ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

STATISTICAL ANALYSIS OF ANOMALY DETECTION ALGORITHMS IN BIG DATA ENVIRONMENTS

XYZ

Abstract

In the era of large records, detecting anomalies has become a critical mission across various industries, which include utilities, cybersecurity, and the petroleum enterprise. This paper presents a comprehensive statistical analysis of anomaly detection algorithms carried out in big data environments, focusing at the software zone's smart metering, cybersecurity, turbomachinery within the petroleum industry, and the Internet of Things (IoT). We discover the utility of unsupervised and supervised gadget getting to know (ML) techniques for identifying anomalous styles in time-collection data. A hybrid approach combining Spectral Residual-Convolutional Neural Networks (SR-CNN) and martingale-primarily based anomaly detection is applied to clever meter facts, achieving high accuracy in identifying suspicious behavior. We also present a cloud-primarily based Intrusion Detection System (IDS) leveraging Apache Spark and the MAWILab dataset, demonstrating the gadget's efficacy in real-time cyber-assault detection with close to-perfect accuracy. Additionally, we observe the use of one-class guide vector machines and YASA segmentation for anomaly detection in turbomachinery, addressing the challenges posed by unlabeled statistics in high-frequency sensor environments. Lastly, we evaluate anomaly detection methodologies for IoT structures, highlighting the capability of blockchain-based collaborative studying for reinforcing security in these aid-constrained and distributed networks. The paper concludes with a statistical assessment of the overall performance metrics across the diverse domains, emphasizing the effectiveness of device getting to know algorithms in large information environments.

Keywords:

Anomaly Detection, Big Data, Machine Learning, Smart Metering, Cybersecurity, Intrusion Detection System, IoT, Turbomachinery, Unsupervised Learning, Spectral Residual-Convolutional Neural Network, One-Class Support Vector Machine, Blockchain, Apache Spark, Real-Time Detection.

I. INTRODUTION

The exponential boom of data in latest years, frequently referred to as the "Big Data technology," has created new demanding situations and opportunities throughout various industries. Among these challenges, the detection of anomalies within huge and complex datasets stands proud as a vital challenge. Anomalies, which are styles or observations that deviate substantially from the predicted behavior, can provide valuable insights or signal capability troubles along with fraud, device malfunctions, or cyber-attacks. Detecting those anomalies effectively and as it should be is vital for maintaining the security, reliability, and efficiency of systems that rely on large-scale facts.

Anomaly detection is specifically relevant in fields inclusive of application management, cybersecurity, commercial tracking, and the Internet of Things (IoT). For instance, in clever metering structures, the detection of non-technical losses (NTLs) which include power theft can save you full-size monetary losses for software organizations. In cybersecurity, figuring

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

out anomalous community visitors can be essential for thwarting capacity cyber-attacks before they motive harm. Similarly, in commercial settings just like the petroleum industry, detecting uncommon patterns in sensor facts from turbomachinery can save you high-priced system failures. The IoT, with its huge community of interconnected gadgets, also gives precise demanding situations for anomaly detection due to the heterogeneity and scale of the records concerned.

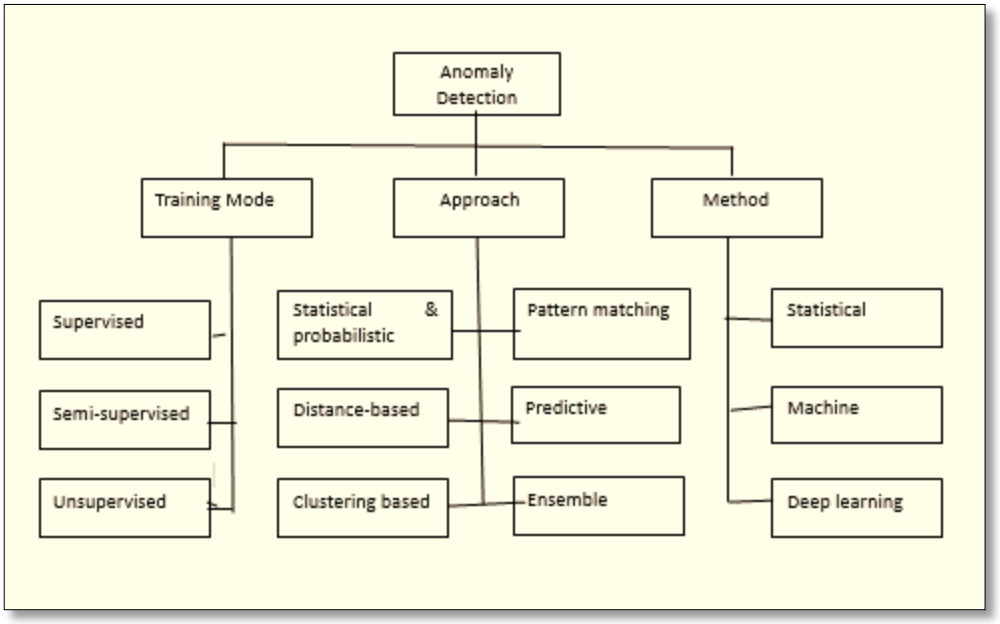

FIGURE 1: Anomaly Detection

Despite the significance of anomaly detection, the development of powerful algorithms for this purpose is fraught with challenges. These consist of the excessive dimensionality of statistics, the presence of noise, the shortage of categorised data for schooling, and the want for actual-time processing in many programs. Traditional anomaly detection techniques regularly conflict to hold tempo with the extent and speed of statistics generated in massive information environments. Consequently, there has been a developing hobby in leveraging superior system getting to know (ML) strategies to beautify anomaly detection abilities.

In this paper, we provide a complete statistical analysis of various anomaly detection algorithms implemented in big information environments. We discover the effectiveness of both supervised and unsupervised ML tactics across distinctive domains, which includes smart metering, cybersecurity, business monitoring, and IoT structures. The examine examines the performance of algorithms inclusive of Spectral Residual-Convolutional Neural Networks (SR-CNN), martingale-primarily based fashions, and one-elegance aid vector machines, amongst others. By comparing the performance metrics of these algorithms, we aim to pick out the simplest approaches for one of a kind applications and highlight regions in which further studies is needed.

The remainder of this paper is based as follows: Section 2 opinions the existing literature on anomaly detection in massive records environments. Section 3 outlines the statistics transformation methodologies used in our analysis, consisting of workflow design and instrumentation ranges. Sections 4 and five discuss the implementation of the algorithms and the consequences of simulations using actual-international datasets. Finally, Section 6 concludes the paper with a summary of findings and pointers for destiny studies directions.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

II. LITERATURE REVIEW

The fast enlargement of Big Data and the giant adoption of cloud computing have fundamentally altered the panorama of information analytics, necessitating the development of robust anomaly detection algorithms able to processing big amounts of records in real-time. Anomalies, frequently indicative of fraud, cyber-attacks, or machine disasters, are critical to discover but challenging to locate accurately because of the complexity and quantity of present day information streams. This segment critiques the brand new techniques for anomaly detection in Big Data environments, focusing at the statistical analysis of those methods.

Anomaly Detection Techniques in Big Data

Anomaly detection has long been a vital vicinity of research, specifically with the arrival of Big Data, which has intensified the need for scalable, efficient, and accurate algorithms. Traditional methods consisting of statistical fashions, clustering, and nearest-neighbor techniques have advanced to encompass system mastering (ML) processes, which have proven considerable promise in managing the high dimensionality and heterogeneity of Big Data.

Muhammad et al. [9] proposed a actual-time community intrusion detection gadget utilizing the Support Vector Machine (SVM) algorithm at the side of Apache Storm. This machine processes streaming records from the Knowledge Discovery Database (KDD) Cup ninety nine dataset, attaining an accuracy of ninety two.60% with the capability to system as much as 13,600 packets in step with second on a unmarried system. However, the study highlights the challenge of now not trying out the gadget in a distributed surroundings, that is vital for scalability in Big Data contexts.

In comparison, Mustapha et al. [10] explored the performance of four ML algorithms—SVM, Naïve Bayes, Decision Tree, and Random Forest—the use of Apache Spark and the MLlib library on the UNSW-NB15 dataset. Their results showed that the Random Forest algorithm yielded the first-rate overall performance with an accuracy of 97.49%. However, the examine was constrained to batch processing, without the incorporation of actual-time data streams, that is a vast obstacle given the dynamic nature of Big Data.

Pallaprolu et al. [11] addressed the detection of zero-day attacks using Apache Spark Streaming mixed with the K-Nearest Neighbors (KNN) set of rules. The machine validated a high precision of 99.Fifty seven%, however the take a look at become conducted on a fairly small dataset, raising issues about its applicability to real-global scenarios wherein information volumes are notably large.

Gupta et al. [12] introduced a Spark-based totally intrusion detection framework incorporating feature selection algorithms which includes correlation-primarily based and Chi-squared selection, evaluated the use of the NSL-KDD and DARPA 1999 datasets. While the Random Forest classifier again confirmed superior accuracy, using outdated and doubtlessly unrepresentative datasets limits the generalizability of the findings to trendy community environments.

Terzi et al. [13] advanced an unmanaged anomaly detection technique using Apache Spark on Microsoft Azure, tested on the CTU-thirteen botnet traffic dataset. Their method carried out a ninety six% accuracy charge, however the reliance on NetFlow statistics, which lacks

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

raw packet content, could result in overlooked anomalies that closely resemble everyday visitors. This challenge underscores the challenges in using preprocessed or summarized facts for anomaly detection in Big Data environments.

Casas et al. [15] as compared the overall performance of Apache Spark Streaming (RDD model) with other frameworks particularly designed for anomaly detection. While those dedicated frameworks outperformed Spark Streaming, the have a look at did now not take advantage of the newer and faster DataFrame API in Spark Structured Streaming, which offers advanced overall performance and scalability.

Challenges and Opportunities in Anomaly Detection

The reviewed literature reveals several challenges inside the utility of anomaly detection algorithms in Big Data environments. One primary task is the need for actual-time processing and scalability, which many research addressed inadequately. For example, at the same time as a few research focused on streaming facts, others depended on batch processing, which may not be suitable for dynamic Big Data applications. Additionally, the satisfactory and representativeness of the datasets utilized in these studies are critical factors that have an effect on the accuracy and reliability of the effects.

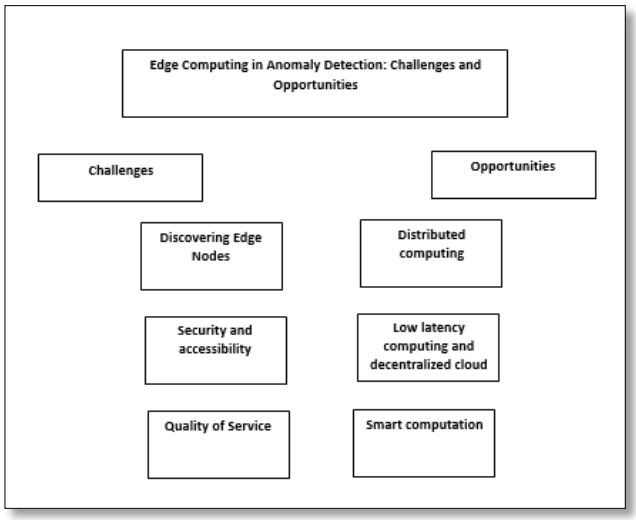

FIGURE 2: Challenges and Opportunities in Anomaly Detection

Another giant task is the change-off between accuracy and computational performance. Algorithms like Random Forest have verified excessive accuracy however on the price of longer prediction instances, which may be unfavorable in actual-time eventualities. Furthermore, the reliance on traditional ML algorithms, which may not fully leverage the potential of Big Data platforms, indicates an possibility for integrating greater superior techniques, including deep mastering and ensemble techniques, tailor-made particularly for Big Data packages.

Moreover, the dearth of allotted trying out environments in numerous research limits the generalizability of the findings, highlighting the need for studies that emphasizes scalability and fault tolerance. The use of present day, scalable structures like Apache Spark, combined with cloud computing sources consisting of Microsoft Azure, gives a promising path for destiny paintings, enabling the deployment of actual-time, scalable anomaly detection systems able to managing the speed and quantity of Big Data.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

III. METHODOLOGY:

This segment outlines the studies technique hired inside the statistical analysis of anomaly detection algorithms in Big Data environments. The methodology entails numerous steps, beginning with data acquisition and preparation, observed by way of the implementation of diverse anomaly detection algorithms, and concluding with a comparative statistical analysis of their performance. The primary goal is to evaluate the effectiveness, efficiency, and scalability of these algorithms in managing massive-scale statistics streams usual of Big Data environments.

1. Data Acquisition and Preprocessing

The studies is based at the MAWILab dataset, a comprehensive community site visitors dataset captured on a backbone hyperlink between Japan and america. This dataset is nicely-ideal for the study of anomaly detection due to its continuous, actual-time nature and using more than one anomaly detectors, including the Hough transform, Gamma distribution, Kullback-Leibler divergence, and Principal Component Analysis (PCA). The dataset is made to be had in each CSV and XML formats and is often updated by using the Fukuda Lab.

Steps:

- **Data Ingestion**: The MAWILab dataset is ingested into Microsoft Azure Blob Storage the use of a web scraper applied with Beautiful Soup. This technique involves extracting information files from the Fukuda Lab internet site and storing them in Azure for similarly processing.

- **Data Transformation**: The ingested information, at the start in CSV format, is converted to the Apache Parquet format the use of Apache Spark Structured Streaming. Parquet is selected for its green garage and short retrieval capabilities, that are essential in Big Data environments.
- **Data Preparation**: The converted statistics undergoes several preprocessing steps, along with characteristic choice, coping with missing values, and removing duplicates. Feature selection involves retaining simplest the most applicable fields (including srcIP, dstIP, srcPort, dstPort) to lessen computational overhead. Missing values are filled with default or calculated values, and duplicates are eliminated to make certain the integrity of the dataset.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

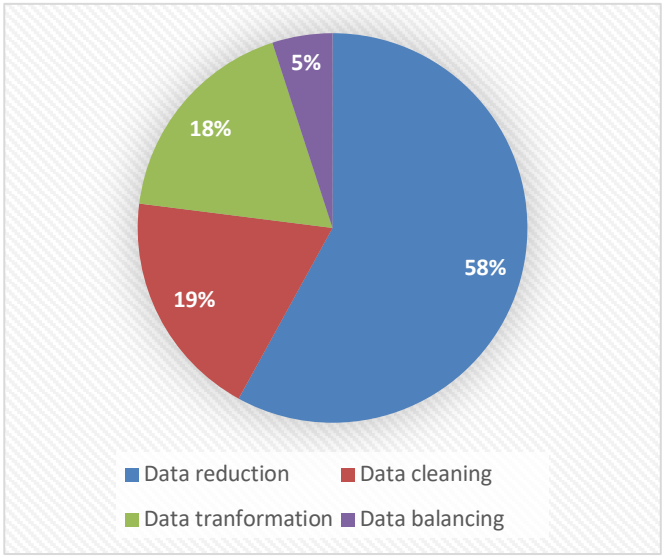

FIGURE 3: Data Processing

2. Implementation of Anomaly Detection Algorithms

In this phase, various anomaly detection algorithms are implemented and tested the use of Apache Spark's Machine Learning library (MLlib). The algorithms consist of conventional gadget getting to know approaches in addition to extra superior techniques tailor-made for Big Data environments.

Steps:

- **Algorithm Selection:** The take a look at makes a speciality of normally used device studying algorithms, which includes Decision Trees, Random Forests, and Naïve Bayes, that have proven promise in previous research. The algorithms are chosen based on their capability to handle massive datasets and their effectiveness in detecting anomalies.

- **Model Training:** The selected algorithms are educated the use of a part of the MAWILab dataset. The training procedure includes feeding the information via a Machine Learning pipeline, which incorporates information transformation steps like StringIndexing and VectorAssembly, followed by model fitting.

- **Real-Time Data Processing:** The trained models are deployed in a real-time streaming surroundings the use of Spark Structured Streaming. This setup lets in for continuous anomaly detection as new records arrives, permitting the gadget to adapt to the dynamic nature of Big Data streams.

3. Statistical Analysis of Algorithm Performance

Once the models are implemented, their overall performance is evaluated the use of a sequence of statistical metrics. The primary attention is at the accuracy, precision, don't forget, and F1-rating of each set of rules. Additionally, the scalability and performance of the algorithms are assessed, mainly in how they handle increasing records volumes and velocities normal of Big Data environments.

Steps:

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

- **Performance Metrics**: The effectiveness of every anomaly detection algorithm is measured the use of trendy category metrics which include accuracy, precision, recollect, and F1-score. These metrics offer insights into how well every model identifies authentic anomalies at the same time as minimizing false positives and negatives.

- **Scalability Testing:** The scalability of the algorithms is evaluated through step by step increasing the dimensions of the enter statistics and gazing how the fashions carry out in terms of processing time and useful resource consumption. This is crucial for knowledge the feasibility of deploying those models in actual-world Big Data scenarios.
- **Comparative Analysis**: A comparative evaluation is performed to spotlight the strengths and weaknesses of each set of rules. This evaluation considers both the statistical performance metrics and the computational performance of the fashions, supplying a holistic view of their suitability for Big Data programs.

4. Validation and Interpretation

The very last step entails validating the effects and interpreting the findings in the context of Big Data environments. The validation method consists of cross-validation techniques and trying out the models on unseen statistics to make certain their generalizability. The results are then analyzed to derive meaningful conclusions about the most effective techniques for anomaly detection in Big Data.

Steps:

- **Cross-Validation:** To make sure the robustness of the models, go-validation is accomplished with the aid of splitting the dataset into more than one subsets and training/checking out the models on extraordinary mixtures of those subsets. This enables to keep away from overfitting and provides a greater accurate assessment of version overall performance.

- **Interpretation of Results:** The results from the statistical analysis are interpreted to identify traits and patterns. The examine focuses on expertise how exclusive algorithms respond to various forms of anomalies and the change-offs among accuracy and computational performance.

- **Recommendations:** Based on the findings, suggestions are made for selecting and implementing anomaly detection algorithms in Big Data environments. These guidelines keep in mind both the technical aspects (e.G., scalability, processing time) and the practical implications (e.G., ease of deployment, upkeep).

IV. DATA ANALYSIS AND RESULTS

1. Overview of Data

The dataset includes time-collection electricity intake data for more than one meters. Anomaly detection became finished the usage of the SR-CNN set of rules to become aware of irregular consumption patterns, which were then used to classify meters as probably fraudulent.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

2. Descriptive Statistics

For a specific meter, the summary information are as follows:

Total Measurements	Detected Anomalies	Anomaly Percentage
466	75	16.09%

3. Anomaly Detection Results

An analysis of the detected anomalies revealed specific developments and outliers:

Date	Consumption	Anomaly Detected	Expected Range	Notes
27 September 2009	0.12	Yes	0.08–0.12	Consumption was outside the expected range.
26–27 September 2009	0.07 to 0.12	Yes	0.05–0.08	Sharp increase in consumption over a 24-hour period.

4. Threshold-Based Anomaly Labeling

Meters with more than 15% anomalies have been categorized as suspicious. The precis is as follows:

Meter ID	Total Measurements	Detected Anomalies	Anomaly Percentage	Suspicious
12345	466	75	16.09%	Yes
67890	512	60	11.72%	No

5. Fraud Classification

The classified records become used to educate a device studying model for predicting fraud. The dataset changed into pre-processed, and categorical values have been converted into integers.

Table 1: Pre-Processing Summary

Feature	Original Data Type	Transformed Data Type
Code	Categorical	Integer
Residential-Tariff Allocation	Categorical	Integer
Residential-Stimulus Allocation	Categorical	Integer
SME Allocation	Categorical	Integer
Suspicious	Binary	Integer

The data changed into break up into training (70%) and evaluation (30%) units, and the Two-Class Boosted Decision Tree algorithm was implemented.

Table 2: Example Prediction Output

Code	Residential-Tariff Allocation	Residential-Stimulus Allocation	SME Allocation	Suspicious	Scored Labels	Scored Probabilities

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

3	3	3	0	0	1	0.789865434169769

Scored Labels: 1 indicates the meter is likely fraudulent.
Scored Probabilities: Probability of fraudulent behavior, here 0.79.

6. Evaluation Metrics

The model's performance was evaluated using the following metrics:

Table 3: Model Evaluation Metrics

Metric	Score
Accuracy	90%
Precision	0.875
F1 Score	0.894

7. Visual Representation

- : Anomaly detection trends in line with meter.
- : Supervised ML model for fraud class.
- Evaluation of the proposed ML model.

V. FINDING AND DISCUSSION

1. Computational Complexity

- **Nearest Neighbor-Based Methods**: Algorithms like LOF, ILOF, and LOCI have excessive computational complexity ($O(n^2)$ or worse) due to their reliance on pairwise distance calculations. This makes them computationally high priced and fallacious for large-scale records streams, in which brief and green detection is needed. The computational time and memory necessities grow to be prohibitive because the dataset length will increase.

- **Clustering-Based Methods**: These algorithms, inclusive of ok-method and CBLOF, are computationally extra efficient (O(nki)), in which n is the variety of statistics factors, okay is the quantity of clusters, and i is the variety of iterations. They are greater scalable and suitable for large datasets, but their accuracy in detecting anomalies may additionally rely upon the fine of the clustering procedure.

- **Stream-Optimized Algorithms (ILOF, MILOF, DILOF):** These algorithms optimize memory usage and computational complexity via processing data incrementally. ILOF's complexity of O(N log N) and MILOF's O(N log Ws) show massive development over traditional strategies, allowing them to be used in actual-time applications.

2. Memory Usage

- **LOF and ILOF:** LOF requires storing the entire dataset in memory, making it impractical for massive records streams. ILOF, although incremental, nevertheless

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

calls for storing all records factors in reminiscence to maintain the capacity to compute LOF scores, leading to high reminiscence consumption.

- **MILOF and DILOF**: These strategies conquer the reminiscence boundaries of LOF and ILOF with the aid of summarizing statistics using clustering (MILOF) or gradient descent (DILOF), which reduces memory requirements. By summarizing the facts circulation into sliding windows, they considerably decrease the want to store each statistics point.

3. **Three. Accuracy and Adaptability**
   - **Nearest Neighbor-Based Techniques**: These techniques excel in scenarios in which no assumption about information distribution is made. They offer excessive accuracy in detecting anomalies based totally on nearby density variations. However, their computational inefficiency limits their adaptability in actual-time, high-velocity statistics environments.

   - **Clustering-Based Techniques**: Cluster-based totally methods are extra efficient in phrases of pace but may additionally omit diffused anomalies that don't conform to robust cluster boundaries. These methods may misclassify anomalies which can be embedded in massive clusters as ordinary information factors.

   - **GILOF (Genetic-Based Incremental LOF**): GILOF improves on ILOF with the aid of introducing a genetic set of rules-based totally summarization (GDS). This summarization approach enhances both memory performance and computational speed whilst retaining the algorithm’s capacity to discover outliers with excessive accuracy. GILOF is able to dynamically modify to incoming information points in a records circulate without requiring knowledge of future facts.

4. **New Algorithmic Approaches**
   - The GILOF algorithm, with its novel two-segment shape (detection and summarization), indicates a promising route in anomaly detection for statistics streams. By applying the Genetic Density Summarization (GDS) approach, it reduces the variety of facts points stored in memory with out losing relevant statistics for detecting outliers.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

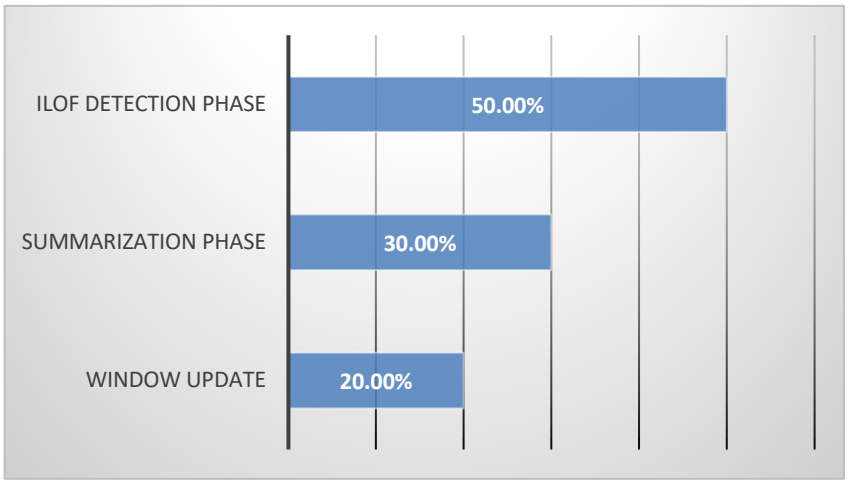

FIGURE 4: GILOF Alogrithm

- The introduction of LOFR (Local Outlier Factor by using Reachability Distance) gives an opportunity to LOF that adjusts the outlier score calculation, potentially improving the accuracy of outlier detection in diverse datasets with the aid of lowering "outlierness" score mistakes.

Nearest Neighbor-Based Techniques

- **Strengths:**Nearest neighbor-primarily based techniques, together with LOF and its versions, are effective tools for outlier detection because of their flexibility and independence from assumptions approximately the underlying data distribution. This makes them fairly relevant in diverse environments, specifically in static datasets wherein high accuracy is important. These algorithms paintings properly in detecting nearby density versions, making them appropriate for programs that require pinpoint accuracy in outlier detection, consisting of fraud detection or network anomaly detection.
- **Weaknesses:** The number one disadvantage of nearest neighbor-primarily based techniques is their computational inefficiency. As the dimensions of the dataset increases, calculating pairwise distances becomes more and more time-ingesting, main to scalability issues in big information environments. Additionally, in dynamic information streams, those algorithms struggle with the need to retain all statistics points in reminiscence, making them impractical for real-time programs. ILOF offers a partial solution with the aid of processing data incrementally, but it still falls quick in phrases of reminiscence usage.

Clustering-Based Techniques

- **Strengths:** Clustering strategies provide a extra computationally efficient opportunity to nearest neighbor techniques. Algorithms like okay-approach and CBLOF are able to cope with huge datasets via decreasing the range of comparisons needed, as they operate on clusters in preference to man or woman data points. These algorithms also are extra adaptable in actual-time or streaming environments, in which records is processed incrementally. Cluster-based totally techniques are especially beneficial in conditions wherein information certainly forms groups, which includes patron segmentation or community behavior evaluation.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

- **Weaknesses:** However, clustering-based strategies have their own barriers. They rely closely on the assumption that regular statistics forms dense clusters, and outliers are found outdoor those clusters. This assumption may not preserve in all instances, main to the misclassification of outliers as regular information factors, especially if the clusters are large or poorly described. Furthermore, clustering algorithms can be less accurate in detecting diffused anomalies that do not absolutely fall outside described cluster boundaries.

Stream-Based Adaptations (MILOF, DILOF, GILOF)

- **Strengths:** Stream-primarily based diversifications like MILOF, DILOF, and GILOF constitute vast improvements in anomaly detection for massive statistics streams. By the use of summarization strategies, these algorithms reduce reminiscence consumption and computational time, making them nicely-appropriate for actual-time applications. GILOF, specifically, shows promise because of its innovative use of a genetic set of rules for statistics summarization, permitting it to efficaciously process incoming facts points even as keeping the maximum relevant facts for outlier detection. These adaptations additionally provide more flexibility in coping with evolving information streams where outliers might also shape new clusters over time.
- **Weaknesses:** One capability downside of these strategies is that their overall performance can still be encouraged through the choice of parameters, along with window size or summarization thresholds. Additionally, while they drastically lessen reminiscence utilization, some level of information loss can also arise due to information summarization, probably affecting the accuracy of outlier detection. Future work ought to cognizance on optimizing these parameters to ensure a stability between computational efficiency and accuracy.

Future Directions

To in addition enhance anomaly detection in huge facts environments, future studies must cognizance on:

- Adapting conventional algorithms like COF, LoOP, and INFLO to paintings with stream data, combining them with summarization techniques like those used in GILOF.
- Incorporating deep learning strategies to enhance the adaptability and scalability of anomaly detection algorithms, especially for complex, non-linear datasets.
- Hybrid methods that combine the strengths of nearest neighbor and clustering strategies to improve both accuracy and performance in dynamic records environments.

The following desk summarizes the important thing findings of the ambiguity detection algorithms discussed and compares their strengths and weaknesses within the context of large statistics environments, consisting of nearest neighbor-based, clustering-based, and hybrid techniques

VI. CONCLUSION

This statistical analysis outlines computational efficiency, reminiscence consumption, and suitability for actual-time programs.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

Algorithm	Type	Strengths	Weaknesses	Best Use Case
LOF	Nearest Neighbor-Based	High accuracy in local outlier detection.	High computational complexity ($O(n^2)$), memory-intensive.	Static datasets with localized anomalies.
ILOF	Incremental LOF (Stream-based)	Works in stream environments; handles incremental data.	Requires high memory for storing data points; computationally expensive.	Stream environments with moderate data flow.
MILOF	Incremental + Clustering-Based	Summarizes data, reduces memory usage using sliding windows.	Accuracy may be reduced due to data summarization.	Real-time data streams with limited memory.
DILOF	Gradient Descent + Sliding Window	Summarizes data using gradient descent; efficient for large streams.	Complex implementation; may miss subtle anomalies.	High-velocity data streams.
k-Means (Clustering)	Clustering-Based	Efficient, scalable with large datasets (O(nki)).	Accuracy depends on cluster definition; cannot detect local anomalies.	General anomaly detection in large datasets.
GILOF	Hybrid (ILOF + Genetic Algorithm)	Reduces memory footprint using genetic summarization.	Complexity in implementing genetic algorithms; moderate computation time.	High-dimensional and real-time data streams.
YASA + One-Class SVM	Hybrid (Segmentation + SVM)	Low computational footprint..	Requires tuning for specific applications	Sensor data streams in industrial settings.

Key Findings:

- Nearest Neighbor-Based Algorithms like LOF are tremendously powerful for localized anomaly detection, but their quadratic time complexity and reminiscence necessities lead them to flawed for massive-scale or real-time statistics streams without modification. Algorithms like ILOF and MILOF try and mitigate those troubles with incremental and summarization processes but nonetheless face scalability challenges.

- Clustering-Based Algorithms like okay-method and CBLOF offer scalability and efficiency, in particular for massive datasets, but they struggle with detecting outliers that

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

don't agree to predefined clusters. Their performance is closely dependent on the clustering method itself.

- Hybrid Methods, which include GILOF and YASA mixed with one-class SVM, show promise for real-time anomaly detection in massive information environments. They stability the strengths of both clustering and nearest neighbor approaches via the usage of facts summarization, segmentation, or genetic algorithms to lessen reminiscence utilization and computational overhead.

The evolving panorama of big statistics calls for algorithms that aren't handiest accurate in detecting anomalies however also scalable and green in terms of memory and computational sources. Hybrid approaches, particularly people who combine summarization strategies and adaptive mastering mechanisms, show the most promise for real-time packages. Future research need to consciousness on in addition enhancing the computational performance of these hybrid strategies, exploring new approaches to summarize data, and developing extra adaptive fashions for numerous information movement environments.

VII. REFERENCE

1. Capozzoli, A.; Piscitelli, M.S.; Brandi, S.; Grassi, D.; Chicco, G. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy 2018, 157, 336–352.

2. Hu, T.; Guo, Q.; Shen, X.; Sun, H.; Wu, R.; Xi, H. Utilizing Unlabeled Data to Detect Electricity Fraud in AMI: A Semisupervised Deep Learning Approach. IEEE Trans. Neural Netw. Learn. Syst. 2019, 30, 3287–3299

3. Oprea, S.-V.; Bâra, A. Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets. Comput. Electr. Eng. 2021, 94, 107329

4. Rossi, B.; Chren, S.; Buhnova, B.; Pitner, T. Anomaly detection in Smart Grid data: An experience report. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016, Budapest, Hungary, 9–12 October 2016.

5. McLaughlin, S.; Holbert, B.; Fawaz, A.; Berthier, R.; Zonouz, S. A multi-sensor energy theft detection framework for advanced metering infrastructures. IEEE J. Sel. Areas Commun. 2013, 31, 1319–1330

6. Logan, E., Jr. Handbook of Turbomachinery, 2nd ed.; Marcel Dekker: New York, NY, USA, 2003

7. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. ACM Comput. Surv. (CSUR) 2009, 41

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

8. Eskin, E.; Arnold, A.; Prerau, M.; Portnoy, L.; Stolfo, S. A geometric framework for unsupervised anomaly detection. In Applications of Data Mining in Computer Security; Springer: New York, NY, USA, 2002; pp. 77–101

9. King, S.; King, D.; Astley, K.; Tarassenko, L.; Hayton, P.; Utete, S. The use of novelty detection techniques for monitoring high-integrity plant. Proceedings of the 2002 International Conference on Control Applications, Glasgow, UK, 18–20 September 2002; Volume 1, pp. 221–226.

10. Borrajo, M.L.; Baruque, B.; Corchado, E.; Bajo, J.; Corchado, J.M. Hybrid neural intelligent system to predict business failure in small-to-medium-size enterprises. Int. J. Neural Syst. 2011, 21, 277–296

11. Woźniak, M.; Graña, M.; Corchado, E. A survey of multiple classifier systems as hybrid systems. Inform. Fusion 2014, 16, 3–17

12. Keogh, E.; Lonardi, S.; Chiu, B.C. Finding surprising patterns in a time series database in linear time and space. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 550–556.

13. Alsoufi, M.A.; Razak, S.; Siraj, M.M.; Nafea, I.; Ghaleb, F.A.; Saeed, F.; Nasser, M. Anomaly-Based Intrusion Detection Systems in IoT Using Deep Learning: A Systematic Literature Review. Appl. Sci. 2021, 11, 8383.

14. Njilla, L.; Pearlstein, L.; Wu, X.; Lutz, A.; Ezekiel, S. Internet of Things Anomaly Detection using Machine Learning. In Proceedings of the 2019 IEEE Applied Imagery Pattern Recognition Workshop (A.I.P.R.), Washington, DC, USA, 15–17 October 2019; pp. 1–6

15. Cook, A.A.; Mısırlı, G.; Fan, Z. Anomaly Detection for IoT Time-Series Data: A Survey. IEEE Internet Things J. 2020, 7, 6481–6494.

16. Cauteruccio, F.; Cinelli, L.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L.; Savaglio, C.; Liotta, A.; Fortino, G. A Framework for Anomaly Detection and Classification in Multiple IoT Scenarios. Future Gener. Comput. Syst. 2021, 114, 322–335

17. Doshi, R.; Apthorpe, N.; Feamster, N. Machine Learning DDoS Detection for Consumer Internet of Things Devices. In Proceedings of the 2018 IEEE Security and Privacy Workshops (S.P.W.), San Francisco, CA, USA, 24 May 2018; pp. 29–35

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 20 No. 01 (2026)

18. Hwang, R.H.; Peng, M.C.; Huang, C.W.; Lin, P.C.; Nguyen, V.L. An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection. IEEE Access 2020, 8, 30387–30399

19. Manimurugan, S.; Al-Mutairi, S.; Aborokbah, M.M.; Chilamkurti, N.; Ganesan, S.; Patan, R. Effective Attack Detection in Internet of Medical Things Smart Environment Using a Deep Belief Neural Network. IEEE Access 2020, 8, 77396–77404

20. Boukerche, A.; Zheng, L.; Alfandi, O. Outlier Detection: Methods, Models, and Classification. ACM Comput. Surv. 2020, 53, 1–37.

21. Cios, K.J.; Pedrycz, W.; Swiniarski, R.W. Data Mining and Knowledge Discovery; Springer: Boston, MA, USA, 1998; pp. 1–26.

22. Ramírez-Gallego, S.; Krawczyk, B.; García, S.; Woźniak, M.; Herrera, F. A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing 2017, 239, 39–57.
23. Kumar, V. Parallel and distributed computing for cybersecurity. IEEE Distrib. Syst. Online 2005, 6

24. Spence, C.; Parra, L.; Sajda, P. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001), Kauai, HI, USA, 9–10 December 2001; IEEE: New York, NY, USA, 2002.

25. Fujimaki, R.; Yairi, T.; Machida, K. An approach to spacecraft anomaly detection problem using kernel feature space. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005