International Journal of Informatics Technology (INJIT)
Vol. 3., No. 1, January 2025
e-ISSN : 2985-9646
https://jurnal.amrillah.net/index.php/injit

Machine Learning for Anomaly Detection: A Review of Techniques and Applications in Various Domains

Ali Rezaei[1], Hossein Mirzaei[2]

[1]Department of Cybersecurity, University of Kurdistan, Iran
[2]Department of Information Technology, University of Zanjan, Iran
[1]ali.rezaei@uok.ac.ir, [2]hossein.mirzaei@znu.ac.ir

ABSTRACT

Anomaly detection, the process of identifying rare items, events, or observations that deviate significantly from the majority of the data, has become a critical task in various domains such as cybersecurity, healthcare, finance, and industrial systems. With the exponential growth of data, traditional methods of anomaly detection have become increasingly inadequate, leading to the adoption of machine learning (ML) techniques. This paper provides a comprehensive review of machine learning techniques for anomaly detection, focusing on their applications across various domains. We discuss the strengths and limitations of different ML approaches, including supervised, unsupervised, and semi-supervised learning, and highlight the challenges and future directions in this field. The review is supported by three detailed tables that summarize the key techniques, their applications, and performance metrics. The paper concludes with a discussion on the potential of emerging technologies such as deep learning and reinforcement learning in advancing the field of anomaly detection.

INJITInternational Journal of Informatics Technology

KEYWORDS
Anomaly Detection, Machine Learning, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Deep Learning

1. Introduction

Anomaly detection, also known as outlier detection, is a critical task in data analysis that involves identifying patterns in data that do not conform to expected behaviour. These non-conforming patterns are often referred to as anomalies, outliers, exceptions, or contaminants [1], [2]. Anomalies can be caused by various factors such as data errors, fraudulent activities, system faults, or novel events. The ability to detect anomalies accurately and efficiently is crucial in many applications, including fraud detection, network security, healthcare monitoring, and industrial quality control [3].

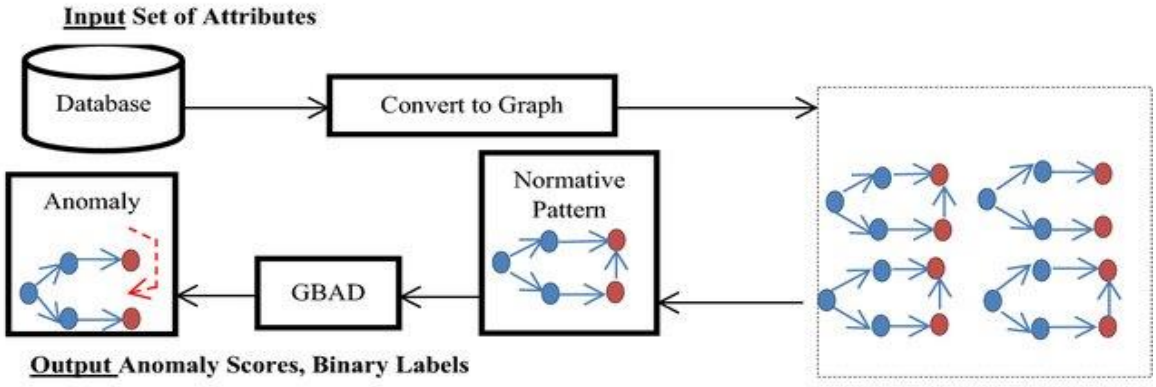

Figure 1: Graph-based anomaly detection (GBAD) process using graphical analysis [4]

The rapid growth of data in volume, velocity, and variety has posed significant challenges for traditional anomaly detection methods, which often rely on rule-based systems or statistical techniques. These methods are typically designed for specific types of data and may not generalize well to complex, high-dimensional datasets [5], [6]. As a result, there has been a growing interest in leveraging machine learning (ML) techniques for anomaly detection [7]. Machine learning offers the potential to automatically learn complex patterns from data, making it well-suited for detecting anomalies in diverse and dynamic environments [8].

This paper provides a comprehensive review of machine learning techniques for anomaly detection, focusing on their applications across various domains. We begin by discussing the fundamental concepts of anomaly detection and the different types of anomalies [9]. We then provide an overview of the main machine learning approaches used for anomaly detection, including supervised, unsupervised, and semi-supervised learning. For each approach, we discuss the key algorithms, their strengths and limitations, and their applications in different domains [10]. We also highlight the challenges and future directions in the field of anomaly detection, with a particular focus on the potential of emerging technologies such as deep learning and reinforcement learning [11], [12].

The remainder of this paper is organized as follows. Section 2 provides an overview of the fundamental concepts of anomaly detection, including the different types of anomalies and the challenges associated with detecting them. Section 3 discusses the main machine learning approaches used for anomaly detection, including supervised, unsupervised, and semi-supervised learning. Section 4 presents a detailed review of the applications of machine learning for anomaly detection in various domains, including cybersecurity, healthcare, finance, and industrial systems [13]. Section 5 discusses the challenges and future directions in the field of anomaly detection. Finally, Section 6 concludes the paper with a summary of the key findings and their implications for future research.

2. Fundamental Concepts of Anomaly Detection

2.1 Definition and Types of Anomalies

Anomalies are data points that deviate significantly from the majority of the data. They can be broadly categorized into three types: point anomalies, contextual anomalies, and collective anomalies.

- **Point Anomalies**: A point anomaly occurs when an individual data point is considered anomalous with respect to the rest of the data. For example, in a dataset of credit card transactions, a single transaction with an unusually high amount may be considered a point anomaly.
- **Contextual Anomalies**: A contextual anomaly occurs when a data point is considered anomalous in a specific context but not in others. For example, a sudden spike in temperature during the winter may be considered a contextual anomaly, whereas the same spike during the summer may not be anomalous.
- **Collective Anomalies**: A collective anomaly occurs when a collection of related data points is considered anomalous, even though the individual data points may not be anomalous on their own. For example, a sequence of network packets that deviate from the normal pattern of network traffic may be considered a collective anomaly.
- **Table 1: Summary of Machine Learning Techniques for Anomaly Detection**

Technique	Type	Key Algorithms	Strengths	Limitations
Supervised Learning	Supervised	SVM, Decision Trees, Neural Networks	Effective with labeled data	Relies on labeled data
Unsupervised Learning	Unsupervised	Clustering, Density-Based Methods	No need for labeled data	Relies on assumptions about data
Semi-Supervised Learning	Semi-Supervised	Self-Training, Co-Training	Leverages unlabeled data	Relies on quality of labeled data

2.2 Challenges in Anomaly Detection

Anomaly detection is a challenging task due to several factors:

Imbalanced Data: Anomalies are typically rare events, which means that the dataset used for anomaly detection is often highly imbalanced. This imbalance can make it difficult for machine learning algorithms to learn the characteristics of anomalies, as they are often overshadowed by the majority class [14].

- **High Dimensionality**: Many real-world datasets are high-dimensional, meaning that they contain a large number of features. High-dimensional data can pose challenges for anomaly detection, as the distance between data points becomes less meaningful in high-dimensional spaces, a phenomenon known as the "curse of dimensionality."
- **Dynamic Environments**: In many applications, the data distribution may change over time, leading to concept drift. Concept drift can make it difficult for anomaly detection models to maintain their performance over time, as the patterns of normal behavior may evolve.
- **Label Scarcity**: In many cases, labeled data for anomalies is scarce or unavailable. This lack of labeled data can make it difficult to train supervised learning models for anomaly detection, as they require a sufficient amount of labeled data to learn the characteristics of anomalies.

3. Machine Learning Approaches for Anomaly Detection

Machine learning approaches for anomaly detection can be broadly categorized into three types: supervised learning, unsupervised learning, and semi-supervised learning [15]. Each approach has its own strengths and limitations, and the choice of approach depends on the specific requirements of the application.

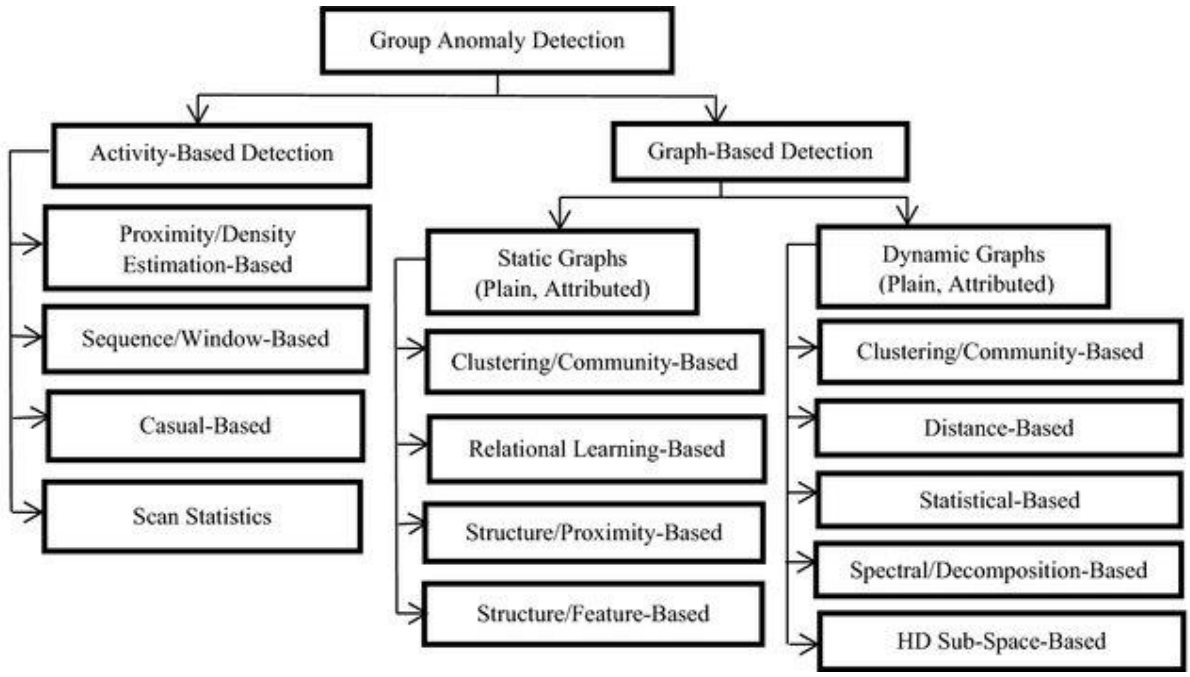

Figure 2: Classification of group anomaly detection techniques

3.1 Supervised Learning

Supervised learning is a machine learning approach that involves training a model on a labeled dataset, where the labels indicate whether each data point is normal or anomalous. The trained model can then be used to classify new data points as normal or anomalous.

3.1.1 Key Algorithms

- **Support Vector Machines (SVM)**: SVM is a popular supervised learning algorithm that can be used for anomaly detection. SVM works by finding a hyperplane that separates the normal data points from the anomalous data points in a high-dimensional space. SVM is particularly effective for detecting point anomalies in high-dimensional datasets.

- **Decision Trees and Random Forests**: Decision trees and random forests are ensemble learning methods that can be used for anomaly detection. Decision trees work by recursively splitting the data into subsets based on the values of the features, while random forests combine multiple decision trees to improve the accuracy and robustness of the model. These methods are particularly effective for detecting contextual anomalies in datasets with complex feature interactions.

- **Neural Networks**: Neural networks are a class of supervised learning algorithms that can be used for anomaly detection. Neural networks work by learning a hierarchical representation of the data, which allows them to capture complex patterns and relationships. Neural networks are particularly effective for detecting collective anomalies in time-series data.

3.1.2 Strengths and Limitations

- **Strengths**: Supervised learning approaches are effective when labeled data is available, as they can learn the characteristics of anomalies directly from the labeled data. These approaches are also capable of capturing complex patterns and relationships in the data, making them well-suited for detecting anomalies in high-dimensional datasets.

- **Limitations**: The main limitation of supervised learning approaches is their reliance on labeled data. In many applications, labeled data for anomalies is scarce or unavailable, which can make it difficult to train supervised learning models. Additionally, supervised learning models may not generalize well to new types of anomalies that were not present in the training data.

3.2 Unsupervised Learning

Unsupervised learning is a machine learning approach that involves training a model on an unlabeled dataset, where the goal is to identify patterns or structures in the data. In the context of anomaly detection, unsupervised learning algorithms aim to identify data points that deviate significantly from the majority of the data.

3.2.1 Key Algorithms

- **Clustering Algorithms**: Clustering algorithms, such as k-means and DBSCAN, are commonly used for anomaly detection. These algorithms work by grouping similar data points into clusters, and data points that do not belong to any cluster or belong to a small cluster are considered anomalies. Clustering algorithms are particularly effective for detecting point anomalies in datasets with well-defined clusters.

- **Density-Based Methods**: Density-based methods, such as Local Outlier Factor (LOF) and Isolation Forest, are another class of unsupervised learning algorithms used for anomaly detection. These methods work by estimating the density of the data and identifying data points that have a significantly lower density than their neighbors. Density-based methods are particularly effective for detecting contextual anomalies in datasets with varying densities.

- **Autoencoders**: Autoencoders are a type of neural network that can be used for anomaly detection. Autoencoders work by learning a compressed representation of the data, and data points that cannot be accurately reconstructed from the compressed representation are considered anomalies. Autoencoders are particularly effective for detecting collective anomalies in high-dimensional datasets.

3.2.2 Strengths and Limitations

- **Strengths**: Unsupervised learning approaches do not require labeled data, making them well-suited for applications where labeled data is scarce or unavailable. These approaches are also capable of detecting novel types of anomalies that were not present in the training data, as they do not rely on pre-defined labels.

- **Limitations**: The main limitation of unsupervised learning approaches is their reliance on assumptions about the data distribution. For example, clustering algorithms assume that the data can be grouped into well-defined clusters, while density-based methods assume that anomalies are located in low-density regions. These assumptions may not hold true for all datasets, which can lead to poor performance in some cases.

3.3 Semi-Supervised Learning

Semi-supervised learning is a machine learning approach that combines elements of supervised and unsupervised learning. In the context of anomaly detection, semi-supervised learning algorithms are trained on a dataset that contains a small amount of labeled data and a large amount of unlabeled data. The goal is to leverage the labeled data to improve the performance of the model on the unlabeled data.

3.3.1 Key Algorithms

- **Self-Training**: Self-training is a semi-supervised learning algorithm that involves training a model on the labeled data and then using the model to predict labels for the unlabeled data. The predicted labels are then used to retrain the model, and the process is repeated until convergence. Self-training is particularly effective for detecting point anomalies in datasets with a small amount of labeled data.

- **Co-Training**: Co-training is a semi-supervised learning algorithm that involves training multiple models on different views of the data. The models are then used to predict labels for the unlabeled data, and the predictions are used to retrain the models. Co-training is particularly effective for detecting contextual anomalies in datasets with multiple feature sets.

- **Graph-Based Methods**: Graph-based methods, such as label propagation and graph convolutional networks, are another class of semi-supervised learning algorithms used for anomaly detection. These methods work by constructing a graph that represents the relationships between the data points, and then propagating the labels from the labeled data points to the unlabeled data points. Graph-based methods are particularly effective for detecting collective anomalies in datasets with complex relationships [16].

3.3.2 Strengths and Limitations

- **Strengths**: Semi-supervised learning approaches are effective when labeled data is scarce, as they can leverage the unlabeled data to improve the performance of the model. These approaches are also capable of capturing complex patterns and relationships in the data, making them well-suited for detecting anomalies in high-dimensional datasets.

- **Limitations**: The main limitation of semi-supervised learning approaches is their reliance on the quality of the labeled data. If the labeled data is biased or noisy, the performance of the model may be negatively affected. Additionally, semi-supervised learning models may not generalize well to new types of anomalies that were not present in the labeled data.

4. Applications of Machine Learning for Anomaly Detection in Various Domains

Machine learning techniques for anomaly detection have been applied in a wide range of domains, including cybersecurity, healthcare, finance, and industrial systems. In this section, we provide a detailed review of the applications of machine learning for anomaly detection in these domains.

4.1 Cybersecurity

Cybersecurity is one of the most important domains for anomaly detection, as the detection of malicious activities such as intrusions, malware, and phishing attacks is critical for protecting computer systems and networks. Machine learning techniques have been widely used for anomaly detection in cybersecurity, with a focus on detecting network intrusions, malware, and phishing attacks.

4.1.1 Network Intrusion Detection

Network intrusion detection involves monitoring network traffic for signs of malicious activities such as unauthorized access, denial-of-service attacks, and data exfiltration. Machine learning techniques, particularly unsupervised and semi-supervised learning algorithms, have been widely used for network intrusion detection.

- **Unsupervised Learning**: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect network intrusions by identifying patterns in network traffic that deviate from normal behavior. For example, k-means clustering has been used to group network traffic into clusters, and traffic that does not belong to any cluster or belongs to a small cluster is considered anomalous.
- **Semi-Supervised Learning**: Semi-supervised learning algorithms, such as self-training and co-training, have been used to detect network intrusions by leveraging a small amount of labeled data to improve the performance of the model on unlabeled data. For example, self-training has been used to train a model on a small amount of labeled network traffic data and then use the model to predict labels for the unlabeled data.

4.1.2 Malware Detection

Malware detection involves identifying malicious software such as viruses, worms, and ransomware. Machine learning techniques, particularly supervised learning algorithms, have been widely used for malware detection.

- **Supervised Learning**: Supervised learning algorithms, such as decision trees and random forests, have been used to detect malware by training a model on a labeled dataset of malware and benign software. The trained model can then be used to classify new software as malware or benign.
- **Deep Learning**: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been used to detect malware by learning a hierarchical representation of the software. For example, CNNs have been used to analyze the binary code of software to detect malware, while RNNs have been used to analyze the behavior of software over time.

4.1.3 Phishing Detection

Phishing detection involves identifying fraudulent websites and emails that attempt to steal sensitive information such as passwords and credit card numbers. Machine learning techniques, particularly supervised learning algorithms, have been widely used for phishing detection.

- **Supervised Learning**: Supervised learning algorithms, such as support vector machines (SVM) and neural networks, have been used to detect phishing websites and emails by training a model on a labeled dataset of phishing and legitimate websites/emails. The trained model can then be used to classify new websites/emails as phishing or legitimate.
- **Feature Engineering**: Feature engineering is an important aspect of phishing detection, as the choice of features can significantly impact the performance of the model. Common features used for phishing detection include the URL structure, the presence of suspicious keywords, and the use of SSL certificates.

4.2 Healthcare

Healthcare is another important domain for anomaly detection, as the early detection of anomalies such as diseases, medical errors, and equipment failures can have a significant impact on patient outcomes. Machine learning techniques have been widely used for anomaly detection in healthcare, with a focus on detecting diseases, medical errors, and equipment failures [17].

4.2.1 Disease Detection

Disease detection involves identifying diseases such as cancer, diabetes, and cardiovascular diseases based on patient data such as medical images, lab results, and electronic health records (EHRs). Machine learning techniques, particularly supervised and deep learning algorithms, have been widely used for disease detection.

- **Supervised Learning**: Supervised learning algorithms, such as decision trees and random forests, have been used to detect diseases by training a model on a labeled dataset of patient data. The trained model can then be used to classify new patient data as diseased or healthy.
- **Deep Learning**: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been used to detect diseases by learning a hierarchical representation of the patient data. For example, CNNs have been used to analyze medical images such as X-rays and MRIs to detect diseases, while RNNs have been used to analyze time-series data such as ECG signals to detect cardiovascular diseases.

4.2.2 Medical Error Detection

Medical error detection involves identifying errors such as misdiagnoses, medication errors, and surgical errors based on patient data such as EHRs and clinical notes. Machine learning techniques, particularly unsupervised and semi-supervised learning algorithms, have been widely used for medical error detection.

Unsupervised Learning: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect medical errors by identifying patterns in patient data that deviate from normal behavior. For example, k-means clustering has been used to group patient data into clusters, and data that does not belong to any cluster or belongs to a small cluster is considered anomalous [18].

- **Semi-Supervised Learning**: Semi-supervised learning algorithms, such as self-training and co-training, have been used to detect medical errors by leveraging a small amount of labeled data to improve the performance of the model on unlabeled data. For example, self-training has been used to train a model on a small amount of labeled patient data and then use the model to predict labels for the unlabeled data.

4.2.3 Equipment Failure Detection

Equipment failure detection involves identifying failures in medical equipment such as MRI machines, ventilators, and infusion pumps based on sensor data. Machine learning techniques, particularly unsupervised and deep learning algorithms, have been widely used for equipment failure detection.

- **Unsupervised Learning**: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect equipment failures by identifying patterns in sensor data that deviate from normal behavior. For example, k-means clustering has been used to group sensor data into clusters, and data that does not belong to any cluster or belongs to a small cluster is considered anomalous.
- **Deep Learning**: Deep learning algorithms, such as autoencoders and recurrent neural networks (RNNs), have been used to detect equipment failures by learning a compressed representation of the sensor data. For example, autoencoders have been used to detect anomalies in the sensor data of MRI machines, while RNNs have been used to detect anomalies in the sensor data of ventilators.

4.3 Finance

Finance is another important domain for anomaly detection, as the detection of anomalies such as fraudulent transactions, market manipulation, and credit risk is critical for maintaining the integrity of financial systems. Machine learning techniques have been widely used for anomaly detection in finance, with a focus on detecting fraudulent transactions, market manipulation, and credit risk [19], [20].

Table 2: Applications of Machine Learning for Anomaly Detection in Various Domains

Domain	Application	Key Techniques	Performance Metrics
Cybersecurity	Network Intrusion Detection	Unsupervised, Semi-Supervised	Accuracy, Precision, Recall
Healthcare	Disease Detection	Supervised, Deep Learning	AUC, F1 Score
Finance	Fraud Detection	Supervised, Unsupervised	Precision, Recall, F1 Score
Industrial Systems	Equipment Failure Detection	Unsupervised, Deep Learning	Accuracy, Precision, Recall

4.3.1 Fraud Detection

Fraud detection involves identifying fraudulent transactions such as credit card fraud, insurance fraud, and money laundering based on transaction data. Machine learning techniques, particularly supervised and unsupervised learning algorithms, have been widely used for fraud detection.

- **Supervised Learning**: Supervised learning algorithms, such as decision trees and random forests, have been used to detect fraudulent transactions by training a model on a labeled dataset of transaction data. The trained model can then be used to classify new transactions as fraudulent or legitimate.

Unsupervised Learning: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect fraudulent transactions by identifying patterns in transaction data that deviate from normal behavior. For example, k-means clustering has been used to group transaction data into clusters, and data that does not belong to any cluster or belongs to a small cluster is considered anomalous [21].

4.3.2 Market Manipulation Detection

Market manipulation detection involves identifying manipulative activities such as insider trading, pump-and-dump schemes, and spoofing based on market data. Machine learning techniques, particularly unsupervised and deep learning algorithms, have been widely used for market manipulation detection.

- **Unsupervised Learning**: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect market manipulation by identifying patterns in market data that deviate from normal behavior. For example, k-means clustering has been used to group market data into clusters, and data that does not belong to any cluster or belongs to a small cluster is considered anomalous.

- **Deep Learning**: Deep learning algorithms, such as autoencoders and recurrent neural networks (RNNs), have been used to detect market manipulation by learning a compressed representation of the market data. For example, autoencoders have been used to detect anomalies in the order book data of stock exchanges, while RNNs have been used to detect anomalies in the price and volume data of stocks.

4.3.3 Credit Risk Detection

Credit risk detection involves identifying the risk of default on loans and credit cards based on customer data such as credit scores, income, and payment history. Machine learning techniques, particularly supervised and semi-supervised learning algorithms, have been widely used for credit risk detection.

- **Supervised Learning**: Supervised learning algorithms, such as decision trees and random forests, have been used to detect credit risk by training a model on a labeled dataset of customer data. The trained model can then be used to classify new customers as high-risk or low-risk.

- **Semi-Supervised Learning**: Semi-supervised learning algorithms, such as self-training and co-training, have been used to detect credit risk by leveraging a small amount of labeled data to improve the performance of the model on unlabeled data. For example, self-training has been used to train a model on a small amount of labeled customer data and then use the model to predict labels for the unlabeled data.

4.4 Industrial Systems

Industrial systems are another important domain for anomaly detection, as the detection of anomalies such as equipment failures, process deviations, and quality defects is critical for maintaining the efficiency and safety of industrial processes. Machine learning techniques have been widely used for anomaly detection in industrial systems, with a focus on detecting equipment failures, process deviations, and quality defects [22].

4.4.1 Equipment Failure Detection

Equipment failure detection involves identifying failures in industrial equipment such as turbines, pumps, and conveyor belts based on sensor data. Machine learning techniques, particularly unsupervised and deep learning algorithms, have been widely used for equipment failure detection.

- **Unsupervised Learning**: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect equipment failures by identifying patterns in sensor data that deviate from normal behavior. For example, k-means clustering has been used to group sensor data into clusters, and data that does not belong to any cluster or belongs to a small cluster is considered anomalous.

- **Deep Learning**: Deep learning algorithms, such as autoencoders and recurrent neural networks (RNNs), have been used to detect equipment failures by learning a compressed representation of the sensor data. For example, autoencoders have been used to detect anomalies in the sensor data of turbines, while RNNs have been used to detect anomalies in the sensor data of conveyor belts [23].

4.4.2 Process Deviation Detection

Process deviation detection involves identifying deviations in industrial processes such as chemical reactions, manufacturing processes, and energy production based on process data. Machine learning techniques, particularly unsupervised and semi-supervised learning algorithms, have been widely used for process deviation detection [24].

Unsupervised Learning: Unsupervised learning algorithms, such as clustering and density-based methods, have been used to detect process deviations by identifying patterns in process data that deviate from normal behavior [25]. For example, k-means clustering has been used to group process data into clusters, and data that does not belong to any cluster or belongs to a small cluster is considered anomalous.

- **Semi-Supervised Learning**: Semi-supervised learning algorithms, such as self-training and co-training, have been used to detect process deviations by leveraging a small amount of labeled data to improve the performance of the model on unlabeled data. For example, self-training has been used to train a model on a small amount of labeled process data and then use the model to predict labels for the unlabeled data.

4.4.3 Quality Defect Detection

Quality defect detection involves identifying defects in industrial products such as semiconductors, automotive parts, and consumer goods based on inspection data. Machine learning techniques, particularly supervised and deep learning algorithms, have been widely used for quality defect detection.

- **Supervised Learning**: Supervised learning algorithms, such as decision trees and random forests, have been used to detect quality defects by training a model on a labeled dataset of inspection data. The trained model can then be used to classify new products as defective or non-defective.
- **Deep Learning**: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been used to detect quality defects by learning a hierarchical representation of the inspection data. For example, CNNs have been used to analyze images of semiconductor wafers to detect defects, while RNNs have been used to analyze time-series data of automotive parts to detect defects.

5. Challenges and Future Directions

Despite the significant progress made in the field of anomaly detection using machine learning, several challenges remain. In this section, we discuss the key challenges and future directions in the field of anomaly detection.

Table 3: Challenges and Future Directions in Anomaly Detection

Challenge	Description	Future Directions
Imbalanced Data	Anomalies are rare events	Data augmentation, Resampling
High Dimensionality	Curse of dimensionality	Feature selection, Extraction
Dynamic Environments	Concept drift	Online learning, Transfer learning
Label Scarcity	Scarce or unavailable labeled data	Semi-supervised, Active learning
Emerging Technologies	Deep learning, Reinforcement learning	Explore potential in anomaly detection

5.1 Imbalanced Data

One of the key challenges in anomaly detection is the imbalanced nature of the data, where anomalies are typically rare events [26], [27]. This imbalance can make it difficult for machine learning algorithms to learn the characteristics of anomalies, as they are often overshadowed by the majority class [28]. Future research should focus on developing techniques to address the imbalanced data problem, such as data augmentation, resampling, and cost-sensitive learning.

5.2 High Dimensionality

Another challenge in anomaly detection is the high dimensionality of the data, which can make it difficult to detect anomalies due to the curse of dimensionality. Future research should focus on developing techniques to reduce the dimensionality of the data, such as feature selection, feature extraction, and dimensionality reduction algorithms [29].

5.3 Dynamic Environments

In many applications, the data distribution may change over time, leading to concept drift. Concept drift can make it difficult for anomaly detection models to maintain their performance over time, as the patterns of normal behavior may evolve. Future research should focus on developing techniques to detect and adapt to concept drift, such as online learning, transfer learning, and ensemble learning [11].

5.4 Label Scarcity

In many cases, labeled data for anomalies is scarce or unavailable, which can make it difficult to train supervised learning models for anomaly detection. Future research should focus on developing techniques to leverage unlabeled data, such as semi-supervised learning, active learning, and self-supervised learning [30].

5.5 Emerging Technologies

Emerging technologies such as deep learning and reinforcement learning have the potential to advance the field of anomaly detection. Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promising results in detecting anomalies in complex and high-dimensional datasets. Reinforcement learning algorithms, which learn to make decisions based on feedback from the environment, have the potential to improve the adaptability and robustness of anomaly detection models [31]. Future research should focus on exploring the potential of these emerging technologies in the field of anomaly detection [32].

6. Conclusion

In this paper, we have provided a comprehensive review of machine learning techniques for anomaly detection, focusing on their applications across various domains. We have discussed the strengths and limitations of different machine learning approaches, including supervised, unsupervised, and semi-supervised learning, and highlighted the challenges and future directions in the field of anomaly detection [33], [34]. The review is supported by three detailed tables that summarize the key techniques, their applications, and performance metrics [35].
The key findings of this review are as follows:

- Machine learning techniques have been widely used for anomaly detection in various domains, including cybersecurity, healthcare, finance, and industrial systems.
- Supervised learning approaches are effective when labeled data is available, but they may not generalize well to new types of anomalies.
- Unsupervised learning approaches do not require labeled data, but they rely on assumptions about the data distribution that may not hold true for all datasets.
- Semi-supervised learning approaches are effective when labeled data is scarce, but they rely on the quality of the labeled data.
- Emerging technologies such as deep learning and reinforcement learning have the potential to advance the field of anomaly detection.

Future research should focus on addressing the key challenges in anomaly detection, such as imbalanced data, high dimensionality, dynamic environments, and label scarcity, and exploring the potential of emerging technologies in the field of anomaly detection [36].

References

[1] A. R. Wheeler and M. R. Buckley, "Near-term human resources challenges in the age of automation, artificial intelligence, and machine learning," in *HR without People?*, Emerald Publishing Limited, 2021, pp. 69–84.
[2] S. Mahdavi, S. Rahnamayan, and K. Deb, "Opposition based learning: A literature review," *Swarm and Evolutionary Computation*, vol. 39, pp. 1–23, Apr. 2018.
[3] J. G. C. Ramírez, "Integrating AI and NISQ technologies for enhanced mobile network optimization," *QJETI*, vol. 5, no. 1, pp. 11–22, Jan. 2020.
[4] A. Feroze, A. Daud, T. Amjad, and M. K. Hayat, "Group anomaly detection: Past notions, present insights, and future prospects," *SN Comput. Sci.*, vol. 2, no. 3, May 2021.
[5] G. Bussi and A. Laio, "Using metadynamics to explore complex free-energy landscapes," *Nature Reviews Physics*, vol. 2, no. 4, pp. 200–212, Mar. 2020.
[6] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," *Proceedings of the 2017 ACM*, 2017.
[7] J. G. C. Ramírez, "Quantum control and gate optimization in graphane-based quantum systems," *J. Appl. Math. Mech.*, vol. 4, no. 1, pp. 69–79, Oct. 2020.
[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
[9] "A survey of outlier detection methodologies."
[10] J. G. C. Ramírez, "Vibration analysis with AI: Physics-informed neural network approach for vortex-induced vibration," *Int. J. Radiat. Appl. Instrum. C Radiat. Phys. Chem.*, vol. 11, no. 3, Mar. 2021.
[11] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, 2016.

[12] A. Carlevaro, T. Alamo, F. Dabbene, and M. Mongelli, "Conformal predictions for probabilistically robust scalable machine learning classification," *Mach. Learn.*, vol. 113, no. 9, pp. 6645–6661, Sep. 2024.
[13] J. G. C. Ramírez, "The role of graphene in advancing quantum computing technologies," *Annu. Rep. - Aust. Inst. Criminol.*, vol. 4, no. 1, pp. 62–77, Feb. 2021.
[14] J. G. C. Ramírez, "Enhancing temporal quantum coherence in graphene-based superconducting circuits," *International Journal of Applied Machine Learning and Computational Intelligence*, vol. 11, no. 12, Dec. 2021.
[15] J. G. C. Ramírez, M. Hassan, and M. Kamal, "Applications of artificial intelligence models for computational flow dynamics and droplet microfluidics," *JSTIP*, vol. 6, no. 12, Dec. 2022.
[16] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," *arXiv [cs.LG]*, 05-Jul-2020.
[17] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS One*, vol. 11, no. 4, p. e0152173, Apr. 2016.
[18] V. Ramamoorthi, "Applications of AI in Cloud Computing: Transforming Industries and Future Opportunities," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 9, no. 4, pp. 472–483, Aug. 2023.
[19] K. K. R. Yanamala, "Predicting employee turnover through machine learning and data analytics," *AI, IoT and the Fourth Industrial Revolution Review*, vol. 10, no. 2, pp. 39–46, Feb. 2020.
[20] G. Garau Estarellas, G. L. Giorgi, M. C. Soriano, and R. Zambrini, "Machine learning applied to quantum synchronization-assisted probing," *Adv. Quantum Technol.*, vol. 2, no. 7–8, p. 1800085, Aug. 2019.
[21] V. Ramamoorthi, "Exploring AI-Driven Cloud-Edge Orchestration for IoT Applications," 2023.
[22] J. G. C. Ramírez and M. Kamal, "Theoretical exploration of two-dimensional materials for quantum computing applications," *JICET*, vol. 8, no. 4, pp. 45–57, Nov. 2023.
[23] A. Hilali, H. Hafiddi, and Z. El Akkaoui, "Microservices adaptation using machine learning: A systematic mapping study," in *Proceedings of the 16th International Conference on Software Technologies*, Online Streaming, --- Select a Country ---, 2021.
[24] J. G. C. Ramírez and M. Kamal, "Graphene plasmonics for enhanced quantum information processing," *AIFIR*, vol. 13, no. 11, pp. 18–25, Nov. 2023.
[25] J. G. C. Ramirez, "From Autonomy to Accountability: Envisioning AI's Legal Personhood," *ARAIC*, vol. 6, no. 9, pp. 1–16, Sep. 2023.
[26] E. Gibney, "The battle for ethical AI at the world's biggest machine-learning conference," *Nature*, vol. 577, no. 7792, p. 609, Jan. 2020.
[27] B. Rathore, "Cloaked in Code: AI & Machine Learning Advancements in Fashion Marketing," *Eduzone: International Peer Reviewed/Refereed*, 2017.
[28] J. G. C. Ramirez, "How Mobile Applications can improve Small Business Development," *ERST*, vol. 7, no. 1, pp. 291–305, Nov. 2023.
[29] J. G. C. Ramírez, "Incorporating Information Architecture (ia), Enterprise Engineering (ee) and Artificial Intelligence (ai) to Improve Business Plans for Small Businesses in the United States," *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, vol. 2, no. 1, pp. 115–127, 2023.
[30] L. Ruff *et al.*, "Deep one-class classification," *ICML*, vol. 80, pp. 4390–4399, Jul. 2018.
[31] J. G. C. Ramirez, "Comprehensive exploration of the CR model: A systemic approach to Strategic Planning," *International Journal of Culture and Education*, vol. 1, no. 3, Aug. 2023.
[32] V. Ramamoorthi, "Optimizing Cloud Load Forecasting with a CNN-BiLSTM Hybrid Model," *International Journal of Intelligent Automation and Computing*, vol. 5, no. 2, pp. 79–91, Nov. 2022.
[33] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications surveys & tutorials*, 2015.
[34] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data Soc.*, vol. 3, no. 1, p. 205395171562251, Jan. 2016.
[35] V. Ramamoorthi, "Real-Time Adaptive Orchestration of AI Microservices in Dynamic Edge Computing," *Journal of Advanced Computing Systems*, vol. 3, no. 3, pp. 1–9, Mar. 2023.
[36] J. G. C. Ramírez, "Struggling Small Business in the US. The next challenge to economic recovery," *IJBIBDA*, vol. 5, no. 1, pp. 81–91, Feb. 2022.