A Novel Machine Learning-Based Approach for Outlier Detection in Smart Healthcare Sensor Clouds

Rajendra Kumar Dwivedi, Madan Mohan Malaviya University of Technology, Gorakhpur, India
https://orcid.org/0000-0001-6682-1942

Rakesh Kumar, Madan Mohan Malaviya University of Technology, Gorakhpur, India

Rajkumar Buyya, The University of Melbourne, Australia

ABSTRACT

A smart healthcare sensor cloud is an amalgamation of the body sensor networks and the cloud that facilitates the early diagnosis of diseases and the real-time monitoring of patients. Sensitive data of the patients, which are stored in the cloud, must be free from outliers that may be caused by malfunctioned hardware or the intruders. This paper presents a machine learning-based scheme for outlier detection in smart healthcare sensor clouds. The proposed scheme is a hybrid of clustering and classification techniques in which a two-level framework is devised to identify the outliers precisely. At the first level, a density-based scheme is used for clustering while at the second level, a Gaussian distribution-based approach is used for classification. This scheme is implemented in Python and compared with a clustering-based approach (mean shift) and a classification-based approach (support vector machine) on two different standard datasets. The proposed scheme is evaluated on various performance metrics. Results demonstrate the superiority of the proposed scheme over the existing ones.

KEYWORDS

Body Sensor Network, Classification, Cloud Computing, Clustering, Healthcare, Internet of Things, Machine Learning, Outlier Detection, Sensor Cloud

1. INTRODUCTION

Nowadays, computing is not just limited to a single system; in fact, every device is being redesigned to be connected via the Internet to provide facilities and on-demand computing. This technology is known as the Internet of Things (IoT). In this emerging technology, various types of computing devices and techniques have been integrated to facilitate the users in many ways (Gubbi et al., 2013). The sensor cloud is another such integration. The sensor cloud is an integration of sensor networks with the cloud. The sensor cloud facilitates its end-users to get data from various sensor networks, just in one click (Dwivedi et al., 2019). It is possible due to the process of virtualization performed in the cloud.

IoT applications are based on sensor networks. There are two types of sensor networks namely general-purpose and special-purpose. Body sensor networks are the special-purpose networks used in healthcare systems. Cloud computing is a technology that is based on pay-per-use policy. Users have to pay only for what they use. Clouds can be categorized into three types, viz., public, private,

DOI: 10.4018/IJHISI.20211001.oa26

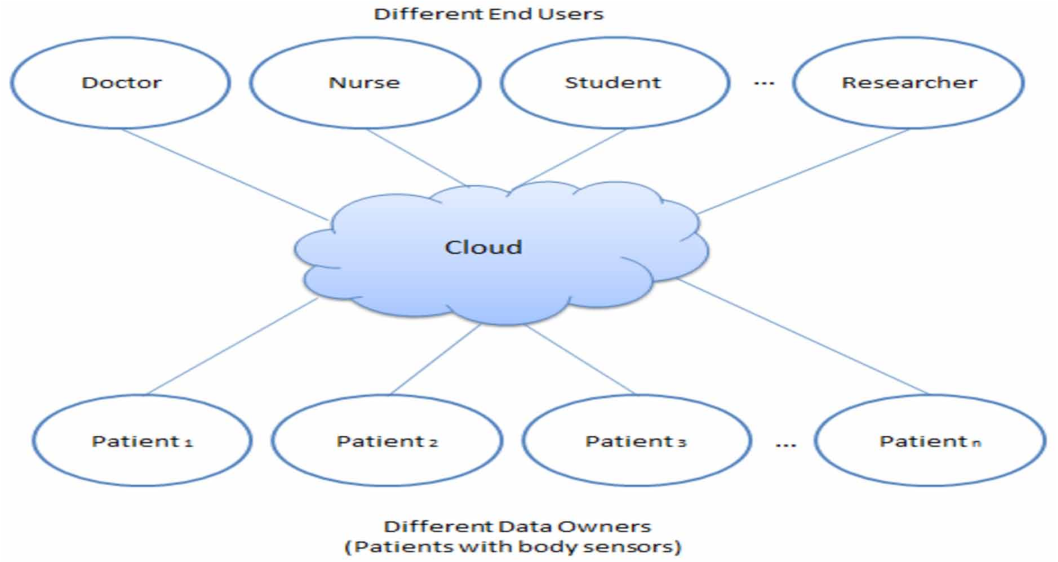
and hybrid. Cloud provides mainly three types of services namely software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS). The sensor cloud architecture enables for one more type of service called sensor-as a-service (SeaaS).

When sensor networks are integrated with the cloud, they facilitate the sensor owner, cloud service provider, and end-users in many ways. This integration also improves the effectiveness of sensor applications. To overcome the storage limitation of sensor networks, the cloud can provide an effective and economical space for storing huge amounts of data. Data analytics tools can be applied to get insights on this huge data.

Sensor clouds find their utility in several information systems such as smart healthcare, forest fire detection, street monitoring, battlefield monitoring, military applications, disaster management (Bessis et al., 2011; Thilakanathan et al., 2014; Lounis et al., 2016). A huge amount of data is generated in such applications. These data are very crucial. Therefore, security becomes a prime concern with these applications (Ahmed et al., 2016; Petrakis et al., 2018).

A smart healthcare system is a system where doctors can monitor the health status of their patients and prescribe the suggestions remotely. The architecture of a smart healthcare sensor cloud is shown in Figure 1. Various real-time facilities can be provided to the patients using this smart system. Emergency cases can be handled quickly and required services can be provided instantly using this system. This system allows remote treatment as well as reduces the treatment cost. In this system, patients are equipped with various body sensors. Several types of patients' data such as SpO2, heart rate, body temperature, and blood pressure can be sensed and transferred to the cloud for further processing. Nurses, junior doctors, medical students, researchers, and other authorized users can access the patients' health data as per their requirements. Patients' health data are very crucial and any false information or outlier can cause various problems. Hence, the sensed information must be precise and accurate. Outliers are the data that are quite different from the other members of a set or group. It might be caused due to any malicious activity by an intruder or some failures in the hardware. This paper focuses on outlier detection in sensor cloud data of the smart healthcare system.

Figure 1. The architecture of a smart healthcare sensor cloud



1.1 Motivation

Healthcare data of any patient are very critical and any incorrect information may mislead the complete diagnosis of the patients. So, any anomaly or outlier must not be present in these data. If the outliers exist in these health records, then they must be detected precisely in time to prevent any serious problem that could be caused due to them. Machine learning techniques can be used to identify them precisely (Xu et al., 2012; Xu et al., 2013; Yenke et al., 2017). There are many classification-based supervised and clustering-based unsupervised machine learning schemes to be used to solve this issue. The existing classification-based technique viz., support vector machine (SVM), and clustering-based algorithm viz., mean-shift detect the outliers well but their accuracy and efficiency are not up to the mark. The advantageous features of clustering and classification techniques can be combined to develop a new method. This motivated us to devise a novel clustering and classification based hybrid machine learning scheme for outlier detection in the healthcare system which enhances the accuracy and efficiency of the system.

1.2 Contributions

A machine learning-based approach for outlier detection in healthcare data is proposed in this paper. We devised a clustering and classification based hybrid approach (CCH). For this, we developed one clustering scheme called a density-based scheme (DBS) and one classification scheme named as Gaussian distribution based approach (GDA). First, we make different clusters using our density-based clustering scheme, then we classify the outliers with the help of our Gaussian distribution based classification approach. DBS helps to create two distinct clusters, one for normal data and other for suspicious data. Thus, normal or genuine data of a dataset falls under one umbrella and now we have to find outliers only from the suspicious data. GDA uniformly distributes the suspicious data and then classifies them into two classes, viz., normal data and outliers. Thus, outliers are identified by a two-level refinement which improves the accuracy and efficiency of the system. The key contributions of the paper are as follows:

- Design of a novel clustering and classification based hybrid scheme for outlier detection in healthcare data
- Two-level of refinement for enhancing the accuracy; firstly, using clustering and secondly using classification techniques
- Performance evaluation on various performance metrics using two different healthcare standard datasets with different sizes, features, and number of outliers
- Analytical validation to justify the implementation results

1.3 Organization of the Paper

The rest of the paper is organized as follows. A brief description of the related work is described in Section 2. The proposed scheme is discussed in Section 3. Section 4 presents a performance comparison of the proposed work with the existing schemes. Section 5 shows the result validation and Section 6 discusses the salient features of the proposed scheme. Finally, Section 7 concludes the work with some future directions.

2. RELATED WORK

Advancement in the body sensors has opened a new direction to the IoT based smart healthcare systems. IT tools have been used in these systems for a variety of applications such as fetching the patients' records, detecting severe diseases, and identifying the outliers in health data. Doctors use the healthcare data for remote treatment of the patients and researchers use them to solve some issues within the system. Various research problems of healthcare systems have been solved using machine learning techniques. Jain et al. (2020) have used a two-phase hybrid classification framework of machine learning to classify chronic diseases. They claimed that the highest classification accuracy of their scheme is 98.5%. Singh et al. (2020) presented a performance analysis of various machine learning techniques on different metrics to detect cervical cancer. They compared the accuracy and computational time of various learning methods on different training sizes.

These days, IoT and other IT enabled techniques have also been used to handle the various issues of the healthcare systems. Jayaraman et al. (2019) have illustrated how IoT could solve various issues of the healthcare supply chain process, viz., product recalls, product shortages, and expiration monitoring. They described the IoT and blockchain technologies to implement and execute these processes securely and efficiently. Ashtari et al. (2019) have illustrated how nurses' perception of performance can be predicted. They applied various linear regression techniques to validate their secure scheme. Wang et al. (2019) proposed an apriori algorithm-based model to predict lifestyle diseases and provide better service to the patients. Users can input some of their basic details to this system for their check-up. This system predicts the disease and provides advice accordingly.

Health data of the patients are very critical, but they need to be accessed for various reasons by different persons. Hence, these data must be accurate and secure. Gope et al. (2016) proposed an IoT based healthcare system for efficient and secure patient monitoring. Sittig et al. (2018) advocated that there should be a shared responsibility of healthcare organizations, IT departments, and vendors towards the safety and security of the smart healthcare systems. They further suggested that if each stakeholder will share the responsibility, then the healthcare system would be safer and progressive.

When we talk about the security of healthcare data then we find one more term called an outlier. Data that deviate significantly from the other data of any group is called outlier or anomaly (Bosman et al., 2017). Malicious activities or abnormal behavior of the sensors can cause such anomalies (Ghorbel et al., 2015; Gil et al., 2016; Dwivedi et al., 2018). Healthcare data contain the patients' health records and any alteration to these data may create serious problems. Therefore, healthcare data must be free from the outliers. Machine learning techniques can help to identify them precisely (Aleksandrova et al., 2019; Ensari et al., 2019; Rath et al., 2019; Sharma et al., 2019). Several authors have devised machine learning-based techniques for outlier detection in healthcare systems.

Hauskrecht et al. (2013) developed a data-driven scheme for outlier detection in the patient monitoring system. It generates an alert if there is any unusual deviation from the past data of the patients. They worked on data of the cardiac patients. This scheme gives the true alert rate in the range of 25% to 66%. Another scheme was developed by Haque et al. (2015) to identify the outliers in medical sensor data. This scheme also detects false alarms very well. Authors claimed that their approach provides a low false-positive rate and a high detection rate. One more machine learning-based model was devised by Hauskrecht et al. (2016) to detect the outliers in the healthcare system with true alert rates in the range of 44% to 71%. However, the performance and accuracy of these schemes can be further enhanced by devising some other novel methods.

There are several clustering and classification based unsupervised and supervised machine learning techniques that can be used for precise outlier detection in healthcare data. In this paper, we focussed on the clustering-based algorithm 'mean-shift' and the classification-based algorithm 'support vector machine'. The mean-shift is an unsupervised machine learning scheme. It is a density-based clustering algorithm. It does not specify the number of clusters in the beginning. Clusters may be of any shape such as elliptical or spherical. This algorithm is robust to outliers, but its computational complexity is high. The support vector machine is a supervised machine learning scheme. It is a classification-based machine learning technique that has a high detection rate with low computational and communication overheads (Snoussi, 2015; Shahid et al., 2012).

Ozertem et al. (2008) described that the mean-shift algorithm gives very good results in various learning situations. They proposed a spectral clustering-based mean-shift scheme. They advocated that their scheme was a well-founded method to enable a probabilistic interpretation of affinity-based clustering. Mattos et al. (2016) presented a study of the mean-shift algorithm to detect the outliers in asymmetric normal regression models. They applied the outlier detection model in breast cancer patients' health records and observed good results. Ahmed et al. (2016) described various types of clustering approaches that can be used for outlier detection. They explained that the mean-shift is a very good clustering method to detect the outliers. Han et al. (2017) have devised an outlier detection scheme using the clustering technique. This scheme uses adaptive mean-shift clustering to identify the outliers. They claimed that their scheme detects the outliers well. However, the efficiency and accuracy of these approaches can be further improved.

Kaplantzis et al. (2014) had introduced a two-class SVM classifier which separates two distinct classes of data. Using various performance measures, they advocated that SVM is a very good classification method to detect the outliers. Zhang et al. (2016) proposed a distributed online outlier detection scheme based on Ellipsoidal SVM. They had shown the effectiveness of their scheme using various performance metrics. Ji et al. (2017) devised the one-class SVM method for outlier detection. This method uses adaptive weights to detect outliers. Deng et al. (2018) discussed some classification methods to detect the outliers in high-dimensional big-sensor-data. They demonstrated how SVM is a very good classifier for outlier detection. Bansal et al. (2018) proposed a novel technique for outlier detection using SVM. This technique uses a classification method to detect anomalies. The feature selection method of this approach helps to identify the outliers well. However, the efficiency and accuracy of these schemes can be further enhanced.

Table 1 presents a comparative study of the machine-learning-based approaches used for outlier detection in healthcare systems. It shows that the accuracy and efficiency of the existing approaches are not up to the mark and needs some enhancements. This study motivated us to design a new outlier detection scheme for healthcare systems that should identify the actual outliers with optimum accuracy and efficiency. To make a secure healthcare system against the outliers, we planned to use a hybrid approach of clustering and classification so that we can grab the benefits of both the clustering and classification techniques of machine learning.

3. PROPOSED WORK

To overcome the issue of identifying the outliers in the healthcare systems with greater accuracy and efficiency, we proposed a clustering and classification based hybrid approach called CCH. Our proposed scheme uses a two-level framework as shown in Figure 2 for achieving better accuracy than the existing schemes. In the first level, by using clustering, we get two clusters: one of the normal data and another of suspicious data which may contain some outliers. In the second level, by using classification on suspicious data, we get the actual outliers. For this purpose, we developed an unsupervised method of machine learning called DBS using clustering and a supervised method of machine learning called GDA using classification. The proposed scheme consists of three main algorithms discussed in the following subsections.

3.1 Clustering and Classification based Hybrid Approach (CCH)

Algorithm 1 describes the main procedure of the proposed scheme in which two different algorithms, viz., Algorithm 2, and Algorithm 3 are nested. First of all, we carried out the preprocessing of the input dataset in which we clean the data by removing the noise and the redundant data. After this, feature selection is done. Then, we divide the dataset into two parts called the training set and the test set. We used 70% of the data for training and 30% of the data for testing as we got the best results at 70:30 proportions. We trained the model with the training set using DBS and GDA. After training the model, we applied testing on the test set. Figure 3 presents the working of the proposed scheme of outlier detection.

Table 1. A comparison of learning-based outlier detection schemes in healthcare IoT

Author (Year)	Approach used	Machine learning technique used	Features	Remarks
Haque et al. (2015)	Regression	Voting based learning scheme	It provides a low false-positive rate and a high detection rate.	Precision and efficiency can be enhanced.
Ozertem et al. (2008)	Clustering	Spectral clustering-based mean-shift scheme	It is a well-founded method to enable the probabilistic interpretation of affinity-based clustering.	Computational overheads can be reduced, and accuracy can be improved.
Mattos et al. (2016)	Clustering	Mean-shift algorithm	It detects the outliers well in asymmetric normal regression models.	Precision can be enhanced, and computational complexities can be reduced.
Han et al. (2017)	Clustering	Adaptive mean-shift clustering	It detects the outliers by creating different clusters of the data.	Accuracy and efficiency need to be optimized.
Hauskrecht et al. (2013)	Classification	Support Vector Machine (SVM) with a linear kernel	It gives the true alert rate in the range of 25% to 66%.	Performance can be enhanced.
Kaplantzis et al. (2014)	Classification	Two-class SVM classifier	It separates two distinct classes of data.	Precision can be enhanced.
Hauskrecht et al. (2016)	Classification	Support Vector Machine (SVM) with a linear kernel	It gives true alert rates in the range of 44% to 71%.	Performance can be enhanced.
Zhang et al. (2016)	Classification	Ellipsoidal SVM	It detects the distributed online outliers.	Accuracy can be improved.
Ji et al. (2017)	Classification	One-class SVM	It uses adaptive weights to detect outliers.	Precision and throughput need improvements.
Bansal et al. (2018)	Classification	Support vector machine	The feature selection method of this approach helps to identify the outliers well.	Efficiency needs to be optimized.

3.2 Density-based Scheme (DBS)

Algorithm 2 presents a density-based scheme for clustering. Here, we select an arbitrary point from the dataset and check whether this point lies within the low-density region or not. If it does not lie there, then, we place this data point in cluster A. It is normal data. If the data point lies within the low-density region, then, we place it in cluster B. This data point might be an outlier. Now, we will apply a classification algorithm on this suspicious cluster B to confirm whether it is an actual outlier or not. The working of DBS is shown in Figure 4.

Figure 2. Overview of the proposed scheme CCH

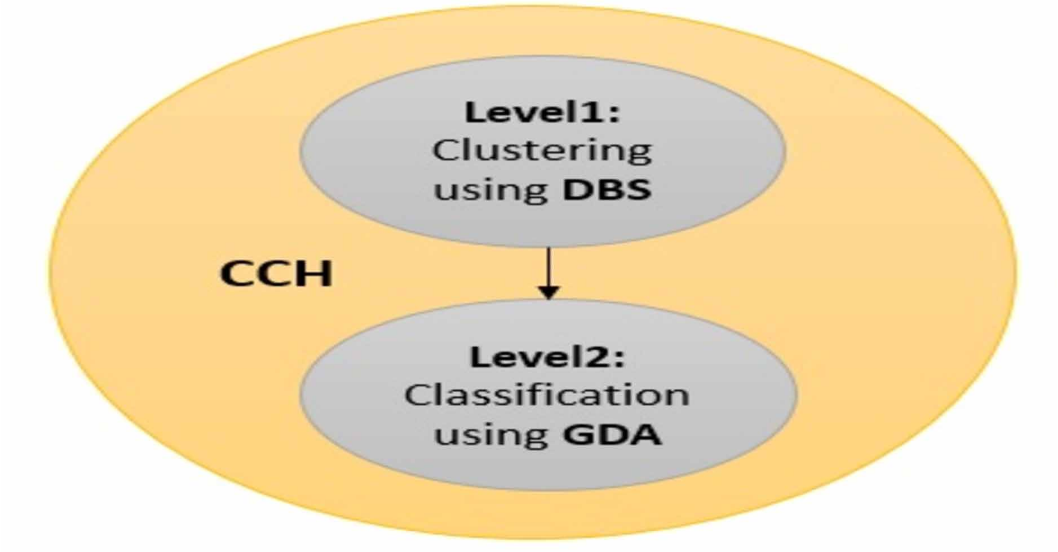
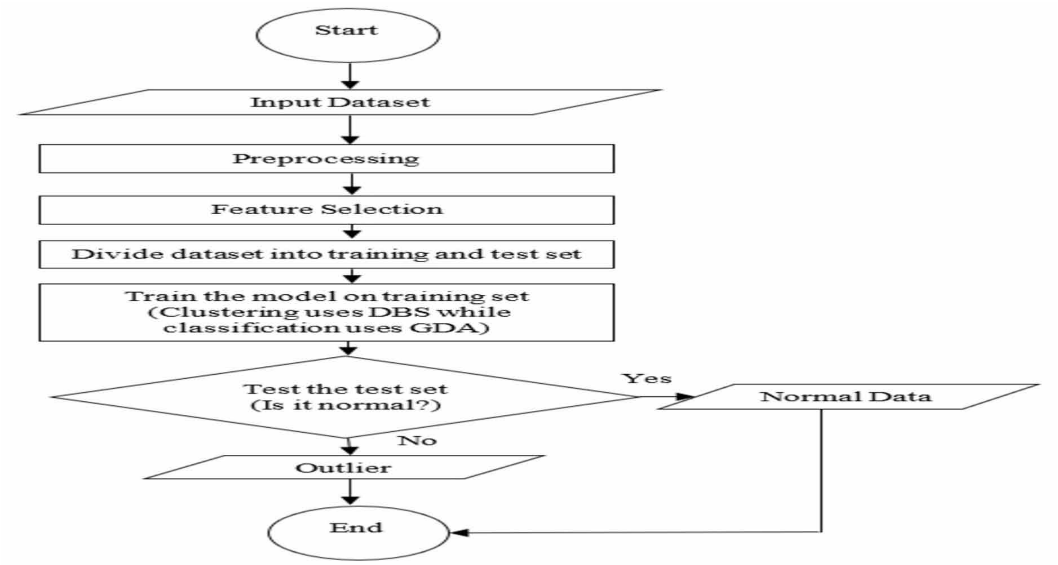


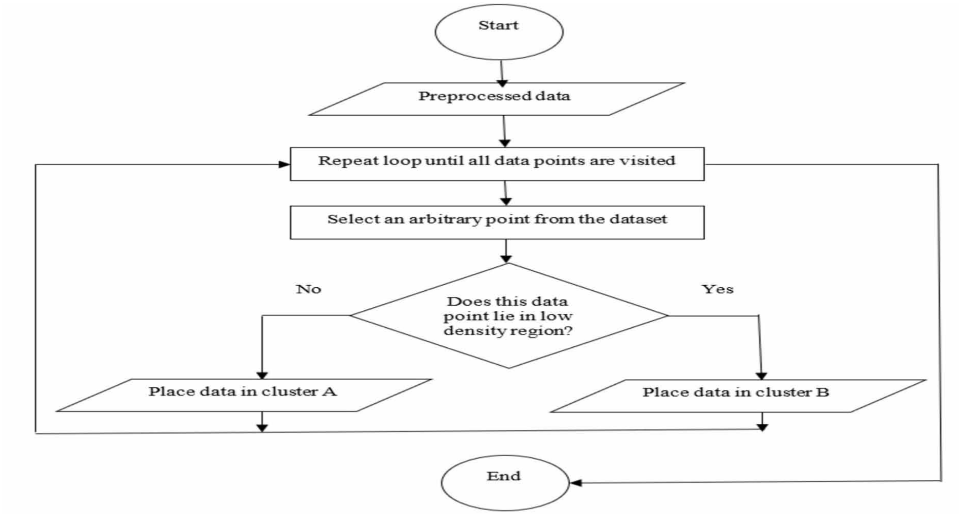
Figure 3. Working of clustering and classification based hybrid approach


Start
Input Dataset
Preprocessing
Feature Selection
Divide dataset into training and test set
Train the model on training set (Clustering uses DBS while classification uses GDA)
Test the test set (Is it normal?)
Yes
Normal Data
No
Outlier
End

Algorithm 1: Outlier Detection With Clustering And Classification Based Hybrid Approach

Input: Dataset
Output: Outliers (Anomalous Data) and Normal Data (Non-Anomalous Data)
Begin
Step 1: Take the input data.
Step 2: Start preprocessing:
i. Remove noise.
ii. Reduce dimensions.
iii. Handle redundancies.
iv. Fill missing values.
v. Normalize data.
Step 3: Select the features.
Step 4: Divide the dataset into the training set and the test set.
Step 5: Train the training data using **CCH:**
i. Create clusters using **Algorithm 2: DBS.**
ii. Implement classification using **Algorithm 3: GDA.**
Step 6: Take test data and apply this learning to get the decision (Normal Data or Outliers).
Step 7: Report the identified outliers.
End

Figure 4. Cluster creation using the density-based scheme


Start
Preprocessed data
Repeat loop until all data points are visited
Select an arbitrary point from the dataset
No
Yes
Does this data point lie in low density region?
Place data in cluster A
Place data in cluster B
End

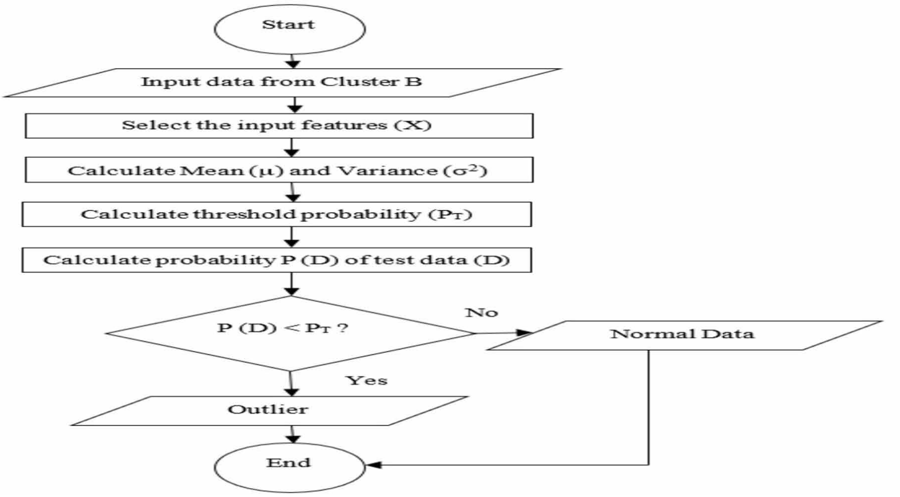
3.3 Gaussian Distribution based Approach (GDA)

Algorithm 3 demonstrates the Gaussian distribution based approach for classification. Here, we set a threshold probability. Then, we compute the probability of the test data using Gaussian distribution. If the probability of the test data is less than the threshold probability, then, it will be an outlier otherwise it will be just normal data. The working of GDA is shown in Figure 5.

Algorithm 2: Density-Based Scheme for Clustering

```
Input: Preprocessed Data
Output: Clusters A and B
Begin
Step 1: Fit the DBS scheme to the training data.
Step 2: Repeat Step 3 to 8 for all input data points of the test data.
Step 3: Find out the maximum distance d between two points of the cluster to become neighbors.
Step 4: Take a new point of the test data as input.
Step 5: Calculate the distance of the new point to all the actual points of the cluster.
Step 6: Check if any point of the cluster is neighbor of the new point of test data as follows:
    If the distance of a new point with the actual cluster points < d then
        it is the neighbor point of the actual cluster point. Place it in cluster A.
    Else
        it may be an outlier. Place it in cluster B.
Step 7: If new behavior of data is observed
        Re-train the model and then test the next data point.
Step 8: Else
        Test the next data point.
Step 9: Report the created cluster B to the next module (GDA) to get the actual outliers.
End
```

Figure 5. Classifying outliers using the Gaussian distribution based approach


Start
Input data from Cluster B
Select the input features (X)
Calculate Mean (μ) and Variance (σ²)
Calculate threshold probability (PT)
Calculate probability P (D) of test data (D)
P (D) < PT ?
No
Normal Data
Yes
Outlier
End

4. PERFORMANCE EVALUATION

The proposed outlier detection algorithm is implemented in Python on x86_64 architecture based Intel Core i7 processor with Windows 10 platform. The scheme is compared with the existing schemes of SVM and mean-shift on two different standard datasets (PhysioNet: 2020; UCI Machine

Algorithm 3: Gaussian Distribution Based Approach for Classification

```
Input: Clustered suspicious data (Cluster B)
Output: Identified outliers
Begin
Step 1: Choose the features ($X_1$, $X_2$,…….$X_m$) that are most likely to be helpful for outlier detection.
Step 2: Find the means ($\mu_1$, $\mu_2$........$\mu_m$) of the features in the test dataset.
Step 3: Find the variance ($\sigma^2{}_1$, $\sigma^2{}_2$,..........$\sigma^2{}_m$) of the features in the test dataset.
Step 4: Compute Threshold probability ($P_T$) using trained dataset:
i. Calculate a list of probabilities of training dataset using P (data, $\mu$, $\sigma^2$).
ii. Calculate predictions using the probability list.
iii. Calculate $f_1$_score using predictions and class data.
// $f_1$_score is a predefined function that returns comparison value between predictions and class //data based on the similarity between them.
iv. Find the probability at which the $f_1$_score is maximum.
//This will be threshold probability
Step 5: Find the probability of the test data D ($d_1$,$d_2$…….., $d_m$)
$P(D)= p (d_1 ; \mu_1; \sigma^2{}_1) * p (d_2 ; \mu_2; \sigma^2{}_2) * ............*p (d_m ; \mu_m; \sigma^2{}_m)$
Step 6: Check the above-calculated probability:
If ($P(D) < P_T$) then it is an Outlier.
Else it is Normal Data.
End
```

Learning Repository: 2020) of the healthcare systems. The first dataset 'dataset1' is small and has six features; while the second dataset 'dataset2' is large and consists of ten features. These datasets contain different numbers of outliers too.

4.1 Experimental Setup

Proposed and the existing outlier detection algorithms are tested on several input data points of two different datasets. Common parameters used in the implementation of these schemes are presented in Table 2.

4.2 Performance Metrics

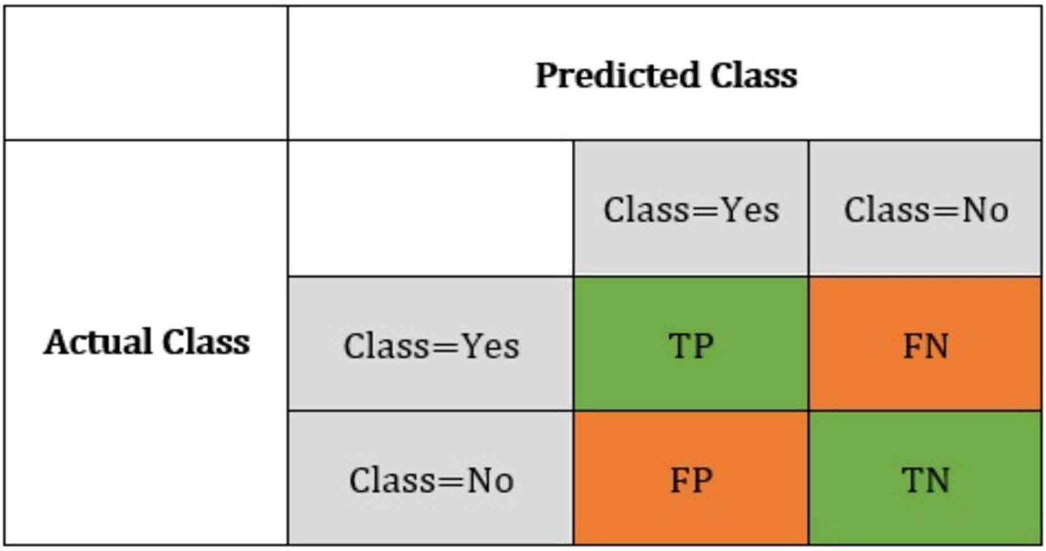
We have evaluated the performance of the proposed scheme on several metrics. Some metrics are based on the confusion matrix and some are evaluated differently. The confusion matrix is a tabular structure that helps to evaluate the performance and accuracy of any machine learning-based algorithm. This matrix is shown in Figure 6. It is represented in terms of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These terms of the confusion matrix are explained below.

TP ¬ positive class data classified into positive class
TN¬ negative class data classified into negative class
FP ¬ negative class data classified into positive class
FN¬ positive class data classified into negative class

Table 2. Parameters and their values

Parameter	Value
Number of datasets used	2
Input data points in dataset1	200-1000
Input data points in dataset2	1000-5000
Features within dataset1	6
Features within dataset2	10
Approach used	Clustering and Classification based Hybrid approach

Figure 6. Confusion matrix



		Predicted Class	
Actual Class		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

FP shows Type1 Error and FN represents Type2 Error. Therefore, we need to minimize them. In the following subsections, we have described four performance metrics (A, P, R, F1) based on the confusion matrix and three performance metrics (C, Th, E) based on some other parameters. These parameters are notified as follows.

N ¬ number of the input data values
n ¬ number of the outliers detected by the system
T ¬ true or actual outliers

4.2.1 Accuracy (A)

It is the measure of precisely classified data and computed according to Eq. (1).

$$\mathrm{A} = \frac{TP + TN}{TP + FP + FN + TN} {*}100 \tag{1}$$

4.2.2 Precision (P)

It is the rate of true positive prediction versus total positive prediction and computed as Eq. (2).

$$\mathrm{P} = \frac{TP}{TP + FP} {*}100 \tag{2}$$

4.2.3 Recall (R)

The recall is also known as Sensitivity. It is the rate of true positive predictions versus the sum of true positive and false negative predictions. It is computed according to Eq. (3).

$$\mathrm{R} = \frac{TP}{TP + FN} {*}100 \tag{3}$$

4.2.4 F1 Score(F1)

It is the combination of both precision and recall and computed as Eq. (4).

$$\mathrm{F1} = \frac{2{*}R{*}P}{R + P} {*}100 \tag{4}$$

4.2.5 Correctness (C)

Correctness is the proportion of actual outliers to the number of detected outliers. It is calculated with T and n as presented in Eq. (5).

$$\mathrm{C}{=}\ \frac{T}{n} {*}\ 100 \tag{5}$$

Now, we calculate the average correctness of any dataset C_{avg} (ds) as given in Eq. (6). Here, C(i) denotes correctness C at particular input i, while z denotes the total number of inputs.

$$C_{avg}\ (\mathrm{ds}) = \frac{\sum_{i=1}^{z} C\left(i\right)}{z} \tag{6}$$

Similarly, we calculate the average correctness of the system C_{avg} as shown in Eq. (7). Here, ds represents a particular dataset, and q is the total number of datasets.

$$C_{avg} = \frac{\sum_{ds=1}^{q} Cavg\left(ds\right)}{q} \tag{7}$$

4.2.6 Throughput (Th)

Throughput means the number of data points that should pass through the system. It is computed with N and n as described in Eq. (8).

Th = (N-n) (8)

Now, we calculate the average throughput of any dataset Th_{avg} (ds) as given in Eq. (9). Here Th(i) denotes throughput Th at particular input i, and z denotes the total number of inputs.

$$\text{Th}_{\text{avg}}\text{ (ds)} = \frac{\sum_{i=1}^{z} Th\left(i\right)}{z} \tag{9}$$

Similarly, we calculate the average throughput of the system Th_{avg} as shown in Eq. (10). Here, ds represents a particular dataset, and q is the total number of datasets.

$$\text{Th}_{\text{avg}} = \frac{\sum_{ds=1}^{q} Thavg\left(ds\right)}{q} \tag{10}$$

4.2.7 Efficiency (E)

Efficiency shows the performance of the system. It is computed with Th and N as written in Eq. (11).

$$\text{E} = \frac{\text{Th}}{\text{N}} *100 \tag{11}$$

Now, we calculate the average efficiency of any dataset E_{avg} (ds) as given in Eq. (12). Here E(i) denotes efficiency E at particular input i, and z denotes the total number of inputs.

$$\text{E}_{\text{avg}}\text{ (ds)} = \frac{\sum_{i=1}^{z} E\left(i\right)}{z} \tag{12}$$

Similarly, we calculate the average efficiency of the system E_{avg} as shown in Eq. (13). Here, ds represents a particular dataset, and q is the total number of datasets.

$$\text{E}_{\text{avg}} = \frac{\sum_{ds=1}^{q} Eavg\left(ds\right)}{q} \tag{13}$$

4.3 **Experimental Results and Analysis**

The proposed and the existing machine learning schemes are tested on two datasets, viz., dataset1 (200 to 1000 input data points), and dataset2 (1000 to 5000 input data points). These datasets are different in terms of size, the number of input features, and the number of outliers therein. The performance of the algorithms is evaluated on various metrics. In the following sub-sections, we discuss various experimental results. The results show that CCH outperforms the existing schemes on both datasets.

4.3.1 Performance Evaluation based on Confusion Matrix Parameters

We compared the proposed scheme with the existing ones by taking the training and test size of the dataset as 70% and 30% respectively. Various performance measures on the samples of 1000 datapoints of dataset1 and 5000 data points of dataset2 are shown in Figure 7 and Figure 8 respectively. Here, we can see that CCH gives the best accuracy, precision, recall, and F1 score on small as well as large datasets as compared to the existing schemes.

4.3.2 Detected Outliers over Various Input Data Points

These plots depict variations in the number of detected outliers against the number of input data points. We observe the different number of outliers over various input data values. On increasing the size of the dataset, an increase in the number of outliers can be seen. It is highly desired that the outliers detected by any scheme should be similar or close to the actual outliers within the dataset. Figure 9 and Figure 10 present the results on dataset1 and dataset2 respectively. Here, we can observe that in the case of CCH, the actual and identified outliers are very close. But, in the case of the existing schemes of SVM and mean-shift, there is a gap between actual and detected outliers. Thus, CCH

Figure 7. Performance evaluation based on confusion matrix parameters on dataset1

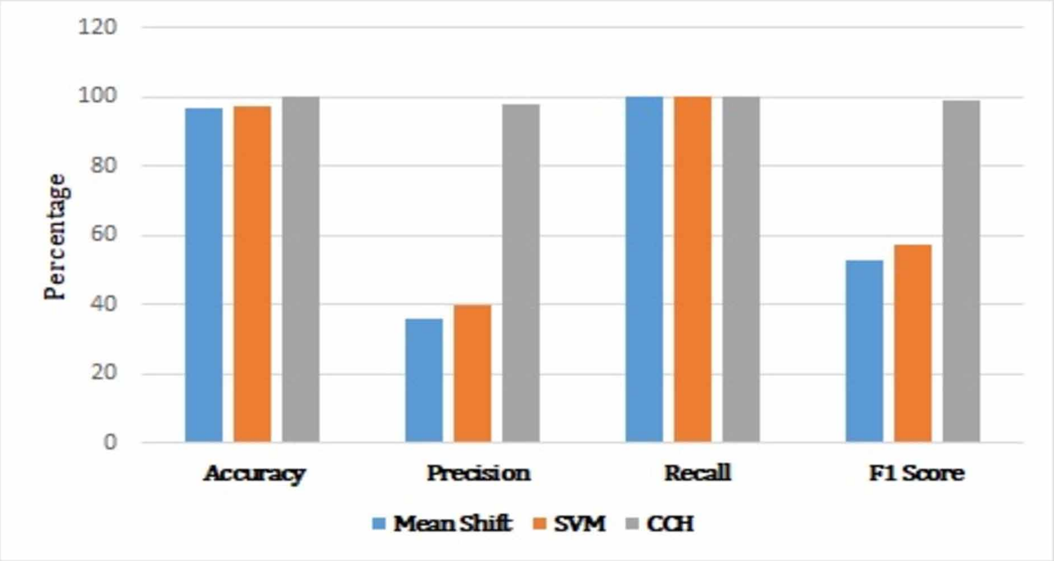


Figure 8. Performance evaluation based on confusion matrix parameters on dataset2

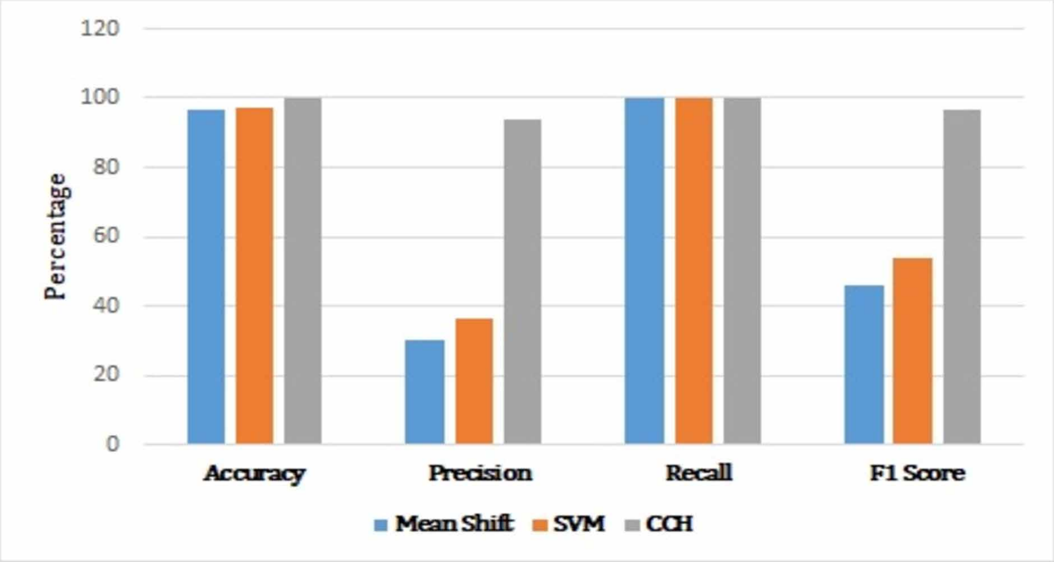


Figure 9. Number of detected outliers over input data of dataset1

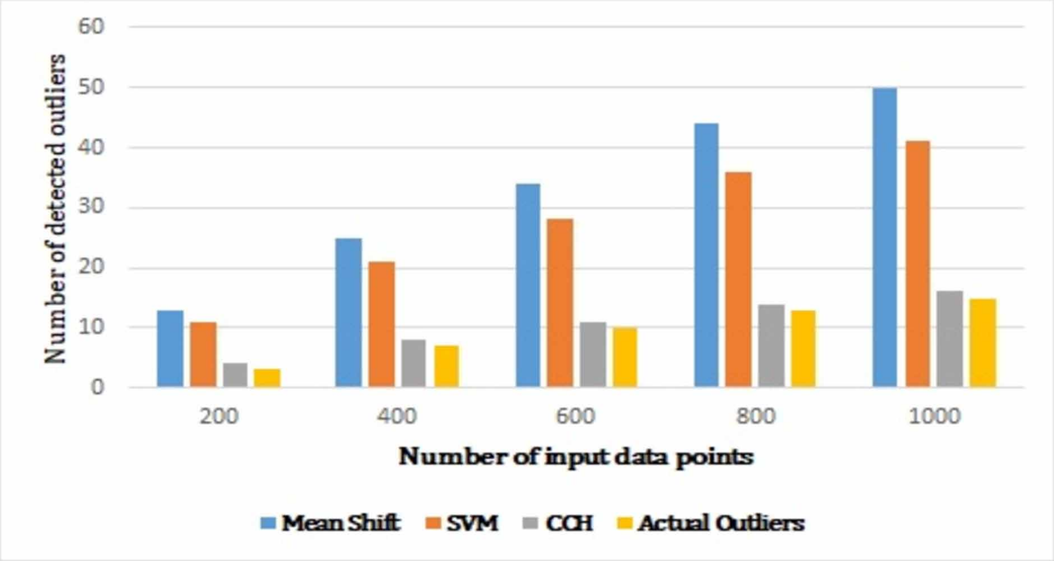
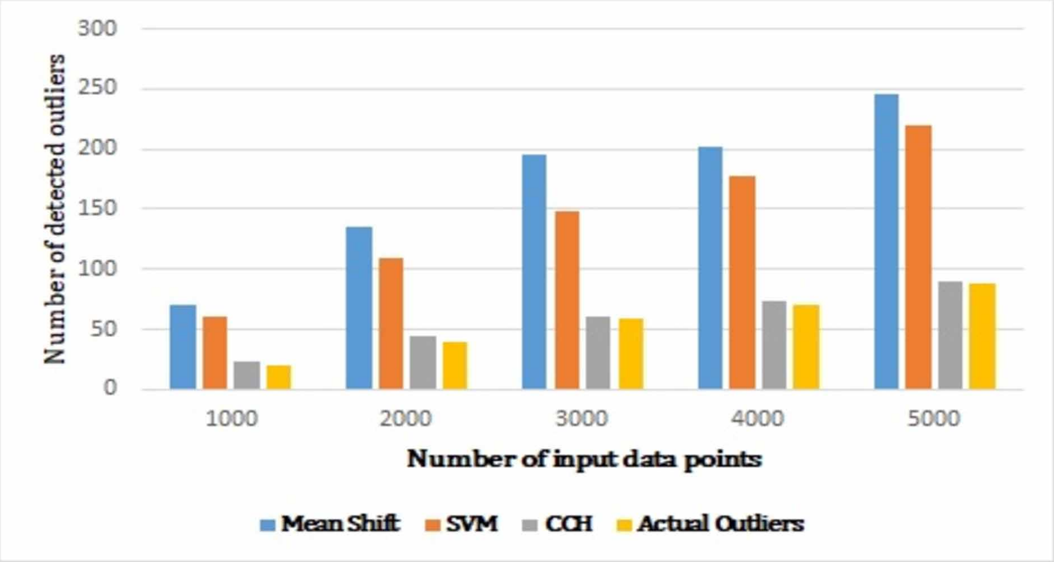


Figure 10. Number of detected outliers over input data of dataset2



identifies the outliers more precisely than the existing schemes. Better precision results in higher throughput and efficiency of the scheme.

4.3.3 Correctness over Various Input Data Points

Figure 11 and Figure 12 depict the comparison of the correctness of CCH, mean-shift, and SVM on dataset1 and dataset2 respectively. Correctness increases on increasing the number of input data values of both datasets in case of the proposed as well as the existing approaches. However, we can see

Figure 11. Correctness over input data of dataset1

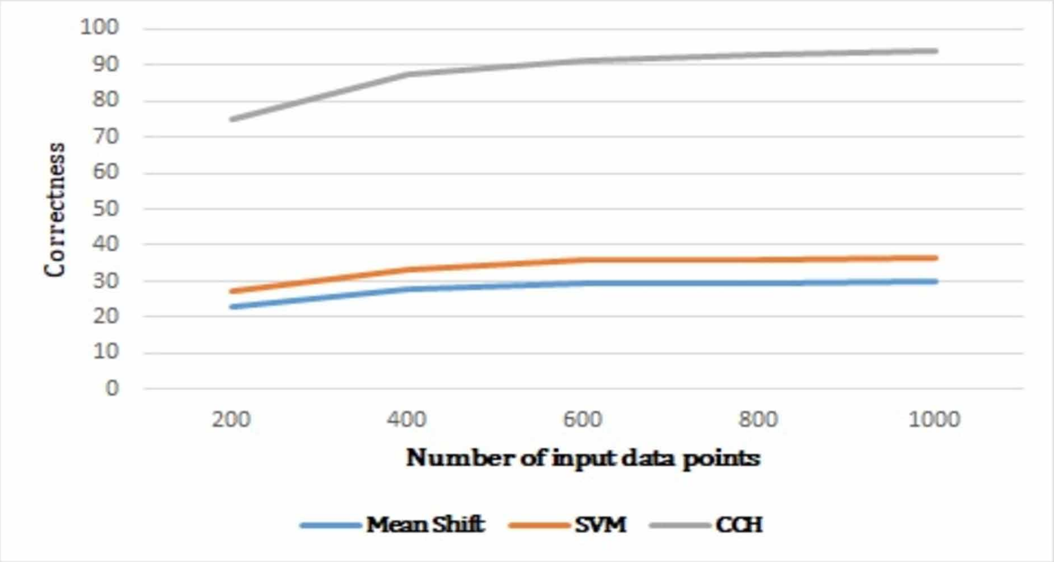
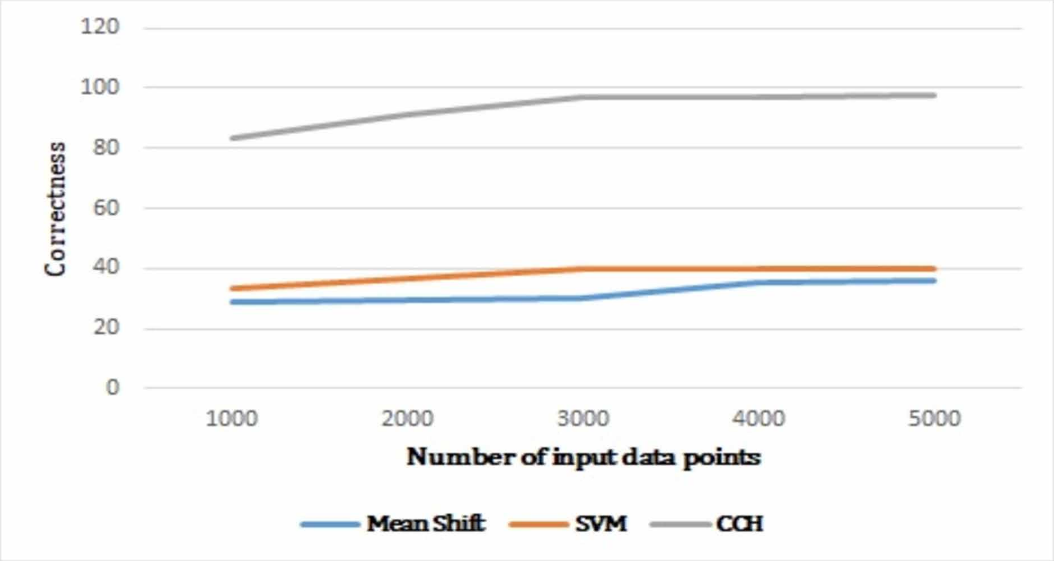


Figure 12. Correctness over input data of dataset2



that the correctness of CCH is better than the existing approaches; as mean-shift and SVM identifies some false outliers while CCH detects almost genuine outliers only.

4.3.4 Throughput over Various Input Data Points

Throughput denotes the actual data points that should be passed through the system. Throughputs of CCH, mean-shift, and SVM approaches are shown in Figure 13 and Figure 14 for dataset1 and dataset2 respectively. Here, we can see that throughput increases on increasing the number of data

Figure 13. Throughput over input data of dataset1

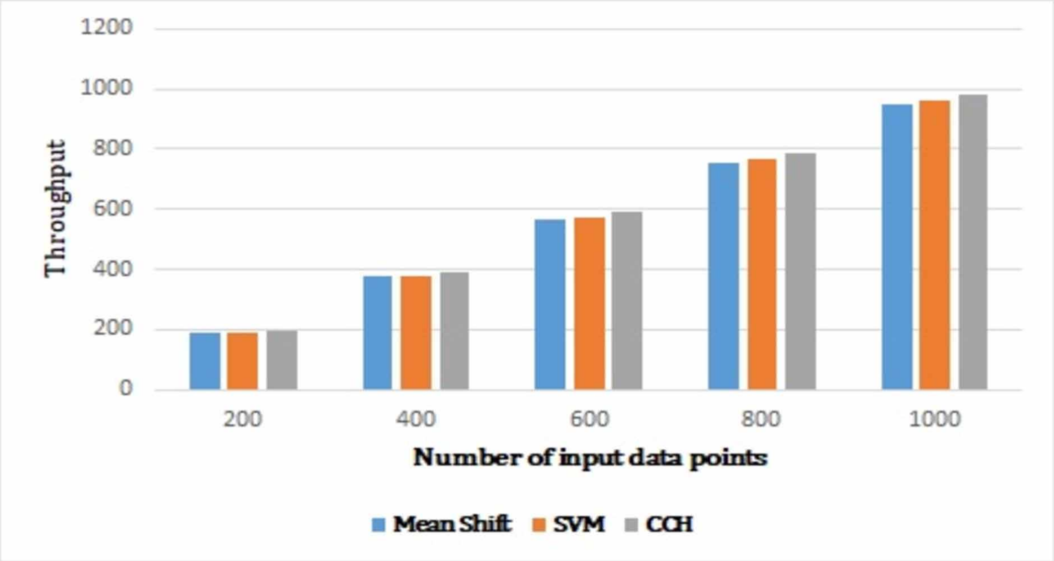
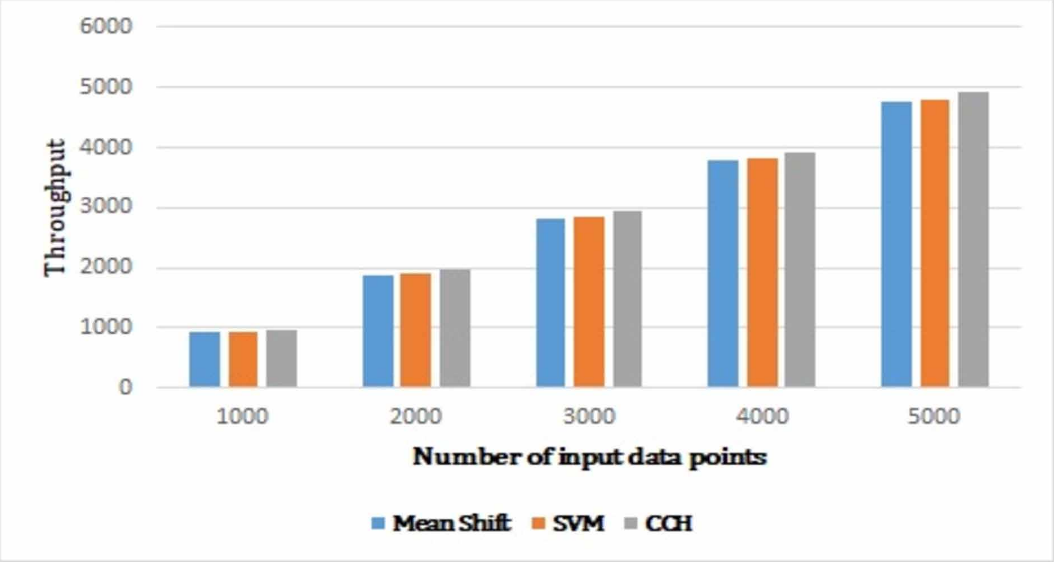


Figure 14. Throughput over input data of dataset2



values of both datasets in case of proposed as well as the existing schemes. However, one can observe that the throughput of CCH is higher than that of the existing schemes. Higher throughput of CCH is due to its higher correctness value.

4.3.5 Efficiency over Various Input Data Points

Good efficiency results in good performance of the system. Figure 15 and Figure 16 compare the efficiency of CCH, mean-shift, and SVM on dataset1 and dataset2 respectively. It is observed that

Figure 15. Efficiency over input data of dataset1

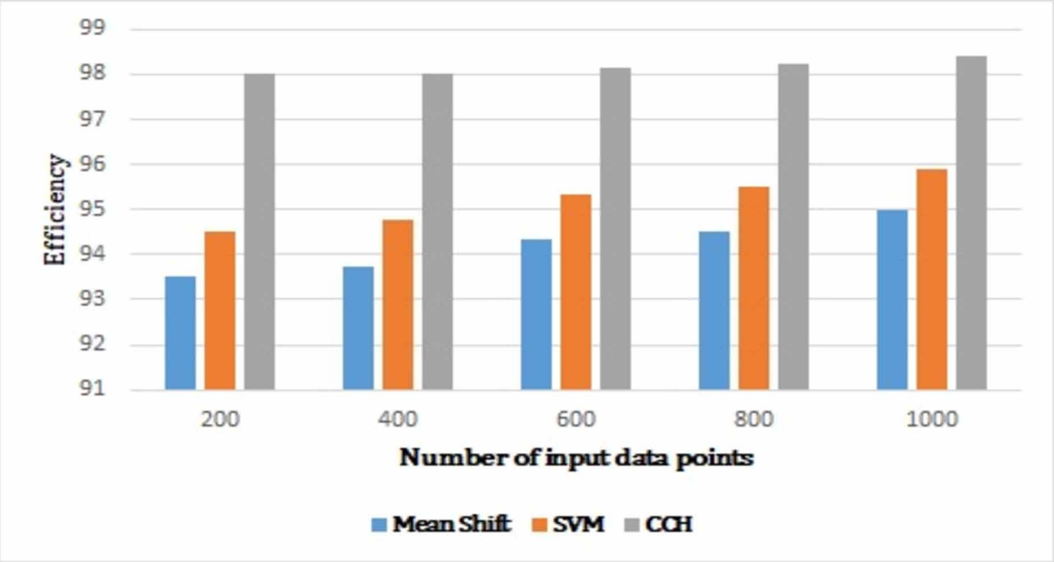
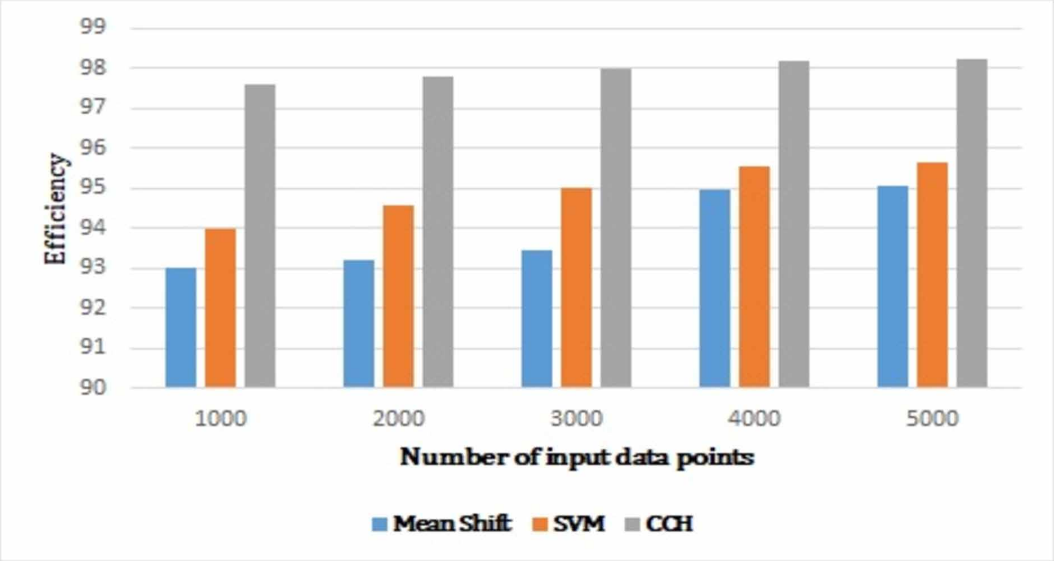


Figure 16. Efficiency over input data of dataset2



the efficiency of these schemes is increasing when we increase the size of datasets. CCH is found more efficient because it identifies the outliers more precisely than the others.

5. RESULT VALIDATION AND DISCUSSION

The algorithms are tested on different input data values of the two standard datasets having different sizes, features, and number of outliers. For 200 input data values of dataset1, the various metrics for mean-shift, SVM, and CCH are calculated mathematically in the following subsections to validate the experimental results.

5.1 Mean Shift Approach

For 200 data points (N=200) of dataset1, the mean-shift scheme identifies 13 outliers (n =13) while actual outliers are 3 (T=3). Here, values of TP, TN, FP, and FN are 3, 187, 10, and 0 respectively. Now, we calculate various performance metrics as follows:

$$\mathrm{A} = \frac{TP+TN}{TP+FP+FN+TN}{*}100 = 95\%$$

$$\mathrm{P} = \frac{TP}{TP+FP}{*}100 = 23.07\%$$

$$\mathrm{R} = \frac{TP}{TP+FN}{*}100 = 100\%$$

$$\mathrm{F1} = \frac{2^{*}R^{*}P}{R+P}{*}100 = 37.49\%$$

$$\mathrm{C} = \frac{T}{n} * 100 = 23.07\%$$

Th= (N-n) = 187

$$E = \frac{Th}{N} *100 = 93.50\%$$

5.2 SVM Approach

For 200 input data values (N=200) of dataset1, SVM identifies 11 outliers`(n=11) while actual outliers are 3 (T=3). Here, values of TP, TN, FP, and FN are 3, 189, 8, and 0 respectively. Now, we compute all the performance measures as follows:

$$A = \frac{TP + TN}{TP + FP + FN + TN} *100 = 96\%$$

$$P = \frac{TP}{TP + FP} *100 = 27.27\%$$

$$R = \frac{TP}{TP + FN} *100 = 100\%$$

$$F1 = \frac{2^* R^* P}{R + P} *100 = 42.85\%$$

$$C = \frac{T}{n} * 100 = 27.27\%$$

Th= (N-n) = 189

$$E = \frac{Th}{N} *100 = 94.50\%$$

5.3 CCH Approach

The proposed approach CCH detects 4 outliers (n=4) while actual outliers are 3 (T=3) for 200 input values (N=200). Here, values of TP, TN, FP, and FN are 3, 196, 1, and 0 respectively. Now, we calculate the various metrics as below:

$$A = \frac{TP + TN}{TP + FP + FN + TN} *100 = 99.5\%$$

$$P = \frac{TP}{TP + FP} *100 = 75\%$$

$$R = \frac{TP}{TP + FN} *100 = 100\%$$

$$F1 = \frac{2^* R^* P}{R + P} *100 = 85.71\%$$

$$C = \frac{T}{n} * 100 = 75.00\%$$

Th= (N-n) = 196

$$\mathrm{E} = \frac{Th}{N} * 100 = 98.00\%$$

Here, we can see that these mathematical results are similar to the experimental results. Similarly, we can compute these metrics (A, P, R, F1, C, Th, and E) for all the input data values of various datasets and the experimental results can be validated with these mathematical results. It is observed that the maximum accuracy of CCH is 99.96% whereas the same of mean-shift and SVM is 96.84% and 97.4% respectively. It shows an improvement of 2.56% and 3.12% in the accuracy of CCH as compared to SVM and mean-shift respectively. Similarly, the average efficiency of mean-shift, SVM, and CCH are noted as 94%, 95%, and 98% respectively. Thus, the average improvement in the efficiency of CCH is 3% and 4% as compared to SVM and mean-shift schemes respectively.

6. SALIENT FEATURES OF THE PROPOSED SCHEME

The proposed scheme CCH outperforms the existing approaches of outlier detection in the healthcare system. It identifies outliers with higher accuracy and efficiency than the existing approaches of the mean-shift and support vector machine. Characteristics of the proposed approach are described as below:

A. *Scalability:* CCH provides scalability according to the requirements. It is suitable for the different sizes of the datasets which means that CCH executes very well irrespective of the size of the datasets.
B. *Integrity*: CCH guarantees the integrity of the system which means that CCH does not produce any modification or loss in the sensed data.
C. *Accuracy*: CCH produces almost accurate results. Outliers detected by CCH are very similar to the actual outliers within the dataset while existing approaches identify some more false outliers. The maximum accuracy of CCH is found 99.96% which is far better than that of the existing machine learning-based outlier detection schemes.
D. *Efficiency:* Efficiency represents the performance of the system. Better efficiency leads to better performance of the system. The average efficiency of CCH is 98% which is better than that of the existing machine learning-based outlier detection schemes.

7. CONCLUSIONS AND FUTURE WORK

The proposed scheme uses a hybrid approach of supervised and unsupervised machine learning techniques for outlier detection. We used a two-level framework in which we applied clustering at first level and then classification at the second level to identify the outliers precisely. We devised a density-based scheme for clustering and a Gaussian distribution based approach for classification.

We compared our approach CCH with one clustering scheme namely mean-shift and one classification scheme called SVM. The performance of the proposed scheme is evaluated for different input data values on the various metrics. We tested these schemes on the two different standard datasets of the healthcare system available in the public domain. These datasets are of different sizes. They are having different numbers of features and outliers too. It is found that actual outliers and the outliers identified by the CCH are very similar while the existing techniques detect some false outliers. Therefore, we get maximum accuracy as 99.96% in the CCH which depicts the improvements of 2.56% and 3.12% as compared to the existing schemes of SVM and mean-shift respectively. The average efficiency of the proposed approach is found 98% which shows the improvements of 3% and

4% as compared to SVM and mean-shift schemes respectively. Analytical validation has been carried out to justify the experimental results.

In the future, we plan to extend this hybrid approach for the highly complex and unstructured healthcare data. The scheme can also be tested on datasets of a variety of other IoT applications, viz., smart farming, smart home, smart vehicles, industrial IoT, and forest fire monitoring system.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENT

This research is partially funded by the Technical Education Quality Improvement Program (TEQIP III).

REFERENCES

Ahmed, E., Bessis, N., & Shahzad, W. (2016). Data Matching: An Algorithm for Detecting and Resolving Anomalies in Data Federation. *Journal of Basic and Applied Scientific Research*, 21–31.

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). *A survey of anomaly detection techniques in financial domain, Future Generation Computer Systems* (Vol. 55). Elsevier.

Aleksandrova, E., & Anagnostopoulos, C. (2019). Adaptive Principal Component Analysis-Based Outliers Detection Through Neighborhood Voting in Wireless Sensor Networks. In I. Comşa & R. Trestian (Eds.), *Next-Generation Wireless Networks Meet Advanced Machine Learning Applications* (pp. 255–285). IGI Global. doi:10.4018/978-1-5225-7458-3.ch011

Ashtari, S., & Bellamy, A. (2019). *Factors Impacting Use of Health IT Applications: Predicting Nurses' Perception of Performance. International Journal of Healthcare Information Systems and Informatics, 14(4).* doi:10.4018/IJHISI.2019100103

Bansal, D. R., & Kishore, B. (2018). Feature selection in support vector machines for outlier detection. *Second IEEE International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 112-115. doi:10.1109/ICECA.2018.8474578

Bessis, N., Asimakopoulou, E., & Xhafa, F. (2011). A next generation emerging technologies roadmap for enabling collective computational intelligence in disaster management. *International Journal of Space-Based and Situated Computing, 1*(1), 76-85.

Bosman, H., Iacca, G., Tejada, A., Wörtche, H. J., & Liotta, A. (2017). *Spatial anomaly detection in sensor networks using neighbourhood information, Information Fusion 33*. Elsevier.

Deng, X., Jiang, P., Peng, X., & Mi, C. (2018). *Support high-order tensor data description for outlier detection in high-dimensional big sensor data, Future Generation Computer Systems* (Vol. 81). Elsevier.

Devadevan, V., & Sankaranarayanan, S. (2017). *Forest Fire Information System Using Wireless Sensor Network. International Journal of Agricultural and Environmental Information Systems (IJAEIS), 8(3).*

Dwivedi, R. K., Saran, M., & Kumar, R. (2019). A Survey on Security over Sensor-Cloud. *9th IEEE International Conference on Cloud Computing, Data Science & Engineering – Confluence*, 31-37.

Dwivedi, R. K., Pandey, S., & Kumar, R. (2018). A study on Machine Learning Approaches for Outlier Detection in Wireless Sensor Network. *8th IEEE International Conference on Cloud Computing, Data Science & Engineering – Confluence*, 189-192.

Ensari, T., Günay, M., Nalçakan, Y., & Yildiz, E. (2019). Overview of Machine Learning Approaches for Wireless Communication. In I. Comşa & R. Trestian (Eds.), *Next-Generation Wireless Networks Meet Advanced Machine Learning Applications* (pp. 123–140). IGI Global. doi:10.4018/978-1-5225-7458-3.ch006

Ghorbel, O., Ayedi, W., Snoussi, H., & Abid, M. (2015). Fast and efficient outlier detection method in Wireless Sensor Networks. *IEEE Sensors Journal*, *15*(6), 3403–3411. doi:10.1109/JSEN.2015.2388498

Gil, P., Martins, H., & Januário, F. (2016). *Detection and accommodation of outliers in Wireless Sensor Networks within a multi-agent framework, Applied Soft Computing 42*. Elsevier.

Gope, P., & Hwang, T. (2016). *BSN-Care: A Secure IoT-Based Modern Healthcare System Using Body Sensor Network. Sensors Journal, 16(5).*

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). *Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions. Future Generation Computer Systems, 29(7).*

Han, G. W., Zhang, Y., Lu, N. Y., Jiang, B., Xu, Z. X., Cao, J. R., & Shi, X. (2017). Incipient anomaly detection for railway vehicle door system based on adaptive mean shift clustering. *IEEE Chinese Automation Congress (CAC)*, 1297-1302. doi:10.1109/CAC.2017.8242967

Haque, S. A., Rahman, M., & Aziz, S. M. (2015). Sensor Anomaly Detection in Wireless Sensor Networks for Healthcare. *Sensors (Basel)*, *15*(4), 8764–8786. doi:10.3390/s150408764 PMID:25884786

Hauskrecht, M., Batal, I., Hong, C., Nguyen, Q., Cooper, G. F., Visweswaran, S., & Clermont, G. (2016). Outlier-based detection of unusual patient-management actions: An ICU study. *Journal of Biomedical Informatics, 64*, 211–221.

Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G.F., & Clermont, G. (2013). Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics, 46*(1), 47-55.

Jain, D., & Singh, V. (2020). *A Novel Hybrid Approach for Chronic Disease Classification. International Journal of Healthcare Information Systems and Informatics (IJHISI), 15(1).*

Jayaraman, R., Salah, K., & King, N. (2019). Improving Opportunities in Healthcare Supply Chain Processes via the Internet of Things and Blockchain Technology. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 14*(2), 49-65.

Ji, M., & Xing, H. (2017). Adaptive-weighted one-class support vector machine for outlier detection. *29th IEEE Chinese Control And Decision Conference (CCDC)*, 1766-1771. doi:10.1109/CCDC.2017.7978802

Kaplantzis, S., Shilton, A., Mani, N., & Sekercioglu, Y. A. (2007). Detecting Selective Forwarding Attacks in Wireless Sensor Networks using Support Vector Machines. *3rd IEEE International Conference on Intelligent Sensors, Sensor Networks and Information*, 335-340. doi:10.1109/ISSNIP.2007.4496866

Kashyap, R. (2019). Machine Learning, Data Mining for IoT-Based Systems. In G. Kaur & P. Tomar (Eds.), Handbook of Research on Big Data and the IoT. IGI Global.

Kashyap, R. (2019). Machine Learning for Internet of Things. In I. Comşa & R. Trestian (Eds.), *Next-Generation Wireless Networks Meet Advanced Machine Learning Applications* (pp. 57–83). IGI Global. doi:10.4018/978-1-5225-7458-3.ch003

Martins, H., Januário, F., Palma, L., Cardoso, A., & Gil, P. (2015). A Machine Learning Technique in a Multi-Agent Framework for Online Outliers Detection in Wireless Sensor Networks. *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 688-693.

Martins, H., Palma, L., Cardoso, A., & Gil, P. (2015). A Support Vector Machine Based Technique for Online Detection of Outliers in Transient Time Series. *10th IEEE Asian Control Conference (ASCC)*, 1-6. doi:10.1109/ASCC.2015.7244794

Mattos, T.B., & Ferreira, C.S. (2016). The Mean-shift Outlier Model under Skew Normal Distribution. *Communications in Statistics - Simulation and Computation, 45*(6), 1905-1917.

Ozertem, U., Erdogmus, D., & Jenssen, R. (2008). Mean shift spectral clustering. *Pattern Recognition, 41*(6), 1924-1938.

Pachauri, G., & Sharma, S. (2015). Anomaly detection in medical wireless sensor networks using machine learning algorithms. *Procedia Computer Science, 70*, 325 – 333.

Petrakis, E. G. M., Sotiriadis, S., Soultanopoulos, T., Renta, P. T., Buyya, R., & Bessis, N. (2018). *Internet of things as a service (iTaaS): challenges and solutions for management of sensor data on the cloud and the fog, Internet of Things* (Vol. 3). Elsevier.

Rath, M., & Mishra, S. (2019). Advanced-Level Security in Network and Real-Time Applications Using Machine Learning Approaches. In M. Khan (Ed.), *Machine Learning and Cognitive Science Applications in Cyber Security* (pp. 84–104). IGI Global. doi:10.4018/978-1-5225-8100-0.ch003

Shahid, N., Naqvi, I. H., & Qaisar, S. B. (2012). Real Time Energy Efficient Approach to Outlier & Event Detection in Wireless Sensor Networks. *2012 IEEE International Conference on Communication Systems (ICCS)*, 162-166. doi:10.1109/ICCS.2012.6406130

Shahid, N., Naqvi, I. H., & Qaisar, S. B. (2012). Quarter-Sphere SVM: Attribute and Spatio Temporal Correlations based Outlier & Event Detection in Wireless Sensor Networks. *IEEE Wireless Communications and Networking Conference (WCNC)*, 2048-2053. doi:10.1109/WCNC.2012.6214127

Sharma, N. V., & Yadav, N. S. (2019). Machine Learning in Wireless Communication: A Survey. In I. Comşa & R. Trestian (Eds.), *Next-Generation Wireless Networks Meet Advanced Machine Learning Applications* (pp. 141–161). IGI Global. doi:10.4018/978-1-5225-7458-3.ch007

Singh, S. K., & Goyal, A. (2020). *Performance Analysis of Machine Learning Algorithms for Cervical Cancer Detection. International Journal of Healthcare Information Systems and Informatics (IJHISI), 15(2).*

Sittig, D.F., Belmont, E., & Singh, H. (2018). Improving the safety of health information technology requires shared responsibility: It is time we all step up. *Healthcare: The Journal of Delivery Science and Innovation, 6*(1), 7-12.

Snoussi, H., Ghorbel, O., Jmal, M., & Abid, M. (2015). Distributed and Efficient One-Class Outliers Detection Classifier in Wireless Sensors Networks. In *13th International Conference on Wired/Wireless Internet Communication (WWIC)*. Malaga, Spain: Springer International Publishing.

Thilakanathan, D., Chen, S., Nepal, S., Calvo, R., & Alem, L. (2014). A platform for secure monitoring and sharing of generic health data in the Cloud. *Future Generation Computer Systems, Elsevier*, *35*, 102–113. doi:10.1016/j.future.2013.09.011

Wang, X., Su, K., & Su, L. (2019). *Research on Improved Apriori Algorithm Based on Data Mining in Electronic Cases. International Journal of Healthcare Information Systems and Informatics (IJHISI), 14(3).*

Xu, L., Yeh, Y., Lee, Y., & Li, J. (2013). A Hierarchical Framework using Approximated Local Outlier Factor for Efficient Anomaly Detection. *Procedia Computer Science, 19*, 1174 – 1181.

Xu, S., Hu, C., Wang, L., & Zhang, G. (2012). Support Vector Machines based on K Nearest Neighbor Algorithm for Outlier Detection in WSNs. *8th IEEE International Conference on Wireless Communications, Networking and Mobile Computing*, 1-4. doi:10.1109/WiCOM.2012.6478696

Yenke, B. O., Aboubakar, M., Titouna, C., Ari, A. A. A., & Gueroui, A. (2017). Adaptive Scheme for Outliers Detection in Wireless Sensor Networks. *International Journal of Computer Networks and Communications Security*, *5*(5), 105–114.

Zhang, Y., Meratinia, N., & Havinga, P.J.M. (2016). Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. *Ad Hoc Networks, 11*(3), 1062-1074.

Zhang, Y., Meratnia, N., & Havinga, P. (2009). Adaptive and Online One-Class Support Vector Machine-based Outlier Detection Techniques for Wireless Sensor Networks. *International Conference on Advanced Information Networking and Applications Workshops*, 990-995. doi:10.1109/WAINA.2009.200

PhysioNet. Dataset. (n.d.). https://www.physionet.org/content/mimicdb/1.0.0/

Machine Learning RepositoryU. C. I. Dataset. (n.d.). https://archive.ics.uci.edu/ml/datasets.php

Rajendra Kumar Dwivedi is Assistant Professor in the Department of Information Technology and Computer Applications at Madan Mohan Malaviya University of Technology, Gorakhpur (U.P.), India. He joined this institute in 2009. He received his B. Tech Degree in 2004 from Pt Ravishanker Shukla University, Raipur and M.Tech. from Indian Institute of Technology, Roorkee in 2015. Currently, he is pursuing his Ph.D. from Department of Computer Science and Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur (U.P.). Before joining MMM Engineering College (under state government of U.P.), he worked in K.V. Lansdowne U.K. (under central government of India). He has supervised a large number of M. Tech. students. He has published a large number of research papers in various international and national journals and conferences of high repute (h-index=10, i-10 index=12, citations=272). He is a member of IEEE and also life member of Institution of Engineers (India). His main research interests lie in Wireless sensor networks, Network security, Cloud computing and Machine learning.

Rakesh Kumar, PhD., is Professor in the Department of Computer Science and Engineering at Madan Mohan Malaviya University of Technology, Gorakhpur (U.P.), India. He received his B. Tech. Degree in 1990 from MMM Engineering College, Gorakhpur and M.E. from SGS Institute of Technology and Science, Indore in 1994. He did his Ph.D. from Indian Institute of Technology, Roorkee in 2011. Before joining MMM Engineering College, he worked in HBTI Kanpur and BIET Jhansi. He was also the principal investigator of a major research project sanctioned from University Grant Commission, New Delhi, India. Dr. Kumar has supervised a large number of M. Tech. Dissertations and guiding several Ph.D. students. He has published a large number of research papers in various international and national journals and conferences of high repute (h-index=13, i-10 index=20, citations=760). He is a member of IEEE, life member of CSI, ISTE and also a Fellow of IETE and Institution of Engineers (India). His main interests lie in mobile ad hoc network, MANET- Internet integration, Sensor network, Network security, Cloud computing and Machine learning.

Rajkumar Buyya, PhD., is a Redmond Barry Distinguished Professor and Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft, a spin-off company of the University, commercializing its innovations in Cloud Computing. He served as a Future Fellow of the Australian Research Council during 2012-2016. He has authored over 625 publications and seven text books including "Mastering Cloud Computing" published by McGraw Hill, China Machine Press, and Morgan Kaufmann for Indian, Chinese and international markets respectively. He also edited several books including "Cloud Computing: Principles and Paradigms" (Wiley Press, USA, Feb 2011). He is one of the highly cited authors in computer science and software engineering worldwide (h-index=149, i-10 index=634, g-index=280, 116862 citations). "A Scientometric Analysis of Cloud Computing Literature" by German scientists ranked Dr. Buyya as the World's Top-Cited (#1) Author and the World's Most-Productive (#1) Author in Cloud Computing. Dr. Buyya is recognized as a "Web of Science Highly Cited Researcher" for three consecutive years since 2016, a Fellow of IEEE, and Scopus Researcher of the Year 2017 with Excellence in Innovative Research Award by Elsevier and recently (2019) received "Lifetime Achievement Awards" from two Indian universities for his outstanding contributions to Cloud computing and distributed systems. Software technologies for Grid and Cloud computing developed under Dr. Buyya's leadership have gained rapid acceptance and are in use at several academic institutions and commercial enterprises in 40 countries around the world. Dr. Buyya has led the establishment and development of key community activities, including serving as foundation Chair of the IEEE Technical Committee on Scalable Computing and five IEEE/ACM conferences. These contributions and international research leadership of Dr. Buyya are recognized through the award of "2009 IEEE Medal for Excellence in Scalable Computing" from the IEEE Computer Society TCSC. Manjrasoft's Aneka Cloud technology developed under his leadership has received "2010 Frost & Sullivan New Product Innovation Award". Recently, Dr. Buyya received "Mahatma Gandhi Award" along with Gold Medals for his outstanding and extraordinary achievements in Information Technology field and services rendered to promote greater friendship and India-International cooperation. He served as the founding Editor-in-Chief of the IEEE Transactions on Cloud Computing. He is currently serving as Co-Editor-in-Chief of Journal of Software: Practice and Experience, which was established ~50 years ago. For further information on Dr.Buyya, please visit his cyberhome: www.buyya.com.