Isolation Forest Based Anomaly Detection Approach for Wireless Body Area Networks

Murad A. Rassam[1,2](✉)

[1] Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
m.qasem@qu.edu.sa

[2] Faculty of Engineering and Information Technology, Taiz University, 6803 Taiz, Yemen

Abstract. Anomalous data detection is an important task for ensuring the quality of data in many real-world applications. Medical healthcare services are one such application where Wireless Body Area Networks (WBAN) is used to track human health situations. Such tracking is achieved by collecting and monitoring the basic physiological vital signs and making them available to the healthcare givers to assess the criticality status of patients, especially in Intensive care units (ICU). Various anomaly detection approaches have been proposed for detecting anomalies collected in WBAN such as statistical, machine learning and deep learning techniques. However, the lack of ground truth data made the job of training such models a difficulty in supervised settings. In this paper, an Isolation Forest-based anomaly detection approach for WBAN (iForestBAN-AD) model is proposed. The iForest technique is fully unsupervised and does not employ any distance measure or density function like most existing techniques and rather detects anomalies based on the concept of isolation. To evaluate the proposed approach, experiments on data samples from real world physiological network records (Physionet) were conducted. The results show the viability of the proposed approach as it achieves around 95% AUC and outperforms many of the existing baseline unsupervised techniques on multivariate dataset samples.

Keywords: Anomaly Detection · Wireless Body Area Network · Isolation Forest · Unsupervised Learning · Internet of Medical Things

1 Introduction

Remote and pervasive vital signs monitoring become a necessity in societies where the average lifetime increases and the number of elderly people who need continuous monitoring are exponentially increasing especially in Europe. Such an increase creates an overload in the healthcare sectors and urges the need for pervasive systems that can monitor large numbers of patients easily. Furthermore, the increase of patients who require ICU admission and monitoring requires automated systems to handle the continuous monitoring of patients in such units and facilitates the decision-making process by doctors and healthcare givers.

A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 1347–1357, 2023.
https://doi.org/10.1007/978-3-031-27409-1_123

Internet of Medical Things (IoMT) is the concept of collecting, analyzing and storing health-related data by tiny sensors that constitute the body area sensor networks. Such data includes many vital signs observations such as blood pressure (BP), oxygen saturation (SPO2), and pulse rate among others [1]. Figure 1 shows different sensors implanted over the human body to measure the vital signs used to monitor the health condition of patients at home or in the ICUs.

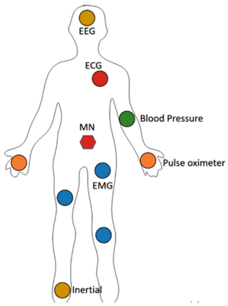

Fig. 1. Wireless Body Area Networks [2]

Ensuring the quality of collected data in WBANs for healthcare monitoring applications is a prominent research area in which the anomaly detection concept is employed to detect anomalous observations that arise due to various reasons. Several anomaly detection approaches for WBANs have been introduced in the literature based on statistical, machine learning and other techniques such as [3–6]. However, such approaches employed techniques that are computationally heavy and therefore require a considerable amount of time that can be critical for the healthcare monitoring situation. In addition, some of the existing approaches have not considered the situation of more than one parameter and monitor individual signs separately.

To this end, this paper considers the situation of detecting anomalous observations in multivariate healthcare data by utilizing the concept of isolation. To achieve this goal, the Isolation Forest (iForest) algorithm is employed where 6 vital signs of healthcare data recorded at the ICU are considered altogether to build a model for efficient detection. The isolation concept as explained and employed in [7] can achieve a low linear time complexity and a small memory requirement because it does not depend on any distance measure calculations.

The contribution of this paper is as follows:

- Proposing a new anomaly detection model for WBAN based on the iForest technique.
- A comparative analysis of the proposed model with baseline existing models in the literature.

The rest of this paper is organized as follows: Sect. 2 reviews and analyzes the literature on anomaly detection for WBAN. Section 3 introduces the proposed model and presents a background on the isolation concept. Section 4 presents the experimental evaluation results and compares the proposed model with existing literature. Section 5 concludes this paper.

2 Related Works

Various anomaly detection schemes have been proposed to detect the abnormal readings collected by WBANs to facilitate accurate decisions by healthcare givers. Such schemes were designed based on different approaches such as machine learning approaches [8, 9], statistical approaches [3, 10–13] and game-based approaches [14] among others. Statistical approaches can be used in two modes: parametric-based and non-parametric-based. Similarly, Machine learning approaches are used in supervised and unsupervised settings. However, the lack of ground truth data to train ML models in supervised modes and the nature of the anomaly detection problems made the unsupervised approaches a choice.

Unsupervised ML models such as [15, 16] have been used to detect anomalous data observations in WBANs. In such models clustering algorithms such as K-Means, hierarchal clustering and fuzzy C-Means clustering techniques are used. However, in such models, authors assumed that the clusters are well distinguishable and therefore a distinct line between normal and anomalous readings is clear. This assumption is not realistic in the case of physiological readings where the abnormal readings for a patient can be considered normal for another.

A study in [3] introduced an approach for detecting continuous changes in readings such as modifications, forgery, and insertions in electrocardiogram (ECG) data. A Markov model with different window sizes (5% and 10%). Using only univariate data, the study reported 99.8% and 98.7% of true negatives with 5% and 10% windows size, respectively. The Markov-based models usually have small time execution but the space complexity is high.

Two types of correlation exist in healthcare data and any other time series data which are temporal and spatial correlation. Temporal correlation refers to the strong relationships between data observations of the same variable according to the time stamp. Spatial correlation refers to the relationship between more than one variable at the same time stamp. Authors in [4, 17] considered the correlation that exists among data observations temporally and spatially. The results of the proposed approaches are found to be enhanced when both types of correlation are utilized. However, those approaches have high computational complexities and cannot be utilized in real-time.

In [13], authors proposed a model for anomaly detection in WBAN by adopting the data sampling approach with the Modified Cumulative Sum (MCUSUM) technique. The proposed approach aimed to enhance the speed of detection by the sampling method, while the use of the MCUSUM algorithm aimed to enhance the security of detection. Although the proposed approach in this work enhances the detection efficiency, the stationary process incurred by MCUSUM made it difficult to detect random and emergent anomalies. In addition, the linear statistical-based approaches are always parametric which makes them unsuitable for real-world applications.

One-class machine learning approaches such as [18, 19] are the most suitable unsupervised learning approaches for anomaly detection in sensor systems. It depends on the availability of normal observations to build a model that can detect any abnormalities in future sensor readings. An example of such approaches. However, most such approaches are implemented to consider only univariate variables of the physiological measurements separately.

Isolation forest-based approaches have been utilized in the literature as good candidates to develop efficient anomaly detection models. In [7] authors employed the concept of isolation to detect anomaly detection efficiently without the need of using distance functions. The experimental evaluation of the iForest proposed in this study using different datasets shows that the iForest approach outperforms the one-class support vector machine (OCSVM) and the local outlier function (LOF) approaches. Furthermore, authors in [20] designed a model for distributed anomaly detection in wireless sensor networks based on the isolation principle. The authors claimed that the proposed isolation principle approach helps to reduce the computational complexity and therefore reduces the energy consumption besides it achieves better detection accuracy results as it utilizes the spatial and temporal correlation in a distributed fashion. Another study [21] utilized the isolation principle in combination with the concept of drifting to detect anomalies in data streams. Both studies [20, 21] evaluated the isolation principle using several real-world datasets to report its viability.

To the best of our knowledge, the isolation principle has not been used for detecting anomalies in the context of the data streams collected by WBAN. Therefore, in this paper, we aim to prove that the isolation forest (iForest) algorithm is a good candidate for detecting anomalies in the context of WBAN more effectively compared to existing models in the literature.

3 Proposed Approach

To discuss the design of the proposed iForestBAN-AD model, we adopt the scenario that several m sensors nodes are placed on multiple different positions of the patient body to collect various physiological observations of the vital signs as in Fig. 2.

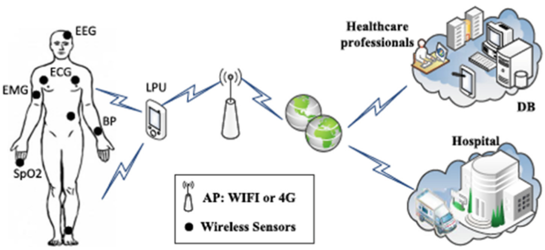

Fig. 2. A scenario for WBAN deployment [12]

As shown in Fig. 2, the collected observations are sent to the Local Processing Unit (LPU) which has enough resources for processing data received from sensors and detects anomalous observations before sending the data to the healthcare professionals or hospital management. If the observations are found to be anomalous, an alarm or any kind of notifications is sent to the healthcare givers to check if it is a sign of health degradation or faulty measurements.

3.1 Principle of Isolation Forest

According to [7], the term, isolation, refers to "separating an instance from the rest of others". The principle of isolation-based anomaly detection models is to measure each data observation's susceptibility to being isolated whereas anomalies are those that have this degree of susceptibility. To model the idea of isolation, a tree structure of the observations that naturally isolates data is modelled. This tree is composed of random binary trees of recursively partitioned observations. To detect anomalous instances, we rely on the fact that anomalies tend to form shorter path trees because anomalies are less in the count and then it results in a smaller number of partitions represented by shorter paths in a tree structure. Moreover, anomalies are those instances with distinguishable feature values which are to be separated early in the partitioning process. As a result, when a shorter path length of the forest of random trees is produced for some points, these points are most likely to be anomalies.

3.2 IForestBAN-AD Model

Figure 3 presents the different phases of the proposed *iForestBAN-AD* model. The details of the proposed model and its different phases are given in the following subsections.

Data collection/loading: in this stage, the real-world physiological dataset samples are collected and loaded to train and evaluate the proposed model. More details on the dataset will be given in Sect. 4.

Data preprocessing: some preprocessing steps are applied to the dataset samples to make them suitable for machine learning operations. Such steps include removing null values of the data observations, scaling the data in the range (0,1) using the min and max functions, and labelling the data observations as 0/1 classes. Such labelling will be used only for the evaluation processes to test the efficacy of the proposed model in terms of accuracy of detection. It is worth mentioning that the iForest algorithm is trained unsupervised, as we will see in the following subsection, to determine the anomaly score that will be used later to decide the abnormality of the data instances.

iForestBAN-AD Engine: the engine of the proposed anomaly detection model based on iForest algorithm consists of two main stages: the training stage and the evaluation stage. The description of the iForest algorithm is adopted from [7].

Let $X = \{x1,\ldots,xn\}$ be the dataset observations. A sample of instances $X' \subset X$ is used to build an isolation tree (iTree). The sample X' is divided recursively by the random selection of an attribute q and a split value p, the process continues until one of the stop criteria is fulfilled which are: (1) one instance only remains in the node, or (2) the node had data items of the same values.

An iTree by definition is a binary tree, in which every node in that tree is composed of zero or two daughter nodes.

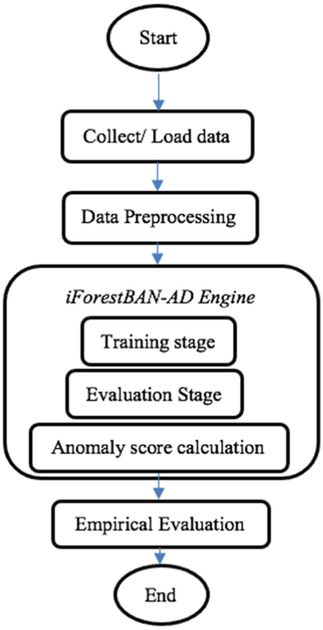

Fig. 3. The *proposed iForestBAN*-AD Model

1. **Training stage**

In this stage, iTrees are formed by dividing a subsample X' recursively until all instances are isolated. A simple abstraction of the training process is given in the pseudocode in Table 1.

The details of iTree(X') procedure of constructing the iTrees can be found in [7]. As the result of the training process, a number of trees (forest) are returned and is ready to be tested.

2. **Evaluation stage**

In the testing stage, as detailed in [7], a single path length h(x) is derived by counting the number of edges e from the root node to an external node as instance x traverses through an iTree.

3. **Anomaly Score Calculation**

The anomaly score s of an instance x is defined as:

$$s(x, \psi) = 2^{\frac{-E(h(x))}{c(\psi)}} \quad (1)$$

Table 1. Pseudocode of Training iForest Algorithm

Algorithm 1 : *iForest(X,t,ψ)*
Inputs: X - input data, t - number of trees, ψ - subsampling size
Output: a set of t iTrees
1: Initialize Forest
2: for i=1 to t do
3: X′ ←sample(X,ψ)
4: Forest← Forest ∪ iTree(X′)
5: end for
6: return Forest

where E(h(x)) is the average of h(x) from a set of iTrees, h(x)single path length, c is the average path length of unsuccessful searches.

4 Experimental Evaluation

After the anomaly score is obtained, a set of experiments using different dataset samples is implemented to verify the efficacy of the proposed model. Before describing experiments and reporting the results, a short description of the dataset is given.

Figure 4 shows the variation of data observations for subject 330 in the MIMIC II. In Fig. 4a, all features are depicted, whereas in Fig. 4b–g the variations in Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Mean Blood Pressure, Pulse Rate, and Oxygen Saturation, respectively. Some features are highly correlated such as HR and Pulse which are identical. Furthermore, a high positive correlation is noticed between ABPDias in Fig. 4d and ABPMean in Fig. 4e.

To evaluate the proposed iForestBAN-AD model, Samples from subjects 330 and 441 are used. The proposed model was developed using various python libraries such as pandas, Numpy, and Sklearn. The settings of the parameters of the iForest algorithm used in this research is reported in Table 2.

The Area under Curve (AUC) scores and the MAE metrics are used to evaluate the proposed model, and to compare its performance with existing unsupervised models in the literature namely One-Class Support Vector Machines (OSVM), KMeans, and the Local Outlier Factor (LOF). For the evaluation of the proposed model as an unsupervised

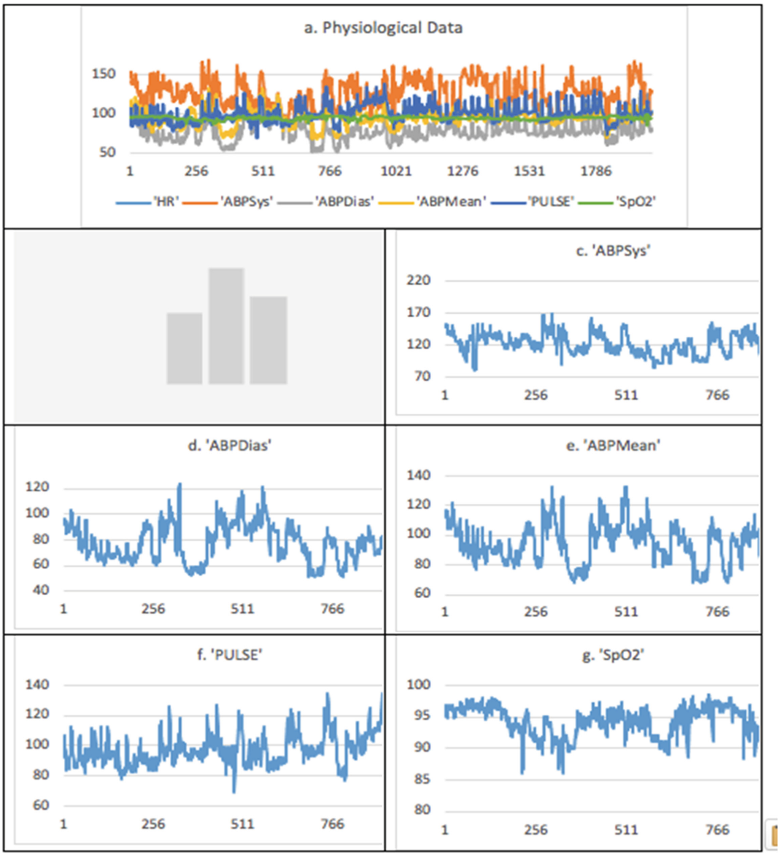

Fig. 4. Sample of physiological data for Subject 330

Table 2. Parameter settings of the iForest algorithm used in this rpaper.

Parameter	Value
n_estimators	100
max_samples	'auto'
contamination	variable
max_features	1.0
random_state	42
verbos	0

model, we assume that anomaly labels are unavailable in the training stage and are only available in the evaluation stage to calculate the evaluation measure, AUC.

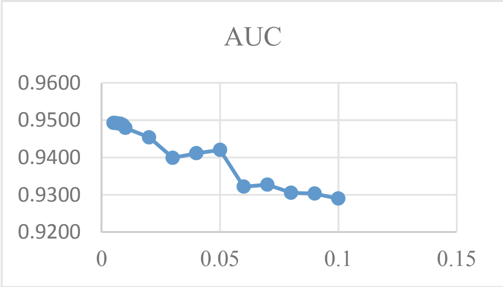

Fig. 5. Accuracy versus contamination ratio for Subject 330 records.

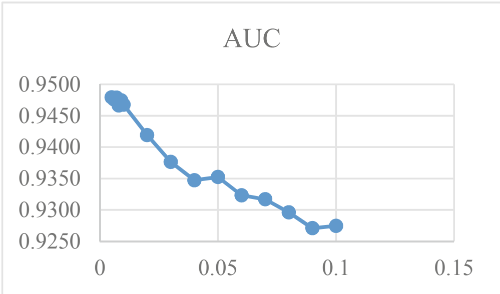

Fig. 6. Accuracy versus contamination ratio for Subject 441 records.

Figures 5 and 6 show the AUC scores of the proposed model on subject 330 and subject 441, respectively. The AUC scores are depicted versus the contamination parameter which is found to be the only parameter that slightly affects the performance. The contamination parameter controls the threshold for deciding when a scored data instance should be considered an anomaly. The experimental evaluation clearly shows that the best achieved AUC score is around 95% for both data subjects which indicates that the performance is stable with different data samples. The results further show that the AUC increases by a decrease in the contamination parameter which controls the ratio of anomalous measurements in the subsample. The number of trees t is made constant and equals 100. The average Mean Absolute Error (MAE) reported is 0.228 and 0.227 for subject 330 and subject 441 records, respectively. It is also noticed that the t value of the MAE decreases with the decrease of contamination parameter for this dataset records.

The performance of the proposed model was empirically compared with 3 existing unsupervised anomaly detection models proposed in the literature named KMeans, LOF and the OCSVM using the subject 330 records as shown in Table 2. The comparison shows that the proposed model outperforms those models as it achieves 95% AUC compared to 72% for OCSVM, 51% for LOF and 41% for KMeans. The OCSVM scored second and the best of the other candidates whereas KMeans is the worst (Table 3).

Table 3. Comparison of AUC on Subject 330 records

	OCSVM	KMeans	LOF	iForest (Proposed)
AUC	0.72	0.41	0.51	0.95

5 Conclusion

Ensuring the quality of vital signs observations collected by WBANs is crucial to facilitate taking timely and accurate decisions by healthcare givers in the IoMT applications. In this paper, the unsupervised iForest algorithm was used to design a model for detecting anomalous data observations in WBANs and therefore ensuring data quality. The experimental evaluation on a real-world physiological data records has proved that the proposed approach outperformed existing baseline unsupervised approaches. Furthermore, the concept of isolation reduces the computational burden that is usually resulted from the employment of machine and deep learning models as it does not require distance measures calculations. In future, the concept drifting of data needs to be investigated together with the isolation concept in order to consider the context of patient in near real time.

References

1. Santos, M.A., Munoz, R., Olivares, R., Rebouças Filho, P.P., Del Ser, J., de Albuquerque, V.H.C.: Online heart monitoring systems on the internet of health things environments: A survey, a reference model and an outlook. Inf. Fusion **53**, 222–239 (2020)
2. Al-Mishmish, H., Alkhayyat, A., Rahim, H.A., Hammood, D.A., Ahmad, R.B., Abbasi, Q.H.: Critical data-based incremental cooperative communication for wireless body area network. Sensors 18, 3661 (2018)
3. Khan, F.A., Haldar, N.A.H., Ali, A., Iftikhar, M., Zia, T.A., Zomaya, A.Y.: A continuous change detection mechanism to identify anomalies in ECG signals for WBAN-based healthcare environments. IEEE Access **5**, 13531–13544 (2017)
4. Mohamed, M.B., Makhlouf, A.M., Fakhfakh, A.: Correlation for efficient anomaly detection in medical environment. In: 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 548–553. IEEE, (Year)
5. Salem, O., Serhrouchni, A., Mehaoua, A., Boutaba, R.: Event detection in wireless body area networks using Kalman filter and power divergence. IEEE Trans. Netw. Serv. Manag. **15**, 1018–1034 (2018)
6. Saneja, B., Rani, R.: An integrated framework for anomaly detection in big data of medical wireless sensors. Mod. Phys. Lett. B **32**, 1850283 (2018)
7. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation-based anomaly detection. ACM Trans. Knowl. Disc. Data (TKDD) **6**, 1–39 (2012)
8. Lau, B.C., Ma, E.W., Chow, T.W.: Probabilistic fault detector for wireless sensor network. Expert Syst. Appl. **41**, 3703–3711 (2014)
9. Saraswathi, S., Suresh, G.R., Katiravan, J.: False alarm detection using dynamic threshold in medical wireless sensor networks. Wirel. Netw. **27**(2), 925–937 (2019). https://doi.org/10.1007/s11276-019-02197-y

10. Zhang, H., Liu, J., Pang, A.-C.: A Bayesian network model for data losses and faults in medical body sensor networks. Comput. Netw. **143**, 166–175 (2018)
11. GS, S., Balakrishnan, R.: A Statistical-based light-weight anomaly detection framework for wireless body area networks. Comput. J. **65**, 1752–1759 (2022)
12. Salem, O., Alsubhi, K., Mehaoua, A., Boutaba, R.: Markov models for anomaly detection in wireless body area networks for secure health monitoring. IEEE J. Selec. Areas Commun. **39**, 526–540 (2020)
13. Boudargham, N., El Sibai, R., Bou Abdo, J., Demerjian, J., Guyeux, C., Makhoul, A.: Toward fast and accurate emergency cases detection in BSNs. IET Wirel. Sensor Syst. **10**, 47–60 (2020)
14. Arfaoui, A., Kribeche, A., Senouci, S.M., Hamdi, M.: Game-based adaptive anomaly detection in wireless body area networks. Comput. Netw. **163**, 106870 (2019)
15. Ahmad, B., Jian, W., Ali, Z.A., Tanvir, S., Khan, M.: Hybrid anomaly detection by using clustering for wireless sensor network. Wirel. Pers. Commun. **106**, 1841–1853 (2019)
16. Qu, H., Lei, L., Tang, X., Wang, P.: A lightweight intrusion detection method based on fuzzy clustering algorithm for wireless sensor networks. Advances in Fuzzy Systems 2018, (2018)
17. Albattah, A., Rassam, M.A.: A correlation-based anomaly detection model for wireless body area networks using convolutional long short-term memory neural network. Sensors **22**, 1951 (2022)
18. Shahid, N., Naqvi, I.H., Qaisar, S.B.: One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh environments. Artif. Intell. Rev. **43**(4), 515–563 (2013). https://doi.org/10.1007/s10462-013-9395-x
19. Zhang, Y., Meratnia, N., Havinga, P.: Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks. In: 2009 international conference on advanced information networking and applications workshops, pp. 990–995. IEEE, (Year)
20. Ding, Z.-G., Du, D.-J., Fei, M.-R.: An isolation principle based distributed anomaly detection method in wireless sensor networks. Int. J. Autom. Comput. **12**(4), 402–412 (2015). https://doi.org/10.1007/s11633-014-0847-9
21. Togbe, M.U., Chabchoub, Y., Boly, A., Barry, M., Chiky, R., Bahri, M.: Anomalies detection using isolation in concept-drifting data streams. Computers **10**, 13 (2021)