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:

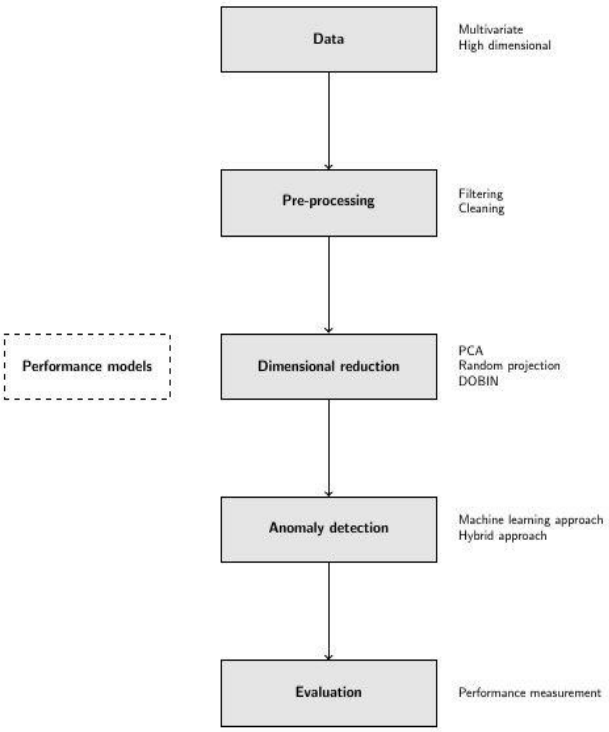

Fig. 2 General framework of anomaly detection [37]

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.

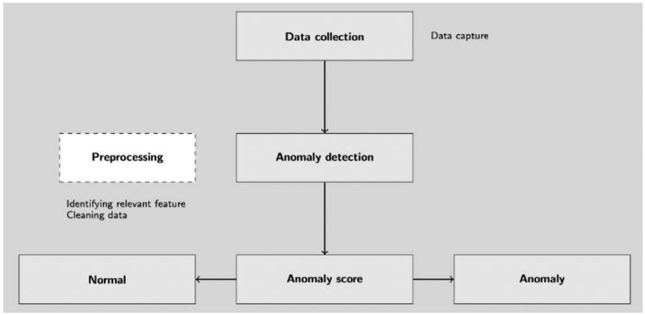

Fig. 3 Common anomaly detection phase

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.