# Multivariate $T^2$ Control Chart Based on James-Stein and Successive Difference Covariance Matrix Estimators for Intrusion Detection

**Muhammad Ahsan[1a], Muhammad Mashuri[1b*], Heri Kuswanto[1c], Dedy Dwi Prastyo[1d] and Hidayatul Khusna[1e]**

[1] Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, INDONESIA. E-mail: ahsan4th@gmail.com[a] ; m_mashuri@statistika.its.ac.id[b] ; heri_k@statistika.its.ac.id[c] ; dedy-dp@statistika.its.ac.id[d] ; khusna16@mhs.statistika.its.ac.id[e]
* Corresponding Author: m_mashuri@statistika.its.ac.id
Received: 21st April 2019 Revised : 19th September 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.3

**ABSTRACT** The intrusion detection is a process to monitor the events taking place in a computer system or network and analyse the monitoring results to find signs of intrusion. The multivariate control chart, which is often used in the intrusion detection system, is Hotelling's $T^2$. In this study, the Hotelling's $T^2$ chart performance for intrusion detection is improved using the successive difference covariance matrix to estimate the covariance matrix and James-Stein estimator to estimate the mean vector. The control limits of the proposed chart are calculated using kernel density estimation. The performance of the proposed method, using $T^2$ based on kernel density estimation control limit, outperforms the other control chart approaches in both training and testing dataset.

**Keywords**: $T^2$ control chart, James-Stein estimator, successive difference covariance matrix, kernel density estimation, intrusion detection.

## 1. INTRODUCTION

The network security systems are needed to be improved because of the rapid development of information exchange processes. In this case, a security system that can prevent and warn of the attacks on the network is needed. One security mechanism that can be used in preventing attacks is the intrusion detection system (IDS). The intrusion detection is a process to monitor the events, which are taking place in a computer system or network and to analyse the monitoring results to find the signs of anomalies in the network (Bace & Mell, 2001). The IDS is an essential tool in modern computing infrastructure to monitor and identify unwanted and suspicious network traffic, related to the unauthorised system access or poorly configured systems (Shenfield, Day, & Ayesh, 2018). The IDS has become an important component in computer network architecture because it has similar purposes to the burglar alarm that gives warning to every malicious event in the network.

The statistical process control (SPC) approach can be performed not only in the industrial field, but also be adopted in network intrusion detection. The SPC is a quality control method, which utilises the statistical methods to monitor and control the process. The control charts, which is one of the favourite tools used in SPC, is a graph employed to study the stability of the process over time. Based on the type of the quality characteristics monitored, the control charts are classified into attribute (Ahsan, Mashuri, & Khusna, 2017; Wibawati et al., 2016; Wibawati et al., 2018) and variable (Page, 1961; Roberts, 1959; Shewhart, 1924) control charts. Meanwhile, based on the number of quality characteristics, the control charts are classified into univariate and multivariate control charts. The univariate control chart only considers one characteristic. On the other

hand, the multivariate control chart is used to control production process with more than one characteristics. The recent development for the multivariate control chart includes Hotelling's $T^2$ control chart (Abu-Shawiesh, Kibria, & George, 2014; Ahsan et al., 2018a; Alkindi, Mashuri, & Prastyo, 2016), MCUSUM control chart (Arkat, Niaki, & Abbasi, 2007; Issam & Mohamed, 2008; Khusna et al., 2018a) and MEWMA control chart (Khusna et al., 2018b; Pirhooshyaran & Niaki, 2015).

The SPC can be used as a capable technique, which can guarantee the system security and stability in network monitoring and intrusion detection process (Bersimis, Sgora, & Psarakis, 2016). The superiority of applying this method to monitor the anomalies in the network does not require the knowledge of information from the prior attacks. This advantage makes the SPC based IDS to be easily executed in the online detection system (Catania & Garino, 2012). There are many studies on SPC, that have been implemented in IDS, for both the univariate and multivariate cases (Park, 2005). Based on the previous research, it can be known that the Hotelling's $T^2$ chart is the most commonly used control chart for the intrusion detection. The Hotelling's $T^2$, which uses the conventional mean and covariance matrix, is reactive to the outlier. As a result, the conventional method is not effective to use for multiple outliers' case, because of the masking effect (Alfaro & Ortega, 2009). The masking effect in the monitoring process happens as a result of the outlier, which cannot be detected by the control chart. To overcome the problem that arises, several robust methods have been proposed to reduce the effect of multiple outliers by substituting the existing estimators with the more robust ones. Moreover, the performance of Hotelling's $T^2$ control chart, in detecting the shift of mean vector, is increasing when the robust covariance matrix estimator is implemented (Williams et al., 2006). The successive difference covariance matrix (SDCM) is one of the robust covariance matrix estimators, that can be used in this instance. The Hotelling's $T^2$ control chart, based on SDCM, is powerful in detecting the shift of mean vector (Sullivan & Woodall, 1996; Vargas, 2003). Moreover, SDCM can also be exploited for auto-correlated processes, such as $T^2$ control chart based on the SDCM for multivariate process, using residuals of vector autoregressive (VAR) model (Wororomi et al., 2014).

Many researchers have proved the effectiveness of Hotelling's $T^2$ control chart based on SDCM. However, the exact distribution for this control chart has not been discovered. Sullivan and Woodall (1996) and Williams et al. (2006) suggested the approximate distribution for Hotelling's $T^2$ control chart based on SDCM. In addition, some studies have been conducted to improve Hotelling's $T^2$ chart control limit, by using nonparametric approaches, to overcome the limited research of Hotelling's $T^2$ based on SDCM distribution. Those studies have been conducted to improve the control limit of Hotelling's $T^2$ using kernel density estimation (KDE) technique (Chou, Mason, & Young, 1999; Phaladiganon et al., 2011; Phaladiganon et al., 2013). Using the same approach, Ahsan et al. (2018b) implemented the idea of KDE into the Hotelling's $T^2$ based on SDCM to monitor the anomalies in the network. The multivariate Hotelling's $T^2$ based on SDCM has a good performance to monitor the anomalies in the network when it is applied to the bootstrap resampling method to calculate the control limit (Ahsan, Mashuri, & Khusna, 2018).

The improvement not only can be done on covariance matrix as stated before, but also can be done on the mean vector. The shrinkage estimators, which have smaller mean squared errors than the traditional estimators, can be practised in this issue (Lehmann & Casella, 2006; Stein, 1956). The James-Stein estimator (James & Stein, 1961), which is the improved estimator of the mean vector, can be employed

to get the better result in estimating the mean vector of the Hotelling's $T^2$ control chart. Wang, Huwang and Yu (2015) found that the performance of the multivariate control chart with the James-Stein estimator is better than that of the existing control charts. Therefore, the aim of this study is to propose Hotelling's $T^2$ control chart based on the hybrid James-Stein and SDCM using the KDE approach. The James-Stein estimator is employed to estimate the mean vector, while the SDCM is adopted to estimate the robust covariance matrix. Theoretically, applying the proposed control chart into IDS can refine the performance in network monitoring system. To prove this statement, the performance of the proposed method is compared with the other methods.

This paper is organised as follows: section 2 describes $T^2$ control chart based on James-Stein and SDCM, while the control limit of Hotelling's $T^2$ control chart using KDE is explained in section 3; section 4 presents the dataset and methodology that were used in this research; the performance comparisons of the proposed IDS with the other control charts are presented in section 5; and finally, section 6 summarises the obtained results and presents a future research.

## 2. HOTELLING'S $T^2$ CONTROL CHART BASED ON JAMES-STEIN AND SDCM ESTIMATORS

In this section, the proposed Hotelling's $T^2$ control chart based on James-Stein and SDCM estimators is presented. The Hotelling's $T^2$ is one of multivariate control charts that can be used to monitor the mean of a multivariate process (Montgomery, 2009). Let $\mathbf{X} = \left[\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'\right]'$, where $i = 1, 2, \ldots, n$ number of observations are identic and independent random vectors, which follow multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Using $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'$, $T^2$ statistic can be calculated as follows (Hotelling, 1974):

$$T_i^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \tag{1}$$

By assuming the data to follow multivariate normal distribution, the control limit of Hotelling's $T^2$ can be obtained with the following equation:

$$CL = \frac{p(n+1)(n-1)}{n^2 - np} F_{(\alpha, p, n-p)}, \tag{2}$$

where $n$ is number of observations, $p$ is number of variables and $\alpha$ is false alarm rate. The process is concluded as in-control, if $T^2$ statistic in Equation (1) is lower than the control limit, $CL$ formulated in Equation (2).

The SDCM is another alternative method to estimate the covariance matrix that was first introduced by Hawkins and Merriam (1974) as well as Holmes & Mergen (1993). This method is constructed by changing the estimated covariance matrix $\mathbf{S}$ with the SDCM. Under in-control condition, the SDCM, $\mathbf{S}_D$ is an unbiased estimator for covariance matrix $\boldsymbol{\Sigma}$ (Sullivan & Woodall, 1996). The $T^2$ based on SDCM can be calculated as follows:

$$T_{D,i}^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \mathbf{S}_D^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right), \tag{3}$$

where

$$\mathbf{S}_D = \frac{1}{(n-1)} \sum_{i=2}^{n} \left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)\left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)'. \tag{4}$$

There are some approaches in constructing the control limit of $T_D^2$ statistics that follow the multivariate normal distribution, such as the control limit based on Sullivan and Woodall ($CL_{SW}$) (Sullivan & Woodall, 1996), control limit based on Mason and Young ($CL_{MY}$) (Mason & Young, 2002) and control limit based on Chi-square distribution ($CL_{\chi^2}$). Those control limits can be calculated using Equations (5) to Equation (7).

$$CL_{SW} = \frac{(n-1)^2}{n} BETA_{(1-\alpha),\frac{p}{2},\frac{(g-p-1)}{2}}, \tag{5}$$

$$CL_{MY} = \frac{(f-1)^2}{f} BETA_{(1-\alpha),\frac{p}{2},\frac{(g-p-1)}{2}} \tag{6}$$

$$CL_{\chi^2} = \chi^2_{(1-\alpha),u} \tag{7}$$

where $BETA_{(1-\alpha),p,g}$ $BETA_{(1-\alpha),p,g}$ is $[1-\alpha]$-th quantile of Beta distribution, $p$ is number of quality characteristics and $g$ is the shape parameter, $\chi^2_{(1-\alpha),u}$ is $[1-\alpha]$-th quantile of Chi-square distribution with $u$ degree of freedom and $g = \frac{2(n-1)^2}{3n-4}$.

In this study, the James-Stein estimator is used to construct better Hotelling's $T^2$ control charts. The basic form of the James-Stein estimator is formulated as follows:

$$\bar{\mathbf{x}}_0^{JS} = \left(1 - \frac{p-2}{n\left(\bar{\mathbf{x}} - \mathbf{v}\right)^T \boldsymbol{\Sigma}^{-1} \left(\bar{\mathbf{x}} - \mathbf{v}\right)}\right)\left(\bar{\mathbf{x}} - \mathbf{v}\right) + \mathbf{v}, \tag{8}$$

where $\mathbf{v}$ is a fixed vector, that contains the target value, by which $\bar{\mathbf{x}}$ will be shrunk. According to Lehmann & Casella (2006), $\mathbf{v}$ can be determined as any $p$-dimensional vector. The James-Stein estimator can be improved to have smaller mean square error (MSE) than the standard James-Stein estimator, as in Equation (9), by Wang et al. (2015). The improved James-Stein estimator can be calculated as follows:

$$\bar{\mathbf{x}}^{JS} = \left(1 - \frac{p-2}{n\left(\bar{\mathbf{x}} - \mathbf{v}\right) \boldsymbol{\Sigma}^{-1} \left(\bar{\mathbf{x}} - \mathbf{v}\right)}\right)\left(\bar{\mathbf{x}} - \mathbf{v}\right) + \mathbf{v}, \tag{9}$$

where the notation $f(x)^+$ is defined as:

$$f(x)^+ = \begin{cases} f(x), \text{ if } f(x) > 0 \\ 0, \text{ otherwise.} \end{cases}$$

The improved James-Stein estimator, as in Equation (9) is utilised to construct the Hotelling's $T^2$ control chart based on James-Stein estimator, which is constructed by replacing the $\bar{\mathbf{x}}$ in Equation (1) with $\bar{\mathbf{x}}^{JS}$. The statistics of the proposed chart is formulated as follows:

$$T^2_{JSD,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}\right)' \mathbf{S}_D^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}\right). \tag{10}$$

The improved James-Stein estimator $\bar{\mathbf{x}}^{JS}$ in Equation (9) can be improved further by replacing the covariance matrix $\boldsymbol{\Sigma}$ with SDCM, $\mathbf{S}_D$ in Equation (4), such that:

$$\bar{\mathbf{x}}_D^{JS} = \left(1 - \frac{p-2}{n\left(\bar{\mathbf{x}} - \mathbf{v}\right)^T \mathbf{S}_D^{-1}\left(\bar{\mathbf{x}} - \mathbf{v}\right)}\right)^+ \left(\bar{\mathbf{x}} - \mathbf{v}\right) + \mathbf{v} \tag{11}$$

The $T^2$ control chart based on James-Stein estimator in Equation (10) is enhanced using the new James Stein estimator $\bar{\mathbf{x}}_D^{JS}$ in Equation (11) as follows:

$$\tilde{T}^2_{JSD,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}_D^{JS}\right)' \mathbf{S}_D^{-1} \left(\mathbf{x}_i - \bar{\mathbf{x}}_D^{JS}\right). \tag{12}$$

Because the distribution of the proposed chart is still unknown, its control limits are calculated using KDE.

## 3. CONTROL LIMIT OF THE PROPOSED CHART BASED ON KDE

The kernel density estimation (KDE) method is a non-parametric method to estimate the probability density function of a random variable. This method was first introduced by Rosenblatt (1956) and Parzen (1962)—the so-called Rosenblatt-Parzen kernel density estimator. Chou, Mason and Young (2001) proposed KDE to estimate the distribution of $T^2$ statistic. Using the same procedure, $\tilde{T}^2_{JSD,i}$ obtained under in-control condition, can be estimated by KDE. The empirical distribution of $\tilde{T}^2_{JSD,i}$ statistic can be calculated using the following kernel function:

$$\hat{f}_h(\tilde{T}^2_{JSD}) = \frac{1}{n}\sum_{i=1}^{n} K\left[\frac{\left(\tilde{T}^2_{JSD} - \tilde{T}^2_{JSD,i}\right)}{h}\right], \tag{13}$$

where $K$ and $h$ define kernel function and smoothing parameter, respectively. The most used kernel is Gaussian kernel; therefore, in this paper, it is used in the analysis. Furthermore, the cumulative distribution function of $\hat{f}_h(\tilde{T}^2_{JSD})$ can be written as:

$$\hat{F}_h(t) = \int_0^t \hat{f}_h(\tilde{T}^2_{JSD}) d\tilde{T}^2_{JSD} \tag{14}$$

The control limit of $\tilde{T}^2_{JSD,i}$ based on KDE could be estimated by taking the percentile of kernel distribution. Hence, the control limit of $T^2$ based on KDE, equal to $\left[100(1-\alpha)\right]$-th percentile of $\tilde{T}^2_{JSD}$ distribution, can be calculated as follows:

$$CL_{KDE} = \hat{F}_h(t)^{-1}(1-\alpha). \tag{15}$$

The cumulative distribution function $\hat{F}_h(t)$ in Equation (14) can be calculated using tables of integral in the closed form distribution. However, the control limit might be inefficient to be calculated, if the distribution is not closed form. To overcome such problem, the kernel control limit is solved by employing trapezoidal rule to calculate the integral. The trapezoidal rule is one of the numerical integration methods to approximate definite value of the integral equation.

## 4. METHODOLOGY

### *4.1 IDS based $T^2$ James-Stein and SDCM Control Chart*

In this section, the procedures of the proposed IDS based on $T^2$ James-Stein and SDCM control chart are described. The algorithm for the proposed IDS, using KDE control limit, can be divided into two phases, as follows:

**Phase I: Building Normal Profile**

The mean vector, the covariance matrix and KDE control limit are calculated from the normal profile of the dataset in this phase. The estimated values are then used in the next phase to monitor the new connection. The procedures for building the normal profile phase are defined as follows:

**Step 1** Form matrix $\mathbf{x}_{normal}$, which is the normal connection data.

**Step 2** Calculate vector $\bar{\mathbf{x}}_{normal}$, which is the mean of each column of the normal connection data $\mathbf{x}_{normal}$.

**Step 3** Calculate the matrix of $\mathbf{S}_{DN}$ as in Equation (4), which is the estimated covariance matrix of the normal connection data $\mathbf{x}_{normal}$.

**Step 4** Calculate vector $\bar{\mathbf{x}}^{JS}_{normal}$, which is the estimated mean vector from James-Stein estimator of normal connection data $\mathbf{x}_{normal}$ using Equation (11).

**Step 5** Calculate statistics $\tilde{T}^2_{JSDN,i}$, as in Equation (12), using the normal connection data $\mathbf{x}_{normal}$.

**Step 6** Determine $\alpha$ and calculate the KDE control limit $CL_{KDE}$, as in Equation (15).

**Phase II: Detection**

The estimated value of $\bar{\mathbf{x}}^{JS}_{normal}$, $\mathbf{S}_{DN}$ and $CL_{KDE}$ from Phase I are then used in this phase. The following steps explain the procedures of the detection phase:

**Step 1** Form matrix $\mathbf{x}_{test}$, which is the new connection data.

**Step 2** Calculate statistics $T^2_{JSDT,i}$ from new connection data $\mathbf{x}_{test}$ as follows:

$$T^2_{JSDT,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}_{normal}\right)' \mathbf{S}^{-1}_{DN} \left(\mathbf{x}_i - \bar{\mathbf{x}}^{JS}_{normal}\right),$$

where $\bar{\mathbf{x}}^{JS}_{normal}$ and $\mathbf{S}_{DN}$ are taken from the normal connection data in phase I.

**Step 3** If $T^2_{JSDT,i} > CL_{KDE}$, then the connection is an intrusion and for $T^2_{JSDT,i} < CL_{KDE}$, the connection is normal.

### *4.2 Performance Evaluation*

The NSL-KDD is the dataset that was used in this study. This dataset was first proposed by Tavallaee et al. (2009), as a solution for outdated KDD-99 dataset (Stolfo, 1999), which has been accessible for more than 15 years. The NSL-KDD dataset consists of 41 variables, with 34 quantitative and 7 qualitative variables. Nevertheless, this study only uses 32 quantitative variables, because the value of the rest of the quantitative variables is equal to zero.

In this study, the NSL-KDD dataset is monitored by using Hotelling's $T^2$ based on James-Stein and SDCM with KDE control limit. The performance of the proposed control chart will be compared with the conventional Hotelling's $T^2$ chart and $T^2$ based on SDCM chart, using various control limits, as stated in Ahsan et al. (2018b). Those various control limits include $F$ distribution control limit according to Equation (2); Sullivan and Woodall control limit approach is based on Equation (5); Mason and Young control limit approach is according to Equation (6); and Chi-square control limit is based on Equation (7). In addition, the performance of each IDS approach is evaluated using a confusion matrix as shown in Table 1 (Ahsan, Mashuri, & Khusna, 2018). The occurrence of false positive (FP) in the network causes a false alarm that can disturb the system. Meanwhile, false negative (FN), taking place in the network, will allow an intrusion in the system.

**Table 1**: Intrusion detection confusion matrix.

| | Prediction | |
|---|---|---|
| | Intrusion | Normal |
| Intrusion | True Positive (TP) | False Negative (FN) |
| Normal | False Positive (FP) | True Negative (TN) |

The level of accuracy used is the hit rate that can be calculated as follows:

$$\text{Hit Rate} = \frac{TP + TN}{TP + TN + FP + FN}$$

Based on the type of error, the level of error in intrusion detection can be divided into two types, namely FP rate and FN rate. The FP and FN rate formulas are calculated as follows:

$$FP \text{ Rate} = \frac{FP}{TN + FP}$$

$$FN \text{ Rate} = \frac{FN}{TP + FN}.$$

## 5. RESULTS AND DISCUSSIONS

The performance comparisons of the proposed IDS, with the other approaches, are presented and discussed in this section. The performance of the proposed IDS (JS-$SDCM_{KDE}$) is compared with the conventional Hotelling's $T^2$ ($T^2$) and Hotelling's $T^2$ based on SDCM, using several control limit approaches, as has been stated in Ahsan et al. (2018b). Those control limits are $F$ distribution control limit ($SDCM_F$), Sullivan and Woodall approach ($SDCM_{SW}$), Mason and Young approach ($SDCM_{MY}$) and chi-square control limit ($SDCM_{CH}$).

### *5.1 Results*

The performance of the proposed chart, using KDE control limit, is compared with the other control chart methods, such as conventional Hotelling's $T^2$ control chart and Hotelling's $T^2$ control chart based on SDCM with various control limits. Table 2 reports the performance comparison between the proposed method and the other control charts for the training dataset. The values of hit rate from Table 4 are presented in a single graphic to simplify the interpretation process. According to Figure 1a, it can be observed that the proposed chart with KDE control limit has the highest hit rate compared to the other approaches. The proposed chart also has the better result in term of FN rate, which is depicted in Figure 2a. However, the proposed chart with KDE control limit has a similar value of FP rate, with Hotelling's $T^2$ control chart based on SDCM with Sullivan and Woodall ($SDCM_{SW}$) control limit. Thus, the proposed chart with KDE control limit has better accuracy to detect anomaly in the network than the other control charts' approach, for the training dataset based on hit rate and FN rate criteria.

**Table 2**: Performance of various IDS for training data.

| IDS | Hit Rate | FN | FP | FN Rate | FP Rate |
|---|---|---|---|---|---|
| $T^2$ | 0.91330 | 5428 | 5494 | 0.0806 | 0.0937 |
| $SDCM_F$ | 0.91338 | 5417 | 5495 | 0.0804 | 0.0937 |
| $SDCM_{SW}$ | 0.91705 | 4280 | 6170 | 0.0636 | 0.1052 |
| $SDCM_{MY}$ | 0.91331 | 5429 | 5492 | 0.0806 | 0.0937 |
| $SDCM_{CH}$ | 0.91332 | 5427 | 5492 | 0.0806 | 0.0937 |
| JS-$SDCM_{KDE}$ | 0.91751 | 4115 | 6277 | 0.0611 | 0.1071 |

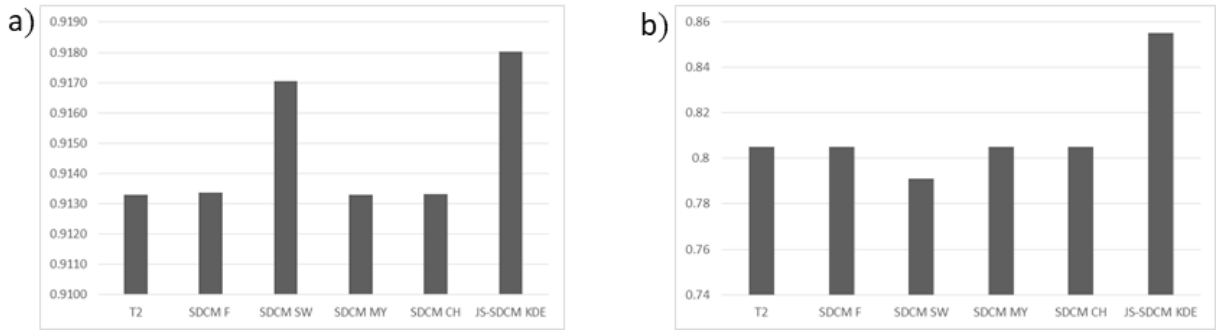


**Figure 1**: Hit rate comparison of various control charts for: a) training dataset b) testing dataset.

Table 3 exhibits the performance comparison between the proposed chart and the other control chart methods for the testing dataset. Similar to the previous result, the performance of the proposed chart with KDE control limit is better than the other approaches based on the hit rate criteria, as shown in Figure 1b. Although the proposed chart with KDE control limit has the higher FN rate compared to the other approaches, its performance outperforms the other methods based on the FP rate criteria, as depicted in Figure 2b. It is noteworthy that the difference between the FN rate of the proposed chart with KDE and the other approaches is not significant.

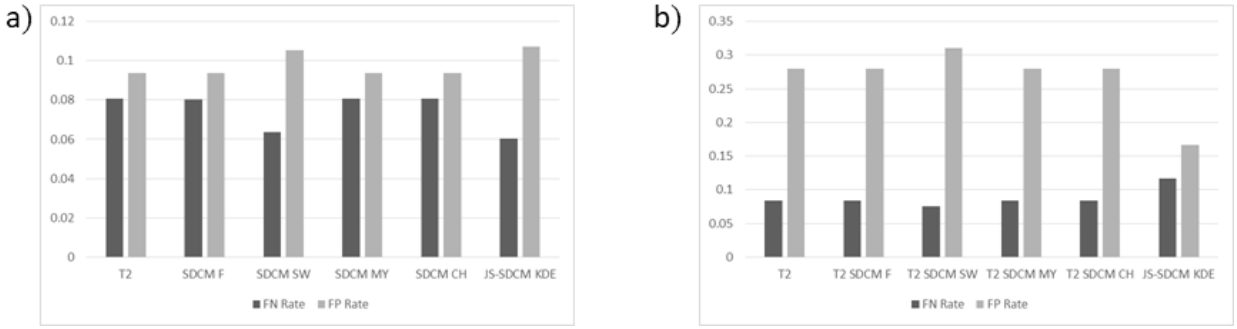


**Figure 2**: FN and FP rate comparison of various control charts for: a) training dataset b) testing dataset.

**Table 3**: Performance of various IDS for testing data.

| IDS | Hit Rate | FN | FP | FN Rate | FP Rate |
|---|---|---|---|---|---|
| $T^2$ | 0.8049 | 814 | 3584 | 0.0838 | 0.2793 |
| $SDCM_F$ | 0.8049 | 814 | 3585 | 0.0838 | 0.2794 |
| $SDCM_{SW}$ | 0.7911 | 731 | 3978 | 0.0753 | 0.3100 |
| $SDCM_{MY}$ | 0.8049 | 814 | 3584 | 0.0838 | 0.2793 |
| $SDCM_{CH}$ | 0.8049 | 814 | 3584 | 0.0838 | 0.2793 |
| $JS\text{-}SDCM_{KDE}$ | 0.8554 | 1127 | 2134 | 0.1160 | 0.1663 |

### *5.2 Discussions*

Based on the performance evaluation from the previous section, it could be shown that the $T^2$ based on James-Stein and SDCM with KDE control limit, has the highest hit rate for the training and testing dataset. For the training dataset, the misdetection happens because of the high value of FP rate produced by the proposed chart. The high value of FP rate from training dataset happens as a result of the control chart oversensitivity to detect an attack, while the attack is not actually happened in the network. However, the proposed IDS is superior to the other approaches in terms of the low value of FN rate criteria. Consequently, IDS constructed by this approach will successfully detect an actual attack in the network, but it has a higher false alarm.

For the testing dataset, similar to the training dataset, the misdetection happens because of the high value of FP rate. These findings indicate that the IDS constructed by using $T^2$, based on James-Stein and SDCM, will produce more false alarm. However, compared to the other approaches, the proposed chart has the lowest FP rate. This

proposed method also will not let the attack occur in the network without any warning shown by the low level of FN rate. Thus, by considering the performance from the training and testing dataset, IDS constructed by the proposed chart with KDE, outperforms the other approaches in terms of detecting the actual attacks in the network.

## 6. CONCLUSION

In this paper, the multivariate Hotelling's $T^2$ control chart is improved by James-Stein and SDCM estimators, while its control limit is calculated using KDE method before it is applied into IDS. The performance of proposed IDS is evaluated and compared with the other control limits by hit rate, FN rate and FP rate criteria. Furthermore, the performance of proposed IDS is also compared with some existing control charts. The performance evaluation results reveal that the proposed IDS using $T^2$, based on James-Stein and SDCM with KDE control limit, outperforms the other approaches in both in training and testing dataset. The proposed method is effective to be applied in IDS, based on its ability to detect anomaly in the network, confirmed by the low value of FN rate. The multiclass detection for each type of attack with an incremental algorithm can be considered as future research.

## 7. ACKNOWLEDGEMENT

This work was supported by Research, Technology, and Higher Education Ministry, the Republic of Indonesia through PMDSU scheme under Grant 128/SP2H//PTNBH/DRPM/2018.

## 8. REFERENCES

Abu-Shawiesh, M. O. A., Kibria, G., and George, F. (2014). A robust bivariate control chart alternative to the Hotelling's T2 control chart. *Quality and Reliability Engineering International*, 30(1): 25–35.

Ahsan, M., Mashuri, M., and Khusna, H. (2017). Evaluation of Laney p' chart performance. *International Journal of Applied Engineering Research*, 12(24): 14208–14217.

Ahsan, M., Mashuri, M., and Khusna, H. (2018). Intrusion detection system using bootstrap resampling approach of T2 control chart based on successive difference covariance matrix. *Journal of Theoretical and Applied Information Technology*, 96(8): 2128–2138.

Ahsan, M., Mashuri, M., Kuswanto, H., Prastyo, D. D., and Khusna, H. (2018a). Multivariate control chart based on PCA mix for variable and attribute quality characteristics. *Production & Manufacturing Research*, 6(1): 364–384. https://doi.org/10.1080/21693277.2018.1517055

Ahsan, M., Mashuri, M., Kuswanto, H., Prastyo, D. D., and Khusna, H. (2018b). T2 control chart based on successive difference covariance matrix for intrusion detection system. In *Journal of Physics: Conference Series*, 1028: 12220.

Alfaro, J. L., and Ortega, J. F. (2009). A comparison of robust alternatives to Hotelling's T2 control chart. *Journal of Applied Statistics*, 36(12): 1385–1396. https://doi.org/10.1080/02664760902810813

Alkindi, Mashuri, M., and Prastyo, D. D. (2016). T2 hotelling fuzzy and W2 control chart with application to wheat flour production process. In *AIP Conference Proceedings*, 1746. https://doi.org/10.1063/1.4953977

Arkat, J., Niaki, S. T. A., and Abbasi, B. (2007). Artificial neural networks in applying MCUSUM residuals charts for AR(1) processes. *Applied Mathematics and Computation*, 189(2): 1889–1901.

https://doi.org/10.1016/j.amc.2006.12.08

Bace, R., and Mell, P. (2001). NIST special publication on intrusion detection systems. *Special Publication (NIST SP) - 800-31*. https://doi.org/10.1016/S1361-3723(01)00614-5

Bersimis, S., Sgora, A., and Psarakis, S. (2016). The application of multivariate statistical process monitoring in non-industrial processes. *Quality Technology and Quantitative Management*, 3703(September): 1–24. https://doi.org/10.1080/16843703.2016.1226711

Catania, C. A., and Garino, C. G. (2012). Automatic network intrusion detection: Current techniques and open issues. *Computers & Electrical Engineering*, 38(5): 1062–1072. https://doi.org/10.1016/j.compeleceng.2012.05.013

Chou, Y.-M., Mason, R., and Young, J. (2001). The control chart for individual observations from a multivariate non-normal distribution. *Communications in Statistics: Theory & Methods*, 30(8-9): 1937-1949. https://doi.org/10.1081/STA-100105706

Chou, Y., Mason, R. L., and Young, J. C. (1999). Power comparisons for a hotelling's t 2 STATISTIC. *Communications in Statistics - Simulation and Computation*, 28(4): 1031–1050. https://doi.org/10.1080/03610919908813591

Hawkins, D. M., and Merriam, D. F. (1974). Zonation of multivariate sequences of digitized geologic data. *Journal of the International Association for Mathematical Geology*, 6(3): 263–269. https://doi.org/10.1007/BF02082892

Holmes, D. S., and Mergen, A. E. (1993). Improving the performance of the T2 control chart. *Quality Engineering*, 5(4): 619–625. https://doi.org/10.1080/08982119308919004

Hotelling, H. (1974). Multivariate quality control. In *Techniques of Statistical Analysis*. New York: McGraw-Hill.

Issam, B. K., and Mohamed, L. (2008). Support vector regression based residual MCUSUM control chart for autocorrelated process. *Applied Mathematics and Computation*, 201(1–2): 565–574. https://doi.org/10.1016/j.amc.2007.12.059

James, W., and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 361–379.

Khusna, H., Mashuri, M., Ahsan, M., Suhartono, S., and Prastyo, D. D. (2018a). Bootstrap based maximum multivariate CUSUM control chart. *Quality Technology & Quantitative Management*. https://doi.org/10.1080/16843703.2018.1535765

Khusna, H., Mashuri, M., Suhartono, Prastyo, D. D., and Ahsan, M. (2018b). Multioutput least square SVR based multivariate EWMA control chart. In *Journal of Physics: Conference Series*, 1028(1): 12221. Retrieved from http://stacks.iop.org/1742-6596/1028/i=1/a=012221

Lehmann, E. L., and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Society for Industrial and Applied Mathematics. Retrieved from http://epubs.siam.org/doi/book/10.1137/1.9780898718461

Montgomery, D. (2009). *Introduction to Statistical Quality Control*. New York: John Wiley & Sons Inc. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Murray Rosenblatt. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27: 832–837. https://doi.org/10.1214/aoms/1177728190

Page, E. S. (1961). Cumulative Sum Charts. *Technometrics*, 3(1): 1–9. https://doi.org/10.1080/00401706.1961.10489922

Park, Y. (2005). *A Statistical Process Control Approach for Network Intrusion Detection*. Georgia Insitute of Technology.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3): 1065–1076. https://doi.org/10.1214/aoms/1177704472

Phaladiganon, P., Kim, S. B., Chen, V. C. P., Baek, J.-G., and Park, S.-K. (2011). Bootstrap-based T2 multivariate control charts. *Communications in Statistics - Simulation and Computation*, 40(5): 645–662. https://doi.org/10.1080/03610918.2010.549989

Phaladiganon, P., Kim, S. B., Chen, V. C. P., and Jiang, W. (2013). Principal component analysis-based control charts for multivariate nonnormal distributions. *Expert Systems with Applications*, 40(8): 3044–3054. https://doi.org/10.1016/j.eswa.2012.12.020

Pirhooshyaran, M., and Niaki, S. T. A. (2015). A double-max MEWMA scheme for simultaneous monitoring and fault isolation of multivariate multistage auto-correlated processes based on novel reduced-dimension statistics. *Journal of Process Control*, 29: 11–22. https://doi.org/10.1016/j.jprocont.2015.03.008

Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, 1(3): 239–250. https://doi.org/10.1080/00401706.1959.10489860

Shenfield, A., Day, D., and Ayesh, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *ICT Express*, 4(2): 95-99.

Shewhart, W. A. (1924). Some applications of statistical methods to the analysis of physical and engineering data. *Bell Labs Technical Journal*, 3(1): 43–87.

Stein, C. (1956). *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution*. United States: Stanford University Stanford.

Stolfo, S. J. (1999). KDD cup 1999 dataset. *UCI KDD Repository. Http://Kdd.Ics.Uci.Edu*, 0.

Sullivan, J. H., and Woodall, W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28(4): 398–408.

Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*. https://doi.org/10.1109/CISDA.2009.5356528

Vargas, N. J. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35(4): 367–376.

Wang, H., Huwang, L., and Yu, J. H. (2015). Multivariate control charts based on the James–Stein estimator. *European Journal of Operational Research*, 246(1): 119–127.

Wibawati, Mashuri, M., Purhadi, and Irhamah. (2016). Fuzzy multinomial control chart and its application. In *AIP Conference Proceedings*, 1718(1): 110004. https://doi.org/10.1063/1.4943351

Wibawati, Mashuri, M., Purhadi, Irhamah, and Ahsan, M. (2018). Performance fuzzy multinomial control chart. In *Journal of Physics: Conference Series*, 1028(1): 12120. Retrieved from http://stacks.iop.org/1742-6596/1028/i=1/a=012120

Williams, J. D., Woodall, W. H., Birch, J. B., and Sullivan, J. O. E. H. (2006). On the distribution of Hotelling's T2 statistic based on the successive differences covariance matrix estimator. *Journal of Quality Technology*, 38: 217–229.

Wororomi, J. K., Mashuri, M., Irhamah, and Arifin, A. Z. (2014). On monitoring shift in the mean processes with vector autoregressive residual control charts of individual observation. *Applied Mathematical Sciences*, 8: 3491–3499. https://doi.org/10.12988/ams.2014.44298