IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 June 2023, accepted 3 July 2023, date of publication 10 July 2023, date of current version 17 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294096

SURVEY

Detecting Outliers in Non-IID Data: A Systematic Literature Review

SHAFAQ SIDDIQI[1], FAIZA QURESHI[2], STEFANIE LINDSTAEDT[1], AND ROMAN KERN[1]

[1]Faculty of Computer Science and Biomedical Engineering, Institute of Interactive Systems and Data Science, Graz University of Technology (TU Graz), 8010 Graz, Austria

[2]Department of Computer Science, Dhanani School of Science and Engineering, Habib University, Karachi 75290, Pakistan

Corresponding author: Shafaq Siddiqi (shafaq.siddiqi@tugraz.at)

ABSTRACT Outlier detection (outlier and anomaly are used interchangeably in this review) in non-independent and identically distributed (non-IID) data refers to identifying unusual or unexpected observations in datasets that do not follow an independent and identically distributed (IID) assumption. This presents a challenge in real-world datasets where correlations, dependencies, and complex structures are common. In recent literature, several methods have been proposed to address this issue and each method has its own strengths and limitations, and the selection depends on the data characteristics and application requirements. However, there is a lack of a comprehensive categorization of these methods in the literature. This study aims to systematically review outlier detection methods for non-IID data published between 2015 and 2023. This study focuses on three major aspects; data characteristics, methods, and evaluation measures. In data characteristics, we discuss the differentiating properties of non-IID data. Then we review the recent methods proposed for outlier detection in non-IID data, covering their theoretical foundations and algorithmic approaches. Finally, we discuss the evaluation metrics proposed to measure the performance of these methods. Additionally, we present a taxonomy for organizing these methods and highlight the application domain of outlier detection in non-IID categorical data, outlier detection in federated learning, and outlier detection in attribute graphs. We provide a comprehensive overview of datasets used in the selected literature. Moreover, we discuss open challenges in outlier detection for non-IID to shed light on future research directions. By synthesizing the existing literature, this study contributes to advancing the understanding and development of outlier detection techniques in non-IID data settings.

INDEX TERMS Outlier detection, non-IID data, anomaly detection, heterogeneous data, data dependency.

I. INTRODUCTION

Outlier detection, also used interchangeably as anomaly detection, is crucial in various domains such as data streams, computer network security, medical diagnosis, and finance [1], [2], [3], [4], [5]. It aims to identify instances in the data that deviate significantly from most instances and are considered unusual or rare. In the era of big data, existing outlier detection methods face two main challenges. First, the non-stationary nature of big data renders the existing detection methods useless to capture the underlying relationship and couplings in the data (non-independent data) [6]. Second, the variety dimension in big data requires careful handling of diverse data types generated from heterogeneous distributions for some specific application domains (non-identically distributed) [7].

Outlier detection is particularly important for non-independent and identically distributed (non-IID) data, due to the fact in real-world applications the data is correlated (i.e., sensors within the same geo-location are likely to have correlated data), heterogeneous (in federated learning different devices can hold different amounts of data with same or different distributions), causally connected (treatments graph data), or skewed in distribution [6], [8], [9]. This type of data presents a considerable challenge for the existing outlier detection methods with independent and identically distributed (IID) assumptions [7].

To address this challenge, various outlier detection methods have been proposed for non-IID data. These methods can be classified into two main categories; supervised and

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

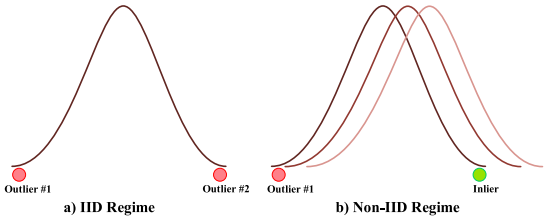


FIGURE 1. Changing definition of outliers in concept drift.

unsupervised [10], [11]. These methods can be further classified into ensemble methods and individual methods [7], [12]. Supervised methods use the labeled data and apply statistical tests to identify distribution shifts in the data and then update the model weights or add a new model into the model pool. The unsupervised methods mostly rely on One-Class Support Vector Machines (OCSVM) and clustering methods to identify normal and abnormal samples [8], [11].

In addition to selecting appropriate outlier detection methods, another critical aspect in detecting outliers in non-IID data is a concept drift adaptation. An important characteristic of non-stationary, non-IID data in the real world is the definition of normal behavior, that keeps changing over time. In other words, the ''concept'' of normal changes with time [13] so does the definition of anomaly. In such cases, the outlier detection models are required to adapt to the changing concept of normal behavior [14]. Hence, the ability of a model to adapt to the changing concept is called concept drift adaptation [15], [16], [17]. Detection and adaptation of concept drift is a challenge in the context of non-IID data, but differentiation between the expected concept drift and outliers makes it more challenging [12], [15]. Figure 1 shows how the definition of an outlier changes with concept drift.

Furthermore, selecting the appropriate performance measure for outlier detection methods for non-IID data is also critical, the commonly used evaluation measures for both IID and non-IID data are accuracy, precision, recall, F-measure, and Area Under The Curve (AUC) and Receiver Operating Characteristics (ROC) curve. Besides them, there are a few specialized metrics for evaluating certain characteristics of non-IID data such as coupling strength, and coupled distance [18], [19], [20].

The main goal of this review is to explore the three main aspects of the outlier detection problem in non-IID data which are 1) characteristics of non-IID data, 2) algorithms developed to detect outliers in non-IID data, and 3) evaluation measures to evaluate the performance of the proposed algorithm. Based on these factors we propose a taxonomy in Figure 2 to guide the literature search and inclusion/exclusion criteria for the searched literature. This study provides a summary of the most recent research in the field and identifies open research challenges and future directions for outlier detection in non-IID data. The main contributions of this paper are:

- This paper provides a systematic review of outlier detection techniques for non-IID data proposed in the last eight years.
- This paper provides a guided review by proposing a taxonomy of outlier detection for non-IID data.
- This paper provides a comprehensive overview of datasets used to evaluate the outlier detection algorithms for non-IID data.
- This paper highlights the challenges and future direction for outlier detection in non-IID data settings.

II. BACKGROUND

A. OUTLIER DETECTION

Outliers are defined as data points that deviate significantly from the expected behavior of a system [10]. Outlier detection is a task of identifying data points or instances in a dataset that are significantly shifted from the majority of the data [1], [21], [22], [23], [24]. It is an important issue in various domains such as finance, network security, and healthcare [25], [26], [27], [28]. Outlier detection plays a crucial role in various domains due to its significance in identifying abnormal or anomalous instances within a dataset. In data streams, where data is continuously generated and analyzed in real-time, outlier detection helps identify sudden changes or anomalies that may indicate important events or anomalies in the underlying process. In computer network security, outlier detection helps identify malicious activities or intrusions by detecting abnormal network behaviors [26]. In medical diagnosis, outlier detection can aid in identifying rare diseases, unusual patient conditions, or outliers in medical imaging data. In finance, outlier detection is valuable for identifying fraudulent transactions, unusual market behaviors, or anomalies in financial data that may indicate potential risks or opportunities [25]. Overall, outlier detection provides valuable insights and helps maintain the integrity, security, and accuracy of data analysis in various domains.

A popular mathematical formulation of outlier detection is based on the concept of the decision boundary. It is a boundary in the feature space that separates the normal data points from the outliers. An outlier score can be assigned to each data point based on its distance from the decision boundary. The data points with high outliers scores (i.e., greater than a threshold value) are considered to be outliers [29]. The decision boundary can be determined by various methods including statistical methods, density-based methods, and distance-based methods. Statistical methods calculate the mean and standard deviation of the data and use these values to determine a threshold value. Then the points outside of this threshold are considered outliers.

In the context of outlier detection, it is important to differentiate between outliers and disturbances. Outliers refer to data points that significantly deviate from the expected patterns or norms, indicating abnormal behavior or events. They can be indicative of critical incidents, anomalies, or fraudulent activities. On the other hand, disturbances refer to temporary or transient fluctuations in the data that do not necessarily indicate abnormal behavior. These disturbances can arise from random noise, measurement errors, or other non-anomalous factors. The distinction between outliers and

disturbances is crucial because outlier detection aims to identify true anomalies that require attention or investigation, while temporary disturbances are expected to not have significant implications. By accurately distinguishing between the two, outlier detection algorithms can minimize false positives and focus on detecting truly exceptional events [27]. In this paper, we focus on those works where the authors proposed a technique for finding true outliers.

The concentration assumption states that normal or non-anomalous data points are more densely clustered or concentrated in a certain region of the data space compared to outliers [30], [31]. This assumption implies that outliers are relatively rare and have a lower density or concentration in the data distribution. By leveraging this assumption, outlier detection algorithms can identify outliers by detecting data points that deviate significantly from the expected concentration of normal data. Density-based methods estimate the probability density of the data and identify data points with low density as outliers. These methods often calculate the distance between each data point and its nearest neighbors/centroids and consider data points with large distances (i.e., non-density reachable) as outliers [32], [33].

Recent popular approaches to outlier detection are based on supervised, semi-supervised or unsupervised machine learning algorithms. In supervised learning, machine learning models are trained on a labeled dataset to learn the patterns of normal and anomalous behavior. Once trained, these algorithms can then be used to detect outliers in new, unlabeled data. Semi-supervised or unsupervised learning is performed when the labels are few or unavailable. For unsupervised learning, the model is trained on unlabeled data without any specific target variable. The goal is to discover patterns, structures, or relationships within the data itself to differentiate between normal and anomalous samples. Semi-supervised learning combines the advantages of both labeled and unlabeled data. It leverages the labeled data to learn from the provided target labels and the unlabeled data to capture the underlying structure or distribution of the data. Common machine learning algorithms for outlier detection include decision trees, random forests, clustering, and one-class support vector machines [7], [11], [34].

B. APPLICATION DOMAINS

In recent years, outlier detection has become crucial to many domains such as healthcare, fraud detection, manufacturing, intrusion, detection, environmental monitoring, supply chain management, marketing, etc. In the healthcare field outlier detection is being used for an early diagnosis of diseases i.e., detecting unusual patterns in a patient's vital signs or medical records can help doctors identify potential health issues. Moreover, it could also be used to detect health insurance fraud [3], [35].

Similarly in fraud detection, outlier detection methods have been widely applied to detect fraudulent transactions. For example, credit card transactions that are far from the usual behavior of a customer are flagged as potential fraud [2], [4], [28]. Outlier detection is also equally being used in network security for detecting cyber-attacks (i.e., detecting unusual network traffic patterns to detect DDoS attacks) [5].

Likewise, outlier detection methods are also applied in the manufacturing industry that can help in identifying defects in products. For example, detecting outliers in the measurements of a component can help in detecting deviations in the manufacturing process [36].

In the same way, it is also being used in environmental monitoring for early warning of natural disasters. For instance, detecting unusual patterns in measuring environmental parameters like temperature, humidity, and wind speed can help predict natural disasters [37].

The prominent role of outlier detection is not limited to the mentioned application domains and portrayed scenarios but it can be concluded that outlier detection is a critical step in many different areas for identifying potential issues or opportunities.

C. NON-IID OUTLIER DETECTION

It is essential to comprehend that real-world systems and data often do not conform to the classic IID assumption; that is data or variables are independent and identically distributed, drawn from a given distribution. Non-IID encompasses various settings beyond the IID assumption, including inter-dependencies, correlations, heterogeneity, and non-stationarity across variables, sources, time, space, and modeling processes, indicating interactions, coupling, and diverse distributions or relationships that can lead to more complex and challenging learning environments.

For example, social networks exhibit inter-dependencies among users. The dependency analysis could help identify outliers (i.e., spam, phishing attacks) if a user suddenly starts receiving unusually high messages from previously unconnected users. Stocks, influenced by market trends, industry-specific factors, and macroeconomic indicators, often exhibit correlations in their price behavior. By incorporating these correlations, an outlier detection algorithm can effectively distinguish between true anomalies and normal market fluctuations. Patient data collected from multiple hospitals or clinics may exhibit heterogeneity in a healthcare setting. Each healthcare facility may have different patient populations, medical protocols, or data collection practices. Outlier detection in this scenario must account for the heterogeneity across healthcare providers and adjust its detection thresholds accordingly. In streaming data from an industrial plant, non-stationarity arises from changing processes due to maintenance, failures, or operating conditions. An effective outlier detection approach should adapt to these dynamic patterns and identify anomalies reflecting the current plant state.

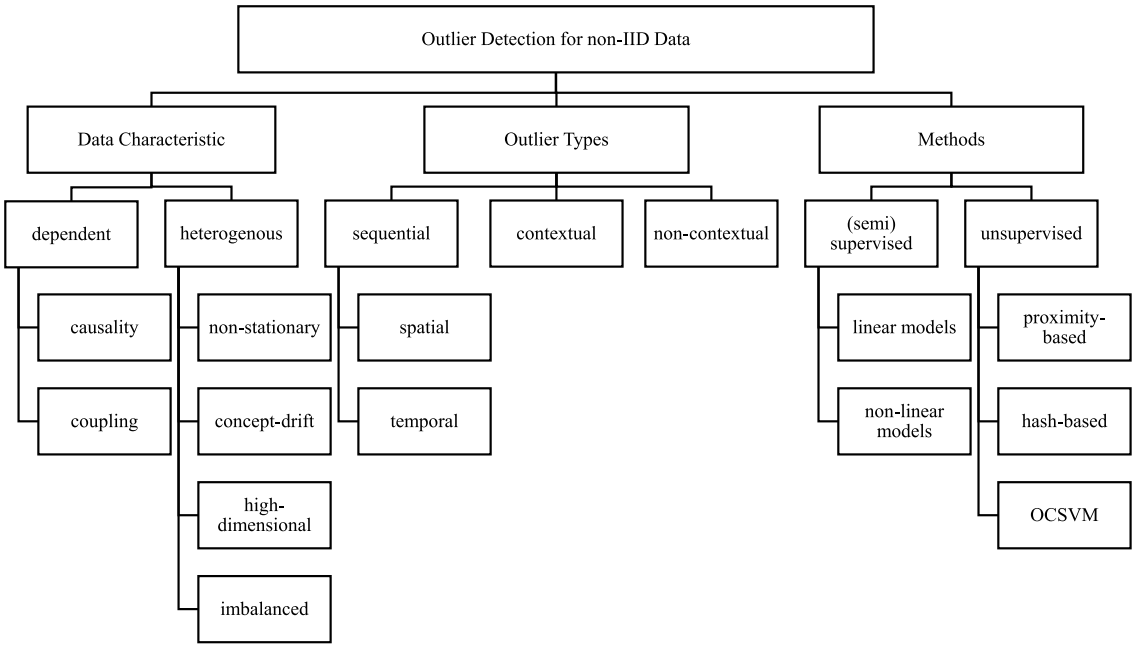


FIGURE 2. Taxonomy for outlier detection in non-IID data.

Non-IID data problem refers to any couplings and heterogeneity that exist within and between two or more aspects, such as entities, objects, inter-attribute, intra-attribute or attribute-value and state of affairs, prior to, during, and after a learning task [38].

The objective of non-IID outlier detection is to recognize and describe the non-IIDness in values, features, labels, or contexts, and to differentiate between inlying and outlying features, objects, or labels. It is also necessary to identify the outlying dynamics during the data analysis process and integrate them into outlier scoring.

Some of the exemplary perspectives in non-IID outlier detection include learning value and feature couplings for outliers scoring. These techniques aim to better understand the relationships between variables and features in non-IID data and use this knowledge to improve the detection of outliers [39]. Considering the importance of data understanding we will discuss the characteristics of non-IID data in detail in Section IV.

D. CONCEPT DRIFT

Concept drift is a major characteristic of non-IID data. It is the change in the data distribution that occurs over time. Given a window $W = [T, (X, Y)]$ with timestamp T, features X and labels Y in data stream S. Where $T = \{t_i, t_{i+1}, \ldots, t_n\}$ and $(X, Y) = \{(x_i, y_i),$
$(x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$, concept drift is defined as following [40],

$$\exists t : p(X_t, Y_t) \neq p(X_{t+1}, Y_{t+1})$$

where $p(X_t, Y_t)$ denotes the joint distribution of data features and labels at time t. Changes in data at time $t + 1$ can be described as changes in prior probabilities of classes $p(y)$, changes in class conditional probability $p(X|y)$, or changes in posterior probability $p(y|X)$. Moreover, the authors in [41] defined three potential sources of concept drift:

1) change in posterior probabilities, an actual drift
2) change in prior probabilities, a virtual drift
3) change in both posterior and prior probabilities or rigorous drift

In literature, there exist several approaches for detecting concept drift, including statistical methods, supervised learning methods, unsupervised learning methods, and ensemble methods.

Statistical methods: Techniques such as hypothesis testing or Time distribution methods can be used to identify significant changes in the data distribution. Time distribution-based methods (such as Kullback–Leibler (KL) divergence, and Jensen–Shannon (JS) divergence) calculate the difference between two probability distributions to detect concept drift.

Supervised learning methods: Supervised learning methods can be used to train a model on a labeled dataset and then monitor its performance over time. If the model's performance drops below a certain threshold, it may indicate that concept drift has occurred.

Unsupervised learning methods: Unsupervised learning methods can be used to identify clusters in the data and track their evolution over time. If the clusters shift significantly, it may indicate concept drift.

Ensemble methods: Ensemble methods such as stacked generalization can be used to combine multiple models trained on different time periods. The performance of the ensemble can be monitored over time, and a drop in performance may indicate concept drift.

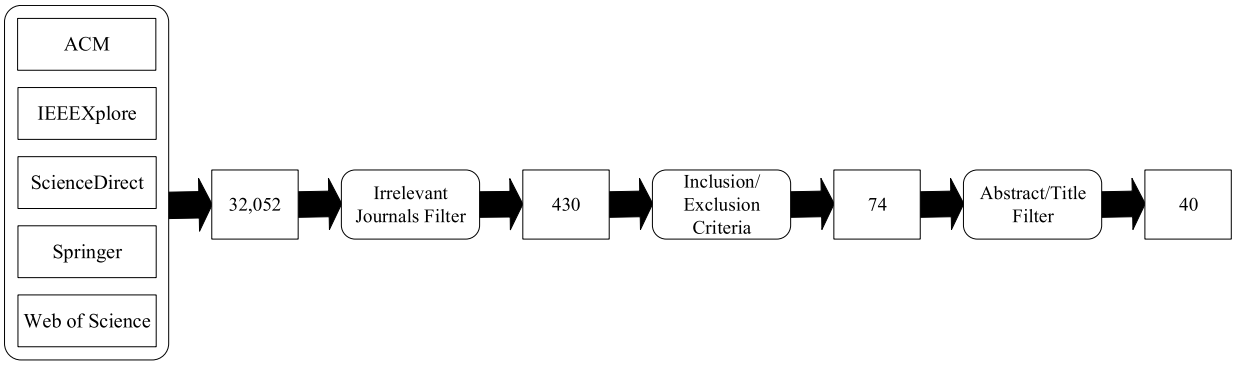


FIGURE 3. Selection process of articles.

Hence, our focus in this study is on the techniques that include the detection and adaption of concept drift for outlier detection in non-IID data.

E. ONLINE AND OFFLINE MACHINE LEARNING

Offline learning, or batch processing, analyzes the entire dataset for comprehensive analysis and accurate outlier detection. It suits static datasets, but struggles with concept drift and real-time detection, being computationally intensive and unsuitable for non-stationary data. While online learning, or streaming/incremental learning, enables real-time detection and adaptability to changing data patterns. It processes data incrementally, making it ideal for streaming and dynamic non-IID data. Online learning efficiently handles large-scale datasets, providing timely outlier detection, although it may lack access to historical data and long-term patterns, being more sensitive to noise and fluctuations. Resource limitations and optimization challenges are also factors. The choice depends on specific requirements. Offline learning ensures thorough analysis but falters in dynamic environments, while online learning excels in real-time detection but sacrifices historical analysis and faces challenges in handling incremental updates. Factors like dataset characteristics, historical data availability, real-time requirements, and adaptability to changing patterns influence the selection process.

III. METHODOLOGY

To explore the recent developments in outlier detection for non-IID data, we conducted a systematic literature review of studies published between 2015 and 2023. This review will help the scientific community and relevant application domain experts to get an in-depth understanding of non-IID outlier detection in terms of its data characteristic, algorithms and evaluation measures. Our approach adheres to the methodology proposed by [42] and [43], which involves retrieving research papers from the existing literature, selecting relevant works, and summarizing them. This systematic literature review process ensures the reproducibility of the results and minimizes selection biases towards specific works in the literature. We also propose a taxonomy in Figure 2 to guide the readers about the prominent aspects of this review. In sections III-A, III-B, and III-C, we present our research questions, search strategy, and study selection criteria, respectively.

A. RESEARCH QUESTION

Following are the research questions that are addressed by this review, and guided the search and selection process of our study:

- Q_1 What are the differentiating characteristics and avaiable sources of non-IID data?
- Q_2 How have researchers addressed the challenges posed by non-IID data in outlier detection (existing methods)?
- Q_3 What are the most commonly used evaluation metrics for outlier detection in non-IID settings?
- Q_4 What are some of the most promising opportunities in this area (open challenges)?

B. SEARCH SOURCES AND METHODS

The literature included in this review comprises research articles sourced from reputable venues in the fields of knowledge discovery, machine learning, and artificial intelligence. These venues include ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, and Web of Science. To gather relevant studies, specific search terms were developed based on the taxonomy and main research question. These queries were then utilized in electronic scientific libraries to retrieve relevant published works. This approach ensured a comprehensive and targeted collection of articles for analysis. Our search focus on the three dimensions related to outlier detection 1) Term 1: we define the keywords for capturing the distinguishing characteristics of non-IID data. 2) Term 2: we use alternatives of the term ''outlier'' and 3) Term 3: we use the terms to define a process, evaluation or detection for outlier detection Table 1 shows the key terms we used for building a search query for literature retrieval.

C. QUERY FORMATION

To construct a search query, we combine the terms listed in Table 1 using ''OR'' and ''AND'' Boolean operators to build conjunction and disjunction queries. First, we select

TABLE 1. Key terms for literature search.

Term1	Term 2	Term 3
Non-IID	Outlier	Detection
Non-stationary	Anomaly	Feature selection
Heterogeneous	Non-parametric	Statistical tests
Value coupling	Data discord	Error analysis
High-dimensional	Novelty	identification
Complex data		Instance selection
Data shifts		
Concept drift		
Imbalanced		

each term from column "Term 1" and join it with a term in the second column "Term 2" using AND boolean operator. Second, we repeat this same step and combine the results of "Term 1" and "Term 2" with "Term 3" using AND boolean operator. All the terms in each column of Table 1 are combined using OR boolean operator. Alongside the exact terms defined in Table 1, we also consider the plural form of these terms. For instance, we replace "drift" with "drifts", "outlier" with "outliers", and so forth. We split the hyphenated compound words using space i.e., "non stationary", "High dimensional", and "non iid". This process results in 315 search queries. Some of the digital libraries (i.e., IEEEXplore, ScienceDirect) provide sophisticated ways to write complex search queries which allows us to combine multiple small search queries into a long query below is an example of a such complex query,

(Title: outlier OR Title: anomaly OR Title: data-discord OR Title: data noise) AND (Title: "non-iid" OR Title: "value-coupling" OR Title: "skewed-data" OR Title: "non-stationary")

D. STUDY SELECTION

The extraction of relevant articles is performed in four steps as shown in figure 3. In the first step, we applied our queries to the digital libraries and we retrieved 32,052 articles in total. To get the most relevant articles from our search, we specifically applied filters on Springer and ScienceDirect libraries journals. This was necessary as the initial search results based on our key terms also included studies from unrelated domains. The list of included and excluded ScienceDirect Journals can be seen in Table 2. This step resulted in 430 articles. Subsequently, we applied the inclusion and exclusion criteria and a total of 74 articles were selected for further analysis. Finally, we went through the title and abstract of the individual articles and this step resulted in 40 articles for the review. We present 30 articles with differentiating techniques in Section V, while we use the rest of them for our understanding of the domain.

We explain our inclusion and exclusion criteria as follows,

1) INCLUSION CRITERIA

We conducted a comprehensive review of the available literature to identify articles that fulfill the following criteria.

- Articles which contained a comprehensive overview of datasets used in the studies of outlier detection for non-IID data
- Articles which discussed the novel methods proposed for outlier detection in non-IID data
- Articles presenting proposed evaluation metrics to assess the performance of outlier detection methods on non-IID data
- Articles that were published between 2015 and 2023
- Works published in top-tier venues including conferences and journals (see Appendix IX for list of all included venues)
- Long research papers including surveys were preferred over short papers (i.e., abstracts, special issues, scope or summary or tutorial).

2) EXCLUSION CRITERIA

Published works that satisfy any of the following exclusion criteria are removed from this study.

- Studies published in languages other than English
- Studies that are not available via open access or institutional access
- Studies addressing areas other than computer science such as Biomedical Signal Processing and Control, Procedia Engineering, and Thermal Science and Engineering Progress
- Doctoral symposiums, theses, abstracts, workshop reports, and books
- Duplicates

IV. DATA CHARACTERISTICS OF NON-IID DATA

Outlier detection on non-IID data is a challenging task due to complex data characteristics. We emphasized the need for understanding data by making it a separate branch in our taxonomy. In this section, the prominent data characteristics of non-IID data presented in the literature are mentioned.

A. DATA HETEROGENEITY

1) NON-STATIONARY/CONCEPT DRIFT

When the mean, covariance, and correlation of data change over time. This can affect the performance of outlier detection methods based on statistical models that assume stationarity. These models are trained on the first few data samples and then applied to subsequent data assuming that the data is generated from the same distribution. However, real-world data streams are non-stationary where the underlying distribution changes over time [19], [44], [45].

2) HIGH DIMENSIONALITY

When dealing with data that has hundreds of dimensions, traditional methods for detecting outliers are ineffective due to several factors referred to as the "curse of dimensionality" [46]. However, there are various sub-space and full-space

TABLE 2. ScienceDirect journals for literature search.

Included	Excluded
Neurocomputing	Mechanical Systems and Signal Processing
Computers & Security	NDT & International
Artificial Intelligence Medicine	Procedia Engineering
Knowledge-Base Systems	Thermal Science and Engineering Processing
Computer Communications	Biomedical Signal Processing and Control
Information Systems	Computers in Industry
Journal of Computer and System Sciences	Future Generation Computer Systems
Computer Science Review	Advanced Engineering Informatics
Machine Learning with Applications	Digital Communications and Networks

outlier detection techniques that are robust against high-dimensional data [47], [48], [49], [50], [51], [52]. The authors of [53] and [54] provide a comparative analysis of outlier detection methods in high dimensional data streams, and text.

3) IMBALANCED DATA

Real-world scenarios exhibit significant variations in collected data among devices due to user preferences and local environments. This poses a critical challenge, especially in IoT anomaly detection, where different devices may encounter diverse types of attacks or anomalies. It is known that even in cases of balanced datasets, local on-device datasets are typically non-IID, leading to a degradation in model performance. For instance, in non-IID datasets, the drop of 11% and 51% in accuracies of MNIST and CIFAR-10 predictions was reported respectively [55]. This degradation becomes more pronounced when handling imbalanced datasets, as the model tends to favor well-represented classes, leading to biased outcomes. The global model further reinforces this bias by prioritizing patterns from majority clients while suppressing anomalous patterns from minority clients [16], [56], [57].

B. DATA DEPENDENCY

1) DATA COUPLING

The coupling mechanisms are natural linkages between observations that can be found in various domains. For example, observations in social networks can be related in terms of order, meaning, and causality. Such couplings can impact the distribution of data over time and domains, and even the features or random variables. In some cases, the data may follow a seasonal trend, such as clothes sales data. It is important to identify and consider such couplings while analyzing data, as they can affect the performance of machine learning algorithms. By accounting for the coupling mechanisms, we can better understand and model the complex relationships within the data [58], [59].

2) CAUSALITY

Causality is central to the understanding of the data generation process. Without an understanding of cause–effect relationship, we cannot use data to answer questions as basic as "Does this treatment harm or help patients?" [60]. In non-IID data, the observations are not independent of each other and may be linked by various coupling mechanisms. Furthermore, causal inference or causal structure learning [61], [62], [63] can be used to address issues of confounding and selection bias in non-IID data. confounding is when one variable (cause) impacts another variable (effect), when at the same time the confounder influences both of them, introducing confounder bias. Selection bias occurs when the sample is not representative of the population being studied [64]. By comprehending the causal relationships among the variables, it may become possible to account for confounding and selection bias. Hence, one can obtain more accurate estimates of the true causal effects [65], [66]. Causal transformers are also widely used for noisy image classification [67], [68], [69]

V. METHODS FOR OUTLIER DETECTION

In prior research, there have been several methods developed specifically for outlier detection in non-IID data. These methods are designed to address the challenges posed by non-IID data (i.e., data heterogeneity, coupling, and data shift). In Table 3, we present a summary of selected methods and how they fit into our taxonomy. We discuss the selected works in their respective categories in subsections V-A and V-B.

A. UNSUPERVISED LEARNING

Table 3 shows the prominent unsupervised approaches commonly found in the literature for unsupervised outlier detection (in the context of non-IID data) such as once-class support vector machines (OCSVM), autoencoders, isolation forest, and specialized measures for measuring various data characteristics to identify outliers. A brief summary of these approaches is discussed in this section.

1) ONE-CLASS SUPPORT VECTOR MACHINES (OCSVM)

A novel methodology for detecting anomalies in non-stationary data using an ensemble-based self-adaptive One-Class Support Vector Machine (OCSVM) algorithm is proposed in [11]. The model is self-adapting because it chooses the parameter setting for kernel bandwidht and regularization $PS = \{\gamma, \nu\}$ automatically using Quick Model Selection (QMS) [82]. The proposed methodology focused

TABLE 3. Methods for outlier detection in non-IID data.

Paper	Year	Outlier Type	Method	Approach	Learning	Data Characteristic	Evaluation Method
[70]	2016	contextual	unsupervised	proximity-based	offline	coupling	AUC
[11]	2016	sequence	unsupervised	OCSVM	online	non-stationary	AUC
[6]	2017	contextual	unsupervised	proximity based	-	high-dimensional	coupling strength
[10]	2017	sequence	supervised	non-linear	online	non-stationary	TP, TN, FP, FN
[15]	2018	contextual	unsupervised	proximity based	offline	high-dimensional	F-measure
[71]	2018	sequence	supervised	non-linear	online	non-stationary	AUC
[12]	2018	non-contextual	semi-supervised	linear	online	concept-drift	AUC
[18]	2019	sequence	unsupervised	OCSVM	online	concept-drift	F-measure
[72]	2019	non-contextual	unsupervised	proximity-based	online	causality/coupling	NeoDis(x_k, x_q)
[19]	2019	sequence	unsupervised	proximity-based	online	non-stationary	F-measure
[45]	2020	sequence	unsupervised	non-linear	online	non-stationary	MSE
[9]	2021	non-contextual	unsupervised	proximity based	-	coupling	AUC
[7]	2021	non-contextual	unsupervised	hash-based	offline	imbalanced	AUC
[73]	2021	non-contextual	unsupervised	proximity-based	online	high-dimensional	F-measure
[20]	2021	non-contextual	supervised	proximity-based	online	imbalanced	TPR, FPR
[74]	2022	sequence	unsupervised	proximity-based	online	concept-drift	F-measure
[75]	2022	sequence	unsupervised	proximity-based	online	non-stationary	AUC
[8]	2022	non-contextual	semi-supervised	proximity-based	online	imbalanced	AUC
[76]	2022	non-contextual	unsupervised	proximity-based	-	imbalanced	F-measure
[44]	2022	contextual	supervised	non-linear	-	non-stationary	MSE
[77]	2022	non-contextual	supervised	non-linear	offline	non-stationary	SMD
[78]	2022	non-contextual	supervised	linear	offline	imbalanced	Accuracy
[79]	2022	sequence	supervised	non-linear	online	concept-drift	AUC
[80]	2023	contextual	supervised	non-linear	-	causality	F-measure
[81]	2023	non-contextual	unsupervised	OCSVM	online	high-dimensional	AUC

on two main points: change detection, and learning and prediction. Change point detection refers to the ability to detect the point after which there is a drift in the data and the model performance will deteriorate. The data was divided into non-overlapping windows of fixed size and change point detection was performed using heuristics such as the number of instances being classified as outliers in the current window and the cumulative sum statistical (CS) test of the current and previous window in a univariate random sequence O_j with mean μ and any sudden or gradual increase in mean $(\mu + \varepsilon)$. If the CS before and after an unknown interval was greater than the threshold then a change point was detected for a window j, where

$$CS_j = max[0, O_j - (\mu + \varepsilon) + CS_{j-1}]$$

In learning and prediction, on the other hand, every time a change point was detected, a new model was trained on new instances and added to the ensemble. When the ensemble was full, the oldest model was removed.

In another approach [18], the authors proposed a method for online anomaly detection in Structural Health Monitoring (SHM) using a combination of one-class support vector machines (OCSVMs) and a concept drift adaptation algorithm. The authors first trained an OCSVM on a set of normal data to identify the baseline behavior of the structural system. Then, during online monitoring, the OCSVM was updated with new data points by computing the similarities between the new data point and the margin support vector (S), error support vector (E), and reserve vector (R) satisfying Lagrange multipliers $\alpha_i \in (0, 1)$, $\alpha_i = 1$ and $\alpha_i = 0$ respectively. These vectors are the categories of training data. The authors introduced notations as gc is the decision of the original model and d_{cR} and d_{cE} are the minimum distance between the new point and reserve vector and error set respectively. The model updates in the following scenarios,

1) if $gc > 0, d_{cR} > dcE$ is gradual normal drift.
2) if $gc < 0, d_{cR} < dcE$ is abrupt normal drift.

In [81], the authors combined the deep Belief Networks (DBN) and quarter-sphere-based one-class SVM (QSOCSVM) with improvements to process high-dimensional data. The input stream is first passed to the DBNs to get a reduced feature set with an order of anomalies distributed on one side of normal data. Then QSOCSVMs – an improvement upon sphere-based OCSVM in terms of handling skewed distribution of data and linear execution time instead of quadratic, are used for computing border support vectors (BSV) in the training phase. A major improvement of this work is that instead of determining hyper-planes or hyper-spheres and then classifying test data with relative positions to those planes. This works computes BSV in the training phase and uses these vectors for a weighted evaluation of test data.

2) PROXIMITY-BASED METHODS

The Glance algorithm [70] is a method designed for detecting anomalies in attributed graphs that possess contextual attributes on their edges. To identify communities or groups of elements within the graph, it leverages the Louvain community detection algorithm, which aims to maximize the modularity score. In the initial stage, the Louvain algorithm

is employed to partition the graph into communities based on their interconnectedness. Each community represents a cohesive group of elements in the graph. In the subsequent stage, the Glance algorithm selects the most pertinent features for each community using the Laplacian Score, a feature selection technique. The Laplacian Score ranks features by considering their variance and similarity among neighboring elements. Once the relevant features for each community are determined, the Glance algorithm calculates an outlierness score for each element within the community using an outlierness score function. This function quantifies the extent to which an element differs from the norm within its community, by comparing it to the mean difference among the community members. The algorithm concludes by providing a ranking of the graph's vertices based on their respective outlierness scores.

SelectVC [6] is a framework for modeling selective value coupling to detect outliers in high-dimensional categorical data. Instead of using independent full-space or feature subspace methods, the authors learn the relationship between inlying and outlying values by using a function ψ to effectively define the outlying values set. Outlying values are infrequent values caused by outliers. To detect outliers, an initial vector of outlier scores was generated for each feature value. Then, an outlier scoring function called ϕ recomputed the outlier score of each value based on its couplings with the selected outlying value set. In condensed space the couplings are modeled using a scoring function ψ, focusing only on the couplings between the single value and the outlying value set rather than the full value set. This re-computation of the outlying value set and computation of outlier values was repeated until the outlier vector converges.

To address the issue of false positives in network anomaly detection, [27] proposed local adaptive multivariate smoothing (LAMS). LAMS works by replacing the output of the anomaly detector with the average anomaly score of similar events that have been previously observed. The similarity between events was determined by a context function called *Kh*. LAMS used a Nadaraya-Watson estimator, which was a non-parametric estimator that performed local averaging of events and converges to the true value of the optimization function, subject to certain assumptions. Long-term structured false positives, which were events confined to a subset of network hosts without a direct relationship to the background, could be removed by LAMS using a different similarity measure for alerts than that used in the anomaly detector. This resulted in decreased scores on false positives identified by LAMS.

The Stepwise framework [15] is designed to address concept drift adaptation and anomaly detection in software Key Performance Indicator (KPI) streams. It consists of three main components: detection, classification, and adaptation. For concept drift detection, the framework integrates two existing techniques: the Improved Singular Spectrum Transform (iSST) to calculate the change score for each interval, and the Extreme Value Theory (EVT) to automatically determine the threshold value. By combining these techniques, the framework can detect concept drift without relying on a fixed threshold. After detecting a concept drift, the framework classifies it as expected or unexpected by conducting a causal impact analysis on the event logs. This classification process involves a semi-automated approach where the operator makes the final decision to differentiate changes caused by software modifications from those caused by system bugs. Finally, the adaptation algorithm extracts two features, denoted as A and B, which represent the iSST scores and median values of each window. These features are then fitted using Robust Linear Models (RLM). A linear transformation is applied to the linear model, which is subsequently used by the anomaly detectors. In summary, the Stepwise framework provides a robust approach for detecting concept drift and adapting to it in software KPI streams. By combining drift detection, classification, and adaptation techniques, it offers a comprehensive solution for addressing concept drift and detecting anomalies in software performance monitoring.

The authors of [72] proposed an approach of structural learning for non-IID attributes using Bayesian networks where nodes represent the attributes and the edges represented the direct dependency. The authors used the MMHC algorithm for discovering the Attribute dependency Graph and Attribute-Value dependency triple. The MMHC (Max-Min Hill Climbing) algorithm is a graphical model structure learning algorithm that aims to discover the underlying causal relationships between variables. It used a hill-climbing search strategy to iteratively explore and refine the structure of a Markov or a Bayesian network based on a given scoring metrics, such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). The algorithm aimed to maximize the score by adding, removing, or reversing edges between variables while considering the constraints of the network structure [83]. Furthermore, it also developed a Non-IID similarity metric consisting of intra- and inter-attribute similarity to capture the relationship between attributes and attribute values based on the attribute structure. This work then introduced several instances of the original framework by using different ML models. For example, KNN with proposed similarity metrics was used to detect outliers in the non-IID data.

The authors of [19] proposed a framework called O-NSD for detecting outliers in non-stationary data streams that incorporate distribution change detection to trigger model updates. The outlier detection model was retrained with the data from the new distribution when a change is detected. The authors introduced a new distance function called Improved Kullback-Leibler (IKL) which is defined as the max of the Kullback-Leibler (KL) divergences between the two distributions with the assumption of being monotonic. The authors provided experimental evidence that the distance between the current window and the reference window monotonically increased at the beginning of the new distribution.

Then the authors experimented with the two variants of IKL; first, keeping the buffer for the distribution changes in current and reference windows. Then using the average value against a threshold to identify the distribution change but this variant requires the careful setting of threshold values. Second, a parameter-free version of IKL measured the longest increasing sub-sequence (LIS) of distance buffers as an indication of distribution change. Whenever a distribution change was detected using any of the variants, the model was retrained on new data to learn the new data characteristics.

[45] proposes an anomaly detection method called Wavelet Auto-encoder Anomaly Detection (WAAD) for non-stationary and non-periodic uni-variate time series. The authors used the sliding window approach and compute Wavelet coefficients on each window and then using autoencoder the reconstruction error of these coefficients for each sliding window was computed. If the reconstruction error of k continuous windows was greater than a predefined threshold then that sub-sequence was identified as an anomalous sub-sequence.

This study [9] investigates the influence of homophily coupling and heterogeneous probability distributions on outlier factors. The authors proposed Coupled Unsupervised OuTlier detection (CUOT) to estimate the outlierness of each value by modeling both intra-feature and inter-feature couplings using outlier factors. The CUOT framework was instantiated into two instances, Coupled Biased Random Walks (CBRW) and multiple-granularity Subgraph Densities-augmented Random Walks (SDRW), to handle noisy features. CBRW introduced an intra-feature outlier factor by normalizing the features and incorporating the interdependence of outlier values across different features through the conditional probabilities of these values. The intra-feature outlier factored and their couplings were mapped onto a directed attributed value-value graph and modeled by biased random walks to estimate the outlierness of all values. SDRW worked on an undirected value graph and instead of using conditional probabilities, it computed a Lift-based outlierness influence vector to learn intra-feature value coupling.

A-Detection [73] is a four-stage technique for detecting anomalies in reliable edge services. A-Detection first collected the reliability data streams from edge services using the Bernoulli test. The definition of reliability is constrained by service requests and maximum completion time. If a service request of length L is completed in M time then the request is successful otherwise a failure. Collected streams were then organized into matrices X_h (history data) and X_n (current data) using the time window technique. Singular Value Decomposition (SVD) was then applied to extract features P_{old} and P_{new} from $X_h and X_n$ respectively. The Jensen Shannon (JS) divergence was calculated to measure the Fractional Distribution Change (FDC) between them. The anomaly detection was performed by identifying FDC peaks.

This research presents a novel outlier detection method called ADD (Average Divergence Difference), as referenced in [76]. ADD is specifically designed to handle data objects with skewed distributions, eliminating the requirement for an artificial parameter k in nearest neighbor algorithms. The ADD algorithm consists of four essential steps. Firstly, it establishes the notions of skewness and local density based on the skewed distribution of the object and its natural neighbors. Subsequently, the algorithm computes the divergence factor (DF) of the object by assessing the ratio between its skewness and local density. This DF serves as an external characterization factor that captures the relationship between the object's skew distribution characteristics and its compactness. The algorithm then calculates the average divergence difference, which serves as an internal characterization factor. It measures the variation in skew distribution characteristics among neighboring data objects by comparing their respective divergence factors. Finally, by employing a threshold value, the algorithm identifies local outliers within the dataset.

Based on the existing works [84], [85], [86], DragStream [74] is an anomaly detection and concept drift detection algorithm. This algorithm is characterized by three building blocks: 1) a cache memory, 2) an incremental clustering algorithm, and 3) concept drift detection. The algorithm started with an empty cache and for any incoming sub-sequence, the sub-sequence was matched with the existing cluster and existing anomalies in cache C. If the z-normalized Euclidean distance between the new sub-sequence and the existing anomaly is less than the threshold then this new sub-sequence was considered normal and the previous subsequence was also removed from the cache. If the new sub-sequence was closer to the existing cluster then the cluster was updated. If this new sub-sequence was not similar to both the cluster and an existing anomaly then this new sub-sequence was classified as an anomaly and added to the cache. The differentiation of clusters and cache captured the normal and abnormal concepts.

The authors of [75] leveraged the Huffman coding to detect anomalies in audio data. The authors used dynamic Huffman coding and performed swapping and reorganizing of nodes in a binary tree to maintain the non-decreasing indices. Fraction weights instead of integer weights were assigned to the nodes. The audios frames were represented as Mel-Frequency Cepstral Coefficients (MFCC), energy, and Zero Crossing Rate (ZCR) features, and cosine similarity was used to find a hit/miss in the tree. When there was a hit the tree weights are updated and when missed a new node was generated. The proposed work did not impose constraints on the length of the tree so a node merge scheme was proposed. Two similar nodes were merged if their similarity score was below the merging threshold. This node merging kept a cap on the tree length and also combined similar events into a single node. The normalized anomaly score was computed by averaging the ratio of matched nodes and root nodes for k continuous frames.

3) HASH-BASED METHODS

Isolation forests [87] have been a popular approach for detecting outliers in data that fulfill the IID assumption. In [7], the authors proposed a framework for detecting anomalies in non-IID data using an ensemble of Isolation Forests and a hash-based indexing method called EDBHiforest. The main contribution of this approach is that it can use any distance measure (metric or non-metric) which generalizes it to both IID and non-IID data. The proposed framework started by partitioning the data into subsets based on the hash value of each data point. The hash function used is the Locality-Sensitive Hashing (LSH) algorithm, which maps closer data points to the same hash bucket with high probability. The authors extended the Distance Based Hashing (DBH) to Extended-DBH (EDBH) and hashed the data into w buckets. A family of EDBH functions was used to compute the indexes for each tuple. This family of functions was treated as a black box and exposed as a parameter for user-defined functions. The EDBHiforest algorithm then generated a random multiway tree and traversed it to compute anomaly scores. The authors identified the limitations of existing work and improved them for better generalization instead of creating a complex novel approach. The datasets used in the above approach were structured and thus we classified the approach as *offline* learning.

B. (SEMI) SUPERVISED LEARNING

Here we briefly discuss the selected approaches present in literature for semi-supervised and supervised outlier detection on non-IID data (mentioned in Table 3) such as Recurrent Neural networks (RNNs), non-linear regression and labeled K-means++.

1) LINEAR MODELS

Reference [12] presented an online anomaly detection approach using Recurrent Neural Networks (RNNs) with concept drift adaptation. The methodology includes four key components: local normalization, multi-step prediction, RNN-based anomaly score computation, and RNN model update. In the local normalization step, the data is divided into fixed-length segments and normalized using mean and standard deviation. The multi-step prediction involves linear units in the input and output layers for future predictions. The RNN model is trained using unsupervised and supervised learning techniques to handle concept drift, where unsupervised learning identifies anomalies and supervised learning updates the model parameters when concept drift is detected. The anomaly score is computed based on the l2 norm over the last predictions, distinguishing between point outliers and concept drift.

In [78] the authors proposed a linear least-square-based approach for detecting poisoning attacks in federated learning. The authors compute the average Euclidean distance between the benign and malicious clients models and global model and monitor the decline in the global model due to poisonous updates while the malicious model shows smooth local convergence. The authors then performed the least-square curve fitting on these distances to predict malicious clients.

2) NON-LINEAR MODELS

The proposed anomaly detection algorithm by [10] is designed for evolving data streams using Hierarchical Temporal Memory (HTM). It involves two key steps to assess anomalies. Firstly, the prediction error is computed as the inverse of the common bits between the actual and predicted vectors, allowing evaluation of HTM accuracy. While this captures shifts in input statistics, it may result in false positives due to inherent noise. To address this, the authors introduce anomaly likelihood as a probabilistic metric. This metric utilizes the distribution of error values, obtained by modeling the rolling normal distribution of prediction errors. By considering the HTM models' prediction history, the anomaly likelihood provides a measure of the degree of anomaly. Notably, it exhibits clear peaks in noisy scenarios, enabling reliable detection. Anomalies can be identified when there is a series of spikes or when a scenario transitions from high unpredictability to complete predictability.

In [71], The authors presented a Growing Neural Gas network approach for evolving data streams, incorporating adaptive learning rates based on local characteristics of neurons and strategies for adding or removing neurons. The adaptive local learning rate, determined by the local error of each neuron, enabled the network to closely track distribution changes and handle local variations effectively. The algorithm improved adaptability to local changes by selecting winning neurons based on their low local error and adjusting their learning rate accordingly. Additionally, a forgetting strategy was introduced to remove aged neurons, considering the impact on neighboring neurons. The authors also presented a dynamic approach to add new neurons based on a probability criterion, ensuring fewer neurons were created during stationary distribution and allowing for more when distribution shifts occurred.

In [20], the authors addressed the problem of data skewness and proposed a novel peer-to-peer algorithm, P2PK-SMOTE, to train supervised anomaly detection machine learning models on non-IID data. This algorithm handles local class-imbalance by synthetically generating the minority class. The authors proposed a P2P environment where each node was holding part of the initial weights and data instances. The proposed method P2PK-SMOTE was a fully decentralized framework that can help participating clients re-balance local datasets for anomaly detection without requiring a cloud or global model. The proposed approach improved SMOTE by adaptively over-sampling the minority class based on k points instead of just one, then n synthetic points were generated based on fixed amounts of nearest neighbors to add complexity and prevent tracing. A small fraction of this synthetic data was shared across the device to reduce the risk of data de-identification. This approach not only solves

the problem of data imbalance but also provided a means to mask the data and not violate privacy in IoT networks. The results on real datasets showed almost 100% performance (i.e., True positive rate) and demonstrated the effectiveness of this approach.

Another approach for handling data skewness is presented in [8], this study proposed two label-aware clustering-based methods to tackle the class imbalance problem in extremely skewed data through majority class undersampling. The authors used kmeans++ and linear vector quantization (LVQ) algorithms to rebalance the highly skewed data. For Kmeans++ the centers equal to the number of minority class instances were defined. After the termination of the algorithm (either by convergence or maximum iterations), the updated centers were used as the majority class samples thus making the data strictly balanced. In the same manner in LVQ the centers were initialized using kmeans++ then the Euclidean distance was used to adjust the centers for each data sample. After the termination of the algorithm, the adjusted samples close to the center were used as the majority class and rest of the majority class samples were discarded.

A novel approach for fault detection in non-stationary industrial processes using deep learning techniques was introduced in [44]. The method combined correlative stacked autoencoder (C-SAE) and correlative deep neural networks (C-DNN) to address the limitations of conventional methods like Principle Component Regression (PCR) and Partial Least-Square (PLS). Two new loss functions were proposed: constructive correlative-SAE and demoting correlative-DNN, which allowed for non-linear correlation analysis and output-related fault detection. The approach relaxed assumptions about linear/non-linear relationships among variables and non-Gaussian distributions with non-stationary behaviors. The C-SAE reconstructed the output-related portion of input features, while the correlative-DNN further decomposed y-unrelated components. This enabled relative output-related decomposition and non-stationary fault detection without requiring total decomposition of process measurements.

In [77], the authors outlined a non-linear regression method for detecting outliers in highly correlated multivariate non-Gaussian data. In the first step, multivariate non-Gaussian random vectors were normalized using a multivariate normalizing transformation into Gaussian random vectors. In the second step, the non-linear regression model was constructed based on the multivariate normalizing transformation. In the third step, the prediction intervals of non-linear regression were built. These intervals were defined by a statistical equation that combined the coefficients of transformation and t-test in the linear regression model [88]. Once the intervals were formed, any predicted value lying outside the interval was classified as an outlier.

ARCUS (Adaptive framework foR online deep anomaly deteCtion Under a complex evolving data Stream) [79] is an online anomaly detection framework that adapts to evolving data streams by managing a compact pool of models. The framework balances accuracy and efficiency by maximizing the accuracy of all models in the pool while keeping the pool as small as possible. It estimates model reliability by comparing learned concepts with the current batch and calculates concept-driven anomaly scores. The statistical significance of score differences between current and previous batches is used to assess model reliability. A forget or merge policy is employed, where model merging utilizes centered kernel alignment (CKA) of latent variables, and models performing poorly on the current distribution are marked for forgetting.

This paper [80] introduced a novel approach for stable detection in Network Intrusion Detection Systems (NIDS). The proposed method combines causal features, weight decorrelation, and a nonlinear mapping. A deep causal stable learning model is employed to address false correlations in NIDS. Causal interference and weighted scores are used to estimate the causal effect between features and labels, amplifying the influence of causal features on the label variables. An optimization approach is applied to eliminate false correlation effects by removing weight and noise factors. Additionally, a weight vector optimization technique evenly distributes weights on the unit hypersphere to reduce feature correlation and improve generalization in neural networks. The weight vector's attributes, module length and direction, were modified using a derivation of Tammes problem [89]. This weight decorrelation technique was applied to both the hidden layer and classification layer of the autoencoder.

C. OUTLIER DETECTION IN APPLICATIONS

Among the 40 papers obtained from our search result, the selected papers are discussed based on their real-life applications. These papers did not propose outlier detection methods but demonstrate the application of existing techniques in complex real-life scenarios.

In [90], the authors trained an autoencoder to identify anomalies in complex ocean acoustic data. The manual approach to identifying faults in the performance of the hydrophone buried in oceans is via visual inspection of power spectral density plots (spectrograms). The authors used these spectrograms to train an autoencoder with a threshold to identify the images with anomalous features.

The authors of [26] developed a tool to visually represent the causal relations for network requests for detecting network anomalies. The design was based on traffic causality analysis to separate the legitimate and abnormal events. The tool provided a special visual locality property that supports different levels of visual-based querying and reasoning required for the sense-making process on complex network data. The high visual locality was provided by grouping the nodes according to their causal relationship. The design prioritized the causality that clusters nodes around their root-triggers and forms separate trees for different requests. Within each tree, the nodes were organized by their temporal and other logical information.

The authors of [91] performed anomaly detection in the domain of dependable systems. For the quantitative assessment, the authors collected 16 heterogeneous attributes to present the Key Performance Attributes (KPI) of an operating system, Linux Red Hat EL5. They applied a random walk algorithm to detect anomalies in the performance. The study showed that the histograms of the first-order time differences of the monitored indicators could be better approximated by a Cauchy and/or Laplace distribution instead of a Gaussian distribution.

The current methods used for detecting anomalies in cosmic ray variations, such as calculating spherical harmonics and employing averaging techniques, are not only time-consuming but also ineffective in capturing all types of anomalous events in cosmic rays. Reference [92] proposed the use of wavelet transform constructions and autoencoder for the analysis of cosmic ray signals. The authors developed two algorithms for anomaly detection. The first one was composed of an under-complete autoencoder followed by continuous wavelet transform (CWT) and a thresholding function to detect anomalies in the narrow spectrums by doing dependency analysis. The second one consisted of orthogonal multiple scale analysis (MSA) of cosmic ray data, its wavelet reconstruction followed by CWT, and the thresholding function for detecting anomalies of various intensities and frequencies. The proposed approach shows promising results on the data from the neutron monitor of the Inuvik station.

VI. EVALUATION MEASURES FOR OUTLIER DETECTION

Evaluation measures play a crucial role in assessing the performance of an algorithm. In this section, we will discuss some of the most commonly used evaluation measures used in literature for outlier detection algorithms on non-IID data.

AUC-ROC Curve: The AUC-ROC Curve is a popular evaluation metric used in binary classification problems. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. AUC stands for Area Under the ROC curve and calculates the area under the entire ROC curve, ranging from (0,0) to (1,1). AUC is known for measuring how well the model's predictions are ranked, rather than their absolute values. This means that AUC is scale and classification-threshold invariant. Many outlier detection algorithms use the AUC-ROC curve to assess their ability to distinguish between actual data and outliers.

Precision-Recall: Precision-Recall is a widely used evaluation metric in outlier detection. It measures the ability of an algorithm to identify outliers in the data while minimizing the number of false positives where normal instances are classified as outliers.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$
$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

Note: True Positive, False Positive, and False Negative refers to the number of correct positive predictions, incorrect positive predictions, and incorrect negative predictions, respectively.

F-measure and F_β-measure: The F-measure is a metric that combines precision and recalls into a single value. F-measure is often used when the number of outliers in the data is low and it is important to identify them accurately.

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The β parameter in F_β-measure is a generalization of the F-measure (assuming $\beta = 1$), allowing adjusting the weights between precision and recall. A smaller β value e.g., $\beta = 0.5$, gives more weight to precision, while a larger value e.g., $\beta = 2.0$, emphasizes recall.

$$F_\beta\text{-measure} = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}$$

Mean squared error (MSE) is a commonly used metric to measure the quality of predictions in regression problems. It calculates the average of the squared differences between the predicted and actual values of the target variable.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where the test set has n observations, y_i is the actual value and $\hat{y}_i$ is the predicted target value for the i-th observation.

MSE is used when identifying outliers using regression analysis or autoencoder. MSE measures the average squared difference between the predicted and actual values, which means that it puts more weight on larger errors. A lower MSE value indicates better prediction accuracy and a perfect MSE score of 0 indicates that the model's predictions exactly match the actual values.

1) SPECIAL METRICS

Following are a few prominent evaluation metrics specifically designed for outlier detection in non-IID data.

Coupling Strength (coup) [6]: It computes the probability of the outlier label given a single feature value to measure their coupling strength. $coup_U$ is defined as the average conditional probability of the outlier class label over all its values in a value set U, i.e.,

$$coup_U = \frac{1}{|U|}\sum_{u \in U} P(O|u)$$

Higher the $coup_U$ stronger the couplings between the outlier class O and the values in U.

NeoDis [72]: Non-IID gEneralized cOupled based Distance (Neo-Dist) is used to measure coupled similarity between two objects x_k and x_q;

$$NeoDis(x_k, x_q) = \sum_{i=1}^{n} \alpha_i \times S^i(v_i^{x_k}, v_i^{x_q})$$

where $\alpha_i \in [0, 1]$ represents the weight of the coupled metric attribute value similarity of an attribute A_i and $\sum_{i=1}^{n} \alpha_i = 1$. The function $S^i(v_i^{x_k}, v_i^{x_q})$ represents inter-attribute or intra-attribute similarity or attribute-value similarity.

Squared Mahalanois Distance (SMD): Squared Mahalanobis distance is a measure of the distance between a point and a distribution. It is computed by taking the difference between the point and the mean of the distribution, scaling the difference by the inverse of the covariance matrix of the distribution, and then squaring the result. The squared Mahalanobis distance is useful in machine learning for anomaly detection and classification tasks where the data is not normally distributed. It is a way to measure how far away a point is from the distribution it came from, taking into account the correlation between the different features. Let $x \in R^p$ generated from a p-variate(probability) distribution $f_X(.)$, $\mu = E(X)$ of the distribution and $\Sigma = E(X - \mu)$ be the covariance matrix. The squared Mahalanobis Distance is defined as:

$$D^2 = (X - \mu)^T \Sigma^{-1} (X - \mu)$$

VII. DATASETS FOR OUTLIER DETECTION IN NON-IID DATA

We compiled a comprehensive list of datasets frequently used to evaluate the performance of outlier detection algorithms for non-IID data in Table 4. We collected the actual source of the dataset and the studies it appears in, the dimensions of the dataset (row × columns) and a brief description from the actual source, highlighting the domain of the data.

While compiling the list of these datasets we observed that the majority of datasets are time-series datasets. Although NICO is a dedicated effort for non-IID image datasets there are still opportunities to explore in the domain of graph, network, spatial and tabular data.

VIII. OPEN DISCUSSION AND CHALLENGES

In this section, we discuss the limitations of existing work and highlight some of the open challenges for outlier detection in non-IID data.

A. LIMITATIONS OF EXISTING WORK

The general limitations of the existing literature are summarized in the following:

1) The literature pays more attention to learning supervised models or ensembles of supervised and unsupervised models but pays little attention to capturing the underlying relationship between data and outliers.
2) There is a gap to explore the domains of DNN-based feature selection and causal linking for outlier detection in non-IID data. For example, causal models can be used to capture the nature of dependencies within the data generation process
3) None of the work focuses on outlier detection in multi-modal data.
4) More attention is needed to high-dimensionality in the context of data dependency and heterogeneity when performing feature selection i.e., scalability of similarity metrics, scoring functions to capture the couplings while considering sampling, and consideration of feature heterogeneity when performing feature selection.
5) Lack of approaches handling missing values, or data with quality issues when dealing with contextual outliers and semi-structured data.
6) Insufficiency of publicly available non-IID datasets for benchmarking outlier detection is a significant challenge. Much of the existing literature evaluates outlier detection algorithms on multi-class datasets, assuming that the minority class represents the outliers (i.e., Arithmiya, BalanceScale and Gas datasets in Table 4. However, this approach oversimplifies the problem of outlier detection for non-IID data and fails to capture its true complexity.

B. OPEN CHALLENGES FOR OUTLIER DETECTION IN NON-IID DATA

This section describes the selected and relevant open challenges encountered in the literature regarding outlier detection in non-IID data.

1) DATA CHARACTERISTICS QUANTIFICATION

The field of data characterization or data complexity quantification aims to gain insights into and measure the inherent characteristics and complexities of data [38]. This understanding is essential for achieving optimal alignment between data and models and for designing and evaluating learning methods [106], [107]. While numerous data indicators have been developed to quantify data complexity in tasks such as classification, sequence analysis, and time-series forecasting, the focus on outlier detection has been limited [25], [108], [109], [110], [111]. Notable studies in this area include [7], [106] and [9].

In [106], various k-nearest-neighbor-based outlier detection methods are evaluated on publicly available datasets. The authors introduce two data indicators, difficulty and diversity, to analyze dataset complexity based on detector performance agreements and conflicts. This approach differs from other works such as [7] and [9], which focus on capturing data relationships with underlying structures and designing diverse data indicators such as feature-value similarity, value-value similarity, etc. These studies contribute to a deeper understanding of data complexity in the context of outlier detection, providing valuable insights for future research and development in this field. However, these approaches are limited to structural or numerical data. Therefore, it is desirable to introduce robust methods which evaluate and benchmark different data characteristics on a variety of structured, semi-structured, high-dimensional, heterogeneous, and coupled data.

TABLE 4. Benchmark datasets for outlier detection in non-IID data.

Dataset	Source & Study	Dimensions	Description
Arrhythmia	[6], [93]	425×279	A sixteen-class classification dataset with linear and nominal attributes used for learning selective value coupling and identifying minority classes as outliers.
BalanceScale	[72], [93]	625×4	A three-class classification dataset. The smallest class is treated as an outlier and the largest class is normal.
Basehock	[6], [94]	$1,993 \times 4,862$	A 20-Newsgroups binary classification dataset used in studies for feature selection in high dimensional data.
Bodyfat	[72], [95]	262×15	A regression dataset used to learn inter/intra attribute-value couplings.
Caleb/CalebA	[6], [96], [97]	$202,599 \times 39$	A thousand-class dataset used to study the scalability of feature value coupling algorithm.
EINino	[19], [72], [93]	$178,080 \times 12$	Oceanographic and surface meteorological measurements gathered from a network of buoys placed across the equatorial Pacific region. This time-series dataset is used to find weather anomalies.
Electricity	[71], [72], [98]	$45,312 \times 5$	Binary classification time-series dataset each instance is an average of 30 minutes and used to predict the increase and decrease in electricity price based on time (context) and demand.
Forest Cover Type	[11], [18], [93]	$581,012 \times 54$	A seven-class classification dataset with all categorical attributes used to study the impact of feature selection and dimensional reduction for detecting outliers.
Gas	[11], [18], [93]	$13,910 \times 128$	A six-class chemical sensor dataset used for drift detection.
GOutRank	[70], [99]	N/A	A subgraph of Disney product from the Amazon co-purchase Amazon Network [99]. This subgraph has 124 nodes and 334 edges and 30 dimensions with hand-labeled local outliers.
HTRU (High Time Resolution Universe Survey)	[72], [93]	$17,898 \times 9$	A binary class earth rotation time-series data with skewed negative samples used to study the relationship between attributes and values for outlier detection.
Hepatocellular Carcinoma	[6], [9], [93]	165×49	A high dimensional binary classification dataset with the majority of categorical variables used to study homophily coupling in higher dimension.
Household Electric Power Consumption	[19], [93]	$2,075,259 \times 9$	Multivariate time-series dataset used to study concept drifts in the data i.e., distribution changes in peak hours.
IBRL (Intel Berkeley Research Lab)	[11], [101]	$2,300,000 \times 8$	Sensor data collected using weatherboards are used to identify concept drifts in timestamp data.
N-BaIoT	[20], [93]	$7,062,606 \times 115$	A multi-class classification dataset used to detect outliers and malicious botnet attacks on IoT devices.
NICO (Non-I.I.D. Image dataset with Contexts)	[80], [102]	N/A	The NICO dataset is designed for non-I.I.D or out-of-distribution image classification, simulating a real-world setting where testing distribution may shift arbitrarily from training distribution. The dataset can support transfer learning or domain adaptation when the testing distribution is known and stable learning or domain generalization when it's unknown. Images are labeled with main concepts and contexts. NICO contains 19 classes, 188 contexts, and nearly 25,000 images, supporting deep convolution networks training with scale.
Numenta Anomaly Benchmark (NAB)	[7], [10], [12]	N/A	NAB (version v1.1) is a benchmark for evaluating algorithms for anomaly detection in streaming, real-time applications. It is composed of over 58 labeled real-world and artificial time-series data files.
Pneumology	[74], [103]	$24,125 \times 1$	Patient respiration dataset used to study data discords in time-series with concept drift.
Rialto	[71], [79], [104]	$82,250 \times 24$	A hand labeled multi-class image classification data. Images of 10 buildings near the famous Rialto bridge in Venice were collected at different times and encoded in RGB histograms. The dataset is used to detect heterogeneous concept drift.
Yahoo Dataset	[12], [105]	N/A	A dataset comprising of real and synthetic time-series data with labeled anomaly points. The dataset includes different types of anomalies such as outliers and concept drifts. The synthetic data includes time-series with varying trends, noise, and seasonality, while the real data consists of Yahoo service metrics time-series.

2) FEATURE SELECTION

Another significant challenge in outlier detection in non-IID data is effectively handling high-dimensional data. High-dimensional data often includes a substantial number of irrelevant or redundant features, which can increase the noise level in the data and make it difficult to detect true outliers. To address this challenge, researchers have proposed a range of subspace [9], [47], [112], [113], [114], [115] and multiple subspace-based [6], [116], [117], [118] approaches for outlier detection in high dimensional data. However, these techniques may not always be effective in high-dimensional non-IID data as the impact of these techniques on non-IID is unknown. Furthermore, research is needed to assess the impact of existing techniques on non-

IID data and develop more sophisticated feature selection methods that can effectively handle the complexity of non-IID data.

3) DEEP LEARNING FOR COMPLEX OUTLIER DETECTION

Deep learning has emerged as a powerful approach for complex data analysis, including the field of outlier detection. While contemporary methods have shown success in detecting point anomalies, they often struggle with conditional or group anomalies. Deep learning models, on the other hand, excel at capturing complex temporal and spatial dependencies, allowing them to effectively learn representations from unordered data points. This capability opens up new possibilities for detecting complex anomalies that exhibit intricate patterns and relationships. Moreover, current deep outlier detection methods primarily focus on single data sources, leaving the realm of multi-modal outlier detection largely unexplored. Bridging the gap presented by multi-modal data is a challenging task for traditional approaches.

However, deep learning has demonstrated remarkable success in learning feature representations from various types of raw data [112], [119], enabling effective anomaly detection across multiple modalities [79], [120]. Deep models can learn unified representations by concatenating representations from different data sources, presenting exciting opportunities for multi-modal anomaly detection.

To fully harness the potential of deep learning in outlier detection, novel neural network layers or objective functions may need to be developed. By leveraging the inherent strengths of deep learning, such as its ability to capture complex dependencies and learn representations from diverse data sources, we can pave the way for more accurate and comprehensive outlier detection in complex datasets.

4) ROBUST EVALUATION METRICS AND ACTIVE LEARNING

A key challenge in outlier detection in non-IID data is the development of robust evaluation metrics that are capable of accurately assessing the performance of outlier detection algorithms. This challenge comes from the fact that non-IID data can exhibit varying levels of noise, outliers, and anomalies, making it difficult to differentiate between anomalous and non-anomalous data. Moreover, it is difficult to differentiate the outliers from concept drift and identify contextual outliers without domain experts. To address this challenge, [6] and [7] proposed specialized similarity measures to differentiate between inlying and outlying values. Reference [12] discussed an RNN-based approach to differentiate between concept drift and outliers. As the definition of outlier changes concerning changes in data (concept drift), more robust measures are needed to adopt these changes. Expert knowledge can help in the reduction of false positives for contextual outliers and when a drift has occurred. Active learning for outlier detection has recently gained prominence [121], [122]. Although this approach seems promising due to the inherited shortcoming of active learning (unavailability of domain experts, the burden of labeling) only a few works have been published in the context of outlier detection and none in the context of non-IID data. For that reason, more investigations are needed in this context.

5) COUPLING AND CAUSALITY

Typically, feature selection/extraction and outlier scoring are performed separately, potentially retaining irrelevant features for outlier detection. To enhance robustness, coupling these processes is crucial but very limited work is published in this context [6], [7]. Therefore, further investigation is required, considering the unique challenges posed by the complexity of non-IID and related challenges such as feature drift and concept drift. Moreover, feature causality or causal learning is also neglected in outlier detection. The authors of [123] presented various use cases to detect outliers by causal modeling of features and explain how the violation of causal relation might be the indicator of anomalous behavior. We have already discussed the application of causal inference in network intrusion detection in [67] but more attention is demanded in incorporating causal modeling for detecting domain-specific outliers. Causal modeling could also be the future for explainable outlier detection [124], [125]. In the same way, the notion of outliers also strongly depends on the domain, causal modeling of domain knowledge appears to be a promising path, especially if combined with machine learning methods making use of the domain knowledge.

6) HYPER-PARAMETER TUNING

Existing approaches in the literature require users to manually specify values for various parameters, such as outlier detection threshold, concept drift or feature drift indicators, hashing function choices, window sizes, autoencoder layer configurations, and more. These parameters play a crucial role in model induction and directly impact the performance of the model. However, tuning these parameters can be challenging, particularly in the case of non-stationary and high-dimensional data. This challenge becomes even more pronounced in online learning models, especially ensembles that evolve, as their input parameters may also need to adapt. Therefore, there is a need for autonomous systems that minimize user-adjustable parameters, especially those that cannot be learned from the data. Such autonomous systems would enhance the practicality and effectiveness of handling dynamic and complex data scenarios.

IX. CONCLUSION

In this paper, we have conducted a systematic literature review on outlier detection in non-IID data. Through an extensive search across various electronic databases, we have identified 32,052 papers, from which we selected 40 relevant studies. The primary objectives of this study were twofold. Firstly, we proposed a taxonomy to provide readers with a comprehensive understanding of the key aspects of outlier detection in non-IID data, including data characteristics,

methods, and evaluation measures. This taxonomy serves as a valuable guide for researchers and practitioners in the field.

Secondly, we focused on the application domain of outlier detection, highlighting its significance in various domains such as finance, healthcare, and computer networks. We also addressed the challenge of concept drift and emphasized the need to differentiate it from outliers, offering insights on adapting outlier detection algorithms for robust performance in the presence of concept drift. Additionally, we compiled a comprehensive list of non-IID datasets specifically designed for evaluating specialized outlier detection algorithms, facilitating further research and benchmarking.

While this study provides valuable insights and advancements in outlier detection on non-IID data, we acknowledge certain limitations of the existing work. These limitations include the need for more advanced methodologies that can handle complex non-IID data characteristics, the development of comprehensive evaluation frameworks, and the consideration of application-specific challenges. Moreover, several open challenges remain in this field, such as addressing the impact of data cleaning, incorporating deep learning and privacy-preserving techniques, and ensuring interpretability of outlier detection models in real-world scenarios.

APPENDIX

Following is the list of all venues (conferences, journals) included in the retrieval process.

1) Conference on Information and Knowledge Management (CIKM)
2) Data Mining and Knowledge Discovery
3) International Joint Conference on Neural Networks (IJCNN)
4) IEEE Intelligent Systems
5) Conference on Data and Application Security and Privacy (CODASPY)
6) Measurement
7) International Conference on Data Science and Management of Data (COMAD)
8) Artificial Intelligence in Medicine
9) International Conference on Scientific and Statistical Database Management (SSDBM)
10) Journal of Computer and System Sciences
11) Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)
12) Neurocomputing
13) Knowledge Discovery and Data Mining (KDD)
14) Applied Soft Computing
15) Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)
16) IEEE Sensors Journal
17) IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)
18) IEEE Transactions on Knowledge and Data Engineering (TKDE)
19) IEEE OCEANS
20) IEEE Transactions on Network and Service Management
21) IEEE International Symposium on Software Reliability Engineering (ISSRE)
22) Expert Systems with Applications
23) IEEE International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)
24) IEEE Access
25) IEEE International Conference on Data Mining (ICDM)
26) Knowledge and Information Systems
27) Journal of Applied Intelligence
28) Neural Computing and Applications
29) IEEE Conference on Applications of Computer Vision
30) AAAI Conference
31) ACM Computing Surveys
32) Advances in Databases and Information Systems (ADBIS)
33) Computer Science Review
34) Special Interest Group on Management of Data (SIGMOD)
35) International Conference on Very Large Data Bases (VLDB)
36) The Journal of VLDB

REFERENCES

[1] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier Analysis*. Cham, Switzerland: Springer, 2017.

[2] Y. Gao, H. Guan, and B. Gong, "CODM: An outlier detection method for medical insurance claims fraud," *Int. J. Comput. Sci. Eng.*, vol. 20, no. 3, pp. 404–411, 2019.

[3] S. G. Jacob, M. M. B. A. Sulaiman, B. Bennet, R. Vijayaraghavan, M. S. Sahayam, N. Thiviyakalyani, S. Shriram, and T. Hameed, "A graphical approach for outlier detection in gene–protein mapping of cognitive ailments: An insight into neurodegenerative disorders," *Netw. Model. Anal. Health Inform. Bioinf.*, vol. 11, no. 1, p. 22, Dec. 2022.

[4] V. Chadysas, A. Bugajev, R. Kriauziene, and O. Vasilecas, "Outlier analysis for telecom fraud detection," in *Digital Business and Intelligent Systems* (Communications in Computer and Information Science), vol. 1598. Riga, Latvia: Springer, Jul. 2022, pp. 219–231.

[5] D.-M. Ngo, D. Lightbody, A. Temko, C. Pham-Quoc, N.-T. Tran, C. C. Murphy, and E. Popovici, "HH-NIDS: Heterogeneous hardware-based network intrusion detection framework for IoT security," *Future Internet*, vol. 15, no. 1, p. 9, Dec. 2022.

[6] G. Pang, H. Xu, L. Cao, and W. Zhao, "Selective value coupling learning for detecting outliers in high-dimensional categorical data," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, Nov. 2017, pp. 807–816.

[7] H. Xiang, J. Wang, K. Ramamohanarao, Z. Salcic, W. Dou, and X. Zhang, "Isolation forest based anomaly detection framework on non-IID data," *IEEE Intell. Syst.*, vol. 36, no. 3, pp. 31–40, May 2021.

[8] J. Li, Y. Tao, H. Cong, E. Zhu, and T. Cai, "Predicting liver cancers using skewed epidemiological data," *Artif. Intell. Med.*, vol. 124, Feb. 2022, Art. no. 102234.

[9] G. Pang, L. Cao, and L. Chen, "Homophily outlier detection in non-IID categorical data," *Data Mining Knowl. Discovery*, vol. 35, no. 4, pp. 1163–1224, Jul. 2021.

[10] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, Nov. 2017.

[11] Z. Ghafoori, S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. A. Leckie, "Anomaly detection in non-stationary data: Ensemble based self-adaptive OCSVM," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2476–2483.

[12] S. Saurav, P. Malhotra, V. T. Vishnu, N. Gugulothu, L. Vig, P. Agarwal, and G. Shroff, "Online anomaly detection with concept drift adaptation using recurrent neural networks," in *Proc. ACM India Joint Int. Conf. Data Sci. Manage. Data*, Jan. 2018, pp. 78–87.

[13] C. Wiwatcharakoses and D. Berrar, "A self-organizing incremental neural network for continual supervised learning," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115662.

[14] S. Huang, J. Lin, and R. Tsaih, "Outlier detection in the concept drifting environment," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 31–37.

[15] M. Ma, S. Zhang, D. Pei, X. Huang, and H. Dai, "Robust and rapid adaption for concept drift in software system anomaly detection," in *Proc. IEEE 29th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2018, pp. 13–24.

[16] C. Lin, D. Deng, C. Kuo, and L. Chen, "Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers," *IEEE Access*, vol. 7, pp. 56198–56207, 2019.

[17] Y. Gao, S. Chandra, Y. Li, L. Khan, and T. Bhavani, "SACCOS: A semi-supervised framework for emerging class detection and concept drift adaption over data streams," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1416–1426, Mar. 2022.

[18] H. Tian, N. L. D. Khoa, A. Anaissi, Y. Wang, and F. Chen, "Concept drift adaption for online anomaly detection in structural health monitoring," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2813–2821.

[19] L. Tran, L. Fan, and C. Shahabi, "Outlier detection in non-stationary data streams," in *Proc. 31st Int. Conf. Sci. Stat. Database Manage. (SSDBM)*, Jul. 2019, pp. 25–36.

[20] H. Wang, L. Muñoz-González, D. Eklund, and S. Raza, "Non-IID data re-balancing at IoT edge with peer-to-peer federated learning for anomaly detection," in *Proc. 14th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, Jun. 2021, pp. 153–163.

[21] S. Sathe and C. C. Aggarwal, "Subspace histograms for outlier detection in linear time," *Knowl. Inf. Syst.*, vol. 56, no. 3, pp. 691–715, Sep. 2018.

[22] A. Abhaya and B. K. Patra, "An efficient method for autoencoder based outlier detection," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118904.

[23] F. M. Pereira and R. C. Sofia, "An analysis of ML-based outlier detection from mobile phone trajectories," *Future Internet*, vol. 15, no. 1, p. 4, Dec. 2022.

[24] J. Shi, Z. Pan, J. Fang, and P. Chao, "RUTOD: Real-time urban traffic outlier detection on streaming trajectory," *Neural Comput. Appl.*, vol. 35, no. 5, pp. 3625–3637, Feb. 2023.

[25] G. Lin, A. Lin, and J. Cao, "Multidimensional KNN algorithm based on EEMD and complexity measures in financial time series forecasting," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114443.

[26] H. Zhang, M. Sun, D. D. Yao, and C. North, "Visualizing traffic causality for analyzing network anomalies," in *Proc. ACM Int. Workshop Secur. Privacy Anal. (IWSPA)*, Mar. 2015, pp. 37–42.

[27] M. Grill, T. Pevný, and M. Rehak, "Reducing false positives of network anomaly detection by local adaptive multivariate smoothing," *J. Comput. Syst. Sci.*, vol. 83, no. 1, pp. 43–57, Feb. 2017.

[28] M. C. Massi, F. Ieva, and E. Lettieri, "Data mining application to healthcare fraud detection: A two-step unsupervised clustering method for outlier detection with administrative databases," *BMC Med. Inform. Decis. Making*, vol. 20, no. 1, p. 160, Dec. 2020.

[29] Z. Niu, S. Shi, J. Sun, and X. He, "A survey of outlier detection methodologies and their applications," in *Proc. 3rd Int. Conf. Artif. Intell. Comput. Intell. (AICI)*. Taiyuan, China: Springer, Sep. 2011, pp. 380–387.

[30] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Mach. Learn. Res.*, vol. 6, pp. 211–232, Jun. 2005.

[31] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K. Müller, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.

[32] Á. Fernández, J. Bella, and J. R. Dorronsoro, "Supervised outlier detection for classification and regression," *Neurocomputing*, vol. 486, pp. 77–92, May 2022.

[33] Y. Wang, X. Cao, and Y. Li, "Unsupervised outlier detection for mixed-valued dataset based on the adaptive k-Nearest Neighbor global network," *IEEE Access*, vol. 10, pp. 32093–32103, 2022.

[34] S. Buschjäger, P.-J. Honysz, and K. Morik, "Randomized outlier detection with trees," *Int. J. Data Sci. Anal.*, vol. 13, no. 2, pp. 91–104, Mar. 2022.

[35] A. Duraj and L. Chomatek, "Supporting breast cancer diagnosis with multi-objective genetic algorithm for outlier detection," in *Advanced Solutions in Diagnostics and Fault Tolerant Control* (Advances in Intelligent Systems and Computing), vol. 635, J. M. Kościelny, M. Syfert, and A. Sztyber, Eds. Sandomierz, Poland: Springer, Sep. 2017, pp. 304–315.

[36] H. Choi, D. Kim, J. Kim, J. Kim, and P. Kang, "Explainable anomaly detection framework for predictive maintenance in manufacturing systems," *Appl. Soft Comput.*, vol. 125, Aug. 2022, Art. no. 109147.

[37] G. de Jesus, A. Casimiro, and A. Oliveira, "Using machine learning for dependable outlier detection in environmental monitoring systems," *ACM Trans. Cyber Phys. Syst.*, vol. 5, no. 3, pp. 29:1–29:30, 2021.

[38] L. L. Cao, P. S. Yu, and Z. Zhao, "Shallow and deep non-IID learning on complex data," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, A. Zhang and H. Rangwala, Eds., Aug. 2022, pp. 4774–4775.

[39] L. Cao, "Beyond i.i.d.: Non-IID thinking, informatics, and learning," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 5–17, Jul. 2022.

[40] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[41] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2346–2363, Dec. 2019.

[42] B. Kitchenham, "Procedures for performing systematic reviews," Dept. Comput. Sci., Keele Univ., Nat. ICT Aust. Ltd., Joint Tech. Rep. TR/SE-0401 and 0400011T.1, 2004.

[43] S. Sousa and R. Kern, "How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1427–1492, 2023, doi: 10.1007/s10462-022-10204-6.

[44] B. Rashidi and Q. Zhao, "Output-related fault detection in non-stationary processes using constructive correlative-SAE and demoting correlative-DNN," *Appl. Soft Comput.*, vol. 123, Jul. 2022, Art. no. 108898.

[45] Y.-L. Li and J.-R. Jiang, "Anomaly detection for non-stationary and non-periodic univariate time series," in *Proc. IEEE Eurasia Conf. IoT, Commun. Eng. (ECICE)*, Oct. 2020, pp. 177–179.

[46] A. Hui and B. J. Gao, "When is nearest neighbor meaningful: Sequential data," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, A. Zhang and H. Rangwala, Eds., Oct. 2021, pp. 3103–3106.

[47] M. Riahi-Madvar, B. Nasersharif, and A. A. Azirani, "Subspace outlier detection in high dimensional data using ensemble of PCA-based subspaces," in *Proc. 26th Int. Comput. Conf., Comput. Soc. Iran (CSICC)*, Mar. 2021, pp. 1–5.

[48] A. Kumar, A. Kumar, A. K. Bashir, M. Rashid, V. D. A. Kumar, and R. Kharel, "Distance based pattern driven mining for outlier detection in high dimensional big dataset," *ACM Trans. Manage. Inf. Syst.*, vol. 13, no. 1, pp. 1–17, Mar. 2022.

[49] F. Kamalov and H. H. Leung, "Outlier detection in high dimensional data," *J. Inf. Knowl. Manag.*, vol. 19, no. 1, pp. 2040013:1–2040013:16, 2020.

[50] T. A. Messaoud, A. Smiti, and A. Louati, "A novel density-based clustering approach for outlier detection in high-dimensional data," in *Hybrid Artificial Intelligent Systems* (Lecture Notes in Computer Science), vol. 11734, H. P. García, L. Sánchez-González, M. C. Limas, H. Quintián-Pardo, and E. S. C. Rodríguez, Eds. Cham, Switzerland: Springer, 2019, pp. 322–331.

[51] D. Popovic, E. Fouché, and K. Böhm, "Unsupervised artificial neural networks for outlier detection in high-dimensional data," in *Advances in Databases and Information Systems* (Lecture Notes in Computer Science), vol. 11695, T. Welzer, J. Eder, V. Podgorelec, and A. K. Latific, Eds. Cham, Switzerland: Springer, 2019, pp. 3–19.

[52] M. Salehi and L. Rashidi, "A survey on anomaly detection in evolving data: [With application to forest fire risk prediction]," *ACM SIGKDD Explor. Newslett.*, vol. 20, no. 1, pp. 13–23, May 2018.

[53] C. H. Park, "A comparative study for outlier detection methods in high dimensional text data," *J. Artif. Intell. Soft Comput. Res.*, vol. 13, no. 1, pp. 5–17, Jan. 2023.

[54] I. Souiden, M. N. Omri, and Z. Brahmi, "A survey of outlier detection in high dimensional data streams," *Comput. Sci. Rev.*, vol. 44, May 2022, Art. no. 100463.

[55] A. Soliman, S. Girdzijauskas, M. Bouguelia, S. Pashami, and S. Nowaczyk, "Decentralized and adaptive k-means clustering for non-IID data using HyperLogLog counters," in *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science), vol. 12084, H. W. Lauw, R. C. Wong, A. Ntoulas, E. Lim, S. Ng, and S. J. Pan, Eds. Singapore: Springer, May 2020, pp. 343–355.

[56] S. Dang, "Learning-based methods for outlier detection imbalanced heterogeneous data," Ph.D. thesis, School Comput. Sci. Eng., Fac. Eng., Univ. of New South Wales, Sydney, NSW, Australia, 2018.

[57] M. Duan, D. Liu, X. Chen, Y. Tan, J. Ren, L. Qiao, and L. Liang, "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications," in *Proc. IEEE 37th Int. Conf. Comput. Design (ICCD)*, Abu Dhabi, United Arab Emirates, Nov. 2019, pp. 246–254.

[58] L. Cao, "Non-IIDness learning in behavioral and social data," *Comput. J.*, vol. 57, no. 9, pp. 1358–1370, Sep. 2014.

[59] H. Xu, Y. Wang, Z. Wu, and Y. Wang, "Embedding-based complex feature value coupling learning for detecting outliers in non-IID categorical data," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5541–5548.

[60] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[61] C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim, "Sensitivity analysis of linear structural causal models," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, Jun. 2019, pp. 1252–1261.

[62] J. A. W. B. Costanzo and O. Yagan, "Data-driven I/O structure learning with contemporaneous causality," *IEEE Trans. Control Netw. Syst.*, vol. 7, no. 4, pp. 1929–1939, Dec. 2020.

[63] M. J. Vowels, N. C. Camgöz, and R. Bowden, "D'ya like dags? A survey on structure learning and causal discovery," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 82:1–82:36, 2023.

[64] E. Bareinboim, J. Tian, and J. Pearl, *Recovering From Selection Bias in Causal and Statistical Inference*, 1 ed. New York, NY, USA: Association for Computing Machinery, 2022, pp. 433–450.

[65] K. Kuang, P. Cui, B. Li, M. Jiang, and S. Yang, "Estimating treatment effect in the wild via differentiated confounder balancing," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2017, pp. 265–274.

[66] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1617–1626.

[67] C. H. Yang, D. I. Hung, Y. Liu, and P. Chen, "Treatment learning causal transformer for noisy image classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 6128–6139.

[68] C. Louizos, U. Shalit, J. M. Mooij, D. A. Sontag, R. S. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6446–6456.

[69] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020, pp. 1–12.

[70] M. A. Prado-Romero and A. Gago-Alonso, "Community feature selection for anomaly detection in attributed graphs," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Lecture Notes in Computer Science), vol. 10125, C. Beltrán-Castañón, I. Nyström, and F. Famili, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-52277-7_14.

[71] M. Bouguelia, S. Nowaczyk, and A. H. Payberah, "An adaptive algorithm for anomaly and novelty detection in evolving data streams," *Data Min. Knowl. Discov.*, vol. 32, no. 6, pp. 1597–1633, 2018.

[72] F. Meng, Y. Gao, J. Huo, X. Qi, and S. Yi, "NeoLOD: A novel generalized coupled local outlier detection model embedded non-IID similarity metric," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining (PAKDD)* (Lecture Notes in Computer Science), vol. 11439. Cham, Switzerland: Springer, 2019, pp. 587–599.

[73] L. Wang, S. Chen, and Q. He, "Concept drift-based runtime reliability anomaly detection for edge services adaptation," *Trans. Knowl. Data Eng. (TKDE)*, early access, Nov. 11, 2021, doi: 10.1109/TKDE.2021.3127224.

[74] A. M. S. N. Bibinbe, A. J. Mahamadou, M. F. Mbouopda, and E. M. Nguifo, "DragStream: An anomaly and concept drift detector in univariate data streams," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2022, pp. 842–851.

[75] P. Kumari and M. Saini, "Anomaly detection in audio with concept drift using dynamic Huffman coding," *IEEE Sensors J.*, vol. 22, no. 17, pp. 17126–17138, Sep. 2022.

[76] Z.-Y. Xiong, Q.-Q. Gao, Q. Gao, Y.-F. Zhang, L.-T. Li, and M. Zhang, "ADD: A new average divergence difference-based outlier detection method with skewed distribution of data objects," *Int. J. Speech Technol.*, vol. 52, no. 5, pp. 5100–5124, Mar. 2022.

[77] S. Prykhodko, N. Prykhodko, L. Makarova, and A. Pukhalevych, "Outlier detection in non-linear regression analysis based on the normalizing transformations," in *Proc. IEEE 15th Int. Conf. Adv. Trends Radioelectron., Telecommun. Comput. Eng. (TCSET)*, Feb. 2020, pp. 407–410.

[78] X. You, Z. Liu, X. Yang, and X. Ding, "Poisoning attack detection using client historical similarity in non-IID environments," in *Proc. 12th Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2022, pp. 439–447.

[79] S. Yoon, Y. Lee, J.-G. Lee, and B. S. Lee, "Adaptive model pooling for online deep anomaly detection from a complex evolving data stream," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 2347–2357.

[80] Z. Zeng, W. Peng, and D. Zeng, "Improving the stability of intrusion detection with causal deep learning," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 4, pp. 4750–4763, Dec. 2022.

[81] Y. Qiao, K. Wu, and P. Jin, "Efficient anomaly detection for high-dimensional sensing data with one-class support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 404–417, Jan. 2023.

[82] Z. Ghafoori, S. Rajasegarar, S. M. Erfani, S. Karunasekera, and C. A. Leckie, "Unsupervised parameter estimation for one-class support vector machines," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining (PAKDD)* (Lecture Notes in Computer Science), vol. 9652. Cham, Switzerland: Springer, 2016, pp. 183–195.

[83] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, Oct. 2006.

[84] C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 1317–1322.

[85] D. Yankov, E. Keogh, and U. Rebbapragada, "Disk aware discord discovery: Finding unusual time series in terabyte sized datasets," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 381–390.

[86] M. Salehi, C. Leckie, J. C. Bezdek, T. Vaithianathan, and X. Zhang, "Fast memory efficient local outlier detection in data streams," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3246–3260, Dec. 2016.

[87] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.

[88] N. Prykhodko and S. Prykhodko, "Constructing the non-linear regression models on the basis of multivariate normalizing transformations," *Electron. Model.*, vol. 40, no. 6, pp. 101–110, 2018.

[89] Z. Wang, C. Xiang, W. Zou, and C. Xu, "MMA regularization: Decorrelating weights of neural networks by maximizing the minimal angles," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020, pp. 1–12.

[90] Z. Engida, H. F. Neto, A. Slonimer, J. Bedard, F. S. Alam, and A. M. Snauffer, "Anomaly detection in complex data: A practical application when outliers are few," in *Proc. OCEANS*, Hampton Roads, VA, USA, Oct. 2022, pp. 1–7.

[91] A. Bondavalli, A. Ceccarelli, F. Brancati, D. Santoro, and M. Vadursi, "Differential analysis of operating system indicators for anomaly detection in dependable systems: An experimental study," *Measurement*, vol. 80, pp. 229–240, Feb. 2016.

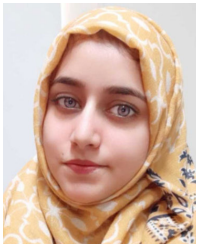
[92] V. Geppener and B. Mandrikova, "Combination of wavelet transform and autoencoder for complex data analysis and anomaly detection," in *Proc. Int. Conf. Inf. Technol. Nanotechnol. (ITNT)*, Sep. 2021, pp. 1–4.

[93] UCI. *Arrhythmia Dataset*. Accessed: Apr. 18, 2023. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/arrhythmia

[94] *Feature Selection*. Accessed: Apr. 18, 2023. [Online]. Available: https://jundongl.github.io/scikit-feature/OLD/datasets_old.html

[95] *BodyFat*. Accessed: Apr. 18, 2023. [Online]. Available: https://www.rdocumentation.org/packages/SIN/versions/0.6/topics/bodyfat

[96] CELAB. *Large-Scale CelebFaces Attributes (CelebA) Dataset*. Accessed: Apr. 18, 2023. [Online]. Available: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[97] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[98] *Elec2*. Accessed: Apr. 18, 2023. [Online]. Available: https://data.hellenicdataservice.gr/dataset/08dad1ab-728d-48e0-afa9-be0ce384b515l
[99] *GOutRank*. Accessed: Apr. 18, 2023. [Online]. Available: https://www.ipd.kit.edu/~muellere/GOutRank/
[100] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, pp. 1–39, 2007.
[101] *IBRl-S9*. Accessed: Apr. 18, 2023. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html
[102] Y. He, Z. Shen, and P. Cui, "Towards Non-I.I.D. image classification: A dataset and baselines," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107383.
[103] *NPRS44*. Accessed: Apr. 18, 2023. [Online]. Available: https://www.cs.ucr.edu/~eamonn/discords/nprs44.txt
[104] V. Losing, B. Hammer, and H. Wersing, "KNN classifier with self adjusting memory for heterogeneous concept drift," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 291–300.
[105] *Yahoo-S5*. Accessed: Apr. 18, 2023. [Online]. Available: https://webscope.sandbox.yahoo.com/catalog.php?datatype=s%5c&did=70
[106] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data Mining Knowl. Discovery*, vol. 30, no. 4, pp. 891–927, Jul. 2016.
[107] Y. Gong, T. Xu, X. Dong, and G. Lv, "e-NSPFI: Efficient mining negative sequential pattern from both frequent and infrequent positive sequential patterns," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 2, Feb. 2017, Art. no. 1750002.
[108] C. Lancho, I. Martín De Diego, M. Cuesta, V. Aceña, and J. M. Moguerza, "Hostility measure for multi-level study of data complexity," *Int. J. Speech Technol.*, vol. 53, no. 7, pp. 8073–8096, Apr. 2023.
[109] F. Lin and J. Chen, "Learning low-complexity autoregressive models with limited time sequence data," in *Proc. Amer. Control Conf. (ACC)*, Seattle, WA, USA, May 2017, pp. 3153–3158.
[110] E. Leyva, A. González, and R. Pérez, "A set of complexity measures designed for applying meta-learning to instance selection," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 354–367, Feb. 2015.
[111] F. J. Baldán and J. M. Benítez, "Complexity measures and features for times series classification," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119227.
[112] D. Chakraborty, V. Narayanan, and A. Ghosh, "Integration of deep feature extraction and ensemble learning for outlier detection," *Pattern Recognit.*, vol. 89, pp. 161–171, May 2019.
[113] D. Yang, Y. Wang, Y. Li, and X. Ma, "A variable Markovian based outlier detection method for multi-dimensional sequence over data stream," in *Proc. 17th Int. Conf. Parallel Distrib. Comput., Appl. Technol. (PDCAT)*, Dec. 2016, pp. 183–188.
[114] N. Y. Almusallam, Z. Tari, P. Bertok, and A. Y. Zomaya, "Dimensionality reduction for intrusion detection systems in multi-data streams—A review and proposal of unsupervised feature selection scheme," in *Emergent Computation* (Emergence, Complexity and Computation), vol. 24, A. Adamatzky, Ed. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-46376-6_22.
[115] D. P. Francis and K. Raimond, "A fast and accurate explicit kernel map," *Int. J. Speech Technol.*, vol. 50, no. 3, pp. 647–662, Mar. 2020.
[116] L. Cao, D. Yang, Q. Wang, Y. Yu, J. Wang, and E. A. Rundensteiner, "Scalable distance-based outlier detection over high-volume data streams," in *Proc. IEEE 30th Int. Conf. Data Eng.*, Mar. 2014, pp. 76–87.
[117] E. Fouché, F. Kalinke, and K. Böhm, "Efficient subspace search in data streams," *Inf. Syst.*, vol. 97, Mar. 2021, Art. no. 101705.
[118] S. Zhao, Z. Yu, T. G. Marbach, G. Wang, and X. Liu, "MSDN: A multi-subspace deviation net for anomaly detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Orlando, FL, USA, Nov. 2022, pp. 1341–1346.
[119] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2016.
[120] F. Zhang, H. Fan, R. Wang, Z. Li, and T. Liang, "Deep dual support vector data description for anomaly detection on attributed networks," *Int. J. Intell. Syst.*, vol. 37, no. 2, pp. 1509–1528, Feb. 2022.
[121] H. Trittenbach and K. Böhm, "One-class active learning for outlier detection with multiple subspaces," in *Proc. Conf. Inf. Knowl. Manag. (CIKM)*, W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, and J. X. Yu, Eds. New York, NY, USA: ACM, 2019, pp. 811–820.
[122] M. J. Bah, J. Zhang, T. Yu, F. Xia, Z. Li, S. Zhou, and H. Wang, "A generative adversarial active learning method for effective outlier detection," in *Proc. IEEE 34th Int. Conf. Tools Artif. Intell. (ICTAI)*, Macao, China, Oct. 2022, pp. 131–139.
[123] S. Vaishampayan, G. K. Palshikar, M. Apte, and V. Z. Attar, "Causality: An overlooked aspect in anomaly detection," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 2413–2417.
[124] K. Budhathoki, L. Minorics, P. Blöbaum, and D. Janzing, "Causal structure-based root cause analysis of outliers," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, MD, USA, vol. 162, Jul. 2022, pp. 2357–2369.
[125] E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, and S. Silvia, "A survey on outlier explanations," *VLDB J.*, vol. 31, no. 5, pp. 977–1008, Sep. 2022.



SHAFAQ SIDDIQI is currently pursuing the Ph.D. degree in computer science with the Graz University of Technology, Austria. Her research focuses on exploiting data characteristics for large-scale data pre-processing. She is also exploring the domains of heterogenous data preprocessing and challenges in non-IID data.

FAIZA QURESHI is currently a Research Assistant with Habib University, Pakistan. During the master's program, she focused on offline handwritten text recognition. She is also exploring the domain of textile defect recognition. Her research interests include machine learning and computer vision.

STEFANIE LINDSTAEDT has been the Managing Director of the Know Center, which is funded as part of the Austrian COMET Program, since 2011. Its mission is to promote data-driven business in Europe, to bring more IT skills into the economy and to promote talent. She is currently the first female Professor of computer science and the Director of the Institute for Interactive Systems and Data Science, Graz University of Technology. Under her leadership, Know Center has grown into one of Europe's leading research centers for the data-driven economy and artificial intelligence, supporting European companies of all sizes and sectors in turning data into value. Together with her team, she develops and improves AI technologies to support the safe and responsible use of data and to promote trust in these new technologies.

ROMAN KERN received the Ph.D. degree from the Graz University of Technology.

He is currently a Computer Scientist and an Associate Professor with the Institute for Interactive Systems and Data Science, Graz University of Technology. He is also the Chief Scientific Officer with the Know-Center Research Centre (Competence Centre for Trustworthy AI). His research interests include natural language processing and machine learning, with a focus on causal data science. He applies these methods to achieve trustworthy AI in fields like digital libraries, intelligent transportation systems, and smart production.