An Overview of Outlier Detection Methods

Dr. Maciej Celi Ski

ABSTRACT

One of the first important steps in achieving informed data analysis is detection of outliers. Even in cases where the final values are often considered to be incorrect calculations or noise, they can still provide very important information in some cases. Therefore, it is very important to detect them before modeling and analysis. In this paper, we present a structured and comprehensive review of residual detection research. There are many different methods, hence the purpose of this article is to help the novice researcher to formulate his ideas and gain an easier understanding of the various lines of research in which research has been conducted on this topic.

Keywords: machine learning, outlier, outlier detection.

Classification: LCC Code: QA76.9.D32

Language: English

Great Britain Journals Press

LJP Copyright ID: 392934
Print ISSN: 2631-8474
Online ISSN: 2631-8482

London Journal of Engineering Research

Volume 24 | Issue 2 | Compilation 1.0

An Overview of Outlier Detection Methods

Dr. Maciej Celi ski

ABSTRACT

One of the first important steps in achieving informed data analysis is detection of outliers. Even in cases where the final values are often considered to be incorrect calculations or noise, they can still provide very important information in some cases. Therefore, it is very important to detect them before modeling and analysis. In this paper, we present a structured and comprehensive review of residual detection research. There are many different methods, hence the purpose of this article is to help the novice researcher to formulate his ideas and gain an easier understanding of the various lines of research in which research has been conducted on this topic.

Keywords: machine learning, outlier, outlier detection.

I. INTRODUCTION

In the era of advanced technology and rapidly evolving digital landscapes, the need for robust anomaly detection techniques has become paramount across various domains. Among these techniques, Isolation Forests (IF) have emerged as a versatile and powerful tool with applications spanning cybersecurity, fault monitoring, big data analysis, fraud detection, and more. This collection of texts delves into the myriad ways Isolation Forests have been harnessed to address specific challenges and enhance various aspects of modern life.

As we delve into the texts that follow, we will uncover the multifaceted role that Isolation Forests play in ensuring the security of computer systems, identifying fraudulent activities, monitoring the health of mechanical systems, and analyzing vast datasets for hidden insights. The texts also highlight the adaptability and effectiveness of Isolation Forests in addressing the evolving demands of these diverse fields.

Our journey begins by exploring the significant role Isolation Forests play in safeguarding the digital realm. We delve into their application in Intrusion Detection Systems (IDS), where they excel in identifying unauthorized access and novel cyber threats, bolstering the security of sensitive data and computer systems.

Next, we venture into the domain of fault monitoring and system diagnostics, where Isolation Forests offer invaluable capabilities in detecting anomalies and irregularities across various mechanical and technical systems. Their rapid anomaly detection and response prove indispensable in industries where system failures can lead to substantial financial losses.

In the context of big data analysis, Isolation Forests provide a lens through which we can uncover hidden patterns and correlations within extensive datasets. This facilitates more informed decision-making in the business world, offering opportunities for comprehensive data exploration.

Our exploration then takes us into the realm of fraud and abuse detection, where Isolation Forests excel in identifying counterfeit reviews, misuse, and fraudulent activities, particularly in the context of customer service. Their role extends not only to consumer protection but also to upholding product and service integrity.

Lastly, we reflect on the broader significance of Isolation Forests, underlining their efficacy and potential in diverse fields. This collection of texts serves as a testament to the ever-evolving nature of Isolation Forests, opening new horizons for researchers and professionals alike.

Together, these texts paint a comprehensive picture of the significance and multifaceted applications of Isolation Forests, offering insights into how this powerful anomaly detection technique continues to shape and protect our digital world.

What is already Known about this Topic:

- Anomaly detection helps in data analysis.
- Anomaly detection saves time needed for subsequent analysis of this data, which is essential.
- Anomaly detection is widely applicable to various types of data.

What this Paper Adds:

- Allows us to notice the diversity of applications of anomaly detection methods.
- Constitutes a valuable collection of information about current methods of detecting anomalies.
- Is particularly important in the search for knowledge in the field of applications of fuzzy methods.

Implications for Practice and/or Policy:

- Is valuable for those wishing to conduct a comprehensive review of the current literature on this topic.
- Allows you to understand current trends in the use of fuzzy methods.
- It allows you to notice the potential of using fuzzy methods in various aspects of human life and its environment.

The purpose of this paper is to provide a comprehensive review of the literature on Outlier Detection and the current outlier detection techniques used in various areas. The methodology for this review presented in this paper is as follows:

- We select articles from 2018-2023 (94 works) from leading online scientific databases such as MDPI, Science Direct, IEEE Explorer, etc.).
- In addition, the work includes some required and particularly important and original articles from before 2018, in the context of the following keywords: outlier detection, anomaly detection.
- In our work we choose the most relevant and up-to-date articles that focus on the subject of our analyses.

The texts can be categorized into several thematic areas related primarily to the applications of Isolation Forests in various domains. These areas include:

1. *Cybersecurity:* Texts in this category focus on the role of Isolation Forests in Intrusion Detection Systems (IDS) and the identification of unauthorized activities and cyberattacks in the field of computer security.
2. *Fault Monitoring and State Diagnostics:* This category contains texts that describe the use of Isolation Forests for monitoring the state of devices and detecting faults in various mechanical and technical systems.
3. *Big Data Analysis:* Texts in this category concentrate on the applications of Isolation Forests in the analysis of large datasets, enabling the identification of unusual patterns and hidden dependencies in data.
4. *Fraud and Abuse Detection:* This category encompasses texts discussing the role of Isolation Forests in detecting fake product reviews, abuse, and fraud, particularly in the context of customer service.
5. *General Category:* In this category, there is a general discussion about the advantages and potential of Isolation Forests in various domains, emphasizing their effectiveness and significance. It's worth noting that the texts in each of these categories address different applications of Isolation Forests and their impact on specific fields.

II. SECTION TWO

2.1 Anomaly Detection Methodology

In the texts below, several important anomaly detection methods, widely utilized across various domains, are described. Here are the most significant of these methods.

Isolation Forests (IF): This is a pivotal method that appears in all the texts. Isolation Forests are an algorithm based on decision trees, exceptionally efficient in identifying outlier observations. This algorithm isolates anomalous points by employing a straightforward logic of partitioning data into smaller subgroups.

Local Outlier Factor (LOF): Another popular method for outlier detection. LOF assesses how isolated each data point is compared to its nearest neighbors. Points with low LOF scores are considered normal, while those with high LOF scores are deemed outliers.

One-Class Support Vector Machine (One-Class SVM): This approach utilizes Support Vector Machines (SVM) for data classification into two categories: normal and anomalies. The One-Class SVM establishes a boundary around the normal data and endeavors to ascertain whether data points situated beyond this boundary qualify as anomalies.

Histogram-Based Outlier Detection (HBOS): This method is based on creating histograms from data and utilizes differences in the distribution of normal and outlier data to identify anomalies.

Decision Trees (DT): Decision trees are used in some texts as part of more complex anomaly detection techniques, such as algorithm combinations.

Logistic Regression (LR): Logistic regression is also used as part of more advanced anomaly detection methods, often for classifying data as normal or outliers.

K-Means: In some texts, K-Means is employed as a data clustering algorithm, which can aid in identifying outlier groups or clusters.

Radial Basis Function Networks (RBF): RBF is used as a model to classify data as normal or outliers based on their distance from the cluster center.

Multi-Layer Perceptron (MLP): MLP is used in the context of data classification and is one of the neural network models employed for outlier detection.

These methods are applied in various contexts, from cybersecurity to big data analysis and fraud detection. Each of these techniques possesses unique advantages and applications, enabling researchers to choose the most suitable method depending on the specific problem and type of data they are currently working with.

Figure 1 below presents the percentage of pseudocode, source code or the lack of these two elements provided by researchers in the works discussed.

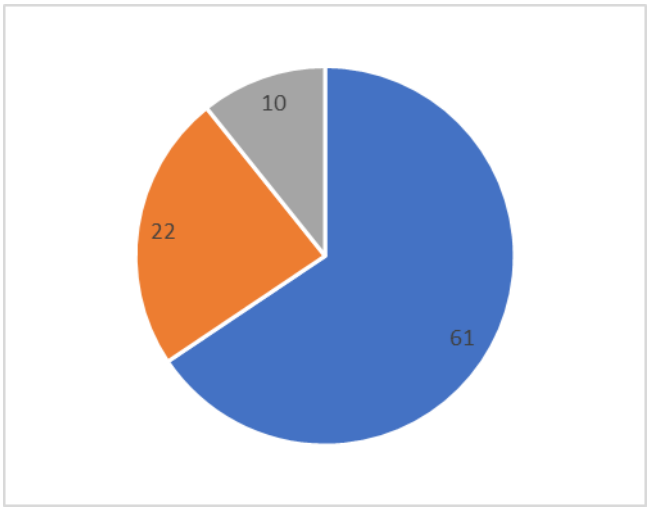

Fig. 1: Gray Color - Source Code Provided, Yellow Color - Pseudocode Included in the Work, Blue - No Source Code or Pseudocode

III. SECTION HEAD

3.1 Outlier Detecion Literature Review

In [1] the paper, the authors introduce an approach to identify and diagnose irregularities through a reduction in their dimensionality, variable selection, and the application of an isolation forest algorithm. The researchers focused their attention on incorporating isolated variables, which might go unnoticed when employing traditional techniques like Principal Component Analysis (PCA). Within the article, the authors delineate a technique for autonomously identifying and diagnosing anomalies using

historical OES data, with a notable emphasis on addressing both dimensionality and aspects associated with their interrelation.

In [2] this study, the researchers introduce ideas for a system designed to identify credit card fraud in real-time, with a strong focus on maximizing its accuracy by capturing as many new transactions as possible. Furthermore, in their research, the authors conduct a comparative analysis of various techniques for detecting credit card fraud, such as LOF and SVM methods. The utilization of Isolation Forest (IF) in detecting fraudulent credit card transactions, as demonstrated through a series of experiments, has proven to be highly efficient in anomaly detection. It's noteworthy that the researchers explicitly state their intention to implement an architecture capable of real-time detection of fraudulent credit card transactions.

In [3] to a varying extent, in the given study, researchers employ techniques such as Isolation Forest (IF) and Long Short-Term Memory (LSTM) to identify anomalies within the dataset.

Subsequently, they utilize neural networks to forecast and rectify irregular data points. In this research, scientists conducted experiments using datasets derived from electromagnetic systems, with the primary objective of refining these datasets for future operational and maintenance purposes. This preparatory phase enables subsequent experiments utilizing the previously mentioned IF and LSTM methods. The experimental outcomes, as presented by the researchers, indicate that the proprietary software for data cleansing related to power supply operations and maintenance, which is based on CiF-AL and incorporates the use of IF and LSTM techniques, has undergone significant optimization, both in terms of anomaly detection and the accuracy of their correlation.

Table 1: Comparison of Corrected RMSE Index for Prediction

Algorithm	RSME
Model based on time series analysis in	0.779
Traditional LSTM neural network	0.292
CiF-AL	0.263

In [4] this study, the researchers introduce an alternative approach to enhancing the security of avionics systems, specifically the ADS-B system, a critical component of most aircraft responsible for ensuring air navigation safety. The primary focus of their work lies in proposing an innovative security solution designed to detect abnormal ADS-B messages. This solution primarily aims to identify falsified or tampered ADS-B data sent either by malicious third parties (attackers) or compromised aircraft. The researchers present an algorithm rooted in Long Short-Term Memory (LSTM) to model an aircraft's flight path by analyzing sequences of crucial data and messages within the ADS-B system. Based on this analysis, the aircraft assesses and evaluates ADS-B messages to spot deviations from the correct flight path. To validate their approach, the researchers conduct a series of experiments using thirteen diverse datasets of flight paths, each containing various types of anomalies. The final result of their research entails a performance comparison between the LSTM algorithm and other frequently employed methods for identifying unusual data patterns within datasets, such as GMM-HMM, DBSTREAM, SVM, LOF, and IF. The research results outlined in the study demonstrate that the proposed approach successfully identifies all injection attacks, with a minor increase in the average false-positive rate by 4.5%. This outcome provides evidence that the effectiveness of the presented algorithm surpasses that of the other methods under examination.

In [5] this work, the authors delve into the application of the Isolation Forest (IF) method, particularly focusing on the critical aspects of selecting the appropriate attribute and determining the optimal split point within that attribute when constructing an IF model. To address these challenges, the authors propose an innovative computational structure, which, in their view, excels at identifying the most

distinguishable attributes and pinpointing the optimal division points for them. The researchers highlight a key limitation of the IF method, namely, its reliance on a binary tree structure, which comes with constraints related to tree depth, leaf nodes, or external nodes. These aspects are pivotal in assessing whether a given instance is normal or anomalous. In this work, the authors present a novel solution aimed at streamlining the process of building a binary tree within the framework of the IF method.

Algorithm GradFindSpli

Input:

X^k: sorted values of the *k*th attribute;

Output:

bestX: the attribute value having the largest value of the separability index;

bestStep: the largest values of the separability index;

1. Initiate step as ceil(Xk|*0.1)
2. Let bestStep be step(Xk,X1) and bestX be (x1+x2)/2 and i=0
3. *while* i<|X^k| do
4. Set i=i+step
5. Set currStep = step(X^k, x_1)
6. If currStep > bestStep *then*
7. Set bestStep = currStep
8. Sete bestX = (x_1+x_{i+1})/2
9. *end if*
10. Update step by following the formulas (6) and (7)
11. end while
12. Return best X and bestStep

In [6] the study discusses the authors' adaptation of the Isolation Forest (IF) method for the detection of network anomalies originating from the network management system. The authors emphasize the limited availability of techniques for identifying network anomalies and propose modifications to the IF method to address this issue. The adapted IF method has been tailored to eliminate recurring patterns, detect peaks and troughs, and fine-tune various metrics, thus increasing its applicability. By incorporating these IF algorithms, the research methodology presented enables the rapid identification of anomalies within extensive datasets containing multidimensional network management data.

Additionally, the paper presents experimental results based on six simulated datasets, showcasing the adaptability of the proposed method to various types of data derived from network performance measurements. Moreover, the authors conduct a comparative assessment of the suitability of five distinct feature extractors, evaluating them based on their characteristics and Area Under the Curve (AUC) scores. Finally, the researchers demonstrate the real-world application of the method in a computer network scenario, utilizing network traffic data for evaluation.

Table 2: Results of Anomaly Detection on Actual Data. TPR Refers to True Positive Ratio, and Fpr Refers to False Positive Ratio

Training Data	Time Range	Anomaly Ratio	TPR	FPR
T1	20171118-20171127	4.4%	97.61%	1.48
T2	20171112-20171121	1.0%	94.42%	0.22%
T3	20171108-20171117	0.0%	98.21%	0.66%

In [7] the researchers primary focus lay in the identification of anomalies within network traffic. While there exist numerous studies in this domain, the novelty of their work lies in the parallel algorithm they introduce, which combines the Isolation Forest (IF) method with Spark. This amalgamation greatly expands the scope of anomaly detection within computer networks, notably by enabling the processing of vast volumes of data. During the computational procedures, the authors distribute tasks across multiple IF and Spark computing nodes, thereby harnessing the potential of handling Big Data. The experimental findings, derived from real-world datasets sourced from network traffic, demonstrate the algorithm's efficiency and its effectiveness in detecting anomalies within network traffic data.

Table 3: AUC and Accuracy of iForest and SPIF

Algorithm	AUC	Accuracy
iForest	0.8831	8,872
SPIF	0.8927	87.144

In [8] their research, the scientists grapple with the challenge of applying the Isolation Forest (IF) method to multidimensional data, where it exhibits instability despite its low time complexity and effective anomaly detection capabilities. They stress that this method is vulnerable to noise in the data, prompting them to develop an enhanced anomaly detection approach rooted in the IF method, leveraging dimensional entropy. In this new method they present, the researchers incorporate dimensional entropy as a crucial component in both the selection of the isolation attribute and the determination of the isolation point. Within this method, the researchers outline a process for choosing isolation attributes and points based on the information entropy associated with each attribute within random samples. The outcomes of their research experiments demonstrate that this presented approach operates significantly faster and offers.

Algorithm 2: EiTree(D_i, e, l, bin, pathL, α)

Inputs: D_i, e, l, bin, pathL, α

Output: An EiTree

1. If e≥l or $|D_i|$≤1 then
2. return exNode
3. Else
4. Calculate {ent_1, ent_2,..., ent_m}
5. End if;
6. If 0< ent_1< α * ent_{max} then record the attribute i into {ch_1,ch_2,..., ch_{chnum}};
7. If {ch_1,ch_2,..., ch_{chnum}} is not empty then
8. Isolation strategy 2 → D_{left} and D_{right}:
9. return inNode {

   pathL = pathL * $pathL_1$

   left → EiTree(D_{left}):

   right → EiTree(D_{right}):}
10. end if.

In [9] the researchers address the limitations of current mobile device user authentication methods, particularly in terms of achieving the necessary security levels. They propose the utilization of an Isolation Forest (IF)-based authentication model as a prototype for enhancing security.Their research centers on behavioral biometrics, specifically continuous user authentication based on the user's interaction with the device's touchscreen. Notably, the presented system's effectiveness in security operations is noteworthy. In the case of continuous authentication, the system was capable of detecting an unauthorized user after an average of seven actions. This achievement holds great unauthorized access.

In [10] to a certain extent, the researchers in this article highlight the connection between variations in vital signs as an indicator of potential health issues requiring attention. Consequently, they suggest that anomalies can be characterized by deviations from typical vital sign patterns, which may signal the necessity for immediate intervention. In their research, the authors introduce a hybrid method combining K-Means clustering and the Isolation Forest (IF) for detecting anomalies. They compare the outcomes of their research with those obtained from the standard IF algorithm and a hybrid technique that combines K-Means clustering with IF. The research findings presented by the authors indicate that the effectiveness of the proposed hybrid technique varies depending on the dataset, yielding either lower or higher error rates.

In [11] their study, the researchers introduce an Isolation Forest (IF) approach for low-fidelity event detection and monitoring, using real-time data from phasor measurement units (PMUs). Their method is capable of identifying early events even within data of subpar quality, and it boasts significantly reduced computational complexity compared to traditional machine learning techniques. Moreover, the suggested method exhibits enhanced reliability when compared to alternative techniques used for detecting and monitoring low-quality events, such as Principal Component Analysis (PCA) or Singular Value Decomposition (SVD).

In [12] the researchers focus their investigation on the Isolation Forest (IF) method, specifically delving into its IForestASD implementation. They employ this method within the scikit-multiflow framework, an open-source machine learning platform designed for data stream handling. The scientists then proceed to conduct experiments using three real datasets to evaluate the predictive performance and resource consumption of the IForestASD method. Additionally, they compare its performance with another data stream anomaly detection method referred to as half-space trees. Based on their research findings, the authors conclude that while IF demonstrates accuracy in traditional batch settings, it may not be well-suited for efficient data stream processing.

In [13] the researchers present an original approach to anomaly detection by incorporating the Isolation Forest (IF) method into streaming data analysis using a sliding window mechanism. They have termed this approach the IForestASD algorithm and have implemented it within Scikit-multiflow, a versatile machine learning platform tailored for handling data with multiple outputs and labels. This innovative solution holds promise for effectively detecting anomalies within streaming platforms, thus offering a substantial contribution to the field of anomaly detection in both industrial and research settings.

In [14] this study, the authors introduce a technique for identifying damage in hydroelectric generators using the Isolation Forest (IF) method. However, they highlight a significant challenge: failures in hydroelectric generators are infrequent, making it difficult to gather failure data. Therefore, the authors stress the importance of automatically constructing models exclusively with normal data and subsequently performing error detection. The researchers evaluate the effectiveness of IF for feature data by comparing it with an error detection method based on multivariate statistical process control (MSPC).

The experimental results validate that the damage detection approach proposed by the researchers, utilizing IF, outperforms the MSPC-based method in terms of accuracy. It's noteworthy that the researchers assess the correctness and functionality of their approach using the Wilcoxon signed-rank test.

In [15] this presented research, the authors introduce a novel method termed "Mondrian Forest Isolation" or iMondrian Forest. The primary objective of this method is to detect anomalies both in batch processing and real-time streaming data. The researchers place particular emphasis on combining two established techniques: Isolation Forest (IF) and Mondrian Forest. They clarify that these methods are already recognized for their efficacy in batch anomaly detection and online random forest applications. The outcome of the researchers' efforts is a fresh data representation format capable of accommodating online data streaming while simultaneously serving as an anomaly detection tool. The conducted experiments reveal that iMondrian Forest outperforms the IF method significantly in batch processing scenarios and either surpasses or matches the performance of other anomaly detection methods designed for both batch and real-time processing. The method proposed by the researchers represents an innovative hybridization of Isolation Forests and Mondrian Forests, harnessing the strengths of both techniques. This combination leverages the isolation-based anomaly detection features and the advantages of online ensemble learning, as underscored by the research findings.

In [16] the researchers address the specific challenge of analyzing datasets that encompass various data types, including temporal, spatial, image, and categorical data. The presence of multiple data types often leads to the emergence of numerous outliers. To address this challenge, the authors introduce modifications to the Isolation Forest (IF) method, which they refer to as the Fuzzy Set-Based Isolation Forest (FSBIF).

The modification introduced by the researchers is based on the concept of measuring the distance from the center of the analyzed cluster during the construction of the binary search tree. Test results indicate that when utilizing the FSBIF approach, all of these data points achieve the highest degree of isolation compared to the conventional IF method, resulting in reduced degrees of isolation.

The research findings presented by the authors suggest that the method proposed in their work outperforms other existing methods in terms of both performance and effectiveness.

Table 4: Execution times of if and FSBIF

Method	Artificial dataset	Taxi Database
IF	11,023 pp	55,997 pp
Fuzzy set-based IF	1,152 pp	7,298 pp

In [17] he authors emphasize the challenges and expenses associated with obtaining damaged industrial samples, which are relatively scarce in practical applications. The researchers address the issue of dealing with imbalanced sample datasets, a common concern in industrial settings.

To tackle this problem, they employ a fusion anomaly detection method that combines Artificial Neural Networks (ANN) with the Isolation Forest (iForest) algorithm. The researchers' primary focus is on expanding the dataset through a technique known as upsampling, which involves increasing the number of samples to maintain a balanced representation of different sample types. In the subsequent step, the researchers utilize Empirical Mode Decomposition (EMD) to extract specific energy features from each sample, aiming to capture the inherent signal characteristics more effectively. In the final stage of their research, the authors extract latent features, which are then input into the Isolation Forest algorithm to detect anomalies within the industrial samples. This comprehensive approach enhances the capacity to identify unusual patterns and anomalies in the data, particularly in scenarios where damaged samples are limited in availability.

Table 5: Details of the Experimental Method and Accuracy

Method	Accuracy (%)
SVM	83.86
Light GBM	93.54
One ClassSVM	80.06
LOF	54.56
Isolation Forest	92.35
ANN	95.6
Proposed method	97.46

The researchers conducted tests on the presented method along with various traditional methods using CWRU bearing datasets. The results demonstrated by the researchers indicate that the method they introduce outperforms the traditional approaches by a significant margin in terms of performance.

In [18] the authors highlight the urgent need for early prediction of diabetes, given that diabetes is currently one of the leading causes of global mortality. Late detection and untreated diabetes can result in severe health complications. To address this challenge, the researchers introduce a prognostic model designed to enable the early prediction of type 2 diabetes in individuals. In their model, the researchers employ the IFo (Isolation Forest-based outlier detection) method and the synthetic minority pooling technique with sequential sampling. These methods are utilized to identify and remove outlier data points while also balancing the data distribution. For learning and predicting type 2 diabetes at an early stage, the researchers utilize ensemble classifiers. To assess the model's reliability, the authors utilize three publicly available datasets and compare the performance of their proposed model with other established models. They employ a 10-fold cross-validation method and evaluate four performance metrics: precision, recall, F1-score, and accuracy.

The research findings indicate that their model significantly outperformed other models, achieving accuracy rates of 93.18%, 98.87%, and 96.09% for datasets I, II, and III, respectively. This suggests the potential effectiveness of their model in early diabetes prediction.

In [19] the researchers propose a multimodal approach for detecting anomalies in embedded systems, especially in situations where there exists a correlation between energy consumption over time and memory access. They utilize time series data that encompass diverse power consumption patterns, L2 cache availability, and memory bus access patterns to train one-class Support Vector Machine (SVM) and Isolation Forest (IF) classifiers. This approach harnesses side channels with the capability to identify anomalies. The authors conducted a series of experiments employing a high-fidelity processor emulator. The experimental results demonstrate that the SVM-based detector, using the power consumption channel, achieved a high level of accuracy (ranging from 95% to 100%, depending on the type of attack). In subsequent attacks where code modifications altered the code's computational workload (e.g., adding more numerical calculations), the IF anomaly detector, which utilized the detected memory side-channel attacks, also achieved high accuracy (ranging from 95% to 100%). In these cases, code modifications led to changes in memory access patterns or the number of memory accesses. The researchers emphasize that the combined utilization of both side channels, in which the presented detectors were employed, leverages their respective strengths and proves highly effective in accurately detecting the considered attacks in their experiments.

In [20] this study, the authors underscore the significance of advancements in technologies for treating type 1 diabetes (T1D), with a particular focus on the field known as the "artificial pancreas" (AP). They highlight the use of mobile and transportable devices equipped with a closed-loop system, designed to administer insulin based on real-time data from a glucose sensor. In this presented work, the researchers shift their attention to the challenge of automatically detecting faults in these insulin pumps using multidimensional data-driven anomaly detection (AD) methods. They emphasize the use of unsupervised methods, which eliminate the need for labeled training data, often difficult to obtain in practical applications. The research outcomes are highly promising, primarily due to the well-designed approach they propose. This approach paves the way for the utilization of these methodologies in AP systems and their potential integration into remote monitoring systems, offering significant benefits in the field of T1D treatment.

In [21] this study, the authors bring attention to the issue of traffic fluctuations in mobile networks, which can lead to inaccuracies in network management and, subsequently, a decrease in service quality. To address this problem, the researchers present an online data mining technique for detecting anomalous data within network operations. Since this data forms the foundation for network operator decisions, including network monitoring and real-world phenomena control, ensuring its accuracy is crucial for minimizing potential harm. The experiments conducted in this study are based on actual traces of cellular communication, employing an automated platform known as STAD. This platform aims to identify spatiotemporal anomalies using a combination of machine learning techniques, including one-class Support Vector Machine (OCSVM), support vector regression (SVR), and recurrent neural networks such as Long Short-Term Memory (LSTM). The researchers validate their findings against a real dataset from Call Detail Records (CDR). The results presented in the study demonstrate significantly higher accuracy compared to methods such as Isolation Forest (IF) and Auto-Regressive Integrated Moving Average (ARIMA). This suggests that the approach proposed by the researchers offers a more effective solution for anomaly detection in mobile network traffic data.

In [22] this study, the authors focus on the detection of very low-frequency electromagnetic anomalies associated with earthquakes. They introduce a multifunctional method for identifying anomalies in AETA ULF (Ultra Low-Frequency) electromagnetic disturbance signals. This method is rooted in the Isolation Forest algorithm and includes an additional step for feature extraction and selection. To assess the effectiveness of their approach, the researchers

employ epoch analysis, abbreviated as SEA. This comprehensive method allows them to evaluate and detect anomalies in ULF electromagnetic disturbance signals related to seismic activity .

Table 6: Results K Comparison of two Methods

Methods	Station Abbreviation					
	GOX	MB	MX	EMS	ZT	MSQS
IQR	6.33	25.83	8.0	4.23	13.42	52.18
i-Forest	52.56	37.76	17.01	28.2	28.97	29.67

The research findings presented by the authors reveal that half of the chosen stations exhibit a correlation between signal anomalies and earthquakes. Moreover, their method demonstrates significantly superior performance compared to the traditional single-function IQR (Interquartile Range) method. The test results suggest that the approach introduced by the researchers holds the potential to be a much more effective choice in the future for identifying global anomaly points, particularly in the context of earthquake-related signal anomalies.

In [23] this study, the authors address the pressing issue of bacterial resistance to antibiotics, recognized as a global health crisis. Their research objective is to leverage machine learning to predict the minimum inhibitory concentration (MIC) for bacterial isolates based on the presence of resistance genes in their DNA. To achieve this, the authors employ a dataset that includes information about the presence of beta-lactamase genes, which were identified using whole genome sequencing. They also examine the correlation between the presence of these genes and the bacteria's sensitivity to beta-lactam antibiotics. The test results indicate that the proposed K-Nearest Neighbors method yields predictions of MIC with a similar level of accuracy as the Random Forest method, achieving predictions within a range of ±12-fold dilution. However, when it comes to predicting MIC, Random Forests achieve better F1-micro scores, demonstrating their effectiveness in this context.

In [24] the researchers in this study focus on employing the Isolation Forest (IF) method for detecting unusual behavior within business processes, encompassing errors and missing values. This led them to develop a method that integrates the degree of consistency within event log functions from multiple perspectives, utilizing IF as a tool for anomaly detection.

In their study, the researchers introduce an algorithm implemented within the Scikit-learn Framework. They assess the algorithm's performance in terms of its ability to quickly recognize anomalous behavior and the quality of its model accuracy. The research findings suggest that the algorithm developed by the researchers is effective in detecting abnormal behavior in this specific type of data and notably enhances the quality of the Isolation Forest (IF) model.

In [25] this research, the authors highlight that the Isolation Forest (IF) method is well-suited for precise data but encounters challenges when dealing with imprecise, incomplete, ambiguous, or uncertain data. In response, they introduce a novel approach known as the Fuzzy Isolation Forest. This method is designed to address the inherent "ambiguity" present in such data. In their work, the researchers utilize an approach based on alpha-cuts, wherein data instances are assigned to tree nodes, and their membership values are compared with an alpha-cut threshold. Through a series of experiments, the authors define this new solution and demonstrate that their method can operate with nearly the same accuracy as when working with clean, well-defined data. Importantly, their approach proves to be as effective as the IF method in detecting anomalies within fuzzy data without significantly increasing processing time.

In [26] this work, the authors highlight an enhancement to the Isolation Forest (IF) algorithm through the use of the Finite Boundary (FB) algorithm. This improvement is particularly advantageous when working with datasets composed of variables that conform to a standard distribution. It provides an alternative to existing

algorithms like IF and Extended Isolation Forest (EIF) in such scenarios.

Table 7: AUC Values for ROC for Credit Card Dataset

Features	IF	EIF	FBIF
(V4, V10, V11, V12, V13, V14, V15)	0.875	0.896	0.925
(V4, V10, V11, V12, V13, V14, V15 CATEGORY AMOUNT)	0.862	0.891	0.911
(V4, V10, V11, V12, V13, V14, V15 CATEGORY AMOINT, CATEGORY TIME)	0.833	0.879	0.908
(ALL NUMERICAL WITHOUT AMOUNT, TIME)	0.868	0.895	0.905
(ALL NUMERICAL, CATEGORY AMOUNT, CATEGORY TIME)	0.863	0.886	0.907

In their study, the authors introduce the FBIF algorithm, which leverages hyperspheres to form a branching boundary aimed at addressing data inconsistencies. This approach is especially effective when dealing with input features that exhibit properties akin to a standard distribution.

The researchers connect these hyperspheres to create a generalized branching decision boundary and assess the data for compliance with a standard distribution of input features.

The results obtained from a series of experiments indicate that the method proposed by the researchers significantly outperforms other methods such as FBIF with EIF and IF. This superior performance holds true even when the input data for the IF and EIF methods consists of a larger number of input features. In summary, their approach proves to be highly effective for anomaly detection, particularly when dealing with data characterized by standard distribution properties.

Table 8: AUC Values for PRC for Credit Card Dataset

Features	IF	EIF	FBIF
(V4,V10,V11,V12,V13,V14,V15)	0.862	0.889	0.931
(V4,V10,V11,V12,V13,V14,V15 CATEGORY AMOUNT)	0.867	0.891	0.920
(V4,V10,V11,V12,V13,V14,V15 CATEGORY AMOINT, CATEGORY TIME)	0.843	0.882	0.922
(ALL NUMERICAL WITHOUT AMOUNT, TIME)	0.874	0.897	0.922
(ALL NUMERICAL, CATEGORY AMOUNT, CATEGORY TIME)	0.871	0.892	0.921

In [27] this study, the researchers aim to predict the accuracy of detecting credit card fraud, specifically through the analysis of sequential transaction data. They compare two methods: Light Gradient Booster and Isolation Forest (IF).

In this work, the researchers employ an algorithm, specifically Light Gradient Booster, with a sample size of 10, and they also use the Isolation Forest (IF) method with the same sample size of 10. Their objective is to provide a highly detailed assessment of the accuracy percentage in detecting fraudulent credit card transactions. The experimental results presented in the study indicate that the Light Gradient Booster algorithm achieves significantly higher accuracy, with a rate of 91.6%, compared to the Isolation Forest method, which achieves an accuracy rate of 81.8%. This suggests that the Light Gradient Booster method is more effective in detecting credit card fraud in the context of sequential transaction data.

Table 8: The Accuracy of Predicting Fraudulent Credit Card Transactions Varies Based on Different Sample Sizes, with the Light Gradient Booster Algorithm Achieving an Accuracy of 91.6% and the Isolation Forest Method Achieving an Accuracy of 81.8%. the Choice of Sample Size Can Have a Significant Impact on the Performance of These Algorithms in Detecting Fraud

S. No.	Random_State	Light GMB	Isolation Forest
1	100	94.50	83.50
2	200	95.00	82.60
3	300	92.00	82.69
4	400	93.00	80.20
5	500	92.61	79.56
6	600	89.25	82090
7	700	91.60	81.80
8	800	90.11	81.90
9	900	88.11	83.00
10	100	90.26	79.50

The results of the conducted research indicate a statistically significant difference between the Light Gradient Booster (LGB) algorithm and the Isolation Forest (IF) algorithm, with a p-value of 0.0001 ($p < 0.05$) based on a two-sided analysis. This statistical significance suggests that there is a meaningful difference between the two algorithms.

Furthermore, the research findings confirm that the presented LGB algorithm outperforms the IF method, particularly in terms of accuracy in detecting fraudulent credit card transactions. Specifically, the LGB algorithm demonstrates a significantly higher accuracy rate for eight out of ten credit card transactions when compared to the IF method.

In [28] the researchers highlight that anomaly detection plays a crucial role in enabling automatic monitoring and identification of abnormal and unusual events. They emphasize that quick actions are essential for preventing potential failures. In this context, they reevaluate the classic Isolation Forest (IF) method, particularly when dealing with datasets that contain an exceptionally large number of data points. The researchers seek an alternative approach to calculate the division value, aiming to avoid dividing areas with very high data point density. This approach is designed to enhance the isolation capability of the method and improve its effectiveness in detecting anomalies within large datasets.

Table 9: Auc-Pr Results. the Second Column Shows the Value Obtained Using the Classic Isolation Forest. the Third Column Shows the Results Achieved Using the Proposed Model. Bold Texts Provide Leading Results

Dataset	Isolation Forest	Proposed Method
Annthyroid	0.3042 0.323	0.3227 0.0195
Breastw	0.9707 0.0043	0.9715 0.0032
Glass	0.1133 0.0106	0.0931 0.0112
Ionosphere	0.8095 0072	0.8021 0.0085
Mammography	0.2211 0.0357	0.2286 0.0476
Pendigits	0.02631 0.0639	0.2843 0.0574
Pima	0.5005 0.0089	0.5024 0.0135
Satellite	06583 0.0237	0.6651 0.0221
Thyroid	0.5257 0.0908	0.5461 0.0380

As previously emphasized, the researchers introduced a modification in the method by changing the approach to selecting the division value. This adjustment was aimed at avoiding

division in areas with high data point density. The research findings underscore that the proposed method consistently outperforms the results achieved using the traditional Isolation Forest (IF) method. Furthermore, it is noteworthy that these improved results are particularly significant when dealing with datasets that contain a large number of instances. Consequently, the comparative analysis between the standard IF method and the novel approach introduced by the researchers consistently indicates that the new method is superior, especially when applied to large datasets.

In [29] the researchers aimed to explore anomaly detection algorithms, with a specific focus on the Isolation Forest algorithm and the LOF algorithm, both renowned in the field of anomaly detection. They highlighted a challenge associated with these algorithms: determining which one is better suited for handling extensive datasets. Their study commenced by delving into the fundamental principles and operational mechanisms of these two algorithms. Subsequently, the researchers conducted practical experiments using a dataset to compare and assess the accuracy and stability of both algorithms in identifying data anomalies. As a result of their investigative work and the insights garnered from their experiments, the researchers concluded that the Isolation Forest algorithm exhibited greater effectiveness in detecting anomalies within data mining tasks. To address some of the limitations inherent to the Isolation Forest algorithm, they proposed several enhancements: Parameter Sensitivity Adjustment: The researchers suggested fine-tuning the algorithm's parameters to enhance its adaptability when processing datasets of varying sizes. This adjustment was aimed at bolstering overall performance and resilience. Drawing from Random Forest: Taking inspiration from the bagging-based Random Forest algorithm, the researchers sought to incorporate similar techniques to bolster the Isolation Forest's capability to effectively handle high-dimensional data. Integration of Simulated Annealing: The researchers recommended integrating a simulated annealing algorithm into the Isolation Forest. This addition would enable the algorithm to selectively filter out trees with superior detection accuracy, ultimately elevating the algorithm's overall accuracy in anomaly detection. In summary, the researchers objective was to not only compare the Isolation Forest and LOF algorithms for anomaly detection but also to propose valuable enhancements to the Isolation Forest algorithm. These refinements aimed to mitigate its limitations, rendering it more versatile and precise, especially when dealing with extensive and intricate datasets.

In [30] the researchers address the challenges associated with GPS data, which can be easily influenced by environmental factors, resulting in uneven data quality and impacting its utility. Their study focuses on utilizing the Isolation Forest (IF) technique to identify anomalies within GPS data. The IF technique operates by constructing a set of trees, initially using subsamples from the training dataset. Additional samples are then integrated into the isolated trees to calculate an anomaly score for each sampled point. For their experiments, the researchers employed real GPS data collected from sanitation vehicles and employed the Local Outlier Factor (LOF) technique for cross-validation. Detecting anomalies in GPS data poses several inherent challenges. Firstly, the likelihood of GPS drift is low, and historical data from moving targets provides limited actual abnormal data samples. Consequently, obtaining complete characteristics of all abnormal data is often unfeasible.

Additionally, GPS data frequently lacks explicit "normal" or "abnormal" labels, making the conversion of extensive GPS data into accurately labeled datasets challenging, particularly when manual labeling is required. To address these challenges, the researchers propose an unsupervised approach using the Isolation Forest method. This method enables the detection of anomalies within GPS datasets that are entirely unlabeled and contain a significantly larger proportion of normal data compared to abnormal data.

In [31] the researchers highlight that existing anomaly detection methods primarily rely on one-dimensional data for identification and

analysis. However, in their study, they introduce an algorithm designed for detecting anomalies within multidimensional data in an isolated forest framework. To achieve this, they employ multidimensional plane segmentation during the hyperplane segmentation construction, optimizing the process of generating an isolated forest.

The algorithm's effectiveness is demonstrated through comparative experiments involving several typical datasets, where it delivers promising results. The presented algorithm significantly enhances the analysis of multidimensional data by incorporating multidimensional plane segmentation, which improves segmentation efficiency and simplifies the identification of anomalies. Furthermore, within the isolated tree generation process, the researchers implement a method for controlling the lengths of the left and right subtrees. This halts growth once specific proportion conditions are met, leading to the selection of an iTree, thereby enhancing the efficiency of the isolated forest algorithm for anomaly detection. In summary, the researchers approach, which centers on improving the construction of isolated forests and the utilization of multidimensional plane segmentation in hyperplane segmentation construction, leads to more efficient segmentation and facilitates the identification of anomalies within multidimensional datasets.

Table 10: Influence of the Division Value on the F1 Result

Dataset	F1 score		
	I Forest	E-iForest	M. D. iForest
Cancer	0.5657	0.5672	0.5855
Carida	0.4126	0.4122	0.4174
Wbc	0.3413	0.3354	0.3972
Prime	0.5399	0.5388	0.5412
Satimage-2	0.0924	0.0862	0.0973

The research findings demonstrate that the algorithm proposed by the researchers outperforms the original iForest algorithm by achieving significantly higher values. As the next step, the researchers plan to focus on enhancing the algorithm's computational efficiency.

In [32] the researchers in this study focus on addressing the challenges posed by the vast amount of data generated by wind farms. They introduce an approach using an autoencoder combined with the Isolation Forest (IF) method to tackle the complexities and incompleteness often encountered when processing wind energy data.

Their methodology involves segmenting wind speed data into bins and creating a distribution of anomalous data, followed by the identification and removal of outliers using the autoencoder-based Isolation Forest. Subsequently, the data is reconstructed and cleaned. To validate the effectiveness of their approach, the researchers conducted experiments using real wind turbine data, utilizing the cleaned data for forecasting and comparing the results with actual forecasts. In summary, their study addresses outlier removal in wind speed-power curves, and the introduced forest cleaning algorithm, featuring an autoencoder and isolation forest, effectively removes outliers by leveraging wind speed intervals and the characteristics of abnormal data distribution.

Table 11: Comparison of Prediction Performance before and after Data Processing

Algorithms	Date status	R2	MSE
LSTM	Unprocessed data	0.7160	342664.4400
	Processed data	0.9850	15178.6981
XGBOOST	Unprocessed data	0.6847	380499.2805
	Processed data	0.9867	14177.0987
LinerRegression	Unprocessed data	0.6785	387913.4428

	Processed data	0.9704	31597.9429
Lasso	Unprocessed data	0.6785	387899.3228
	Processed data	0.9704	31617.9951
SVR	Unprocessed data	0.6825	383181.6014
	Processed data	0.9831	17980.5811
Ridge	Unprocessed data	0.6785	387913,2304
	Processed data	0.9704	31598.2600
RandomForest	Unprocessed data	0.6377	438113.0878
	Processed data	0.9844	16592.2207

Through a series of tests conducted on the Yalova wind energy dataset in Turkey, the researchers validate the effectiveness of their proposed method. They achieve this by comparing the prediction results using the data before and after the cleaning process. This verification is crucial for subsequent tasks such as wind power forecasting and modeling wind speed-power curves. Additionally, it provides valuable technical support for integrating wind energy into the grid and mitigating some of the adverse effects associated with wind energy generation.

In [34] this research, the authors highlight the significance of addressing fraud and anomalies within the telecommunications sector, emphasizing the potential of machine learning-based tools for detection. To tackle this challenge, they leverage Call Detail Record (CDR) data, comprising a substantial 417,000 call records across 217 different countries.

Initially, the researchers analyze telephone traffic, employing K-means clustering to categorize multi-valued categorical variables effectively. In the subsequent phase, they explore various models, including XGBoost, Extra Trees, Random Forest, and the Isolation Forest, as well as a novel model known as the Mixture of Experts. This new model calculates the average prediction probabilities of supervised models, contributing to anomaly detection. In the final step of their methodology, the researchers aggregate the results by summing the predictions from the five models, ultimately labeling connections based on their anomaly scores. The research findings reveal that approximately 1% of calls raise suspicion of fraudulent activity, aligning with industry reports on this issue. In another aspect of their work, the authors turn their focus to Fog Computing Networks (FCNs), taking into account factors such as user dispersion, real-time data, and user privacy. They introduce a federated learning framework that incorporates malicious models.

This framework comprises a two-layer blockchain, consisting of the primary blockchain and a directed acyclic graph chain, which enhances data security within the network. To ensure the security of global models, the authors present an isolation forest-based malicious model detection algorithm. This algorithm effectively filters out malicious local models, facilitating global aggregation using the Stochastic Gradient Descent algorithm. The presented research results, including simulation outcomes, demonstrate the effectiveness of the proposed algorithm in enhancing security and privacy in Fog Computing Networks.

In [34] the authors of this study focus on the context of fog computing networks (FCNs), taking into careful consideration factors like user distribution, real-time data processing, and user privacy. In their research, they introduce a federated learning framework that hinges on the concept of malicious models. This innovative framework incorporates a two-layer blockchain system, consisting of the primary blockchain and an additional directed acyclic graph chain, seamlessly integrated into the network structure. The primary purpose of this blockchain architecture is to fortify data security within the FCNs, addressing concerns related to user privacy and secure data transmission.

Within their work, the researchers introduce an isolation forest-based algorithm designed to detect malicious models effectively. This algorithm plays a pivotal role in filtering out potentially malicious local models within the FCNs. Furthermore, it enables the global

aggregation of data while maintaining security standards. This aggregation process is facilitated through the application of the Stochastic Gradient Descent algorithm, ensuring the integrity and security of the global model. The research findings, which encompass results from extensive simulations, convincingly demonstrate the effectiveness of the proposed algorithm. It offers a robust solution to enhance security and privacy within fog computing networks, contributing to the advancement of this technology in various application domains.

In [35] researchers are paying attention to identifying places for in situ research in robot and human exploration. In the presented work, researchers analyze the usefulness of machine learning algorithms for outlier detection in order to identify interesting samples in field exploration sites based on remotely sensed observations. For the presented analyses, the authors used two sites of the Cucomungo alluvial fan in Death Valley in California and the Jezero crater on Mars. The study results indicated that outliers identified by the isolation forest algorithm in the satellite datasets appeared to correspond to unique surface compositions or properties. The results indicate that this is a method that can support scientists by guiding the selection of a field exploration site.

In [36] the researchers are concerned with ensuring the reliability of sensor data in Internet of Things (IoT) applications, especially in detecting unusual conditions or anomalies like intrusions. They highlight the potential of machine learning-based anomaly detection in achieving this. In their study, they focus on addressing the issue of sensor data consistency within the context of hotspot detection using sensor pairs in a commercial IoT system designed for monitoring grain in large horizontal silos. The researchers aimed to assess the performance of traditional machine learning algorithms for anomaly detection, including the Localization Coefficient, Isolation Forest, and Single Class Support Vector Machine, in this specific context.

The study's findings reveal that the deep learning model with long-term memory (LSTM) introduced by the researchers outperforms the conventional machine learning methods [36].

Table 12: Results of the Traditional and Proposed Models

Model	Features	Accuracy	Precision	Recall	F1 Score
iForrest	Differences	85.2% (0.001)	84.2% (0.002)	38.9% (0.002)	0.532% (0.001)
	2 Sensors Readings	77.77% (0.001)	46.6% (0.002)	21.5% (0.002)	0.294 (0.001)
	2 Sensors Reading & Difference	85.2% (0.001)	84.2% (0.002)	38.9% (0.002)	0.532 (0.001)
LOF	Differences	53.9% (0.0008)	61.6% (0.001)	54.8% (0.003)	0.850 (0.001)
	2 Sensors Readings	57.6% (0.0006)	59.9% (0.001)	80.6% (0.002)	0.688 (0.001)
	2 Sensors Reading & Difference	57.8% (0.0001)	57.9% (0.001)	*99.5% (0.001)*	0.732 (0.001)
OneClass SVM	Differences	52.4% (0.002)	*100% (0.001)*	2.0% (0.001)	0.404 (0.002)
	2 Sensors Readings	51.5% (0.005)	54.9% (0.002)	1.2% (0.001)	0.231 (0.002)
	2 Sensors Reading & Difference	51.7% (0.005)	64.1% (0.002)	1.4% (0.001)	0.266 (0.002)
LSTM	*2 Sensors Readings*	90% (0.003)	88% (0.002)	90% (0.003)	*0.89 (0.004)*

In [37] the authors of the study draw attention to a significant surge in the volume of data being transmitted across publicly accessible computer networks. This proliferation of data has been accompanied by an escalating demand for novel and robust methods to safeguard against cyber threats and intrusions. In response to this pressing need, the researchers have directed their focus towards the crucial task of anomaly detection within the domain of cybersecurity. To address this challenge, the authors have undertaken a comprehensive analysis of various anomaly detection methodologies, including DBSCAN, One-class SVM, LSTM, and Isolation Forest. Their objective is to evaluate the efficacy of these methods in identifying and mitigating cybersecurity threats effectively. The experimental results presented in their study yield valuable insights. They suggest that the individual classification algorithms examined may not be

well-suited for routine cybersecurity operations, thereby underscoring the imperative to explore avenues for potential performance enhancements in this critical area.

In [38] this work, the authors highlight the crucial role of flight data recorders (FDR) and cockpit voice recorders (CVR) in aviation accidents investigations, emphasizing their mandatory usage in aircraft. However, the authors propose a novel approach termed I-FDR (Intelligence Flight Data Recorder) that aims to enhance the utility of FDRs by providing continuous data mining capabilities. This innovation seeks to empower flight crews with improved situational awareness through real-time data analysis. Similar to traditional FDRs, I-FDR records essential flight parameters, but its distinctive feature lies in its ability to process and analyze this data continuously. To assess the effectiveness of I-FDR in assisting flight crews during critical situations, the authors explore the applicability of three unsupervised machine learning algorithms: DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Isolation Forest, and LSTM (Long Short-Term Memory). The research results presented in the study provide valuable insights into the potential of these algorithms to contribute to flight crew management in hazardous scenarios

In [39] the authors emphasize the significance of Software-Defined Networking (SDN) as a solution for enhancing network security. SDN offers a holistic network view via a logically centralized component known as an SDN controller, effectively separating the control plane from the data plane. This division provides increased network control and introduces new possibilities for addressing emerging security threats, such as zero-day attacks. In their research, the authors propose a comprehensive machine learning (ML) framework designed for SDN environments. This framework incorporates unsupervised ML techniques and features a scalable method for collecting and selecting relevant network data attributes to facilitate the rapid detection of security threats. The key components of this framework include a Data Flow Collector (DFC) responsible for efficiently gathering network data features using the sFlow protocol, an Information Gathering Feature (IGF) selection module that identifies the most informative features, thus reducing training and testing complexities, and an innovative unsupervised ML module utilizing a novel outlier detection technique known as Isolation Forest (ML-IF) to swiftly and efficiently detect network security threats within the SDN context.

Table 13: Performance Metrics of Our Proposed Framework and State-of-the-Art ML/DL Models Based on the Available Measures Using the UNSW-NB15 Datase

Methods	Accuracy	DR	Time (second)
SSL	0.86	0.85	ON
RF	0.93	0.92	ON
FeedWSN	0.92	0.91	ON
OGM	0.95	0.94	ON
MHMM	0.96	0.95	ON
ML-IF	0.97	0.96	38.33

The experimental results conducted on the UNSW-NB15 public network security dataset clearly demonstrate the superiority of the framework introduced in this research. This framework exhibits substantial enhancements in terms of both detection accuracy and processing speed when compared to contemporary solutions.

Furthermore, it achieves these improvements while effectively reducing computational complexity, making it a highly promising and efficient approach to network security threat detection [39].

In [40] the authors of this study focus on chronic kidney disease (CKD), a complex condition that impacts kidney functions and structures, affecting millions of individuals worldwide and ranking among the leading causes of illness and death. In their research, the authors introduce an

innovative diagnostic approach for CKD, leveraging a 1D convolutional neural network (1D CNN) to address the limitations of existing methods while notably enhancing diagnostic accuracy. The outcomes of this study reveal that the model developed by the researchers achieved an impressive accuracy rate of 99.21%, surpassing current state-of-the-art techniques.

In [41] the authors of this study emphasize the significance of the industrial Internet of Things (IoT) and its successful integration with machine learning techniques in recent years. They address a notable limitation in many machine learning approaches, which is their ability to make predictions without providing explanations. This lack of interpretability poses a challenge for decision-makers who may struggle to comprehend and trust these predictions, potentially hindering the adoption of machine learning tools for practical use. To overcome this limitation, the authors propose an unsupervised anomaly detection system capable of not only identifying anomalies but also explaining the predictions, facilitating root cause analysis. They achieve this by combining the Isolation Forest method with a fast, model-independent interpretation technique known as Accelerated-agnostic Explanations (AcME). Their research findings highlight the effectiveness of AcME in providing explanations for predictions. The authors adapt AcME for anomaly detection, replacing the interpretation of Anomaly Detection with explanations for the anomaly results predicted by the Anomaly Detection algorithm, including the Isolation Forest. This approach proves to be both efficient and suitable for integration with decision support systems. In contrast, SHAP (Shapley additive explanations) is deemed impractical due to its high computational complexity.

In summary, the authors' work offers a promising solution for enhancing the interpretability of machine learning-based anomaly detection systems, particularly in the context of the industrial IoT, where actionable insights from predictions are crucial for informed decision-making.

In [42] their work, the authors introduce SPiForest as an innovative approach to outlier detection. They address a common issue faced by existing methods, where performance diminishes when dealing with outliers that appear similar to normal data in specific subspaces. To tackle this challenge, they propose SPiForest, which leverages principal component analysis (PCA) to identify principal components and estimate the probability density function (pdf) of each component. Additionally, they use inv-pdf, the inverse of the estimated pdf based on the dataset, to create support points across all attributes.

These support points define hyperplanes used to isolate outliers effectively. SPiForest offers two key advantages: it isolates outliers with fewer hyperplanes, leading to improved accuracy, and it excels at detecting outliers that may blend into subspaces due to their "few and different" nature. Comprehensive testing and comparative analyses demonstrate that SPiForest significantly enhances the area under the curve (AUC) compared to state-of-the-art methods. Notably, it improves AUC by up to 17.7% when compared to iF-based algorithms.

In [43] the authors highlight the significance of phasor measurement units (PMUs) in monitoring the state of smart grid systems. In their research, they explore various unsupervised learning techniques for detecting false data injection (FDI) attacks by identifying outliers. Initially, the study reveals that the moving average (MA) and density-based spatial clustering of applications with noise (DBSCAN) methods can accurately identify over 90% of the injected data points, while the isolation forest method detects almost 60% of these points. However, the research findings indicate that the implementation of the Amelia II imputation method, which imputes current data for multiple missing durations, yields impressive results. Specifically, the mean absolute percentage error (MAPE) for Amelia II is just 0.67, surpassing the performance of standard imputation methods.

Table 14: Results of DBSCAN in Identifying Injected Data

Data set	% Total Inj. Pts Identified/Number of Inj. pts	Pts number. Detected Outside Injection Interval / Total Number of Data Pts
W1, 6 Sec.	94.3% / 300	31/4080
W2, 6 Sec.	96.6%/300	65/4080
W1, 10 Sec.	98.4%/500	61/4080
W2, 10 Sec.	93.8%/500	65/4080
W1, 21 Sec.	88.9%/1050	22/4080
W2, 21 Sec.	91.4%/1050	37/4080

In [44] the authors of this study address pressing contemporary health concerns, highlighting that currently, one in every four individuals aged over 25 is at risk of experiencing a stroke. Their primary research focus revolves around the prediction of stroke occurrence in patients, a crucial aspect for physicians to determine prognosis and offer targeted therapy within a limited timeframe. In their research, the authors have constructed an ensemble model, considering various basic, bagging, and boosting classifiers, which include support vector machine, Naive Bayes, decision tree, logistic regression, artificial neural network, random forest, XGBoost, LightGBM, and CatBoost. The experimental results presented in their work indicate that the final ensemble model, employing the Max Voting approach, achieved an impressive accuracy rate of 95.76%.

In [45] the authors of this study address the issue of assessing fatigue in athletes or patients, a task traditionally performed using costly laboratory equipment. In their research, they focus on investigating whether Inertial Measurement Units (IMUs) placed on the lower limbs, either individually or in combination, can effectively distinguish between states of fatigue and non-fatigue and predict the degree of fatigue. To accomplish this, the researchers created a multi-IMU based running fatigue dataset by recording inertial data during running sessions that led to fatigue. They conducted experiments to validate the performance of a Random Forest (RF) model and a Support Vector Machine (SVM) for classifying running fatigue and determining fatigue levels. The research findings reveal that the RF model outperformed the SVM in classification accuracy. Moreover, as the number of sensors increased, the classification accuracy improved. Notably, the RF model achieved an accuracy of 87.21% when using IMU tibial data alone, while the highest classification accuracy of 91.10% was attained when combining tibial and femoral IMUs. The authors suggest that inertial sensors have the potential to objectively assess fatigue levels during running by detecting subtle biomechanical deviations in lower limb movements.

In [46] the authors of this study address the critical issue of intrusions into computer networks, which can disrupt network functionality and pose a significant threat to communication systems. Cyberattacks represent a substantial challenge in this context, compromising the privacy, authenticity, and availability of network resources. Intrusion Detection Systems (IDS) play a crucial role in identifying and mitigating these unauthorized actions or attacks. In their research, the authors propose a decision tree-based method for detecting network intrusions while enhancing data quality. They introduce their own model, which achieves impressive accuracy rates of 99.98% and 99.82% when tested on the CICIDS 2017 and NSL-KDD datasets, respectively. In comparison to existing models, this novel approach offers numerous advantages, particularly in terms of reducing the false alarm rate (FAR), improving detection rate (DR), and enhancing accuracy (ACC).

Table 15: Computation of Time

Criterion	Decision Tree	Proposed
Anomaly computation time	120ms	30ms
Signature computation time	150ms	50ms
Preprocessing time	50ms	15 ms

In [47] the authors highlight the significance of IPv6 target generation in the context of rapid IPv6 scanning for online surveys. They identify a critical issue, which is the low hit rate resulting from improper space partitioning attributed to outlier addresses and myopic partition pointers. To address this challenge, the authors introduce 6Forest, a novel approach to IPv6 target generation based on ensemble learning. This approach offers global coverage and robustness against address outliers. Their experimental findings, based on eight extensive candidate datasets, demonstrate the effectiveness of 6Forest.

It outperforms state-of-the-art IPv6 scanning methods on a global scale, achieving an impressive up to 116.5% enhancement for low-cost scanning and a remarkable 15x improvement for high-cost scanning.

In [48] the researchers focused their attention on the common method for estimating Gaussian mixture models (GMM). They identified a challenge with GMM estimation, which is its susceptibility to outliers. This sensitivity to outliers can result in subpar estimation performance, depending on the dataset being analyzed. In their study, the authors suggest incorporating an outlier detection step into the Expectation-Maximization (EM) algorithm. This step assigns an anomaly score to each data sample in an unsupervised manner. The experimental results demonstrate the effectiveness of this proposed enhancement compared to other established imputation techniques, highlighting its potential benefits in GMM estimation.

In [49] the researchers have put forward an enhanced version of the Isolation Forest (IF) method by introducing two key elements. The first element involves assigning weights to the path taken by each feature, thereby generating a more informative anomaly score. The second element modifies the aggregation function used to combine the results from individual trees within the forest. The researchers conducted extensive testing of their proposed method on various datasets. The outcomes of these tests and research highlight the significance of the first element, which enhances results at the individual tree level through the utilization of additional information. In contrast, the second contribution to the IF method involves a novel approach to aggregating results at the forest level, deviating from the original anomaly result and adopting a probabilistic tree interpretation.

In [50] the authors of this study explored various variants of the Isolation Forest (IF) method, including N-ary (NIF), fuzzy membership function-based (MIF), k-means clustering-based (KIF), two fuzzy clusters-enabled (CIF), and two fuzzy clusters with the addition of distance to the cluster center (prototype) (C2DIF). Their research and evaluation of these IF-based methods focused on detecting road anomalies in real-world data, which is a critical challenge for maintaining modern economies and infrastructure. The results of the study demonstrated significant improvements in accuracy and a reduction in the false-positive rate compared to other state-of-the-art methods. These enhancements led to a 100% increase in sensitivity, showcasing the effectiveness of the proposed IF-based methods for road anomaly detection.

Table 16: Accuracy, Sensitivity and False Positive Rate of the Compared Algorithms

Algorithm	Accuracy	Sensitivity	False Positive Rate
C2DIF	94.94%	100%	5.16%
KIF	94.94%	100%	5.16%
C2IF	95.24%	100%	4.86%
MIG	95.24%	100%	4.86%
NIF	95.50%	100%	4.60%
IF	95.47%	96.77%	4.56%
F-THRESH	94.16%	45.1%	5.83%
MOD-Z-THRESH	93.26%	68.33%	5.67%

The research results clearly demonstrate the usefulness and effectiveness of IF-based techniques in a series of experiments. Among these techniques, NIF stands out with the highest accuracy of 95.5%, surpassing currently employed methods like F-THRESH (slightly above 94%) and MOD-Z-THRESH (approximately 93%). This signifies that IF-based techniques perform significantly better in terms of false positives, with the current rate being 4.5% compared to 6% with the previously used methods. The presented methods and their research results conclusively establish their superiority over the original IF method, making them more valuable and applicable to the provided datasets.

In [51] the researchers have tackled the challenge of dealing with a large volume of alerts generated by network security devices, which can lead to alert fatigue and slow response times. They have proposed a solution to alleviate this issue by using Extended Isolation Forest to identify and highlight anomalous alerts. This approach has significantly improved the quality of alerts for monitoring purposes.

While acknowledging that no algorithm is entirely foolproof, the research results demonstrate that their model effectively reduces the number of alerts received in the Security Operations Center (SOC) by an impressive 82.15%. This means that Security Analysts only need to focus on monitoring 17.85% of the 50,000 total alerts received from the Intrusion Detection System (IDS) system. This substantial reduction in the volume of alerts can lead to more efficient and effective threat detection and response in cybersecurity operations.

In [52] the researchers have directed their attention to the crucial task of land monitoring in agriculture to provide early warnings about land conditions and anomalies for farmers. They have introduced an anomaly detection model for terrain monitoring systems. In their approach, they utilized raw data from site monitoring and employed Isolation Forest (IF) to transform unlabeled data into labeled data. Subsequently, they developed an anomaly detection model using a Long Short-Term Memory (LSTM) autoencoder. The experimental results reveal promising outcomes for their method. The LSTM autoencoder demonstrated an accuracy of 0.95, precision of 0.96, recall of 0.99, and F1-score of 0.97. This translates to an overall accuracy of 95.72% for the proposed anomaly detection method. However, the researchers acknowledge certain limitations. They note that the quality of the area-monitoring data is not optimal, as illustrated in the PCA function graph. Additionally, the ROC curve showed less than satisfactory results, indicating that the anomaly detection method has high false positive and false negative rates. In light of these limitations, the researchers collectively agree that there is a need to re-evaluate and refine the proposed model to enhance its performance and robustness.

In [53] the researchers have emphasized the importance of detecting anomalies in the context of Industry 4.0, particularly during production processes. Their study involves a comparison of various algorithms designed for detecting anomalies in time series data collected from automotive sensors. In their investigation, the researchers evaluated several algorithms,

including the interquartile range (IQR), isolation forest, particle swarm optimization (PSO), and k-means clustering. They specifically focused on identifying anomalies within automotive systems, both during the initial phase of a vehicle's usage and throughout its entire lifecycle. The authors highlighted the significance of the vast amount of data generated by sensors inside vehicles, which amounts to over a gigabyte per second. These sensors are interconnected through the vehicle's network, consisting of electronic control units (ECU) and the Controller Area Network (CAN bus). Each ECU processes input from its sensors, executes specific commands, and monitors the vehicle's normal state while detecting any abnormal behavior. Based on their research findings, the authors concluded that, for the task of unsupervised anomaly detection in time series data from vehicle sensors, the isolation forest algorithm outperformed the IQR and PSO+K-Means algorithms in terms of accuracy and effectiveness.

In [54] the researchers address the challenge of anomaly detection in the context of heterogeneous and correlated multivariate data, and they do so without assuming prior knowledge of statistical correlation. Notably, they employ a copula-based approach to assess the statistical correlation between different data modalities. In their methodology, they utilize an unsupervised learning (UL) framework to identify anomalies by working with data points extracted from a copula-based joint distribution. Specifically, they explore various Gaussian copula techniques, including R-Vine, D-Vine, and C-Vine, in combination with anomaly detection algorithms such as isolation forest, one-class SVM, local outlier factor, elliptic envelope, and UL autoencoder. The results of their experiments indicate that the proposed framework significantly outperforms direct training methods in terms of detection accuracy, showcasing its effectiveness in addressing the problem of anomaly detection in correlated and heterogeneous multivariate data.

In [55] the researchers focus on the detection of anomalies in time series data, as this data type is crucial for analyzing historical behavior and forecasting future trends. They emphasize that in domains such as satellite data, taxi rides, stock exchanges, online transactions, and more, the volume of generated data is so vast that manual processing becomes infeasible. In their study, the researchers employ deep learning models, including ARIMA, Isolation Forest, and LSTM-based autoencoders, to identify anomalies in datasets. The dataset used in their research pertains to daily closing prices, aiming to determine whether these prices are correct or not. The dataset is subjected to analysis using the aforementioned models. The test results reveal that ARIMA, LSTM, and Isolation Forest achieved accuracy rates of 90.13%, 84.98%, and 88.88%, respectively. The researchers suggest that further improvements can be made to these models by incorporating multidimensional time series data and optimizing them using various parameters for enhanced anomaly detection in time series datasets.

In [56] the researchers focus their attention on the issue of fraud in credit and debit card transactions, aiming to develop methods for effectively detecting online fraud when using credit cards. In their study, they explore various aspects of machine learning for fraud detection, including techniques such as the Local Outlier Factor (LOF), Isolation Forest (IF), and Convolutional Neural Networks (CNN). The research findings reveal that the authors achieved a high accuracy rate of 99% for both deep learning (CNN) and supervised machine learning techniques. Notably, the Isolation Forest method detected 73 errors, while the Local Outlier Factor identified 97 errors. Isolation Forest achieved an accuracy rate of 99.74%, and LOF achieved an accuracy rate of 99.65%. When comparing key metrics such as F1 score, precision, and recall across the three models (CNN, LOF, and IF), it became evident that the Convolutional Neural Network (CNN) outperformed the other methods.

Therefore, the conclusion drawn is that the Convolutional Neural Network method excels in identifying fraud cases, demonstrating that neural networks outperform traditional machine learning models in this context.

In [57] the researchers introduce a novel similarity search method called "Random Separations" designed to detect unknown variants of known malware in network traffic, with a focus on identifying threats. This method demonstrates superior performance when compared to other approaches, including unsupervised methods like isolation forest and lightweight online anomaly detectors, supervised approaches like random forest, and traditional similarity search algorithms such as kNN. The study involves the evaluation of eight high-risk malware families in various known-to-unknown ratios. The authors' proposed method, Random Separations, incorporates elements inspired by Random Forest and Isolation Forest. Notably, this algorithm introduces innovative features that contribute to its effectiveness in identifying threats in network traffic data.

In [58] the authors introduce a novel condition monitoring technique for wind turbine pitch systems, with a primary focus on reducing data storage requirements and computational intensity. This technique leverages the isolation forest anomaly detection model. In their approach, researchers train the new technique using data from a specific time period for each turbine. Subsequently, they evaluate each following month of data individually. The study's findings suggest that, in most cases, one month of data was sufficient for anomaly detection, while some instances required 3 or 4 months of data.

Remarkably, this technique demonstrated the capability to detect impending failures up to 3 months earlier and identify abnormal activity approximately 10 to 12 months before the actual failure occurs.

In [59] the authors introduce a versatile approach to constructing a gallery of multiple shots for observed reference identities. This method involves L2 norm descriptor matching for gallery retrieval, utilizing descriptors generated by a closed-set re-identification system. The gallery is continuously updated by replacing outliers with newly matched descriptors. To identify outliers, the authors employ the Isolation Forest algorithm, enhancing the gallery's resilience to incorrect assignments. Experimental results demonstrate that this approach surpasses state-of-the-art methods, as evidenced by improved TTR/FTR metrics. Additionally, the researchers note that this method performs effectively in controlled environments.

In [60] the authors emphasize the critical role of data accuracy from sensors in maintaining the functionality of mechanical equipment, particularly in practical industrial settings. In their work, researchers introduce an outlier detection algorithm based on a 1D neural network that offers depth separation through extended convolution. The research results demonstrate that this proposed method can enhance the accuracy of outlier identification and boasts state-of-the-art capabilities when compared to existing methods, including principal component analysis, k-means, isolation forest, local outlier, and one-class support vector machine.

Table 17: Prediction Results of Different Methods

	Precision (%)	Accuracy (%)	AUC
PCA	87.50`	99.26	0.9375
iForest	100.00	99.26	0.9961
k-means	62.50	97.78	0.8125
LOF	50.00	93.33	0.7303
OCSVM	50.00	57.78	0.4828
The Proposed Method	100.00	100.00	1.00

In [61] the authors highlight the significance of outlier detection in data mining, a crucial task applicable in various domains such as fraud detection, malicious behavior monitoring, and health diagnostics. In their work, the authors introduce PIF (Privacy-preserving Isolation Forest), designed to detect outlier values across multiple distributed data providers while offering both high performance and accuracy.

Additionally, PIF provides certain security guarantees. Research results indicate that PIF can achieve an average AUC (Area Under the Curve, a common metric for model performance) comparable to the existing iForest method while maintaining linear time complexity, thus preserving privacy without sacrificing efficiency.

Table 18: Comparison of Runtime (S) per Dataset and Algorithm

	H-Solution			V-Solution			iForest			LOF
	Train	Test	Total	Train	Test	Total	Train	Test	Total	
http (KDDCUP99)	0.014	7,189	7.203	0.034	69.486	69,520	0.008	7,495	7,503	ON
ForestCover	0.014	3,761	3.77	0.034	34,474	34,508	0.009	3.774	3,782	ON
Smtp	0.012	1.266	1.178	0.034	12,995	130.29	0.007	1,320	1.327	ON
shuttle	0.014	0.695	0.709	0.036	7,133	7,169	0.007	0.709	0.715	126,943
mammography	0.013	0.145	0.158	0.034	1.487	1.521	0.006	0.150	0.153	64,938
Satellite	0.021	0.091	0.112	0.032	0.909	0.0941	0.007	0.090	0.097	23,231
Pima	0.013	0.012	0.025	0.030	0.102	0.133	0.006	0.011	0.017	0.231
Breastw	0.014	0.009	0.023	0.029	0.096	0.125	0.006	0.009	0.015	0.240
Arrhythmia	0.078	0.006	0.084	0.030	0.062	0.092	0.006	0.006	0.012	0.475
Ionosphere	0.018	0.005	0.024	0.035	0.049	0.084	0.007	0.005	0.012	0.07

In [62] the authors address security concerns related to malicious users gaining unauthorized access to Virtual Machines (VMs) in cloud computing environments. They propose a proactive monitoring and anomaly detection model for VM resource usage. To develop this model, they utilize machine learning algorithms such as Isolation Forest and OCSVM (One-Class Support Vector Machine), training and testing the model on sampled VM load traces, which include various resource metrics. The research results indicate that OCSVM achieved an average F1 score of 0.97 for hourly time series and 0.89 for daily time series, while Isolation Forest achieved an average F1 score of 0.93 for hourly time series and 0.80 for daily time series. Both algorithms show promise for the presented model, but OCSVM exhibited a higher classification success rate compared to Isolation Forest.

In [63] the authors of this study conducted a comparative analysis of fault detection techniques, specifically Local Outlier Factor and Isolation Forest, and introduced a new methodology called Standardized Mahalanobis Distance. Their focus was on detecting faults in bearings and rotating machinery using data from vibration sensors. The research results indicate that the Standardized Mahalanobis Distance methodology outperforms both the Local Outlier Factor and Isolation Forest in detecting voltage drop faults in rotating machinery, particularly when the abnormal voltage value is not close to the nominal value. Additionally, this methodology demonstrated the capability to identify outliers before the occurrence of outer race errors in the bearing, making it a valuable tool for early fault detection.

In [64] the authors of this study highlighted the emerging field of Artificial Intelligence of Things (AIoT) as a notable trend in the context of Industry 4.0, which refers to the fourth industrial revolution characterized by the integration of digital technologies into various industrial processes. Additionally, they emphasized the importance of data privacy within this context, which is a critical consideration as more and more devices and systems become interconnected and generate vast amounts of data. Protecting the privacy of this data is crucial for maintaining the integrity and security of AIoT systems.

The authors of this study were particularly interested in addressing the issue of malicious models in the context of Federated Artificial Intelligence of Things (AIoT) with learning capabilities. They proposed a model called D2MIF, which is based on an isolation forest (iforest) and is designed to detect malicious models within the federated AIoT system. Their research findings demonstrated that the D2MIF model effectively identifies and mitigates the presence of malicious models, leading to a

significant improvement in the overall accuracy of the global model in the federated AIoT system with learning capabilities.

In [65] this study, the authors conducted a comparative analysis of methods for detecting and isolating damage in a chemical process, focusing on strengthening and bagging techniques. These techniques involve training multiple weak classifiers and combining their predictions to enhance the accuracy of damage detection. The study applied boosting methods like AdaBoost and gradient boosting, as well as bagging using a random forest approach, with decision trees as the base classifier. The research findings indicated that the strengthening and bagging methods outperformed conventional techniques when applied to the benchmark Tennessee Eastman process. Both boosting and bagging methods demonstrated significant improvements in performance. Specifically, the random forest (RF) algorithm stood out for its ease of implementation, parameter tuning, and predictive power assessment, making it a practical choice for damage detection in chemical processes. The RF algorithm's out-of-bag (OOB) accuracy provided a reliable measure of its predictive capabilities without requiring separate test samples or cross-validation.

In [66] the authors of this study focus on the problem of errors in electronic spreadsheets, which are widely used in organizations for data analysis and decision-making tasks. They are particularly interested in predicting whether a specific part of a spreadsheet contains errors. To address this issue, they propose a novel approach that combines a wide range of spreadsheet metrics with modern machine learning algorithms. In their approach, the authors employ supervised learning algorithms to create error predictors. These predictors utilize data from various spreadsheet metrics included in their catalog. The experimental results presented in the study demonstrate that, in many cases, random forest classifiers are effective at predicting whether a given spreadsheet formula contains errors with high accuracy. This highlights the potential of their method for identifying errors in spreadsheets.

In [67] the authors of this study address the growing security threats and risks associated with the rapid development of IoT (Internet of Things) in smart regions and cities, particularly in environments like smart healthcare and smart homes/offices. Their specific focus is on detecting tampering with IoT security sensors in an office setting.

To tackle this problem, the authors employ machine learning techniques in two ways: They use real-time traffic patterns to train an isolation forest-based unsupervised machine learning method for anomaly detection. They create labels based on traffic patterns and apply a supervised decision tree method within their anomaly detection system using machine learning (AD-ML). The research results demonstrate that both proposed models achieve a high accuracy rate of 84% with the isolation forest silhouette metric. Additionally, the supervised machine learning model, based on 10 cross-validations for decision trees, achieved the highest classification accuracy of 91.62% with the lowest false positive rate. These findings highlight the effectiveness of their approaches in detecting sensor tampering in office IoT environments.

In [68] the authors of this study acknowledge that machine learning (ML) and artificial intelligence (AI) methods have been widely applied to enhance research outcomes in building energy management, particularly when dealing with large datasets. Their specific focus in this work is on conducting a comparative study of various unsupervised fault detection approaches for heating, ventilation, and air conditioning (HVAC) systems, with a particular emphasis on scenarios involving a limited number of faulty data points. In this research, the authors evaluate three distinct methods: Isolation Forest, One-Class Support Vector Machine (OCSVM), Long Short-Term Memory (LSTM) Autoencoders. The experimental results indicate the following: LSTM outperformed all other techniques in detecting faulty data points, achieving an average precision of around 80%. Isolation Forest performed well when dealing with a small number of erroneous data points. However, its precision decreased as the number of data points increased. OCSVM

exhibited a different pattern, where precision increased with an increasing number of faulty data points until it reached 96 points. Afterward, a regressive trend was observed. These findings suggest that LSTM autoencoders are particularly effective for fault detection in HVAC systems, especially when dealing with limited faulty data points. However, the choice of method may depend on the specific characteristics of the dataset and the number of erroneous data points involved.

In [69] the authors of this study tackle the evolving challenges associated with chronic kidney disease (CKD), which has varying incidence rates, prevalence, and progression patterns, especially in regions with diverse social determinants of health. Their primary objective is to develop an intelligent system capable of efficiently classifying patients into two categories: "CKD" or "Non-CKD." Such a system could significantly aid healthcare professionals in expeditiously diagnosing patients, especially when dealing with a substantial caseload. In their research endeavor, the authors employ a range of unsupervised machine learning algorithms, including: K-Means Clustering, DB-Scan, Isolation Forest (I-Forest), Autoencoder. Additionally, they integrate these algorithms with various feature selection methods to optimize their performance. Notably, the integration of feature reduction methods with the K-Means Clustering algorithm yields remarkable results, achieving an impressive overall classification accuracy of 99% in distinguishing between CKD and Non-CKD clinical data. This research signifies the potential of their intelligent system, which combines K-Means Clustering with feature selection techniques, to significantly enhance the accuracy of patient classification in the context of chronic kidney disease diagnosis.

Table 19: Validation Scores for Reduced Features

	Kmeans	Dbscan	Autoencoder	Iforest
Recall	1	0.956	0.556	0.952
Precision	0.984	1	0.965	0.859
F1 score	0.992	0.978	0.706	0.903
Accuracy score	0.99	0.973	0.71	0.873
Mutualinfo score	0.926	0.84	0.258	0.441
Adjustedrand score	0.96	0.893	0.171	0.55
Vmeasure score	0.926	0.84	0.258	0.441
Silhouette score	0.321	0.327	0.211	0.178
Calinskiharabasz score	195,001	2001,258	138.11	109,964
Daviesbouldin score	1.055	1.115	1,396	1,223
TP	250	239	139	238
TN	146	150	145	111
FB	4	0	5	39
FN	0	11	111	12

In [70] the authors highlight the transformative impact of Internet of Things (IoT), cloud computing, and artificial intelligence (AI) on the healthcare sector, ushering in the era of smart healthcare. Their research centers on introducing an innovative disease diagnosis model that harnesses the convergence of AI and IoT, with a specific focus on diagnosing heart diseases and diabetes. In this novel approach, the authors leverage the CSO-CLSTM model, which stands for "Cascaded Long Short-Term Memory" and incorporates the Crow Search Optimization (CSO) algorithm. Additionally, they employ the isolation forest technique (iForest) to effectively eliminate outliers from the dataset. The research outcomes are noteworthy, demonstrating the CSO-LSTM model's impressive diagnostic capabilities. It achieves a remarkable accuracy rate of 96.16% in diagnosing heart diseases and an even higher accuracy rate of 97.26% for diabetes diagnosis. These results underscore the potential of the proposed CSO-LSTM model as a valuable tool for

disease diagnosis within intelligent healthcare systems.

In [71] the authors highlight the critical issue of faults occurring in heating, ventilation, and air conditioning (HVAC) chiller systems, emphasizing the negative consequences such as energy wastage, user discomfort, reduced equipment lifespan, and system unreliability. They emphasize the importance of early anomaly detection to prevent these issues and conserve energy. In their research, the authors introduce a data-driven approach designed to identify common faults in chiller systems. The proposed method utilizes Kernel Principal Component Analysis (KPCA) to capture the system's normal operating conditions. KPCA proves to be highly effective in handling non-linear phenomena, thanks to the incorporation of a Gaussian kernel. Moreover, a self-tuning procedure ensures optimal accuracy while maintaining strong generalization capabilities. Experimental findings underscore the superiority of the KPCA approach compared to linear PCA. Furthermore, the KPCA method out performs other anomaly detection techniques, including local outlier factor, one-class support vector machine, and isolation forest. These results demonstrate the effectiveness of KPCA in fault detection within HVAC chiller systems.

Table 20: Imulated Dataset—Classification Results: Comparison Between Linear and Kernel PCA

			Normal classified (%)	Faults classified (%)	BA
Rcaf	PCA	Normal	89.3%	10.7	53.55
		Faults	82.2	17.8	
	KPCA	Normal	88.2	11.8	85.50
		Faults	17.2	82.8	
revf	PCA	Normal	92.3	7.7	57.7
		Faults	76.9	23.11	
	KPCA	Normal	91.1	8.9	93.20

In [72] the authors underscore the importance of security in the transmission of information through Wireless Sensor Networks (WSNs). They emphasize that effective anomaly detection is a crucial element in ensuring the security of IoT systems. To address these concerns, they introduce a novel method named BS-iForest, which is based on an isolation forest variant designed specifically for wireless sensor networks.

One distinctive aspect of this approach is its utilization of a subset of data filtered by a boxplot for training and constructing trees. Notably, the authors deliberately do not select isolation trees with higher accuracy during the training phase to build the foundational forest anomaly detector.

Subsequently, this base forest anomaly detector is employed to assess data outliers in subsequent periods, facilitating efficient anomaly detection in the WSN context.

In [73] the authors introduce an innovative approach to enhance anomaly detection in datasets where conventional methods, such as the isolation forest (IF), struggle due to the unique characteristics of the data. Their solution involves leveraging a neural network trained for data reconstruction, coupled with the IF algorithm, which is known for its ability to identify outliers in datasets. In this method, an autoencoder is developed, where the neural network learns to generate a compact representation of the input data. Subsequently, the IF algorithm is applied to the reconstructed data to identify anomalies. This combined approach effectively improves the process of detecting anomalies in multidimensional datasets, addressing the challenges posed by the specific context and nature of the data.

In [74] the authors present modifications to the attention-based Isolation Forest (ABIF), which is based on Nadaraya-Watson regression. These modifications aim to enhance anomaly detection. The central idea behind their approach is to assign attention weights to each tree path using learnable parameters that depend on both the

instance being evaluated and the trees themselves. Importantly, the authors propose this solution without relying on gradient-based algorithms. Their numerical experiments involve synthetic and real datasets, demonstrating that ABIForest outperforms other methods. Notably, the test results confirm that ABIForest can detect additional anomalies that the original Isolation Forest failed to identify. The authors' straightforward attention-based model mechanism represents an initial step toward incorporating various forms of attention mechanisms into Isolation Forest. This inclusion holds promise for advancing research in anomaly detection.

Table 21: F1- Score Measures of the Original iForest and ABIForest as Functions of the Number of Anomalous Instancesnanom

The Ionosphere Dataset					
nano	10	20	40	50	60
iForest	0.655	0.960	0.694	0.687	0.681
ABIForest	0.692	0.709	0.711	0.695	0.687

In [75] the authors highlight the global shift towards distributed energy resources, particularly solar energy generated from photovoltaics, as a key renewable energy source. However, they stress the importance of detecting anomalies in individual photovoltaic panels due to their potential impact on performance and safety, including fire hazards. To address this, the researchers have developed techniques for accurately and effectively identifying anomalies in photovoltaic systems. Their work primarily focuses on the performance analysis of the isolation forest technique for anomaly detection in photovoltaic systems and the use of a rule-based fault localization technique to pinpoint faulty panel events. Through a series of experiments, the isolation forest method successfully detected around 453 anomalies among 45,740 observations, with approximately six panels indicating system failures. The researchers emphasize that any unusual behavior in the system can trigger the rule-based fault detection method. The research results demonstrate the stability and effectiveness of the presented approach in various challenging situations. The accuracy of the fault detection method is notably high, approximately 0.9886, indicating its suitability for identifying faults in photovoltaic systems.

In [76] the authors emphasize the growing significance of the Internet of Things (IoT) and the need for improved cybersecurity measures to counter emerging cyberattacks in this context. Their research aims to develop a malware traffic detection architecture that operates by analyzing summarized statistical packet data, rather than processing information from entire packets. This approach offers the advantage of enabling the analysis of a vast number of IoT devices while reducing the requirements for data storage and computational power. To achieve this goal, the authors employ two machine learning techniques, namely isolation forest and k-means clustering, in their architecture to identify malware traffic patterns. The research results aim to provide insights into the effectiveness of these techniques in detecting and mitigating malware threats within the specified architecture.

Table 22: Comparison Table

	Isolation	K-means
TPR	ABOUT	x
FPR	ABOUT	ABOUT
MCC	ABOUT	ABOUT
C&C detection	ABOUT	ABOUT
Host scan detection	ABOUT	ABOUT
DoS detection	ABOUT	ABOUT
speed	ABOUT	ABOUT

In [77] the authors emphasize the significance of the Building Internet of Energy (BIoE) in optimizing energy consumption, cost reduction, and enhancing building transformations. They stress that combining artificial intelligence with BIoE is essential for effectively analyzing big data and enabling intelligent decision-making. In their research, the authors introduce an intelligent approach for detecting anomalies in energy consumption patterns. Unlike traditional methods that focus on consumption values, this approach analyzes the shape of daily energy consumption curves. The authors employ a two-stage hybrid approach that combines supervised and unsupervised learning techniques. They utilize eXtreme Gradient Boosting (XGBoost) to create a regression model, enabling the classification of consumption anomalies on weekdays using rule-based algorithms and residuals. Subsequently, they employ the isolation forest algorithm for unsupervised anomaly detection. The results of their study demonstrate that this approach achieves an anomaly detection accuracy rate of 95.93%.

In [78] the authors focus on the critical role of networks in contemporary life and the increasing importance of digital security research. Their work centers on the development of an effective machine learning-based Intrusion Detection System (IDS) for detecting HTTP/2.0 slow Denial of Service (DoS) attacks. They begin by utilizing real attack-based datasets to understand the model better. In their study, the researchers explore the capabilities of three single-class classifier algorithms: Support Vector Machine (SVM), Isolation Forest (IF), and Minimum Covariant Determinant (MCD). The test results indicate that one of these classifier algorithms outperforms the others across various measure-ments, including accuracy (0.99), sensitivity (0.99), and specificity (0.99).

In [79] the researchers are primarily concerned with identifying and addressing corrupted data, which can result from various unethical and illegal sources. They emphasize the importance of developing a highly effective method for detecting and properly handling corrupted data within a dataset. In their work, they introduce PAACDA, which stands for Proximity-Based Adamic Adar Corruption Detection Algorithm. Their approach focuses specifically on detecting corrupted data rather than outliers, and they consolidate the results accordingly.

In their study, the authors introduce a novel PAACDA algorithm, which demonstrates superior performance compared to other unsupervised learning benchmarks. They conducted evaluations against 15 popular baselines, including K-means clustering, isolation forest, and LOF (Local Outlier Factor). The results indicate an accuracy of 96.35% for clustered data and an impressive 99.04% for linear data.

In [80] their research, the authors present a theoretical framework that examines the effectiveness of isolation methods from a distributional perspective. They propose that algorithms capable of fitting a mixture of distributions, where the average path length of observations approximates the mixture coefficient, can provide valuable insights. Building on this concept, they introduce a Generalized Isolation Forest (GIF), which goes beyond the traditional use of average path length when combining random trees. Through an extensive evaluation of over 350,000 experiments, the researchers demonstrate that GIF outperforms other methods across various datasets while maintaining comparable runtime. Their work deepens the theoretical understanding of isolation -based techniques and introduces a novel algorithm for improved outlier detection.

Table 23: ROC AUC Score for Isolation-Based Methods (EIF, GIF, IF, and SciF) and Proximity-Based Methods (iNNE, LeSiNN, aNNE). Results From [4] Are Also Included Which Evaluate 12 Additional Proximity-Based Methods. Bold Highlights the Best Isolation-Based Out

	Gif	Eif	If	Science fiction	Anne	Other	Lesinn	[3]
Annthyroid	0.6474	0.6263	0.708	0.518	0.5980	0.5363	0.5903	*0.7666*
Cardiotocography	*0.92 85*	0.7923	0.7944	0.7066	0.6546	0.8172	0.6601	0.8279
Crcditfraud	0.9421	0.9522	0.9513	0.9133	0.9447	0.9516	0.9529	0.9624
Forcstcovcr	0.9408	0.92	0.9236	0.7013	0.7474	0.9568	0.7988	0.8847
KDDCup99	0.9831	*0.9909*	0.9895	0.9905	0.9778	0.9859	0.9152	0.9904
mammography	*0.8706*	0.8119	0.8175	0.5922	0.7345	0.8018	0.7683	0.8435
PageBlocks	0.9342	0.9271	0.9281	0.8616	0.8408	0.9461	0.8904	0.96
PenDigils	0.9644	0.9205	0.9154	0.8803	0.9369	0.8236	0.9413	0.9903
Pi has	*0.8355*	0.7453	0.7415	0.6415	0.6021	0.7272	0.6805	0.7759
Satellite	*0.8575*	0.741	0.7277	0.5893	0.6791	0.745	0.6602	0.7315
ShuUle	0.85%	0.802	0.868	0.9847	0.7051	0.8074	0.7259	0.9484
Spambase	0.7508	0.7889	0.8147	0.8216	0.7038	0.7409	0.7160	0.7914
Waveform	*0.9115*	0.7289	0.7192	0.7115	0.6701	0.7578	0.6527	0.7713
Veela	0.5683	0.4042	0.5188	0.4796	0.5282	0.5363	0.5903	*0.7858*

In [81] their research, the authors introduce an enhanced SA-iForest method for data anomaly detection. Their primary objective is to address the issues of low accuracy and execution efficiency associated with general data anomaly detection algorithms based on the Isolation Forest (IF) method. Their approach focuses on selective integration, prioritizing precision and value differentiation. To achieve this, the authors utilize the simulated annealing algorithm to select isolation trees with a high capability for detecting anomalies, optimizing the forest in the process. Their proposed approach results in substantial improvements in algorithm performance by enhancing the construction of the isolated forest.

Table 24: Execution Time of Different Data Sets

Date set name	SA-iForest	iForest	LOF
Breast	0.11	0.23	1.14
Diabest	0.24	0.29	2.07
Unblanched	0.27	0.38	3.63
Http	8.83	32.67	9186.15
Shuttle	3.21	9.36	736.08
Pendigits	1.4	4.88	376.12
Messidor_features	0.89	1.16	4.54

The research findings indicate that both SA-iForest and iForest algorithms exhibit significantly lower execution times compared to the LOF algorithm. This reduction in execution time is attributed to the fact that SA-iForest and iForest do not involve distance or density calculations, which are not utilized for anomaly detection. Consequently, eliminating distance and density-based methods results in reduced computational costs. The researchers highlight that SA-iForest achieves a lower computational complexity constant. In the SA-iForest algorithm, tree selection is based on simulated annealing for improved detection performance. As a result, it does not construct the entire set of trees used for anomaly detection, unlike the original iForest, which builds all the trees for detection purposes. By removing some detection efficiency and employing the fast convergence of the simulated annealing optimization algorithm, SA-iForest enhances network intrusion anomaly detection performance. The presented test results affirm that SA-iForest exhibits lower execution times than iForest and demonstrates significant improvements in overall generalization and prediction performance.

In [82] the authors of this research focus on enhancing the security of computer networks,

particularly by detecting intrusions at an early stage to minimize their impact, which is critical for an institution's financial well-being and reputation. They emphasize the two primary intrusion detection systems (IDS): signature-based and anomaly-based IDS. Signature-based IDS: These systems identify intrusions by comparing network traffic patterns to a database of known intrusion signatures. If a match is found, it signals an intrusion. Anomaly-based IDS: In contrast, anomaly-based IDS systems identify intrusions by establishing a baseline of normal network behavior. Deviations from this baseline are flagged as potential intrusions. In their research, the authors employ real-time traffic data from the University of Virginia network to evaluate the performance of two anomaly-based intrusion detection methods: Local Outlier Factor (LOF) and Isolation Forest (iForest). They aim to understand the similarities and differences in the results produced by each approach. The results of the research indicate that iForest scores exhibit greater specificity across all data points compared to LOF scores. This suggests that, especially when anomalies are considered rare and distinct data points, the Isolation Forest method performs better at identifying anomalies than LOF.

In [83] the authors of this research highlight the growing volume of hydrological data and the challenges associated with applying anomaly detection algorithms efficiently. They point out that current algorithms often suffer from low time efficiency and produce an excessive number of anomaly points, making it difficult for decision-makers to extract meaningful insights. To address these issues, the researchers present an application of the Isolation Forest algorithm for hydrological pattern representation data. They further propose an Isolation Forest-based hydrological time series anomaly pattern detection algorithm, which they validate through a series of experiments. One significant challenge they tackle is the difficulty in determining the partition threshold in the Isolation Forest and deriving the top-k anomalies. To overcome this, they integrate the k-means clustering algorithm and the nearest neighbor algorithm into the Isolation Forest method. This enhancement reduces the subjectivity associated with manually setting thresholds and improves result stability.

The authors apply these algorithms to data collected from the Chuhe River Basin and compare their improved Isolation Forest method with other anomaly detection algorithms in terms of accuracy and time complexity. The results demonstrate that the improved Isolation Forest algorithm is well-suited for characterizing large hydrological time series and achieves high accuracy in anomaly detection.

Table 25: Comparison of Four Algorithms AUC and Run Time

Algorithm name	AUC	Time/ms
Improved iForest	0.9239	39
NLOF	0.9089	278
LOF	0.8891	475
Improved k-means	0.7267	1136

In [84] the authors of this research highlight the significant volume of electricity data collected from the distribution network and the potential for incorrect data to occur, which can adversely impact data analysis. In response, they propose an Attribute-Linked Isolation Forest algorithm designed to detect anomalous electrical data. In their algorithm, attributes are regrouped based on attribute associations, and they modify the partition generation method. By using data samples to traverse the isolation forest, abnormal data is identified based on the anomaly score.

Furthermore, they suggest that these anomalies can be corrected using the modified Wavelet Neural Network method. The authors emphasize that their method effectively and quickly detects incorrect data and demonstrates high accuracy in both identifying and correcting such data anomalies, addressing a critical issue in working with electrical data.

Table 26: Anomaly Data and Correction Results

Original Anomaly Data			Corrected Data		
Active Pow Wher (k)	Reactive Power (kVar/h)	Voltage	Active Power (kW/h)	Reactive Power (kVar/ bh)	Voltage (V)
10.314	0.007	234.84	4.216	0.418	234.84
5.36	3,663	233.63	5.36	0.436	233.63
5,388	0.502	543.70	5,388	0.502	233.74
3.52	0.522	544.18	3.52	0.522	235.02
3.702	0.02	33.32	3.702	0.52	235/09
3.7	4.34	235.22	3.7	0.52	235.22
0.067	0.51	233.99	3.668	0.51	233.99
3.662	0.01	233.86	3.662	0.51	233.86
4.448	0.498	657.55	4.448	0.498	232.86
5.412	0.47	352.66	5.412	0.47	232.78
1,670	0.478	232.99	5.224	0.478	232.99
5.268	3,398	232.91	5.268	0.398	232.91
4,054	0.422	499.08	4,054	0.422	235.24
3,384	2,677	237.14	3,384	0.282	237.14
6.27	0.152	236.73	3.27	0.152	236.73
6.43	0.156	237/06	3.43	0.156	237/06
3.266	0	437.13	3.266	0.237	437.13
3,728	0	677.10	3,728	0.235	435.84

The results presented by the researchers show that their proposed method detected 15 anomalies, with a false detection rate of 0.08%. In comparison, using the original Isolation Forest (IF) method, a total of 21 anomalous data points were detected, resulting in a higher false detection rate of 0.16%. This indicates that the researchers' method outperforms the traditional IF approach, achieving higher efficiency and accuracy in anomaly detection while mitigating the issues of redundancy.

In [85] the researchers focused on software defect prediction (SDP) in software engineering, which aims to assess the quality and reliability of software. They introduced an innovative SDP method based on the isolation forest (IF) algorithm, with a particular emphasis on feature selection. Their experiments were conducted on five real NASA datasets. The researchers found that the selection of features for building an isolation tree had a significant impact on prediction performance. Carefully chosen features were more effective than randomly selected ones.

The results indicated that their proposed method, which leverages the isolation forest algorithm, could help mitigate issues associated with unbalanced training data in software defect prediction.

In [86] the researchers addressed the challenge of ensuring proper identification in modern applications, especially when dealing with unlabeled data. Their focus was on practical applications, and they aimed to compare various unsupervised anomaly detection techniques using performance metrics like precision, recall, F-score, and the area under the curve (AUC). They experimented with techniques such as One-Class Support Vector Machine (OneClassSVM), Local Outlier Factor (LOF), Isolation Forest (IF), and Elliptic Envelope (EE) using datasets from shuttles and satellites. The research results showed that these unsupervised anomaly detection techniques had varying levels of performance on the given datasets. However, the specific findings or performance metrics achieved were not provided in the summary.

Table 27: Area Under the Curve for Shuttle Dataset

Alhorithm	AUC
One Class SVM	0.95
LOF	0.87
IF	0.99
EE	0.94
SVM	0.85

In [87] the researchers addressed the challenge of analyzing a vast amount of raw data from social media, which contains ironic statements. Manual analysis of this data is challenging due to its rapid growth and limitations posed by character constraints and typographical errors on social media platforms. Traditional classification methods were found to be insufficient for this task. Therefore, the researchers treated irony detection in online social media as a classification problem and assessed the performance of various supervised machine learning methods on real data. The machine learning algorithms applied for irony detection included Bayesian Network (BayesNet), OneR, Stochastic Gradient Descent (SGD), Logistic Model Tree (LMT), Multi-Layer Perceptron (MLP), Radial Basis Function Networks (RBF), Voted Perceptron, IBk, Randomizable Filtered Classifier (RFC), Isolation Forest, Fuzzy Lattice Reasoning (FLR), and Bagging algorithms. The researchers conducted comprehensive evaluations of these methods. The results of the experiments varied depending on the evaluation criteria and the allocation of training and test data. The Voted Perceptron, IBk, and RFC algorithms performed best across various evaluation metrics. Specifically, when the entire dataset was used as training data, IBk and RFC achieved the best results. However, when 70% of the dataset was allocated as training data and the remainder as test data, Voted Perceptron outperformed the other algorithms in terms of most evaluation metrics, except for accuracy. The researchers also conducted 10-fold cross-validation and reported findings, although specific details about those results were not provided in the summary.

Table 28: Results of Experiment III

Classification algorithms	Evaluation Metrics			
	Accuracy	Precision	Recall	F-measure
BayesNet	0.669	0.670	0.669	0.669
OneR	0.674	0.674	0.674	0.674
SGD	0.674	0.674	0.674	0.674
LMT	0.671	0.671	0.671	0.671
MLP	0.671	0.671	0.671	0.671
RBF	0.660	0.660	0.660	0.660
VotedPerceptron	0.676	0.676	0.676	0.676
IBk	0.668	0.668	0.668	0.667
RFCs	0.669	0.668	0.669	0.667
IsolationForest	0.446	0.448	0.446	0.428
FLR	0.504	0.506	0.505	0.505
Bagging	0.672	0.672	0.672	0.672

In [88] the researchers highlight the importance of monitoring the production process and detecting anomalies to maintain a high level of quality in the final product. In their study, they propose an integrated stack-based anomaly detection method called SBIOD. This method combines five different learning components: HBOS (histogram-based outlier detection), LOF (local outlier factor), iForest (isolation forest), DT (decision tree), and LR (logistic regression). In addition to these components, the researchers introduce a feature normalization step, and they incorporate the MLP (Multi-Layer Perceptron) classifier to analyze feature values. The goal is to improve the performance of anomaly detection while maintaining practical value.bThe presented

research results indicate that the SBIOD method significantly enhances the AUC (Area Under the Curve) metric, which is a common measure of the performance of classification models. This suggests that the SBIOD approach is effective in detecting anomalies in the production process and is suitable for practical applications where maintaining product quality is crucial.

Table 29: Results from Different Models for Sugar

Algorithms	Auc before cross validation	Auc after cross validation	Recall
HBOS	0.634	0.623474	0.04
LOF	0.462	0.409413	0.01
iFores1	0.534	0.479122	0.26
DT	0.956	0.994589	0.92
Logistic Regression	0.764	0.970663	0.54
SBIOD	0.95	0.9828	0.94

In [89] the researchers focus on addressing abuses by call center agents and present a machine learning model for abuse detection. They base this model on features extracted from audio recordings of phone calls. To identify the most effective approach, they train, evaluate, and compare three different models: one-class support vector machine, isolation forest, and multivariate Gaussian models. The study's results reveal that a combination of the recursive SVM feature elimination scheme and the isolation forest algorithm performs best for detecting three out of four types of abuse in the phone call recordings. This indicates that this specific combination of techniques is the most suitable for identifying instances of abuse in call center interactions.

In [90] the authors emphasize the challenges in processing ticket purchase applications due to their sheer volume, making management and analysis difficult. To address this issue, they propose the use of both unsupervised and supervised learning techniques to identify invalid tickets. These techniques include autoencoders, angle-based outlier detection, gradient boosting, isolation forests, and neural networks. The authors note that by identifying outliers within the ticket applications, government entities can significantly improve their efficiency in handling these requests, ultimately enhancing their relationship with city residents. The study's results indicate that unsupervised approaches were most effective when examining the dataset for reports with unusual attribute combinations. Conversely, supervised approaches performed best in identifying tickets with unreasonable turnaround times. Furthermore, the authors highlight that the presence of true outliers can lead to better evaluations of various anomaly detection algorithms, thereby improving their overall functionality. This, in turn, allows for more accurate detection of future outliers in similar datasets.

In [91] the authors highlight the significance of product reviews and comments on popular auction websites, as they play a crucial role in influencing buying and selling decisions. They specifically address the issue of fake reviews, which can mislead consumers with fraudulent information, leading to financial losses. In their research, the authors propose a method for detecting fake reviews based on the review records associated with products. Their approach involves extracting product review records into a temporal feature vector. They then develop an isolation forest algorithm to detect outlier reviews, with a focus on identifying differences in product review patterns to pinpoint outlier reviews. This method aims to enhance the authenticity and reliability of product reviews on auction websites.

In [92] the authors introduce a novel outlier detection method based on machine learning in their research. They utilize the isolation forest algorithm to compute the outlier coefficient. In the subsequent step, they establish an outlier detection model, essentially transforming the problem of outlier detection into a binary classification learning model. According to the results presented in their research, this new

method for detecting outliers is competitive with the Z-score method, as it achieves superior results in terms of accuracy and effectiveness. This suggests that their approach offers a promising alternative for outlier detection in various applications.

In [93] their research, the authors focus on the application and analysis of four machine learning (ML) techniques for diagnosing failures in vehicle fleet tracking modules. They specifically compare various sampling methods during the training and testing process using real data from DDMX, a company involved in vehicle fleet tracking. The authors create 16 models using Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) techniques. Their results indicate that these techniques achieve high precision rates of 99.76% and 99.68%, respectively, in detecting and isolating errors in the provided dataset. They suggest that the models developed in their work can serve as prototypes for remote fault diagnosis, showcasing their potential for practical applications in vehicle fleet management and maintenance.

Table 30: Time to Train and Evaluate the Models (S)

		RF		N.B		SVM		MLP	
		Train	Test	Train	Test	Train	Test	Train	Test
Set 1	U.S	0.77	0.31	0.04	0.31	269.00	0.50	3.06	0.36
	AXIS	1.00	0.20	0.01	0.31	565.00	0.54	3.27	0.35
Set 2	U.S	0.53	0.31	0.08	0.32	102.00	0.55	0.54	0.28
	AXIS	1.21	0.38	0.11	0.47	0.865	0.4	2.51	0.27

In [94] their research, the authors highlight the application of machine learning techniques in developing intelligent solutions for detecting anomalies in various computer and communication systems. Their specific focus in this study is to assess the performance of both supervised (such as K-Nearest Neighbors - KNN and Support Vector Machine - SVM) and unsupervised (like Isolation Forest and K-Means) algorithms for intrusion detection, using the UNSW-NB12 dataset. The research results presented in their study show that the supervised SVM Gaussian fine algorithm achieved an impressive accuracy rate of 92%. This high accuracy indicates the algorithm's capability to effectively classify normal and abnormal data.

However, they also note that unsupervised algorithms like the K-Means algorithm are proficient in grouping data and determining the appropriate number of clusters. Nevertheless, they point out that the UNSW-NB12 dataset used in their study exhibits high data density, which can pose challenges for clustering algorithms.

IV. CONCLUSION

In this paper it has been done a brief review of existing outlier detection methods that have been developed in the last 5 years. Methods based on outlier detection are widely applicable to various data sets. Because there are many anomaly detection models.

From the analysis of the above texts emerges a fascinating picture of the role of Isolation Forests in the domain of anomaly detection and outlier observation. This tool, based on decision trees, demonstrates its potential across various contexts and domains, making a valuable contribution to the fields of data analysis and cybersecurity. In the realm of computer security, Isolation Forests play a pivotal role in Intrusion Detection Systems (IDS). Research suggests that it is a promising tool for identifying unauthorized activities and attacks. Importantly, Isolation Forests have the ability to detect new types of threats that have not yet been classified in databases, which is crucial for protecting sensitive data and maintaining the integrity and reliability of computer systems. In the area of fault diagnostics and device monitoring, Isolation Forests are equally valuable.

They enable the detection of errors and abnormalities in various mechanical and technical systems. Industries where failures can lead to significant financial losses and decreased efficiency benefit from the quick detection and

response to anomalies that this tool provides. In the field of large-scale data analysis, Isolation Forests open doors to discovering atypical patterns and outlier observations. They allow for the identification of hidden dependencies within data, which is essential for making informed business decisions. This tool offers the opportunity for in-depth data exploration and a better understanding of its structure, which is invaluable in today's world where data is a source of competitive advantage. Concerning fraud detection and abuse prevention, Isolation Forests prove useful in identifying fake product reviews or abuses in customer service. This is important not only from the perspective of consumer protection but also for companies aiming to maintain the integrity and quality of their products and services. It provides an avenue for safeguarding a company's reputation and building trust with consumers. It's worth emphasizing that Isolation Forests come with several advantages, such as higher precision, efficiency, and speed compared to other anomaly detection techniques. However, it's essential to remember that the effectiveness of this algorithm may depend on the specific data and problem context. Therefore, ongoing research and experimentation are crucial to tailor this technique to various applications and challenges.

In conclusion, Isolation Forests remain an incredibly valuable tool in data analysis, cybersecurity, and outlier detection across multiple domains. Research and development in this technique have the potential to bring further enhancements and broaden its applications in the future. This tool continues to evolve, opening new possibilities for researchers and professionals engaged in data analysis and cybersecurity. Isolation Forests serve as a foundation for innovative advancements in the field of anomaly detection, enhancing our ability to understand and protect the digital world.

REFERENCES

1. Puggini, L., & McLoone, S. (2018). An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. Engineering Applications of Artificial Intelligence, 67, 126-135.
2. Ounacer, S., Ait El Bour, H., Oubrahim, Y., Ghoumari, M. Y., & Azzouazi, M. (2018). Using Isolation Forest in anomaly detection: the case of credit card transactions. Periodicals of Engineering and Natural Sciences, 6 (2), 394-400.
3. Li, X., Cai, Y., & Zhu, W. (2018). Power Data Cleaning Method Based on Isolation Forest and LSTM Neural Network. W X. Sun, Z. Pan i E. Bertino (Red.), Cloud Computing and Security (s. 11067). Springer. doi: 10.1007/978-3-030-00018-9_47.
4. Habler, E., & Shabtai, A. (2018). Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages. Computers & Security, 78, 155-173.
5. Liu, Z., Liu, X., Ma, J., & Gao, H. (2018). An Optimized Computational Framework for Isolation Forest. Mathematical Problems in Engineering, 2018, Article ID 2318763, 13 pages. doi: 10.1155/2018/2318763.
6. Chun-Hui, X., Chen, S., Cong-Xiao, B., & Xing, L. (2018). Anomaly Detection in Network Management System Based on Isolation Forest. W 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC) (s. 56-60). IEEE. doi: 10.1109/ICNISC.2018.00019.
7. Tao, X., Peng, Y., Zhao, F., Zhao, P., & Wang, Y. (2018). A parallel algorithm for network traffic anomaly detection based on Isolation Forest. International Journal of Distributed Sensor Networks, 14(11). doi: 10.1177/1550147718814471.
8. Liao, L., & Luo, B. (2018). Entropy Isolation Forest Based on Dimension Entropy for Anomaly Detection. Communications in Computer and Information Science.
9. Filippov, A. I., Iuzbashev, A. V., & Kurnev, A. S. (2018). User authentication via touch pattern recognition based on isolation forest. W 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) (s. 1485-1489). IEEE. doi: 10.1109/EIConRus.2018.8317378.
10. Kurnianingsih, L. E., Nugroho, E., Widyawan, L., Lazuardi, L., & Prabuwono, A. S. (2018).

Detection of Anomalous Vital Sign of Elderly Using Hybrid K-Means Clustering and Isolation Forest. W TENCON 2018 - 2018 IEEE Region 10 Conference (s. 0913-0918). IEEE. doi: 10.1109/TENCON.2018.8650457T.
11. Wu, Y. -JA Zhang, & X. Tang. (2018). Isolation Forest Based Method for Low-Quality Synchrophasor Measurements and Early Events Detection. 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 1-7. DOI: 10.1109/SmartGridComm.2018.8587434.
12. Togbe, M. U. et al. (2020). "Anomaly Detection for Data Streams Based on Isolation Forest Using Scikit-Multiflow." In: et al. Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science (), vol 12252. Springer, Cham. DOI: 10.1007/978-3-030-58811-3_2.
13. Togbe, M.U., Chabchoub, Y., Boly, A., Chiky, R.: "Etude comparative des méthodes de détection d'anomalies." Revue des Nouvelles Technologies de l'Information Extraction et Gestion des Connaissances, RNTI-E-36, 109–120 (2020).
14. Hara, Y., Fukuyama, Y., Murakami, K., Iizaka, T., & Matsui, T. (2020). Fault Detection of Hydroelectric Generators using Isolation Forest. 2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Chiang Mai, Thailand. DOI: 10.23919/SICE48898.2020.9240331.
15. Ma, H., Ghojogh, B., Samad, M. N., Zheng, D., & Crowley, M. (2020). Isolation Mondrian Forest for Batch and Online Anomaly Detection. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada. DOI: 10.1109/SMC42975.2020.9283073.
16. Karczmarek, P., Kiersztyn, A., & Pedrycz, W. (2020). Fuzzy Set-Based Isolation Forest. 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK. DOI: 10.1109/FUZZ48607.2020.9177718.
17. Liu, S., Ji, Z., & Wang, Y. (2020). Improving Anomaly Detection Fusion Method of Rotating Machinery Based on ANN and Isolation Forest. 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China. DOI: 10.1109/CVIDL51233.2020.00-23.
18. Fitriyani, N. L., Syafrudin, M., Alfian, G., Fatwanto, A., Qolbiyani, S. L., & Rhee, J. (2020). Prediction Model for Type 2 Diabetes using Stacked Ensemble Classifiers. 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain. DOI: 10.1109/DASA51403.2020.9317090.
19. Park, J., Surabhi, V. R., Krishnamurthy, P., Garg, S., Karri, R., & Khorrami, F. (2020). Anomaly Detection in Embedded Systems Using Power and Memory Side Channels. 2020 IEEE European Test Symposium (ETS), Tallinn, Estonia. DOI: 10.1109/ETS48528.2020.9131596.
20. Meneghetti, L., Terzi, M., Del Favero, S., Susto, G. A., & Cobelli, C. (2020). Data-Driven Anomaly Recognition for Unsupervised Model-Free Fault Detection in Artificial Pancreas. IEEE Transactions on Control Systems Technology, 1, 33-47. doi: 10.1109/TCST.2018.2885963.
21. Dridi, A., Boucetta, C., Hammami, S. E., Afifi, H., & Moungla, H. (2021). STAD: Spatio-Temporal Anomaly Detection Mechanism for Mobile Network Management. IEEE Transactions on Network and Service Management, 18(1), 894-906. doi: 10.1109/TNSM.2020.3048131.
22. Liu, C., Yong, S., Wang, X., & Zhang, X. (2020). A Multi-feature Anomaly Detection Method Based on AETA ULF Electromagnetic Disturbance Signal. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (pp. 1103-1108). Chongqing, China. doi: 10.1109/ITNEC4 8623.2020.9085032.
23. Kromer-Edwards, C., Castanheira, M., & Oliveira, S. (2020). Year, Location, and Species Information In Predicting MIC Values with Beta-Lactamase Genes. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1383-1390). Seoul, Korea (South). doi: 10.1109/BIBM49941.2020.9313331.

24. Fang, N., Fang, X., & Lu, K. (2022). Anomalous Behavior Detection Based on the Isolation Forest Model with Multiple Perspective Business Processes. Electronics, 11, 3640. doi: 10.3390/electronics11213640.
25. Chater, M., Borgi, A., Taieb, M. T., Sfar-Gandoura, K., & Landoulsi, M. I. (2022). Fuzzy Isolation Forest for Anomaly Detection. Procedia Computer Science, 207, 916-925. doi: 10.1016/j.procs.2022.09.147.
26. Choudhury, J., & Shi, C. (2022). Enhanced Performance of Finite Boundary Isolation Forest (FBIF) for Datasets with Standard Distribution Properties. In 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-5). Prague, Czech Republic. doi: 10.1109/ICECET 55527.2022.9873022.
27. Reddy, P. R., & Kumar, A. S. (2022). Credit Card Fraudulent Transactions Prediction Using Novel Sequential Transactions by Comparing Light Gradient Booster Algorithm Over Isolation Forest Algorithm. In 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM) (pp. 563-567). Gautam Buddha Nagar, India. doi: 10.1109/ICIPTM 54933. 2022.9754211.
28. Marcelli, E., Barbariol, T., Savarino, V., Beghi, A., & Susto, G. A. (2022). A Revised Isolation Forest Procedure for Anomaly Detection with a High Number of Data Points. In 2022 IEEE 23rd Latin American Test Symposium (LATS) (pp. 1-5). Montevideo, Uruguay. doi: 10.1109/ LATS57337.2022.9936964.
29. Fan, L., Ma, J., Tian, J., Li, T., & Wang, H. (2021). Comparative Study of Isolation Forest and LOF Algorithm in Anomaly Detection of Data Mining. In 2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR) (pp. 1-5). Shanghai, China. doi: 10.1109/ICBAR55169. 2021.00008.
30. Feng, F., Liu, Z., & Zhang, J. (2022). Detection of GPS Abnormal Data of Sanitation Vehicles Based on Isolation Forest Algorithm. In 2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI) (pp. 460-463). Zakopane, Poland. doi: 10.1109/ICDACAI57211.2022.00097J.
31. Su and J. Li, "An Anomaly Detection Algorithm for Multi-dimensional Segmentation Plane Isolation Forest," 2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET), Beijing, China, 2022, pp. 89-93, doi: 10.1109/CCET55412.2022.9906369.
32. Yu, P., & Jia, L. (2022). Wind Power Data Cleaning Based on Autoencoder-Isolation Forest. In 2022 7th International Conference on Power and Renewable Energy (ICPRE) (pp. 803-808). Shanghai, China. doi: 10.1109/ICP RE55555.2022.9960342.
33. Kilinc, H. H. (2022). Anomaly Pattern Analysis Based on Machine Learning on Real Telecommunication Data. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 43-48). Diyarbakir, Turkey. doi: 10.1109/UBMK558 50.2022.9919564.
34. Huang, X., Ren, Y., He, Y., & Chen, Q. (2022). Malicious Models-based Federated Learning in Fog Computing Networks. In 2022 14th International Conference on Wireless Communications and Signal Processing (WCSP) (pp. 192-196). Nanjing, China. doi: 10.1109/WCS P55476.2022.10039266.
35. Kerner, H. R., & Adler, J. B. (2022). Guiding Field Exploration on Earth and Mars with Outlier Detection. In IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium (pp. 5333-5336). Kuala Lumpur, Malaysia. doi: 10.1109/IGARSS468 34.2022.9884366.
36. Zualkernan, I., Ahmed, N., Elmeligy, A., Abdelnaby, A., & Sheta, N. (2022). IoT Sensor Data Consistency using Deep Learning. In 2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS) (pp. 198-203). BALI, Indonesia. doi: 10.1109/IoTaIS56727.2022.9975955.
37. Gajda, J., Kwiecień, J., & Chmiel, W. (2022). Machine learning methods for anomaly detection in computer networks. In 2022 26th International Conference on Methods and Models in Automation and Robotics (MMAR)

(pp. 276-281). Międzyzdroje, Poland. doi: 10.1109/MMAR55195.2022.9874341.
38. Malaek, S. M., & Alipour, E. (2022). Intelligent Flight-Data-Recorders; a Step Toward a New Generation of Learning Aircraft. In 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 1545-1549). Istanbul, Turkey. doi: 10.1109/CoDIT55151.2022.9804136.
39. El Houda, Z. A., Hafid, A. S., & Khoukhi, L. (2021). A Novel Machine Learning Framework for Advanced Attack Detection using SDN. In 2021 IEEE Global Communications Conference (GLOBECOM) (pp. 1-6). Madrid, Spain. doi: 10.1109/GLOBECOM46510.2021.9685643.
40. Kafi, H. M., Miah, A. S. M., Shin, J., & Siddique, M. N. (2022). A Lite-Weight Clinical Features Based Chronic Kidney Disease Diagnosis System Using 1D Convolutional Neural Network. In 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE) (pp. 1-5). Gazipur, Bangladesh. doi: 10.1109/ICAEEE54957.2022.9836398.
41. Anello, E., et al. (2022). Anomaly Detection for the Industrial Internet of Things: an Unsupervised Approach for Fast Root Cause Analysis. In 2022 IEEE Conference on Control Technology and Applications (CCTA) (pp. 1366-1371). Trieste, Italy. doi: 10.1109/CCTA49430.2022.9966158.
42. Yang, X., Zhuang, Y., Shi, M., Cao, X., Chen, D., & Tang, Y. (Year). SPiForest: An Anomaly Detecting Algorithm Using Space Partition Constructed by Probability Density-Based Inverse Sampling. IEEE Transactions on Neural Networks and Learning Systems. doi: 10.1109/TNNLS.2022.3223342.
43. Sugunaraj, N., Vrtis, J., Snyder, J., & Ranganathan, P. (Year). False Data Injection Modeling & Detection for Phasor Measurment Units. In 2022 North American Power Symposium (NAPS) (pp. 1-6). Salt Lake City, UT, USA. doi: 10.1109/NAPS56150.2022.10012137.
44. Premisha, P., Prasanth, S., Kanagarathnam, M., & Banujan, K. (Year). An Ensemble Machine Learning Approach for Stroke Prediction. In 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE) (pp. 165-170). Colombo, Sri Lanka. doi: 10.1109/SCSE56529.2022.9905215.
45. Wang, G., Mao, X., Zhang, Q., & Lu, A. (Year). Fatigue Detection in Running with Inertial Measurement Unit and Machine Learning. In 2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB) (pp. 85-90). Hangzhou, China. doi: 10.1109/ICBCB55259.2022.9802471.
46. Singh, R., & Deorari, R. (Year). Enhancing Collaborative Intrusion detection networks against insider attack using supervised learning technique. In 2022 IEEE 2nd Mysore Sub Section International Conference (Mysuru Con) (pp. 1-6). Mysuru, India. doi: 10.1109/MysuruCon55714.2022.9972599.
47. Yang, T., Cai, Z., Hou, B., & Zhou, T. (Year). 6Forest: An Ensemble Learning-based Approach to Target Generation for Internet-wide IPv6 Scanning. In IEEE INFOCOM 2022 -IEEE Conference on Computer Communications (pp. 1679-1688). London, United Kingdom. doi: 10.1109/INFOCOM48880.2022.9796925.
48. Mouret, F., Albughdadi, M., Duthoit, S., Kouamé, D., & Tourneret, J.-Y. (Year). Robust Estimation of Gaussian Mixture Models Using Anomaly Scores and Bayesian Information Criterion for Missing Value Imputation. In 2022 30th European Signal Processing Conference (EUSIPCO) (pp. 827-831). Belgrade, Serbia. doi: 10.23919/EUSIPCO55093.2022.9909815.
49. Mensi, A., Bicego, M. (Year). Enhanced anomaly scores for isolation forests. Pattern Recognition, Volume 120, December 2021, 108115. doi: doi.org/10.1016/j.patcog.2021.108115.
50. Badurowicz, M., Karczmarek, P., & Montusiewicz, J. (Year). Fuzzy Extensions of Isolation Forests for Road Anomaly Detection. In 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-6). Luxembourg, Luxembourg. doi: 10.1109/FUZZ45933.2021.9494469.

51. Ahmed, T., Shah, A., Kolla, M., & Yellasiri, R. (Year). Reduction of Alert Fatigue using Extended Isolation Forest. In 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS) (pp. 1-5). Bengaluru, India. doi: 10.1109/FABS52071.2021.9702617.
52. Priyanto, C. Y., Hendry, & Purnomo, H. D. (Year). Combination of Isolation Forest and LSTM Autoencoder for Anomaly Detection. In 2021 2nd International Conference on Innovative and Creative Information Technology (ICITech) (pp. 35-38). Salatiga, Indonesia. doi: 10.1109/ICITech50181.2021.9590143.
53. Derse, C., Baghdadi, M., Hegazy, O., Sensoz, U., Gezer, H. N., & Nil, M. (Year). An Anomaly Detection Study on Automotive Sensor Data Time Series for Vehicle Applications. In 2021 Sixteenth International Conference on Ecological Vehicles and Renewable Energies (EVER) (pp. 1-5). Monte-Carlo, Monaco. doi: 10.1109/EVER52347.2021.9456629.
54. Damodaran, S., Padmanabhan, R., Maahin, R., & Gurugopinath, S. (Year). A Copula-Driven Unsupervised Learning Framework for Anomaly Detection with Multivariate Heterogeneous Data. In 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). Gold Coast, Australia. doi: 10.1109/MLSP52302.2021.9596359.
55. Karthik, S., Supreetha, H. V., & Sandhya, S. (Year). Detection of Anomalies in Time Series Data. In 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) (pp. 1-5). Bangalore, India. doi: 10.1109/CSITSS54238.2021.9683715.
56. Lopes, A. P., Parshionikar, S., Kale, A., Sharma, N., & Varghese, A. A. (Year). Comparative Analysis of Deep Learning Techniques For Credit Card Fraud Detection. In 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3) (pp. 1-5). Mumbai, India. doi: 10.1109/ICAC353642.2021.9697205.
57. Komárek, T., Brabec, J., Škarda, Č., & Somol, P. (Year). Threat Hunting as a Similarity Search Problem on Multi-positive and Unlabeled Data. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 2098-2103). Orlando, FL, USA. doi: 10.1109/BigData52589.2021.9671958.
58. McKinnon, C., Carroll, J., McDonald, A., Koukoura, S., & Plumley, C. (Year). Investigation of anomaly detection technique for wind turbine pitch systems. The 9th Renewable Power Generation Conference (RPG Dublin Online 2021) (pp. 277-282). Online Conference. doi: 10.1049/icp.2021.1401.
59. Witzig, P., Upenik, E., & Ebrahimi, T. (Year). Open-Set Person Re-Identification Through Error Resilient Recurring Gallery Building. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 245-249). Anchorage, AK, USA. doi: 10.1109/ICIP42928.2021.9506241.
60. Chen, R., Yang, Y., & Xia, M. (Year). Anomaly Detection of Sensor Data Based on 1D Depth Separable Dilated Convolution Neural Network. In 2021 International Conference on Networking, Communications and Information Technology (NetCIT) (pp. 240-244). Manchester, United Kingdom. doi: 10.1109/NetCIT54147.2021.00055.
61. Lu, G., Duan, C., Zhou, G., Ding, X., & Liu, Y. (Year). Privacy-Preserving Outlier Detection with High Efficiency over Distributed Datasets. IEEE INFOCOM 2021 - IEEE Conference on Computer Communications (pp. 1-10). Vancouver, BC, Canada. doi: 10.1109/INFOCOM42981.2021.9488710P.
62. Ntambu and S. A. Adeshina, "Machine Learning-Based Anomalies Detection in Cloud Virtual Machine Resource Usage," 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, 2021, pp. 1-6, doi: 10.1109/ICMEAS52683.2021.9692308.
63. Kotsiopoulos, T., "Fault Detection on Bearings and Rotating Machines based on Vibration Sensors Data," (2021) 2021 IEEE International Conference on Progress in Informatics and Computing (PIC), Shanghai, China, pp. 474-483. doi: 10.1109/PIC53636.2021.9686999.

64. Liu, W., "D2MIF: A Malicious Model Detection Mechanism for Federated-Learning - Empowered Artificial Intelligence of Things," (2023) IEEE Internet of Things Journal, 10(3), 2141-2151. doi: 10.1109/JIOT.2021. 3081606.
65. Shrivastava, R., "Comparative study of boosting and bagging based methods for fault detection in a chemical process," (2021) 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, pp. 674-679. doi: 10.1109/ ICAIS50930.2021.9395905.
66. Koch, P., Schekotihin, K., Jannach, D., Hofer, B., & Wotawa, F., "Metric-Based Fault Prediction for Spreadsheets," (2021) IEEE Transactions on Software Engineering, 47(10), 2195-2207. doi: 10.1109/TSE.2019.2944604.
67. Pathak, A. K., Saguna, S., Mitra, K., & Åhlund, C., "Anomaly Detection using Machine Learning to Discover Sensor Tampering in IoT Systems," (2021) ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, pp. 1-6. doi: 10.1109/ ICC42927.2021.9500825.
68. Alghanmi, A., Yunusa-Kaltungo, A., & Edwards, R., "A Comparative Study of Faults Detection Techniques on HVAC Systems," (2021) 2021 IEEE PES/IAS PowerAfrica, Nairobi, Kenya, pp. 1-5. doi: 10.1109/Power Africa52236.2021.9543158.
69. Antony, L., "A Comprehensive Unsupervised Framework for Chronic Kidney Disease Prediction," (2021) IEEE Access, 9, 126481-126501. doi: 10.1109/ACCESS.2021.3109168.
70. Mansour, R. F., Amraoui, A. E., Nouaouri, I., Díaz, V. G., Gupta, D., & Kumar, S., "Artificial Intelligence and Internet of Things Enabled Disease Diagnosis Model for Smart Healthcare Systems," (2021) IEEE Access, 9, 45137-45146. doi: 10.1109/ACCESS.2021. 306 6365.
71. Simmini, F., Rampazzo, M., Peterle, F., Susto, G. A., & Beghi, A., "A Self-Tuning KPCA-Based Approach to Fault Detection in Chiller Systems," (2022) IEEE Transactions on Control Systems Technology, 30 (4), 1359-1374, July. doi: 10.1109/TCST. 2021. 3107200.
72. Chen, J., Zhang, J., Qian, R., Yuan, Junfeng, & Ren, Y. (2023). An Anomaly Detection Method for Wireless Sensor Networks Based on the Improved Isolation Forest. Applied Sciences, 13, 702. doi: 10.3390/app13020702.
73. Almansoori, M., & Telek, M. (2023). Anomaly Detection using a combination of Autoencoder and Isolation Forest. In 1st Workshop on Intelligent Infocommunication Networks, Systems and Services (WI2NS2) (pp. 25-30). Budapest. doi: 10.3311/WINS2023-005.
74. Utkin, L., Ageev, A., Konstantinov, A., & Muliukha, V. (2023). Improved Anomaly Detection by Using the Attention-Based Isolation Forest. Algorithms, 16, 19. https:// doi.org/10.3390/a16010019
75. Kabir, S., Shufian, A., & Zishan, M. S. R. (2023). Isolation Forest Based Anomaly Detection and Fault Localization for Solar PV System. 2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, pp. 341-345. doi: 10.1109/ICREST57604. 2023.10 070033.
76. Baviskar, P. V., Singh, G., & Patil, V. N. (2023). Design of Machine Learning-Based Malware Detection Methodologies in the Internet of Things Environment. 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, pp. 1-6. doi: 10.1109/ICONAT57137.2023.10080517.
77. Himeur, Y., Fadli, F., & Amira, A. (2022). A Two-Stage Energy Anomaly Detection for Edge-based Building Internet of Things (BIoT) Applications. 2022 5th International Conference on Signal Processing and Information Security (ICSPIS), Dubai, United Arab Emirates, pp. 180-185. doi: 10.1109/ICS PIS57063.2022.10002641.
78. Anand, N., & Saifulla, M. A. (2023). An efficient IDS for slow rate HTTP/2.0 DoS Attacks using one class classification. 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, pp. 1-9. doi: 10.1109/I2CT57861.2023. 10126162.
79. Bannur, C., Bhat, C., Singh, K., Kulkarni, S. A., & Doddamani, M. (2023). PAACDA: Comprehensive Data Corruption Detection Algorithm. IEEE Access, 11, 24908-24934. doi:10.1109/ ACCESS.2023.3253022.

80. Buschjäger, S., Honysz, P. -J., & Morik, K. (2020). Generalized Isolation Forest: Some Theory and More Applications Extended Abstract. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, pp. 793-794. doi: 10.1109/DSAA49011.2020.00120.
81. Xu, D., Wang, Y., Meng, Y., & Zhang, Z. (2017). An Improved Data Anomaly Detection Method Based on Isolation Forest. 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, pp. 287-291. doi: 10.1109/ISCID.2017.202.
82. Wielgosz, M., Skoczen, A., & Wiatr, K. (2018). Looking for a Correct Solution of Anomaly Detection in the LHC Machine Protection System. 2018 International Conference on Signals and Electronic Systems (ICSES), Kraków, Poland, pp. 257-262. doi: 10.1109/ICSES.2018.8507291Y.
83. Qin, Y., & Lou, Y. (2019). Hydrological Time Series Anomaly Pattern Detection based on Isolation Forest. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, pp. 1706-1710. doi: 10.1109/ITNEC.2019.8729405.
84. Luo, S., Luan, L., Cui, Y., Chai, X., Wang, Z., & Kong, Y. (2019). An Attribute Associated Isolation Forest Algorithm for Detecting Anomalous Electro-data. 2019 Chinese Control Conference (CCC), Guangzhou, China, pp. 3788-3792. doi: 10.23919/ChiCC.2019.8866495.
85. Ding, Z., Mo, Y., & Pan, Z. (2019). A Novel Software Defect Prediction Method Based on Isolation Forest. 2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2 MSE), Zhangjiajie, China, pp. 882-887. doi: 10.1109/QR2MSE46217.2019.9021215.
86. Shriram, S., & Sivasankar, E. (2019). Anomaly Detection on Shuttle data using Unsupervised Learning Techniques. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, pp. 221-225. doi: 10.1109/ICCIKE47802.2019.9004325.
87. Baloglu, U. B., Alatas, B., & Bingol, H. (2019). Assessment of Supervised Learning Algorithms for Irony Detection in Online Social Media. 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, pp. 1-5. doi: 10.1109/UBMYK48245.2019.8965 580.
88. Shu, H., Zhao, X., Luo, H., & Li, C. (2019). Research on Stacking-Based Integrated Algorithm of Anomaly Detection in Production Process. 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), Shenzhen, China, pp. 85-90. doi: 10.1109/HPBDIS.2019.8735447.
89. Iheme, L.O., & Ozan, Ş. (2019). Feature Selection for Anomaly Detection in Call Center Data. 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, pp. 926-929. doi: 10.23919/ELECO47770.2019.8990454.
90. Komárek, T., Brabec, J., Škarda, Č., & Somol, P. (2021). Threat Hunting as a Similarity Search Problem on Multi-positive and Unlabeled Data. 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, pp. 2098-2103. doi: 10.1109/BigData52589.2021.9671958.
91. Liu, W., He, J., Han, S., Cai, F., Yang, Z., & Zhu, N. (Year). A Method for the Detection of Fake Reviews Based on Temporal Features of Reviews and Comments.
92. Lv, Y., Cui, Y., Zhang, X., Cai, M., Gu, X., & Xiong, Z. (2019). A New Outlier Detection Method Based on Machine Learning. 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, pp. 1-7. doi: 10.1109/ICSIDP47821.2019.9173217.
93. Sepulvene, L. H. M., et al. (2019). Analysis of Machine Learning Techniques in Fault Diagnosis of Vehicle Fleet Tracking Modules. 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), Salvador, Brazil, pp. 759-764. doi: 10.1109/BRACIS.2019.00136.
94. Portela, F. G., Mendoza, F. A., & Benavides, L. C. (2019). Evaluation of the performance of

supervised and unsupervised Machine learning techniques for intrusion detection. 2019 IEEE International Conference on Applied Science and Advanced Technology (iCASAT), Queretaro, Mexico, pp. 1-8. doi: 10.1109/iCASAT48251.2019.9069538