JOIV : Int. J. Inform. Visualization, 7(1) - March 2023 122-130

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

A Systematic Review of Anomaly Detection within High Dimensional and Multivariate Data

Syahirah Suboh [a,1], Izzatdin Abdul Aziz [a], Shazlyn Milleana Shaharudin [b,2], Saidatul Akmar Ismail [c,3], Hairulnizam Mahdin [d]

[a] Department of Computer and Information Sciences,Universiti Teknologi PETRONAS, Seri Iskandar, Perak, Malaysia
[b] Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak, Malaysia
[c] Faculty of Information Management, Universiti Teknologi MARA, Selangor branch, Puncak Perdana Campus, 40150 Shah Alam, Selangor, Malaysia
[d] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Johor, Malaysia
Corresponding author: [1]syahirah_19001683@utp.edu.my, [2]shazlyn@fsmt.upsi.edu.my, [3]saidatulakmar@uitm.edu.my

***Abstract*—In data analysis, recognizing unusual patterns (outliers' analysis or anomaly detection) plays a crucial role in identifying critical events. Because of its widespread use in many applications, it remains an important and extensive research brand in data mining. As a result, numerous techniques for finding anomalies have been developed, and more are still being worked on. Researchers can gain vital knowledge by identifying anomalies, which helps them make better meaningful data analyses. However, anomaly detection is even more challenging when the datasets are high-dimensional and multivariate. In the literature, anomaly detection has received much attention but not as much as anomaly detection, specifically in high dimensional and multivariate conditions. This paper systematically reviews the existing related techniques and presents extensive coverage of challenges and perspectives of anomaly detection within high-dimensional and multivariate data. At the same time, it provides a clear insight into the techniques developed for anomaly detection problems. This paper aims to help select the best technique that suits its rightful purpose. It has been found that PCA, DOBIN, Stray algorithm, and DAE-KNN have a high learning rate compared to Random projection, ROBEM, and OCP methods. Overall, most methods have shown an excellent ability to tackle the curse of dimensionality and multivariate features to perform anomaly detection. Moreover, a comparison of each algorithm for anomaly detection is also provided to produce a better algorithm. Finally, it would be a line of future studies to extend by comparing the methods on other domain-specific datasets and offering a comprehensive anomaly interpretation in describing the truth of anomalies.**

***Keywords*— Anomaly detection; high-dimensional data; multivariate data; information science.**

Manuscript received 18 Oct. 2022; revised 14 Jan. 2023; accepted 25 Jan. 2023. Date of publication 31 Mar. 2023.

I. INTRODUCTION

Anomaly detection has attracted much attention due to its importance in many areas, including network intrusion, credit card fraud, energy management, finance, statistics, process control, signal processing as well as machine learning [1], [2] [3]. Anomaly detection contributes to the early detection of irrelevant patterns or unusual events. The statement was supported by Ayesha, Hanif, and Talib [4], who stated that anomaly detection is useful for data pre-processing and cleaning for finding suspect data. Anomaly detection remains extensively researched, and by identifying anomalies, researchers can obtain vital knowledge which helps in getting a better understanding of the data [5]. Furthermore, it is good to have a fundamental understanding of the anomalies that could lead to better analysis and, at the same time, avoid any irrelevant effect on the data quality. Credit card fraud detection and processing loan applications are the most common anomaly detection applications. Subsequently, detecting and identifying anomalies help translate significant information for the application listed above [6], [7] [8].

In today's world, plenty of data is generated every minute, every second, due to the advancement of technology. All these lead to a big data era where the data is growing rapidly, and the recent developments also contribute to the huge data volume. An anomaly detection problem is generally not easy to solve and is difficult, especially within multivariate high-dimensional data. Various domains such as biomedical, web, education, medicine, business, and social media have been

found to apply multivariate high-dimensional data [9]. Previously, anomaly detection was conducted using statistical methods. Advanced to technology, huge demand on using machine learning because of the large data condition that traditional methods unable to cope.

However, if statistical and machine learning methods are blindly used on data containing anomalies, these methods may adversely affect the results obtained, such as model misspecification, biased parameter estimation, and eventually misleading results [10]. For many years, it has been recognized that there is an issue concerned with finding anomalies, specifically handling them [11]. It is crucial to be observant to grasp why anomalies should be discovered and what they represent. What is more, modern data are often high-dimensional, and traditional anomaly detection may face difficulties in handling high-dimensional data. Thu numerous machine learning methods have been developed to identify anomalies such as distance-based, clustering-based, density-basedand classification-based techniques [12], [13].

In addition, most of these methods also address the multivariate high dimensionality problem in anomaly detection, which poses a serious issue, leading to huge computational complexity, producing invalid results, and simultaneously taking the task more challenging [14]. The systematic review's development is based on the main research question: How does anomaly detection perform within high-dimensional and multivariate data?

This systematic review focuses on providing a detailed understanding of the problem of anomaly detection within high-dimensional and multivariate data. The keywords will be revolved around the problem (high dimensional and multivariate data) and techniques (Anomaly detection). For review and study, all works that are closely related to the keywords are examined. Hence, the contributions of this systematic review are twofold:

- Review the techniques of anomaly detection, specifically within high-dimensional and multivariate data.
- Discussing the recent research on managing the problems associated with high-dimensional and multivariate data.

The following is a breakdown of the paper's structure: Section 2 describes the difference between multivariate and high-dimensional conditions. Section 3 presents the introduction of high-dimensional and multivariate problems in anomaly detection. Section 4 discusses the material and methods with the illustrative diagram. In section 5, a discussion of the theoretical background and the formulation is given. Next, in section 6, the result and discussion for related methods are discussed. The last section concludes the paper with a discussion on the direction of future research.

A. Difference between Multivariate and High-dimensional

1) Multivariate: Multivariate data include two or more variables or features [15], [16]. It is quite difficult because multivariate data requires understanding the relationships between many variables, and usually, the human brain is overwhelmed by the sheer bulk of the data. On top of that, multivariate requires more mathematics than univariate to make an inference.

2) High dimensional: Dimensionality refers to a few variables, features,r attributes within a dataset greater than the number of observations [17]. High dimensional or the increase in dimension can lead to sparsity of data, resulting that the data have many counterintuitive properties which are more scattered and more isolated and poses a significant challenge for data analysis. This issue widely known as "curse of dimensionality" [18].

B. Anomaly Detection in High-Dimensional and Multivariate Data

Because of anomalies are unusual by definition and can differ significantly from one another, they present different concerns and challenges than regularly supervised classification [19], [20]. Despite this, anomaly detection algorithms have been effectively implemented in a variety of domains. On the other hand, numerous different methods have already been developed because each application area has its definition of abnormality and application limitations [21], [22]. The first half of the challenge is identifying anomalies; the second is interpreting anomalies that have been discovered. Often, real-world datasets will have a condition where some points behave differently from the rest of the datasets. It is very important to be able to detect anomalies, which may spoil the resulting analysis or may also contain valuable information. The statement was supported by research from Rousseeuw and Hubert [23, which emphasizes that errors may cause anomalies but could also belong to unusual circumstances. It is also implied that we should somehow investigate and understand them from various standpoints rather than remove them.

With the world increasingly data-driven and at the same time expanding of new technologies, the data collected gradually become huge in size and dimensionality. Most traditional anomaly detection methods are unable to cope with high-dimensional data. Research from Ayesha, Hanif, and Talib [4] also supported the claim that analyzing high-dimensional data has become a complex process. As dimensionality increases, the data becomes more sparse, causing difficulty in detecting and analyzing anomalies. When involving high-dimensional data, it is important to make a proper interpretation. The reason is that it helps the users evaluate an abnormal sample for each additional piece of information to understand the data fully.

Most recent anomaly detection techniques were developed for low-dimensional data sets and face difficulties as the dimensions increase. At the same time, as the dimension of data increases, the existing methods require high computational costs [24]. In addition, direct applications may produce invalid results. Numerous algorithms have been developed in numerical high-dimensional data over the years. Even though various techniques have been developed, it is vital to be aware that traditional anomaly detection is less significant as the dimension keeps increasing [25]. Besides, research from Kandanaarachchi and Hyndman [26] states that a feasible strategy for handling high dimensional data is by applying dimensional reduction methods to improve anomaly detection. Nowadays, anomalies are detected by using machine learning algorithms.

Alternatively, the performance of machine learning algorithms is negatively affected by high-dimensional data.

Even though machine learning algorithms are capable of task prediction, their performance is often restricted and sometimes will produce poor results to the quality of data representation, especially with high-dimensional data and more features condition. In addition, most real-world data have more than one feature, variables, and attributes widely acknowledged as multivariate data. Anomaly detection in multivariate data is increasingly important, especially in research. For instance, it is very important in some domain applications such as healthcare planning, factory systems, and transportation systems. When dealing with multivariate data, the data point may be inconsistent with the pattern of the main data. Thus, the anomaly may not be perceivable by an inspection. That kind of anomaly is identifiable using a statistical tool.

On the contrary, Statisticians developed various algorithms for anomaly detection, but most of the techniques only apply to univariate cases [27]. The process of determining anomaly is more complicated in multivariate datasets compared to univariate datasets. Even though many studies have explored anomaly detection methods, it only focuses on univariate datasets, and only a few have considered multivariate datasets. This leads to increased difficulty in anomaly detection.

Many attempts have been made to comprehend the issues of anomaly detection. However, it is not easy to detect it when anomalies are within multivariate and high-dimensional data. The statement is supported by research from Chen et al. [28], who state that finding anomalies in multivariate and high-dimensional data is becoming extremely difficult. The research work by Kim and Park [24] suggests that in multivariate high-dimensional data, considering the distance of an observation from the centroid as well as the shape of the data is required. Besides that, it is necessary to be aware of the number of features that need to be considered (univariate or multivariate); otherwise, eliminating anomalies of correct data might cause significant information loss. The research work from Foorthuis [29] also states that each variable in the multivariate dataset should be analyzed together to consider their relationship. This is another nature of anomalies in a multivariate condition which depends on the relationship between variables. On top of that, it cannot be easily detected by visualization techniques such as histograms, box plots, or scatter plots. The limitation of visualization techniques is only useful for up to 3D spaces and are not beneficial for dimensionalities more than 3D spaces [24].

II. Material and Methods

This section consists of four parts: (i) PRISMA, (ii) resource inclusion and exclusion criteria, (iii) the systematic review, and (iv) data abstraction and analysis for current research development. The details are henceforth.

A. PRISMA

PRISMA, or Preferred Reporting Items for Systematic Reviews and Meta-Analyses, serves as a "standard diagram tool" for developing systematic literature reviews. Research from Shaffril [30] also states that PRISMA is a well-established method for conducting a systematic literature review. On the other side, the diagram will help guide the researchers in filtering any irrelevant papers and only take the papers closely related to their focus into consideration for review.

B. Resources

The literature review of this study is conducted using various reliable databases, namely Scopus, Web of Science, Science Direct, Taylor Francis, and Hindawi. Accordingly, Scopus indexes 23715 journals corresponding to anomaly detection in the computer science field, followed by 4914 journals for Web of Science. Meanwhile, Science Direct and Taylor Francis have published a total of 955 and 1724 journals and articles directly related to anomaly detection.

C. The Systematic Review Process for Selecting the Articles

1) Identification: In the first stage of literature identification, several relevant articles for the review process are selected. It includes employing keywords and finding closely related articles based on terms, past works, and keyword identification. Currently, 254 articles from Scopus and Web of Science were retrieved from the research study's search string (Refer to Table I). A manual search using precise terms was also carried out on other databases such as Hindawi and Taylor Francis. Altogether, 298 articles were considered in the first stage of the systematic review process.

TABLE I
THE SEARCH STRING

Criterion	Database search string
Scopus	TITLE-ABS-KEY (("Anomaly detection" OR "Outlier detection") AND ("Multivariate") AND ("High Dimensional"))
WoS	TS=(("Anomaly detection" OR "Outlier detection") AND ("Multivariate") AND ("High Dimensional"))
Taylor and Francis	("Anomaly detection" OR "Outlier detection") AND ("Multivariate") AND ("High Dimensional")
Hindawi	("Anomaly detection" OR "Outlier detection") AND ("Multivariate") AND ("High Dimensional")

2) Screening: There were seven duplicate articles in the first stage and are now excluded before processing to the second stage. The resulted articles with a total of 291 were screened in the second stage to retain only the reliable and relevant ones based on the combination of inclusion and exclusion criteria obtained. The first criterion was the type of literature, with an emphasis on journals (research articles) as the primary source for constructing the comprehensive systematic review. Following that, it implies that any publication in the form of review, conference paper, book, book chapter were all excluded in the study. Next, the second criteria would focus on the articles published in English for this study. Other than that, five years period (2015-2020) were chosen for the timeline we considered reviewing in this study.

Moreover, only articles published in the subject area of computer science are selected for the retrieving process. These criteria ensure that the selection process is objective and help to limit unnecessary articles (Refer to Table II). Overall, 137 articles were removed according to the mentioned criteria.

TABLE II
THE INCLUSION AND EXCLUSION CRITERIA

Criterion	Eligibility	Exclusion
Literature type	Journal (research articles)	Book, book series, chapter in book, conference proceeding
Language	English	Non- English
Timeline	Between 2015 and 2020	Less than 2015
Subject area	Computer Science	Other than Computer Science

3) Eligibility: Finally, eligibility is the last stage, and 154 articles are prepared for the following process. At this stage, the articles would thoroughly inspect the titles, abstract, and main contents to ensure that they fit the requirements needed and that the current objective can be achieved. Unfortunately, a total of 132 articles were removed because the articles did not highlight the main term, which is anomaly detection within multivariate and high dimensional data. At the same time, the articles are more general and not specific, following our necessary related terms. Finally, a total of 22 articles have been prepared for analysis.

4) Data Abstraction and Analysis: All 22 articles will be analyzed to develop the appropriate findings based on the depth analysis through the framework below, as illustrated in Fig. 1. There will be an analysis from those 22 articles to extract information to answer the research questions in the first stage. Next, in the second phase, the information will be categorized based on the groups following the nature of the data. The researchers generally convert raw data into meaningful data through related information, specifically about ideas, concepts, and themes. Following that, the analysis shows three main themes: dimensional reduction approach, machine learning approach, and hybrid approach. In addition, there will be seven sub-themes named PCA [4], Random projection [10], DOBIN [26], Stray algorithm [31], ROBEM [32], DAE-KNN [33] as well as OCP method [34].

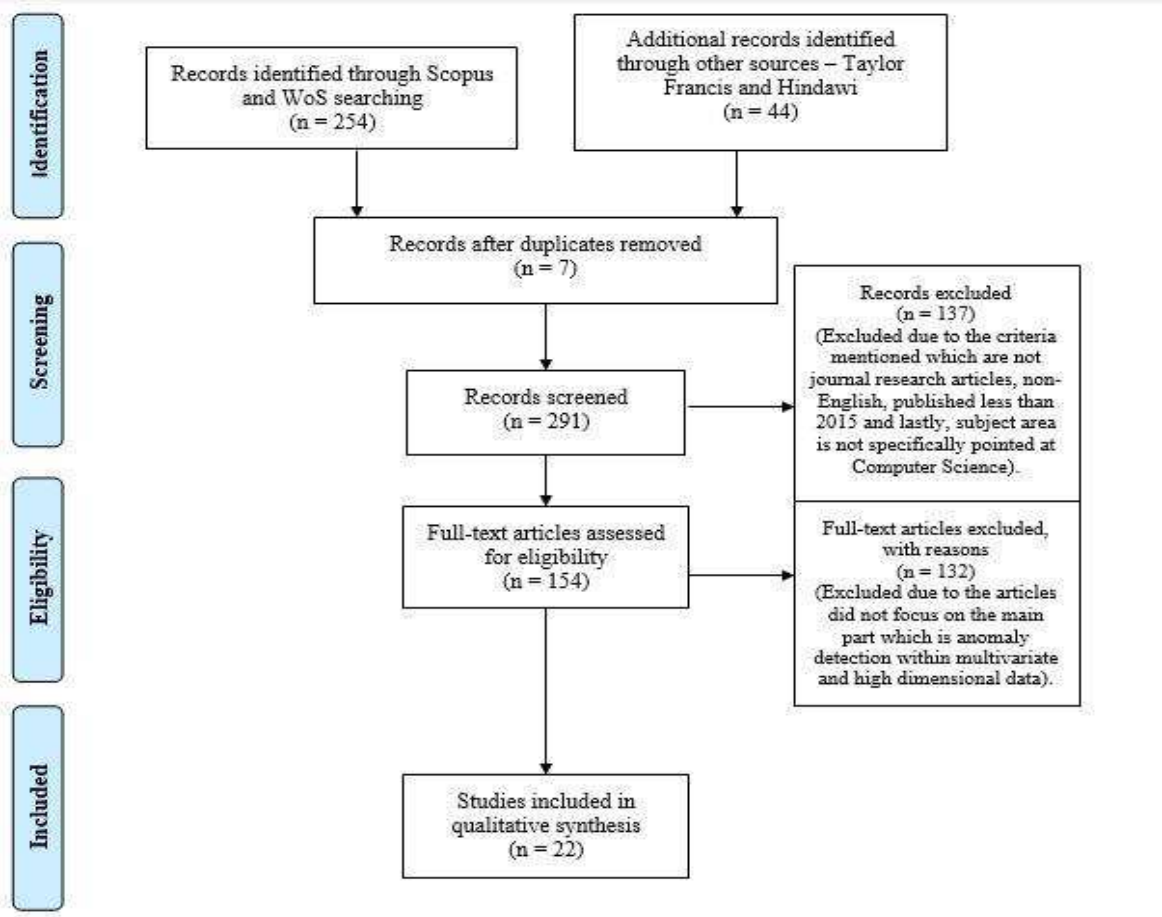

Fig. 1 Flow diagram of study, adapted from [35]

III. RESULT AND DISCUSSIONS

As discussed in "Anomaly Detection in High Dimensional and Multivariate Data", the two features that most have a significant effect on anomaly detection problems are "high dimensional" and "multivariate". The problem of high-dimensional and multivariate data not only makes it difficult to recognize anomalies but it also brings new obstacles such as computational cost [24], irrelevant results if direct applications are performed [33], inconsistent points with the primary data as well as sparsity of data [17]. Methods that address the problem of anomaly detection within high dimensional and multivariate data is summarized in Table III. Each method has advantages and disadvantages when it comes to fix various issues depending on the nature of the data.

TABLE III
COMPARISON OF THE PERFORMANCE OF ANOMALY DETECTION ALGORITHMS IN MULTIVARIATE AND HIGH DIMENSIONAL

Algorithm	Advantages	Disadvantages
PCA	Widely used due to the simplicity and efficiency	In high dimensional situations, evaluation usually difficult; the presence of anomaly can affect the performance of PCA
Random Projection	Any combination of sample sizes and dimensions can be used	There is no clear guideline on the number of preferred projections.
DOBIN	Helps in assisting the detection of anomaly using fewer components	Sensitive
Stray algorithm	Applicable for both one dimensional and high dimensional data, and the model building process does not need the use of training datasets	Must produce optimization on the best value of K
ROBEM	It uses critical value to detect anomaly. Thus, it leads to a successful performance concerning anomaly detection	Slowest algorithm
DAE-KNN	Reduces the computational cost and improves detection efficiency when compared to a single anomaly detector	Constructing the DAE is time-consuming if the data set is huge
OCP method	It is not necessary to estimate covariance, ideally suited to high dimensional data	Computational time is higher

In general, the strategies for tackling anomaly detection problems can be classified into several categories: dimensional reduction, machine learning, and hybrid.

A. Dimensional Reduction Approach

The process of finding low-dimensional features in high-dimensional data to remove high-dimensional data's barriers. It assisted in the reduction of the dataset's number of input variables. In simple words, converting a high-dimensional data representation into a low-dimensional data representation while keeping as much of the data's original meaning as feasible. Several applicable techniques can be used to reduce dimensionalities, such as principal component analysis (PCA), feature selection, genetic algorithm, linear discriminant analysis, and machine learning. Following that, Aremu et al. [36] research also state that data is converted using dimensional reduction to make a conducive data representation to accurately generate machine learning algorithm performance in other fields of study. The ability of dimensional reduction approaches to transform it becomes more straightforward from such complex data making the methods widely used for analyzing and visualizing high dimensional data [4].

1) Principal Component Analysis: The oldest and most popular approach has been proposed by Song et al. [33]. It is also known as one of the approaches capable of handling the high dimensionality problem. The statement is supported by research from Aremu et al. [36], which implies that PCA methods are frequently used to overcome the curse of dimensionality. PCA aims to extract all relevant factors from a data set and combine them into new orthogonal variables known as principal components [32]. These are linear combinations of correlated variables with fewer components than the original.

The first principal components represent a large amount of original data variance, following the second PC, which holds the second large variance. The method implies that the first PC holds the large variation at the start and reduces the dimension from p to k. Besides, they can be computed as a linear weighted combination of features.

2) Random Projection: Most methods in detecting anomalies within multivariate and high dimensional require the information of the covariance matrix. However, as the dimension of data increase, the more complex the estimation of the matrix become. Research from [10] proposed anomaly detection using random projection to avoid having to estimate the matrix. The proposed method employs projections as a technique for dimensionality reduction. In a way, it does not have to estimate mean and standard deviation.

3) Distance-Based Outlier Basis using Neighbours (DOBIN): Research from the paper Kandanaarachchi and Hyndman [26] proposed DOBIN that acts as a pre-processing strategy that can be applied by any anomaly detection method. It is common to use PCA to detect anomalies when high dimensional data. However, from the analysis results, DOBIN is preferred over PCA. What is more, DOBIN has two usages, first is making an easy way by only considering fewer components for anomaly detection. Secondly, another use of DOBIN is to help detect anomalies in the form of visualization. The basis construction for DOBIN is by maximizing K nearest neighbor (knn) distances.

In summary, DOBIN's key steps:

- Determine the Y space for a given dataset.
- Construct the basis.
- Transform the original space Z

B. Machine Learning Approach

Previously, anomaly detection was conducted using statistical methods. Following the big-data phenomenon, machine learning is a widely used technique due to the massive amount of data those traditional methods cannot handle. The excessive dimensionality of data can cause problems for machine learning models, such as accurate categorization, pattern identification, and presentation. Examples of machine learning approaches include linear regression, autoencoder-decoder, and clustering-based approaches.

1) Stray Algorithm: Stray taken from words Search and TRace AnomalY is proposed to overcome the limitation and enhance the capabilities of another anomaly detection method, HD outliers. The stray algorithm is a distance-based type of approach that uses Euclidean distances on the k-nearest neighbor searching. For each individual observation, compute the k-nearest neighbor distances of KNN, where i=1, 2…, k. After that, calculate the consecutive differences between distances. Then, take the k-nearest neighbor distance with the largest gap.

C. Hybrid Approach

Combining the machine learning approach with other techniques, such as statistical or other applicable techniques, is called the hybrid approach. In the early stages of anomaly detection, simple data analyses such as descriptive statistics may be performed to help identify anomalous observations to obtain insight into the data, which could eventually lead to modifications, including a combination of other techniques.

1) DAE with Ensemble KNN: Deep Autoencoder (DAE) is created using the Deep Belief Network (DBN) derived from RBM. On the other hand, RBM is an undirected graphical model made up of visible units v and hidden units h that represent observations and features. DAE tries to map high-dimensional data into a lower-dimensional feature space. The final decision will be on abnormal sample if it indicates 1 and normal sample if it indicates -1.

2) One Class Peeling (OCP) Method: The OCP approach is a flexible framework for detecting abnormalities in multivariate data that integrates statistical and machine learning methods. Kernel density and statistical distance techniques are incorporated into the strategy. Furthermore, it does not involve the computation of the covariance matrix. The OCP technique then incorporates a kernel distance measure between each observation and the center and robustly predicts the center. The formulation is given by determining the center of the multivariate data using an iterative peeling method based on boundaries derived from SVDD. A finite sample replacement breakdown point (FSRBP) is often used for robustness estimation. In summary, the key steps for OCP method:

- Determine threshold value, h.
- Compute the robust estimation using the SVDD with the Gaussian kernel function.
- Calculate the kernel distance between each observation vector and estimation of robustness on the data's center,
- Scale the distances.
- Mark observations larger than has a potential anomaly.

3) Robust Expectation Maximization (ROBEM): Many machine learning and statistical techniques have been developed to find anomalies. One way of identifying anomalies is through clustering. The clustering method is compelling in the field of machine learning. Research by Öner and Bulut [32] proposed a new clustering algorithm by combining EM clustering algorithm as well as robust principal component analysis (ROBPCA). Furthermore, the proposed method consists of two stages: 1) Anomalies are detected using the ROBPCA algorithm and 2) Dataset available is clustered using EM clustering algorithm. Following stage 1, the ROBPCA algorithm will take place to calculate principal component scores and orthogonal distances. In summary, key steps for ROBEM method:

- In stage 1, the anomaly detection takes place with the ROBPCA algorithm. Anomalies are defined as observations that exceed critical values for both score and orthogonal distances (as calculated from ROBPCA) and are sent to the anomaly cluster. In comparison, the cleaned data contains all remaining observations.
- Clustering occurred during the stage where observations in cleaned data were clustered using the EM algorithm.

D. Overview of Anomaly Detection

This detailed out the framework of the flow of anomaly detection within multivariate and high-dimensional data. The research framework, which comprises the following phases as outlined in Fig. 2:

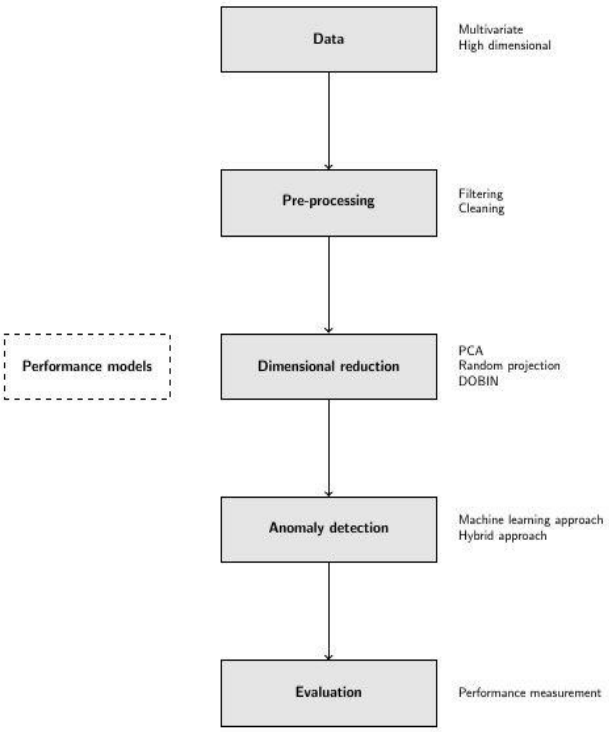

Fig. 2 General framework of anomaly detection [37]

1) Data: The data preparation phase where appropriate datasets are selected for anomaly detection. In this case, both multivariate and high-dimensional are considered.

2) Data Pre-processing: In this phase, multivariate and high-dimensional datasets were cleaned and filtered to make sure that there were no uncertainties and further divided into training and testing datasets.

3) Dimensional Reduction: The process of seeking low dimensional features of high dimensional data. Assisting in clearing the obstacles of high-dimensional data as most of the existing methods cannot perform well under high-dimensional conditions.

4) Anomaly Detection: The goal of anomaly detection is to investigate if there are anomalies in the data. The forms of output would be in the forms of scores and labels. Technically, the scores are sorted, and a threshold is chosen to designate anomalies. Meanwhile, labels are through a binary decision on whether the algorithm is an anomaly or not.

5) Evaluation: The model is integrated through the final phase. This phase is a critical step as it tests the reliability and generalizability of the model. Mostly, the performance will be measured by the area under the receiver operator characteristics (AUC), outlier detection rates (ODR), faulty classification rates (FCR), as well as the ROC curve, especially for classification tasks.

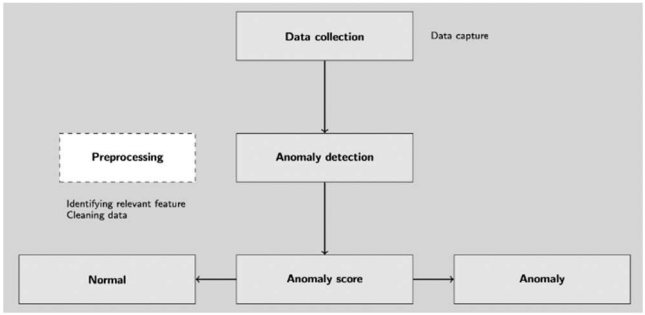

Fig. 3 Common anomaly detection phase

The common anomaly detection phase is stopped until the process of identifying abnormal and normal data [7]. There is no further explanation on whether the points are classified as anomaly which are meant to be removed, and large normal observations (extreme) as outlined in Fig.3. There are several previous approaches to anomaly detection as listed and extracted on "Result and Discussions". However, one crucial difference between some of those approaches and the case we are interested in is that there is no further explanation of the difference between anomaly and extreme observations [38], [39].

Different researchers have done many experimental tests to measure anomaly detection performance within multivariate and high-dimensional data [40]. Various performance metrics have been chosen to compare the performance respectively. A comprehensive comparative evaluation of various methods based on anomaly detection is presented in Table IV. For an extensive review, four characteristics are analyzed in detail for this review: learning rate, effective usage, efficiency, and resource requirement. The learning rate indicates the degree to which the proposed method is effective in learning. Meanwhile, effective usage describes the application domain of the technique, whether only applicable to multivariate, high dimensional, or both. Next, efficiency refers to the performance of the proposed method in contrast to the traditional one, and resource requirement refers to the computational requirements of the proposed method.

TABLE IV
COMPARATIVE ANALYSIS OF THE PROPOSED METHOD

	Learning Rate	**Effective Usage**	**Efficiency**	**Resource Requirements**
PCA	High	High dimensional	N/A	Low
Random projection	Mid	Both	More stable when the dimension varies	High
DOBIN	High	High dimensional	Better as a dimension reduction tool as compared to PCA.COV4	N/A
Stray algorithm	High	Both	Outperforms HDoutliers in terms of accuracy and computational time	Low
ROBEM	Mid	Both	More successful as compared to the existing one	High
DAE-KNN	High	High dimensional	Accurate as compared to standalone algorithms.	High
OCP method	Mid	Multivariate	Up to 88% more accurately on correctly classified	High

PCA, DOBIN, Stray algorithm, and DAE-KNN have a high learning rate that shows a perfect result and has been proven compared to Random projection, ROBEM, and OCP methods. Furthermore, most of the methods applicable for both conditions are multivariate and highly dimensional as these two conditions relate to each other and are interchangeable. If the methods are inefficient, they take too much time to detect anomalies. Based on the research reported, most methods have shown an excellent ability to tackle the curse of dimensionality and multivariate features in anomaly detection. Lastly, most of the methods also are very time-consuming. However, we believe that each method has its benefits regardless of the problem in time complexity.

IV. CONCLUSION

Overall, the study's focus is to review and discuss the recent research related to anomaly detection methods within multivariate and high-dimensional data. In addition, it also provides advantages and disadvantages of each method respectively so that a more reliable method can be developed. As summarized in a section of "Result and Discussion", it can be shown that each method serves the purpose rightfully. However, two problems in anomaly detection algorithms have been identified in this study. First, choosing a suitable reduction technique based on the data is essential in some

dimensional reduction approaches because sometimes vital information can be lost during the dimension reduction process. For instance, there are some shortcomings of PCA when there is noise. However, PCA and its modified variants, such as robust PCA and sparse PCA, are still widely used on many applications due to their simplicity and efficiency.

Meanwhile, for Random projection and DOBIN, the techniques act as dimensional reduction tools in data pre-processing to help any anomaly detection algorithm find anomalies. The development of the techniques is due to the lack of interpretation coming from traditional dimensional reduction techniques. On the other hand, the OCP method combines statistical and machine learning, focusing on detecting an anomaly in multivariate conditions. The last one would be the DAE-KNN, ROBEM, and Stray algorithm, a machine learning approach that applies to detect anomalies in multivariate and high dimensional conditions. The researcher established these methods not only to identify anomalies but also to enhance the capabilities of existing techniques by incorporating them into them. For example, for DAE-KNN, by combining autoencoder and K-nearest neighbor, ROBEM based on the ROBPCA and EM clustering algorithm, and lastly, Stray algorithm aims to improve the abilities of HDoutliers further. Second, most methods tackle identifying anomalies very well, but there is no proper test provided to know whether anomalies found are real anomalies or just large normal values. Following that, we should not somehow remove them but maybe investigate them properly, as anomalies are not necessarily errors.

After a study to compare the different methods for anomaly detection problems within multivariate and high dimensional data, the researchers continued to the next step. The next step is to formulate a more reliable anomaly detection algorithm that can perform well in multivariate and high dimensional data and can properly distinguish between anomalies that can have a poor impact on the data or anomaly that contains valuable information. Then, evaluate the proposed anomaly detection algorithm with the existing ones to compare when it comes to efficiency and accuracy. Implementing the proposed anomaly detection algorithms can help decision-making, improve performance, and solve various complex problems.

ACKNOWLEDGMENT

This research is supported by Universiti Pendidikan Sultan Idris (UPSI) through a Grant from Penyelidikan Universiti Fundamental (GPUF) 2020 (2020-0172-103-01).

REFERENCES

[1] M. Çelik, F. Dadaşer-Çelik, and A. Ş. Dokuz, "Anomaly detection in temperature data using dbscan algorithm," in *2011 international symposium on innovations in intelligent systems and applications*, 2011, pp. 91–95.

[2] R. Alguliyev, R. Aliguliyev, and L. Sukhostat, "Anomaly detection in Big data based on clustering," *Statistics, Optimization & Information Computing*, vol. 5, no. 4, pp. 325–340, 2017.

[3] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 131–146.

[4] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.

[5] A. Ukil, S. Bandyoapdhyay, C. Puri, and A. Pal, "IoT healthcare analytics: The importance of anomaly detection," in *2016 IEEE 30th international conference on advanced information networking and applications (AINA)*, 2016, pp. 994–997.

[6] L. Basora, X. Olive, and T. Dubot, "Recent advances in anomaly detection methods applied to aviation," *Aerospace*, vol. 6, no. 11, p. 117, 2019.

[7] M. A. Hayes and M. A. M. Capretz, "Contextual anomaly detection framework for big sensor data," *J Big Data*, vol. 2, no. 1, p. 2, 2015.

[8] A. Sreenivasulu, "Evaluation of cluster based Anomaly detection." 2019.

[9] X. Yang, Z. Wang, and X. Zi, "Thresholding-based outlier detection for high-dimensional data," *J Stat Comput Simul*, vol. 88, no. 11, pp. 2170–2184, 2018.

[10] P. Navarro-Esteban and J. A. Cuesta-Albertos, "High-dimensional outlier detection using random projections," *TEST*, pp. 1–27, 2021.

[11] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *Ieee Access*, vol. 7, pp. 107964–108000, 2019.

[12] N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, "Anomaly detection," *Computers, Materials and Continua*, vol. 14, no. 1, pp. 1–22, 2009, doi: 10.1145/1541880.1541882.

[13] D. Samariya and A. Thakkar, "A Comprehensive Survey of Anomaly Detection Algorithms," *Annals of Data Science*. Springer Science and Business Media Deutschland GmbH, 2021. doi: 10.1007/s40745-021-00362-9.

[14] Y. Yang, L. Chen, and C. Fan, "ELOF: fast and memory-efficient anomaly detection algorithm in data streams," *Soft comput*, vol. 25, no. 6, pp. 4283–4294, 2021.

[15] E. Uzabaci, I. Ercan, and O. Alpu, "Evaluation of outlier detection method performance in symmetric multivariate distributions," *Communications in Statistics-Simulation and Computation*, vol. 49, no. 2, pp. 516–531, 2020.

[16] R. A. Johnson, D. W. Wichern, and others, *Applied multivariate statistical analysis*, vol. 6. Pearson London, UK:, 2014.

[17] S. Thudumu, P. Branch, J. Jin, and J. J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *J Big Data*, vol. 7, no. 1, pp. 1–30, 2020.

[18] H. Liu, X. Li, J. Li, and S. Zhang, "Efficient Outlier Detection for High-Dimensional Data," *IEEE Trans Syst Man Cybern Syst*, vol. 48, no. 12, pp. 2451–2461, Dec. 2018, doi: 10.1109/TSMC.2017.2718220.

[19] V. S. L'vov, A. Pomyalov, and I. Procaccia, "Outliers, extreme events, and multiscaling," *Phys Rev E*, vol. 63, no. 5, p. 56118, 2001.

[20] X. Xu, H. Liu, and M. Yao, "Recent progress of anomaly detection," *Complexity*, 2019.

[21] K. Malik, H. Sadawarti, and K. G S, "Comparative analysis of outlier detection techniques," in *IJCA*, 2014, vol. 97, no. 8, pp. 12–21.

[22] D. Ghosh and A. Vogt, "Outliers: An evaluation of methodologies," in *Joint statistical meetings*, 2012, vol. 2012.

[23] P. J. Rousseeuw and M. Hubert, "Anomaly detection by robust statistics," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 2, p. e1236, 2018.

[24] J. M. Kim and C. S. Park, "Elimination of multidimensional outliers for a compression chiller using a support vector data description," *Sci Technol Built Environ*, vol. 27, no. 5, pp. 578–591, 2021.

[25] G. Horváth, E. Kovács, R. Molontay, and S. Nováczki, "Copula-based anomaly scoring and localization for large-scale, high-dimensional continuous data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–26, 2020.

[26] S. Kandanaarachchi and R. J. Hyndman, "Dimension reduction for outlier detection using DOBIN," *Journal of Computational and Graphical Statistics*, vol. 30, no. 1, pp. 204–219, 2021.

[27] S. Suboh and I. A. Aziz, "Anomaly Detection with Machine Learning in the Presence of Extreme Value-A Review Paper," in *2020 IEEE Conference on Big Data and Analytics (ICBDA)*, 2020, pp. 66–72.

[28] X. Chen, B. Zhang, T. Wang, A. Bonni, and G. Zhao, "Robust principal component analysis for accurate outlier sample detection in RNA-Seq data," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–20, 2020.

[29] R. Foorthuis, "On the nature and types of anomalies: a review of deviations in data," *Int J Data Sci Anal*, vol. 12, no. 4, pp. 297–331, 2021.

[30] H. A. M. Shaffril, A. A. Samah, S. F. Samsuddin, and Z. Ali, "Mirror-mirror on the wall, what climate change adaptation strategies are practiced by the Asian's fishermen of all?," *J Clean Prod*, vol. 232, pp. 104–117, 2019.

[31] P. D. Talagala, R. J. Hyndman, and K. Smith-Miles, "Anomaly detection in high-dimensional data," *Journal of Computational and Graphical Statistics*, vol. 30, no. 2, pp. 360–374, 2021.

[32] Y. Öner and H. Bulut, "A robust EM clustering approach: ROBEM," *Communications in Statistics-Theory and Methods*, vol. 50, no. 19, pp. 4587–4605, 2021.

[33] H. Song, Z. Jiang, A. Men, and B. Yang, "A hybrid semi-supervised anomaly detection model for high-dimensional data," *Comput Intell Neurosci*, vol. 2017, 2017.

[34] W. G. Martinez, M. L. Weese, and L. A. Jones-Farmer, "A one-class peeling method for multivariate outlier detection with applications in phase I SPC," *Qual Reliab Eng Int*, vol. 36, no. 4, pp. 1272–1295, 2020.

[35] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and others, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *Int J Surg*, vol. 8, no. 5, pp. 336–341, 2010.

[36] O. O. Aremu, R. A. Cody, D. Hyland-Wood, and P. R. McAree, "A relative entropy based feature selection framework for asset data in predictive maintenance," *Comput Ind Eng*, vol. 145, p. 106536, 2020.

[37] S. Anitha and M. Metilda, "An efficient and robust cluster based outlying points detection in multivariate data sets," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 2881–2885, 2018.

[38] V. Yepmo, G. Smits, O. Pivert, and V. Yepmo Tchaghe, "Anomaly Explanation : A Review Anomaly Explanation: A Review," 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03449887

[39] B. Rad, F. Song, V. Jacob, and Y. Diao, "Explainable anomaly detection on high-dimensional time series data," in *DEBS 2021 - Proceedings of the 15th ACM International Conference on Distributed and Event-Based Systems*, Jun. 2021, pp. 142–147. doi: 10.1145/3465480.3468292.

[40] T. Fujiwara, N. Sakamoto, J. Nonaka, K. Yamamoto, K.-L. Ma, and others, "A visual analytics framework for reviewing multivariate time-series data with dimensionality reduction," *IEEE Trans Vis Comput Graph*, vol. 27, no. 2, pp. 1601–1611, 2020.