hoc networks (VANETs). IEEE Access, 9, 102747–102767.

[11] Chalapathy, R., & Chawla, S. (2015). Anomaly detection for IoT security using supervised learning approaches. In Machine Learning and Data Mining in Pattern Recognition (pp. 12–23). Springer.

[12] He, Q., Li, N., & Luo, W. J. (2001). A survey on popular classification data mining algorithms. Journal of Computer Research and Development, 38 (11), 1376–1384.

[13] Zhou, Z. H. (2016). Machine learning (pp. 1–10). Tsinghua University Press.

[14] Ester, M., Kriegel, H. P., Sander, J., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (pp. 226–231). AAAI Press.

[15] Breunig, M. M., Kriegel, H. P., Ng, R. T., et al. (2000). LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93–104). ACM.

[16] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations (pp. 1–14). ICLR.

[17] Zong, B., Song, Q., Min, M. R., et al. (2018). Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In International Conference on Learning Representations (pp. 1–19). ICLR.

[18] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9 (8), 1735–1780.

[19] Goodfellow, I., Poullot, J., Mirza, M., et al. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems 27 (pp. 2672–2680). MIT Press.

[20] Sharafaldeen, I., Belkadi, I., Alharidi, S., et al. (2021). Developing a real time intrusion detection system for Industry 4.0. Computers & Security, 101, 102017.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

© 2026 The Author(s). Authors retain copyright in their articles.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.