


# Gaussian Distribution-Based Machine Learning Scheme for Anomaly Detection in Healthcare Sensor Cloud

Rajendra Kumar Dwivedi, Madan Mohan Malaviya University of Technology, India

 <https://orcid.org/0000-0001-6682-1942>

Rakesh Kumar, Madan Mohan Malaviya University of Technology, India

Rajkumar Buyya, The University of Melbourne, Australia

## ABSTRACT

Smart information systems are based on sensors that generate a huge amount of data. This data can be stored in cloud for further processing and efficient utilization. Anomalous data might be present within the sensor data due to various reasons (e.g., malicious activities by intruders, low quality sensors, and node deployment in harsh environments). Anomaly detection is crucial in some applications such as healthcare monitoring systems, forest fire information systems, and other internet of things (IoT) systems. This paper proposes a Gaussian distribution-based supervised machine learning scheme of anomaly detection (GDA) for healthcare monitoring sensor cloud, which is an integration of various body sensors of different patients and cloud. This work is implemented in Python. Use of Gaussian statistical model in the proposed scheme improves precision, throughput, and efficiency. GDA provides 98% efficiency with 3% and 4% improvements as compared to the other supervised learning-based anomaly detection schemes (e.g., support vector machine [SVM] and self-organizing map [SOM], respectively).

## KEYWORDS

Anomaly Detection, Gaussian Distribution Approach, Machine Learning, Outlier, Self-Organizing Map (SOM), Supervised Learning, Support Vector Machine (SVM), Wireless Sensor Network (WSN)

## 1. INTRODUCTION

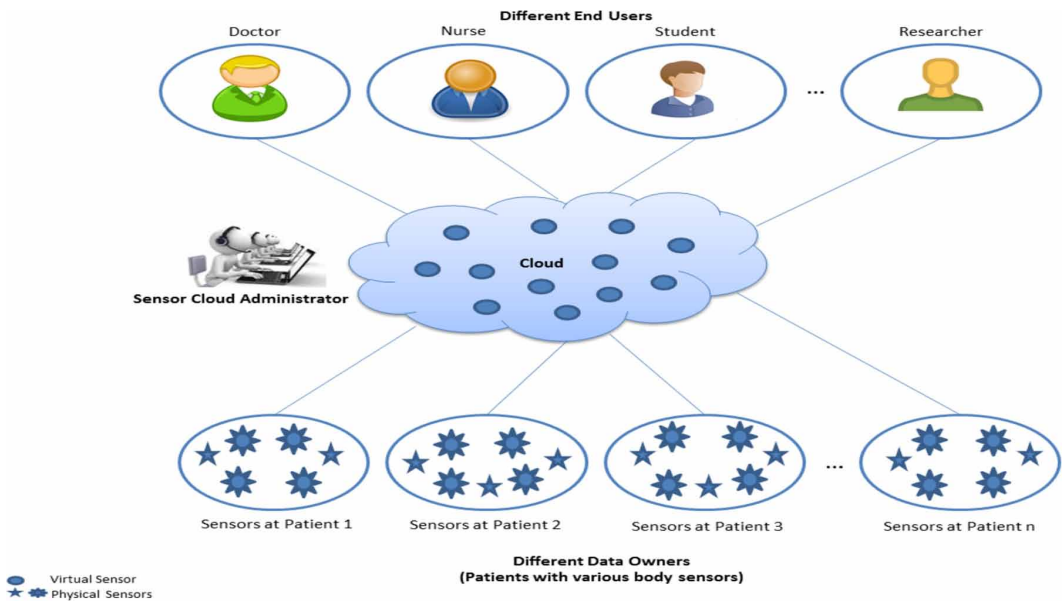
These days, various smart systems have been developed to facilitate monitoring and management of physical and human environments in many ways. Such smart systems are also known as Internet of Things (IoT) systems. Sensor based IoT systems have various applications such as healthcare monitoring, battlefield monitoring, street monitoring, disaster management, military applications, forest fire detection, unmanned vehicles and manufacturing industries (Bessis 2011; Lounis et al. 2016). Such IoT applications generate a huge amount of data that is usually stored at cloud to increase usefulness of the resources (Thilakanathan et al. 2014). Sensor networks are integrated with cloud to improve the effectiveness of the applications. This integration is termed as sensor cloud which is beneficial for both sensor networks and cloud. Various sensor networks store their sensed data at the cloud. These physical sensors are mapped with virtual sensors at cloud. Sensor cloud administrator integrates the sensed data from various sensors into the unified standard with help of virtualization

DOI: 10.4018/IJCAC.2021010103

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

at cloud. Thus, cloud can provide sensor as a service with help of virtualization to the multiple users according to their choice and demand. Any genuine end user can access the data of one or all authorized sensor networks just in one click with help of this integration (Dwivedi et al. 2019). Figure 1 presents a healthcare monitoring system where each human body behaves as a sensor network. Here, data from various wearable body sensors of many patients have been stored at cloud through base station such as mobile phone. Different types of authorized users viz., doctors, nurses, medical students and researchers can access the health records of the patients using their credentials. Doctors can provide medical support to the patients anytime and from anywhere with this system. They can help the patients instantly if the emergency case is observed. Data owners may also earn money for providing their data at cloud in some cases. Cloud can provide sensor as a service to the authorized students and researchers by providing them various types of data. Thus, legitimate end users can get data of one or more patients easily and quickly. Doctors, nurses, students, researchers and patients may belong to either same or different hospitals in this healthcare system. In this way, everyone is benefitted with this sensor cloud integration.

Figure 1. Healthcare monitoring sensor cloud



There are several research issues and challenges in sensor networks viz., node failure, network lifetime, load balancing, routing, data aggregation, localization, power efficiency, QoS, security, outliers and anomaly detection etc. (Ahmed et al. 2016; Petrakis et al. 2018). By resolving these issues the performance of this network can be increased. This paper focuses on the issue of anomaly detection in sensor cloud of healthcare system. There are many medical cases in which continuous monitoring of health conditions is required which allow doctors to know the health status of the patients regularly or when required. IoT based smart healthcare system is very helpful in such situations which minimizes the healthcare treatment cost and allows the mobility of patients too. In such systems, various body sensors are applied at the patients for the purpose of continuous monitoring of their health status. These body sensors are wearable devices worn by patients that can collect various body data such as blood pressure, body temperature and heart beat rate. Data collected by these sensors

are sent to gateways via wireless communication medium and from gateways finally transferred to the cloud for storage and processing. Medical data of the patients collected by various body sensors are very crucial. Any alteration or loss in the medical data of the patients may result in negative health conditions or sometimes lead to very serious situations. Hence, it must be accurate. Some false alarms might be generated due to various reasons such as malicious activities performed by intruders or malfunctioning sensors. Such anomalies must be detected so that information received by the user could be correct and accurate. This situation can be handled by performing data analysis using machine learning techniques on the sensed data where outliers or anomalies can be easily detected and removed. Research shows that use of supervised machine learning techniques like Bayesian Belief Network (BBN), K Nearest Neighbors (KNN), Self Organizing Map (SOM) and Support Vector Machine (SVM) offer efficient solution for anomaly detection (Xu et al. 2012; Xu et al. 2013; Yenke et al. 2017). However, this efficiency can be further improved. Therefore, a new model of anomaly detection needs to be developed which should reduce the computational complexities and improve the efficiency. In this direction, this paper proposes a Gaussian distribution based supervised machine learning technique for anomaly detection in smart healthcare system and named as GDA. Proposed approach gives 98% efficiency that is an improvement of 3% and 4% as compared to SVM and SOM based schemes respectively. The major contributions of the paper are as follows:

- Design of a Gaussian distribution based supervised machine learning scheme for anomaly detection
- Analytical validation to justify the implementation results

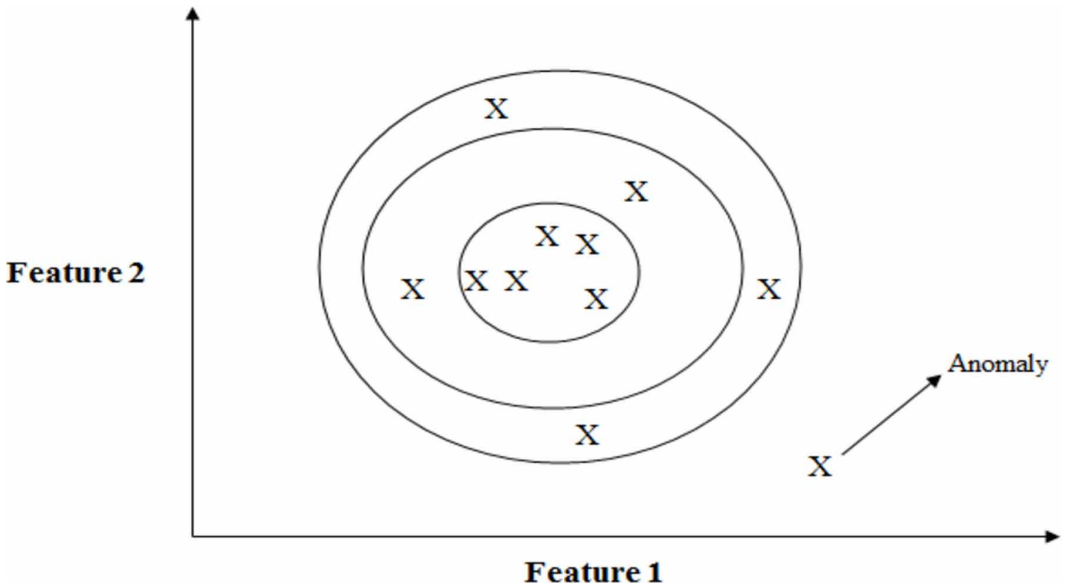
The rest of the paper is organized as follows. A brief survey of the related work is described in section 2. The proposed scheme is discussed in section 3. Section 4 presents performance evaluation of the proposed work. It also compares the proposed scheme with other existing schemes. Finally, section 5 concludes the work with some future directions.

## **2. RELATED WORK**

Anomalies can occur due to some malicious actions performed by the intruders, malfunctioning sensors or abnormal behaviour of the sensor nodes (Ghorbel et al. 2015; Gil et al. 2016; Dwivedi et al. 2018). In healthcare systems, a malfunctioning body sensor may generate false alarms by producing out of range data while actual data of the patient is of normal range. There may be several other instances like this. Deviation of the data from the actual family of data can be seen in two forms viz., outliers (anomalies) and missing data. Outliers are the data generated by some nodes which deviates so much from the data generated by neighbor nodes (Bosman et al. 2017). There is also a possibility that the data at the node is unable to be detected by the other nodes. This data is termed as missing data. Figure 2 describes the anomalous data. Focus of this paper is on outlier or anomaly detection.

Machine Learning is an emerging area of this generation. It helps the machine to learn from the environment (Ensari et al. 2019). The machine itself enhances the performance in future because the experience of the machine is enhanced (Alsheikh et al. 2014; Ayadi et al. 2017; Aleksandrova et al. 2019). Supervised learning is a part of machine learning in which there is a labelled trained data set and known output. There are some input points and exact output for those points. We also have an idea about the relationship of input with output (Forster et al. 2011; Fawzy et al. 2013). These learning techniques can be used for anomaly detection in sensor data. This section presents taxonomy of machine learning algorithms used for anomaly detection along with a comparative analysis of the related schemes on basis of their properties and complexity.

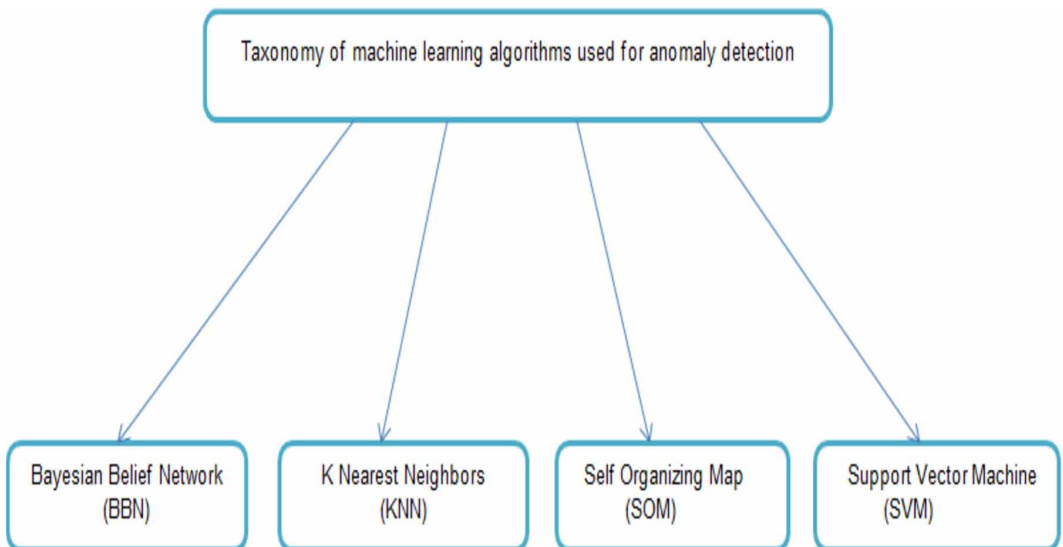
Figure 2. Anomalous data



### 2.1 Taxonomy of Machine Learning Algorithms for Anomaly Detection

Figure 3 presents the taxonomy of machine learning algorithms used for anomaly detection. These algorithms can be categorised as BBN, KNN, SOM and SVM which are discussed below with a brief literature survey.

Figure 3. Taxonomy of machine learning algorithms used for anomaly detection



### 2.1.1 Bayesian Belief Network (BBN)

BBN is a directed graph. The nodes of graph represent variables (discrete or continuous) and the arcs represent causal relationships between the variables. There are mainly three steps in the process of anomaly detection using BBN viz., construct the BBN, learn the BBN and infer from BBN.

*Kirk et al. (2014)* considered the BBNs in order to design an algorithm for outlier detection. Most of the neighbour sensors have close readings. This method presents conditional dependencies within all the sensor readings. Bayesian Belief Network uses the conditional relationships using probabilistic approach among the values coming from nodes to detect any outliers present in the collected data. This method can also be used for the treatment of missing data at the sensor nodes.

*Janakiram et al. (2006)* uses “Naive Bayes” assumption for training the spatial data or the series time data. It models the spatial data using some neighbors only. The “Naive Bayes” assumption states that the naive based values are independent. It uses the sensor nodes of homogeneous type.

### 2.1.2 K Nearest Neighbors (KNN)

Outlier detection is the most essential part in any application where data are processed. The robustness, accuracy and the correctness of the deployed model enhances very steadily by preprocessing the data and fixing the outliers or the missing values. In this method, the data are exchanged with mean of K nearest nodes residing in the network.

*Branch et al. (2006, 2013)* have used an anomaly detection method which is based on the KNN algorithm. The KNN based model requires large space to store the information gathered from the nodes placed in the environment. The scheme does not predict missing data and has moderate complexity.

*Sheng et al. (2007)* developed a technique in order to discover outliers which considered a threshold value and distance between their k nearest neighbors. If it exceeds that threshold or the topmost distance to the nearest neighbors, it comes into outlier range. Every sensor node contains a histogram type summary and the sink node plays an important role by collecting and querying those summaries which are needed to determine the outliers correctly.

### 2.1.3 Self Organizing Map (SOM)

There are two types of attacks viz., internal and external attacks. The internal attacks originate within the network which means that source of the attack is from a particular network. External attacks are the attacks which arrive from the different sources. These attacks can be detected using SOM algorithm. It is not suitable in the large networks.

*Avram et al. (2007)* focused on detection of the attacks in networks using SOM. Weights of various sensors of the network are computed by statistical analysis of input data features. Determining input weights of sensor nodes is the major drawback of this proposed scheme. This scheme has moderate complexity.

*Puttini et al. (2016)* came up with an anomaly detection mechanism which takes into account a behavioral model. It uses multiple profiles for outlier detection. This method does not predict any missing data and has moderate complexity.

### 2.1.4 Support Vector Machine (SVM)

The main issues in outlier detection are type of attributes, size of data, dimensionality and high detection rate (Rath et al. 2019; Sharma et al. 2019). SVM is a method having low memory usage, less communication overheads and small computational complexity (Snoussi 2015; Shahid et al. 2012). SVM is used in various applications such as fault diagnosis, intrusion detection, medical imaging and predicting the protein structure etc.

*Kaplantzis et al. (2014)* had introduced binary (two-class) SVM classifiers which incorporates the theory of structural risk minimization and kernel-based methods. By separating the two different

Table 1. Comparative analysis of machine learning algorithms used for anomaly detection

| Machine Learning Algorithm used      | Author                    | Objective                            | Approach                             | Predicting missing data | Complexity of the algorithm        |
|--------------------------------------|---------------------------|--------------------------------------|--------------------------------------|-------------------------|------------------------------------|
| <b>BBN (Bayesian Belief Network)</b> | Kirk et al., 2014         | Outlier Detection                    | Conditional probability              | Yes                     | Moderate                           |
|                                      | Janakiram et al., 2006    | Outlier detection                    | Naive Bayes Assumption               | No                      | Moderate                           |
| <b>KNN (K Nearest Neighbors)</b>     | Branch et al., 2006, 2013 | Distributed outlier detection        | Nearest nodes using area             | Yes                     | Moderate, uses large memory        |
|                                      | Sheng et al., 2007        | Outlier detection and missing data   | Histogram data, length between nodes | Yes                     | High computations, high complexity |
| <b>SOM (Self Organizing Map)</b>     | Avram et al., 2007        | Attack detection                     | Statistical analysis                 | No                      | Moderate                           |
|                                      | Puttini et al., 2016      | Anomaly detection                    | Probability based                    | No                      | Moderate                           |
| <b>SVM (Support Vector Machine)</b>  | Kaplantzis et al., 2014   | Outlier detection                    | Two class SVM                        | No                      | It depends on implementation       |
|                                      | Zhang et al., 2016        | Distributed online outlier detection | Ellipsoidal SVM                      | No                      | Moderate complexity                |

classes of data in the feature space, binary SVM matches the highest margin hyper plane. This scheme does not predict the missing data.

Zhang et al. (2016) used relation among the sensor data and have proposed a distributed online outlier detection technique. His technique works on an Ellipsoidal SVM. It takes into account the spatial-temporal correlation for anomaly detection and updates the SVM model of sensor data for future outlier detection.

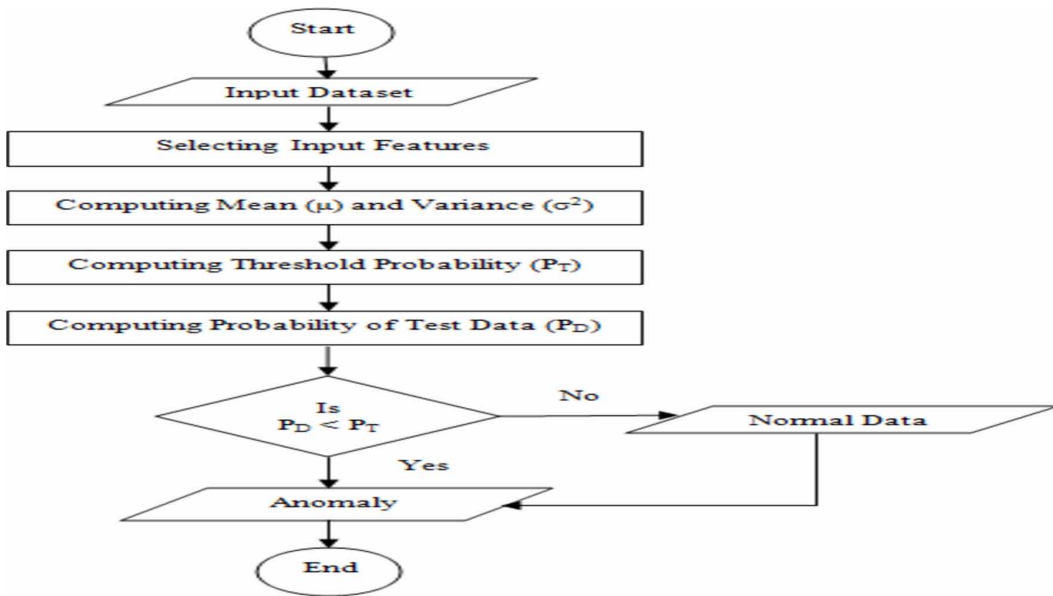
## 2.2 Comparison of Machine Learning Based Anomaly Detection Schemes

Existing machine learning based algorithms of anomaly detection can be compared on the basis of their objective, approach, behaviour and complexity. We have done this comparison and shown the summary in Table 1 which depicts that SOM and SVM based approaches are the latest approaches of anomaly detection. These schemes give outperforming results with moderate complexities. However, their efficiency can be further improved. This gives us motivation to design a new algorithm of anomaly detection using supervised machine learning with Gaussian distribution based scheme for outperforming results. We have used classification approach in our scheme.

## 3. PROPOSED WORK

The malfunctioning sensor nodes may send anomalous data. This anomalous data should not be used in the computation or in crucial decision making. They must be detected and fixed. This paper proposes a Gaussian distribution based machine learning scheme for anomaly detection (GDA). In this scheme, supervised machine learning is used for classification of the data. Here, previously labeled trained data is provided to the machine and the machine learns on basis of that data. After that, probability of the newly arrived data is computed. If that data is having probability less than the threshold probability, then it will be anomalous data. Otherwise the data will be non-anomalous.

Figure 4. Working of Gaussian distribution based approach for anomaly detection



Thus, it can be decided whether the newly coming data from a sensor node is anomalous or non-anomalous. Figure 4 presents working of the proposed Gaussian distribution based machine learning scheme for anomaly detection.

This scheme consists of the following three main algorithms. Algorithm 1 shows the procedure of anomaly detection. Input data is divided into training data and test data. First of all, system is trained with training data and then this learning of the system is applied on the test data to get the decisions on the data. Algorithm 2 presents data preprocessing method. Data preprocessing is necessary before starting the anomaly detection process. In the procedure of preprocessing, data is cleaned by removing noise, reducing dimensions, handling redundancies, filling missing values, normalizing data and selecting features. Then, this preprocessed data is passed to the system for predictive analysis. Algorithm 3 describes the proposed Gaussian distribution based anomaly detection scheme. In this scheme, labeled trained set of data is provided for classification of the elements into two classes anomalous and non-anomalous. The probability of the test data is calculated with help of Gaussian distribution. We compute variance ( $\sigma^2$ ) and mean ( $\mu$ ) of the features of test data set. Then we calculate the probability of the test case. If the probability is lesser than the threshold probability then data will be anomalous, otherwise it will be non-anomalous. Computation of the threshold probability should consider all input features of the healthcare data. It can be decided on basis of normal range values of the particular feature as well as out of range but possible values of that feature. On basis of the periodic data received by the sensors, machine can verify correctness of the sensed data.

Algorithm 1. Procedure of anomaly detection

Input: Healthcare Dataset

Output: Identified Anomaly

Begin

**Step 1:** Take input data

**Step 2:** Start preprocessing and select the features

**Step 3:** Train the training data using GDA

**Step 4:** Take test data and apply this learning for getting decision  
End

#### Algorithm 2. Data preprocessing

Input: Dataset  
Output: Preprocessed Data  
Begin  
**Step 1:** Take input data  
**Step 2:** Start preprocessing  
    i. Remove noise  
    ii. Reduce dimensions  
    iii. Handle redundancies  
    iv. Fill missing values  
    v. Normalize data  
    vi. Feature selection  
**Step 3:** Pass this data for prediction computation  
End

#### Algorithm 3. Gaussian distribution based learning scheme for anomaly detection

Input: Dataset  
Output: Identified Anomaly  
Begin  
**Step 1:** Choose the features that are most likely to be helpful for detecting the anomalies.  
**Step 2:** Find the means ( $\mu_1, \mu_2, \dots, \mu_m$ ) of the features in the test data set.  
**Step 3:** Find the variance ( $\sigma^2_1, \sigma^2_2, \dots, \sigma^2_m$ ) of the features in the test data set.  
**Step 4:** Compute the Threshold probability  
**Step 5:** Find the probability of the test data  $D(d_1, d_2, \dots, d_m)$   
$$P(D) = p(d_1; \mu_1; \sigma^2_1) * p(d_2; \mu_2; \sigma^2_2) * \dots * p(d_m; \mu_m; \sigma^2_m)$$
  
**Step 6:** **If** ( $P(D) < \text{Threshold probability}$ )  
    It is Anomalous Data.  
    **Else**  
    It is Non Anomalous Data.  
End

## 4. PERFORMANCE EVALUATION

Proposed algorithm is implemented in Python on x86\_64 architecture based Intel core i7 processor with Windows 10 platform. The proposed anomaly detection algorithm is compared with the schemes of SOM and SVM on two different datasets. We have taken three features in dataset1 and eight features in dataset2.



**Table 2. Parameters and their values**

| Parameter                               | Value                                                 |
|-----------------------------------------|-------------------------------------------------------|
| Number of datasets used                 | 2                                                     |
| Number of input data points in dataset1 | 500-2500                                              |
| Number of input data points in dataset2 | 1000-5000                                             |
| Features taken in dataset1              | 3 (BP, Sugar, Body Temperature)                       |
| Features taken in dataset2              | 8 (Age, BP, Sugar, Urea, RBC, WBC, SpO2, Haemoglobin) |
| Approach used                           | Classification                                        |

### 4.1 Experimental Setup

The anomaly detection algorithms are tested on various input data points of two different datasets. Classification approach is used for anomaly detection. Parameters used during implementation of these algorithms are shown in Table 2.

### 4.2 Performance Metrics

We are using following three metrics for performance analysis of the anomaly detection schemes. Here, N is total number of input data points, n is number of top potential anomalies identified by the detection method and A is true or actual anomalies.

(i) Precision (P)

Precision is an evaluation measurement which is defined as the proportion of the true anomalies to the top potential anomalies detected by the system. It is calculated using A and n as shown in eq. (1):

$$P = (A/n) * 100 \tag{1}$$

(ii) Throughput (T)

Throughput measures the amount of actual data points which should be passed through any system. It is calculated using N and n as shown in eq. (2):

$$T = (N-n) \tag{2}$$

(iii) Efficiency (E)

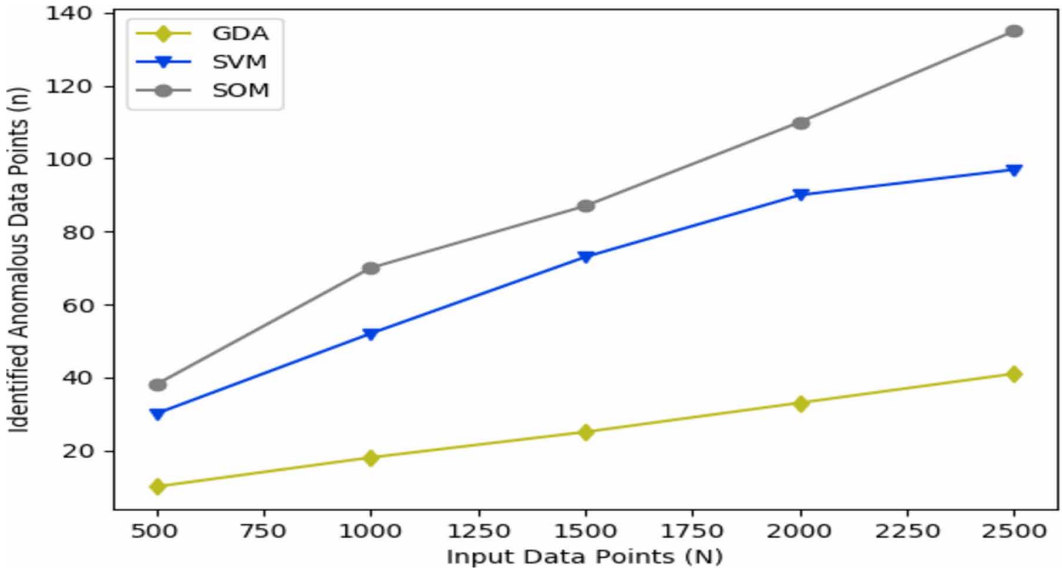
Efficiency is a metric which tells how efficient the system is. It reflects the performance and complexity of the system. If the efficiency of any system is good then its performance will also be good and complexity will be less. Efficiency is calculated with help of T and N as shown in eq. (3):

$$E = (T/N)*100 \tag{3}$$

### 4.3 Results and Analysis

Proposed and existing algorithms are executed on two datasets: dataset1 with 500 to 2500 data points and dataset2 with 1000 to 5000 data points. These datasets have different number of input data points, features and anomalies. Therefore, we get different results which are discussed and analyzed in this section. These algorithms are compared on basis of various performance metrics viz., precision, throughput and efficiency which depicts that proposed Gaussian distribution approach (GDA)

Figure 5. Number of detected anomalies over input data of dataset1



outperforms the approaches of self organizing map (SOM) and support vector machine (SVM). Now, we are going to discuss various results which are obtained on varying the input data points.

#### 4.3.1 Detected Anomalies Over Various Input Data Points

When we vary the input data points, we get variation in detected anomolous data points. Actually, when we increase size of the data set, we identify more number of outliers. Figure 5 and figure 6 show the results on dataset1 and dataset2 respectively. Now, it must be ensured that the detected anomalies and actual anomalies or similar or not. Gaussian distribution fits many natural phenomena and it gives the best model approximation. Therefore, Gaussian distribution based approach detects almost genuine data points as anomalies while SOM and SVM reports few more data points as anomalous data which in turns affects its precision, throughput and efficiency. Actual anomalies present in the input datasets and the identified anomalies by these detection schemes can be seen in table 3 and table 4.

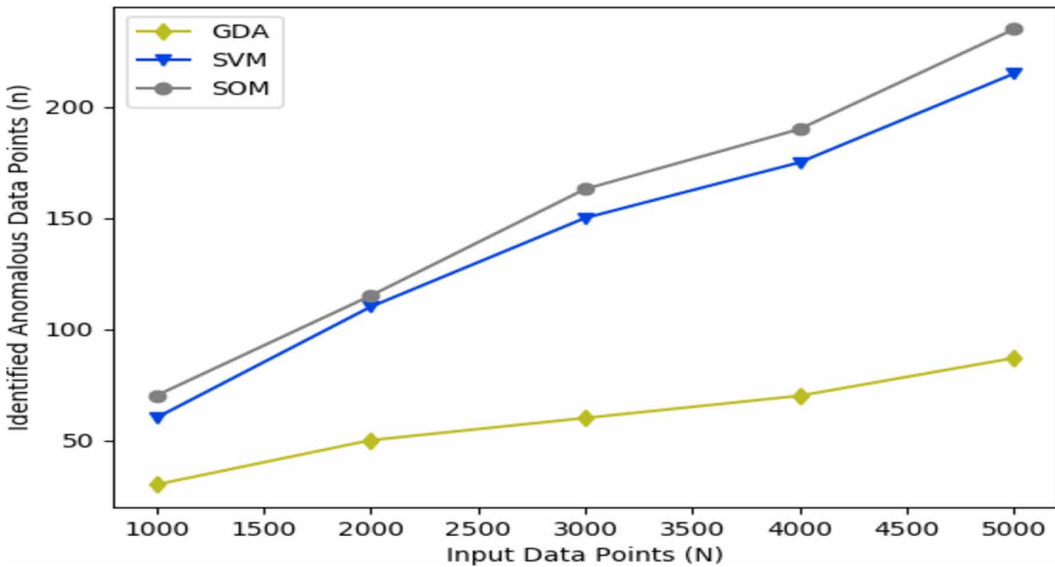
#### 4.3.2 Precision Over Various Input Data Points

When input data points are increased, we get increased precision values. Figure 7 and 8 are showing the comparison of precision values of GDA, SVM and SOM on dataset1 and dataset2 respectively. Here, it is observed that GDA provides the best precision among these anomaly detection approaches. In the input data set, some points tends to be anomolous points but are not true anomalies. SOM and SVM identifies some of them as anomalies. In this way, SOM and SVM detects few more points as anomalies which are not true or actual anomalies. GDA identifies almost all genuine anomalous data points. Therefore, GDA shows higher precision than SOM and SVM approaches.

#### 4.3.3 Throughput Over Various Input Data Points

Throughput refers to the actual data points that should be passed through the system. Throughput increases as we increase the number of data points in all the schemes for both datasets. Throughputs of Gaussian distribution, self organizing map and support vector machine based approaches are compared in figure 9 and figure 10 for dataset1 and dataset2 respectively. It is found that throughput of Gaussian distribution based method is higher than that of SOM and SVM based approaches. Higher

Figure 6. Number of detected anomalies over input data of dataset2



throughput of Gaussian distribution based approach over SOM and SVM schemes is due to its higher precision value for anomaly detection.

#### 4.3.4 Efficiency Over Various Input Data Points

Performance of any system is decided by its efficiency. Efficiency of the existing and proposed schemes is tested on both datasets. It is found that efficiency of the proposed and existing algorithms increases with the increase in number of input data points for both datasets. Figure 11 and figure 12 compare the efficiency of proposed Gaussian distribution based method with the existing SOM and

Figure 7. Precision over input data of dataset1

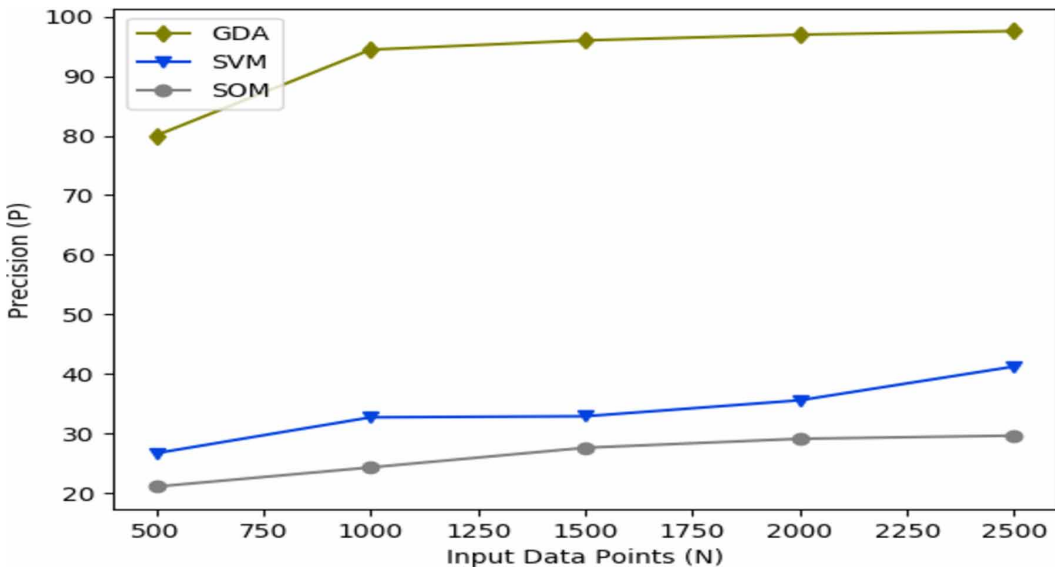
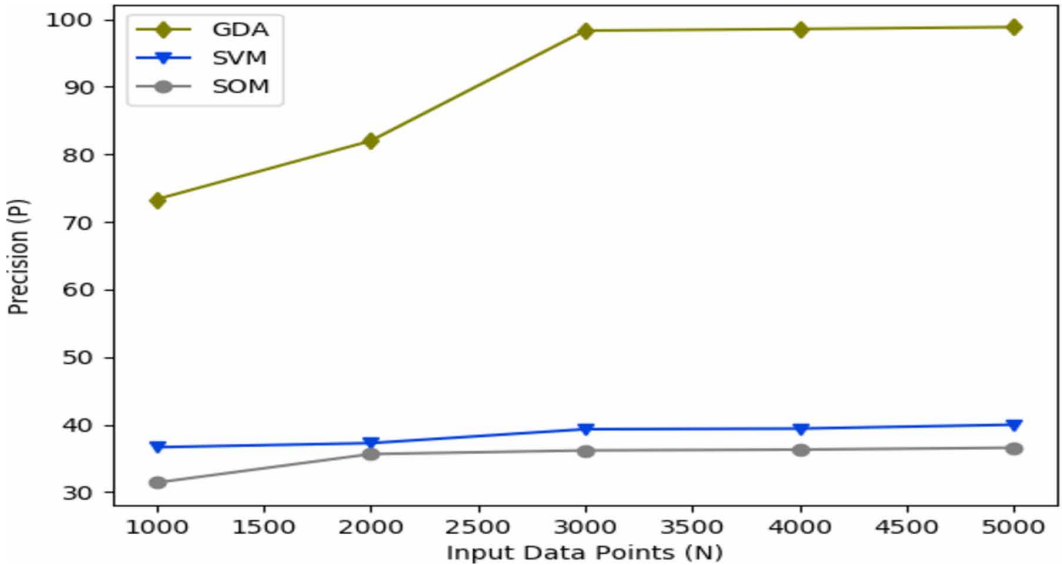


Figure 8. Precision over input data of dataset2



SVM based schemes on dataset1 and dataset2 respectively. It depicts how efficient these approaches are. Gaussian distribution based method is found more efficient because it detects almost genuine outliers or anomalous data with a high precision rate.

#### 4.4 Execution Time

Till now, we have compared our methodology with the existing schemes on basis of some performance metrics. However, it is interesting to observe the execution time of these approaches. Figure 13 and figure 14 present the execution time of these anomaly detection schemes for dataset1 and dataset2

Figure 9. Throughput over input data of dataset1

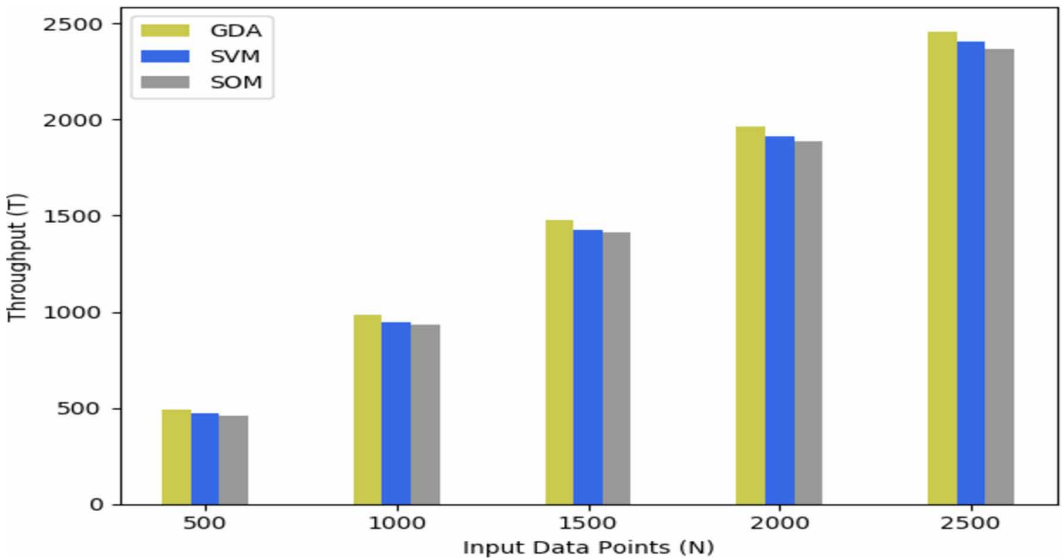
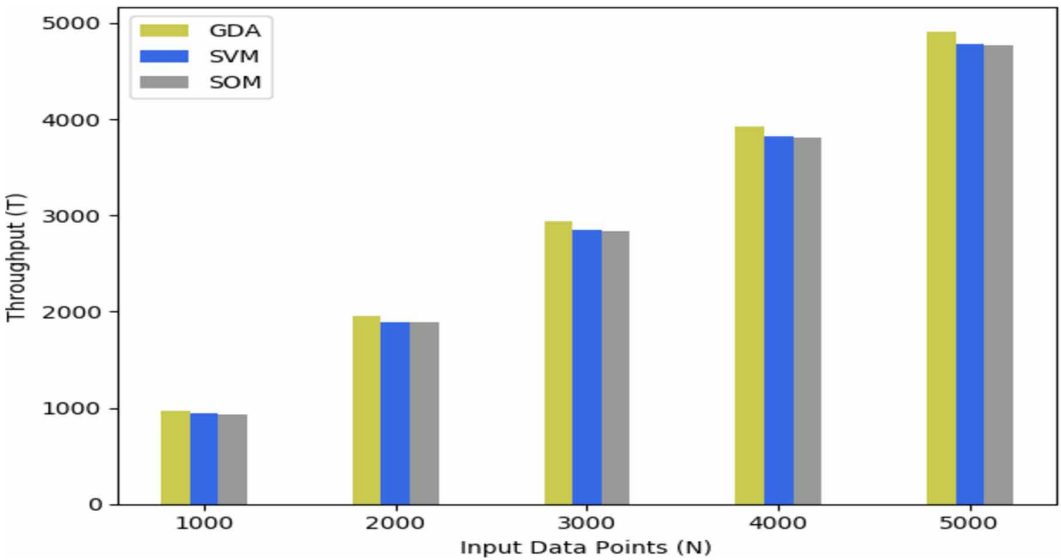


Figure 10. Throughput over input data of dataset2



respectively. We can see that GDA has the smallest execution time than the other existing schemes. Thus, GDA outperforms SOM and SVM approaches for identifying the anomalies.

#### 4.5 Result Validation

The code is tested on various input data points of two different datasets having different features. For 500 input data points of dataset1, the various metrics of proposed (GDA) and existing approaches (SOM and SVM) are computed mathematically as follows:

- (i) SOM Approach

Figure 11. Efficiency over input data of dataset1

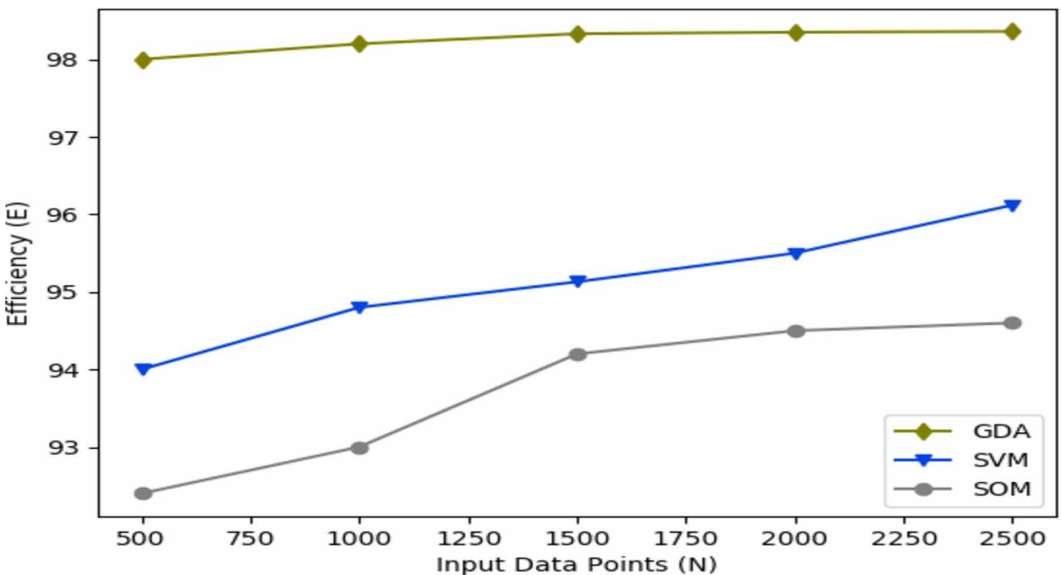


Figure 12. Efficiency over input data of dataset2

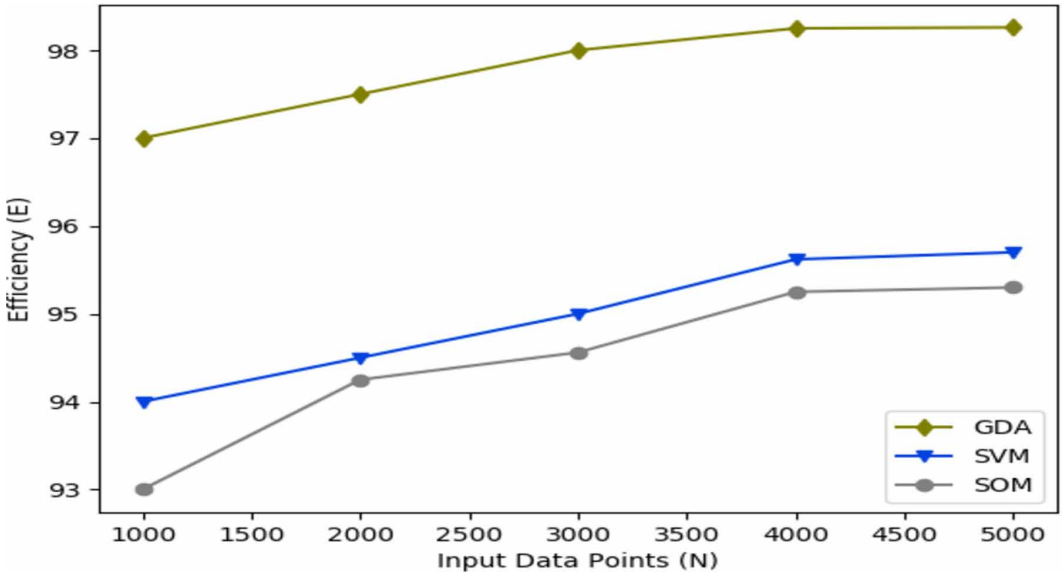
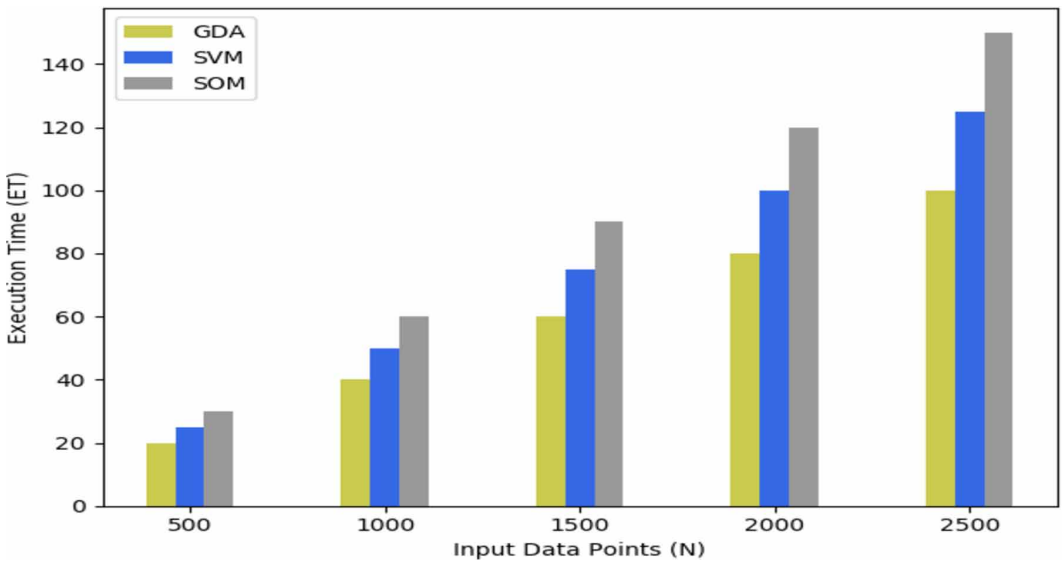


Figure 13. Execution time over input data of dataset1



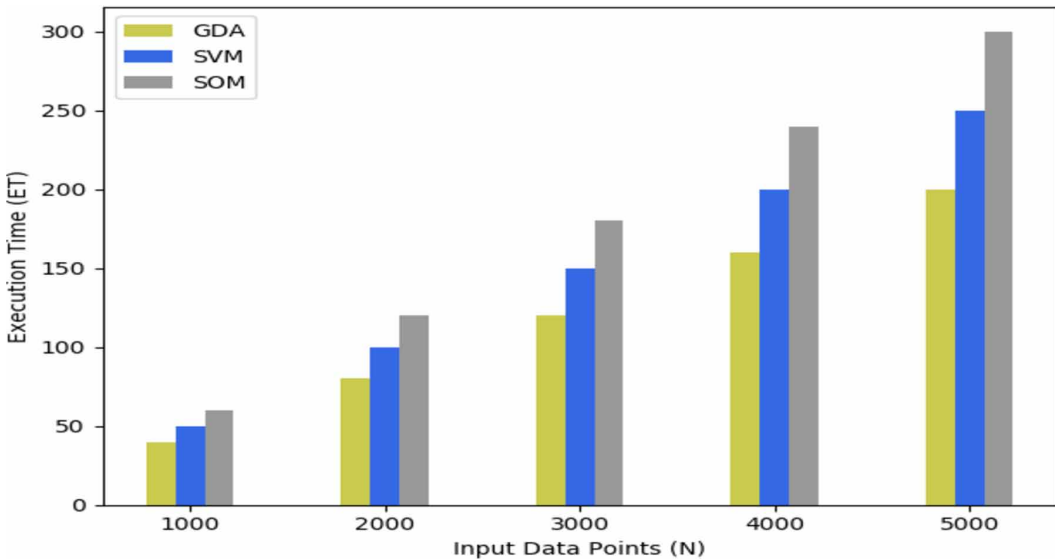
For 500 data points (N=500) of dataset1, SOM identifies 38 anomalous data values (n =38) while actual anomalies are 8 (A=8). Now, we calculate Precision (P), Throughput (T) and Efficiency (E) as below:

$$P = (A/n) * 100$$

$$= 21.05\%$$

$$T = (N-n)$$

Figure 14. Execution time over input data of dataset2



$$= 462$$

$$E = (T/N) * 100$$

$$= 92.40\%$$

(ii) SVM Approach

For 500 input data values (N=500) of dataset1, SVM identifies 30 anomalous data values (n=30) while true anomalies are 8 (A=8). Now, we calculate Precision (P), Throughput (T) and Efficiency (E) as follows:

$$P = (A/n) * 100$$

$$= 26.66\%$$

$$T = (N-n)$$

$$= 470$$

$$E = (T/N) * 100$$

$$= 94.00\%$$

(iii) GDA Approach

Proposed approach GDA detects 10 anomalies (n=10) while actual anomalies are 8 (A=8) for 500 input values (N=500) of dataset1. Now, we calculate Precision (P), Throughput (T) and Efficiency (E) as below:

$$P = (A/n) * 100$$

$$= 80.00\%$$

$$T = (N-n)$$

$$= 490$$

$$E = (T/N) * 100$$

$$= 98.00\%$$

Table 3. Comparison of proposed scheme with existing schemes on dataset1

| Total input data points (N) | True anomalies (A) | Total detected anomalies (n) |     |           | Precision (P in %) |       |              | Throughput (T) |      |             | Efficiency (E in %) |       |              |
|-----------------------------|--------------------|------------------------------|-----|-----------|--------------------|-------|--------------|----------------|------|-------------|---------------------|-------|--------------|
|                             |                    | SOM                          | SVM | GDA       | SOM                | SVM   | GDA          | SOM            | SVM  | GDA         | SOM                 | SVM   | GDA          |
| 500                         | 8                  | 38                           | 30  | <b>10</b> | 21.05              | 26.66 | <b>80.00</b> | 462            | 470  | <b>490</b>  | 92.40               | 94.00 | <b>98.00</b> |
| 1000                        | 17                 | 70                           | 52  | <b>18</b> | 24.28              | 32.69 | <b>94.44</b> | 930            | 948  | <b>982</b>  | 93.00               | 94.80 | <b>98.20</b> |
| 1500                        | 24                 | 87                           | 73  | <b>25</b> | 27.58              | 32.87 | <b>96.00</b> | 1413           | 1427 | <b>1475</b> | 94.20               | 95.13 | <b>98.33</b> |
| 2000                        | 32                 | 110                          | 90  | <b>33</b> | 29.09              | 35.55 | <b>96.96</b> | 1890           | 1910 | <b>1967</b> | 94.50               | 95.50 | <b>98.35</b> |
| 2500                        | 40                 | 135                          | 97  | <b>41</b> | 29.62              | 41.23 | <b>97.56</b> | 2365           | 2403 | <b>2459</b> | 94.60               | 96.12 | <b>98.36</b> |

Here, we can see that these mathematical results are same as the experimental results. Similarly, these metrics (P, T and E) can be computed for other input data points of all the datasets and the experimental results can be validated with these mathematical results.

#### 4.6 Discussion

We have observed different results because of variation in the number of input data points, anomalous points and input features. Table 3 presents comparison of the proposed Gaussian distribution based approach (GDA) with the existing schemes of SOM and SVM on dataset1. Similarly, Table 4 describes this comparison on dataset2. We can see in these tables that precision in the anomaly detection, throughput of the system and efficiency of the scheme are significantly improved in the proposed method as compared to the existing schemes. Detected anomalies by Gaussian distribution based method are very near to the true or actual anomalies. But, in case of SVM and SOM, detected anomalies and true anomalies are not very close. In this way, precision of GDA becomes much better than that of SVM and SOM which also causes better throughput and efficiency in GDA. It can be noted that average efficiency for both datasets of the proposed scheme GDA is 98% whereas it is 95% and 94% in case of existing schemes SVM and SOM respectively. Thus, average improvement in efficiency of GDA is 3% and 4% as compared to SVM and SOM respectively.

#### 4.7 Salient Features of Proposed Scheme

Proposed approach detects anomalies with lesser computational complexities than the existing approaches of SOM and SVM. Various characteristics of the proposed scheme are discussed as follows:

1. *Integrity*: Proposed method ensures integrity of the system. Proposed approach does not create any loss or modification in the sensed data during training or testing procedure.

Table 4. Comparison of proposed scheme with existing schemes on dataset2

| Total input data points (N) | True anomalies (A) | Total detected anomalies (n) |     |           | Precision (P in %) |       |              | Throughput (T) |      |             | Efficiency (E in %) |       |              |
|-----------------------------|--------------------|------------------------------|-----|-----------|--------------------|-------|--------------|----------------|------|-------------|---------------------|-------|--------------|
|                             |                    | SOM                          | SVM | GDA       | SOM                | SVM   | GDA          | SOM            | SVM  | GDA         | SOM                 | SVM   | GDA          |
| 1000                        | 22                 | 70                           | 60  | <b>30</b> | 31.42              | 36.66 | <b>73.33</b> | 930            | 940  | <b>970</b>  | 93.00               | 94.00 | <b>97.00</b> |
| 2000                        | 41                 | 115                          | 110 | <b>50</b> | 35.65              | 37.27 | <b>82.00</b> | 1885           | 1890 | <b>1950</b> | 94.25               | 94.50 | <b>97.50</b> |
| 3000                        | 59                 | 163                          | 150 | <b>60</b> | 36.19              | 39.33 | <b>98.33</b> | 2837           | 2850 | <b>2940</b> | 94.56               | 95.00 | <b>98.00</b> |
| 4000                        | 69                 | 190                          | 175 | <b>70</b> | 36.31              | 39.42 | <b>98.57</b> | 3810           | 3825 | <b>3930</b> | 95.25               | 95.62 | <b>98.25</b> |
| 5000                        | 86                 | 235                          | 215 | <b>87</b> | 36.59              | 40.00 | <b>98.85</b> | 4765           | 4785 | <b>4913</b> | 95.30               | 95.70 | <b>98.26</b> |



2. *Scalability*: Proposed model confirms scalability as per the requirements. It is convenient with the change in size of the datasets. It means that proposed work performs well even if size of the dataset is changed.
3. *Precision*: Proposed scheme gives almost accurate results. It detects outliers which are very close to the true or actual outliers while SOM and SVM based schemes reports few more false outliers.
4. *Efficiency*: Efficiency describes the performance of any scheme. If any system is efficient then its performance will be good. Efficiency of the proposed approach is 98% which is better than SOM and SVM based schemes.

## 5. CONCLUSION AND FUTURE DIRECTIONS

Sensors are used for data collection from various types of environments and this data can be kept over cloud to facilitate end users in many ways. Healthcare monitoring sensor cloud is an integration of various body sensors of different patients with cloud. There is always a possibility of anomalies in data due to some malicious activities or malfunctioning sensor nodes. Sometimes, the nodes may behave abnormally and they might send wrong data. The collected data are very crucial and used for various types of analysis and decision making. So, this data must be precise and accurate. Proposed anomaly detection scheme (GDA) uses supervised machine learning approach in which Gaussian statistical model is used for detection of anomalous behavior of sensor nodes. Some labeled training data are provided for the computation of threshold probability that is used further for computing the behavior of newly coming data from sensor nodes. The work is compared with the other schemes of supervised machine learning viz., self organizing map (SOM) and support vector machine (SVM) on performance metrics namely precision, throughput and efficiency by varying the input data points. These schemes are tested on two different datasets. We observed that true (actual) and detected (identified by scheme) anomalies are very close in the proposed scheme GDA as compared to the existing schemes SVM and SOM. Thus, significant improvement in precision of GDA has been observed than that of SVM and SOM, resulting in better throughput and efficiency of the proposed algorithm. Analytical validation has also been carried out to justify the experimental results. It can be observed that average efficiency of the proposed technique is 98% which shows an improvement of 3% and 4% in average efficiency as compared to the existing techniques of SVM and SOM respectively.

In future, other supervised, unsupervised and semi supervised machine learning algorithms can be applied on different datasets to find out the best approach for anomaly detection. These techniques could be tested on datasets of the various applications of wireless sensor networks and Internet of Things (IoT) such as underwater monitoring system, underground applications, agricultural IoT, smart farming system, smart healthcare information system, forest fire detection system and military applications. This paper has focused on identifying the outliers or anomalies. Missing data handling may be part of the future research. We can also work on some highly complex healthcare data in future. We have used Gaussian distribution based approach in this paper. In future, the work can be extended for the data with non-normal distribution. Here, we have used classification approach of machine learning. In future, clustering approach might be tried for detecting the anomalies.

## ACKNOWLEDGMENT

This research is partially funded by 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.
- 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
- Alsheikh, M. A., Lin, S., Niyato, D., & Tan, H.-P. (2014). Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. *IEEE Communications Surveys and Tutorials*, 16(4), 1996–2018. doi:10.1109/COMST.2014.2320099
- Avram, T., Oh, S., & Hariri, S. (2007). Analyzing Attacks in Wireless Ad Hoc Network with Self-Organizing Maps. *5th Annual Conference on Communication Networks and Services Research (CNSR '07)*, 166-175 doi:10.1109/CNSR.2007.15
- Ayadi, A., Ghorbel, O., Bensaleh, M. S., Obeid, A., & Abid, M. (2017). Performance of outlier detection techniques based classification in Wireless Sensor Networks. *13th IEEE International Wireless Communications and Mobile Computing Conference (IWCMC)*, 687-692. doi:10.1109/IWCMC.2017.7986368
- Ayadi, H., Zouinkhi, A., Boussaid, B., & Abdelkrim, M. N. (2015). Machine Learning Methods: Outlier detection in WSN. *16<sup>th</sup> IEEE international conference on Sciences and Techniques of Automatic control & computer engineering – STA'2015*, 722-727. doi:10.1109/STA.2015.7505190
- 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. In *Information Fusion 33* (pp. 41–56). Elsevier.
- Branch, J. W., Giannella, C., Szymanski, B., Wolff, R., & Kargupta, H. (2006). In-network outlier detection in wireless sensor networks. *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, 51-51. doi:10.1109/ICDCS.2006.49
- Branch, J. W., Giannella, C., Szymanski, B., Wolff, R., & Kargupta, H. (2013). In-network outlier detection in wireless sensor networks. *Knowledge and Information Systems*, 34(1), 23–54. doi:10.1007/s10115-011-0474-5
- Devadevan, V., & Sankaranarayanan, S. (2017). *Forest Fire Information System Using Wireless Sensor Network*. *International Journal of Agricultural and Environmental Information Systems*, 8(3).
- 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.
- 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. doi:10.1109/CONFLUENCE.2019.8776897
- 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
- Fawzy, A., Mokhtar, H. M. O., & Hegazy, O. (2013). Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal*, 157–164.
- Forster, A., Venyagamoorthi, G. K., & Kulkarni, R. V. (2011). Computational Intelligence in Wireless Sensor Networks: A Survey. *IEEE Communications Surveys and Tutorials*, 13(1), 68–96. doi:10.1109/SURV.2011.040310.00002
- 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. In *Applied Soft Computing* 42 (pp. 204–214). Elsevier. doi:10.1016/j.asoc.2015.12.042
- Janakiram, D., Reddy, V. A. M., & Kumar, P. (2006). Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks. IEEE conference on communication system software and middleware, 1-6. doi:10.1109/COMSWA.2006.1665221
- 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. doi:10.4018/978-1-5225-7432-3.ch018
- 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
- Kirk, A., Legg, J., & El-Mahassni, E. (2014). Anomaly detection and Attribution using Bayesian networks. *DSTO-TR*, 2975, 1–22.
- Lounis, A., Hadjadj, A., Bouabdallah, A., & Challal, Y. (2016). Healing on the cloud: Secure cloud architecture for medical wireless sensor networks. *Future Generation Computer Systems, Elsevier*, 55, 266–277. doi:10.1016/j.future.2015.01.009
- Machine Learning Repository, U. C. I. Data Sets. (n.d.). <https://archive.ics.uci.edu/ml/datasets.php>
- 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. doi:10.1109/IECON.2015.7392180
- 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
- Pachauri, G., & Sharma, S. (2015). Anomaly detection in medical wireless sensor networks using machine learning algorithms. *Procedia Computer Science*, 70, 325 – 333. doi:10.1016/j.procs.2015.10.026
- 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. In *Internet of Things* (Vol. 3, pp. 156–174). Elsevier.
- Puttini, R., Hanashiro, M., Miziara, F., de Sousa, R., García-Villalba, L. J., & Barenco, C. J. (2016). On the Anomaly Intrusion Detection in Mobile Ad Hoc Network Environments. In *Lecture Notes in Computer Science: Vol. 4217. Personal Wireless Communications* (pp. 182–193). Springer. doi:10.1007/11872153\_16
- 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
- Sheng, B., Li, Q., Mao, W., & Jin, W. (2007). Outlier Detection in Sensor Networks. *Proceedings of the 8th ACM International Symposium on Mobile and Ad Hoc Networking and Computing (MobiHoc)*, 219-228.

- 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
- 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. doi:10.1016/j.procs.2013.06.168
- 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

*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=7, i-10 index=1, citations=115). 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 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=9, i-10 index=8, citations=287). 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 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=132, i-10 index=524, g-index=280, 92544+ 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](http://www.buyya.com).*