IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 20, 2020, accepted December 3, 2020, date of publication December 9, 2020,
date of current version December 31, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3043421*

# Large-Scale Outlier Detection for Low-Cost $PM_{10}$ Sensors

**YUANYUAN WEI[1], JULIAN JANG-JACCARD[1], FARIZA SABRINA[2], (Member, IEEE), AND HOOMAN ALAVIZADEH[1], (Member, IEEE)**

[1]Cyber Security Laboratory, School of Natural and Computational Sciences, Massey University, Auckland 0632, New Zealand
[2]School of Engineering and Technology, Central Queensland University, Sydney, NSW 2000, Australia

Corresponding author: Yuanyuan Wei (y.wei1@massey.ac.nz)

This work was supported by the Ministry of Business, Innovation, and Employment (MBIE) of New Zealand as parts of Catalyst Strategy Funds under Grant UOCX1720 and Grant MAUX1912.

**ABSTRACT** Evaluating the air quality of classrooms is important as children spend a large amount of time at school. Massey University (NZ) led the development of a low-cost and affordable Indoor Air Quality (IAQ) platform called SKOMOBO that was deployed on a large scale across the classrooms of primary schools in New Zealand. When the data from SKOMBO units were collected, it was important to detect any unexpected high air pollution events. To address this concern, we propose a study of outlier detection for $PM_{10}$ dataset from SKOMOBO units using MSD-Kmeans. MSD-Kmeans combines the statistical method of Mean and Standard Deviation (MSD) with the machine learning clustering algorithm K-means where the former eliminates as many noisy data to minimize the inference on clustering while the latter is able to achieve better local optimal clustering. We compare the performance of MSD-Kmeans with other similar outlier detection algorithms. Our experimental results illustrate that MSD-Kmeans outperforms the majority of performance indicators (*e.g.,* TPR, FPR, Accuracy, F-measures) compared to other similar methods. We conclude that it is feasible to use MSD-Kmeans as an effective outlier detection tool on large scale datasets.

**INDEX TERMS** Environment monitoring, indoor air quality (IAQ), outlier detection, machine learning, sensors, statistical analysis.

## I. INTRODUCTION

Children spend the second-largest proportion of their time at school which illustrates that providing a good quality environment where children can breathe in the fresh air in their classrooms is essential for their health and well-being. A low ventilation rate can lead to a build-up of pollutants, moisture, mold, and bacteria that could impact children's health and performance. International studies have demonstrated links between low ventilation rates and lower children's performance and attention [1], and similarly between low ventilation rates and higher absenteeism [2].

Investigating the indoor air quality (IAQ) in large numbers of classrooms, unfortunately, can be very costly as standard IAQ monitoring units are too expensive. It is not achievable to use such expensive tools for many budget strapped public schools. To address such concern, a team of researchers at Massey University developed a low-cost monitoring suite called SKOol MOnitoring BOx (SKOMOBO) with the financial support of the Building Research Levy and in technical collaboration with the National Institute of Water and Atmospheric Research (NIWA). A small box, approximately the size of $100 \times 100 \times 100$ *mm*, was designed to house a number of low-cost sensors which could capture particulate matter ($PM_{2.5}$ and $PM_{10}$), temperature, relative humidity, carbon dioxide ($CO_2$) and human occupancy in classrooms. Since September 2017, 165 units of SKOMOBO have been deployed in the 11 schools on 45 classrooms across the North and South islands. The schools have been strategically chosen to cover the broad representation of the climatic variation in New Zealand and also to include both provincial towns and large cities.

One of the major challenges in such large-scale deployment is to ensure the stability of the sensor detection by detecting data outliers, such as reflecting in unexpected high air pollution events. For example, walking close to the sensor box or vacuuming the carpet may increase the contamination concentration of particles in the air which would cause

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

outliers in the PM sensor. Data outliers, which are occasional sensor readings that do not follow the normal patterns of the majority of sensor data, can incur due to various factors that include unexpected events (*e.g.,* software bug, data transmission error, *etc*), sensor fault or environmental interference [3]. Because the operation of SKOMOBO relies on sensor data (*i.e.,* IAQ parameters), such data outliers could significantly affect the interpretation of data or even mislead them into undesirable states [4].

In this paper, we are especially concerned with detecting outliers for $PM_{10}$ collected by SKOMOBO units. The reports [5], [6] from Ministry for the Environment and Stats New Zealand (NZ) has indicated that there has been growing concern on $PM_{10}$ rates especially in winter with the emission of heater etc. [7] and children being more sensitive to air contamination due to their underdeveloped immune system [8].

Towards this goal, we use the hybrid MSD-Kmeans [9] for $PM_{10}$ outlier detection. The hybrid MSD-Kmeans takes the advantages of two approaches, Mean and Standard Deviation (MSD) and K-means, respectively. Using the insight of MSD which provides a more accurate idea of data distribution, extreme values (e.g., noisy samples) deviate from the standard mean is eliminated. Here, MSD is used as a data normalization mechanism in our approach. The normalized data is then fed into the K-means clustering algorithm to cluster similar data points and eliminate outlier that is more closely aligned with normal data points thus difficult to identify. This combination is proven to be more effective in detecting outliers where the proximities of the outliners are very close to the normal values.

We evaluate the MSD-Kmeans and demonstrate its effectiveness in the context of detecting possible anomalies from a real-life application such as SKOMOBO units. The main contributions of this paper are as follows.

- We propose the use of a hybrid MSD-Kmeans outlier detection algorithm that combines the statistical method of Mean and Standard Deviation MSD) to detect global outliers and the machine learning method of K-means to detect local outliers.
- We apply the MSD-Kmeans on $PM_{10}$ Data dataset from the SKOMOBO units to identify outliers from the real-life application and illustrates that MSD-Kmeans is an efficient outlier detection algorithm that provides high accuracy in finding abnormal clusters of data.
- We compare the performance of MSD-Kmeans with other outlier detection algorithms and demonstrate that MSD-Kmeans outperforms in terms of many accuracy measures.

The rest of this paper is structured as follows. Section II introduces related works in the field of environmental monitoring and state-of-the-art anomaly techniques in the area. Section III provides the background material for the design, development, and deployment of SKOMOBO units. Section IV discusses the brief details of MSD-KMeans. Section V illustrates the experimental setup and analysis of results for applying the proposed MSD-Kmeans as an outlier algorithm on $PM_{10}$ date from SKOMOBO. Section VI concludes the paper and introduces some future work.

## II. RELATED WORK

Outlier Detection methods can be categorized into two. Statistical methods were developed first as a way to measure how each individual piece of data deviates from the statistical norm or average values. The effectiveness largely depended on the model design and process of data analysis [10]–[12]. Machine Learning algorithms were later developed to assist in data analysis and became a popular technique for detecting outliers. Two most popular techniques used for outlier detection using Machine Learning algorithms include the cluster-based K-means [13] and the density-based Local Outlier Factor (LOF) [14].

K-means has become a popular method for outlier detection because many found it easy to implement [15]. However, K-means can be sensitive to noisy data when used to detect outlier [16]. A few studies have proposed several improvements in K-means for outlier detection. In [17], the Network Data Mining (NDM) method was used to extract the features from the packet and data flow captured from a network then performed a clustering task based on a distance-based K-means algorithm. In [17], the authors proposed a method which processes both classification and outlier detection concurrently to improve its scalability as a tool to use in real-time detection. However, their method required additional work manually having to determine the optimum number of clusters. In [18], the authors proposed a method that utilizes K-means with an optimization technique. Their method was applied to the air quality time-series datasets by the use of an optimization technique that involves weight-based center and standard deviation approaches to enhance the efficiency of the proposed K-means algorithm.

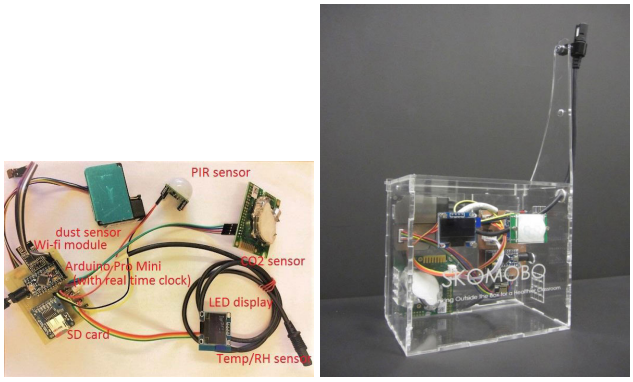
Often, K-means has been combined with other methods for better detection outcomes. In [19], the Density Based Improved K-means Clustering (Dbkmeans) algorithm was proposed to combine K-means and Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to get the advantage of both algorithms. Although [19] could better handle clusters of circularly distributed data points and slightly overlapped clusters, the study used synthetically created data. The outcome of this hybrid methodology showed a higher precision in outlier detection. In [20], it was proposed to improve K-means by applying density-based detection methods and adding the discovery and processing steps of the noise data to the original algorithm. The extra pre-processing step in [20] to exclude the interference of outliers was proven to be more time-consuming when applied to larger datasets which were limiting the scalability and applicability of this algorithm. In [21], a hybrid algorithm named the Gravitational Search Algorithm and K-means (GSA-KM) was designed to combine GSA and K-means for better clustering, but it required a minimum number of function evaluations to reach the optimal solution. In [22], the Triangle Area-based Nearest Neighbours (TANN) method was proposed to use

K-means to acquire centroids of each cluster, before using triangle area from each cluster centroid to get new datasets and applying the K-NN classification method to classify attacks. Although the implementation achieved higher accuracy and detection rates and lower false negative rates, the study did not discuss whether K-means was the optimal clustering technique for TANN.

While the previous work made practical applications of outlier detection, they have not proven their scalability on larger datasets based on real-life scenarios or demonstrated resistance to noisy data.

## III. SKOol MOnitering BOx (SKOMOBO)

With the support of the Building Research Levy and in collaboration with the National Institute of Water and Atmospheric Research (NIWA), a team of researchers at Massey University has developed a low-cost monitoring suite called SKOol MOnitoring BOx (SKOMOBO). The SKOMOBO units have been designed in such a way to monitor the main IAQ parameters such as particular matters (but also classroom temperature, relative humidity, and carbon dioxide) at a fraction of the cost when compared to the commercial counterparts. For example, a professional grade of a Particular Matter (PM) monitor cost around NZ$10,000 as of December 2017 while our SKOMOBO unit cost only around NZ$250 [23], [24]. This has allowed us to investigate the indoor climate of New Zealand classrooms on a large scale to ensure whether our children are learning in healthy environments. Fig. 1 (a) shows the sensors and other auxiliaries connected to SKOMOBO motherboard while Fig. 1 (b) demonstrates a complete SKOMOBO unit with an enclosure.




(a) SKOMOBO with sensors and other auxiliaries (b) A SKOMOBO unit with enclosure

**FIGURE 1.** SKOol MOnitoring BOx.

### A. CIRCUIT BOARD AND MICROCONTROLLER

We developed our own circuit board, evolved from the PCB of the NIWA PACMAN [25], to mount different IAQ sensors which allowed us flexibility and customizability. Our own circuit board gave us the opportunity to test different sensors and compare the price competitiveness without sacrificing quality. It also allowed us to check the compatibility among different sensors and to choose the combination that worked best. In the center of the circuit board, we had an Arduino Pro Mini as the microcontroller module. Our PCB also had the locations for four sensors which include a temperature and relative humidity sensor, a $CO_2$ sensor, a PM sensor, and a motion sensor. These main sensors were supported by other modules such as a real-time clock module, an SD card as primary storage, and a Wi-Fi interface. Fig. 2 illustrates the printed circuit board.

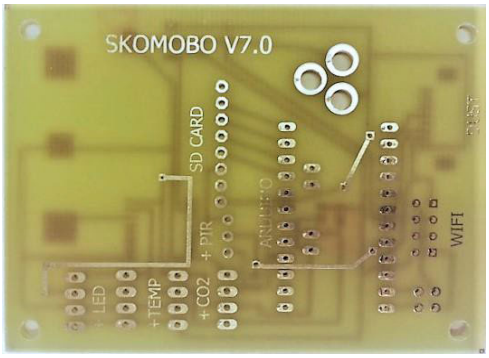



**FIGURE 2.** Blueprint for SKOMOBO circuit board.

### B. PARTICULAR MATTER (PM) SENSOR

A PM measurement sensor, also informally know as a dust sensor, was used to collect the level of dust densities in the classroom. The dust sensor was expected to show the quality of air the students breathe in. We tested three low cost dust sensors: Sharp model GP2Y1010AU0F, Shinyei model PPD42NS, and Samyoung model DSM501A against a commercial aerosol monitor (SidePak, TSI Incorporated). We found a strong correlation between these three sensor measurements and the commercial equivalent measurements ($R^2 > 0.89$). However, we found from the literature that the Sharp sensor showed limited accuracy over long-term monitoring [26]. A linear correlation was found between the Shinyei sensor and the TSI Aerosol Particle Sizer model 3321 when $PM_{2.5}$ level was lower than 50 $\mu g/m^3$. However, when $PM_{2.5}$ level was higher than 800 $\mu g/m^3$ there was no linear regression between measurements of Shinyei and TSI commercial equipment [27]. Similar results were reported by [28] and [29]. Another dust sensor, named PMS3003 (Plantower, China) was recommended after laboratory testing by our colleagues from NIWA [28]. NIWA tests showed a strong linear correlation ($R^2 = 0.99$) between PMS3003 and SidePak (TSI Incorporation) for $PM_{10}$ (ranged from 0 $\mu g/m^3$ to 350 $\mu g/m^3$) and under a temperature between 18°C and 28°C. We selected the PMS3003 dust sensor for our SKOMOBO platform development.

### C. DEPLOYMENT

The SKOMOBO platform deployment was carried out in two phases. For the first phase, we located one primary school close to Massey University, Auckland campus to facilitate the logistic. This first deployment was undertaken to check any issues before sending the SKOMOBO units to other

parts of the country. Once the first Auckland deployment was validated, the second phase was carried out in three areas - Hawkes' Bay in the North Island, Christchurch, and Dunedin which is both located in the South Island. The latitudes of these locations were 36°S, 39°S, 43°S, and 46°S for Auckland, Hawke's Bay, Christchurch, and Dunedin respectively. These areas were broadly representative of the climatic variation in NZ and also covered a mix of provincial towns and larger cities. Fig. 3 shows the deployment of the SKOMOBO platform among 11 schools in four locations:

- $1^{st}$ deployment: Auckland (North Island) – One school
- $2^{nd}$ deployment: Hawke's Bay (North Island) – Two schools (H1 and H2), Christchurch (South Island) – Four schools (C1, C2, C3 and C4), Dunedin (South Island) – Four schools (D1, D2, D3 and D4).

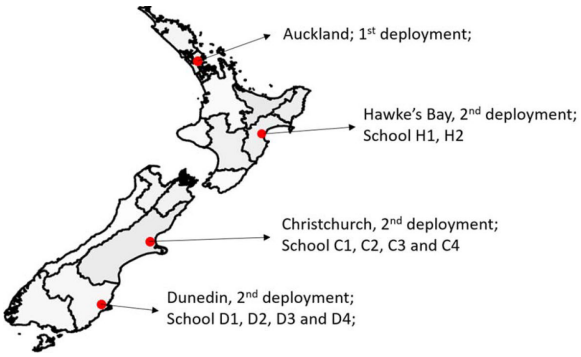



**FIGURE 3.** SKOMOBO deployment.

### D. DATA TRANSMISSION AND STORAGE

At each school, 2-4 SKOMOBO units were usually deployed across multiple classrooms along with a Wi-Fi dongle [23]. Strategically deployed SKOMOBO units collected PM data using the dust sensor described earlier. Connecting to the Internet with a built-in Wi-Fi interface, the PM data was sent as a part of HTTP request header from the SKOMOBO units to the Wi-Fi hotspot. The Wi-Fi hotspot routed the HTTP request header over the Internet using a mobile wireless connection. On a web server (*i.e.,* Microsoft IIS) running at a virtual machine at Massey University continuously listened to any incoming HTTP requests. Once an HTTP request arrived, the HTTP server forwarded the HTTP payload to Node.js application server. Node.js examined the HTTP payload and processed the sensor data. The sensor data was read, formatted, and then logged onto a MariaDB database. The deployment scenario of SKOMOBO is presented in Fig. 4.

### E. DATA VISUALIZATION

We have developed a dashboard that contains the indicators to present different IAQ parameters. The dashboard allows three different types of data users to view data for different purposes.

- The teacher and the student are presented with a set of visual data representations to see if they are exposed to a healthy environment.

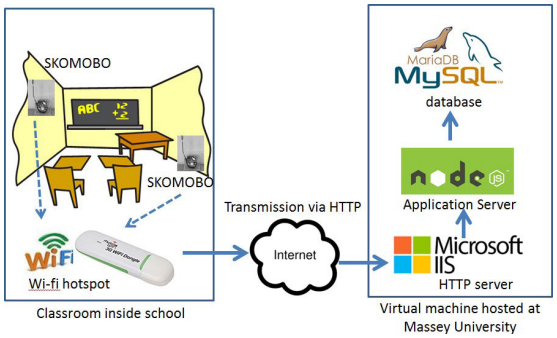



**FIGURE 4.** SKOMOBO deployment.

- The school principals and the Board of Trustee (BOT) members are able to visualize and easily export data for further discussion and meeting preparation.
- The researchers are able to run data mining queries to better understand the impact of different factors on the classroom environment.

Fig. 5 illustrates the preliminary dashboard screenshot to use by the researchers. Using the dashboard, the researchers are able to select the period of duration, the IAQ parameters, and produce statistics for this selected period which includes the mean, standard deviation, percentage of time in the comfort zone, etc.

**FIGURE 5.** The preliminary dashboard for the researchers.

We have also developed a Healthy Classroom Index (HCI) to facilitate data visualization for occupants and facility managers. Fig. 6 shows an example of HCI with three color code (*i.e.,* similar to traffic light system) to visualize the classrooms which are exposed to comfort level (green), or very close to comfort level (orange), or not comfortable and will need attention (red).

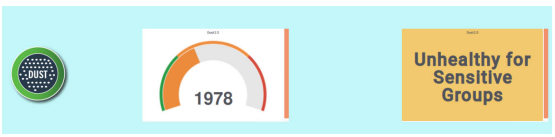



**FIGURE 6.** A Healthy Classroom Index (HCI) panel to visualize the data.

## IV. MSD-KMEANS

Outlier detection techniques can be classified into two categories: global and local outlier detection [30]. Global outliers can be defined as the observations that are different

from all the data points in the dataset which typically have very high or low values. Local outliers are defined as the observations that typically hold the data points that are only slightly different from the normal range. In this paper, firstly we introduce our proposed method MSD-Kmeans which can detect both global outliers (Step 1 in Algorithm 1) and local outliers (Step 2 in Algorithm 1) in the SKOMOBO sensor dataset. The proposed algorithm MSD-Kmeans employs the statistical method of MSD to eliminate most sensitive global outliers (*i.e.,* extreme values) to minimize their interference on efficient clustering by K-means. Since the number of normal data points is generally greater than the number of outliers, if the extreme value can be eliminated before clustering via K-means, the efficiency and accuracy and local optima can be improved. This is what is happening in the first phase of MSD-Kmeans. In the second phase, utilizing the remaining normal data from the MSD method, the K-means algorithm is applied to partition data into clusters.

The first phase of outlier detection undergoes the following steps:

1) Calculating the maximum extreme value $o_{max}$:
   Let define $\chi = \{x_1, x_2, x_3, \ldots, x_i\}$ as a set of the $PM_{10}$ data records in the dataset. Then, the maximum extreme value of $\chi$ ($o_{max}$) can be obtained as $o_{max} = \mu + \sigma$, where $\mu$ is the mean value and $\sigma$ is the standard deviation of the dataset $\chi$, respectively.
2) Calculating the Minimum extreme value $o_{min}$:
   Similarly, we compute the minimum extreme value of $\chi$ ($o_{min}$) based on both mean and standard deviation as $o_{min} = \mu - \sigma$.
3) We can then figure out both maximum and minimum extreme values $\mathbb{N}$ and $\mathbb{S}$ based on Formulas 1 and 2, respectively.

$$\mathbb{N} = \{x_j | x_j \in \chi, x_j < o_{max} \wedge x_j > o_{min}\} \quad (1)$$

$$\mathbb{S}_G = \{x_j | x_j \in \chi, x_j < o_{min} \vee x_j > o_{max}\}, \quad (2)$$

   where $0 < j \leq i$, and $\mathbb{N}$ is a set of normal values of data in $\chi$, and $\mathbb{S}_G$ represents a set of global outliers determined by Formula 2.
4) K-means clustering based on normal dataset $\mathbb{N}$.

The size of confidence intervals can be considered as the data deviating from the mean plus or minus two or three coefficient times from the standard deviation [31]. The levels of confidence intervals can also depend on the different application scenarios [31], [32]. Outliers often can be defined as the data out of those confidence levels. Note that, in our implementation of MSD-Kmeans, we use one standard deviation and the mean value to fence in the normal values as this was close to the ground zero knowledge we learned from similar scenarios [9], [33].

The second phase starts after eliminating the global outliers using the MSD algorithm, the remaining normal data and local outliers were grouped into two clusters by applying the K-means clustering algorithm. If the dataset assumed that all data points in each clustering are close to each other, outliers can be detected in each cluster based on the threshold of each cluster. The threshold in our proposal is calculated based on the intra-cluster distance of each cluster. Intra-cluster distance is the Euclidean Distance (ED) calculated from each data point to the centroid of the cluster. The outlier threshold is calculated as the sum of the mean value and 1.5 times the standard deviation of intra-clustering distance. The approach by

**Algorithm 1** MSD-Kmeans Algorithm

**Input:** $\chi = \{x_1, x_2, x_3, \ldots, x_i\}$
**Output:** $\mathbb{S}_G$,$\mathbb{S}_L$

```
    /* step 1: MSD for global outliers */
1:  Calculate μ and σ of χ
2:  for each x_j ∈ χ do
3:      if (x_j < σ − μ) || (x_j > σ + μ) then
4:          Add (x_j, S_G)
5:  end for
    /* step 2: K-means for clustering */
6:  for each c_i ∈ C do
7:      c_i ← x_k ∈ χ − S_G
    /* C is a set of cluster centroids
    determined with number of clusters,
    where |C|=k */
8:  end for
9:  for each x_s ∈ (χ − S_G) do
10:     l(x_u) ← ArgMinDistance (x_u, c_v), v ∈ {1 . . . k}
    /* l(xu): Making label for each
    cluster, note that k is the number
    of clusters
11: end for
12: changed ← false
13: while (changed == false) do
14:     for each c_i ∈ C do
15:         UpdateClusters(c_i)
16:     end for
17:     for each x_u ∈ (χ − S_G) do
18:         dist ← ArgMinDistance (x_u, c_v), v ∈ {1 . . . k}
19:         if (dist ≠ l(x_u)) then
20:             l(x_u) ← dist
21:             changed ← true
22:     end for
23: end while
    /* step 3: K-means for local
    outliers */
24: Calculate μ and σ of χ − S_G
25: for each x_s ∈ (χ − S_G) do
26:     o = DistanceFromCentroid (x_s)
27:     if (o > μ + 1.5 ∗ σ) then
28:         Add (x_s, S_L)
    /* SL is a set of local outliers */
29: end for
30: return S_L
```

using K-means to find outliers is also named Top-n outliers, meaning the n data points which have the largest distances from their $K^{th}$ nearest neighbors [34]. Here, top-n outlier data points are defined as those when the intra-cluster distance of each data point is greater than the calculated threshold.

As we mentioned in previous research [9], K-means is suitable for dealing with spherical-like shaped clusters (*i.e.* distribution of data on a plot looks like a sphere) [16]. We applied K-means in the second stage to detect local and the remaining global outliers among PM data. According to [35] and [9], $k = 2$ appeared to have produced the best clustering results compared to when $k = 3$ or more were used. According to step 2 in Algorithm 1, the intra-cluster distance, from each data point to the centroid of the cluster it belongs to is calculated. All intra-cluster distances are sorted into descending orders in each cluster. Finally, the threshold of outlier values is calculated as the sum of the mean value and 1.5 times the standard deviation of intra-cluster distances in each cluster.

## V. EXPERIMENT RESULTS

In this section, we describe the details of our experimental results to understand the feasibility of the MSD-Kmeans to be used in a real dataset such as the $PM_{10}$ dataset from SKOMOBO units.

### A. EXPERIMENT SETUP

We run our experiments in the following system to measure the performance in Table 1.

**TABLE 1. Implementation environment specification.**

| Unit | Description |
|---|---|
| Processor | 3.4GH$_z$ Inter Core i5 |
| RAM | 16GB |
| OS | MacOS Mojave |

**TABLE 2. The number of SKOMOBO data records collected per month.**

| Date | The number of Data Records |
|---|---|
| 2017-08 | 3,200 |
| 2017-09 | 1,209,165 |
| 2017-10 | 2,251,576 |
| 2017-11 | 2,269,150 |
| 2017-12 | 1,865,934 |
| 2018-01 | 1,711,932 |
| 2018-02 | 819,308 |
| 2018-03 | 1,016,665 |
| **Total** | 11,218,930 |

#### 1) DATA ANALYSIS

PM and other data in SKOMOBO were stored in every 1-second interval. The typical amount of the data and their distribution on monthly basis can be seen in Table 2. The amount of the daily collection of data was varying from day to day. For example, there were only approximately 3,000 records in the 10 days during the $21^{th}$ to

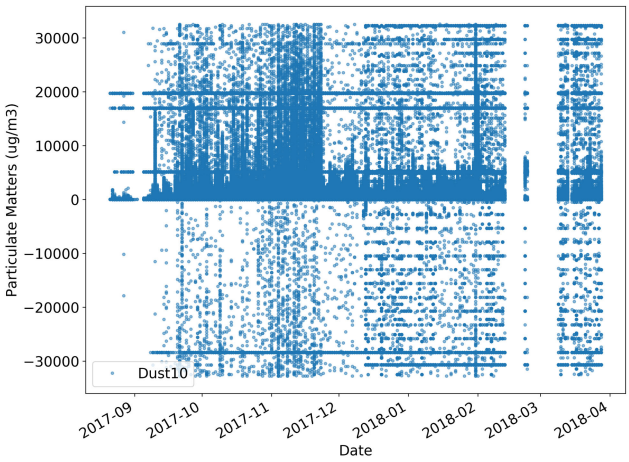



**FIGURE 7. Distribution of all $PM_{10}$ data.**

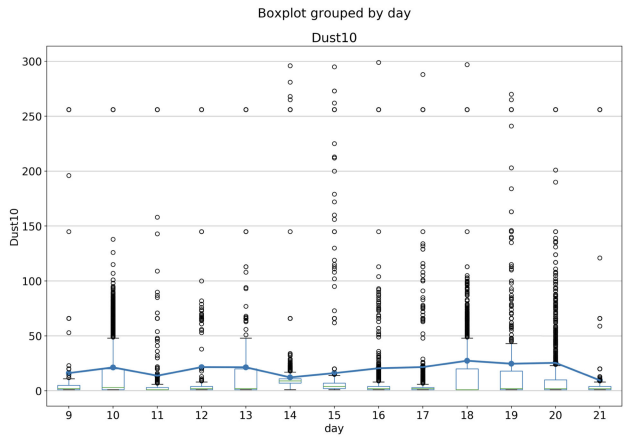



**FIGURE 8. Boxplot of outliers from September 9, 2017 to September 21, 2017.**

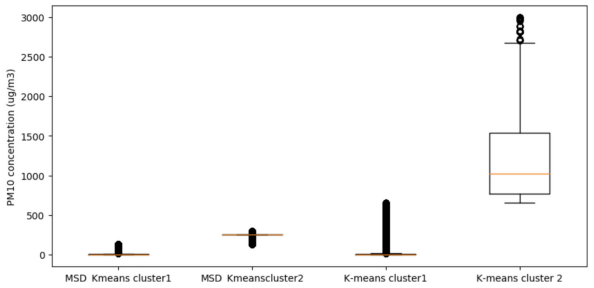
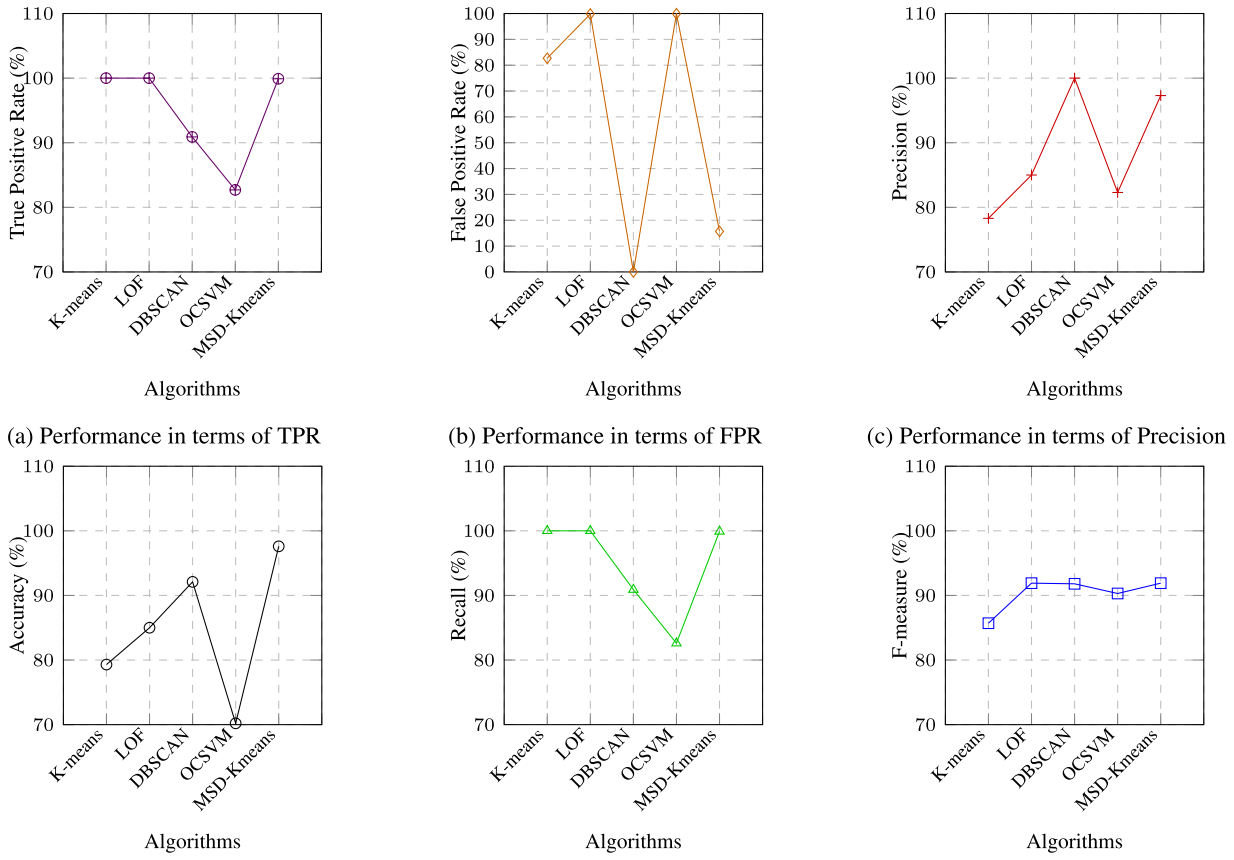



**FIGURE 9. Normal and outliers $PM_{10}$ distribution using MSD-Kmeans and K-means algorithm.**

$31^{st}$ August 2017 period while there were approximately 819,000 records in the 23 days of February 2018. We found that SKOMOBO units on some dates were not turned on or the server running the database was shut down by Massey network administrator for the maintenance purposes, *etc*. Fig. 7 shows the data when it appeared that the full sets of data were available for $PM_{10}$.

The boxplots in Fig. 8 shows the outliers during the 13 days from September 9 to September 21 in 2017 based on Tukey's test [36]. All small circles represent outliers that sit outside of a threshold in which depicted by a square box at the bottom of




(a) Performance in terms of TPR (b) Performance in terms of FPR (c) Performance in terms of Precision

(d) Performance in terms of Accuracy (e) Performance in terms of Recall (f) Performance in terms of F-measure

**FIGURE 10.** **Performance comparison of outlier detection algorithms.**

the graph. A small blue circle that is filled above the threshold square box represents a mean value for the day. All mean values are connected via a blue line. As we can observe, the position of the mean values is sensitive to classifying outliers.

We selected the values of $PM_{10}$ from which we applied MSD to detect global outliers. This may include the dust concentration of a extreme value such as -32768 ($ug/m^3$). This could be a programming error as it is unlikely that $PM_{10}$ reading from a PM sensor would reach such a negative value when a typical acceptable PM level would normally sit in the range from 0 ($ug/m^3$) to 50 ($ug/m^3$).

### B. DETECTION RATES

According to National Environment Standards for Air Quality (NESAQ) [37], it defines the standard range $PM_{10}$ is under 50 micrograms per cubic meter ($ug/m^3$) on average across 24-hour. This is the recommended value to provide a healthy indoor environment for people living in New Zealand. Based on this recommendation, we used the criteria of the range 0 ($ug/m^3$) to 50 ($ug/m^3$) as the normal data whereas all below and above this range is treated to be outliers.

Fig. 9 shows the results of only using the K-means algorithm as a main outlier detection mechanism based on the entire dataset of about 11 million records. The result

**TABLE 3.** **Confusion matrix.**

| Total Poputlation | | Predicted Condition | |
|---|---|---|---|
| | | Normal | Outlier |
| Actual Condition | Normal | TP | FN |
| | Outlier | FP | TN |

illustrated that any data that were further away from the centroid were more likely considered as outliers.

In other words, depending on the distance of intra-cluster it can affect the threshold which again is sensitive to the range of outlier outcome. For example, the threshold of intra-cluster distance was 116 when the centroid value was 17 ($ug/m^3$) in cluster 1. The threshold of intra-cluster distance was 913 when the centroid value was 1294 ($ug/m^3$) in cluster 2. Here the value of centroid is relatively high due to the large distribution over the input dataset which includes extreme values. This would result in the threshold that is further away from the centroid. This would adversely affect the outlier outcomes where it is unable to detect hidden outliers [16].

### C. PERFORMANCE COMPARISON

The performance of an outlier detection algorithm can be evaluated using six possible performance indicators.

**TABLE 4.** Performance comparison of outlier detection algorithms using SKOMOBO dataset.

| Outlier Detection Algorithm | TPR (%) | FPR (%) | Precision (%) | Accuracy (%) | Recall (%) | F-measure (%) | Execution Time (MS) |
|---|---|---|---|---|---|---|---|
| K-means [13] | 100 | 82.7 | 78.3 | 79.3 | 100 | 85.7 | 8,125 |
| LOF [14] | 100 | 99.8 | 85.0 | 85.0 | 100 | 91.9 | 612,799 |
| DBSCAN [39] | 90.9 | 0 | 100 | 92.1 | 90.9 | 91.8 | 42,129 |
| OCSVM [40] | 82.7 | 99.9 | 82.3 | 70.2 | 82.6 | 90.3 | 2,073,348 |
| MSD-Kmeans | 99.9 | 15.7 | 97.3 | 97.6 | 99.9 | 91.9 | 12,364 |

This includes; True Positive Rate (TPR), False Positive Rate (FPR), Precision, Accuracy, Recall, and F-measure [38].

We used a confusion matrix to reflect the classification results of a certain classification model as in Table 3, in which True Positive (TP) shows the normal data correctly classified as normal, True Negative (TN) denotes outliers correctly classified as outliers, False Positive (FP) indicates outliers incorrectly classified as normal data, and finally False Negative (FN) shows normal data incorrectly classified as outliers.

Using the confusion table, we compute the indicators. True Positive Rate (*TPR*) is also known as sensitivity or recall, which denotes the proportion of samples correctly classified as normal samples. It can be calculated as Equation 3.

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

False Positive Rate (*FPR*) is computed using Equation 4 which denotes the proportion of samples correctly classify as an outlier.

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

Then, *Precision* (also called as positive predictive value) and *Accuracy* can be computed using Equation 5 and Equation 6, respectively. Precision indicates correctly predicting the normal samples and Accuracy is the percentage of correct prediction and shows the proportion of the number of normal correctly classified samples to total samples for a given dataset.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

We then compute the *recall* (also known as sensitivity) indicator based on Equation 7 which measures the proportion of true positive samples correctly classified from true positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Finally, we compute $F - measure$ which is a measure of the harmonic mean of the precision and the true positive rate based on Equation 8.

$$F - measure = \frac{2 * TP}{2 * TP + FP + FN} \tag{8}$$

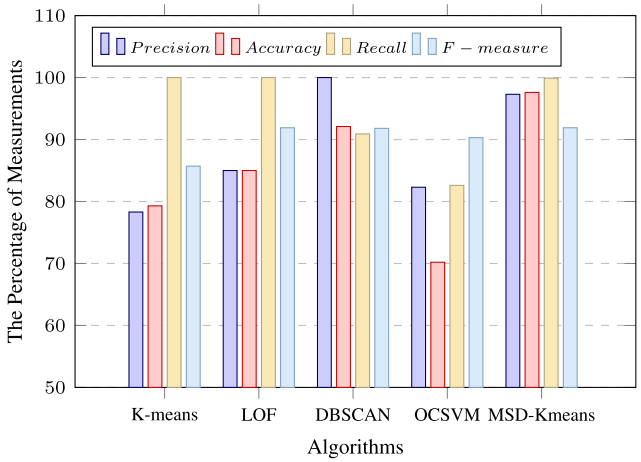



**FIGURE 11.** Comparison of the percentage of outlier detection algorithms.

We have calculated our results in these performance indicators and compared them with those of other outlier detection algorithms by applying the same SKOMOBO dataset.
The overall results across all performance indicators and execution time are shown in TABLE 4. Examining the details of each performance indicator, MSD-Kmeans and other algorithms show relatively high TPR above 90% except One-Class SVM (OCSVM) as shown in Fig. 10 (a). MSD-Kmeans resulted the lowest FPR value at 15.7% (as shown in Fig. 10 (b)) while Precision score reached the highest at 97.3% (shown in Fig. 10(c)). Similarly, MSD-Kmeans had the highest Accuracy at 97.6% (in Fig. 10 (d)) and F-measure 91.9% (in Fig. 10 (f)) compared to other machine learning algorithms. Observing the rate of Recall, all algorithms showed relatively high rates above 90% except OCSVM as shown in Fig. 10 (e).

In terms of the execution time (MS) as shown in TABLE 4, MSD-Kmeans resulted in one of the best execution times due to its ability to remove extreme values in the earlier stage. Other similar machine learning approaches such as LOF, DBSCAN, and OCSVM took a significantly longer time to execute their respective algorithms.

Fig. 11 shows the comparison of all performance indicators of different outlier detection algorithms. Our proposed method of MSD-Kmeans is clearly well-performing in almost all aspects of performance indicators of accurate outlier detecting compared to other similar methods. This indicates that our MSD-Kmeans is reliable to use as an effective outlier

detection algorithm for real-life dataset such as SKOMOBO $PM_{10}$ as shown in here.

## VI. CONCLUSION

We proposed the use of MSD-Kmeans as an outlier detection to identify outlying $PM_{10}$ rates. The $PM_{10}$ data we used for this feasibility study is produced from the real-life application that deploys 165 SKOMOBO units across New Zealand primary schools that monitor the healthiness of the classrooms for our children. The hybrid method MSD-Kmeans applies the MSD algorithm to eliminate as many extreme values as possible, before applying the K-means clustering algorithm to cluster normalized datasets in different groups. Our experimental result demonstrated that MSD-Kmeans achieved the best scores for many important performance indicators such as accuracy and F-measure, with the lower false positive rate, compared to other outlier detection algorithms applied on the same dataset. We believe that MSD-Kmeans is a promising algorithm in outlier detection that could benefit the processing of sensor data from networked IoT devices.

In the future, we plan to extend MSD-Kmeans along with multivariate value analysis to understand its strengths or weaknesses in detecting outliers with multiple attributes where relationships among them could affect the outlier outcomes.

## REFERENCES

[1] D. Shendell, R. Prill, W. Fisk, M. Apte, D. Blake, and D. Faulkner, "Associations between classroom $CO_2$ concentrations and student attendance," Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. lbnl-53586, 2003.

[2] S. Gaihre, S. Semple, J. Miller, S. Fielding, and S. Turner, "Classroom carbon dioxide concentration, school attendance, and educational attainment," *J. School Health*, vol. 84, no. 9, pp. 569–574, Sep. 2014.

[3] T. Yu, X. Wang, and A. Shami, "Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 2207–2216, Dec. 2017.

[4] N. Javed and T. Wolf, "Automated sensor verification using outlier detection in the Internet of Things," in *Proc. 32nd Int. Conf. Distrib. Comput. Syst. Workshops*, Jun. 2012, pp. 291–296.

[5] (2019). *Ministry for the Environment Annual Report*. [Online]. Available: https://www.mfe.govt.nz/publications/about-us/ministry-environment-annual-report-2019

[6] (2019). *Monitoring New Zealand's Environmental Health*. [Online]. Available: http://www.ehinz.ac.nz

[7] J. Morton. (2018). *NZ's Hidden Health Threat: Poor Indoor Air Quality*. [Online]. Available: https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=12146670

[8] P. Taptiklis and R. Phipps, "Indoor air quality in new zealand homes and schools: A literature review of healthy homes and schools with emphasis on the issues pertinent to new zealand," Building Res. Assoc. New Zealand, Porirua, New Zealand, Tech. Rep., 2017.

[9] Y. Wei, J. Jang-Jaccard, F. Sabrina, and T. McIntosh, "MSD-kmeans: A novel algorithm for efficient detection of global and local outliers," 2019, *arXiv:1910.06588*. [Online]. Available: http://arxiv.org/abs/1910.06588

[10] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[11] C. C. Aggarwal, "Probabilistic and statistical models for outlier detection," in *Outlier Analysis*. New York, NY, USA: Springer, 2017, pp. 35–64.

[12] Y. S. Prasad and G. R. Krishna, "Statistical anomaly detection technique for real time datasets," *Int. J. Comput. Trends Technol.*, vol. 6, no. 2, pp. 89–94, 2013.

[13] L. Duan, L. Xu, Y. Liu, and J. Lee, "Cluster-based outlier detection," *Ann. Oper. Res.*, vol. 168, no. 1, pp. 151–168, 2009.

[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Rec.*, 2000, vol. 29, no. 2, pp. 93–104.

[15] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[16] O. A. Abbas, "Comparisons between data clustering algorithms," *Int. Arab J. Inf. Technol.*, vol. 5, no. 3, pp. 1–8, 2008.

[17] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in *Proc. GI/ITG Workshop MMBnet*, 2007, pp. 13–14.

[18] N. Kant and M. Mahajan, "Time-series outlier detection using enhanced k-means in combination with pso algorithm," in *Engineering Vibration, Communication and Information Processing*. Singapore: Springer, 2019, pp. 363–373.

[19] K. Mumtaz and K. Duraiswamy, "A novel density based improved k-means clustering algorithm–Dbkmeans," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 213–218, 2010.

[20] J. Wang and X. Su, "An improved k-means clustering algorithm," in *Proc. IEEE 3rd Int. Conf. Commun. Softw. Netw.*, May 2011, pp. 44–46.

[21] A. Hatamlou, S. Abdullah, and H. Nezamabadi-Pour, "A combined approach for clustering based on K-means and gravitational search algorithms," *Swarm Evol. Comput.*, vol. 6, pp. 47–52, Oct. 2012.

[22] C.-F. Tsai and C.-Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern Recognit.*, vol. 43, no. 1, pp. 222–229, Jan. 2010.

[23] R. Weyers, J. Jang-Jaccard, A. Moses, Y. Wang, M. Boulic, C. Chitty, R. Phipps, and C. Cunningham, "Low-cost indoor air quality (IAQ) platform for healthier classrooms in New Zealand: Engineering issues," in *Proc. 4th Asia–Pacific World Congr. Comput. Sci. Eng. (APWC CSE)*, Dec. 2017, pp. 208–215.

[24] Y. Wang, M. Boulic, R. Phipps, C. Chitty, A. Moses, R. Weyers, J. Jang-Jaccard, G. Olivares, A. Ponder-Sutton, and C. Cunningham, "Integrating open-source technologies to build a school indoor air quality monitoring box (SKOMOBO)," in *Proc. 4th Asia–Pacific World Congr. Comput. Sci. Eng. (APWC CSE)*, Dec. 2017, pp. 216–223.

[25] O. Gustavo. (2017). *Dust-Acorn Pacman*. [Online]. Available: https://github.com/niwa/dusty-acorn/tree/master/PACMAN

[26] M. Budde, M. Busse, and M. Beigl, "Investigating the use of commodity dust sensors for the embedded measurement of particulate matter," in *Proc. 9th Int. Conf. Netw. Sens. (INSS)*, Jun. 2012, pp. 1–4.

[27] E. Austin, I. Novosselov, E. Seto, and M. G. Yost, "Laboratory evaluation of the shinyei PPD42NS low-cost particulate matter sensor," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0137789.

[28] O. Gustavo. (2016). *From Sharp to Plantower—What we Learned About Cheap Dust Sensors*. [Online]. Available: https://figshare.com/articles/From_Sharp_to_Plantower_What_we_learned_about_cheap_dust_sensors/4213815

[29] K. K. Johnson, M. H. Bergin, A. G. Russell, and G. S. Hagler, "Using low cost sensors to measure ambient particulate matter concentrations and on-road emissions factors," *Atmos. Meas. Techn. Discuss.*, vol. 2016, pp. 1–22, 2016.

[30] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1369–1382, May 2015.

[31] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Experim. Social Psychol.*, vol. 49, no. 4, pp. 764–766, Jul. 2013.

[32] V. M. van Zoest, A. Stein, and G. Hoek, "Outlier detection in urban air quality sensor networks," *Water, Air, Soil Pollut.*, vol. 229, no. 4, p. 111, Apr. 2018.

[33] (2017). *Taxi in New York City—Yellow Cabs*. [Online]. Available: https://loving-newyork.com/taxi-in-new-york-city-yellow-cabs/

[34] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2000, vol. 29, no. 2, pp. 427–438.

[35] M. S. G. Karypis, V. Kumar, and M. Steinbach, "A comparison of document clustering techniques," in *Proc. TextMining Workshop KDD*, May 2000.

[36] J. W. Tukey, "Exploratory data," in *Analysis*. Reading, MA, USA: Addison-Wesley, 1977.

[37] (2018). *New Zealand's Environmental Reporting Series: Our Air 2018*. [Online]. Available: https://www.mfe.govt.nz/sites/default/files/media/Air/our-air-2018.pdf

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[39] M. Celik, F. Dadaser-Celik, and A. S. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *Proc. Int. Symp. Innov. Intell. Syst. Appl.*, Jun. 2011, pp. 91–95.
[40] M. Amer, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in *Proc. ACM SIGKDD Workshop Outlier Detection Description (ODD)*, 2013, pp. 8–15.

**YUANYUAN WEI** received the master's degree in information technology from Massey University, New Zealand, where she is currently pursuing the Ph.D. degree with the School of Natural and Computational Sciences. Her research interests include AI-powered anomaly detection, machine learning, and deep learning.

**JULIAN JANG-JACCARD** received the M.Sc. and Ph.D. degrees from The University of Sydney, Australia. She is currently an Associate Professor and also the Lead of the Cyber Security Laboratory, Massey University, New Zealand. She has published more than 70 papers in the leading conferences and journal venues, including IEEE and ACM. Her research interests include cybersecurity, intrusion detection, artificial intelligence, machine learning, deep learning, data anonymization, and privacy-preservation techniques. She was also a recipient of many multi-million dollar research awards both from Australian and NZ governments while collaborating with the top international ICT companies and universities around the world.

**FARIZA SABRINA** (Member, IEEE) received the Master of Engineering degree (by research) in electrical and information engineering from The University of Sydney, Australia and the Ph.D. degree in computer science and engineering from the University of New South Wales. She is currently a Senior Lecturer in ICT with the School of Engineering and Technology, Central Queensland University, Australia. She has authored/coauthored and published many papers in top-ranking conferences and journals. Her research interests include cybersecurity, the Internet of Things, network security, blockchain, artificial intelligence, machine learning, and deep learning.

**HOOMAN ALAVIZADEH** (Member, IEEE) received the M.Sc. degree in computer science from Eastern Mediterranean University (EMU), Cyprus, and the Ph.D. degree in cybersecurity from Massey University, Auckland, New Zealand. He is currently a Postdoctoral Fellow with the School of Natural and Computational Sciences, Massey University. His research interests include moving target defense (MTD), cybersecurity situation awareness, AI-based attack and defense evaluation, network security modeling and analysis, cloud computing security, and cryptography.