

Received November 19, 2019, accepted December 9, 2019, date of publication December 13, 2019, date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959739

HLMCC: A Hybrid Learning Anomaly Detection Model for Unlabeled Data in Internet of Things

NUSAYBAH ALGHANMI¹, REEM ALOTAIBI¹, AND SEYED M BUHARI¹

Department of Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Nusaybah Alghanmi (nalghanmi0020@stu.kau.edu.sa)

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (DG 40-611-1440).

ABSTRACT The Internet of Things (IoT) is a network of distributed devices or sensors connected through the internet to allow gathering and sharing of data. The data generated by these devices is affected by anomalies or abnormal behaviour due to attack issues, or breakdown in devices, as examples. However, most current anomaly detection systems rely on labelled data, while the class labels for IoT data are usually unavailable. Furthermore, the manual labelling task is expensive and time-consuming to perform due to the need for domain experts. More importantly, the volume of data in the IoT is growing rapidly, creating a need to predict the classification labels for future data. This study proposes a Hybrid Learning Model which uses both Clustering and Classification methods (HLMCC) to automate the labelling process and detect anomalies in IoT data. The model consists of two practical phases, automatic labelling and detecting anomalies. First, the HLMCC groups the data into normal and anomaly clusters by adopting Hierarchical Affinity Propagation (HAP) clustering. Second, the labelled data obtained from the clustering phase is used to train the Decision Trees (DTs) and to classify future unseen data. The results show that the HLMCC is able to automate the labelling of data, which is beneficial to minimize human involvement. Moreover, HLMCC outperforms the DTs on the originally labelled datasets and the state-of-the-art model over a wide range of evaluation metrics based on the average ranks. HLMCC produces the highest average ranks against other models in terms of False Positive Rate (FPR), recall, precision and the Area Under the Precision-Recall curve (AUCPR) with 1.8, 1.6, 1.8 and 1.8, respectively.

INDEX TERMS Anomaly detection, Internet of Things, machine learning, unlabelled data, sensor data.

I. INTRODUCTION

Many organizations and academic sectors need to pay attention to the Internet of Things (IoT), owing to the current potential to automate our lives “smartly”. The IoT is a network of distributed devices or sensors connected through the internet to allow the gathering and sharing of data. The IoT has been found in several application domains such as smart homes, wearable devices, smart cities, health care, agriculture, transportation, and industrial sectors of industry.

IoT devices generate data that may behave inconsistently owing to abnormal or anomaly behaviour as a result of attack issues or breakdown in devices, as examples. An anomaly, in this context, means an abnormality in the data that differs from the predicted pattern [1]. The characteristics of an anomaly are: different from the norm and occurring rarely in

the data [2]. Anomaly detection is the technique of identifying rare observations which do not follow the expected behaviour. It can be applied within different domains and diverse industries such as intrusion detection, fraud, and fault detection, as examples.

The common technique for performing anomaly detection involves the use of machine learning algorithms. This helps to improve the performance of the system by learning from and using data from previous experiences. There are three types of machine learning task, which are supervised, unsupervised, and semi-supervised learning. Supervised learning trains the model based on predefined labelled data, while unsupervised learning finds similarities between unlabelled data. However, semi-supervised learning deals with partially labelled data to build the model.

Most current anomaly detection systems rely on labelled data which may not be available or it is time-consuming and expensive to produce. In addition, the data collected from IoT

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés¹.

devices usually lack the class label and form as unlabelled data. Moreover, the volume of IoT data is growing at an increasingly rapid rate, creating a need to predict, detect, and classify any anomaly for future unseen data. To overcome these limitations, this paper proposes a Hybrid Learning anomaly detection Model in IoT that employs Clustering and Classification approaches called HLMCC.

The HLMCC model consists of two functional phases: automatic labelling and detecting anomalies. In the automatic labelling phase, Hierarchical Affinity Propagation (HAP) clustering is applied to automate the labelling process, which helps to address the issue of unlabelled data and can be helpful in reducing human intervention. In detecting anomalies, the obtained labelled data is used to train the Decision Trees (DTs) to detect and classify future unseen data. The HLMCC model is evaluated over different clustering and classification validation techniques such as silhouette coefficients and recall, respectively.

In particular, the main contributions of this article are as follows:

- To propose the HLMCC model based on clustering and classification approaches to automate data labelling and to detect anomalies in IoT.
- Label the data by employing the HAP clustering algorithm.
- Compare HAP with other clustering algorithms such as Inverse Weight Clustering (IWC).
- Compare HLMCC against the DTs on the originally labelled data and the existing model.
- For reproducibility purposes, the implementation of the proposed HLMCC model is available in GitHub.¹

This paper represents an extension of the published paper in the Proceedings of the 6th Swiss Conference on Data Science [3]. This extended version improves the results in detecting anomalies for the proposed model against DTs on the originally labelled data, and the state-of-the-art model, applies the Synthetic Minority Over-sampling technique (SMOTE) by adding synthetic data to the original data, and describes the proposed algorithm in more detail. Also, it presents expanded experimentation that focuses on detecting positive samples correctly, with new experiment settings and results, and using additional different evaluation metrics such as the Area Under the Precision-Recall curve (AUCPR).

The paper is organized as follows: Section II provides a review of the machine learning workflow for anomaly detection in IoT. It then explains the existing anomaly detection approaches in IoT. Sections III, IV and V present the proposed HLMCC model, the experimental set-up, and discuss the experiment results, respectively. Finally, Section VI presents the conclusion of the paper.

A. MOTIVATION

The motivation for this work is the widespread use of day-to-day devices and sensors, such as weather sensors, in IoT, meaning more data can be collected and analysed. This data is

the foundation of any system for making intelligent decisions on future actions and plans. The availability of labelled data is considered one of the major challenges in anomaly detection systems [1], while the class labels (normal/anomaly) for IoT data are usually unavailable.

Generally speaking, machine learning algorithms (supervised, unsupervised or semi-supervised) provide effective methods for detecting, identifying, and classifying anomalies. The main benefit of all approaches is the ability to learn from data that is widely available in the IoT environment. Therefore, it is a challenging task to build an anomaly detection model that relies entirely on supervised learning algorithms, due to the cost of producing the labelled data [1], [4].

Furthermore, Cisco estimates that the data generated from the IoT will reach 847 Zettabytes (ZB) by 2021, but the stored amount will be relatively small, at approximately 7.2 ZB {Brazdil, 2000 #150} [5]. It means that the IoT data is growing quickly and there is a need to predict anomalies in new and future data, which, from a practical perspective, are limited in unsupervised learning algorithms.

II. LITERATURE REVIEW

To construct a machine learning model, there are diverse elements which should be considered, such as datasets, the type of learning algorithm, feature selection and evaluation techniques. For anomaly detection, the data is collected from IoT devices and placed into data storage. Then, machine learning techniques (supervised, unsupervised or semi-supervised) are implemented. Finally, validation techniques are used to evaluate the performance of the model.

The existing solutions based on machine learning algorithms (supervised, unsupervised or semi-supervised) for anomaly detection in IoT are described in the following sub-sections.

A. SUPERVISED LEARNING

Supervised learning builds a model based on predefined labelled data. Training and testing are the two phases for supervised learning [6]. In the training phase, we build the model using the training data, while in the testing step, the trained model provides the class label for unseen data. There is a wide range of learning algorithms such as Neural Networks (NNs), Support Vector Machines (SVMs) and K-nearest neighbours (KNNs).

Different learning approaches such as single, ensemble or hybrid models are explained by Tsai *et al.* [7], that are used to classify the data into either normal or abnormal classes. Single models consist of a single classifier such as KNNs, SVMs, NNs, whereas ensemble models improve the system performance by including diverse weak classifiers, while hybrid models consist of more than two classifiers in a model such as the neuro-fuzzy model. Two main steps are performed in hybrid models [8]: firstly, the model uses the data to produce the intermediate results. Secondly, the intermediate results are used as input to output the final results.

¹https://github.com/nrghanmi/HLMCC_AD_IoT

The following sections present solutions that are based on single, ensemble, and hybrid models.

1) SINGLE AND ENSEMBLE MODELS

Single models consist of only one method for the classification process. For securing the IoT network, the paper [9] suggested a solution based on NNs to 4,000 observations collected by simulating ten sensors. The model consists of two experiments with different features to train NNs. The first experiment includes two parameters, which are device ID and sensor value, plus the delay value in the second experiment. In terms of accuracy, both tests performed at over 99%, with less than a 1% negative rate in the second experiment.

Jain and Shah [10] proposed a model to extract the desired information from the Citypulse project for Aarhus, Denmark [11], [12], by comparing three algorithms, namely NNs, binary SVMs, and multiclass SVMs. In both SVMs cases, various types of kernels were applied, such as linear, polynomial, and Radial Basis Function (RBF). The results showed that both cases of SVMs performed better than NNs. More precisely, the RBF kernel was better than the linear kernel in terms of accuracy for binary SVMs. In multi-class SVMs, the polynomial kernel performed better than the linear kernel and RBF kernel in terms of accuracy.

Furthermore, Pachauri and Sharma [13] detected anomalous behaviour in the medical domain by using the MIMIC dataset [14], [15]. They applied the Random Forest (ensemble model), and KNN and J48 decision trees (single models). In terms of the performance, the results showed that the ensemble model was better than single models.

2) HYBRID MODELS

Hybrid models consist of more than two machine learning algorithms for the classification process. Pajouh *et al.* [16] proposed a model for intrusion detection based on anomaly detection in the IoT network. The model consists of two phases, namely Two-layer Dimension Reduction and the Two-tier Classification (TDTC) were applied to the NSL-KDD dataset [17]. To reduce the complexity of the datasets, the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were used as the first step. For classification propose, they have applied two classifiers, Naive Bayes (NB) and KNN. The data is input using the NB method to detect the anomaly behaviours. Since the NB is a weak classifier [18], the normal data is passed to the certainty factor version of KNN (CF-KNN) to detect the class of normal and abnormal behaviours. The results revealed that the model achieved the highest accuracy of 84.86%.

Alghuried [19] combined two algorithms - the IWC and C4.5 decision tree - that was applied to the Intel Lab datasets [20]. The IWC algorithm was used to cluster the data into normal and abnormal. It then used the results to train the C4.5 decision tree. The results showed that the accuracy of the model was 97% and the recall was 98.3%. In addition, the application of density-based spatial clustering of applications with noise (DBSCAN) and SVMs were proposed by

Emadi and Mazinani [21] to the Intel Lab datasets [20]. The selection of the features (temperature, humidity, and voltage) was the first step in the model. In DBSCAN, the Coefficient Correlation (CC) was applied to adjust the values of Epsilon and MinPts. It then grouped the data to normal and anomaly clusters, based on the density. Finally, the SVMs was trained by the labelled data that gained from the clustering step. The accuracy of the model over different experiments was over 94%.

B. UNSUPERVISED LEARNING

Unsupervised learning involves dealing with unlabelled data by finding the similarity among the data points to cluster them. There are many unsupervised learning algorithms, such as K-means, DBSCAN clustering, and PCA. Morrow *et al.* [22] made a comparison between three algorithms: K-means, gaussian kernel density estimation, and DBSCAN. The proposed model used a dataset from a Cray supercomputing facility featured in Usenix's Computer Failure Data Repository [23]. The results showed that DBSCAN was able to effectively detect the anomaly and avoid false positives. Furthermore, based on location density, DBSCAN ranked only truly anomalous data.

Four machine learning algorithms are compared in [24] namely Mahalanobis Distance (MD), Local Outlier Factor (LOF), hierarchical clustering, and one-class SVM (OC-SVM). The model used the data of the network state that was obtained over a period of 14 days in the streets of Barcelona. Different evaluation metrics were used to evaluate the models such as True Positive Rate (TPR), False Positive Rate (FPR), and F-score. Also, Feature Vector 1 (FV1), Feature Vector 2 (FV2) and Feature Vector 3 (FV3) were selected as different feature selection cases. FV1 includes the sensor reading and the timestamp, and FV2 contains FV1, the sequence number of the application packet, and the battery level. FV3 includes FV2, and other features such as the number of received MAC ACK and CTS. In terms of TPR, the results revealed that OC-SVM with FV2 (less than 5% in FPR) was achieved over 75%.

Martí *et al.* [25] combined two methods, which were YASA and OC-SVM to overcome a large amount of unlabelled data collected from sensors. YASA is a method to segment the data into blocks, which is useful to overcome certain limitations in the data such as inconsistency; the segmented data is then passed to OC-SVM. The model applied to the data was obtained from 64 sensors in the operational system over a period of six months in 2012. The McNamara statistical test showed that the proposed model outperformed the other solutions such as OC-SVM, and also statistical Confidence Intervals (CIs). Inoue *et al.* [26] dealt with the complex system by applying Deep Neural Networks (DNNs) and OC-SVM. The model used the data collected from the cyber-physical system called the Secure Water Treatment (SWaT) dataset [27], [28]. In terms of precision and F-score, DNNs were better than OC-SVM, and OC-SVM was slightly better in the recall.

C. SEMI-SUPERVISED LEARNING

Semi-supervised learning deals with a small volume of labelled data and a large volume of unlabelled data to build the model.

Meng et al. [29] proposed a LogClass to detect and classify anomalies from switch device logs. The proposed model was applied to data collected from tens of millions of switch logs over different data centres. The LogClass model used the bag-of-words method to show the patterns of word combinations. It then used the positive samples (anomalies) and unlabelled data to build a PU learning model (learning from Positive and Unlabelled data). In this step, the results showed that the LogClass achieved a 99.515% F-score. Later on, the anomaly logs were used to build a multi-class model to classify anomalies in their appropriate categories. The results indicated that the LogClass achieved 95.32% and 99.74% for Macro-F1 and Micro-F1 scores, respectively.

Zhang et al. [30] proposed a model for Anomaly Detection with partially Observed Anomalies (ADOA), which consisted of two phases. First, observed anomalies were clustered into k clusters, and the unlabelled data were grouped into potential anomalies and reliable normal observations, according to the isolation degree and similarity score. Second, each observation was weighted, and the weighted multi-class model was built to differentiate the anomalies from normal observations. The proposed model used synthetic data, different benchmark datasets, and collected data for malicious URLs. In terms of the Area Under the Curve (AUC) and accuracy, the results showed that the ADOA was better than unsupervised, supervised, and PU learning.

In general, previous studies have almost exclusively focused on building a model based on either the supervised or unsupervised mode. The former requires predefined trained labelled data, while the latter deals with unlabelled data by finding similarities within data to group them. The approach proposed by Alghuried [19] cannot be considered conclusive because IWC clustering is required to determine the optimal number of running k -means to obtain the best cluster results.

In this study, we employ HAP clustering to label the data, which considers all data points as exemplars until a set of good exemplars is selected. This means there is no need to determine the optimal number for running the algorithm, as in IWC [19]. Moreover, we use different validation techniques to measure the quality of the clustering results, such as the silhouette coefficient.

To provide an effective solution for labelling and detecting anomalies, we propose a hybrid learning model called HLMCC, for anomaly detection in IoT using both clustering and classification algorithms. The HLMCC model applies HAP clustering to automate the data labelling, which helps to reduce human intervention and address the issues of unlabelled data. The classification approach is then applied to the obtained labelled data to train the DTs to help with anomaly detection for future unseen data.

III. THE PROPOSED HLMCC

The conceptual phases of the HLMCC model are shown in Figure 1. The HLMCC model consists of two phases:

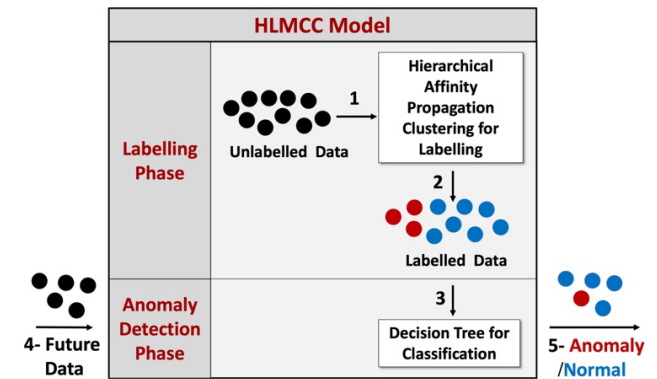


FIGURE 1. The HLMCC model.

1. Automatic Labelling: Employing HAP clustering to classify the data label into normal and abnormal clusters.
2. Detecting Anomalies: The labelled data obtained from the clustering is used to train DTs.

The HLMCC model applies Algorithm 1, which receives unlabelled data as input and classifies the data into normal and abnormal classes as output. The two phases are explained in detail in Sections A and B.

Algorithm 1 HLMCC algorithm

Input: D : Unlabelled dataset

Output: Classified data as anomaly or normal

- 1: Employ HAP algorithm to cluster D into two groups,
- 2: Label D using the clusters, cluster 1: anomaly and cluster 2: normal,
- 3: Split D into two partitions D_{tr} for training and D_{st} for test,
- 4: Train the DTs using D_{tr} ,
- 5: Classify D_{st} as an anomaly or a normal label using DTs.
- 6: end**

A. AUTOMATIC LABELLING: CLUSTERING

In automatic labelling, clustering is used to group the data points that are similar to each other in clusters. However, the characteristics of the application domain play an important role in selecting the appropriate clustering algorithms from a wide range of algorithms. Partitioning clusters such as K -means assumes that each cluster has a spherical shape [31]. For anomaly detection in IoT, an algorithm that produces clusters with no assumptions regarding their shape is needed. Most importantly, which clusters are normal or anomalous must be assumed. It is then assumed that the clusters with a large number of instances are considered normal, whereas those with a small number of instances are considered abnormal [32].

One promising direction to enhance the results of hierarchical clustering methods is to integrate them with other clustering techniques, known as multiple-phase (or multiphase) clustering [31]. HAP applies Affinity Propagation (AP) as a first step before employing agglomerative clustering to join clusters together. AP [33] considers all data points as exemplars and exchanges messages between them until a set of good exemplars is selected. One of the main advantages of AP is that there is no need to determine the number of clusters beforehand; other K-centres are sensitive to an initial number of K.

AP works by accepting two inputs (similarity matrix and preferences). Preferences are used to measure how a data point is suited to be an exemplar (set to the median of the similarity matrix). Two types of messages are exchanged between the data points. First, Responsibility: $r(i,k)$; this is sent from a data point i to a candidate exemplar point k to measure how appropriate k is to be an exemplar for i , considering other potential exemplars. Second, Availability: $a(i,k)$; this is sent from a candidate exemplar point k to a point i to measure how appropriate it would be for i to choose k as an exemplar, taking into account the support from other data points showing that k should be an exemplar. AP produces flat clusters and can be extended to be HAP, which works by finding exemplars at each layer via sending information up and down [34].

B. DETECTING ANOMALIES: CLASSIFICATION

In detecting anomalies, a Decision Tree (DT) is used to classify the data as an anomaly or a normal class label. A DT is a simple type of machine learning algorithm, which works by splitting the data continuously until reaching leaf nodes. The leaf node contains the class label. There are different DT algorithms, such as classification and regression tree (CART), ID3, and C4.5 (J48). In our experiment, CART was used.

IV. EXPERIMENTAL SET-UP

The following sections introduce the datasets and describe data pre-processing.

A. DATASETS

This study is focused on IoT data, so this experiment was applied to two available IoT datasets. All datasets have a ground truth, which helps to later validate the model. The description of each dataset is given below.

1) LWSNDR DATASET

The Labelled Wireless Sensor Network Data Repository (LWSNDR) is a real dataset gained from physical sensors (indoor and outdoor) that comprises four datasets over different scenarios [35], [36]. Two scenarios for each type of sensors (indoor and outdoor) are applied: single-hop and multi-hop cases. The single-hop case consists of a sensor and head station, plus a router in the multi-hop case.

Four input features are used for each dataset, Reading#, Mote ID, Humidity and Temperature. Over six hours, with a five-second interval, both humidity and temperature readings were gathered. LWSNDR is designed for binary classification task which consists of two classes: “0” for the normal data and “1” for the anomaly data. Both temperature and humidity were increased by using hot water to introduce anomalous behaviour.

2) LANDSAT SATELLITE DATASET

The satellite dataset reflects the intensity values for the images collected from satellite observations. The original classification task for this dataset was a multi-class task. However, a new version of this dataset was delivered by Goldstein and Uchida [2] for a binary classification task. They considered two classes: first the normal classes which were “red soil”, “grey soil”, “damp grey soil” and “very damp grey soil”. Second, the anomaly classes referred to non-soil cases, namely “cotton crop” and “soil with vegetation stubble”. The dataset consisted of 36 attributes in the range [0,255] that reflected four spectral bands with nine values for each band.

Table 1 summarizes the datasets used in the experiments. These datasets have a class imbalance, with one category (normal) representing the overwhelming majority of the data points.

TABLE 1. Lwsndr and satellite datasets.

Dataset	No. Instances	No. Attributes	Normal Instances	Anomaly Instances
Single-hop indoor	4,417	4	4,300	117
Single-hop Outdoor	5,041	4	5,009	32
Multi-hop Indoor	4,690	4	4,590	100
Multi-hop Outdoor	4,690	4	4,632	58
Satellite	5,100	36	5,025	75

B. DATA PRE-PROCESSING

Data pre-processing involves transforming raw data into an understandable format. Raw data is often adversely affected by issues such as incompleteness, inconsistency or noise. Different techniques are applied to improve and enhance the quality of the data, such as data cleaning and transformation. This experiment scaled the data into the range [0,1] after ensuring there were no missing values in the data. The normalisation is a helpful task for techniques that are realised on distance measurements like clustering algorithms [31]. The results of the pre-processing phase were used as inputs to the clustering algorithm.

TABLE 2. Adjusted rand index scores for clustering methods. HAP, PAM, HC and IWC stand for hierarchical affinity propagation, partitioning around medoids, agglomerative hierarchical clustering, and inverse weight clustering, respectively.

	LWSNDR dataset				Satellite dataset
	Multi-hop indoor	Single-hop indoor	Single-hop outdoor	Multi-hop outdoor	
HAP	0.1258	0.150110	0.854987	0.126150	0.57618
PAM	0.9018	-0.00929	-0.001130	-0.00551	0.00507
HC	0.074	0.077883	0.913841	0.002247	0.0055
IWC	0.8565	0.010471	0.000268	-0.00562	0.00295

V. PERFORMANCE EVALUATION

R Language was employed in our experiment to program the model and produce the results. After clustering the data, a clustering label was assigned to each instance as a class label. Then, each dataset was divided into two parts: 70% for training and 30% for testing. We ran the experiment five times with fixed seeds using stratified sampling. Moreover, SMOTE was used as a further step to enhance the performance by adding synthetic data to the original data.

We compared three models using different evaluation metrics such as FPR, recall, precision, AUCPR and F-score. The first model is CART, which applies DTs on the originally labelled datasets. The second model is the HLMCC (HAP + CART), which uses the dataset labelled by HAP. The third model (IWC + CART), proposed by Alghuried [19], labels the data using IWC.

Sections A and B present the results of the clustering and discuss the comparison with the state-of-the-art model.

A. EVALUATION OF AUTOMATIC LABELLING

In this experiment, we applied three clustering algorithms in addition to HAP to compare the quality of the results. These algorithms were, namely, Partition Around Medoids (PAM), agglomerative Hierarchical Clustering with complete linkage (HC), and IWC (existing solution) [19]. Both PAM and HC have a similar function to HAP; PAM deals with real points as exemplars, while HC groups the data points as a tree of clusters. The IWC is based on K-means, which finds the best means to assign the data points to clusters. This section explains the results of the clustering algorithms over a variety of validation techniques on the same datasets.

The Adjusted Rand Index (ARI) [37] was used to evaluate the label agreements between actual and predicted labels, ignoring the difference in the label names. For example, if 0 and 1 are actual labels and 1 and 2 are predicted labels, the ARI finds the similarity regardless of the label names. The range of ARI is $[-1, +1]$; $+1$ indicates the labelling is similar, while -1 indicates that the labelling is not similar. However, the ground truth is required to apply ARI.

As shown in Table 2, HAP has the highest value of ARI compared to PAM, HC and IWC labels, but there are some exceptions in PAM, HC, and IWC. PAM has the highest value for multi-hop indoor dataset, while three datasets have negative values, meaning that there is a difference between the true label and the predicted label. This also applies to HC; four of the datasets are less than HAP with around double

the score, this means that the labelling process is not similar. Finally, in IWC, the scores are unsteady between negative to around zero values.


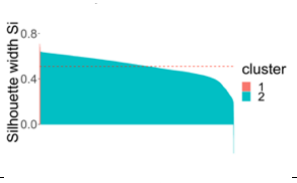
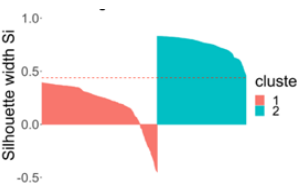
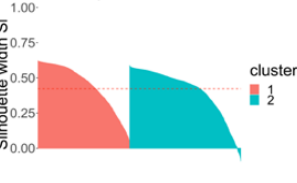
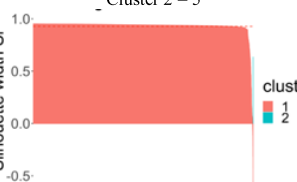
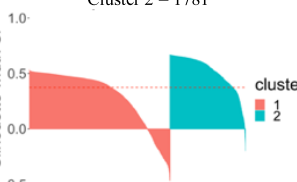
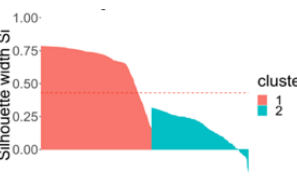
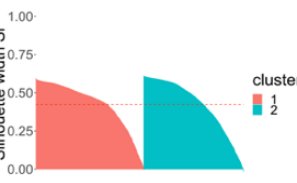
Moreover, silhouette coefficients [38] was also used to validate the quality of different clustering algorithms. Silhouette coefficients measure the compactness or cluster cohesion, meaning how close to each other the data points in a cluster are. Also, it measures the separation among clusters, in terms of how a cluster is separated from other clusters. The range of silhouette is between $[-1, +1]$. $+1$ indicates good clustering and -1 is poor clustering. The silhouette is calculated using distance measurements such as Euclidean distance.

In silhouette coefficient plots, positive values indicate that the clusters are well separated, whilst negative values (under zero) mean inappropriate separation. As shown in Table 3, the silhouette plots clearly indicate that the data points are well separated, with fewer noticeable negative values in HAP. However, the PAM algorithm reveals that there is instability between the datasets, such as the single-hop indoor dataset. For the single-hop indoor dataset, cluster 1 has negative values, meaning that the data points are assigned to the wrong cluster as a result of the wrong separation. However, the HC provides good results for the single-hop indoor dataset, but then leads to a dramatic change in the satellite dataset with negative values (under zero). Finally, the IWC results show that silhouette coefficient plots produce good-quality results among datasets, with only the single-hop indoor dataset showing a noticeable negative value in cluster 2.

As shown in Table 3, the number of data points belonging to the clusters can play an important role in deciding whether the clustering methods are appropriate. Firstly, in HAP, all clusters have a similar percentage of data points 10:4407 and 31:5069. Secondly, in PAM and IWC, some clusters have almost an equal number of data points, such as 2493:1924 and 2310:2790, which is logically not acceptable since this study is concerned with imbalanced datasets. Finally, in HC, the results fluctuated between unbalanced to almost equal data points in clusters, such as 5:4412 to 3319:1781.

Table 4 presents scatterplots for the results of the clustering algorithms; the classes of data points are grouped by colour: red for abnormal and blue for normal cases. In some datasets, the numbers of data points almost equal; therefore, the smaller amount of data points are indicated as anomalous cases. As shown in Table 4, the anomalous points in HAP are grouped close to each other similar to the original

TABLE 3. Silhouette plots and the number of data points in each cluster for Single-hop indoor and Landsat Satellite dataset. HAP, PAM, HC and IWC stand for hierarchical affinity propagation, partitioning around medoids, agglomerative hierarchical clustering, and inverse weight clustering, respectively. Cluster 1 and Cluster 2 mean the number of data points belong to the cluster.

Clustering Methods	LWSNDR dataset	Landsat Satellite dataset
	Single-hop indoor	
HAP	Cluster 1 = 10 Cluster 2 = 4407	Cluster 1 = 31 Cluster 2 = 5069
		
PAM	Cluster 1 = 2493 Cluster 2 = 1924	Cluster 1 = 2310 Cluster 2 = 2790
		
HC	Cluster 1 = 4412 Cluster 2 = 5	Cluster 1 = 3319 Cluster 2 = 1781
		
IWC	Cluster 1 = 2359 Cluster 2 = 2058	Cluster 1 = 2652 Cluster 2 = 2448
		

dataset (ground truth). While, the anomaly points in other clustering algorithms such as PAM, HC, and IWC are not always close to each other as shown in the multi-hop outdoor dataset.

In conclusion, the silhouette plots show that HAP performs quite well across all datasets and is more stable than PAM, HC, and IWC. Moreover, concerning ARI scores, HAP outperforms PAM, HC, and IWC. Furthermore, the scatterplots show that HAP grouped the data points similarly to the ground truth, and not as PAM, HC, and IWC. The results and the plots for the remaining datasets are presented in Tables 5 and 6.

B. EVALUATION OF ANOMALY DETECTING

Due to the imbalanced dataset, detecting correctly positive samples (anomaly) is our primary focus. Therefore, to

evaluate the model we used different evaluation metrics such as FPR, recall, precision, AUCPR and F-score. We used average ranks [39] to compare HLMCC with other models, where the best performance is ranked with a 1. CART uses the original dataset (ground truth), whereas IWC + CART [19] and HLMCC use clustering to label the data.

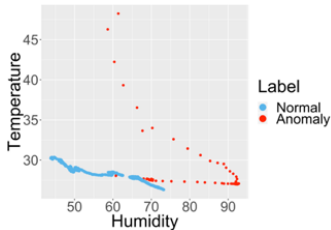
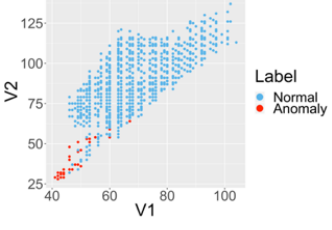
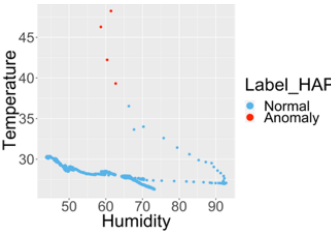
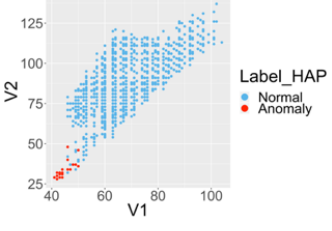
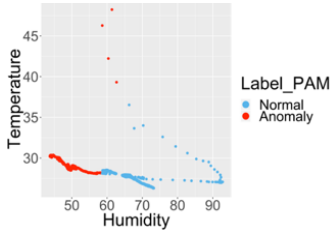
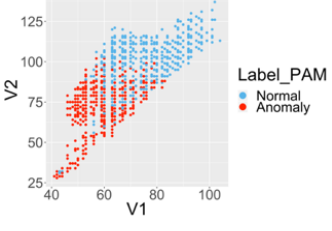
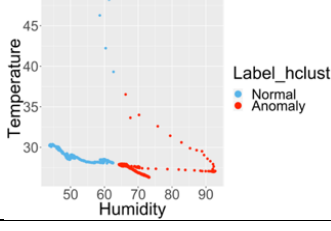
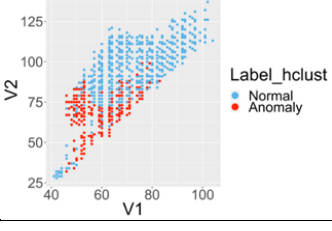
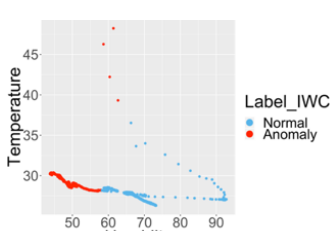
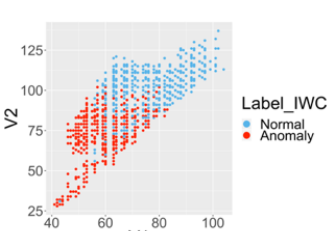
1) FALSE POSITIVE RATE (FPR)

FPR measures how often an anomaly can be predicted while it is not.

$$FPR = FP / (FP + TN), \tag{1}$$

where False Positive (FP) refers to normal cases that are predicted as anomalous and True Negative (TN) to normal cases that are predicted as normal. The lowest scores for FPR

TABLE 4. The clustering results for Multi-hop outdoor and Landsat Satellite datasets over the original space: blue for normal and red for anomaly classes.

	LWSNDR dataset Multi-hop outdoor	Landsat Satellite dataset
Original Datasets		
HAP		
PAM		
HC		
IWC		

indicate a good score, which means that there are no normal cases predicted as anomalous.

Table 7 shows the results in terms of FPR. IWC + CART [19] has the highest FPR with around 0.06 compared to HLMCC and CART, meaning that FPR values have a clear difference among models in satellite datasets. Also, Table 7 reports the average ranks for each model across the datasets. HLMCC performs better than CART and IWC + CART [19] with average ranks 1.8.


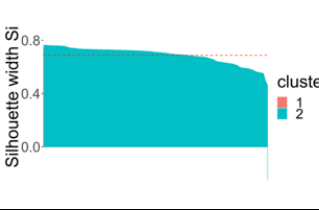
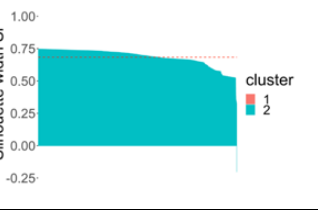
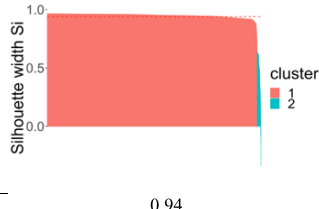
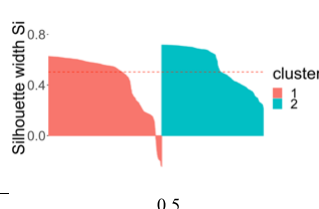
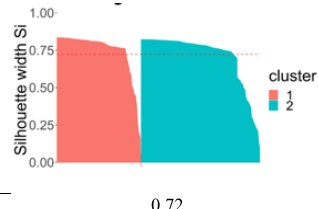
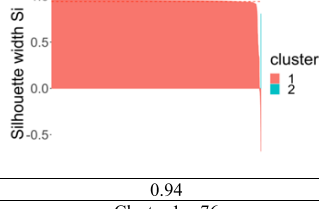
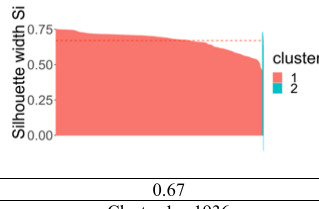
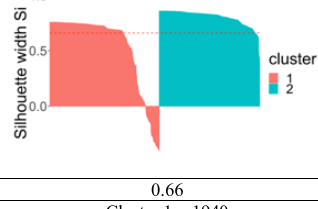
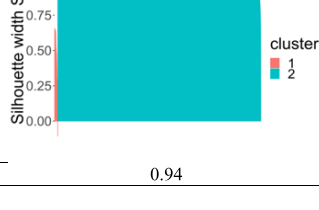
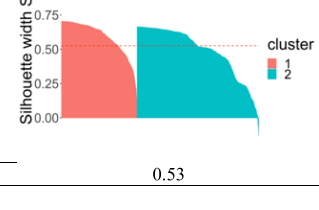
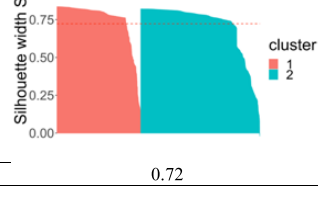
2) RECALL

The recall (sensitivity, TPR) is defined as how many anomalous classes are predicted correctly.

$$Recall = TP / (TP + FN), \tag{2}$$

where True Positive (TP) refers to anomalous cases that are predicted as anomalous and False Negative (FN) to anomalous cases that are predicted as normal. The highest scores for recall indicate a good score, which relates to a low false

TABLE 5. Silhouette plots and the number of data points in each cluster for LWSNDR datasets. HAP, PAM, HC and IWC stand for Hierarchical Affinity Propagation, Partitioning Around Medoids, Agglomerative hierarchical clustering, and inverse weight clustering, respectively. Cluster 1 and Cluster 2 mean the number of data points belong to cluster.

Clustering Methods	LWSNDR datasets		
	Multi-hop indoor	Single-hop outdoor	Multi-hop outdoor
HAP	Cluster 1 = 7 Cluster 2 = 4683	Cluster 1 = 24 Cluster 2 = 5017	Cluster 1 = 4 Cluster 2 = 4686
			
	0.93	0.69	0.68
PAM	Cluster 1 = 4607 Cluster 2 = 83	Cluster 1 = 2655 Cluster 2 = 2386	Cluster 1 = 1949 Cluster 2 = 2741
			
	0.94	0.5	0.72
HC	Cluster 1 = 4686 Cluster 2 = 4	Cluster 1 = 5014 Cluster 2 = 27	Cluster 1 = 2445 Cluster 2 = 2245
			
	0.94	0.67	0.66
IWC	Cluster 1 = 76 Cluster 2 = 4614	Cluster 1 = 1936 Cluster 2 = 3105	Cluster 1 = 1940 Cluster 2 = 2750
			
	0.94	0.53	0.72

negative rate. Table 8 shows the results in terms of recall. For satellite datasets, both CART and IWC + CART [19] have the lowest values compared to HLMCC with 0.7 and 0.95559, respectively. Moreover, HLMCC performs better than CART and IWC + CART [19] with average ranks 1.6.

3) PRECISION

The precision is defined as the proportion of correct anomalous cases among all available anomalous cases. The highest scores for precision indicate a good score, resulting in a low FPR.

$$Precision = TP / (TP + FP), \tag{3}$$

Table 9 shows the results in terms of precision. In satellite datasets, both CART and IWC + CART [19] have the lowest

values compared to HLMCC with 0.82795 and 0.93575, respectively. In average ranks, HLMCC performs better than CART and IWC + CART [19] with average ranks 1.8.

4) AREA UNDER THE PRECISION-RECALL CURVE (AUCPR)

The precision-recall curves present the trade-off between precision and recall. The high area under curve means that both precision and recall are high. Table 10 shows the results in terms of AUCPR. In multi-hop indoor, single-hop outdoor and satellite datasets, HLMCC performs better than other models with 0.9942, 1 and 0.9922, respectively. In average ranks, HLMCC performs better than CART and IWC + CART [19] with average ranks 1.8.

TABLE 6. The clustering results for LWSNDR datasets over the original space: blue for normal and red for anomaly classes.

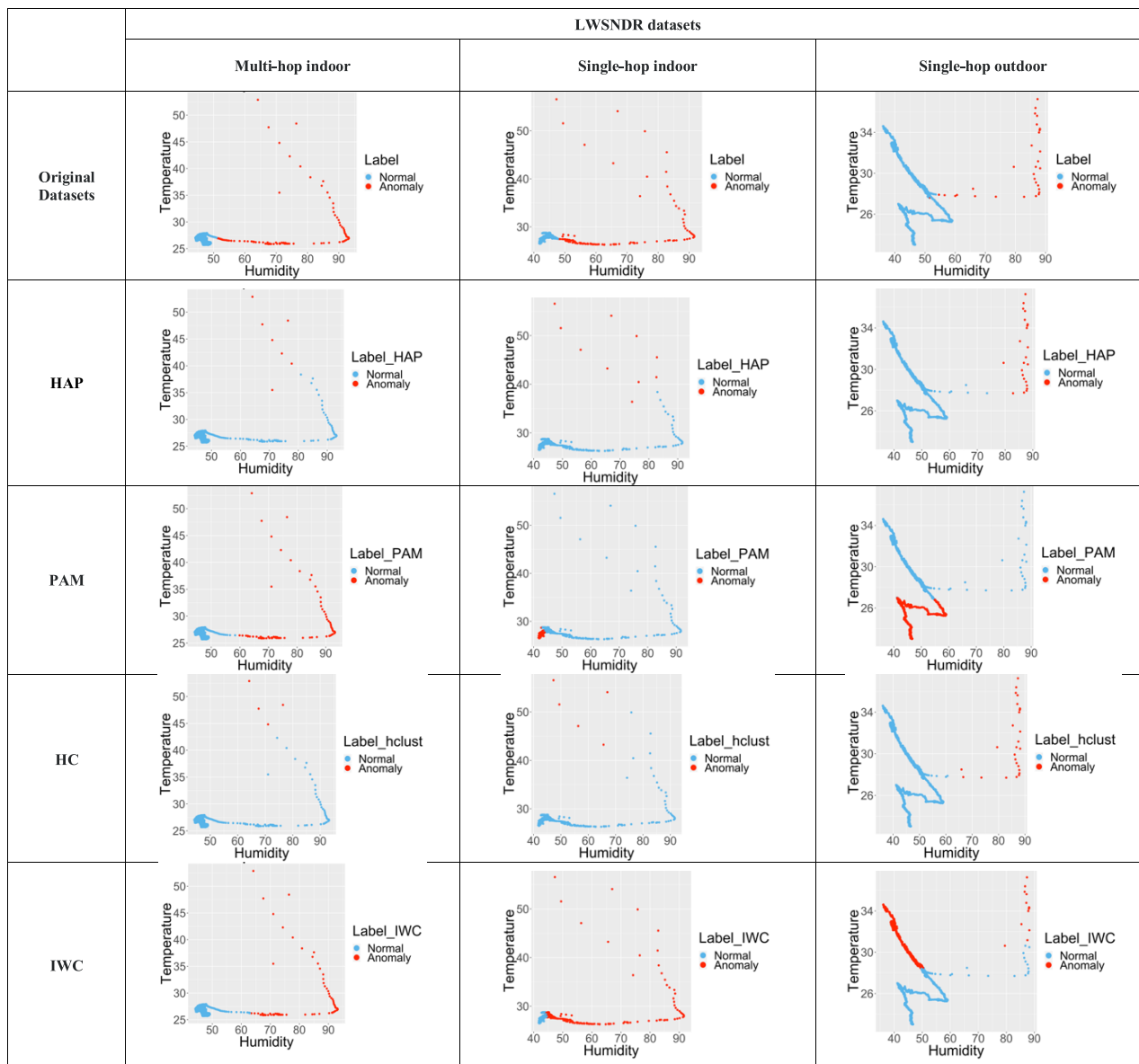


TABLE 7. Comparison of the proposed model with the state-of-the-art based on FPR. CART uses the original datasets. FPR and SD stand for the False Positive Rate value and Standard deviation, respectively. The average rank is based on the FPR value.

Dataset	Models								
	CART			IWC+CART [19]			HLMCC		
	FPR	SD	Rank	FPR	SD	Rank	FPR	SD	Rank
Multi-hop indoor	0.000145	± 0.0002904866	2	0	0	1	0.0006	± 0.0003220612	3
Single-hop indoor	0.000155	± 0.0003100775	1	0.00736	± 0.00105846	3	0.0015	± 0.0009832971	2
Single-hop outdoor	0	0	1.5	0.00344	± 0.001424973	3	0	0	1.5
Multi-hop outdoor	0.00144	± 0.001821326	3	0	0	1.5	0	0	1.5
Landsat Satellite	0.003052	± 0.003267029	2	0.06063	± 0.006627391	3	0.0007	± 0.0004723342	1
Average Ranks			1.9			2.3			1.8

5) F-SCORE

F-score is the harmonic mean of both precision and recall.

$$F - score = (2 * Recall * Precision)/(Recall + Precision), \tag{4}$$

Table 11 shows the results in terms of F-score. It reports the average ranks for each model across the datasets. HLMCC and IWC + CART [19] perform with the same average ranks 1.9.

In conclusion, HLMCC performs better than CART and IWC + CART [19] over a wide range of evaluation metrics:

TABLE 8. Comparison of the proposed model with the state-of-the-art based on the recall. CART uses the original datasets. Recall and SD stand for the recall value and Standard deviation, respectively. The average rank is based on the recall value.

Dataset	Models								
	CART			IWC+CART [19]			HLMCC		
	Recall	SD	Rank	Recall	SD	Rank	Recall	SD	Rank
Multi-hop indoor	1	0	1.5	0.96364	± 0.03401507	3	1	0	1.5
Single-hop indoor	0.994	± 0.01142857	3	0.996	± 0.004741893	2	1	0	1
Single-hop outdoor	0.866667	± 0.1295767	3	1	0	1.5	1	0	1.5
Multi-hop outdoor	0.670588	± 0.1421535	2	0.99931	± 0.0008417491	1	0.5757	± 0.03050594	3
Landsat Satellite	0.7	± 0.1336085	3	0.95559	± 0.003510654	2	0.9986	± 0.002702703	1
Average Ranks			2.5			1.9			1.6

TABLE 9. Comparison of the proposed model with the state-of-the-art based on the precision. CART uses the original datasets. Precision and SD stand for the precision value and Standard deviation, respectively. The average rank is based on the Precision value.

Dataset	Models								
	CART			IWC+CART [19]			HLMCC		
	Precision	SD	Rank	Precision	SD	Rank	Precision	SD	Rank
Multi-hop indoor	0.993548	± 0.01290323	3	1	0	1	0.9942	± 0.002877698	2
Single-hop indoor	0.994444	± 0.01111111	1	0.99161	± 0.001194122	2	0.9866	± 0.00862712	3
Single-hop outdoor	1	0	1.5	0.99452	± 0.002264232	3	1	0	1.5
Multi-hop outdoor	0.894737	± 0.1331485	3	1	0	1.5	1	0	1.5
Landsat Satellite	0.827953	± 0.1611546	3	0.93575	± 0.006457321	2	0.9933	± 0.004216505	1
Average Ranks			2.3			1.9			1.8

TABLE 10. Comparison of the proposed model with the state-of-the-art based on AUCPR. CART uses the original datasets. AUCPR and SD stand for the area under the precision-recall curve value and Standard deviation, respectively. The average rank is based on the AUCPR value.

Dataset	Models								
	CART			IWC+CART [19]			HLMCC		
	AUCPR	SD	Rank	AUCPR	SD	Rank	AUCPR	SD	Rank
Multi-hop indoor	0.993548	± 0.0129	2	0.96595	± 0.0318734	3	0.9942	± 0.002877698	1
Single-hop indoor	0.989	± 0.01314	1.5	0.989	± 0.0028482	1.5	0.9866	± 0.00862712	3
Single-hop outdoor	0.87007	± 0.12641	3	0.99452	± 0.0022642	2	1	0	1
Multi-hop outdoor	0.767881	± 0.10807	2	0.99974	± 0.0003177	1	0.6651	± 0.02556201	3
Landsat Satellite	0.633401	± 0.10179	3	0.92326	± 0.0059592	2	0.9922	± 0.00467766	1
Average Ranks			2.3			1.9			1.8

TABLE 11. Comparison of the proposed model with the state-of-the-art based on F-score. CART uses the original datasets. F-score and SD stand for the F-score value and Standard deviation, respectively. The average rank is based on the F-score value.

Dataset	Models								
	CART			IWC+CART [19]			HLMCC		
	F-score	SD	Rank	F-score	SD	Rank	F-score	SD	Rank
Multi-hop indoor	0.997	± 0.006557377	1.5	0.98117	± 0.0177602	3	0.997	± 0.001444043	1.5
Single-hop indoor	0.994	± 0.007001161	1.5	0.994	± 0.002451152	1.5	0.993	± 0.004364863	3
Single-hop outdoor	0.923235	± 0.07696896	3	0.99725	± 0.001137864	2	1	0	1
Multi-hop outdoor	0.744402	± 0.0648392	2	0.99966	± 0.0004212364	1	0.7303	± 0.02427372	3
Landsat Satellite	0.732841	± 0.0617579	3	0.94555	± 0.00309722	2	0.996	± 0.002519708	1
Average Ranks			2.2			1.9			1.9

FPR, recall, precision and AUCPR. While both HLMCC and IWC + CART [19] perform similarly for F-score. Most importantly, CART and IWC + CART [19] have fluctuations in average ranks among evaluation metrics; for example, CART obtains the second ranks in FPR whereas it obtains the third ranks in recall and precision.

VI. CONCLUSION

The data in IoT is inconsistent for varying reasons, such as attack issues, or a breakdown in devices. Anomaly detection is the technique of finding abnormal patterns in the data, which is found in different application domains, such as fault or fraud detection. One of the popular anomaly

detection techniques involves using machine learning algorithms.

This paper has proposed the hybrid learning model HLMCC, which uses clustering and classification approaches for anomaly detection in the IoT. The HLMCC consists of two functional phases, automatic labelling and detecting anomalies. First, the HLMCC employed HAP clustering to automate labelling of data, which can be helpful to minimize human involvement and address the issue of unlabelled data. Second, the obtained data was trained by DTs to predict and detect the class labels for unseen future data.

The results found that the HLMCC was able to overcome the absence of labelled data by automating the labelling process. Moreover, the HLMCC outperformed the DTs on originally labelled datasets and the state-of-the-art model in different evaluation metrics such as FPR, recall, precision and AUCPR.

In future work, we aim to improve the model and address certain limitations such as applying and testing different classifiers, which may help to improve the classification process. Additionally, selecting features to enhance the detection process and reduce the dimensions without the loss of critical information, especially on datasets like the Landsat satellite dataset, would be recommended.

ACKNOWLEDGMENT

The authors would like to thank DSR technical and financial support.

REFERENCES

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.
- [2] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, 2016, Art. no. e0152173.
- [3] N. Alghanmi, R. Alotaibi, and S. M. Buhari, "TCMD: A two-tier classification model for anomaly-based detection in IoT," in *Proc. 6th Swiss Conf. Data Sci. (SDS)*, Jun. 2019, pp. 130–135.
- [4] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016.
- [5] "Cisco global cloud index: Forecast and methodology, 2016–2021," Cisco Syst., Inc., San Jose, CA, USA, White Paper, 2018.
- [6] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.
- [7] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 11994–12000, 2009.
- [8] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Upper Saddle River, NJ, USA: New Jersey: Prentice-Hall, 1997.
- [9] J. Cañedo and A. Skjellum, "Using machine learning to secure IoT systems," in *Proc. 14th Annu. Conf. Privacy, Secur. Trust (PST)*, Dec. 2016, pp. 219–222.
- [10] R. Jain and H. Shah, "An anomaly detection in smart cities modeled as wireless sensor network" in *Proc. Int. Conf. Signal Inf. Process. (IconSIP)*, Oct. 2016, pp. 1–5.
- [11] *Pollution Data, Citypulse Project*, Univ. Surrey, Guildford, U.K., 2014.
- [12] M. I. Ali, F. Gao, and A. Mileo, "CityBench: A configurable benchmark to evaluate RSP engines using smart city Datasets," in *The Semantic Web - ISWC*. Cham, Switzerland: Springer, 2015, pp. 374–389.
- [13] G. Pachauri and S. Sharma, "Anomaly detection in medical wireless sensor networks using machine learning algorithms," *Procedia Comput. Sci.*, vol. 70, pp. 325–333, Dec. 2015.
- [14] *PhysioNet*. Accessed: Oct. 12, 2019. [Online]. Available: <https://archive.physionet.org/cgi-bin/atm/ATM>
- [15] L. G. Ary, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, H. E. Stanley, "Physiobank, physiokit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [16] H. H. Pajouh, R. Javidan, R. Khayami, D. Ali, and K.-K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 2, pp. 314–323, Apr./Jun. 2019.
- [17] M. Tavallae, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.
- [18] H. H. Pajouh, G. Dastghaibfard, and S. Hashemi, "Two-tier network anomaly detection model: A machine learning approach," *J. Intell. Inf. Syst.*, vol. 48, no. 1, pp. 61–74, 2017.
- [19] A. Alghuried, "A model for anomalies detection in Internet of Things (IoT) using inverse weight clustering and decision tree," Ph.D. dissertation, Dublin Inst. Technol., Dublin, Ireland, 2017.
- [20] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, and R. Thibaux, "Intel lab data," Reza Int. Univ., Mashhad, Iran, Tech. Rep., 2004.
- [21] H. S. Emadi and S. M. Mazinani, "A novel anomaly detection algorithm using DBSCAN and SVM in wireless sensor networks," *Wireless Pers. Commun.*, vol. 98, no. 2, pp. 2025–2035, 2018.
- [22] A. Morrow, E. Baseman, and S. Blanchard, "Ranking anomalous high performance computing sensor data using unsupervised clustering," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2016, pp. 629–632.
- [23] B. Schroeder and G. A. Gibson, "The computer failure data repository (CFDR)," in *Proc. Workshop Rel. Anal. Syst. Failure Data (RAF)*, Cambridge, U.K., 2007.
- [24] V. Garcia-font, C. Garrigues, and H. Rifà-Pous, "A comparative study of anomaly detection techniques for smart city wireless sensor networks," *Sensors*, vol. 16, no. 6, p. 868, Jun. 2016.
- [25] L. Martí, N. Sanchez-Pi, J. M. Molina, and A. C. B. Garcia, "Anomaly detection based on sensor data in petroleum industry applications," *Sensors*, vol. 15, no. 2, pp. 2774–2797, Jan. 2015.
- [26] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, and J. Sun, "Anomaly detection for a water treatment system using unsupervised machine learning," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 1058–1065.
- [27] J. Goh, S. Adepur, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in *Critical Information Infrastructures Security*. Cham, Switzerland: Springer, 2017, pp. 88–99.
- [28] (2017). *Secure Water Treatment (SWaT)*. [Online]. Available: <https://itrust.sutd.edu.sg/dataset/>
- [29] W. Meng, Y. Liu, S. Zhang, D. Pei, H. Dong, L. Song, and X. Luo, "Device-agnostic log anomaly classification with partial labels," in *Proc. IEEE/ACM 26th Int. Symp. Qual. Service (IWQoS)*, Jun. 2018, pp. 1–6.
- [30] Y.-L. Zhang, L. Li, J. Zhou, X. Li, and Z.-H. Zhou, "Anomaly detection with partially observed anomalies," presented at the Companion Proc. Web Conf., Lyon, France, 2018.
- [31] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
- [32] L. Portnoy, *Intrusion Detection With Unlabeled Data Using Clustering*. Portland, OR, USA: Columbia Univ., 2000.
- [33] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.
- [34] I. E. Givoni, C. Chung, and B. J. Frey, "Hierarchical affinity propagation," presented at the Proc. 27th Conf. Uncertainty Artif. Intell., Barcelona, Spain, 2011.
- [35] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks," in *Proc. 6th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process.*, Dec. 2010, pp. 269–274.
- [36] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami. (2010). *Labelled Wireless Sensor Network Data Repository (LWSNDR)*. [Online]. Available: http://issnip.unimelb.edu.au/research_program/downloads

- [37] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [38] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
- [39] P. B. Brazdil and C. Soares, *A Comparison of Ranking Methods for Classification Algorithm Selection*. Berlin, Germany: Springer, 2000, pp. 63–75.

NUSAYBAH ALGHANMI received the B.S. and M.S. degrees in information technology from King Abdulaziz University, Jeddah, Saudi Arabia. She is currently a Teaching Assistant with the Faculty of Computing and Information Technology, University of Jeddah, Khulais, Saudi Arabia. Her research interests include wireless sensor networks, machine learning, and data mining.

REEM ALOTAIBI received the Ph.D. degree in computer science from the University of Bristol, Bristol, U.K., in 2017. She is currently an Assistant Professor with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include machine learning, data mining, and multilabel classification.



SEYED M BUHARI is currently an Associate Professor with the Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University. His areas of interests are data mining, grid computing, and wireless sensor networks.

...