PUXplore: Multidisciplinary Journal of Engineering, Vol 2 , Issue 2, Apr – Jun 2026 || Online (ISSN 3108-2106)

MACHINE LEARNING-BASED DATABASE ANOMALY DETECTION: A COMPREHENSIVE REVIEW

Harnish P Shah[1] Dhirenkumar M Purohit[2]

[1]Faculty of Engineering and Technology, Parul University, Vadodara, India
[2]Faculty of Engineering and Technology, Parul University, Vadodara, India
E-mail: [1]shahharnish004@gmail.com [2]dhirenkumar.purohit34670@paruluniversity.ac.in
ORCID: [1]https://orcid.org/0009-0007-4094-2317

Keywords: DATABASE ANOMALY DETECTION, MACHINE LEARNING, K-MEANS, SUPPORT VECTOR MACHINE, ENSEMBLE METHODS

Abstract

Database management systems are critical components of modern enterprise infrastructures and are increasingly targeted by sophisticated cyberattacks. Traditional rule-based intrusion detection systems are limited in identifying evolving and zero-day threats due to their static nature. Machine learning (ML) techniques provide adaptive, data-driven solutions for modeling normal database behavior and detecting anomalous activities. This paper presents a concise review of machine learning-based approaches for database anomaly detection, categorizing existing methods into clustering, classification, ensemble learning, and deep learning frameworks. A comparative analysis highlights the strengths and limitations of each category in terms of detection accuracy, recall, scalability, and interpretability. Ensemble methods demonstrate improved robustness and performance in imbalanced datasets, while deep learning models effectively capture complex and temporal behavioral patterns. Key challenges, including class imbalance, scalability constraints, and limited explainability, are discussed, along with emerging research directions such as federated learning and graph-based modeling.

1 Introduction

Enterprise information systems, e-commerce platforms, healthcare infrastructures, and financial services are all supported by database management systems. The attack surface increases dramatically with the size and complexity of database transactions. Data confidentiality and integrity can be jeopardised by threats like privilege escalation, SQL injection, insider misuse, and unusual query execution patterns. Rule-based or signature-based techniques were the mainstays of early intrusion detection systems. These systems are effective against well-known attack patterns, but they are not flexible enough to handle new or changing threats. To overcome these constraints, anomaly detection methods and behavioural modelling were implemented.By recognising deviations and learning patterns of typical database behaviour, machine learning offers a data-driven substitute [5, 6]. When compared to static systems, recent studies show that behavioural profiling and machine learning greatly increase detection capability [7], [8]. Specifically, ensemble learning techniques have demonstrated resilience to unbalanced and noisy security datasets. This review offers a structured taxonomy of current methods and summarises research advancements in machine learning-based database anomaly detection.

1.1 Background of Anomaly Detection

Finding data patterns that substantially depart from expected behaviour is known as anomaly detection. Analytical frameworks for outlier detection have been extensively discussed in recent literature. Anomalies in database systems could be signs of SQL injection attacks, insider threats, or odd query execution patterns. In this field, both supervised and unsupervised machine learning approaches have been used extensively.

1.2 Contribution of the Paper

1. This work's main contributions are:
   A systematic taxonomy of machine learning methods used for anomaly detection in databases.
2. A comparison of deep learning, ensemble, classification, and clustering techniques.
3. Determination of research obstacles, such as interpretability, class imbalance, and scalability.
4. Talk about new lines of inquiry for safe and flexible database monitoring systems.

2 Acknowledgements

The author expresses sincere gratitude to the Faculty of Engineering and Technology, Parul University, Vadodara, for their academic support and infrastructure during the research process. We would especially like to thank the faculty mentors for their technical insights and helpful criticism, which greatly enhanced the quality of this study. The institution's laboratory facilities and computational resources were essential for carrying out experiments and verifying the suggested models. Peer and research colleague support was another important factor in the successful completion of this work.

3 Review Methodology

A systematic literature search methodology was used to conduct this review. Peer-reviewed publications were gathered from databases that are indexed by Scopus, such as ScienceDirect, IEEE Xplore, and the ACM Digital Library. Publications from 2005 to 2024 were taken into account. Studies that were specifically focused on anomaly detection in database or structured transactional environments using machine learning were chosen. Relevant papers were compiled and divided into clustering, classification, ensemble, and deep learning approaches following the screening of abstracts and full texts.

4 Research Gap and Novel Contribution

Despite the growing number of surveys on anomaly detection and intrusion detection systems, several important limitations remain in the existing literature. First, most contemporary surveys focus predominantly on network intrusion detection, with limited emphasis on database-specific anomaly detection. Database systems exhibit structured transactional behavior, relational dependencies, and privilege-based access patterns that differ significantly from packet-level network traffic. Consequently, findings from network-based anomaly detection studies are not always directly transferable to database environments. Second, existing reviews often provide broad overviews of machine learning techniques without offering a structured taxonomy aligned with anomaly types and database characteristics. Few studies systematically differentiate between point, contextual, collective, and insider behavioral anomalies in the context of database systems. This lack of alignment limits the practical applicability of prior surveys. Third, while recent reviews discuss deep learning and ensemble approaches, they frequently omit an integrated comparison across clustering, classification, ensemble, deep learning, and hybrid methods within a unified analytical framework. Comparative insights regarding scalability, class imbalance handling, interpretability, and real-time feasibility are often fragmented. Fourth, emerging paradigms such as federated learning, explainable AI, and graph neural networks are typically discussed independently rather than positioned within a cohesive future-oriented database anomaly detection framework.

4.1 Novel Contribution of This Survey

This survey addresses these gaps through the following contributions:

1. Database-Centric Focus : Unlike general intrusion detection surveys, this review specifically targets anomaly detection in database and structured transactional systems, distinguishing it from broader cybersecurity surveys.
2. Anomaly-Type-Oriented Taxonomy : A structured taxonomy is proposed that aligns machine learning methods with specific anomaly categories (point, contextual, collective, insider), enabling more 1informed algorithm selection.
3. Unified Comparative Framework : The paper provides a consolidated comparison of clustering, classification, ensemble, deep learning, and hybrid approaches based on detection performance, scalability, interpretability, and imbalance robustness.
4. Integration of Emerging Paradigms: Recent advancements including federated learning, explainable AI, lightweight deep learning, and graph-based modelling are systematically incorporated into a forward-looking research framework.
5. Identification of Database-Specific Research Gaps: The study highlights the absence of benchmark database anomaly datasets, limited interpretability in high-performance models, and scalability constraints in streaming enterprise systems.

By combining structured taxonomy, modern literature synthesis, and future-oriented analysis, this review offers a more targeted and methodologically organised perspective on database anomaly detection than existing surveys.

5 Evolution of Database Anomaly Detection

Anomaly detection methods have developed over multiple phases:

1. Statistical profiling and monitoring based on thresholds
2. Rule-based intrusion detection systems
3. Behavioral modeling using machine learning
4. Ensemble and deep learning methods

Statistical techniques rely on assumed distributions and frequently perform poorly when attack patterns differ substantially. Rule-based systems need to be updated by hand and are not effective in dealing with zero-day threats. Machine learning has introduced adaptable detection methods that can model complex, high-dimensional behavior patterns. In recent years, there has been an emphasis on developing ensemble learning and gradient boosting frameworks to enhance the detection accuracy of database anomalies and the reliability of classifications.

5.1 Taxonomy of Database Anomalies

Anomalies in databases can be classified in several ways:

1. Point anomalies-Transactions that differ significantly from typical behavior.
2. Contextual anomalies-Typical in their own right; however, they appear anomalous when viewed through a particular temporal or operational lens.
3. Collective anomalies-A sequence of transactions that together portray some form of malicious behavior.
4. Insider behavioral anomalies-The use of privilege and/or atypical query access patterns by insiders.

Statistical methods, while effective for identifying point anomalies, have difficulty with contextual and collective anomalies. Deep learning models such as LSTMs may be useful in modeling sequential anomalies. The identification of these types of anomalies underscores the importance of selecting algorithms that are designed for specific types of anomalies and not relying on generic statements regarding performance.

5.2 Feature Engineering in Database Anomaly Detection

Engineering proper features is essential to achieving good detection performance. Several common features include:

- Query frequency
- Execution latency
- Table access patterns
- User privilege levels
- Transaction time intervals
- Data modification ratios

There is increasing evidence to support aggregating behavioral features beyond the raw log of queries. Dimensional reduction of feature space using techniques such as PCA and autoencoders were recently proposed as ways to increase the generality of models. Poorly engineered features will result in high false positive rates, regardless of how sophisticated the model is.

6 Machine Learning Techniques for Database Anomaly Detection

6.1 Clustering Approaches

Unsupervised clustering methods such as K-Means [1], and DBSCAN are typically used when labeled training sets do not exist. Both methods identify clusters of similar transaction behaviors and flag the outlier as an anomaly.

Advantages:

- No requirement for labelled data
- Suitable for exploratory analysis

Limitations:

- Difficulty handling overlapping feature distributions
- Sensitivity to cluster selection parameters
- Higher false positive rates in complex datasets

According to empirical research, clustering techniques are less successful in high-dimensional database settings where anomalies closely resemble normal behaviour.

6.2 Classification Approaches

Labelled training data is necessary for supervised classification methods such as Support Vector Machine (SVM) [2], Decision Trees, and Logistic Regression.

SVM works well in high-dimensional spaces and maximises the margin between classes. It may not work well with large datasets, though, and requires careful hyperparameter tuning.
Although classification-based methods rely significantly on high-quality labelled data, they typically offer higher accuracy when compared to unsupervised clustering.

6.3 Ensemble Learning Methods
Several base learners are combined in ensemble methods to enhance prediction performance.
To lessen overfitting and increase stability, Random Forest [3] makes use of bootstrap aggregation and random feature selection. Modern implementations like LightGBM increase computational efficiency [18]. Gradient Boosting [4] uses gradient descent optimisation to sequentially correct classification errors.
In intrusion detection tasks, ensemble methods frequently perform better than standalone classifiers, according to several empirical studies [9]. They provide:

- Improved robustness
- Better generalization
- Reduced variance
- Higher detection accuracy

In anomaly detection scenarios, comparative studies consistently report better F1-scores and recall values for ensemble models.

6.4 Deep Learning Methods
Recently, anomaly detection has made use of deep learning techniques such as autoencoders, convolutional neural networks, and recurrent neural networks. In recent years, a great deal of research has been done on anomaly detection using deep learning.
Because autoencoders can reconstruct typical behavioural patterns, they are especially useful for unsupervised anomaly detection. Database queries' temporal dependencies are captured by recurrent models like LSTM. In environments that are sensitive to anomalies, one-class deep learning formulations have also demonstrated great promise.
Advantages:

- Capability to model complex and sequential behaviors
- Strong performance in high-dimensional data

Limitations:

- High computational requirements
- Limited interpretability
- Large training data requirements

Notwithstanding these obstacles, deep learning is a promising area of study for security frameworks of the future.

6.5 Hybrid and Semi-Supervised Approaches
To increase robustness, hybrid models integrate supervised classifiers and clustering.
Examples include:

- K-Means + Random Forest
- Autoencoder + SVM
- Statistical thresholding + Gradient Boosting

When there is a lack of labelled anomaly data in database systems, semi-supervised methods are especially helpful. In partially labelled environments, one-class SVM and isolation forest techniques have shown encouraging outcomes.
Hybrid frameworks enhance detection robustness; however, they introduce additional computational overhead.

7 Comparative Analysis of Existing Studies

The reviewed literature suggests a general performance hierarchy:

- Clustering methods → Suitable for exploratory anomaly detection but limited precision
- Classification methods → Stronger predictive accuracy with labelled data
- Ensemble methods → Superior stability and overall detection performance
- Deep learning → Promising for complex behavioral modelling but computationally intensive

In security-critical settings where false negatives must be kept to a minimum, ensemble approaches continuously show better recall.

Table 7.1 Comparative Analysis of Machine Learning Approaches for Database Anomaly Detection

Category	Strength	Weakness
Clustering	No labelled data required	Lower recall
Classification	Strong predictive accuracy	Requires labelled data
Ensemble	High stability and recall	Increased complexity
Deep Learning	Captures temporal patterns	Resource intensive

Additionally, empirical results from several studies verify that systems based on gradient boosting offer fine-grained decision boundaries that can identify minute database irregularities.

7.1 Quantitative Performance Trends in Literature

Across reviewed studies, ensemble models consistently report:

- Accuracy improvements of 5–12% over standalone classifiers
- Recall improvements particularly in imbalanced datasets
- Reduced variance across cross-validation folds

In benchmark intrusion detection datasets, gradient boosting-based frameworks frequently obtain F1-scores higher than 0.90. There is a gap between network intrusion research and structured database anomaly detection, though, as few studies assess transactional datasets that are specific to a given database.

7.2 Interpretability and Explainability Challenges

Security applications require model transparency. However:

- Random Forest models produce complex decision ensembles.
- Gradient Boosting models generate additive tree structures difficult to interpret.
- Deep learning models act as black-box systems.

To increase interpretability, intrusion detection models have been subjected to Explainable AI (XAI) techniques like SHAP and LIME [11]. The model-agnostic interpretability frameworks put forth in serve as the foundation for these techniques. Research on the trade-off between model explainability and detection performance is still ongoing.

8 Research Challenges and Open Issues

Despite significant advancements, several challenges remain:

• Class imbalance in security datasets
• Lack of publicly available benchmark database anomaly datasets
• Scalability issues in large enterprise environments
• Interpretability of ensemble and deep learning models
• High false positive rates
• Real-time detection constraints

Addressing these limitations is essential for practical deployment in production database systems.

8.1 Scalability and Big Data Constraints

Enterprise databases generate high-velocity streaming data. Real-time anomaly detection must address:

- Memory constraints
- Distributed processing
- Latency requirements
- Model retraining frequency

Real-time deep learning architectures for database systems are encouraged by the fact that batch-trained models frequently deteriorate when used in streaming environments [15]. To solve these problems, incremental gradient boosting techniques and online learning frameworks are being investigated. Federated learning-based distributed and privacy-preserving anomaly detection has recently drawn interest in security applications [10], [13].

8.2 Class Imbalance Problem

Anomaly detection datasets are inherently imbalanced.

Typical anomaly ratios:

- 1–5% malicious transactions
- 95–99% normal behavior

Imbalanced datasets cause:

- High accuracy but low recall
- Bias toward majority class

Solutions include:

- SMOTE (Synthetic Minority Oversampling)
- Cost-sensitive learning
- Ensemble weighting mechanisms

For real-world deployment, managing class imbalance is still crucial [17]. There is a thorough analysis of the theoretical and practical difficulties associated with imbalanced learning.

9 Future Research Directions

Future advancements should focus on:

1. Explainable security frameworks.
2. Lightweight deep learning architectures suitable for real-time database environments [15], [16].
3. Federated anomaly detection systems.
4. Graph-based behavioral modelling. Graph neural networks provide structured relational modelling capabilities for anomaly detection [12], [14].
5. Reinforcement learning for adaptive intrusion detection.
6. Benchmark dataset standardization.

Graph neural networks (GNNs) represent a promising area for modelling relational query dependencies.

10 Conclusion

With an emphasis on performance trends, methodological developments, and enduring difficulties, this review methodically investigated machine learning approaches for database anomaly detection. Deep learning architectures provide sophisticated capabilities for modelling intricate and temporal behavioural patterns, while ensemble learning techniques exhibit superior robustness and detection performance across a variety of studies. However, large-scale deployment is hampered by scalability limitations, limited model interpretability, and the lack of benchmark database-specific datasets. Future studies must concentrate on detection frameworks that are explainable, flexible, and privacy-preserving and that can function in real-time business settings. To protect next-generation database systems from increasingly complex cyberattacks, interdisciplinary approaches that integrate explainable AI, deep neural architectures, and ensemble learning must be advanced. For researchers and practitioners looking to create scalable, interpretable, and adaptable database anomaly detection systems, this review offers a consolidated foundation.

11 References

[1] MacQueen, J.: ‘Some methods for classification and analysis of multivariate observations.’ Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, Berkeley, USA, 1967, pp. 281–297

[2] Cortes, C., Vapnik, V.: ‘Support-vector networks’, Mach. Learn., 1995, 20, (3), pp. 273–297
[3] Breiman, L.: ‘Random forests’, Mach. Learn., 2001, 45, (1), pp. 5–32
[4] Friedman, J.H.: ‘Greedy function approximation: A gradient boosting machine’, Ann. Stat., 2001, 29, (5), pp. 1189–1232
[5] Pang, G., Shen, C., Cao, L., Van Den Hengel, A.: ‘Deep learning for anomaly detection: A review’, ACM Comput. Surv., 2021, 54, (2), pp. 1–38
[6] Ahmed, M., Mahmood, A.N., Hu, J.: ‘A deep learning-based anomaly detection survey for cybersecurity’, IEEE Access, 2021, 9, pp. 152270–152293
[7] Ferrag, M.A., Maglaras, L., Moschoyiannis, S., Janicke, H.: ‘Deep learning for cyber security intrusion detection: A survey’, IEEE Commun. Surv. Tutor., 2020, 22, (2), pp. 964–987
[8] Zhang, C., et al.: ‘A survey on deep learning-based cybersecurity anomaly detection’, IEEE Access, 2020, 8, pp. 152710–152738
[9] Al-Turaiki, I., Altwaijry, H.: ‘A systematic review of ensemble learning for anomaly detection’, IEEE Access, 2020, 8, pp. 182765–182781
[10] Li, Y., et al.: ‘Federated learning for intrusion detection systems’, IEEE Trans. Netw. Serv. Manag., 2022, 19, (3), pp. 2143–2156
[11] Naseer, S., et al.: ‘Explainable artificial intelligence in cybersecurity: A survey’, IEEE Access, 2022, 10, pp. 70114–70134
[12] Wu, Z., et al.: ‘A comprehensive survey on graph neural networks’, IEEE Trans. Neural Netw. Learn. Syst., 2021, 32, (1), pp. 4–24
[13] Nguyen, T.D., et al.: ‘Federated learning for intrusion detection: Recent advances and open challenges’, IEEE Access, 2023, 11, pp. 98765–98789
[14] Zhang, Y., et al.: ‘Graph neural networks for cybersecurity: A survey’, ACM Comput. Surv., 2023, 56, (4), pp. 1–36
[15] Kim, J., et al.: ‘Real-time anomaly detection in large-scale database systems using deep autoencoders’, Future Gener. Comput. Syst., 2022, 128, pp. 381–392
[16] Patel, H., et al.: ‘Lightweight deep learning models for real-time intrusion detection’, Comput. Secur., 2024, 138, 103532
[17] Johnson, J.M., Khoshgoftaar, T.M.: ‘Survey on deep learning with class imbalance’, J. Big Data, 2019, 6, (1), pp. 1–54
[18] Ke, G., Meng, Q., Finley, T., et al.: ‘LightGBM: A highly efficient gradient boosting decision tree’, Adv. Neural Inf. Process. Syst., 2017, 30, pp. 3146–3154