Anomaly Detection to Network Instrusion:A Systematic Literature Review

[1]Tapasvi M I,[2]Yashaswini B C,[3]Yuktha S Gowda,[4]Yogeesh M Under The Guidence:Ravi Kumar D
Computer and Science and Engineering, Malnad College of Engineering
Hassan-573202, India

Email id: tapasvigowda46@gmail.com,yashaswinibc01@gmail.com, yukthasgowda512@gmail.com, yogeeshy39@gmail.com

***Abstract* -In networking, anomaly detection is essential to preserving network infrastructure performance, security, and dependability. The increasing volume of network traffic and the dynamic and complicated character of contemporary networks pose serious problems for conventional anomaly detection techniques. The application of machine learning (ML) approaches to identify network traffic anomalies, such as malicious activity, performance deterioration, and configuration problems, is examined in this study. Supervised, unsupervised, and semi-supervised machine learning models provide improved accuracy in detecting trends and anomalies in real-time data streams.**

The study compares the efficacy of many machine learning techniques, including decision trees, support vector machines, k-means clustering, and neural networks, in identifying various anomalies across a range of network contexts. The difficulties in handling imbalanced datasets, training models with labeled data, and lowering false positives are also highlighted in the research. It also talks about how anomaly detection systems can be integrated with intrusion detection systems and network monitoring tools to create a stronger security architecture. All things considered, using machine learning to networking anomaly detection is a promising strategy for risk reduction and guaranteeing the seamless operation of contemporary networked systems.

Keywords*—*Machine learning, Deep learning ,Local Outlier Factor ,One-Class Support Vector Machine,Autoencoder

I. INTRODUCTION

As networked systems expand quickly and become more complex, maintaining security, performance, and dependability has become increasingly difficult. Conventional anomaly detection techniques, including rule-based or threshold-based methods, frequently find it difficult to successfully handle dynamic and new network threats. Operations can be seriously disrupted by network abnormalities such as hardware failures, congestion, denial-of-service assaults, and security breaches. A proposed method for more accurately and adaptably identifying network anomalies is machine learning (ML). ML models may learn to identify typical behavior and spot deviations that might indicate possible problems by examining network records, traffic patterns, and user activities. ML can find previously unidentified attack pathways and flaws, in contrast to established rules.

There are three types of ML-based anomaly detection methods: supervised, unsupervised, and semi-supervised. Semi-supervised techniques blend tiny labeled datasets with large unlabeled ones, supervised techniques use labeled data to train models, while unsupervised techniques identify abnormalities without labeled data. However, real-time implementation is complicated by issues including high false positive rates, uneven data, and the requirement for enormous datasets. To improve defense, intrusion detection systems (IDS) and network monitoring tools must be integrated effectively. In order to ensure security and operational efficiency in complex networks, this study analyzes machine learning (ML) techniques for network anomaly detection, assesses their advantages and disadvantages, and investigates how they can improve contemporary network management tactics.

II. RELATED WORKS

A. PaperTitle: Machine Learning for Anomaly Detection: A Systematic Review

Description: This systematic review investigates the use of machine learning (ML) techniques for anomaly detection. It evaluates 290 research articles published between 2000 and 2020. The study focuses on anomaly detection applications, ML techniques, performance metrics, and classification types. The goal is to provide a comprehensive guide for researchers in the field.. .

Methodology: The writers used the systematic review methodology developed by Kitchenham and Charters. They developed research questions, applied quality evaluation guidelines, and used inclusion/exclusion criteria to choose pertinent papers. Questions concerning ML methods, performance, and applicability were addressed by extracting

data. Analysis was done using both narrative and quantitative synthesis.

Limitations: The study is restricted to journal and conference papers, potentially missing non-academic sources. Some articles lacked clarity on performance metrics or ML techniques. The reliance on older datasets and incomplete classification types are notable limitations. Despite its breadth, the review suggests further investigation is needed.

Key Insights: 27% of studies on anomaly detection are dominated by unsupervised learning. The most widely used machine learning technology is Support Vector Machines (SVMs). PCA and CFS are two often utilized feature selection techniques. Many studies use narrow metrics, such as accuracy, to evaluate models insufficiently.. .

Citation: Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Fatima Mohamad Dakalbab, "Machine Learning for Anomaly Detection: A Systematic Review," IEEE Access, Vol. 9, 2021, pp. 78658–78692. DOI: [10.1109/ACCESS.2021.3083060](https://doi.org/10.1109/ACCESS.2021.3083060)

B. Paper Title: Automatic Bug Fixing

Description: This paper explores automating bug fixing by integrating it into an existing automatic debugging tool. The goal is to identify and apply fixes for regression bugs to make failing tests pass. The paper highlights challenges, lessons learned, and recommendations based on internal trials.

Methodology: The automatic bug-fix process involves running a test suite, debugging failures, validating fixes, and optionally committing them. The authors implemented this flow in the PinDown tool and conducted internal trials to analyze its effectiveness. Key steps include validating fixes locally and ensuring no conflicts with concurrent changes. .

Limitations: The approach is limited by challenges such as human-tool race conditions, fault oscillation, and issues with partial implementations. Randomized test failures further complicate automated fixes. The study recommends local fixes without

Key Insights: Automating bug fixing is feasible but should focus on local fixes and validated bug reports. Committing fixes automatically can lead to inefficiencies and conflicts. Continuous integration works well for small, directed tests, but larger, randomized suites benefit from automated local fixes and manual validation.

Citation: Daniel Hansson, "Automatic Bug Fixing," 16th International Workshop on Microprocessor and SOC Test and Verification (MTV), 2015, DOI: [10.1109/MTV.2015.21](https://doi.org/10.1109/MTV.2015.21).

C. Paper title: A Systematic Review of Anomaly Detection Using Machine and Deep Learning Techniques

Description: This paper reviews anomaly detection techniques employing machine and deep learning models from 2019 to 2021. It discusses various applications, including traffic surveillance and healthcare, and highlights the challenges such as limited training data and environmental variations. It aims to provide insights into anomaly detection methods, their performance metrics, and real-world dataset usage.

Methodology: The study uses a two-stage systematic review process. First, it collects relevant research papers using search engines and digital libraries like IEEE and Springer. Boolean operators are applied to refine results. The second stage evaluates the selected papers based on accuracy, performance, and dataset applicability, categorizing techniques by machine and deep learning methods. .

Limitations: Challenges include the unavailability of labeled datasets for real-world scenarios, high noise in data, and limited adaptability of detection algorithms to dynamic environments. Moreover, most methods struggle with high time complexity and are unsuitable for datasets with unknown anomaly characteristics.

Key Insights:: The review emphasizes the evolution of anomaly detection methods, noting significant improvements in machine and deep learning techniques. However, it identifies a need for benchmark datasets and interdisciplinary collaboration. Improved models with cross- validation, reduced false positives, and low-cost implementation are crucial for future advancements.

Citation: Sarfaraz Natha, Mehwish Leghari, Muhammad Awais Rajput, Syed Saood Zia, Jawaid Shabir. (2022). A Systematic Review of Anomaly Detection Using Machine and Deep Learning Techniques. QUEST Research Journal, 20(01), 83–94. DOI: 10.52584/QRJ.2001.11.

D. Paper title: Real-Time Deep Anomaly Detection Framework for Multivariate Time-Series Data in Industrial IoT

Description: This paper proposes a hybrid deep anomaly detection framework designed for Industrial IoT (IIoT) environments. Using convolutional neural networks (CNNs) and long short-term memory (LSTM)-based autoencoders (AEs), the framework identifies anomalies and rare events in multivariate time-series data. The model supports real- time inference on edge devices, enabling efficient and accurate anomaly detection.

Methodology: The framework integrates a CNN to extract spatial features and a two-stage LSTM AE to capture temporal dependencies. It employs an unsupervised learning

approach for detecting short- and long-term variations in time-series data. The system was trained using TensorFlow and optimized for deployment on edge devices through quantization and performance tuning. .

Limitations: The study identifies challenges in scalability and latency when handling large-scale datasets. Additionally, the model's performance may decline with increased data size or noisy datasets. Future research is needed to address communication overhead, privacy issues, and scalability for diverse IIoT applications.

Key Insights: The proposed model outperforms state-of-the-art anomaly detection methods in accuracy, precision, and inference time. It demonstrates the potential for real-time anomaly detection on resource-constrained edge devices, offering robust solutions for IIoT scenarios. Optimization for edge deployment and low-latency processing are emphasized as critical advancements.

Citation: Hussain Nizam, Samra Zafar, Zefeng Lv, Fan Wang, and Xiaopeng Hu. (2022). "Real-Time Deep Anomaly Detection Framework for Multivariate Time-Series Data in Industrial IoT." IEEE Sensors Journal, 22(23), 22836–22848. DOI: 10.1109/JSEN.2022.3211874.

E. Paper titleHybrid Statistical-Machine Learning for Real-Time Anomaly Detection in Industrial Cyber-Physical Systems

Description: This paper introduces a hybrid anomaly detection model for Industrial Cyber-Physical Systems (ICPS), combining statistical methods and machine learning. The framework integrates Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long Short-Term Memory (LSTM) networks to identify network traffic anomalies. It effectively detects cyberattacks, malicious behaviors, and system irregularities in real-time while maintaining computational efficiency.

Methodology: The SARIMA model predicts short-term patterns and establishes dynamic thresholds, while LSTM captures long-term dependencies in network traffic. Real-time and offline analyses are combined to improve detection accuracy. A custom ICPS testbed simulates industrial systems such as power grids and gas pipelines to evaluate the model's performance

Limitations: Scalability to diverse ICPS setups and adaptability to new threats remain challenging. The computational efficiency, while improved, could limit deployment in highly resource-constrained environments. Benchmarking against more state-of-the-art techniques and larger datasets is suggested for future validation.

Key Insights: The hybrid model achieves high detection accuracy with low computational complexity, identifying anomalies such as cyberattacks with minimal false positives. It highlights the importance of combining statistical models and machine learning to balance precision and efficiency in anomaly detection.

Citation: Hao, Weijie, Tao Yang, and Qiang Yang. "Hybrid Statistical-Machine Learning for Real-Time Anomaly Detection in Industrial Cyber-Physical Systems." IEEE Transactions on Automation Science and Engineering, vol. 20, no. 1, Jan. 2023, pp. 32–45. DOI: 10.1109/TASE.2021.3073396

F. Paper title: A Hybrid Machine-Learning Ensemble for Anomaly Detection in Real-Time Industry 4.0 Systems

Description: This study addresses the challenge of real-time anomaly detection in complex Industry 4.0 systems. It proposes a hybrid machine-learning ensemble combining Local Outlier Factor, One-Class Support Vector Machine, and Autoencoder models. Using weighted averages, the ensemble improves detection accuracy. Tested on industrial air-blowing machines, it demonstrated high F1-scores and adaptability for predictive maintenance in real-time settings.

Methodology: The pipeline consists of two stages: a manufacturing stage, where the model is trained using sensor data and preprocessed for quality control, and an operation stage, where the trained model is deployed for real-time anomaly detection. Feature selection, clustering, and dimensionality reduction (PCA) are applied. Three individual machine-learning models are integrated into a hybrid model using weighted averaging, with weights derived from individual F1-scores. .

Limitations: - Performance relies on the quality and completeness of the training data.

- The Autoencoder model demonstrated relatively low performance compared to the others.

- Limited to single-type anomaly detection; fault classification requires further research.

- Real-time calibration and retraining in industrial settings present computational challenges.

Key Insights: - The hybrid ensemble outperformed individual models in F1-scores and AUC metrics across three test machines.

- Demonstrated feasibility for real-time applications with computational efficiency on standard hardware.

- Vertical scalability allows integration of additional models to enhance detection performance.

- Future enhancements include addressing multi-class anomaly detection and incorporating explainable AI techniques.

Citation: Velásquez, D., Pérez, E., Oregui, X., Artetxe, A., Manteca, J., Mansilla, J. E., Toro, M., Maiza, M., & Sierra, B. (2022). A Hybrid Machine-Learning Ensemble for Anomaly Detection in Real-Time Industry 4.0 Systems. IEEE Access, 10, 72024-72036. DOI:

[10.1109/ACCESS.2022.3188102](https://doi.org/10.1109/ACCESS.2022.3188102).

III METHODOLOGY

With two stages—manufacturing and operation—the suggested methodology presents a hybrid machine-learning pipeline for real-time anomaly identification in industrial systems. During quality control procedures in the manufacturing stage, industrial sensors capture variables such as pressure, power, temperature, and flow rate. By eliminating invalid items and standardizing ranges using Min-Max scaling, preprocessing guarantees the quality of the data. Using algorithms and domain knowledge as a guide, feature selection finds key variables and removes duplicates. By retaining 90% of the data variance, Principal Component Analysis (PCA) reduces dimensionality and complexity. DBSCAN finds outliers in the steady-state data, while K-Means clustering distinguishes between transient and steady stages of operation.

While steady-state data is designated as normal (1), outliers and transient states are designated as anomalies (-1). Normal data is used to train three machine learning models: Autoencoder, One-Class Support Vector Machine (OCSVM), and Local Outlier Factor (LOF). In order to determine model weights for merging outputs into a hybrid ensemble model using a weighted average, validation sets calculate F1-scores. This hybrid model analyzes real-time sensor data throughout the operation stage to categorize points as normal or abnormal, sending out notifications when anomalies are found. By labeling data during stable conditions and retraining the system to adjust to machine degradation, operators may periodically recalibrate the model, guaranteeing continued accuracy.

With F1-scores of 0.904, 0.890, and 0.887 for each of the three industrial air-blowing devices tested, the technique outperformed individual models in terms of precision and recall. The system met industrial criteria for real-time detection with a maximum reaction time of 190 milliseconds. This scalable and reliable pipeline is a workable solution for Industry 4.0 applications since it allows for the inclusion of extra algorithms

III. ALGORITHMS USED

A. One-Class Support Vector Machine (OCSVM)

One-Class Support Vector Machine (OCSVM) is a specialized machine learning algorithm used primarily for anomaly detection, especially in scenarios where labeled data for abnormal instances is scarce or unavailable. OCSVM operates under the assumption that the majority of the data represents normal behavior. It works by learning a decision boundary in the feature space that encapsulates most of the normal data points. Any new data point that falls outside this learned boundary is classified as an anomaly. The algorithm employs kernel functions, such as the Radial Basis Function (RBF), to transform input data into a higher-dimensional space, making it easier to distinguish between normal and anomalous points even in complex datasets. OCSVM is particularly effective in high-dimensional spaces and with non-linear data distributions. It is widely used in applications like network intrusion detection, fraud detection, and predictive maintenance. During training, OCSVM adjusts its parameters to maximize the separation between normal data and outliers while minimizing the number of false positives. This makes it suitable for applications where anomalies are rare but critical, and the cost of missed detection is high. However, its performance can be sensitive to hyperparameter selection, such as the kernel type and regularization parameters, requiring careful tuning for optimal results. Additionally, OCSVM assumes a relatively balanced representation of normal data, which might necessitate preprocessing steps like normalization or oversampling in imbalanced datasets. Despite these challenges, its unsupervised learning nature and adaptability make OCSVM a powerful tool for real-world anomaly detection scenarios.

B. Smote Algorithm

The SMOTE (Synthetic Minority Oversampling Technique) algorithm is a solution for handling class imbalance in machine learning tasks, particularly useful in anomaly detection. Class imbalance occurs when one class (e.g., anomalies) has significantly fewer samples compared to the majority class (e.g., normal data), which often leads to biased model predictions. SMOTE addresses this by generating synthetic data points for the minority class.

In the context of network anomaly detection, anomalies such as Distributed Denial of Service (DDoS) attacks or unauthorized intrusions form the minority class. SMOTE enhances the model's ability to detect such anomalies by creating synthetic samples for these rare events. This is achieved by identifying nearest neighbors of the minority class samples and generating new synthetic data points through interpolation between these samples. The process involves three main steps: identifying minority class data, applying SMOTE to generate synthetic data, and retraining the machine learning model on the augmented dataset. By doing so, the algorithm balances the dataset, enabling the model to better learn the characteristics of both normal and anomalous classes.

After retraining with SMOTE-enhanced data, the model's performance is evaluated using metrics like Precision, Recall, F1-Score, and ROC-AUC. These evaluations often

show significant improvements, especially in detecting low-frequency anomalies. SMOTE is thus a critical preprocessing step in imbalanced data scenarios, ensuring that machine learning models can more effectively identify and classify anomalies in applications like network intrusion detection.

C. Local Outlier Factor(LOF)

The Local Outlier Factor (LOF) algorithm is a density-based anomaly detection method designed to identify data points that deviate significantly from their local neighborhood. LOF measures the local density of a point compared to its neighbors, using the concept of k-nearest neighbors (k-NN). The key idea is to calculate the "local reachability density" of a point, which represents the density of points in its neighborhood, and compare it to the densities of its nearest neighbors. If the local reachability density of a point is significantly lower than that of its neighbors, the point is considered an outlier. LOF assigns an anomaly score to each data point, where a score close to 1 indicates the point is similar to its neighbors (normal), and higher scores suggest potential outliers. This algorithm is particularly useful for datasets with varying densities, as it adapts to the local structure of the data. LOF is non-parametric and does not assume a global distribution, making it effective for detecting anomalies in complex and heterogeneous datasets. However, it requires careful selection of parameters, such as the number of neighbors (k), which can influence its sensitivity and accuracy. LOF is widely used in applications like fraud detection, network security, and industrial monitoring, where local anomalies need to be identified efficiently.

E. Autoencoder

Autoencoders are a type of unsupervised neural network architecture widely used in anomaly detection to identify unusual patterns in data by learning an efficient representation of the input. They consist of two main components: an encoder, which compresses input data into a lower-dimensional latent representation, and a decoder, which reconstructs the input from this representation. The key metric in autoencoder-based anomaly detection is the reconstruction error, which measures the difference between the input and its reconstructed output. During training, autoencoders are typically fed only normal data, allowing the network to learn to reconstruct these inputs with minimal error. When encountering anomalous data during inference, the reconstruction error increases, signaling an anomaly. The architecture of autoencoders includes input layers for high-dimensional data, hidden layers for encoding and decoding, and a latent space that captures the essential features of the data. Hyperparameters such as the number of hidden layers, the size of the latent space, and activation functions are critical for optimizing performance. Autoencoders can be further enhanced with regularization techniques to prevent overfitting and improve generalization. In hybrid ensemble models for anomaly detection, autoencoders complement other approaches like Local Outlier Factor and One-Class SVM by providing a reconstruction-based perspective. While effective, their performance depends heavily on proper hyperparameter tuning and their ability to handle time-series data.

IV. RESULTS

The final results of the study highlight the effectiveness of a hybrid machine-learning ensemble model for real-time anomaly detection in Industry 4.0 systems. The ensemble model integrates three machine-learning techniques: Local Outlier Factor (LOF), One-Class Support Vector Machine (OCSVM), and Autoencoder, using a weighted average approach based on their F1-scores. Tested on three industrial air-blowing machines, the ensemble demonstrated improved performance over individual models. For Machine A, the hybrid model achieved an F1-score of 0.904 for anomalies and 0.944 for normal data, with an AUC of 0.913. For Machine B, it reached 0.890 and 0.946 for anomalies and normal data, respectively, with an AUC of 0.905. For Machine C, the scores were 0.887 for anomalies, 0.889 for normal data, and an AUC of 0.897. Despite slightly higher computation times, the ensemble processed data within real- time thresholds, confirming its suitability for deployment. The research concludes that the hybrid model enhances anomaly detection accuracy and robustness, making it viable for dynamic industrial environments. Future work involves testing additional algorithms, addressing system degradation over time, and expanding datasets for generalization across diverse industrial machines.

V. CONCLUSION

The hybrid ensemble method improved anomaly detection accuracy and real-time response for Industry 4.0 systems. Its flexible architecture supports vertical scaling with additional models. Future work includes retraining cost studies, degradation handling, and testing on larger machine datasets. Explainable AI could enhance fault categorization in subsequent iterations.

VI. REFERENCES

[1] Aburomman AA, Reaz MBI. A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers and Security. 2017;65:135-152.

[2] Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques. Procedia Computer Science. 2015;60:708-713.

[3] Ahmad S, Lavin A, Purdy S, Agha Z. Unsupervised real-time anomaly detection for streaming data. Neurocomputing. 2017;262:134-147.

[4] Singh RP, Javaid M, Haleem A, Suman R. "Internet of things (IoT) applications to fight against COVID-19 pandemic," Diabetes and Metabolic Syndrome: Clinical Research and Reviews. 2020;14:521-524.

[5] Stoyanova M, Nikoloudakis Y, Panagiotakis S, Pallis E, Markakis EK. "A survey on the internet of things (IoT) forensics: Challenges, approaches, and open issues," IEEE Communications Surveys and Tutorials. 2020;22:1191-1221.

[6] Hassija V, Chamola V, Saxena V, Jain D, Goyal P, Sikdar B. "A survey on IoT security: application areas, security threats, and solution architectures," IEEE Access. 2019;7:82721-82743.

[7] Adat V, Gupta B. ”Security in Internet of Things: Issues, challenges, taxonomy an architecture,Telecommunication Systems.2018;67:423-441.

[8]. Hassija V, Chamola V, Saxena V, Jain D, Goyal P, Sikdar B. "A survey on IoT security: application areas, security threats, and solution architectures," IEEE Access. 2019;7:82721-82743.

[9]. Adat V, Gupta B. "Security in Internet of Things: Issues, challenges, taxonomy and architecture," Telecommunication Systems. 2018;67:423-441.

[10]. Anderson JP. Computer security threat monitoring and surveillance. Technical Report, Fort Washington, PA, James P. Anderson Co; 1980.

[11]. Bock T. Displayr blog. https://www.displayr.com/what-is-hierarchical-clustering.

[12]. Ashfaq RAR, Wang XZ, Huang JZ, Abbas H, He YL. Fuzziness-based semi-supervised learning approach for intrusion detection system. Information Sciences. 2017;378:484-497.

[13]. Aung YY, Min MM. An analysis of K-means algorithm- based network intrusion detection system. Advances in Science, Technology and Engineering Systems Journal. 2018;3(1):496-501.

[14].Bauer FC, Muir DR, Indiveri G. Real-time ultra-low power ECG anomaly detection using an event-driven neuromorphic processor. IEEE Transactions on Biomedical Circuits and Systems. 2019;13:1575-1582

[15]. Bhati BS, Rai CS, Balamurugan B, Al-Turjman F. An intrusion detection scheme based on the ensemble of discriminant classifiers. Computers and Electrical Engineering. 2020;86:106742..Bhattacharyya DK, Kalita JK. Network anomaly detection: A machine learning perspective.
CRC Press; 2013

[16]. Blanco R, Malagón P, Briongos S, Moya JM. Anomaly detection using Gaussian mixture probability model to implement intrusion detection system. In: International Conference on Hybrid Artificial Intelligence Systems. Springer;2019.p.648-659.