An Anomaly Detection Model Using Principal Component Analysis Technique for Medical Wireless Sensor Networks

Nabeel Abdulrazaq Yaseen
Faculty of Basic Education, University of Missan, Iraq
nabilalrashy78@gmail.com

Abbas Abd-Alhussein Hadad
Faculty of Basic Education, University of Missan, Iraq
abbas@uomisan.edu.iq

Mustafa Sabah Taha
Missan Oil Training Institute,HSE Center, Ministry of Oil, Iraq
mustafa@moti.oil.gov.iq

2021 International Conference on Data Science and Its Applications (ICoDSA) | 978-1-6654-4303-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICODSA53588.2021.9617547

***Abstract*—A Wireless Medical Sensor Network (WMSN) connects autonomous nodes such as (sensors and actuators) existing on the body, or under a person's skin. The network usually extends over the entire human body and nodes are connected via a wireless communication channel. WMSN sense human physiological signs and monitor a patient's health status. The framework and preliminary experiment on developing an anomaly detection model for ubiquitous patient and healthcare monitoring in medical wireless sensor networks (MWSNs)were presented in this study. The architecture is a combination of an improved data mining method and machine learning algorithms using modern fusion methods. Being that MWSNs are highly susceptible to failures due to certain limitations, such as low energy resources, poor reliability, low computational resources, and considerable susceptibility to post-deployment security attacks, the proposed model is an anomaly detection method for MWSNs for the detection of anomalies in an adaptive manner with high accuracy while maintaining resource constraints using two phases. First, an anomaly-based detection model was created using Principal Component Analysis (PCA) technique to reduce dimensionality, prevent overfitting, and increase detection accuracy. Second, the detection accuracy of the proposed model was evaluated and compare before and after PCA integration. The experimental study showed that the proposed model can rapidly identify sensor anomalies with high accuracy.**

Keywords— medical wireless sensor network, anomaly detection, principal component analysis, detection accuracy, detection rate, data reduction

I. INTRODUCTION

With the continued increase in the average human lifespan, the number of old people has continued to increase, causing increases in healthcare-related costs and low patient-to-doctor ratio due to the ever-increasing demand for medical care [1]. Caregivers and healthcare providers have devised the remote monitoring method to cater to the increasing number of old persons and this has raised the interest in the use of WSNs in the healthcare sector [2]. Scientists and researchers have developed MWSNs as a network of wireless sensors consisting of several miniaturized sensors that can execute wireless transmission of data from their zones of deployment (connected or implanted) [3][4]. Hence, physicians rely on these devices to remotely monitor the vital signs of their patients even when they are far from the hospital; the sensed data is communicated to the related professional via a Control Processing Unit (CPU), such as smartphone, laptop, or tablet that has more processing power, larger batteries, and a wider range of transmission than the individual MWSN nodes. So, the CPU must have the capacity to process the received signals in real-time, as must be capable of raising medical alarms for caregivers when patients' health deteriorates, allowing them to respond immediately by taking relevant actions [5,6]. The processed data can also be sent to distant databases (DB) by the CPU for onward long-term analysis and storage. The advantages of MWSNs include helping healthcare givers in monitoring patients irrespective of their location, improving the efficiency and accuracy of diagnosis, and reducing the overall health-care related costs while allowing constant monitoring of patients [5-7]. The rate of disease detection can be improved by using MWSN and it can reduce the risk and impacts on people's lives during detection of dangerous diseases. Many medical WSN systems are commercially available, including Tmote Sky, MICAz, MICA2, IRIS, Imote2, TelosB, & Shimmer for the monitoring of vital signs like heart rate (HR), pulse, oxygen saturation (SpO_2), respiration rate (RR), body temperature (BT), electrocardiogram (ECG), blood pressure (BP), electromyogram (EMG), blood glucose levels (BGL), etc. However, these systems are essentially devices for collecting and reporting crucial data, and may not guarantee data security; hence, they must be equipped with intrusion detection systems to guarantee data security. Furthermore, data reduction techniques that aid in increasing battery life are not offered in these systems.

While MWSNs have several advantages, they often have several drawbacks, such as low reliability, limited energy resources, low computational power, and considerable susceptibility to security attacks upon their deployment. Therefore, an adaptive data reduction solution is required to suit the resource constraints demands of medical sensor devices in terms of computational complexity while incurring lower approximation error of original data. To improve the strengths of these devices and minimize the chances of any weakness, it is important to first investigate the weaknesses in further detail to discover some ways of mitigation. MWSN sensor nodes are vulnerable to data dynamic changes that can affect their efficiency. In general,

the MWSN has problems such as sensor calibration, faulty components, dislocation, and battery exhaustion [8,9].

This paper aims to develop a highly efficient dimensionality reduction method that can be used to find accurate AD model for MWSN to detect anomalies with high accuracy while maintaining resource constraints. To prove the research hypothesis, the following objectives are presented. To prove the research hypothesis, the following objectives are presented:

a) To propose and develop an anomaly-based detection model by employing Principal Component Analysis (PCA) technique for dimensionality reduction (this technique is able to reduce the dimensionality of the medical data and increases detection accuracy).
b) To evaluate the detection accuracy of the suggested method by implementing the model and compare the results before and after integrating PCA.

The rest of this article is arranged as follows: Section 2 presents the proposed method, while Section 3 discusses the results of the experiments. Section 4 presents the conclusion of the study.

II. RESEARCH METHOD

The proposed study begins with a review of the literature. The goal of investigating the state-of-the-arts is to look for approaches that have been implemented in handling the problem of AD in MWSN in the existing works. It has been determined that PCA is a highly efficient dimensionality reduction method that can be used to find accurate AD solutions. This research involves three phases, each of which contributes to the next. Figure 1 depicts a high-level detail of the entire framework. In Phase 1, a literature review is undertaken to determine the research problem. The outcome of this phase is the research gaps that will be addressed in the subsequent phases. Based on the identified problem and gaps, Phase 2 carries out the first objective of this study which is dimensionality reduction in healthcare data using the PCA technique. The reduced vital data is the result of this phase. Phase 2's reduced medical data will be employed as an input for Phase 3. In Phase 3, the reduced medical data will be applied to train the MWSN anomaly detection model, and then, the evaluated accuracy will be compared to the existing methods. The output of this phase is a comparison of the proposed model's accuracy with PCA and without PCA technique.

There are two stages to the implementation of the dimensionality reduction proposed in the first objective; these are offline training and online implementation. Data was collected via measurements throughout the training period. Several pre-processing operations were performed on the collected medical data, including standardization, noise removal, normalization, and imputation. The PCA technique was used in the reduction procedure to get the reduction parameters, such as eigenvectors and eigenvalues. To ensure that the dimension of the incoming data from the node is reduced, these parameters are conserved in the sensor before the implementation phase. They are also used to calculate the measurements in the cluster head (CH).

During the online implementation phase, PCA relies on the stored eigenvectors and eigenvalues determined during the training phase to performs real-time dimensionality reduction on the medical data. Then, the low dimensional medical data is forwarded to the CH where it is restored using an approximation in its original form based on the same parameters used during the reduction step. If the observed changes in medical parameters are beyond a set threshold, the medical data is recalculated. The readings of the reduced medical data that replaced the original data during the detection stage are the outcome of this phase. Consequently, the communication overhead, memory usage, and computational complexity of the suggested AD model were dramatically decreased during the real-time process. This improves the worthiness of the suggested AD models as it is trained in limited medical data. Section 3 is the implementation specifics of the PCA technique and its incorporation into the suggested model for AD.

For the second objective, PCA was incorporated into the AD model for MWSN with two stages - offline training, and online detection. During offline training, the training medical data were collected from all sensors that constitute the MWSN. These medical data underwent the pre-processing step for standardization, normalization, and noise removal. After that, the PCA technique proposed in the first objective was used to carry out the dimensionality reduction. logistic regression algorithm was trained using the reduced data to construct the classifier that distinguishes between the normal behaviour and the malicious one based on the normal reference inferred from the medical data. This normal reference was then stored at each node for later use during online detection The implementation of this model is discussed in detail in section 3.

Procedures	Action
1- Study the Literature	Determine the related problems
2- Design and implement the PCA	➢ Design the proposed dimensionality reduction technique.
3- Train the MWSN anomaly detection model	➢ Training the model. ➢ Validation. ➢ Testing. ➢ Coding.
4- Experiments and evaluation	➢ Result. ➢ Discussion.

Fig. 1. The high- level detail of the entire proposed framework

For the online detection phase, the new vital data observed by any node in the MWSN are collected. Similar to the training phase, the new observation undergoes pre-processing activities like standardization, normalization, imputation, and noise removal. PCA technique was also used for dimensionality reduction. After that, the observation was compared against the normal reference stored at the offline training in the respective node to identify whether it is normal. It is worth noting that both training and detection phases took place locally at the node nodes level. The outcome of this objective is the efficient online anomaly detection model that trained the logistic regression algorithm using reduced data from the proposed first objective. The

design and implementation of this model were elaborated in section 3.

III. Result Discussion and Evaluations

This section discussed the proposed model's design and implementation, as well as the facts acquired as a result of its execution. The proposed model was described in Section 3.1. The suggested PCA-based anomaly detection model's design was discussed in Section 3.2, while Section 3.3 discussed the evaluation metrics. Finally, the experimental results and comparison were detailed in Section 3.4

3.1. PCA Technique-Based Anomaly Detection Model

This study implemented a similar transformation procedure as found in most previous studies. The value of each feature was normalized to a range of 0 and 1. Note that the transformation and standardization of all medical data occur at the CH in the centralized scenario. But in the distributed scenario, a summary of the collected data by each node is sent to the CH and this information contains the number, linear sum, & linear sum of squares of the local data vectors. Furthermore, each node transmits the minimum and maximum values of each feature to the CH for the computation of the global max, global min, global mean, and global variance after receiving this data. The sensor nodes receive these global values and use them for their local data pre-processing. Having a reduced number of features in a dataset improves the performance of the anomaly detector via speeding up the detection process and improving its accuracy.

The hierarchical or cluster-based network structure was used at the present study, as illustrated in Figure 2. The network was divided into clusters in this design, with every cluster having an strategy node called the CH; this CH has extra processing power and energy and is responsible for data processing and propagation from the other sensors to the base station. The use of these clusters and the assignment of specific tasks to the CHs significantly improves the scalability, energy efficiency, and network lifetime of the system. After applying PCA locally, other sensor nodes (SNs) participate in the dimensionality reduction process, as well as provide only a summary of their reduced data to CH. The CH, which operates as a sensor node, performs an approximation of the original data before sending it to the sink. Each cluster's nodes are assumed to be static, homogeneous, and time-synchronized.

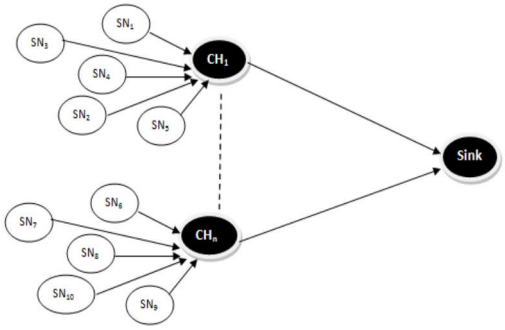

Fig. 2. The hierarchical or cluster-based sensor network structure

3.2. The design of the suggested PCA-based AD model

In this phase, the PCA technique was applied in each node to extract the principal components that represent the medical data in the feature space. This phase was carried out offline after collecting sufficient data from the sensors. The features extracted by PCA were then fed to a machine learning classifier and train the detection model using the logistic regression algorithm that can distinguish between normal and anomalous medical data. The same procedure was applied online as the technique is lightweight in terms of computation and requires no intensive operations. The eigenvectors and eigenvalues are the products of the PCA process and were used as the features for the training of the ML classifier model. The pseudocode of the initialization step of the model is shown in Algorithm 1 while Figure 3 depicted the proposed model's design.

Algorithm 1. The pseudocode of the initialization step of the model

Input: Data collected from sensors
Output: The classification of the data as normal or anomalous

##Training Phase (offline)
1: Collect data from sensors.
2: Do data normalization and Standardization
3: Extract the raw features from data
4: Apply PCA on the raw data and features
5: select the best n eigenvector as PCA features.
6: feed the selected PCA features into logistic regression classifier.
##Testing phase (online):
7: read the new measurement (data) from the sensor
8: apply PCA to newly obtained data
9: Feed the data with PCA features into the classifier
10: The classifier will determine whether the measurement is normal or anomalous.

3.3. Evaluation Metrics

The performance of the proposed PCA-based dimensionality reduction technique was evaluated in terms of the approximation error, and approximation accuracy; these were the popular performance evaluation metrics used by most of the previous studies to determine the effectiveness of novel PCA-based dimensionality reduction techniques [11],[12]. Furthermore, the effectiveness of the suggested AD system for MWSN were evaluated based by several metrics, such as the detection rate (DR), false-positive rates (FPR), detection accuracy (DA), as well as false-negative rates (FNR); these metrics were adopted from previous studies [13]. The efficiency of the new method was also evaluated based on the communication overhead, memory usage, and computational complexity as obtainable in the related [14]. The DA, DR, precision, and F-measure were calculated using Equations 1, 2, 3, & 4, respectively based on the confusion matrix and related evaluation metrics for evaluating anomaly detection models.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$DR = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

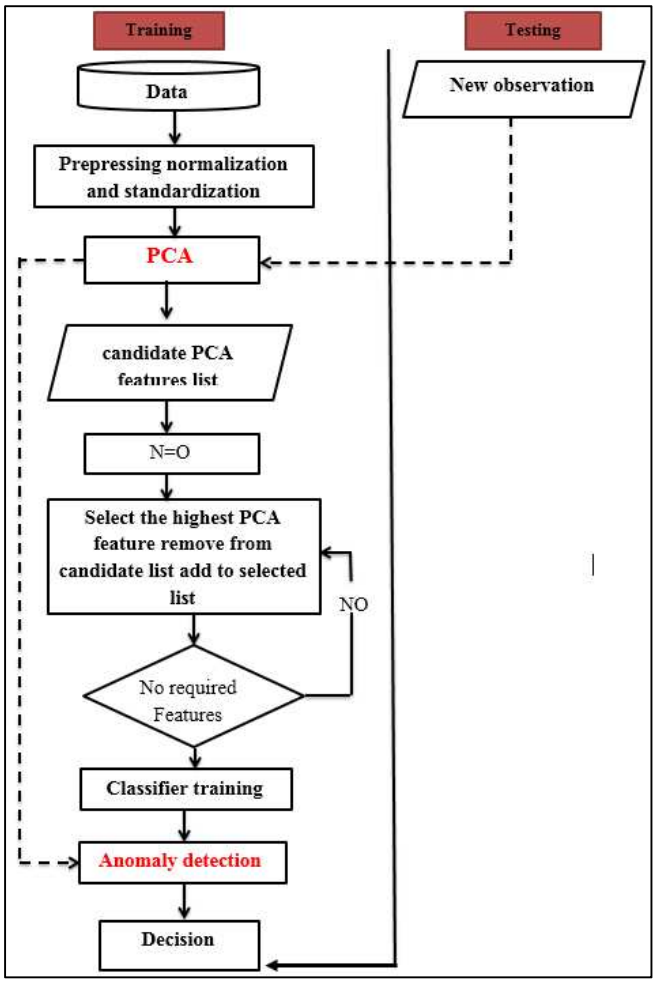

Fig. 3. The proposed model's design

4. Result Analysis and Comparison

Figures 4, 5, and 6 showed the distribution of the medical data gathered from the sensor nodes. The medical data were not normally distributed as seen in Figures 4, 6, and 7 due to the randomness of medical data as body temperature, heart rate, and blood pressure were all shaped by a multitude of variables. These variables are uncontrollable; furthermore, these three sensors captured medical data that are relatively associated. As a result, variability in one of these critical facts causes randomness in the others. The medical data collected by the body temperature sensor, on the other hand, was distributed normally due to the measurement's predictability and the restricted range of values for body temperature.

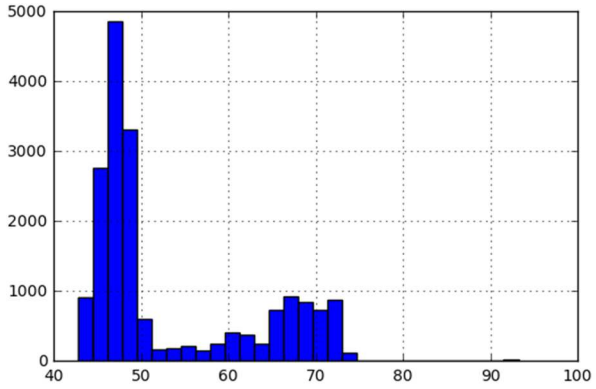

Fig. 4. The distribution of medical data collected by the Oxygenation ratio sensor.

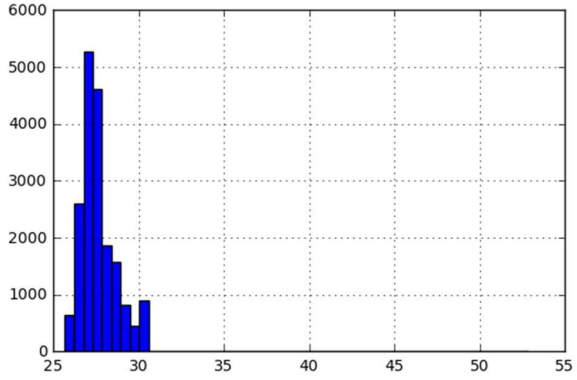

Fig. 5. The distribution of medical data gathered by the body temperature node

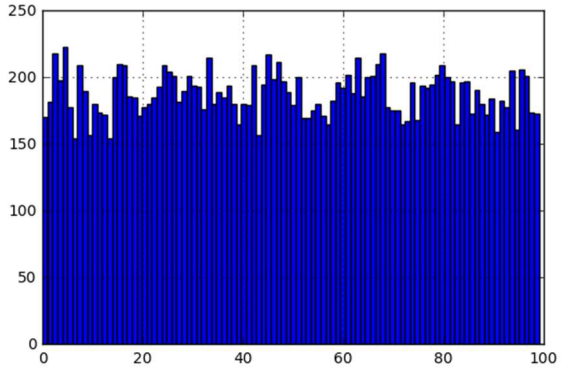

Fig. 6. The distribution of medical data gathered by the blood pressure node

The performance of the suggested AD method with and without the PCA technique was shown in Figure 8. As we can see, the suggested PCA-based system performed better with PCA as the values of the precision and recall of the proposed F1 model were greater than the values for the model without PCA. This improved performance is due to the data dimensionality capability of PCA as it selected the most

relevant features for use by the detection model. The impact of this data dimensionality reduction capability on the model performance is that the problem of overfitting which deteriorates DA is prevented. Furthermore, the performance of the model in terms of accuracy was slightly higher with PCA than without PCA because of the issue of high false alarms associated with the traditional AD techniques. The extent of improvement achieved by the new model over the conventional models was determined by checking for the t-test value at the α value of 0.05. Observably, the model achieved a p-value of 0.03 which was <0.05 [15], suggesting the significance of the suggested PCA-based AD model.

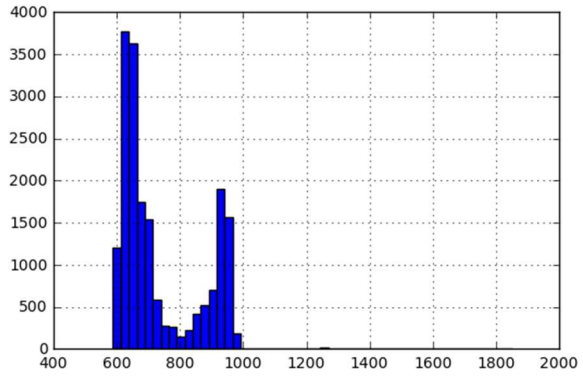

Fig. 7. The distribution of medical data gathered by the heart rate node

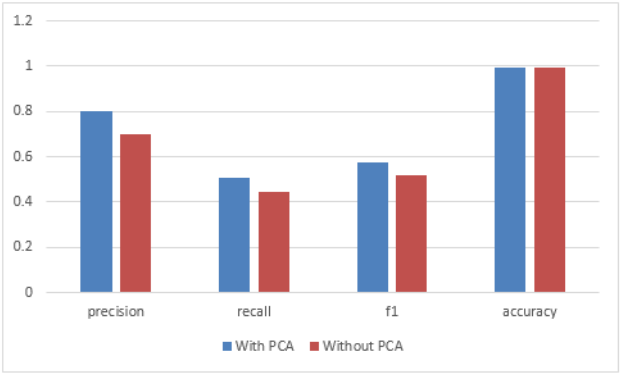

Fig. 8. The comparison of the performance of the proposed model with and without the PCA-based AD technique.

5. Conclusion

This study designed and implemented a PCA-based AD model for MWSNs. The model was built in two main phases which are the training and testing phases. In the training phase, data were collected from different sensors and pre-processed. During the pre-processing step, several procedures were implemented on the data, such as data normalization and standardization. PCA was implemented on the data to reduce the dimensionality and prevent overfitting. The features extracted by PCA were fed to a machine learning classifier for the training of the detection model using the logistic regression algorithm. The developed model with and without the PCA-based dimensionality reduction step was evaluated and compared in terms of performance using various performance metrics.

Acknowledgment

We would like to express our very great appreciation to Dr. Hiyam N. Khakid for her valuable and constructive suggestions during the planning and development of this research work. Her willingness to give her time so generously has been very much appreciated.

We would also like to thank the staff of the following institutions for enabling us to visit their Labs to do our operations:

- University of Misan
- Missan Oil Training Institute

References

[1] M. A. M. El-Bendary, H. Kasban, A. Haggag, and M. A. R. El-Tokhy, "Investigating of nodes and personal authentications utilizing multimodal biometrics for medical application of WBANs security," *Multimed. Tools Appl.*, vol. 79, no. 33, pp. 24507–24535, 2020.

[2] G. Shanthi and M. Sundarambal, "FSO–PSO based multihop clustering in WSN for efficient medical building management system," *Cluster Comput.*, vol. 22, no. 5, pp. 12157–12168, 2019.

[3] S. Izza, M. Benssalah, and K. Drouiche, "An enhanced scalable and secure RFID authentication protocol for WBAN within an IoT environment," *J. Inf. Secur. Appl.*, vol. 58, p. 102705, 2021.

[4] A. S. H. Altamimi, O. R. K. Al-Dulaimi, A. A. Mahawish, M. M. Hashim, and M. S. Taha, "Power minimization of WBSN using adaptive routing protocol," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 2, pp. 837–846, 2020.

[5] M. Salayma, A. Al-Dubai, I. Romdhani, and Y. Nasser, "Wireless body area network (WBAN) a survey on reliability, fault tolerance, and technologies coexistence," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 1–38, 2017.

[6] Taha, M. S., Rahim, M. S. M., Hashim, M. M., & Khalid, H. N. (2020). Information Hiding: A Tools for Securing Biometric Information. *Technology Reports of Kansai University*, *62*(04), 1383-1394. [7] M. S. Hajar, M. O. Al-Kadri, and H. K. Kalutarage, "A survey on wireless body area networks: architecture, security challenges and research opportunities," *Comput. Secur.*, p. 102211, 2021.

[8] S. J. Hussain, M. Irfan, N. Z. Jhanjhi, K. Hussain, and M. Humayun, "Performance enhancement in wireless body area networks with secure communication," *Wirel. Pers. Commun.*, vol. 116, no. 1, pp. 1–22, 2021.

[9] S. GS and R. Balakrishnan, "A Statistical-Based

Light-Weight Anomaly Detection Framework for Wireless Body Area Networks," *Comput. J.*, 2021.

[10] A. El Aalaoui and A. Hajraoui, "Energy efficiency of organized cluster election method in wireless sensor networks," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 1, pp. 218–226, 2020.

[11] Alrikabi, Hanan Ali, et al. "Using FFNN classifier with HOS-WPD method for epileptic seizure detection." *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*. IEEE, 2019.

[12] Y.-A. Le Borgne, S. Raybaud, and G. Bontempi, "Distributed principal component analysis for wireless sensor networks," *Sensors*, vol. 8, no. 8, pp. 4821–4850, 2008.

[13] M. Xie, J. Hu, S. Han, and H.-H. Chen, "Scalable hypergrid k-NN-based online anomaly detection in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 8, pp. 1661–1670, 2012.

[14] M. Moshtaghi *et al.*, "Clustering ellipses for anomaly detection," *Pattern Recognit.*, vol. 44, no. 1, pp. 55–69, 2011.

[15] J. Wang, J.-M. Wei, Z. Yang, and S.-Q. Wang, "Feature selection by maximizing independent classification information," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 4, pp. 828–841, 2017.

AUTHOR BIOGRAPHY

Mustafa Sabah Taha is a senior researcher at Missan Oil Training Institute. He earned his Ph.D. degree in Information Security from University Technology Malaysia (UTM) in 2020. During his Ph.D. study, he was accorded several honorable awards as recognition for his level of excellence and tenacity, such as GOT (Graduate-on-Time) award, and the Best Researcher award from UTM. His research work has been published in several reputable academic journals, book chapters, and refereed conference proceedings. His main research interest is in Image Processing, Information Security, Wireless Sensor Networks, Wireless Body Area Network and Internet of Things (IoT). He is a member of IEEE since 2017.